# Bimodal sound source tracking applied to road traffic monitoring

THÈSE N$^{O}$ 5618 (2013)

PRÉSENTÉE LE 15 FÉVRIER 2013

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
LABORATOIRE D'ÉLECTROMAGNÉTISME ET ACOUSTIQUE
PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Patrick MARMAROLI

acceptée sur proposition du jury:

Prof. P. Vandergheynst, président du jury
Prof. J. R. Mosig, Dr H. Lissek, directeurs de thèse
Prof. J. Antoni, rapporteur
Dr G. Dutilleux, rapporteur
Dr J.-M. Vesin, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2013

# Remerciements

Le dénouement de cette aventure n'aurait sans doute pas été aussi heureux sans l'aide, la confiance et la bienveillance de nombreuses personnes que j'aimerais remercier en préambule à ce mémoire.

Tout d'abord mes deux co-directeurs de thèse: Prof. Juan Mosig, directeur du LEMA, et Dr. Hervé Lissek, chef du Groupe Acoustique. Merci Juan d'avoir accepté de co-diriger ce travail. Merci pour l'ambiance que tu apportes à nos sorties, réunions et autres soirées événementielles. Ta bonne humeur contagieuse maintient le laboratoire dans une cohésion et un esprit d'équipe très appréciable au quotidien. Quand à toi, Hervé, merci de m'avoir accueilli au sein de ton équipe, et pour ainsi dire, de m'avoir ouvert les portes de la recherche. Tes qualités d'encadrant, tant au niveau humain que scientifique, m'ont véritablement tiré vers le haut et chacun des conseils et autres choix stratégiques que tu as fait à mon égard se sont avérés payants. Tout en m'accordant les libertés d'action, de réflexion et d'expérimentations nécessaires, tu as su t'assurer que je ne m'égare jamais trop longtemps. En bref, merci pour ta confiance.

Je tiens à adresser des remerciements tout particulier au Dr. Xavier Falourd, avec qui j'ai eu le plaisir de partager mon bureau pendant ces quatre années. Merci Xavier pour tout ce temps passé à m'enseigner, me conseiller, m'écouter, me relire, me réorienter. Nous avons collaboré sur des projets aussi variés que passionnants, et tu n'as jamais ménagé ton énergie pour me faire progresser dans tous les domaines que ce soit (programmation, mesure expérimentale, analyse et présentation des résultats etc.), merci infiniment.

Je remercie chaleureusement les membres du jury: Prof. Pierre Vandergheynst, Prof. Jérôme Antoni, Dr. Jean-Marc Vesin et Dr. Guillaume Dutilleux pour le temps qu'ils ont consacré à l'expertise de ce mémoire et l'intérêt porté à mon travail. Merci pour vos encouragements, conseils, corrections et critiques constructives qui auront permis une amélioration notable de la qualité scientifique (et grammaticale) de ce document.

J'exprime toute ma gratitude envers les anciens et actuels membres du LEMA que j'ai eu la chance de côtoyer et en particulier: Prof. Mario Rossi, Syddharta Berns,

Dr. Romain Boulandet, Cédric Monchâtre, Etienne Rivet, Anne-Sophie Moreau, Dr. François Aballéa, Philippe Martin, Dr. Pierre-Jean René, Lukas Rohr, Patrick Roe, Dr. Michaël Mattes, Dr. Benjamin Fuchs et Dr. Roberto Torres. Merci d'avoir partagé votre expérience et d'avoir répondu présent pour m'aider à surmonter les difficultés techniques et scientifiques du quotidien. Une pensée affective et pleine d'encouragement pour nos doctorants fraîchement arrivés au labo: Gilles Courtois et Hussein Seyyed. Merci à David Desscan, et avant lui, Sébastien Halouze-Lamy, pour leurs multiples dépannages informatiques. Un énorme merci à notre secrétaire Eulalia Durussel pour son sourire, sa disponibilité et sa compétence tout simplement irremplaçables. Merci aux étudiants stagiaires que j'ai eu le plaisir de co-encadrer entre 2009 et 2012: Yvonne Blaszczyk, Andreas Weishaupt, Stephan Hesse, Samuel Egli, Lukas Doméjean, Dorian Cazau, Lionel Velut, Vincent Kuenlin. Je garde un souvenir impérissable de ces collaborations, véritables sources de ma progression.

J'aimerais remercier chaleureusement les Prof. Dacorogna (EPFL), Prof. Vincent Martin (Institut Jean le Rond d'Alembert, Saint Cyr), Dr. Mikael Carmona (CEA-LETI, Grenoble), Dr. Alain Dufaux (EPFL) et Dr. Jean-Marc Odobez (IDIAP, Martigny) pour leur aide précieuse sur divers points mathématiques précis et le partage de leur expertise vis-à-vis de mes travaux d'une manière générale. Merci également aux membres de l'atelier d'electromécanique de l'EPFL, et particulièrement Roland Dupuis et Jean-Paul Brugger, pour leurs conseils experts et leurs réalisations techniques de grande qualité sans lesquelles bon nombre de mesures expérimentales ayant parsemées ce doctorat n'auraient pu être effectuées.

Enfin, mes remerciements les plus sincères et profonds à ma famille, et tout particulièrement mes parents, Claude et Monique. Mon accès aux études est le fruit de vos longs efforts. Merci pour votre soutien, vos conseils et l'ensemble de votre éducation qui m'ont mené tout droit à la réussite de ce doctorat. Merci infiniment à mon épouse, et mère de mes deux (bientôt trois) beaux enfants, Sofia. Ton sourire rayonnant et ton réconfort sans faille m'ont permis de surmonter les moments les plus difficiles de cette aventure. Ta contribution à l'obtention de ce diplôme est difficilement quantifiable tant elle est importante. Merci pour ta patience, ta compréhension, ton amour. Quand à l'énergie nécessaire pour effectuer ce travail, je l'ai sans nul doute puisée dans les éclats de rires de mes deux petits princes, Malik et Naïm. Merci de m'apporter tant chaque jour. La rédaction de ce mémoire vous aura volé beaucoup du temps de ma présence qu'il me tarde de rattraper. Je vous dédie ce document.

*Lausanne, le 14 janvier 2013* Patrick

# Abstract

The constant increase of road traffic requires closer and closer road network monitoring. The awareness of traffic characteristics in real time as well as its historical trends, facilitates decision-making for flow regulation, triggering relief operations, ensuring the motorists' safety and contribute to optimize transport infrastructures.

Today, the heterogeneity of the available data makes their processing complex and expensive (multiple sensors with different technologies, placed in different locations, with their own data format, unsynchronized, etc.). This leads metrologists to develop "smarter" monitoring devices, *i.e.* capable of providing all the necessary data synchronized from a single measurement point, with no impact on the flow road itself and ideally without complex installation.

This work contributes to achieve such an objective through the development of a passive, compact, non-intrusive, acoustic-based system composed of a microphone array with a few number of elements placed on the roadside. The proposed signal processing techniques enable vehicle detection, the estimation of their speed as well as the estimation of their wheelbase length as they pass by. Sound sources emitted by tyre/road interactions are localized using generalized cross-correlation functions between sensor pairs. These successive correlation measurements are filtered using a sequential Monte Carlo method (particle filter) enabling, on one hand, the simultaneous tracking of multiple vehicles (that follow or pass each other) and on the other hand, a discrimination between useful sound sources and interfering noises.

This document focuses on two-axle road vehicles only. The two tyre/road interactions (front and rear) observed by a microphone array on the roadside are modeled as two stochastic, zero-mean and uncorrelated processes, spatially disjoint by the wheelbase length. This *bimodal* sound source model defines a specific particle filter, called *bimodal particle filter*, which is presented here. Compared to the classical (unimodal) particle filter, a better robustness for speed estimation is achieved especially in cases of harsh observation. Moreover the proposed algorithm enables the wheelbase length estimation through purely passive acoustic measurement. An innovative microphone array design methodology, based on a mathematical expression of the observation and the tracking

methodology itself is also presented.

The developed algorithms are validated and assessed through *in-situ* measurements. Estimates provided by the acoustical signal processing are compared with standard radar measurements and confronted to video monitoring images. Although presented in a purely road-related applied context, we feel that the developed methodologies can be, at least partly, applied to rail, aerial, underwater or industrial metrology.

**Key-words**: road traffic monitoring, sound source tracking, particle filtering, generalized cross-correlation functions, microphone array processing.

# Résumé

L'accroissement constant du trafic routier impose une surveillance de plus en plus étroite des voies de circulation. La connaissance en temps réel des caractéristiques du trafic, ainsi que leurs tendances sur le long terme, facilitent la prise de décision pour la régulation du flux, le déclenchement des opérations de secours, la sécurité des usagers et contribuent à l'amélioration des infrastructures du transport.

Aujourd'hui, l'hétérogénéité des données à disposition rend leur traitement complexe et coûteux (multiples capteurs aux technologies différentes, placés en des lieux différents, ayant leur propre format de données, désynchronisés etc.). Ce constat pousse les métrologues à développer des systèmes de surveillance plus "intelligents", c'est-à-dire, capables de fournir toutes les données nécessaires, synchronisées, provenant d'un même point de mesure, sans impact sur le flux routier lui-même et idéalement sans installation complexe.

Ce travail de thèse s'inscrit comme une contribution aux développements de tels capteurs via l'élaboration d'une station de mesure acoustique passive, compacte et non-intrusive, composée d'un réseau à faible nombre de microphones plaçé en bord de voie. Les techniques de traitement proposées autorisent la détection, l'estimation en vitesse et l'estimation en empattement des véhicules au passage. Les sources de sons émises par les intéractions pneus-chaussée sont localisées à l'aide de fonctions d'inter-corrélations généralisées entre paires de capteurs. Ces mesures de corrélations successives sont filtrées par une méthode séquentielle de Monte Carlo (filtrage particulaire) permettant d'une part, le suivi simultané de plusieurs véhicules (qui se suivent ou se croisent) et d'autre part, une discrimination entre sources sonores d'intérêts et sources sonores parasites.

Le seul cas des véhicules à deux essieux est traité dans ce document. Les deux interactions pneus-chaussée observées (avant et arrière) sont modélisées par deux processus aléatoires centrés et décorrélés, séparés d'une distance fixe au cours du temps (l'empattement). Ce modèle *bimodal* de source sonore définit un filtre particulaire dédié, baptisé filtrage particulaire bimodal, que nous présentons ici. Par rapport au filtre particulaire classique (unimodal), nous obtenons d'une part, une meilleure robustesse dans l'estimation en vitesse pour les conditions d'observations difficiles et d'autre part, une estimation automatique de l'empattement des véhicules au passage. Le filtrage proposé associé à une

expression mathématique de l'observation constituent également la base d'une stratégie innovante de dimensionnement du réseau microphonique.

Les algorithmes développés sont validés et qualifiés par des mesures *in-situ*. Les estimations fournies par le traitement des signaux acoustiques sont comparées aux mesures radar normalisées et confrontées aux images de surveillance vidéo. Bien que présentées dans un cadre strictement routier, nous pensons que les méthodologies développées dans ce document peuvent en partie s'appliquer à la métrologie ferroviaire, aérienne, sous-marine et industrielle.

**Mots-clés** : surveillance du trafic routier, suivi de sources sonores, filtrage particulaire, fonctions d'inter-corrélations généralisées, traitement d'antenne microphoniques.

# Zusammenfassung[1]

Die Verkehrszunahme erfordert eine immer engere Überwachung des Strassennetzes. Die Kenntnis in Echtzeit der Charakteristiken des Verkehrsflusses und dessen langfristige Entwicklung sind eine Entscheidungshilfe bei der Verkehrsregelung, der Auslösung von Notfallmassnahmen und der Sicherheit der Verkehrsteilnehmer und sie ermöglichen eine Verbesserung der Transportinfrastrukturen.

Die Heterogenität der verfügbaren Daten erschwert heutzutage ihre Auswertung und macht sie unnötig komplex und teuer (Verwendung von multiplen Sensoren und Technologien, die an unterschiedlichen Orten platziert werden, asynchron und mit proprietären Datenformaten arbeiten, etc.). Diese Feststellung drängt die Metrologen "intelligentere" Überwachungssysteme zu entwickeln, die alle erforderlichen Daten synchron, von einem einzelnen Messpunkt ausgehend, ohne Einwirkung auf den Verkehrsfluss und idealerweise ohne komplizierte Installation liefern.

Diese Arbeit soll einen Beitrag zu der Entwicklung solcher Sensoren leisten mit der Ausarbeitung einer passiven, nicht intrusiven, kompakten akustischen Messstation. Sie besteht aus einem kleinen Netzwerk von Mikrofonen und wird am Rand der Fahrbahn aufgestellt. Die zugehörige Signalverarbeitung ermöglicht die automatische Erkennung und eine Schätzung der Geschwindigkeit und des Achsabstandes von vorbeifahrenden Fahrzeugen. Die Schallquellen, die durch die Wechselwirkung zwischen Belag und Reifen erzeugt werden, werden mittels paarweiser generalisierter Interkorrelationsfunktionen zwischen den Sensoren lokalisiert. Diese aufeinanderfolgenden Korrelationsmessungen werden durch eine sequenzielle Monte-Carlo Methode gefiltert, die einerseits die gleichzeitige Verfolgung von mehreren (sich folgenden oder kreuzenden) Fahrzeugen und andererseits die Diskriminierung zwischen nützlichen und unerwünschten Schallquellen ermöglicht.

Einzig der Fall von zweiachsigen Fahrzeugen wird in diesem Dokument behandelt. Die beiden beobachteten Wechselwirkungen zwischen Belag und Reifen (Vorder- und Hinterachse) werden durch zwei zentrierte und unkorrelierte stochastische Prozesse modellisiert, die räumlich durch eine fixe Distanz getrennt sind (Achsabstand). Dieses *bimodale* Modell einer Schallquelle definiert einen dedizierten Partikel-Filter, den wir hier vorstellen. Im

[1] *The translation from french to german was kindly performed by Lukas Rohr (LEMA)*

Gegensatz zum klassischen (unimodalen) Ansatz eines Partikel-Filters, erhalten wir einerseits eine grössere Robustheit bei der Schätzung der Geschwindigkeit unter schwierigen Messbedingungen und andererseits eine automatische Schätzung des Achsabstandes von vorbeifahrenden Fahrzeugen. Die Verknöpfung des vorgeschlagenen Filters mit einer mathematischen Näherung der Messung dient ebenfalls als Grundlage zu einer innovativen Dimensionierungsstrategie für das Mikrofonnetzwerk.

Alle entwickelten Algorithmen werden durch Feldversuche bestätigt und charakterisiert. Die Schätzungen aus der akustischen Signalverarbeitung werden mit normalisierten Radarmessungen verglichen und Videoüberwachungsbildern gegenübergestellt. Auch wenn der Rahmen der hier aufgeführten Methodologien sich nur auf den Strassenkontext beschränkt, denken wir, dass sie teilweise auch auf andere Anwendungsfelder wie Schiene, Flugverkehr, Industrie und Unterwassermetrologie angewandt werden können.

**Schlüsselwörter**: Strassenverkehrsüberwachung, Verfolgung von Schallquellen, Partikelfilter, generalisierte Korrelationsfunktionen, Mikrofonantennen-Verarbeitung.

# Resumen[2]

El crecimiento del tráfico hace necesaria una mayor monitorización de la red de carreteras. El conocimiento en tiempo real de las características del tráfico, así como de sus estadísticas a largo plazo, facilita la toma de decisiones relativas a su regulación, a la activación de operaciones de socorro, a la seguridad de los usuarios y puede contribuir a una mejora de las infraestructuras de transporte.

A día de hoy, la heterogeneidad de los datos disponibles complica y encarece su explotación (empleo de numerosos sensores con tecnologías diversas, en diferentes ubicaciones, no sincronizados, cada uno con su propio formato de datos, etc.). Este hecho obliga a los metrólogos a desarrollar sistemas de monitorización más "inteligentes", es decir, capaces de suministrar todos los datos necesarios, de forma sincronizada, provenientes de un mismo punto de medida, sin interferir con el flujo circulatorio e, idealmente, sin necesidad de una instalación compleja.

Esta Tesis constituye una contribución de cara al desarrollo de tales sensores mediante la elaboración una estación de medida acústica pasiva, compacta y no intrusiva, compuesta por un pequeño array de micrófonos situado en el borde de la calzada. El tratamiento de señal asociado posibilita la detección, la estimación de la velocidad y la estimación de distancia entre ejes de los vehículos en tránsito. Las fuentes del sonido emitido por la interacción entre neumáticos y asfalto son localizadas con la ayuda de funciones de correlación cruzada generalizadas entre sensores. Estas medidas sucesivas de correlación son filtradas en base a un método de Monte Carlo, lo que permite, por un lado, el seguimiento simultáneo de varios vehículos (que se siguen o se cruzan) y, por otro lado, la discriminación entre fuentes sonoras significativas y parásitas.

En este documento se trata únicamente el caso de vehículos con dos ejes. Las dos interacciones neumático-asfalto que se observan (ejes delantero y trasero) son modeladas como una pareja de procesos aleatorios centrados e incorrelados, separados por una distancia que es fija en el tiempo (distancia entre ejes). Este modelo *bimodal* de fuente sonora define un filtro de partículas de propósito específico, al que se bautiza como *filtrado de partículas bimodal*, que aquí presentamos. Con respecto al filtrado de partículas clásico

[2] *The translation from french to spanish was kindly performed by Dr. Roberto Torres (LEMA)*

(unimodal), obtenemos, por un lado, una estimación más robusta para la velocidad en condiciones de observación difíciles y, por otro lado, una estimación automática de la distancia entre ejes de los vehículos que circulan. La técnica de filtrado y la aproximación matemática de la observación que lleva asociada constituyen la base de una estrategia innovadora para el dimensionamiento del array de micrófonos.

El conjunto de los algoritmos que aqui se desarrollan son validados y cualificados con medidas in-situ. Las estimaciones proporcionadas por el tratamiento de la señal acústica son comparadas con medidas radar normalizadas y cotejadas con imágenes de monitorización video. Aunque presentadas en un contexto estrictamente automovilístico, pensamos que las metodologías desarrolladas en este documento también pueden aplicarse, en parte, a la metrología ferroviaria, aérea, submarina e industrial.

**Palabras-clave** : monitorización del tráfico rodado, seguimiento de fuentes sonoras, filtrado de partículas, funciones de correlación cruzada generalizadas, tratamiento de antenas a base de micrófonos.

# Sommario[3]

L'aumento costante di traffico stradale richiede un sempre piú accurato monitoraggio della rete stradale. La conoscenza delle caratteristiche del traffico in tempo reale e del suo andamento nel passato facilita i processi decisionali per il controllo del flusso e l'attivazione di interventi di soccorso, assicurando la sicurezza degli automobilisti e contribuendo ad ottimizzare le infrastrutture di trasporto.

Oggi la natura eterogenea dei dati disponibili rende la loro elaborazione complessa e costosa (vari sensori con diverse tecnologie, posti in luoghi diversi, ciascuno con il prorio formato dati, assenza di sincronizzazione, ecc.). Ció ha indotto i metrologi allo sviluppo di sistemi di moniraggio piu "intelligenti", ossia in gradio di fornire tutti i dati necessari sincronizzati da un singolo sito di misura, senza impatto sul flusso stradale stesso ed idealmente senza installazioni complesse.

Questo studio contribuisce a conseguire questo obiettivo tramite lo sviluppo di un sistema acustico passivo, compatto, non invasivo composto da una schiera di microfoni con un basso numero di elementi collocati sul ciglio stradale. Le qui proposte tecniche di analisi dei segnali consentono la rilevazione dei veicoli, la stima della loro velocità e del loro passo interasse mentre passano. Le fonti sonore emesse dall'interazione pneumatico/strada sono localizzate utilizzando funzioni di cross-correlazione tra coppie di sensori. Questa serie di misure di correlazione sono filtrate usando un metodo Monte Carlo sequenziale (filtro a particelle) consentendo, da un lato, di tracciare più veicoli simultaneamente (che si susseguono o si sorpassano) e, dall'altro, una discriminazione tra fonti utili e rumore interferente.

Questo documento tratta solo vecoli stradali a due assi. Le due interazioni pneumatico/strada (anteriore e posteriore) osservate da una schiera di microfoni sul ciglio stradale sono modellizate da due processi stocastici scorrelati ed a media nulla, separati spazialmente d una distanza pari all'interasse. Questa fonte acustica *bimodale* definisce uno specifico filtro a particelle, chamato *filtro a particelle bimodale*, che è qui presentato. Rispetto al filtro a particelle classico (unimodale), una migliore robustezza e velocità di stima sono ottenute, specialmente in caso di osservazioni in ambiente ostile. Inoltre l'algoritmo

[3] *The translation from french to italian was kindly performed by Michele Tamagnone (LEMA)*

proposto consente la stima dell'interasse tramite misure acustiche puramente passive. Vengono inoltre presentate un'innovativa metodologia di progetto di schiere di microfoni, basata su un'espressione matematica dell'osservazione, e la strategia di tracciamento stessa.

Gli algoritmi sviluppati sono stati validati e collaudati tramite misure in-situ. Stime fornite dall'analisi dei segnali acustici sono confrontate con misure radar standard e con immagini di monitoraggio video. Anche se le metodologie sviluppate sono presentate nel contesto applicato di musure su strada, riteniamo che esse possano, almeno in parte, essere applicate alla metrologia indstriale, ferroviaria, aerea e subaquea.

**Parole-chiave** : monitoraggio traffico stradale, tracciamento fonti sonore, filtro a particelle, funzioni di cross-correlazione generalizzate, analisi di schiere di microfoni.

# Contents

# List of Figures

# List of Tables

# List of Symbols and Acronyms

## Mathematical operators

| | |
|---|---|
| $(.)*$ | complexe conjugate operator |
| $\mathbf{T}$ | transpose operator |
| $i$ | imaginary unit: $i = \sqrt{-1}$ |
| $\|.\|$ | absolute value operator ($l^1$-norm) |
| $\|\|.\|\|$ | Euclidean distance operator ($l^2$-norm) |
| $\lfloor . \rfloor$ | floor function |
| $\delta(.)$ | Dirac delta function |
| $\mathbb{E}\{.\}$ | statistical expectation operator |
| $\mathbf{Re}\{.\}$ | real part operator |
| $(.)^T$ | transpose operator |
| $(.)^H$ | transpose hermitian operator |
| $\mathcal{N}(\mu, \sigma)$ | Gaussian density with mean $\mu$ and standard deviation $\sigma$ |
| DFT[.] | discrete Fourier transform operator |
| IDFT[.] | inverse discrete Fourier transform operator |

## Specific notations and parameters

| | |
|---|---|
| $\mathbf{r}_k^s$ | coordinate of the $k^{th}$ source in the Cartesian coordinate system |
| $\mathbf{r}_j^m$ | coordinate of the $j^{th}$ microphone in the Cartesian coordinate system |
| $\mathbf{r}^{(n)}$ | coordinate of the $n^{th}$ potential sound source position |
| $d$ | inter-sensor distance |
| $D$ | distance between the microphone array and the closest point of approach of the vehicle |
| $x_0$ | distance between the front axle and the closest point of approach |
| $B_w$ | bandwidth (Hz) |
| $f_c$ | central frequency (Hz) |
| $c$ | acoustic wave propagation velocity |
| $\delta_{jk}$ | time of flight of an acoustic wave between the receiver at $\mathbf{r}_j^m$ and the source at $\mathbf{r}_k^s$ |

| | |
|---|---|
| $R(\tau)$ | continuous cross-correlation function, $\tau$ denotes the time lag |
| $R^{gcc}(\tau)$ | continuous generalized cross-correlation function, $\tau$ denotes the time lag |
| $\tau_{12}$, $\tau_p$ | TDOA of an acoustic wave between sensors 1 and 2, or those belonging to the $p^{th}$ pair |
| $T, N$ | duration of a recording, $T$ in seconds, $N$ in samples |
| $N$ | may also refer to the sensor number (the context should make clear in which sense $N$ is used) |
| $M$ | number of microphones in the array |
| $P$ | number of sensor pairs in the array |
| $t, n$ | denotes the time index for continuous, respectively discret, signals |
| $f, k$ | denotes the frequency index for continuous, respectively discrete, Fourier transform |
| $y_j(t)$ | analogical signal acquired at position $\mathbf{r}_j^m$, $0 \leq t \leq T$ |
| $N_w$ | number of audio frames in a recording |
| $N_s$ | length of an audio frame, in samples |
| $N_o$ | overlap between two successive audio frames, in samples |
| $y_j[n]$ | value of the $n^{th}$ sample |
| $\mathbf{y}_j^q[n]$ | $q^{th}$ audio frame, $1 \leq q \leq N_w$, $1 \leq n \leq N_s$ |
| $\mathbf{Y}_j$ | discrete Fourier transform of $\mathbf{y}_j$ |
| $\alpha_t$ | state vector at time $t$ |
| $\alpha_t^{(n)}$ | state of the n$^{th}$ particle at time $t$ |
| $w_t$ | weighting vector at time $t$ |
| $w_t^{(n)}$ | weight of the n$^{th}$ particle at time $t$ |
| $\beta$ | observation vector |
| $N_p$ | number of particles |

## List of acronyms

# 1 Introduction

## 1.1 General context of this thesis

According to the 2011 report of the National Institute for Health and Welfare [1] and relayed in the 2012 report of the World Health Organization (WHO) [2], transportation noise is the third environmental stressor having the highest impact on the European people's health, just after air pollution and second-hand smoking. Noise from traffic, rail or aircraft affects a great number of people as it may cause sleep disturbance as well as annoyance, potentially leading to high blood pressure and increase risk of myocardial infarction [3]. The WHO estimates that at least one million healthy life years are lost every year from traffic-related noise in western European countries. Since 2002, European environmental directives have forced cities with more than 100 000 inhabitants to establish acoustic maps of their territory, identify and reduce hot points and preserve quiet places. In Switzerland, measurement and protection against noise are ruled by the Noise Protection Ordonnance (NPO) [4]. As well as for all of the 168 countries having ratified the protocol of Kyoto, Switzerland is also committed to reduce $CO_2$ emissions due to road traffic. As a consequence, mobility is listed as one of the top priorities of the Swiss environmental research plan for years 2013-2016 [5] in which the need of information systems and traffic management is highlighted.

Traffic data collecting and processing are what road traffic monitoring (RTM) refers to. Real time knowledge of the network characteristics (number of vehicles per hour, average speeds etc.) plays a key role in ensuring road safety, regulating the traffic or improving the reactivity of rescue teams. Also, long-term data and historical trends (daily average traffic density, rush hours etc.) enable the future infrastructure investments to be optimized. For more than seven decades, RTM is one of the most basic administrative request in the US and the EU [6, 7].

Equipments dedicated to RTM have been investigated through many comparative tech-

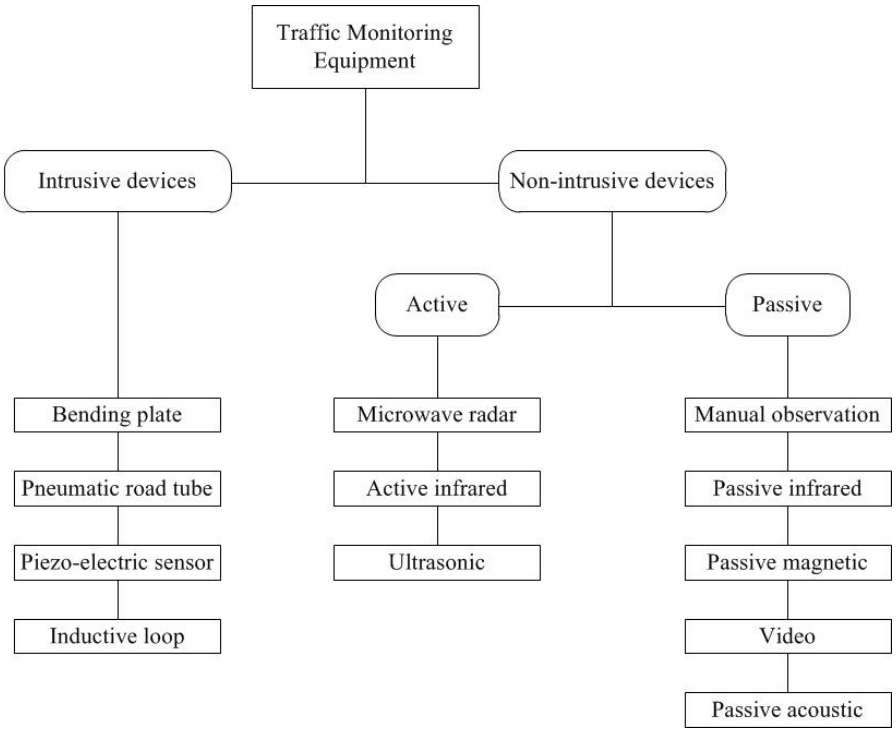
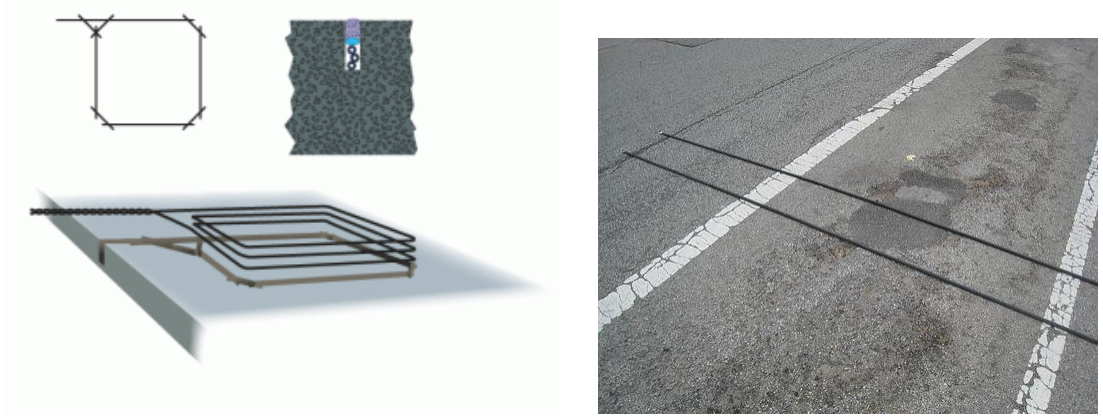



Figure 1.1: Classification of traffic monitoring equipment.

nical studies in the last decade [7, 8, 9, 10, 11]. These reports provide pros, cons, prices and limitations technologies currently available on the market. These technologies can be divided into two categories: intrusive and non-intrusive, Fig. 1.1. Intrusive devices are the most common, they involve the installation of sensors on top or into the lanes to be monitored. Despite their high reliability, safety considerations, damage risks or installation costs may limit their use. For instance, inductive loops consist of a metallic wire coiled to form a loop placed into the road pavement which senses the magnetic variations due to the presence of a metallic mass, see Fig. 1.2a. Its installation requires to cut and re-surface the road pavement, causing disruption of trafic and making the maintenance quite difficult and expensive. Another example is the one of pneumatic road tubes, generally used for short-term traffic counting. Two tubes are placed on the road lane, both perpendicular to the traffic flow direction, sensing the pressure variations when a moving body drives through, see Fig. 1.2b. These detectors are exposed to vandalism and damage caused by busy traffic. As an anecdote, a collaborator at LEMA used such a device during winter 2010 and it was destroyed after a snowplow passage.

As an alternative, one can use non-intrusive detectors which are placed on the roadside or in height. Their installation and maintenance do not need any traffic interruption making their deployment more secure than intrusive detectors. Non-intrusive detectors can be active or passive. Active ones emit a deterministic signal and measure the echoes produced by interactions with vehicles. They can handle detection, counting, speed and vehicle length estimation problems. Their main drawback concerns human or animal



(a) (b)

Figure 1.2: (a) inductive loop and (b) pneumatic road tubes detectors. Sources: www.diamondtraffic.com and www.bikecommuters.com

safety [12], especially regarding emission of optic rays, ultrasonic or electromagnetic waves. Conversely, passive technologies are those which do not emit any wave: the estimation procedure is based on an environmental sensing only. Microphone array belong to the latter category. It is a totally inoffensive solution that has the advantage of providing different kinds of data depending on the developed processing algorithm on the basis of the same physical measurement. Despite the significant progress made in the field of audio and video research the last decade, passive technologies suffer from the reputation of not being as efficient as active or intrusive ones, which was almost true in the 1990s because of the limited computation ressources. But based on the power of modern-day computing, a large community of acoustic researchers are working on the challenge of equalling, or even outperforming , the performance of active and/or intrusive technologies.

## 1.2 Primary statement of the thesis

The presented work specifically focuses on road vehicles monitoring by means of acoustic sensing. For the sake of clarity, the scenario of interest is shown schematically in Fig. 1.3. A section of road, with one or two lanes of circulation, is monitored by a few number of microphones placed on the roadside. Vehicles enter and leave the monitored zone according to an unknown law. Interfering noises may occur (aircraft landing, pedestrians speaking, tractor machinery, other vehicles etc.). A detection step, based on acoustic or other kind of sensors, returns an alert each time a new vehicle enters the monitored area. The tracking step is then activated. The objective is to estimate the “hidden states” of each detected vehicle as it passes by, namely, position, speed and wheelbase length (if two-axle vehicle).

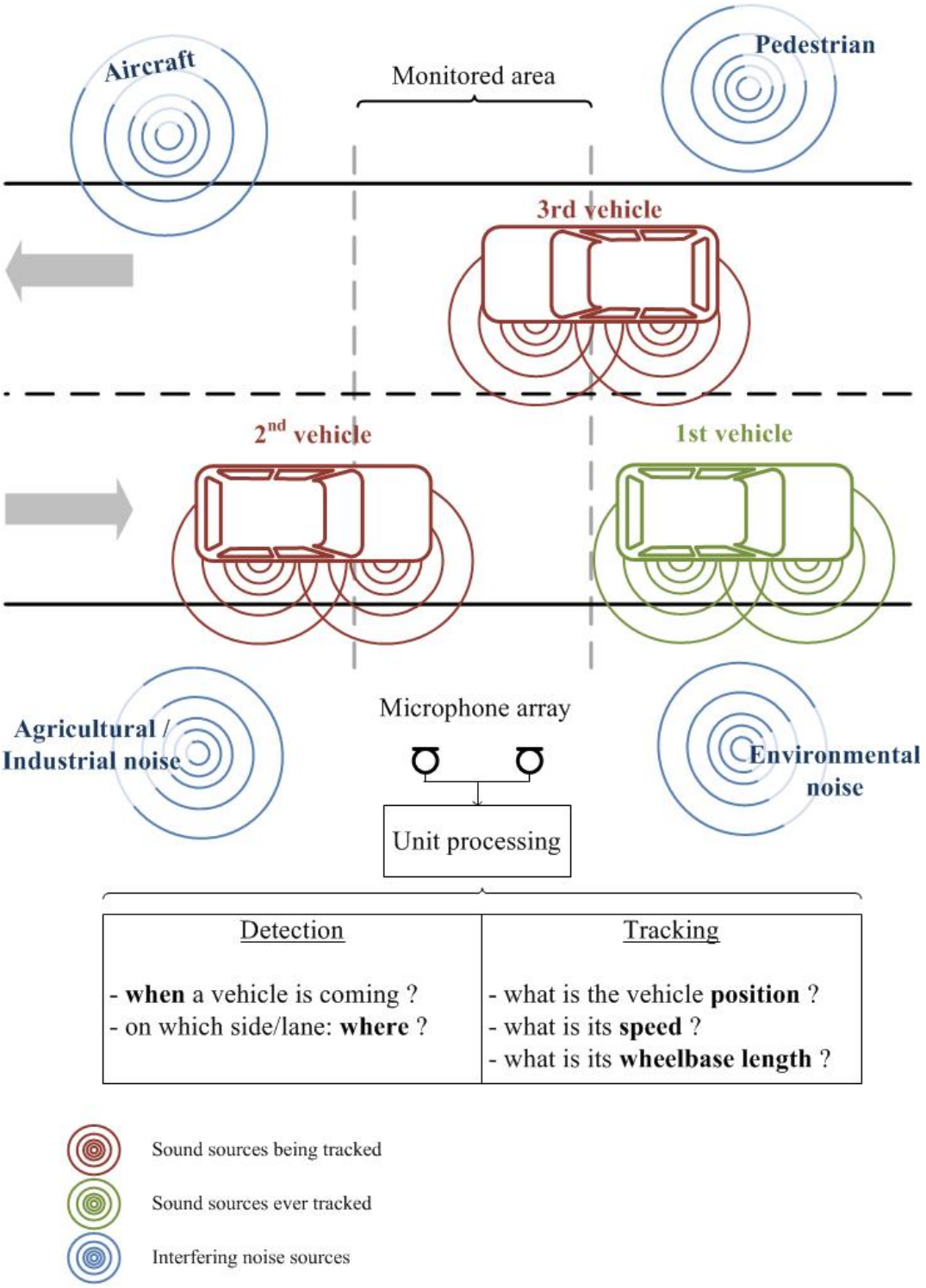



Figure 1.3: Objective of the thesis. A microphone array composed of a limited number of sensors, easily movable and of small aperture, is disposed on the roadside as a standard sound pressure level meter, the acoustic recordings are processed in real time to deliver number, position, speed and wheelbase length of vehicles as they pass-by.

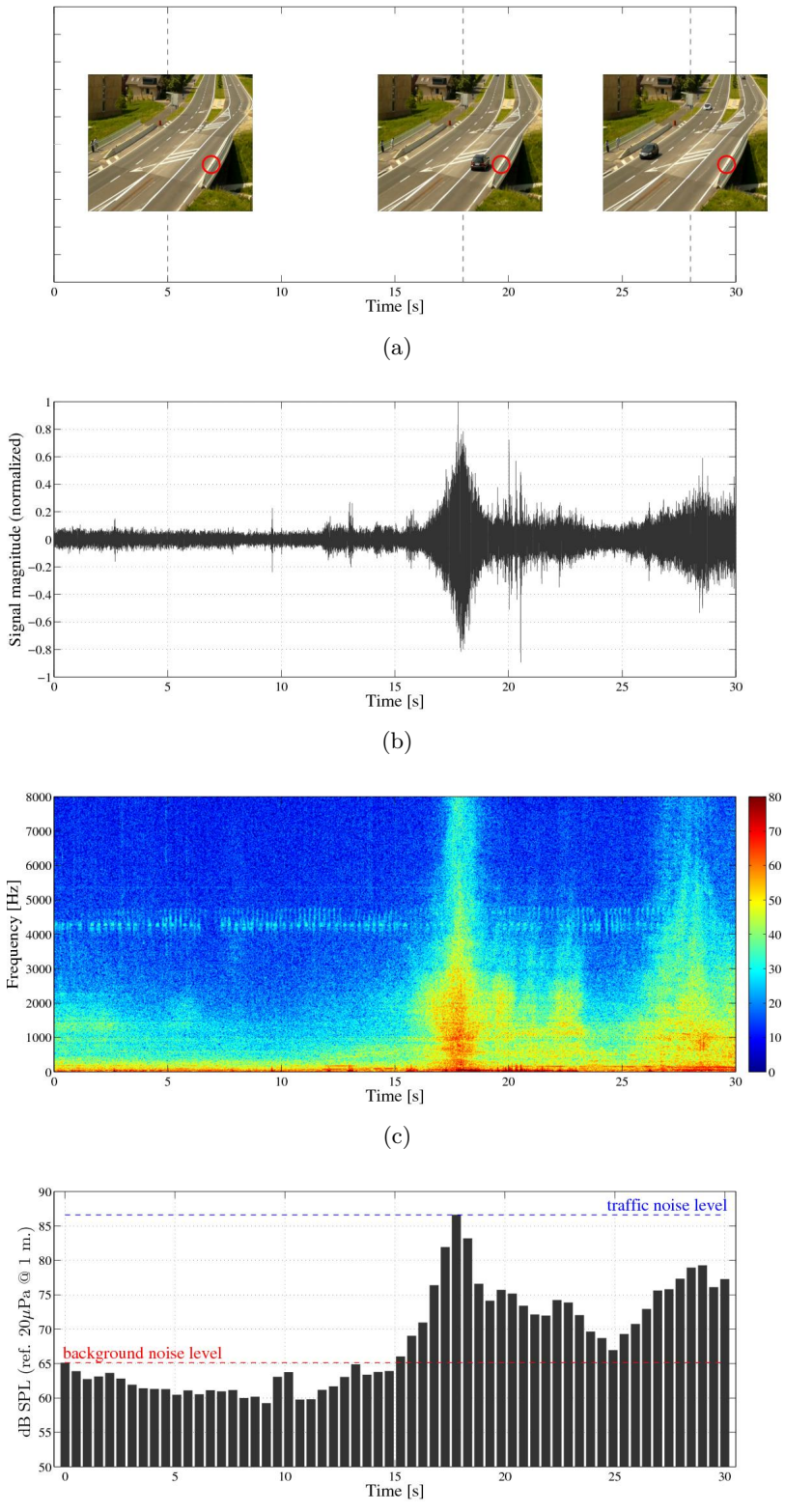



Figure 1.4: A typical *in-situ* audio recording. (a): pictures show the placement of the microphone (red circle) and the positions of vehicles as a function of time, (b): temporal waveform, (c): spectrogram and (d): equivalent sound pressure level. The visible patterns in the spectrogram between 4000 Hz and 4500 Hz are due to cricket chirps.

In the present case, the *signal of interest* is what is commonly called the *pass-by noise*, that is, a combination of mechanical, aerodynamic and tyre/road noises produced by vehicles in movement and perceived from an external and static observer. For readers familiar with temporal and spectro-temporal representations of audio signals, a typical recording of several pass-by noises is depicted in Fig. 1.4. Photos in Fig. 1.4a depict the environmental conditions of the measurement. The location of the roadside microphone is represented by a red circle. The audio excerpt lasts 30 seconds, in which three different cases occurs: no pass-by (until the $15^{th}$ second), one pass-by on the nearest lane (between 15 and 20 seconds), and one pass-by on the opposite lane (since the $25^{h}$ second). The temporal waveform is depicted in Fig. 1.4b. The matching spectrogram in dB SPL is depicted in Fig. 1.4c. The broadband nature of the pass-by noise is clearly visible on this plot. Fig. 1.4d depicts the equivalent continuous sound pressure level (Leq SPL), averaged over a half second, as a function of time. It indicates what can be typically expected in term of signal to noise ratio (SNR), that is, the difference in dB SPL, between the useful sound pressure level and the background noise. In this example, a peak-to-peak difference indicates a SNR of more than 25 dB for the nearest lane, and more than 15 dB for the farthest one.

## 1.3 Motivation

Passive acoustic monitoring has no direct impact on the environment. However, it is needed in any planned action dedicated to reduce the environmental impact of transport. A smart acoustic station should be able to establish a diagnostic of noise and in the same time extract some additional information (number, speed, vehicle types etc.) in order to deduce the energy consumption and emissions of pollutants on a road leg, to assess new facilities, or to map the acoustic noise to help cities in their territorial facilities policy. These reasons constitute the *environmental* motivation of this work.

If conventional sensors are not so expensive at the scale of a city, the collection and processing of data can be considered as expensive and time consuming: each sensor has its own data format, its own location, and is not necessarily synchronized with the others. Thus, the "ideal" sensor for the operator is the one that can be placed on the roadside, without cables, and which automatically provides all the required data. Such an *all-in-one* sensor must replace heavy and expensive current technologies requiring several technical skills and materials. This constitutes the *technical* motivation of this work.

Finally, as recently pointed out by Perez-Lorenzo *et al.* [13], it is a general trend in microphone array processing to use a high number of microphones both in the research community and in industry. But in the RTM context, the demand consists of low-cost, robust and versatile sensor systems able to automatically monitor road sections. This is mainly because current solutions are expensive to produce, install, repair and because they consume too much time to process data. The philosophy of this work has always


(a) (b)

Figure 1.5: Acoustic traffic-actuated signal light of Charles Adler (1928). (a) *stop, sound horn to clear signal*, (b) *to obtain signal, stop blow horn*. Source: http://www.rolandpark.org/ThenAndNowNorthwest

been to extract the maximum of information with the minimum of sensors. This led us to design a compact microphone array providing sparse observations which have to be compensated by advanced and still affordable signal processing techniques. This last point constitutes the *scientific* motivation of the thesis.

## 1.4 Acoustic sensing for road monitoring: a state of the art

The long story of road vehicle detectors actually began with acoustics. In 1928, Charles Adler Jr. developed the first traffic light system designed to manage vehicles at crossroads. Motorists approaching the intersection, facing a red light, were advised to blow their horns. A microphone then transmitted the sound to a call box, which caused the light to change [9, 14], see Fig. 1.5.

Surprising as it may seem, non-intrusive technologies were largely predominant in the first half of the $20^{th}$ century, magnetic, ultrasonic and microwaves sensors were used until the 1960s. Inductive loops and pneumatic road tubes largely replaced non-intrusive methods afterward [7]. The renewed interest for innovative techniques took off in the 1990s, corresponding to the political will of cities, especially in the U.S.A., to reduce the

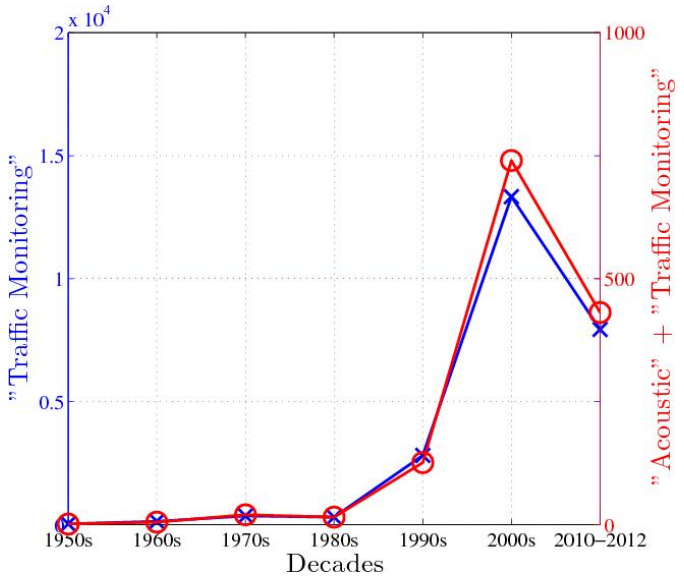



Figure 1.6: Number of articles and conference papers focusing on "Traffic Monitoring" (blue) and "Acoustic" + "Traffic Monitoring" (orange) as a function of the decades except for the last value (three years only). These results come from the Google Scholar search bar and were carried out in October 2012.

construction of new roads while exploiting the existing network at best. Non-intrusive technologies which were discarded hitherto took advantages of advances in computer and signal processing sciences. This phenomenon clearly appears in Fig. 1.6 which illustrates the quantity of articles/conference papers dealing with "Traffic Monitoring" regardless of the technology as a function of decades (in blue, left ordinate axis). Those dealing with "Acoustic" + "Traffic Monitoring" (in orange, right ordinate axis) also increase exponentially, obviously, to a lesser extent, but following the same trend.

Nowadays, engineering traffic noise measurements are reduced to a normative sound pressure level as a function of time using an omnidirectional microphone (sound level meter). But to establish relevant analysis, number and types of vehicles are generally required. In case of short term measurements, this classification is done manually. Otherwise, one resort to pneumatic road tubes and/or microwave Doppler radar to obtain additional data. Practitioners must be careful when using both sound level meter and pneumatic road tubes at the same time because of the "plops" sounds emitted when a vehicle travels through the tubes. This is why tubes are disposed at nearly 100 meters of the microphone in practice, making the post-processing rather complex because of the spatial and temporal incoherence between the two sensors. Similarly, the use of radar needs a meticulous positioning and calibration process. Missed detections may also occur because of the masking effect between vehicles in case of high traffic. Once again, radar and acoustic data are not synchronized and require a manual post-processing. In this context, a microphone array appears as a good candidate as it can provide sound level measurements, and also, handle counting, classification, speed estimation problems, each

at the same location and with the same time clock. Moreover, a microphone array can act as a spatial filter by dissociating sounds coming from the road from those coming from other directions, providing much more relevant result than with the standard sound level meter which integrates all the surrounding sound sources without any distinction. There has been a growing interest in passive acoustic-based systems for vehicle monitoring since the mid 1990s. In 1996, vehicle classification using wavelet decomposition of audio signals were investigated by Choe *et. al* in [15]. Automatic classification has then been investigated by numerous researchers, especially for the military context. In 1997, Chen *et al.* [12] and Forren *et al.* [16] independently investigated the road vehicle detection problem using cross-correlation functions between sensor pairs. The counting problem was also handled by Brockman *et al.* in 1997 [17] and Kuhn *et. al* [18] in 1998 which respectively deployed an auto-regressive algorithm based on a pass-by spectrum model (one sensor) and a beamforming-based technique (80 sensors) to detect vehicle presence. Other modern techniques have emerged for early queue detection, manage crossroads, estimate vehicular traffic density and so on [19, 20, 21, 22, 23, 24, 25, 26]. The speed estimation problem has also been addressed extensively, for instance in [27, 28, 29, 30, 31, 32, 33, 34, 35, 36]. Recently, a trend consists in seeing the pass-by noise as a measure of the energy consumption: in 2011, Can *et al.* successfully showed the correlation between emitted airborne pollutant and road traffic noise near a highway [37].

## 1.5 Outlines and original contributions of the thesis

This section summarizes the contents and the original contributions of the following chapters.

In chapter 2, key concepts involved in localization of static and wideband sound source are recalled. Due to the acoustical conditions of observation, it is demonstrated that the localization problem here can be turned into a time-delay estimation problem. Most common time-delay estimators are discussed through theoretical development and experimental measurements, in particular, we largely argue in favor of generalized cross-correlation functions, making the first contribution of the thesis.

The case of moving sound source is discussed in chapter 3. After recalling the conventional techniques used in acoustics the case of harmonic moving sources speed estimation, the less conventional but more suitable Bayesian theory for broadband source tracking is introduced, with an emphasis on the particle filtering algorithm. The contribution of this chapter is twofold, first a state of the art of Bayesian-based tracking methods is established, discarding the Kalman filter and its variants in particular. Secondly, we processed to a preliminary measurement which, in addition to validate the proposed method, allows the reader to figure out how implement a particle filter in practice, and see the relationships between the physical problem and the Bayes' statistical way of thinking.

The gap between the theoretical developments in the chapters above and practical road vehicle monitoring problem is bridged in chapter 4. Time-delay estimation and tracking techniques are both improved to match with the monitoring of two-axle road vehicles at best. A closed-form expression of the observation is derived, constituting the first contribution of this chapter. Moreover, a Bayesian model of two-axle vehicles is proposed, defining an improved particle filter allowing the estimation of wheelbase length, rather rarely addressed in the acoustic community, constituting a second contribution.

Given that the performance of any tracking algorithm is related to the observation quality, a specific methodology of microphone array design is presented in chapter 5. It consists in optimizing the inter-sensor distance in order to feed the tracking algorithm at best depending on the geometrical and spectral characteristics of the scenario.

Experimental results of the thesis are presented and discussed in chapter 6. Both aspects of tracking and detection strategies are assessed. Besides the promising results themselves, one contribution of this chapter is the share of our experience about in-situ measurements.

Chapter 7 needs to be considered as a freelance investigation of an unaddressed problem in compact microphone array processing: the estimation of the number of axles and the separation of their sound contribution. A research approach based on the subspace-based theory is investigated for the pure tonal case. First results highlight interesting mathematical difficulties to overcome in the future.

A summary of the key findings of the work achieved during this thesis is presented in chapter 8, suggesting some lines for future research.

# 2 Airborne sound source localization

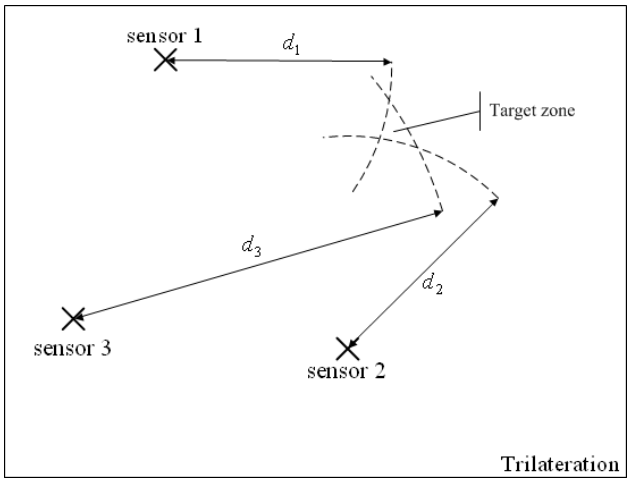
## 2.1 Introduction

The airborne sound source localization (SSL) problem consists in estimating position (coordinates) or bearing (angle) of an active point emitter through sound pressure measurements of the radiated wavefield. These measurements are carried out using microphones placed at different points of space, forming a *microphone array* with known geometry. Recordings are processed by a *localization algorithm* which delivers the sound source position estimate. SSL is addressed in a plethora of applications and research works, for instance marine mammals localization [38], human-computer interactions improvement [39], speaker localization and identification [40], hearing aid improvement [41] to list a few. Localization algorithms are numerous but rely on three main principles: trilateration, triangulation or multilateration.

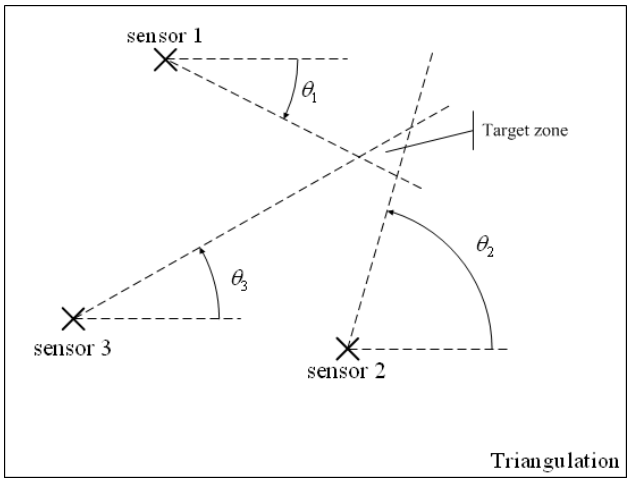
A famous example of trilateration-based algorithm is the Global Positioning System (GPS) one which equips most of the cars and smartphones. It consists in acquiring, on a receiver, signals broadcasted by satellites in orbit with known position, comparing the times of arrival of each signal, deducing the distance between the receiver and each satellite, and finally estimating the receiver position. Hence, the trilateration principle relies on *absolute distances* between the object and reference points, as depicted in Fig. 2.1a.

When absolute distances are not available, another technique consists in measuring *angles* between a reference direction and the target direction and repeating the procedure for several space locations. All the measured directions should therefore intersect at the actual target position in the Cartesian plan. This is the so-called triangulation principle, depicted in Fig. 2.1b.

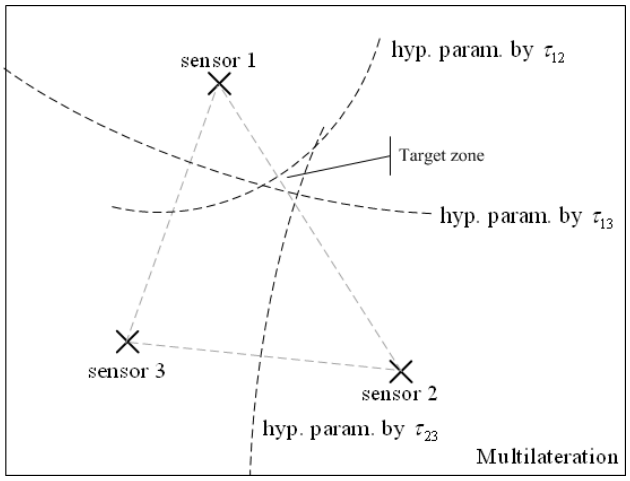
When neither distances nor angles are available, as it is the case when using microphones, one can rely on the multilateration principle. This technique is based on *signal spatial*




(a)




(b)




(c)

Figure 2.1: (a) Trilateration: sensors deliver absolute distances, (b) Triangulation: sensors delivers angles, (c) Multilateration: sensor pairs deliver hyperboloids of solutions.

*differences* between sensors. Namely, due to the bounded speed of sound, a wavefront coming from an active point emitter arrives at different time instances on each microphone. The time for the sound wave to travel from a microphone $m_1$ to another one $m_2$ is called time-delay of arrival (TDOA) and is denoted $\tau_{12}$ hereafter. The set of solutions for a given delay $\tau_{12}$ is an hyperboloid whose foci are the microphones and whose shape is totally parameterized by the speed of sound, the inter-sensor distance and the delay itself. Using several sensor pairs therefore yields an estimate of the object position by solving an hyperboloids intersection problem, as depicted in Fig. 2.1c in the 2 dimensional (2D) plane.

## 2.2 Direct methods

SSL problems are traditionally addressed through *direct methods* or one-step procedures. The general idea consists in finding which position or direction of arrival (DOA), among a set of candidates, explains the observation delivered by all sensors at best. The most classical one-step procedure is the delay-and-sum beamformer (DSB). A beamformer "steers" the acquired signals into one desired direction by numerically compensating the physical delays inherent to this direction. The summation of these delayed signals is coherent and of maximal power if the steering direction corresponds to the actual sound source one, if not, the summation is incoherent and the additive effect of signals which are not in phase produces a lower response power.

To avoid spatial aliasing, the sensors of a DSB should be spaced less than half the smallest wavelength of interest, $d \leq \lambda_{min}/2$. Under the plane wave hypothesis and using a uniform linear array, the closed-form expression of a DSB beampattern is [42] page 57:

$$|b(f,\theta)| = \left| \frac{sin([\pi f M d(sin\theta - sin\theta_0)]/c)}{sin([\pi f d(sin\theta - sin\theta_0)]/c)} \right| \tag{2.1}$$

where $M$ is the number of sensors, $f$ the frequency (Hz), $\theta_0$ is the actual source DOA and $\theta$ is the steering direction. Beampatterns of an array composed of $M = 2$ sensors spaced by $d = 3.5$ cm (corresponding to $\lambda_{min}/2$), and $d$=20 cm are respectively depicted in Fig. 2.2a and Fig. 2.2b. In both cases, $\theta_0 = 0$ and $\theta$ vary between -90° and +90°. In the first case, no spatial aliasing occurs: the maximum power corresponds to the actual DOA without ambiguity whatever the frequency. However, the resolution is very low regarding the global power (summation over frequencies). The angular resolution at -3dB is 120°. With the larger array, the resolution is better but spatial aliasing occurs above 1715 Hz. Above this frequency, the maximal power may correspond to multiple DOA.

Much higher performances can be achieved with *subspace-based methods*. These include the Capon beamformer [43], the multiple signal classification algorithm (MUSIC) technique [44, 45] or the estimation of signal parameters via rotational invariant techniques (ESPRIT) [46]. Each relies on the singular value decomposition of the acquired signals covariance matrix. The subspace theory is explained in more detail in chapter 7.

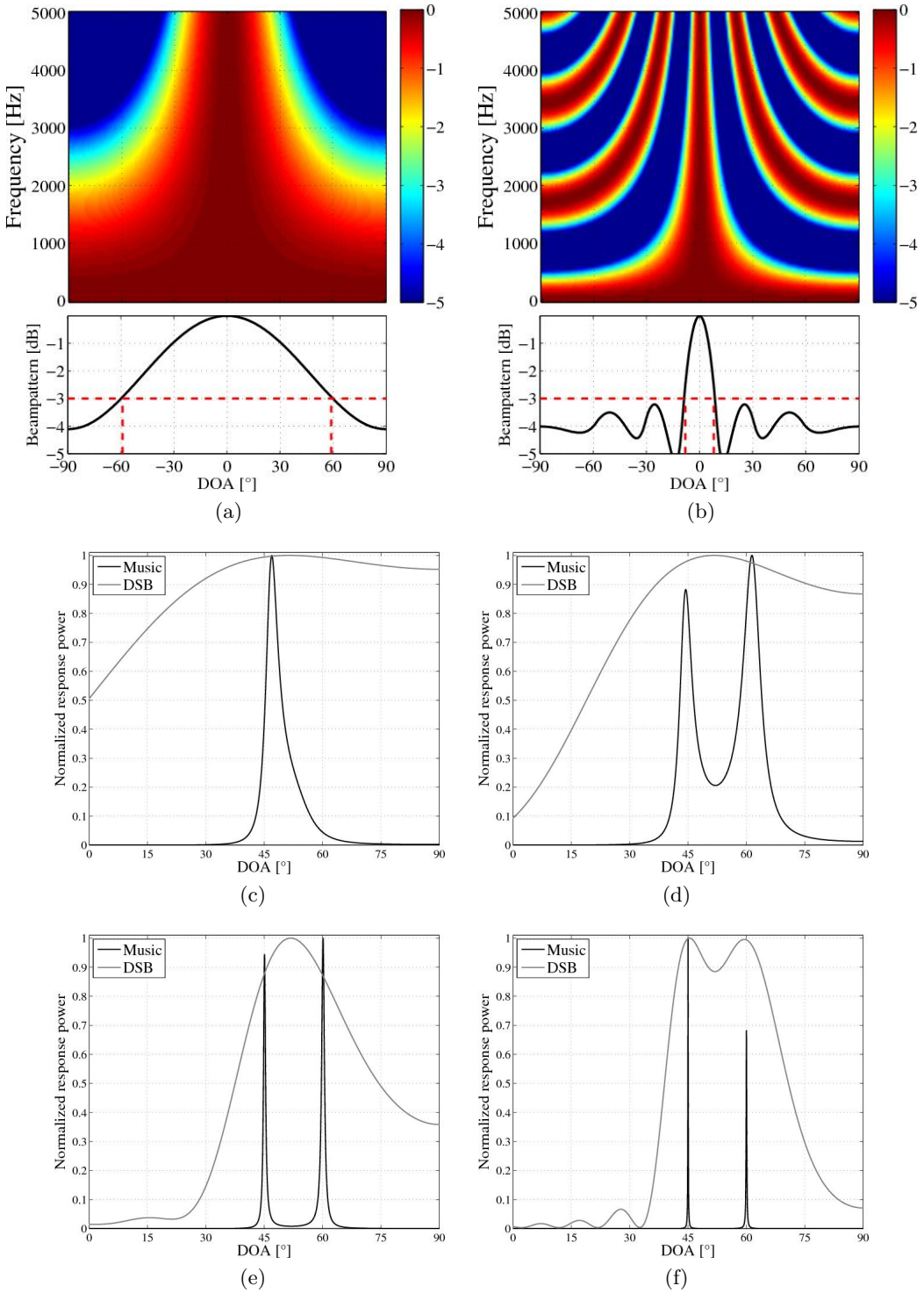



Figure 2.2: Top: spectral and global beampattern of a delay-and-sum beamformer composed of two microphones spaced by (a) 3.5 cm and (b) 20 cm. Below: comparison between delay-and-sum beamformer spectrum and music spectrum; two sound sources (white noise) are in the far field, the microphone number is equal to 3 in (c), 5 in (d), 15 in (e) and 30 in (f).

For now, let us consider the following simulation: two microphones spaced by $d$=20 cm, acquire the wavefronts of two zero mean Gaussian and uncorrelated process located at 45° and 60° in the far field, the 0° reference being the end-fire DOA, the SNR equals +10 dB. The number of snapshots is 8192 and the grid of research is uniform with 0.01° sampling. DSB and MUSIC spectra are compared in Fig. 2.2c to Fig. 2.2f. Four different microphone numbers are tested: 3 in Fig. 2.2c, 5 in Fig. 2.2d, 15 in Fig. 2.2e and 30 in Fig. 2.2f. It is clear that the MUSIC algorithm outperform the DSB one regarding the much sharper peaks that have been obtained by MUSIC.

The price to pay is that such a method requires i) more sensors than sources, ii) a wave propagation model matching well with reality, and iii) a high number of snapshots. Regarding the applied context of this thesis, point i) is at odds with the objective of developing a small, light and easily movable microphone array, point ii) seems unrealistic to ensure in outdoor conditions, and point iii) acts as a hindrance to the development of a real time road traffic monitoring device.

Moreover, both beamforming and subspace-based methods have all been initially designed for *narrowband* signals, *i.e.* sounds having their spectrum centered around a central frequency and a bandwidth which does not exceed one octave [47]. But from what has been discussed in section 1.2, sounds of interest in this work are rather *broadband*, meaning they spectral bandwidth is rather large and flat. In everyday life, speech, road traffic noise, aircraft noise are all examples of broadband sources, which are very different from pure tone signals usually processed in underwater acoustics, sonar or electromagnetism and for which these one-step procedures have been initially designed.

As a consequence, one needs to rely on another framework, namely, the two-step procedures or *indirect methods* consisting in estimating the source position only after estimating the energy or phase differences between sensor pairs. In the following, indirect methods are introduced by firstly describing the signal model on which they are built.

## 2.3 Signal modeling

Let $\mathbf{r}^s$ be the coordinates of the sound source to locate and let $\mathbf{r}_1^m$ and $\mathbf{r}_2^m$ be the coordinates of the microphones. Without loss of generality, let the first microphone be the reference sensor. Under the assumption of an ideal non reverberant, non dispersive and homogeneous medium, the signals acquired by the two microphones $y_1(t)$ and $y_2(t)$ are attenuated and delayed versions of the original signal $s(t)$ such that:

$$y_1(t) = a_1 s(t - \delta_{11}) + n_1(t), \tag{2.2}$$

$$y_2(t) = a_2 s(t - \delta_{11} - \tau_{12}) + n_2(t), \tag{2.3}$$

where $a_1$ and $a_2$ are attenuation factors due to the propagation effects, $\delta_{11}$ is the time of flight (TOF) that the sound wave needs to travel from $\mathbf{r}^s$ to $\mathbf{r}_1^m$, $n_m$ is an additive noise due to the $m^{th}$ channel of the acquisition device, considered as a stochastic, stationary,

zero-mean Gaussian signal, uncorrelated both with the signals and noise at other sensors, and $\tau_{12}$ is the TDOA between the two sensors. According to the model (2.2)-(2.3), what differs between $y_1(t)$ and $y_2(t)$ are the amplitude and phase information. Both may be used as $\mathbf{r}^s$ estimation features.

### Energy-based methods

Methods that exploit the amplitude differences between signals acquired at different positions are called *energy-based methods* [48, 49, 50]. This approach rely on the fact that the amount of source energy attenuation at a sensor is proportional to the square of the distance between the source and the sensor. Such techniques are commonly used for military, bioacoustics or underwater acoustics problem due to the large distances between sensors. In the present case, because a small aperture array is used, inter-sensor distances compared to distances between sensors and sound sources make unnoticeable the magnitude differences within the array. Thus, it is assumed that $a_1$=$a_2$ throughout this document, definitely discarding this kind of methods.

### Time-delay-based methods

Methods that exploit the time-delay of arrivals between signals are called *time-delay-based methods*. They are based on the estimation of the TDOA $\tau_{12}$ which is related to the microphone positions and sound source position through the relation:

$$\tau_{12} = \frac{||\mathbf{r}^s - \mathbf{r}_1^m|| - ||\mathbf{r}^s - \mathbf{r}_2^m||}{c}, \tag{2.4}$$

where $c$ is the speed of sound (in m/s). Considering $\mathbf{r}^s$ as the variable turns (2.4) into the expression of a half-hyperboloid in 3D (hyberbola in 2D) with foci at coordinates $\mathbf{r}_1^m$ and $\mathbf{r}_2^m$. Consequently, an infinity of positions can explain one single time-delay measurement. This is why a set of delays, coming from different sensor pairs, is required to properly estimate the source coordinates. One solution for solving the hyperbola intersection problem in the 2D case is derived analytically in Appendix A.1.

In the case of an array aperture much smaller than the distance between the array and the source ($c\delta_{11} >> c\tau_{12}$), the successive incoming wavefronts are quasi-planar, as depicted in Fig. 2.3. One says that the sound source is in the *far-field* of the array. Such a propagation model enables the array to return a bearing estimation only, thanks to the relation:

$$\tau_{12} = \frac{d}{c}\sin\theta, \tag{2.5}$$

where $d$ is the inter-sensor distance, defined by:

$$d = ||\mathbf{r}_1^m - \mathbf{r}_2^m||\,, \tag{2.6}$$

and $\theta$ is the sound source DOA, also called sound source bearing.

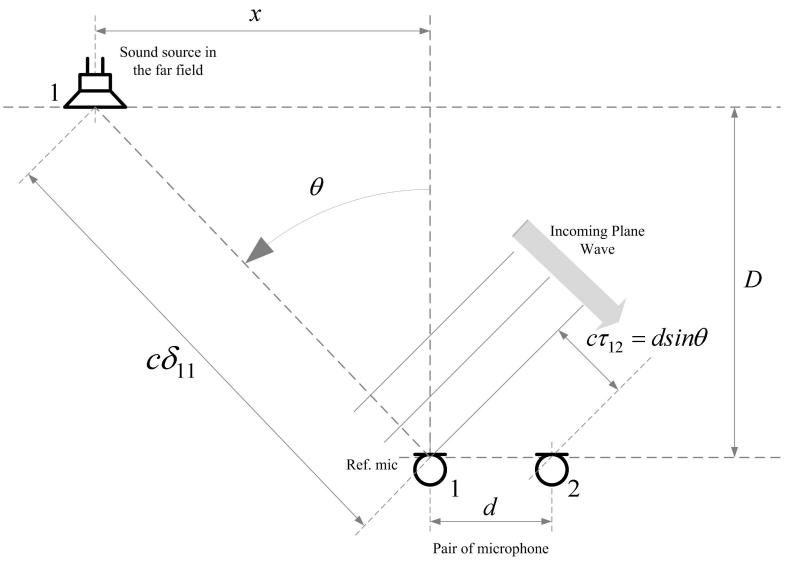



Figure 2.3: Far-field hypothesis: waves impinging at sensors are planar, one can only estimate the bearing of the source.

Now suppose that the sound source of Fig. 2.3 is a road vehicle whose position is constrained by a straight road at a known distance $D$ to the sensor. The abscissa of the source $x$ then becomes estimable through the relation:

$$x = D\tan\theta \tag{2.7}$$

$$= D\tan\left(\arcsin\left(\frac{c\tau_{12}}{d}\right)\right), \tag{2.8}$$

$$= D\frac{c\tau_{12}/d}{\sqrt{1-(c\tau_{12}/d)^2}}. \tag{2.9}$$

Relations (2.5) and (2.9) are both depicted in Fig. 2.4, the former in red with $d = 34$ cm and $c = 343$ m/s, the latter in blue with same $d$ and $c$ but with two different $D$: 1 m and 10 m. What is important to note for the following is the non-linearity between TDOA and DOA (in red), and between TDOA and abscissa (in blue). This non-linearity partly justifies the choice of a particle filtering-based tracking method introduced in chapter 3. Another remark concerns the bijective nature of the relationship between TDOA and DOA within the range [-90°,+90°]. This range also corresponds to vehicle DOAs when it is constrained by a straight road and observed by microphones placed in parallel to the road lane. One can derive the fact that in such a scenario, the localization problem is reduced to a time-delay estimation problem since only one pair is sufficient to locate the vehicle without ambiguity. Therefore, we concentrate our efforts on time-delay estimation procedures. Most of the famous techniques are addressed in the next section.

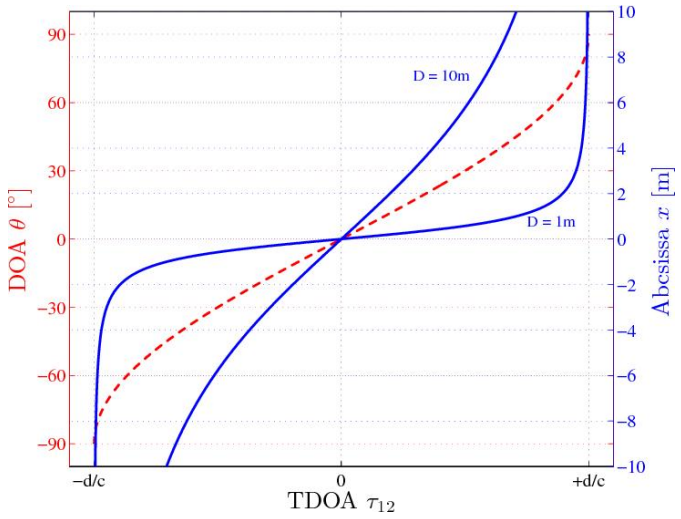



Figure 2.4: red dashed line: DOA as a function of the TDOA [Eq. (2.5)] with $d$ = 34 cm and $c$ = 343 m/s ; blue lines: abscissa as a function of the TDOA [Eq. (2.9)] for two different value of $D$: 1m and 10m.

## 2.4 Time-delay estimation

This section focuses on time-delay estimators between two broadband signals.

### 2.4.1 The cross-correlation function

It is a well-known result that in presence of a single source, the optimal estimator of $\tau_{12}$ is the lag corresponding to the maximum value of the cross-correlation between $y_1(t)$ and $y_2(t)$ [51]. In that case, one can also give an explicit expression of the Cramer-Rao lower bound (CRLB), which depends on the spectral bandwidth of the source and on the signal-to-noise ratio.

The cross-correlation (CC) function is the most straightforward method to estimate the delay between two broadband signals. It is particularly well adapted to the case of constant delay, stationary processes and long observation interval (CDSPLOT) [52]. It is defined by:

$$R(\tau) = \mathbb{E}\left[y_1(t)y_2(t-\tau)\right], \tag{2.10}$$

where $\tau$ is the time lag and $\mathbb{E}[.]$ is the statistical expectation operator. The value of $\tau$ that maximises Eq. (2.10) provides an estimate of the actual time-delay $\tau_{12}$:

$$\hat{\tau}_{12} = \arg\max_{\tau} R(\tau). \tag{2.11}$$

According to the Wiener-Kintchine theorem, the CC function may also be expressed in the Fourier domain:

$$R(\tau) = \int_{-\infty}^{+\infty} S_{y_1 y_2}(f) e^{i2\pi f \tau} df, \tag{2.12}$$

where $S_{y_1y_2}(f)$ denotes the cross-spectral density (CSD) of the signals such that:

$$S_{y_1y_2}(f) = Y_1(f).Y_2^*(f), \tag{2.13}$$

and $Y_j(f)$ is the Fourier transform of $y_j(t)$ such that:

$$Y_j(f) = \int_{-\infty}^{+\infty} y_j(t)e^{-i2\pi ft}dt,\ j \in [1,2]. \tag{2.14}$$

For the specific case where $y_2(t) = y_1(t - \tau_{12})$, that is, $a_1 = a_2$ and $n_1 = n_2 = 0$ in (2.2)-(2.3), one gets:

$$Y_2(f) = Y_1(f)e^{-j2\pi f\tau_{12}}. \tag{2.15}$$

From Eq. (2.12),Eq. (2.13) and Eq. (2.15), it appears that the shape of the CC is closely related to the spectral content of the acquired signal. A flat spectrum produces a delta function with its singular point at $\tau_{12}$. Conversely, a narrower spectrum produces a more sinusoidal shaped CC.

While being suboptimal, various techniques permit to accentuate the peak of the CC when applied to real world signals. They are the generalized cross-correlations, that are presented below.

### 2.4.2 The generalized cross-correlation functions

The goal of the generalized cross-correlation (GCC) functions is to accentuate the peak of the cross-correlation associated to the actual delay by filtering signals upstream the correlation. The expression of the GCC is given by:

$$R^{gcc}(\tau) = \int_{-\infty}^{+\infty} \psi_g(f)S_{y_1y_2}(f)e^{i2\pi f\tau}df, \tag{2.16}$$

where $\psi_g(f)$ is called the *weighting function*. Note that the basic cross-correlation function is a particular case of the generalized one with $\psi_g(f) = 1\ \forall f$. For more than four decades, many weighting functions have been proposed in the literature. The most famous of them are introduced below.

#### Phase Transform

The Phase Transform (PHAT) processor is given by [53]:

$$\psi_{phat}(f) = \begin{cases} \frac{1}{|S_{y_1y_2}(f)|} & \text{if } |S_{y_1y_2}(f)| \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2.17}$$

This processor was originally developed as an “ad-hoc” technique by Knapp and Carter

in the mid 1970s but remains today one of the most commonly used time-delay estimator in the SSL community. Reasons for its success are numerous: its implementation is straightforward, no *a priori* knowledge on signal and noise is required, it is more consistent than some other GCC members when the characteristics of the source change over time [54]. Also it has been found to perform very well under everyday life acoustical conditions. Recently, Zhang *et al.* proved that in case of high signal to noise ratio, the GCC-PHAT function is the optimal time-delay estimator in a maximum likelihood sense, regardless of the amount of reverberation in the environment [55]. Indeed, many practical comparative studies confirm its robustness in presence of multipath distorsion [56, 57, 58, 59].

**Maximum-Likelihood (or Hannan-Thomson) processor**

From a statistical point of view, the weighting derived by Hannan and Thomson in 1971 [60] is the optimal one under CDSPLOT conditions, without reverberation, in the sense that its variance can achieve the CRLB. It is known as the *Maximum-Likelihood* (ML) or Hannan-Thomson (HT) processor and is expressed by:

$$\psi_{ml}(f) = \begin{cases} \frac{\gamma^2_{y_1y_2}(f)}{1-\gamma^2_{y_1y_2}(f)} \frac{1}{|S_{y_1y_2}(f)|} & \text{if } |S_{y_1y_2}(f)| \neq 0 \\ 0 & \text{otherwise,} \end{cases} \tag{2.18}$$

where $\gamma^2_{y_1y_2}(f)$ is the coherence function between $y_1(t)$ and $y_2(t)$. It is given by:

$$\gamma^2_{y_1y_2}(f) = \frac{|S_{y_1y_2}(f)|^2}{S_{y_1y_1}(f)S_{y_2y_2}(f)}. \tag{2.19}$$

The coherence can be considered as a measure of the linear dependence between two signals. The ML estimator weights the cross-spectrum according to the SNR (term in $\gamma^2/(1\text{-}\gamma^2)$), giving more weight to the phase in regions of the frequency domain where coherence is large. This coherence term has the effect of canceling artefacts due to the band-limited characteristics of real-world signals [47].

**Roth processor**

In 1971, Peter Roth proposed to normalize the CSD by the auto-spectrum of one of the two signals, considered as the input to the system, the other signal being considered as the output, such that [61]:

$$\psi_{roth}(f) = \begin{cases} \frac{1}{|Y_1(f)Y_1^*(f)|} & \text{if } |S_{y_1y_1}(f)| \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2.20}$$

This procedure reduces the spectral components for which the auto-spectrum is large, and consequently, remove the effects of the input for more accurate delay estimation. Regarding applications, this approach does not hold since the spectrum of the vehicle cannot be measured directly which makes the input signal inevitably those acquired by

one of the two sensors. When microphones and acquisition system used are of good quality, and the transmission path between sensors implies only a delay, the auto-spectra of the two channels are similar,. Consequently, removing the effect of one microphone drastically deteriorates the final cross-correlation in the present case.

#### Smoothed Coherence Transform

In 1973, Carter *et. al* proposed the Smoothed Coherence Transform (SCOT) processor [62] expressed by:

$$\psi_{scot}(f) = \begin{cases} \frac{\gamma_{y_1 y_2}(f)}{|S_{y_1 y_2}(f)|} & \text{if } |S_{y_1 y_2}(f)| \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2.21}$$

In the SCOT method, the cross-spectra is normalized by the square root of the product of the auto-spectra of $y_1$ and $y_2$. In addition to suppressing the cross-spectral estimate in regions of the spectrum with low signal to noise ratio, high signal to noise ratio are also suppressed in order to deemphasize strong components such as pure tones in the broadband observations.

#### The relay cross-correlation

The relay cross-correlation proposed by Madala and Ivakhnenko in [63] has also been tested. Their idea was to exploit only the sign of the acquired signals. This one-bit quantification presents the advantage of drastically simplifying the computation of the cross-correlation, this is particularly used when it comes to implement it on an embedded apparatus for instance.

$$R^{relay}(\tau) = \mathbb{E}\left[sign(y_1(t))sign(y_2(t-\tau))\right]. \tag{2.22}$$

#### Other processors

Many other processors, optimal or suboptimal, have been proposed in the literature like the Wiener processor [64], the Eckart Filter [65], the Modified CPSP [66], Hassab-Boucher transform [67] and so on. Applying such processors on the signals used in this work did not lead to any significant improvement compared to the PHAT one. Moreover, one important advantage of the PHAT processor over other ones is that its closed-form expression can be easily derived as shown in chapter 4.

### 2.4.3 Others estimators

Many other time-delay estimators have been assessed in the RTM context. None of them gave better satisfaction that GCC-based ones but they are briefly listed below for the sake of completeness.

#### Least-mean-square method

The least-mean-square (LMS) estimator proposed by Reed *et. al* [68] considers the signal

of one channel as the finite-impulse-response filtered version of the signal of the other channel. Reed proposed to recursively estimate this filter by beginning with a candidate impulse response and minimizing the mean-square error between the reference channel and the filter output.

In our experience, based on real measurements, an estimate of the impulse response much more accurate than the basic cross-correlator one is effectively achieved [69], but the price to pay is the computation time which is totally unadapted in the context of an *in-situ* monitoring application. Anyway, getting the exact impulse response between two signals is not the objective here since only time-delays are of interest. Furthermore, the LMS algorithm requires a feedback coefficient that controls the convergence rate and which is delicate to properly adjust.

#### Higher-order statistics

The higher-order statistics (HOS) technique exploits the fact that, for Gaussian processes, moments and cumulants of order greater than two are null. Estimating the signal parameters in the higher statistical domain is therefore a big advantage in case of Gaussian noise, even if this one is correlated with the signal. This supposes that signal and noise are respectively non-Gaussian and Gaussian. This technique was originally developed for underwater passive sonar applications, where "listened" signals often come from complicated mechanical systems with strong periodic (or quasi-periodic) components, and therefore considered as non-Gaussian [70].

After having implemented and checked *in-silico* the validity of this method, multiple unsuccessful attempts using real data led us to conclude that the non-Gaussianity assumption of the source of interest does not hold, definitely discarding this method.

#### Other methods

Many other time-delay estimators have been proposed in the literature. They are classically compared regarding their variance as a function of the signal to noise ratio (SNR), reverberation, or number of sensors. A reference article is that of Chen *et. al* [54] in which GCC, Multi-Channel LMS, Blind Channel Identification, Adaptive Eigenvalue Decomposition (AED) and others techniques are introduced and compared. Comparison of time-delay estimators has also been the subject of the Ph.D thesis of Björklund in 2003 [71].

Time-delay estimators have been studied thoroughly in the last decades since they find applications in various fields like radar, ultrasonics, communications or seismology. In the acoustic processing community, one research field of growing interest consists in counteracting the effects of reverberation as in underwater acoustics or room acoustics. Indeed, it is known that GCC-based estimators tend to break down in presence of a too large multipath distorsion. However, in this thesis, reverberation (in the sense of multipath distorsion) has never been an issue, considering the measurements which have been carried out. On the other hand, objects being dynamic and sometimes numerous, it

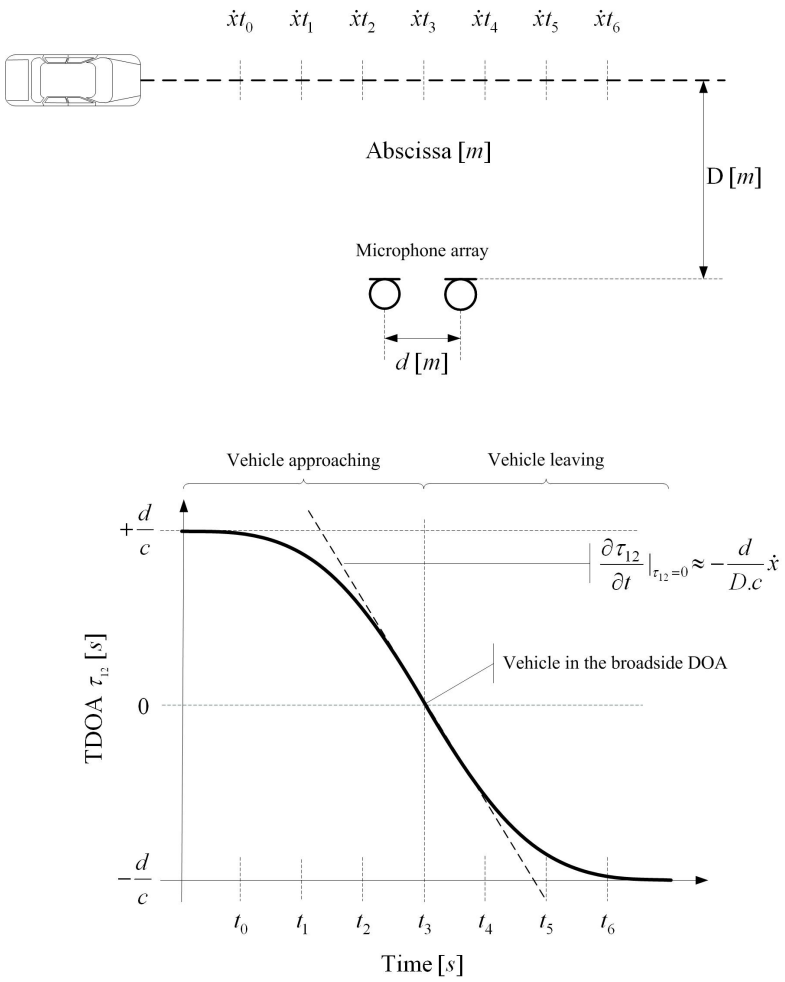


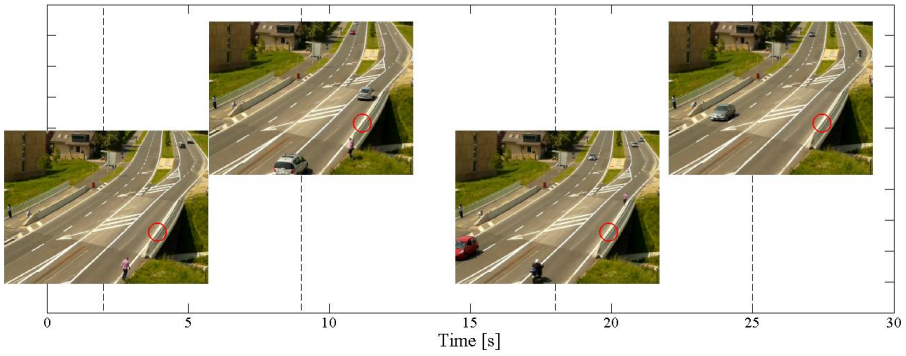

Figure 2.5: Typical cross-correlation time series (CCTS) for a vehicle running at a constant speed in a straight line and which sound is acquired by two microphones placed in parallel to the trajectory.

is required to have fast TDOA estimate updates. CC-based methods are ideal in this regard.

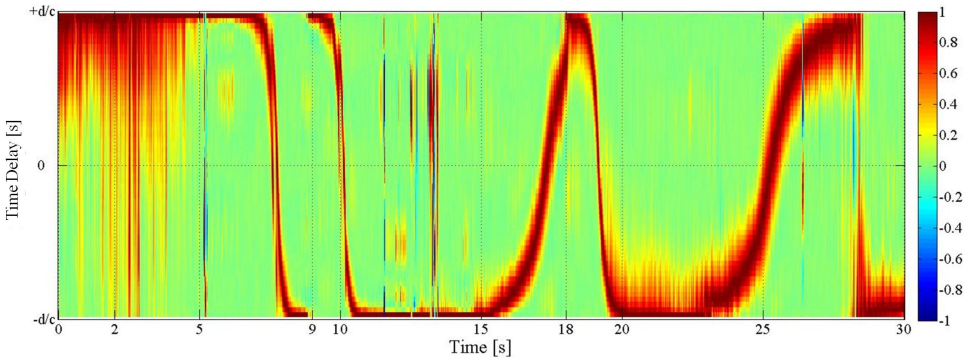
## 2.5 Cross-correlation time series

Estimating motion parameters of a sound source requires to get estimates on its position repeatedly. In practice, cross-correlation measurements are performed on short audio frames (30 - 40 ms) with overlap. A convenient way to observe the time evolution of TDOA consists in plotting the concatenation of successive cross-correlation measurements, introducing the notion of CCTS in two dimensions: TDOA versus time.

Consider a vehicle running at a constant speed $\dot{x}$ on a straight road monitored by two microphones, placed in parallel to the road lane, at a distance $D$ of the closest point of approach (CPA). The concatenation of the correlation measurements yields a




(a)




(b)

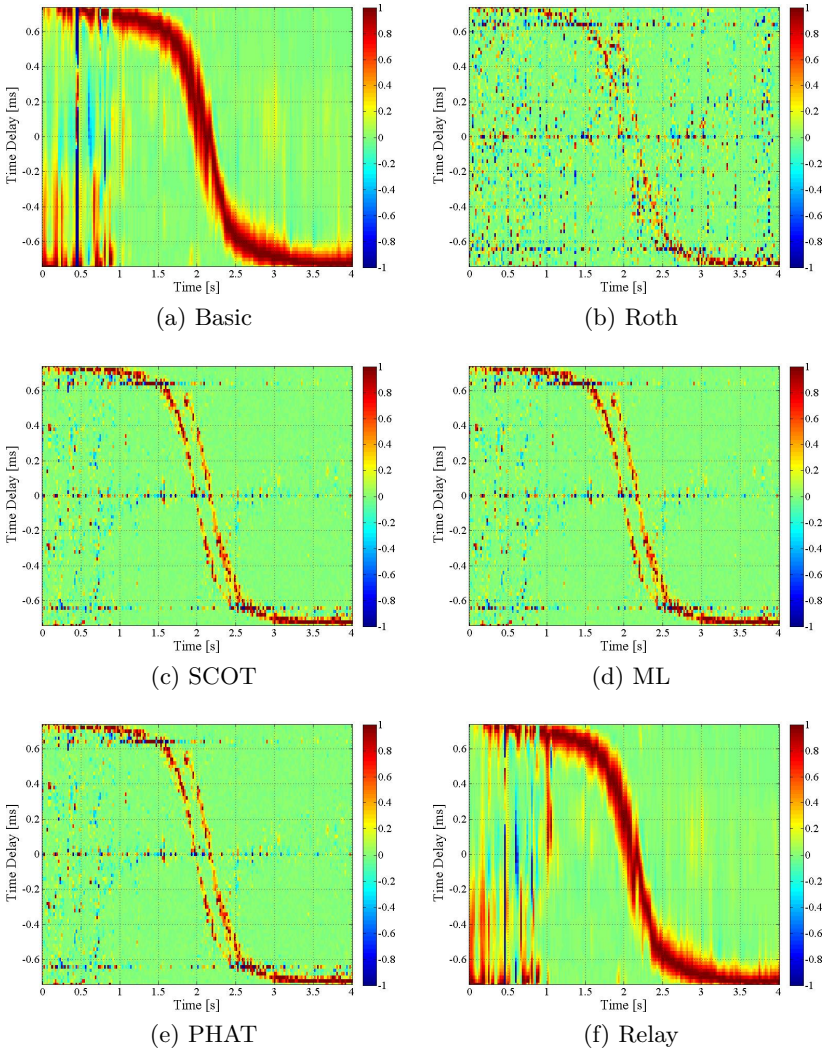
Figure 2.6: Example of a real CCTS over 30 seconds of signal.

typical graph whose shape is directly related to $\dot{x}$, $D$, and the inter-sensor distance $d$ as schematically explained by Fig. 2.5. This trace is bounded by $\pm\, d/c$ which theoretically corresponds to a wavefront coming from an angle of 180° or 0° respectively (*endfire* DOA). At the opposite, when $\tau$ is close to zero, namely when the source wavefront is captured at the same time instants on both microphones, meaning the vehicle is just in front of the array, the DOA is equal to 90° (*broadside* DOA).

Thus, similarly to spectrogram for time-frequency analysis, CCTS is a convenient tool for auditory scene visualisation. It allows the practitioner to count the number of vehicles present in a recording, to know their direction, to compare their speed and with a little practice, to discriminate vehicle types. An example of CCTS over a 30-second recording is depicted in Fig. 2.6. Nothing happens until the fifth second, then two vehicles follow one another (from left to right). At second 15, a vehicle is detected in the other lane (from right to left), followed by a motorbike between seconds 18 and 20 in the closer lane, and a last vehicle in the remote lane starting at second 23.

## 2.6 Comparison between different weighting functions

Fig. 2.7 depicts different CCTS, obtained on the same audio recording, but using different filters $\psi_g(f)$. The audio signal corresponds to the pass-by of an unknown vehicle moving at nearly 60 km/h, acquired by two sensors placed at a distance $D = 2.5$ m from the CPA with a sampling rate $f_s = 50$ kHz. Correlation measurements were performed on successive audio frames of size $N_s = 2048$ samples (41 ms) with an overlap of 75% (31 ms): the correlation measure was updated every 10 ms. The unitary weighting giving the basic cross-correlation time series is depicted in Fig. 2.7a. Other transforms introduced in section 2.4.2 are depicted from Fig. 2.7b to Fig. 2.7e.




(a) Basic (b) Roth

(c) SCOT (d) ML

(e) PHAT (f) Relay

Figure 2.7: Different cross-correlation time series (CCTS) using (a) classic, (b) Roth, (c) SCOT, (d) ML, (e) PHAT and (f) Relay weighting.

The unitary weight in Fig. 2.7a provides an image with high contrast but a low TDOA resolution in comparison with SCOT-CCTS, ML-CCTS and PHAT-CCTS, Fig. 2.7c,

2.7d and 2.7e respectively, where more spurious peaks appear but high resolution is time-delay is achieved. The two visible traces are representative of the front and rear axles trajectories, each producing a tyre/road noise dissociated in space. As a consequence, in addition to the vehicle speed, these GCCs make possible the estimation of the wheelbase through the separation between the two traces, except for the Roth processor which provides a very noisy image in the present application as expected from the theory. The relay cross-correlation give quite the same results as the basic cross-correlation in this case, and it is not as accurate as the PHAT, SCOT or ML-based methods. It has been finally decided to opt for the PHAT process in the reminder of the work.

## 2.7 Conclusion

In this chapter, key concepts in passive sound source localization have been presented with an emphasis on time-delay-based techniques using a pair of sensors. The scenario we are interested in involves wideband sound sources (i) monitored by a limited number of sensors (ii) in a non-reverberant (iii) and non-dispersive (iv) medium. Statements (i) and (ii) definitely discard traditional one-step procedures such as delay-and-sum beamforming or subspace-based methods. On the other hand, it has been shown that statements (iii) and (iv) turn the localization problem into a simpler time-delay estimation problem.

Some of the most successful time-delay estimators belong to the family of generalized cross-correlation functions. This family contains multiple members, characterized by different weighting functions. The best known ones have been described and compared on a real pass-by noise audio recording. In accordance with the existing literature, we found that the phase transform (PHAT) weighting is certainly the most relevant one because of its temporal resolution: it makes possible the observation of front and rear axles, its ease of implementation, its rapidity of execution, its efficiency with a limited number of sensors and, its robustness to model errors and weather conditions. In addition, we will see in chapter 4 that the analytic expression of the GCC-PHAT for the one source and multiple sources case can be derived.

The concept of cross-correlation time series (CCTS) has also been introduced. It consists in the concatenation in time of several correlation measures. CCTS of a pass-by noise produce a typical trace whose shape is related to the vehicle position, speed and aperture of the array. In particular, the concatenation of GCC-PHAT observations, giving a “PHAT-CCTS”, enables the observation of the trajectories of front and rear axles separately, and therefore a potential way of estimation of the wheelbase length of pass-by vehicles. This point is investigated in more detail in chapter 4, some experimental results are provided in chapter 6.

In the next chapter, we investigate how to exploit such cross-correlation based measurements to automatically estimate the speed of a moving sound source. In order to counteract possible interfering noises in the observation, the Bayesian theory is introduced.

# 3 Moving sound source detection and tracking

## 3.1 Introduction

The state of the art on moving sound source speed estimation through acoustic sensing extensively exploits the well-known *Doppler effect*, for instance in [28, 29, 72, 73, 74]. The Doppler effect explains the apparent change in the frequency of a wave caused by relative motion between the source of the wave and the observer. Consider the case when the source of sound moves with speed $v_s$ and emits a sound of frequency $f_0$, the frequency $f$ perceived by a static receiver is:

$$f = \frac{c}{c + v_s} f_0, \tag{3.1}$$

where $c$ is the speed of sound, and $v_s$ is positive or negative depending if the source is approaching or moving away from the receiver. This technique requires that the target ideally emit a pure tone wave in order to study its evolution in the time-frequency domain. But the stochastic nature of the pass-by noise makes the observation of the Doppler effect quite difficult using a single sensor. This aspect is exemplified in Fig. 3.1 where spectrograms of an fire truck pass-by noise and a classical car pass-by noise are depicted. Both signals were acquired at the same location (Lat. 46°36'27.22"N, Long. 6°32'34.38"E) in an interval of a few minutes using a single microphone on the roadside. The Doppler effect on the fire truck siren is clearly visible in Fig. 3.1a, but with a classic road car, Fig. 3.1b, extracting any Doppler information seems more challenging.

Using a pair of sensors allows one to study the *relative Doppler effect*. This is based on the fact that, when a moving sound source is recorded by two spatially distant sensors, the spectrum of one acquired signal is a stretched version of the other [75] p. 64. The degree of stretching is related to the vehicle speed. However, this technique requires a strongly shaped spectrum and gives poor performance for flat sound source spectrum. It is interesting to note that in developing countries, the extensive use of vehicular honks

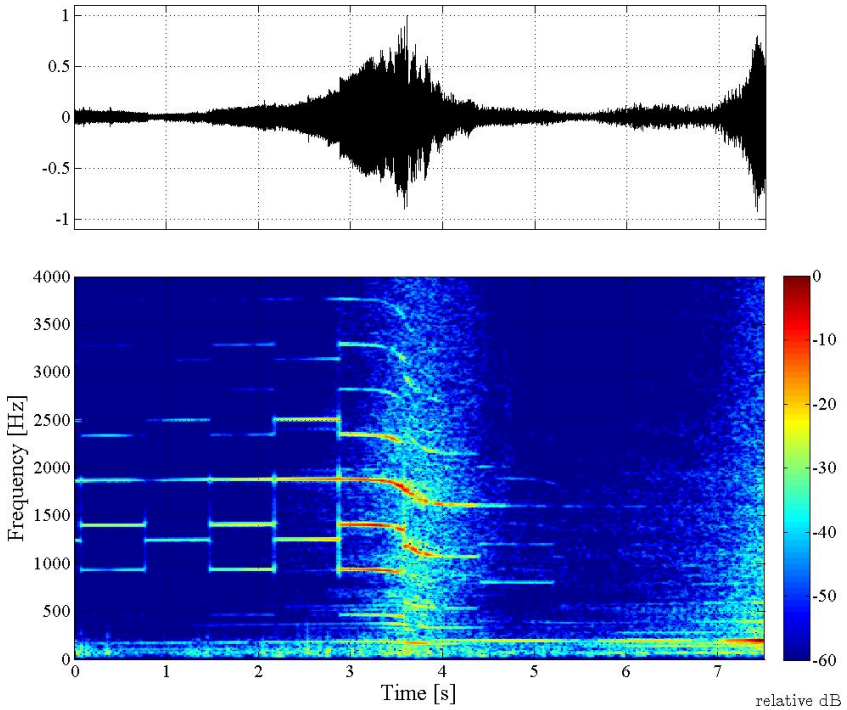
enable the vehicle speed estimation from honks differential Doppler shift. Sen. *et al.*, for instance, proposed such a system for the India network [76].

In the 2000s, maximum likelihood approaches were proposed for vehicle motion and size estimations by López-Valcarce *et al.* and Cevher *et al.*, the first using a pair of microphones on a 6.5 m tall pole, the second using a single microphone on the roadside [77, 30, 31, 32, 34, 35, 36]. As for Doppler-effect-based methods, such techniques require rather clean signals and the presence of multiple vehicles or interfering noises in the monitored area may limit their applicability.

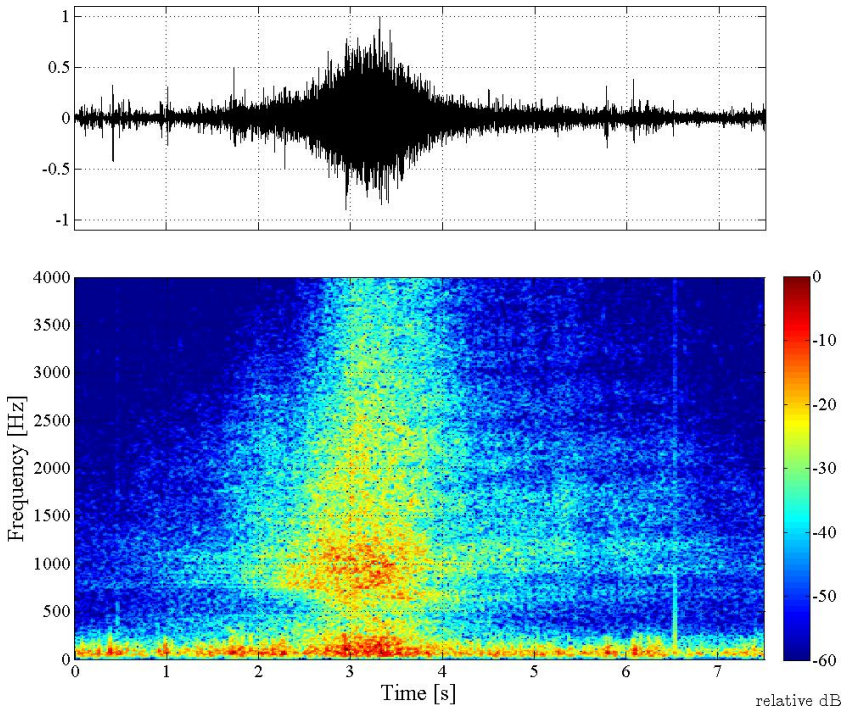
The proposed approach is inspired from works of S. Chen *et al.* [12] and J.F. Forren *et al.* [16] who in the mid of 1990s both independently showed the relevance of CCTS for *in-situ* road monitoring. In 2001, S. Chen *et al.* processed large-scale correlation measurements from the center of London, over six months from winter to spring, and proved the robustness of the cross-correlation-based methods against bad weather conditions [19]. But at this time, no automatized process was proposed to extract the motion parameters of vehicles. This is the point investigated in this chapter.

Whatever the procedure - one step or two steps - the result of a SSL estimator is a *localization function*, like CCTS in [19] or those previously depicted in Fig. 2.7. This function contains a mode (peak) whose the argument is - or is related - to the source position. When the source is moving, estimating its trajectory simply consists in concatenating successive SSL estimates by looking at the evolution of this argument using a peak picking procedure for instance. But in the real world, such a basic process can be strongly affected by a plethora of errors due to noise in the measurement procedure, mismatches between modeled and actual recordings, data missing due to an interruption of the observation, apparition of spurious modes due to acoustic phenomena unrelated to the source of interest etc. In particular, spurious peaks are big issues in the sense that their amplitude can be much greater than the peak due to the actual source, especially for measurements in environmental conditions or in a reverberant room, as pointed out in [78, 79, 80].

One solution therefore consists in dissociating “true” from “false” peaks by discriminating those that follow a well-established dynamical model from those which do not have any temporal consistency. That supposes to take into account all the previous observations to make the distinction between noise and signal at time $t$. This is the strong idea brought by Bayesian theory, forming the basis of most tracking algorithms, and that we propose to apply in the traffic flow monitoring context.




(a)




(b)

Figure 3.1: (a): pass-by noise of a fire truck, (b): pass-by noise of a standard car.

## 3.2 State-space model of a moving object

Dynamic systems are generally modeled by a system of equations, called a *state-space model*, in which the actual states $\alpha$ and their observations $\beta$ are related by:

$$\alpha_t = \mathscr{T}_t(\alpha_{t-1}, \mathbf{u}_t), \tag{3.2}$$

$$\beta_t = \mathscr{M}_t(\alpha_t, \mathbf{v}_t). \tag{3.3}$$

Eq. (3.2) is the transition equation or *dynamic model*. It describes the temporal evolution of the target state through the *transition function* $\mathscr{T}_t(.)$. Eq. (3.3) is the measurement equation or *observation model*. It describes the relationship between state and observation through the *measurement function* $\mathscr{M}_t(.)$. Both the transition and measurement functions are supposed to be known. The quantities $\mathbf{u}_t$ and $\mathbf{v}_t$ are respectively called *state noise* and *measurement noise*, independent from the states, the observations and from each other. They are described by known probability density functions PDF: $\mathbf{u}_t \sim p_u$ and $\mathbf{v}_t \sim p_v$. The state noise models the uncertainties one has on the actual dynamical characteristics and the measurement noise models the errors which may affect the measurement procedure.

Variables $\alpha_0, \alpha_1, ..., \alpha_t$ denote the *state vector* at times $0, 1, ..., t$. They are modeled by a first order Markov process (the present state depends only on the previous state). Due to the random noise $\mathbf{u}_t$ in the transition equation 3.2, the state at time $t$ is drawn from a *transitional prior distribution* linking past and present states. One can write:

$$\alpha_t | \alpha_{t-1} \sim p(\alpha_t | \alpha_{t-1}). \tag{3.4}$$

And the initial distribution is denoted $p(\alpha_0)$.

Variables $\beta_0, \beta_1, ..., \beta_t$ denote the available observations at times $0, 1, ..., t$. We assume that each observation $\beta_t$ depends only on the state $\alpha_t$, in other words, the $\beta_i$ are *conditionally independent provided that states are known.* As the state is generally not directly observable, it is qualified as "hidden". Generally, observations provide only partial information on the state.

The objective of tracking is to recursively estimate $\alpha_t$ out of the observations $\beta_t$. Knowledge of the two models are required to make inference about the dynamic system. In the statistical literature, the dynamic transition and observation models are both available in a probabilistic form. This is particularly convenient for the Bayesian approach and, in a sense, a more general and rigorous way for solving the problem [81].

## 3.3 The sequential Bayesian approach

The sequential Bayesian approach consists in recursively estimating the *posterior PDF* of the state vector each time a new observation is received, without having to reprocess

previous observations. The reason is that from a statistical point of view, the state to be estimated is precisely the argument of the maximal value of the posterior. In the filtering context, this is denoted $p(\alpha_t|\beta_{1:t})$ where $\beta_{1:t}$ stands for all measurements acquired until time $t$[1]. The posterior is also known as *filtering distribution*. Unfortunately, the posterior density is unavailable in practice. However, according to Bayes' theory, if this law is known at time $t$-1, one can find that at time $t$ through a *prediction* step and an *update* step. Assuming that the initial PDF of the state vector $p(\alpha_0|\beta_0) = p(\alpha_0)$ is available and under standard assumptions (first order dynamical model, conditional independence of observations given the states), the equations to solve are [82]:

$$p(\alpha_t|\beta_{1:t-1}) = \int p(\alpha_t|\alpha_{t-1})p(\alpha_{t-1}|\beta_{1:t-1})d\alpha_{t-1}, \tag{3.5}$$

$$p(\alpha_t|\beta_{1:t}) = \frac{p(\beta_t|\alpha_t)p(\alpha_t|\beta_{1:t-1})}{\int p(\beta_t|\alpha_t)p(\alpha_t|\beta_{1:t-1})d\alpha_t}. \tag{3.6}$$

In the prediction step, Eq. (3.5), the dynamic model $p(\alpha_t|\alpha_{t-1})$ is used to propagate the posterior distribution $p(\alpha_{t-1}|\beta_{1:t-1})$ at time $t$-1 to provide the predictive distribution $p(\alpha_t|\beta_{1:t-1})$. This is the Chapman-Kolmogorov equation.

In the update step, Eq. (3.6), the predictive distribution is combined with the likelihood $p(\beta_t|\alpha_t)$ to obtain the new posterior distribution $p(\alpha_t|\beta_{1:t})$ at time $t$. This is Bayes' rule.

The state estimation at time $t$ entails different aspects, depending on the observations used:

- prediction : observations available from time 0 to time $t$-$m$ ($m > 0$);
- filtering : observations available from time 0 to time $t$;
- smoothing : observations available from time 0 to time $t$+$m$ ($m > 0$).

This thesis focuses on the filtering aspects. Correlation measurements are available until time $t$ and the objective is to perform position, speed and wheelbase estimations as vehicles pass by.

In principle, both an optimal estimate of the state (with respect to any criterion) and a measure of the accuracy of the estimate may be obtained from the posterior distribution. The recurrence relations (3.5) and (3.6) form the basis of the optimal Bayesian solution for recursive filtering [81]. But this recursive propagation of the posterior density is only a conceptual solution. This is because, in general, it cannot be determined analytically, except in a restrictive set of cases in which solutions do exist and can be handled by an *optimal* filter (*e.g.* Kalman or grid-based filters). In all other cases, the solution is approximated by a *suboptimal* filter (*e.g.* extended, unscented Kalman filters, particle filter). These algorithms are described hereafter.

[1]Note that in this document, notions of *observation* and *measurement* are not dissociated to simplify notation.

## 3.4 Optimal filters

### 3.4.1 Kalman Filter

The Kalman filter (KF) [83] is the optimal solution for the filtering problem if the following assumptions hold:

- $\mathscr{T}_t(\alpha_{t-1}, \mathbf{u}_t)$ is known and linear with respect to $\alpha_{t-1}$ and $\mathbf{u}_t$;
- $\mathscr{M}_t(\alpha_t, \mathbf{v}_t)$ is known and linear with respect to $\alpha_t$ and $\mathbf{v}_t$;
- $\mathbf{u}_{t-1}$ and $\mathbf{v}_t$ are drawn from known Gaussian distributions.

In such conditions, Eq. (3.2) and Eq. (3.3) can be rewritten as:

$$\alpha_t = T_t\alpha_{t-1} + \mathbf{u}_t, \tag{3.7}$$
$$\beta_t = M_t\alpha_t + \mathbf{v}_t, \tag{3.8}$$

where the initial state $\alpha_0$ is Gaussian with mean $\hat{\alpha}_0$ and covariance $P_0$, denoted $p(\alpha_0) = \mathscr{N}(\alpha_0; \hat{\alpha}_0, P_0)$ in the following. $\mathbf{u}_t$ and $\mathbf{v}_t$ are statistically independent and their covariance is respectively denoted $\Sigma_{u,t}$ and $\Sigma_{v,t}$. Noise parameters $\Sigma_{u,t}$ and $\Sigma_{v,t}$, state matrix $T_t$ and measurement matrix $M_t$ may be time dependent. Because of linearity, $\alpha_t$ and $\beta_t$ are Gaussian:

$$p(\alpha_{t-1}|\beta_{1:t-1}) = \mathscr{N}(\alpha_{t-1}; \hat{\alpha}_{t-1|t-1}, P_{t-1|t-1}), \tag{3.9}$$
$$p(\alpha_t|\beta_{1:t-1}) = \mathscr{N}(\alpha_t; \hat{\alpha}_{t|t-1}, P_{t|t-1}), \tag{3.10}$$
$$p(\alpha_t|\beta_{1:t}) = \mathscr{N}(\alpha_t; \hat{\alpha}_{t|t}, P_{t|t}), \tag{3.11}$$

where $\hat{\alpha}_{t-1|t-1}$ (respectively $P_{t-1|t-1}$) denotes the mean state value (respectively covariance) at time $t-1$ and $\hat{\alpha}_{t|t-1}$ (respectively $P_{t|t-1}$) denotes the predicted mean state value (respectively predicted covariance). From the recursion (3.9), the posterior to estimate is totally characterized by its two first moments. They are obtained using the following equations:

$$\hat{\alpha}_{t|t-1} = T_t\hat{\alpha}_{t-1|t-1} + \mathbf{u}_t, \tag{3.12}$$
$$P_{t|t-1} = T_tP_{t-1|t-1}T_t^T + \Sigma_{u,t-1}, \tag{3.13}$$
$$\tag{3.14}$$

and

$$\hat{\alpha}_{t|t} = \hat{\alpha}_{t|t-1} + K_t(\beta_t - M_t\hat{\alpha}_{t|t-1}), \tag{3.15}$$
$$P_{t|t} = P_{t|t-1} - K_tM_tP_{t|t-1}, \tag{3.16}$$
$$S_t = M_tP_{t|t-1}M_t^T + \Sigma_{v,t}, \tag{3.17}$$
$$K_t = P_{t|t-1}M_t^TS_t^{-1}. \tag{3.18}$$

$S_t$ is the covariance of the *innovation* term $\beta_t - M_t\hat{\alpha}_{t|t-1}$ and $K_t$ is the Kalman gain. From the above equations, if the Kalman gain increases as the covariance matrix of measurement noise $\Sigma_{v,t}$ tends to the null matrix, measurement is favored with respect to prediction. On the opposite, if the predicted covariance approaches zero, then the gain approaches zero too and prediction is favored with respect to measurement. In other words, the Kalman gain decides what is the "weight" of the measurement in the new state estimate.

### 3.4.2 Grid-based methods

Considering the same assumptions as in KF, grid-based methods (GBM) provide the exact posterior density if the state space is discrete and finite [81]. GBM does not intend to propagate the two first moments as KF, but to estimate directly the posterior $p(\alpha_t|\beta_{1:t})$ - which is the primary objective - with a deterministic grid of the state space.

Let $\alpha^i_{t-1}$, $i = 1, ..., N$ be the discrete states constituting the state space at time $t$-1 and $w^i_{t-1|t-1}$ their associated conditional probability, given measurements up to time $t$-1, that is:

$$w^i_{t-1|t-1} = p(\alpha_{t-1} = \alpha^i_{t-1}|\beta_{1:t-1}). \tag{3.19}$$

Then, the posterior PDF at $t$-1 becomes:

$$p(\alpha_{t-1}|\beta_{1:t-1}) = \sum_{i=1}^{N} w^i_{t-1|t-1}\delta(\alpha_{t-1} - \alpha^i_{t-1}), \tag{3.20}$$

where $\delta(.)$ is the Dirac delta function. Substituting (3.20) into (3.5) and (3.6) gives new prediction and update equations:

$$p(\alpha_t|\beta_{1:t-1}) = \sum_{i=1}^{N} w^i_{t|t-1}\delta(\alpha_t - \alpha^i_t), \tag{3.21}$$

$$p(\alpha_t|\beta_{1:t}) = \sum_{i=1}^{N} w^i_{t|t}\delta(\alpha_t - \alpha^i_t), \tag{3.22}$$

where

$$w^i_{t|t-1} \triangleq \sum_{j=1}^{N} w^j_{t-1|t-1}p(\alpha^i_t|\alpha^j_{t-1}), \tag{3.23}$$

$$w^i_{t|t} \triangleq \frac{w^i_{t|t-1}p(\beta_t|\alpha^i_t)}{\sum_{j=1}^{N} w^j_{t|t-1}p(\beta_t|\alpha^j_t)}. \tag{3.24}$$

The above assumes that $p(\alpha^i_t|\alpha^j_{t-1})$ and $p(\beta_t|\alpha^j_t)$ are known without any restriction on their form.

KF and GBM are said optimal in the sense that they minimize the variance of the estimates for linear cases. But as highlighted in the previous chapter the state of interest (abscissa of the sound source) is a non linear function of the observation (TDOA). As a consequence, both optimal methods presented can not handle the *bearing-only target tracking* problem properly [82, 84, 85, 86]. Hence, suboptimal, but more adapted techniques have been deployed and the three most famous of them are reviewed below.

## 3.5 Suboptimal filters

### 3.5.1 Extended Kalman filter

The Extended Kalman Filter (EKF) handles the case where $\mathscr{T}_t$ and/or $\mathscr{M}_t$ are nonlinear. That is the Gaussianity of the posterior is not ensured anymore, and consequently, not totally characterized by its first two moments. The key idea consists in deriving a first-order Taylor expansion to locally linearize $\mathscr{T}_t$ and $\mathscr{M}_t$ around an estimate of the current mean and covariance. As a result, the EKF provides the optimal linear, or Linear Minimum Mean Square Error (LMMSE) solution [87].

$$
\begin{aligned}
p(\alpha_{t-1}|\beta_{1:t-1}) &\approx \mathscr{N}(\alpha_{t-1}; \hat{\alpha}_{t-1|t-1}, P_{t-1|t-1}), && (3.25)\\
p(\alpha_t|\beta_{1:t-1}) &\approx \mathscr{N}(\alpha_t; \hat{\alpha}_{t|t-1}, P_{t|t-1}), && (3.26)\\
p(\alpha_t|\beta_{1:t}) &\approx \mathscr{N}(\alpha_t; \hat{\alpha}_{t|t}, P_{t|t}), && (3.27)
\end{aligned}
$$

where:

$$
\begin{aligned}
\hat{\alpha}_{t|t-1} &= \mathscr{T}_t(\hat{\alpha}_{t-1|t-1}) + \mathbf{u}_t, && (3.28)\\
P_{t|t-1} &= \hat{T}_t P_{t-1|t-1} \hat{T}_t^T + \Sigma_{u,t-1}, && (3.29)\\
\hat{\alpha}_{t|t} &= \hat{\alpha}_{t|t-1} + K_t(\beta_t - \mathscr{M}_t(\hat{\alpha}_{t|t-1}) - \mathbf{v}_t), && (3.30)\\
P_{t|t} &= P_{t|t-1} - K_t \hat{M}_t P_{t|t-1}, && (3.31)\\
& && (3.32)
\end{aligned}
$$

and

$$
\begin{aligned}
\hat{T}_t &= \left.\frac{d\mathscr{T}_t(x)}{dx}\right|_{x=\hat{\alpha}_{t-1|t-1}}, && (3.33)\\
\hat{M}_t &= \left.\frac{d\mathscr{M}_t(x)}{dx}\right|_{x=\hat{\alpha}_{t|t-1}}, && (3.34)\\
S_t &= \hat{M}_t P_{t|t-1} \hat{M}_t^T + \Sigma_{v,t}, && (3.35)\\
K_t &= P_{t|t-1} \hat{M}_t^T S_t^{-1}. && (3.36)
\end{aligned}
$$

The same kind of equations may be written for higher-order linearizations. The higher the order, the better the results but also the higher the computation complexity so order

1one is widespread used one. For the last 30 years, EKF has been a standard Bayesian state-estimation algorithm for nonlinear systems [88] but, despite its wide use, EKF is not suitable in case of too strong non-linearities or highly non-Gaussian conditional PDFs. Moreover, if the functions governing the system are not differentiable, the implementation of the Jacobian is impossible. In order to deal with highly non-linear cases, another approach has been proposed in the 1990s, called the Unscented Kalman Filter.

### 3.5.2 Unscented Kalman filter

The Unscented Kalman Filter (UKF) is based on the idea that it is easier to approximate a Gaussian by using a cloud of points rather than linearizing a function [89]. Therefore, no calculations of Jacobians are required, and the posterior density is represented by a set of deterministically chosen points called *sigma points*. These points totally estimate the mean and covariance of the posterior given the real non-linear transition and measurement functions and the prior mean and covariance. Once again, the posterior is supposed to be Gaussian so that equations (3.25), (3.26) and (3.27) should remain valid.

The *unscented transform* is a method for calculating the statistics of a random variable that undergoes a nonlinear transformation. It returns a set of 2N sigma points $\zeta_t^{(n)}$ with corresponding weights $W_i$ given a state vector $\hat{\alpha}$ of length $N$ such that [89]:

$$\begin{aligned}
\zeta^{(0)} &= \hat{\alpha} && (3.37)\\
\zeta^{(n)} &= \hat{\alpha} + \left(\sqrt{(N+\kappa)P}\right)_n \ , n = 1, ..., N && (3.38)\\
\zeta^{(n)} &= \hat{\alpha} - \left(\sqrt{(N+\kappa)P}\right)_n \ , n = N+1, ..., 2N && (3.39)\\
W^{(0)} &= \kappa/(N+\kappa) && (3.40)\\
W^{(n)} &= 1/(2(N+\kappa)), i = 1, ..., 2N. && (3.41)\\
& && (3.42)
\end{aligned}$$

where $\kappa$ is a scaling parameter which determines the spread of the sigma-points distribution around $\hat{\alpha}$ and $\left(\sqrt{(N+\kappa)P}\right)_n$ is the n$^{th}$ row or column of the matrix square root of $(N+\kappa)P$. $W^{(n)}$ is the weight corresponding to the n$^{th}$ sigma point such that $\sum_{n=0}^{2N} W^{(n)} = 1$. Sigma points are then propagated through the real nonlinear transition function:

$$\zeta_{t|t-1}^{(n)} = \mathscr{T}_t(\zeta_{t-1}^{(n)}) \ , n = 0, ..., 2N \qquad (3.43)$$

The statistics (mean and covariance) of the state vector at time $t$ are estimated as follows:

$$\hat{\alpha}_{t|t-1} = \sum_{n=0}^{2N} W^{(n)} \zeta_{t|t-1}^{(n)}, \tag{3.44}$$

$$P_{t|t-1} = \sum_{n=0}^{2N} W^{(n)} [\zeta_{t|t-1}^{(n)} - \hat{\alpha}_{t|t-1}][\zeta_{t|t-1}^{(n)} - \hat{\alpha}_{t|t-1}]^T. \tag{3.45}$$

The predicted measurement is then given by:

$$\hat{\beta}_{t|t-1} = \sum_{n=0}^{2N} W^{(n)} \zeta_{t|t-1}^{(n)}, \tag{3.46}$$

$$P_{t|t-1} = \sum_{n=0}^{2N} W^{(n)} [\zeta_{t|t-1}^{(n)} - \hat{\alpha}_{t|t-1}][\zeta_{t|t-1}^{(n)} - \hat{\alpha}_{t|t-1}]^T. \tag{3.47}$$

UKF appears to perform better compared to EKF in cases of higher non-linearities, in terms of state estimation and robustness to noise measurement [90]. But as EKF, UKF always approximates $p(\alpha_t|\beta_{1:t})$ to be Gaussian. Even if UKF is able to approximate heavy-tailed distribution better than EKF [87], this may be a restrictive assumption in the real world; a more critical point is this assumption does not permit these methods to estimate the posterior PDF if the is multi-modal (presence of several modes to track). As no assumption on the distribution of the prior or of linearity cab be made for the problem at hand, the particle filtering has been investigated in this work.

### 3.5.3 Particle filter

The particle filter (PF), also called Sequential Monte Carlo (SMC) method, is a nonparametric filter, in the sense that no *a priori* functional form of the posterior is required. The PF is quite similar to the UKF in that they both generate points about the mean estimate but in the case of UKF, the sampling of the sigma points is deterministic, while in PF the "particles" are randomly distributed. Hence, PF is effective under the following (non-restrictive) assumptions:

- $\mathscr{T}_t(\alpha_{t-1}, \mathbf{u}_t)$ is known and may be non-linear with respect to $\alpha_{t-1}$ and $\mathbf{u}_t$;
- $\mathscr{M}_t(\alpha_t, \mathbf{v}_t)$ is known and may be non-linear with respect to $\alpha_t$ and $\mathbf{v}_t$;
- $\mathbf{u}_{t-1}$ and $\mathbf{v}_t$ are independent stochastic processes and their distribution are not necessarily Gaussian.

The idea behind PF consists in representing the posterior density as a finite summation of Dirac distributions at points called *particles* (or state hypotheses), $\alpha_t^{(1)}$, $\alpha_t^{(2)}$,..., $\alpha_t^{(N_p)}$, weighted by coefficients called *weights*, $w_t^{(1)}$, $w_t^{(2)}$,..., $w_t^{(N_p)}$, $N_p$ being the number of

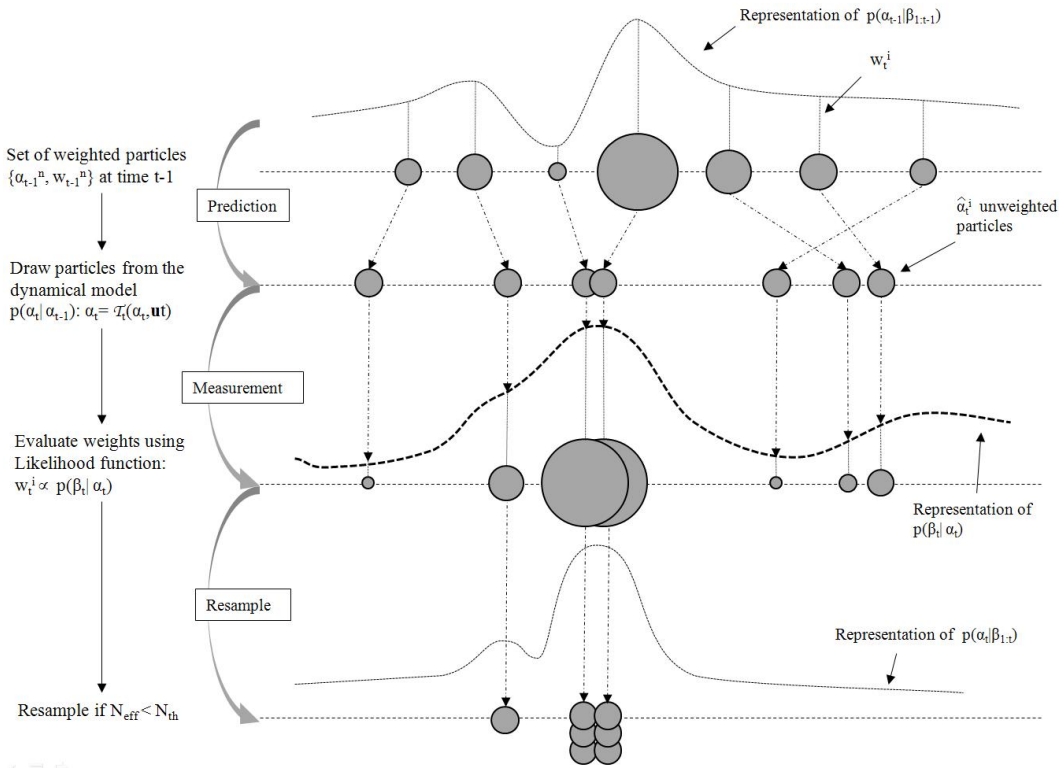



Figure 3.2: Generic particle filter algorithm.

particles, such that:

$$p(\alpha_t|\beta_{1:t}) \approx \sum_{n=1}^{N_p} w_t^{(n)} \delta(\alpha_t - \alpha_t^{(n)}), \tag{3.48}$$

The way to choose the weights is a crucial point in PF design and is the topic of many theoretical papers. The optimal solution is given in [91] but is very difficult, if not impossible, to obtain in practice. A very common, intuitive and simple way of updating weights is:

$$w_t^{(n)} \propto w_{t-1}^{(n)} p(\beta_t|\alpha_t^{(n)}), \tag{3.49}$$

which expresses that the new weights depend on the old weights and the new particle position.

From (3.49), the better a particle matches with the observation, the heavier its weight. Replacing $w_t^{(n)}$ by its expression in (3.48) permits to recursively update the posterior, and then to estimate the current state by looking at the mean or mode of the posterior.

The generic PF algorithm is summarized in algorithm 1. The various steps are also graphically represented in Fig. 3.2, inspired from a graph in [88].

One well-known problem with PF is that particles may quickly degenerate so that a single particle dominates after few iterations. This is called the *degeneracy effect*. To

**Algorithm 1** Algorithm of the generic particle filter.

**Initialisation**

- Initialize the particles from a Gaussian distribution around the *a priori* state vectors: $\alpha_0 \sim p(\alpha_0)$;
- Attribute the same weight to all particles: $\forall n \in [1, 2, ..., N_p]$, $w_0^{(n)} = 1/N_p$ ;

**For** $t = 1, 2, ...$
*Prediction*

- Predict the new set of particles by propagating the last set according to the dynamical source model: $\alpha_t \sim p(\alpha_t|\alpha_{t-1})$;

*Update*

- Weight the new particles: $\tilde{w}_t^{(n)} = w_t^{(n)} p(\beta_t|\alpha_t^{(n)})$, where $p(\beta_t|\alpha_t^{(n)})$ is the conditional likelihood of the observation obtained from raw data;
- Normalize the weights: $\forall n \in [1, 2, ..., N_p]$, $w_t^{(n)} = \tilde{w}_t^{(n)} / \sum_{n=1}^{N_p} \tilde{w}_t^{(n)}$;

*Resampling*

- Calculate $N_{eff}$ using (3.50);
- If $N_{eff} < N_{th}$, resample the particles according to their weights;

**endfor**
*Output of the algorithm*

- Estimate the posterior using Eq. (3.48);
- Deduce the current state $\alpha_t$ (mean or mode of the posterior).

counteract it, the updated, weighted particles can be *resampled* to yield a new set of equally weighted points. Inversely, if the resampling step is systematic, all weights remain equal and no convergence occurs. This is called the *sample impoverishment problem*. Traditionally, the resampling step is executed only when the degeneracy is too important. A suitable measure of the degeneracy is the effective sample size (ESS) introduced in [92], which can be estimated as [93]:

$$N_{eff} = \frac{1}{\sum_{n=1}^{N_p} (w_t^{(n)})^2}. \tag{3.50}$$

$N_{eff}$ takes values between 1 and $N_p$. When the ESS is below a threshold $N_{th}$ such as $N_p/2$ , the resampling procedure is activated. There are a number of algorithms for performing resampling: multinomial, stratified, systematic [94]. The multinomial resampling method is the simplest approach and the one used in this work. Its implementation may be found

in [94], section 2.1.

For more theoretical details about PF, one can advise the excellent tutorials [81, 95, 96, 97] and also [98] for French readers.

**Remark** The methods presented above have not been extensively compared by us during this thesis, for mainly two reasons: the first one is that an extensive literature points in favor of the PF method over Kalman-based ones for bearing-only tracking problems, for instance in [99, 100]; secondly the multimodal nature of the target itself, described in more detail in chapter 4, imposed us to rely on PF because it is the only method, among those presented, which handles the multiple-peaks tracking in a relatively simple and intuitive way.

## 3.6 An experimental measurement in semi-anechoic conditions

As a proof of concept, a PF algorithm has been designed and applied to audio recordings coming from an *in-lab* experiment in semi-anechoic conditions[2].

The setup consists of a small loudspeaker mounted on a slot car and running at a constant speed of $\dot{x}$=3.5 m/s on a linear path. Two microphones separated by $d$=56 cm, both placed at $D$=82 cm from the track, measure the sound pressure generated by the car passing by. A pair of infrared diodes (emitter and receiver) is placed on each side of the track, facing each other, to detect the slot car at the beginning and at the end of the track, to initialize and stop the tracking. The first pair of diodes is placed more than one meter after the starting line of the slot car to ensure that the acceleration phase of the car is over when particles are launched. The mobile speaker is fed with a white noise and the acquisition is done at a sampling rate of $f_s$= 50 kHz. A schematic representation of the setup is proposed in Fig. 3.3.

### 3.6.1 Target model

The target (or state) model $\alpha_t$ is the abstract representation of the object we are interested in. The slot car is modeled here by an active point emitter in the x-y plane moving with a constant speed on the $x$ axis. Therefore, the state vector $\alpha_t$ of the target at time $t$ is composed of three parameters which are the abscissa $x_t$, the ordinate $y_t$ and the speed $\dot{x}_t$:

$$\alpha_t = [x_t, y_t, \dot{x}_t]^T. \tag{3.51}$$

[2]Data have been kindly provided by Dr. Meritxell Genesca i Francitorra which carried out this measurement during her PhD thesis published in 2008 [75] in the Acoustic and Mechanical Engineering Laboratory (LEMA) of the Universitat Politècnica de Catalunya (UPC).

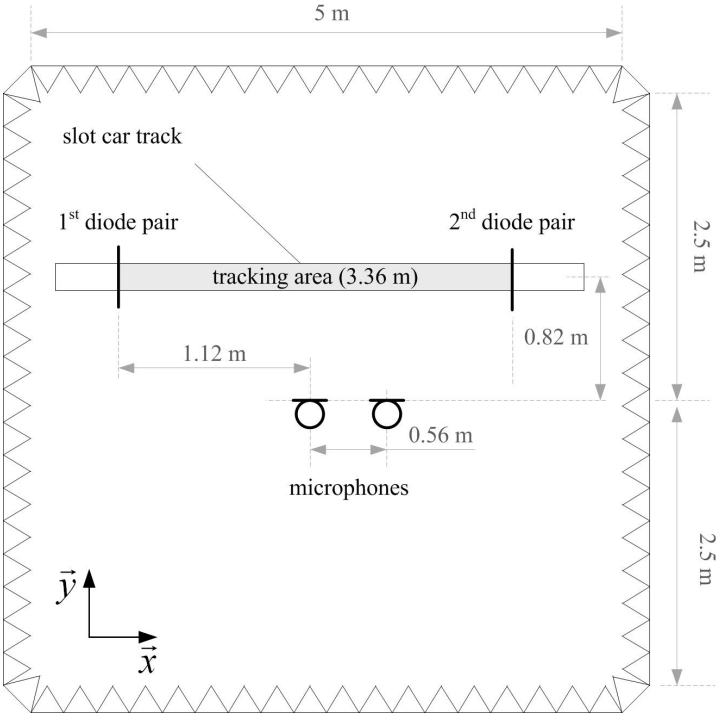



Figure 3.3: Scheme showing the semi-anechoic room dimensions (in m), the slot track and the microphone array.

### 3.6.2 Dynamical model

The dynamical model $p(\alpha_t|\alpha_{t-1})$ governs the temporal evolution of the state, that is, the mathematical relation between state vectors taken at successive times $t-1$ and $t$.

In the example considered, the sound source is expected to move at a constant speed on the $x$ axis. A Gaussian noise is added to model the possible speed variations of the target, giving:

$$p(\alpha_t|\alpha_{t-1}) = \mathcal{N}\left(\mathbf{F}\alpha_{t-1}, \mathbf{V}\right), \tag{3.52}$$

where the prediction matrix $\mathbf{F}$ and the noise covariance $\mathbf{V}$ are given by:

$$\mathbf{F} = \begin{pmatrix} 1 & 0 & \Delta T \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } \mathbf{V} = \begin{pmatrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_y^2 & 0 \\ 0 & 0 & \sigma_{\dot{x}}^2 \end{pmatrix}$$

,

and where $\Delta T$ is the time interval between two successive observations and $\sigma_x$ (respectively $\sigma_y$ and $\sigma_{\dot{x}}$) denotes the standard deviation (STD) of the zero-mean noise added to the state $x_t$ (respectively $y_t$ and speed $\dot{x}_t$).

### 3.6.3 Likelihood model

The likelihood model $p(\beta_t|\alpha_t)$ measures the adequacy of the data given the proposed configuration of the tracked object. This is the core of the PF algorithm as it discriminates good from bad particles, that is, particles that well explain the observation from those which do not. The likelihood model is determined by the practitioner and depends on the kind of available measurements.

Consider a particle $n$ with coordinates $\mathbf{r}_n^p = [x_t^{(n)}, y_t^{(n)}]^T$ at time $t$. The current TDOA $\tau_{12,t}^{(n)}$ of an hypothetic wavefront coming from the $n^{th}$ particle position is given by the relation:

$$\tau_{12,t}^{(n)} = \frac{d}{c} \sin\left(\arctan\left(\frac{x_t^{(n)}}{y_t^{(n)}}\right)\right), \tag{3.53}$$

where $c$ is the speed of sound and $d$ is the inter-sensor distance. What we propose is to consider the correlation measure at time lag $\tau_{12,t}^{(n)}$ as the likelihood of the particle $n$, such that:

$$p(\beta_t|\alpha_t^{(n)}) = R_{s_1 s_2}^{phat}\left(\tau_{12,t}^{(n)}\right). \tag{3.54}$$

In essence, candidate positions with the highest cross-correlation measures are the most likely candidates.

### 3.6.4 Initialisation and stopping conditions

Initialisation and stopping conditions are the two rules governing the birth and death of particles.

In this example, tracking begins when the slot car is detected by the first diode pair. At this instant, the initial state vector is drawn from a Gaussian distribution and initial weights are all equal and normalized, such that, for $n \in [1, 2, ..., N_p]$:

$$\alpha_0^{(n)} \sim \mathcal{N}\left(\begin{bmatrix} \mu_{x,0} \\ \mu_{y,0} \\ \mu_{\dot{x},0} \end{bmatrix}, \begin{bmatrix} \sigma_{x,0}^2 & 0 & 0 \\ 0 & \sigma_{y,0}^2 & 0 \\ 0 & 0 & \sigma_{\dot{x},0}^2 \end{bmatrix}\right), \tag{3.55}$$

$$w_0^{(n)} = \frac{1}{N_p}, \tag{3.56}$$

where the means $\mu_{.,0}$ denotes the *a priori* knowledge of the object state vector, and the noise variances $\sigma_{.,0}^2$ denotes the uncertainty in this knowledge.

As soon as the second diode pair detects the slot car, the algorithm is stopped.

### 3.6.5 Experiments

Two experiments are carried out. For both, the Cartesian position of the target is supposed to be almost perfectly known at initialisation as it is delivered by the first

| Actual states | Part. mean ($t = 0$) | Part. STD ($t = 0$) | State noise STD ($t > 0$) |
|---|---|---|---|
| $x_0$ = -1.96 m | $\mu_{x,0}$ = -1.96 m | $\sigma_{x,0}$ = 0.4 m | $\sigma_x = 1e^{-2}$ m |
| $y$ = 0.82 m | $\mu_{y,0}$ = 0.82 m | $\sigma_{y,0} = 1e^{-2}$ m | $\sigma_x = 5e^{-4}$ m |
| $\dot{x}$ **= 3.5 m/s** | $\mu_{\dot{x},0}$ **= 5 m/s** | $\sigma_{\dot{x},0}$ = 3 m/s | $\sigma_{\dot{x}} = 1e^{-4}$ m/s |

Table 3.1: Parameters of the particle filter for the first experiment: the speed *a priori* is higher than the actual one.

| Actual states | Part. mean ($t = 0$) | Part. STD ($t = 0$) | State noise STD ($t > 0$) |
|---|---|---|---|
| $x_0$ = -1.96 m | $\mu_{x,0}$ = -1.96 m | $\sigma_{x,0}$ = 0.4 m | $\sigma_x = 1e^{-2}$ m |
| $y$ = 0.82 m | $\mu_{y,0}$ = 0.82 m | $\sigma_{y,0} = 1e^{-2}$ m | $\sigma_x = 5e^{-4}$ m |
| $\dot{x}$ **= 3.5 m/s** | $\mu_{\dot{x},0}$ **= 2 m/s** | $\sigma_{\dot{x},0}$ = 3 m/s | $\sigma_{\dot{x}} = 1e^{-4}$ m/s |

Table 3.2: Parameters of the particle filter for the second experiment: the speed *a priori* is lower than the actual one.

diode pair, *i.e.* $\mu_{x,0}$ and $\mu_{y,0}$ from Eq.3.55 are close to reality and the uncertainties on the initial coordinates $\sigma^2_{x,0}$ and $\sigma^2_{y,0}$ are set rather low. However, the *a priori* speed $\mu_{\dot{x},0}$ overestimates reality in the experiment 1, Table 3.1, and underestimates reality in the experiment 2, Table 3.2. The objective is of course to determine if the PF retrieves the actual speed $\dot{x}$ well through the PHAT-CCTS image.

### 3.6.6 Results

One run per experiment is depicted in Fig. 3.4. As the same pair of recordings is used in both cases, the observations (CCTS) are actually the same as depicted in Fig. 3.4a and Fig. 3.4c. For each case, the observation is confronted to the *a priori* state model represented by a black dashed line, that is, the theoretical evolution of the TDOA as a function of time if the initial conditions were actually true. *A priori* speed (black dashed line) and actual speed (black full line) are also confronted in Fig. 3.4b and Fig. 3.4d. In each picture, the 95% confidence interval (CI95) of the particle states (TDOA or speed) are represented by red dashed lines.

On these examples, particles successfully converge towards the actual state vector relatively quickly. The sound source is in the broadside DOA at nearly 0.45 seconds. Before this time, particles look for the actual speed and after it refine the estimation and reduce their CI95.

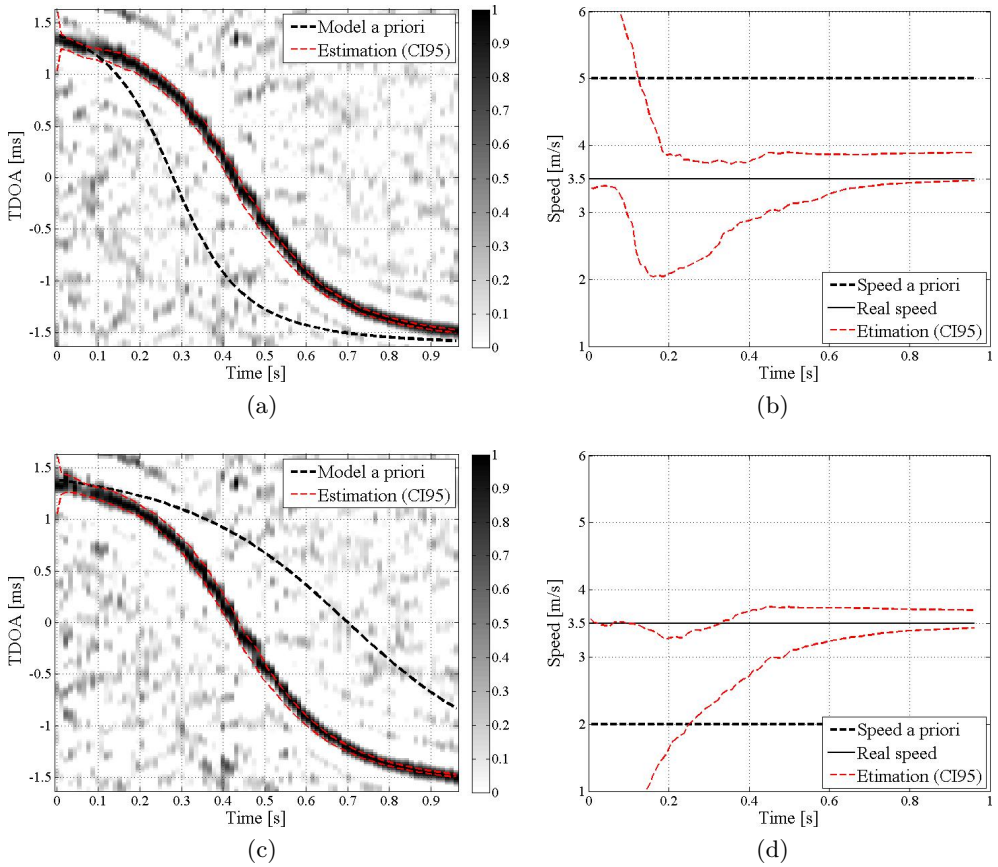



Figure 3.4: Experimental result of a particle filtering algorithm in semi-anechoic conditions. On the left: confrontation between the observation (black and gray) and the third state (speed) particles trajectories (CI95 of the cloud represented by red lines) initialized with a false *a priori* model (black dashed line). On the right: evolution of the particle distribution, (CI95 of the cloud represented by red lines) initialized with an *a priori* value (black dashed line) quite different from the actual one (full black line).

## 3.7 The detection problem

A specificity of the application targeted in this work is that the number of targets (vehicles) to monitor is unknown and may be larger than one. Moreover, according to preliminary developments and experimental measurements discussed in chapters 4 and 6, a pretty good knowledge of the initial position of each target is required to ensure good tracking performances (*i.e.* precision and accuracy).

Dealing with multiple targets at the same time is called a multiple target tracking (MTT) problem. If the number of targets is unknown, particle-filter-based methods, or any other Bayesian technique, are not reliable solutions [101]. But a workaround consists in turning the MTT problem into a single target tracking one by launching in parallel as many particle clouds as the number of sources, each cloud evolving independently from the others. This solution is the one that has been retained in this work. It requires

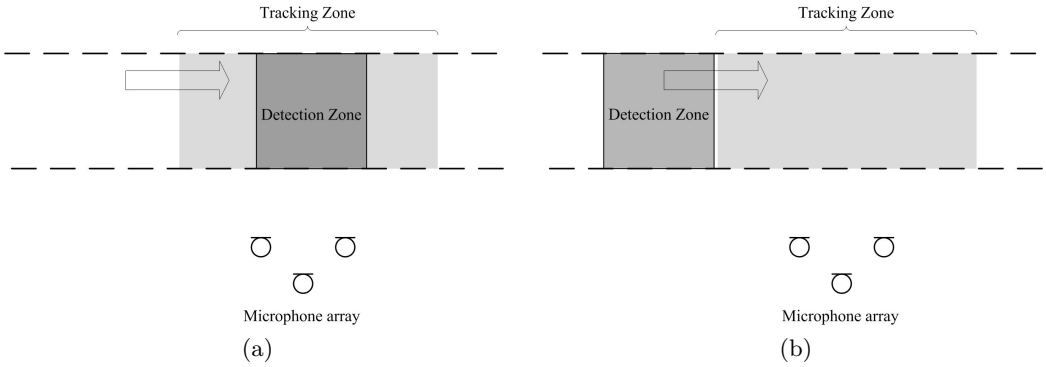



Figure 3.5: Two proposed detection strategies. In (a), the detection zone is in the broadside direction as well as the tracking zone (broadside detection strategy), in (b), the detection zone is upstream the tracking zone (endfire detection strategy).

a detection step answering the two questions: is there any new target to track now (*i.e.* in the current audio frame) ? If yes, what is its position ? During this thesis, the detection problem has been considered as totally separated from the tracking one. For this reason, experimental-based tracking algorithms have been assessed with the help of video- or infrared-based detectors to initialize the particles. However, two ad-hoc and purely acoustic-based detection techniques have been developed and evaluated by experimental measurements also. Results are presented in chapter 6, and the description of the two proposed solutions follows.

Let us divide the road section into a tracking zone and a detection zone. The latter, also called "region of interest" in [102, 103], is continuously monitored to issue an alarm if a new vehicle is detected. Two detection strategies are proposed: the *broadside detection strategy* and the *endfire detection strategy*. In the former, the detection zone is placed in front of the array. In the latter, the detection zone is placed far from the array. In both strategies, the tracking zone is placed in front of the array, see Fig. 3.5.

### 3.7.1 Broadside detection

The first approach, probably the easier one, consists in detecting vehicles when they are in front of the microphone array.

Audio recordings are partitioned into short audio frames. In its simplest form, the detection problem may be seen as a comparison between a *classifier* $\mathcal{D}[q]$, built from audio features extracted from the $q^{th}$ audio frame, and a *threshold* $\Lambda$, above (respectively below) which the hypothesis $H_1$ holds: *at least one road vehicle is in the detection zone* (respectively the hypothesis $H_0$ holds: *all other situations*). This is what is called the *likelihood-ratio test*, expressed by:

$$\mathcal{D}[q] \lessgtr_{H_1}^{H_0} \Lambda. \tag{3.57}$$

In the present work, we did not focus on techniques aiming at building the classifier $\mathcal{D}$ regardless of the number of features and their performance. For this purpose, many methods have been proposed in the literature: Maximum a Posteriori (MAP) [104], Logistic Regression [105], Decision Tree (C4.5 Algorithm) [106], Maximum Distance Approach (MPP), K-Nearest Neighbor Search, Neural Network (Multi-Layer, Artificial), Gaussian Mixture Model (GMM), Support Vector Machine (SVM) [107] to list a few. These methods propose different solutions to combine the features extracted from the raw signal in order to return a probability of belonging to one of both classes.

Actually, we chose to focus on optimizing each feature so as to build the simplest classifier $\mathcal{D}$, *i.e.*, with the lowest possible number of features. In this view, each feature has been derived by considering a specific octave band as well as considering the raw signal. The performance in detection of each sub-feature has been assessed using a receiver operating characteristics (ROC) analysis.

The ROC analysis [108] permits to assess the performance of a classifier after calculation of its false positive rate (FPR) and true positive rate (TPR), expressed by:

$$\text{TPR} = \frac{TP}{P} = \frac{TP}{TP+FN}, \tag{3.58}$$

$$\text{FPR} = \frac{FP}{N} = \frac{FP}{FP+TN}, \tag{3.59}$$

where:

- $P$: *positives* (actual number of frames in class 1);
- $N$: *negatives* (actual number of frames in class 0);
- $TP$: *true positives* (number of frames classified as 1 and belonging to class 1);
- $FN$: *false negatives* (number of frames classified as 0 and belonging to class 1);
- $FP$: *false positives* (number of frames classified as 1 and belonging to class 0);
- $TN$: *true negatives* (number of frames classified as 0 and belonging to class 0).

Thus, a perfect classifier is one for which $TPR$ equals one and $FPR$ equals zero.

As underlined in [109], most of the literature dedicated to audio signal classification concerns speaker recognition, music classification or musical instrument recognition. Studies on environmental sound recognition are few in comparison. Experimental results using this approach are presented in section 6.5.1.

### 3.7.2 Endfire detection

The second approach that has been investigated is the endfire detection strategy. It consists in monitoring a zone upstream the tracking one, as exemplified in Fig. 3.5b. The objective is to return an alert if and only if a vehicle leaves the detection zone and enters the tracking zone.

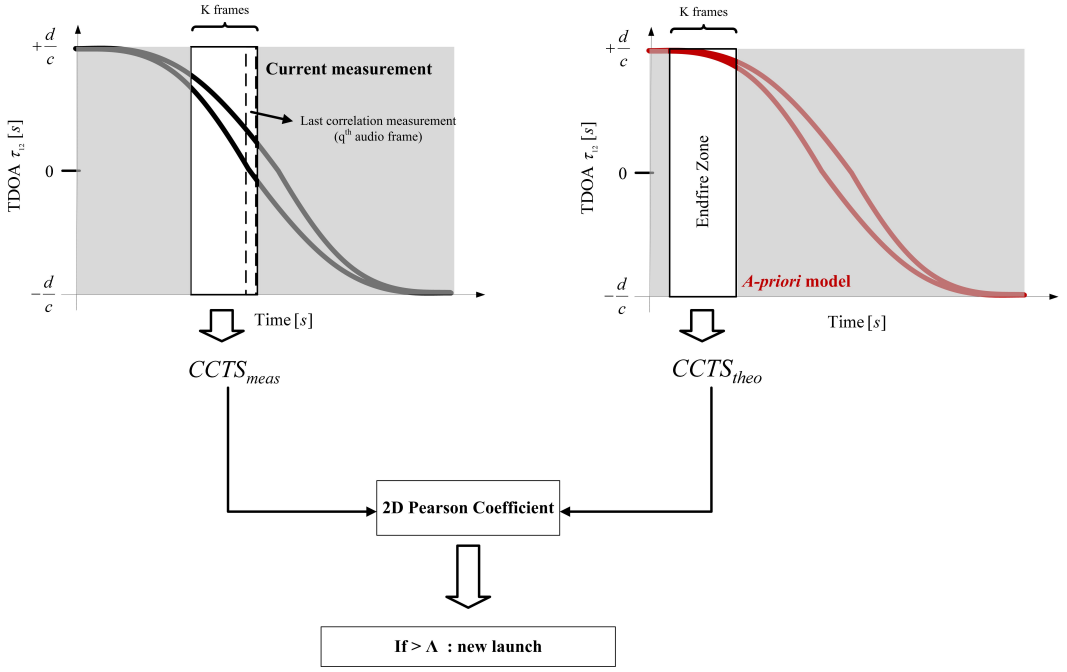



Figure 3.6: Principle of the endfire detection (CCTS matching test).

The advantage of detecting vehicles as early as possible is twofold. Firstly, it enables the simultaneous tracking of vehicles that pass each other in front of the array. Secondly, it allows a pseudo real-time tracking: the particles are launched as soon as the vehicle enters the tracking zone, then the correlation measurements are filtered as soon as they are updated. No measurement storage is required. The problem is that, the further the detection zone is, the lower the signal-to-noise ratio. The key point therefore consists in designing the two zones in order to ensure, at the same time, a sufficient observation time interval for the tracking, and the lowest missed detection as possible. Ideally, the detection zone is far from the array, in order to begin the tracking as early as possible, but also to make it as short as possible, in order to avoid that multiple vehicles that follow each other too closely be considered as only one.

The endfire detection technique that we propose is schematically depicted in Fig. 3.6. It consists in comparing two cross-correlation time series (CCTS) of same size one being the concatenation of the $K$ last cross-correlation measurements, called $CCTS_{meas}$, the other being a theoretic CCTS corresponding to the expected trajectory of a vehicle running in the detection zone, called $CCTS_{theo}$.

A simple way to compare the two matrices is to compute the 2D Pearson coefficient $r$. To simplify notations, let us consider two matrices A and B of size M$\times$ K. Then $r$ is

given by:

$$r = \frac{\sum_{k=1}^{K}\sum_{m=1}^{M}\left(A[m,k]-\bar{A}\right)\left(B[m,k]-\bar{B}\right)}{\sqrt{\sum_{k=1}^{K}\sum_{m=1}^{M}\left(A[m,k]-\bar{A}\right)^2\sum_{k=1}^{K}\sum_{m=1}^{M}\left(B[m,k]-\bar{B}\right)^2}}, \tag{3.60}$$

Considering $r$ as the classifier in (3.57) solves the detection problem under the condition of finding the optimal threshold $\Lambda$. Again, this is achieved using the ROC analysis through a training database. Experimental results are provided in section 6.5.2.

## 3.8 Conclusion

This chapter focused on moving sound source detection and tracking. The Bayesian theory has been introduced and most classical Bayesian-based tracking algorithms have been reviewed. For linear/Gaussian systems, no estimator can outperform the optimal methods (Kalman Filter, Grid-based method). But, for tracking since bearing-only measurements, such assumptions do not hold. Suboptimal methods like extended or unscented Kalman filter have been proposed, but according to previous published works, on applications similar to ours, both Kalman filter extensions are outperformed by the particle filtering technique. A small-scale experimental measurement validated the reliability of particle filtering such a technique for speed estimation even in case of highly false initial speed.

The detection problem has also been discussed. The rationale of the detection algorithm is to properly initialize the particles in abscissa and ordinate. Two original methods have been proposed, one for detecting vehicles in the broadside direction, another for detecting vehicles in the endfire direction. These methods will be assessed through real audio recordings in the chapter 6.

Now that the methods for localization and tracking have been defined, one has to investigate how apply them in practice. This is what is dealt within the next chapter.

# 4 Bimodal sound source model: application to the monitoring of two-axle vehicles

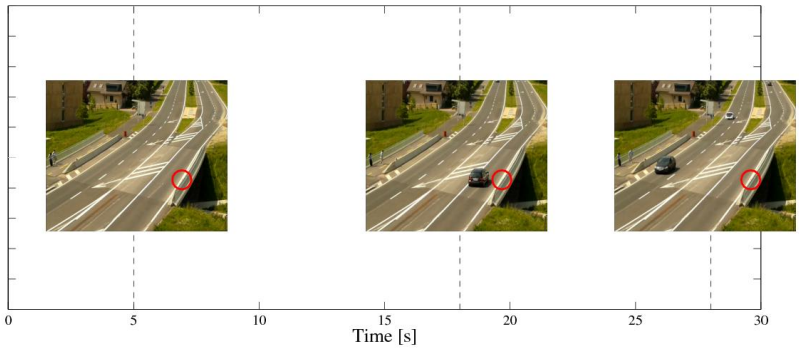
## 4.1 Introduction

In this chapter, the concept of bimodal sound source tracking is introduced, namely, the method allowing to track of a couple of statistically independent but mechanically constraints sound sources simultaneously. The specific case of two-axles road vehicles is considered as a direct application of the method.

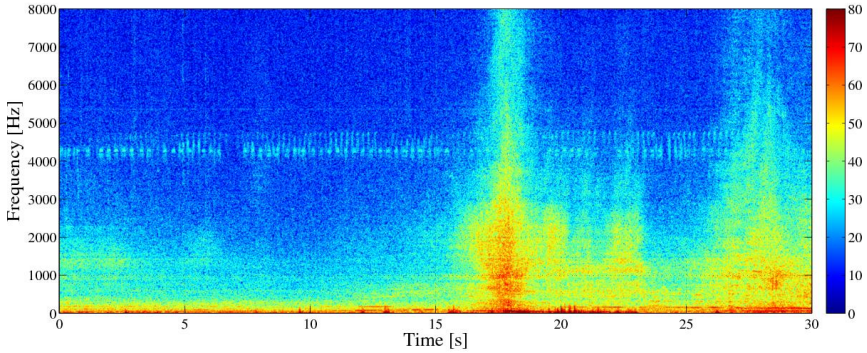
According to chapter 3, the particle filtering algorithm requires i) an observation of the acoustic environment regularly updated to govern particle resampling and ii) a model defining the target state vector, the relationship between the particle likelihood and the measurements, the dynamic model which is expected and the conditions about particle birth and death. The point i) was already discussed in chapter 2 and corresponds to the cross-correlation time series (CCTS) between a pair of sensors placed in parallel to the road lane. Although the GCC-PHAT function was found to be one of the most effective to observe a pass-by, this chapter discusses of a potential improvement to it using the spectral content of the pass-by noise. Moreover, a closed-form expression of the observation is proposed, making the assessment of the tracking methods through simulations possible. In order to address point ii), the model of section 3.6 is updated to match with real pass-by measurements. A new model, adapted to the most common class of vehicles, *i.e.* two-axle vehicles, is proposed.

From a physics viewpoint, the pass-by noise is mainly composed of three different components, namely [110, 34]:

- the mechanical noise, including transmission and exhaust system;
- the rolling noise (or tyre/road noise), due to the interaction between tyres and asphalt ;
- the aerodynamic noise, due to the air flow generated by the boundary layer of the vehicle.




(a)

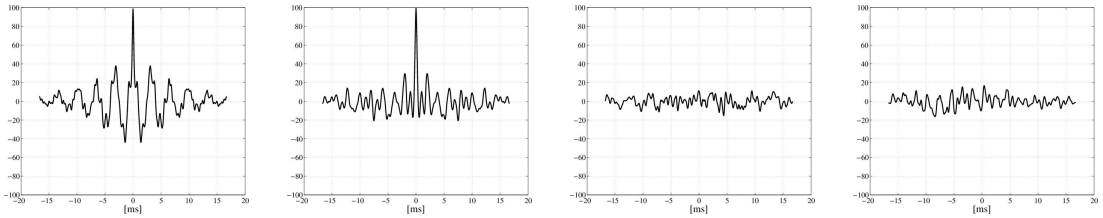



(b)

Figure 4.1: A typical in-situ audio recording. (a): pictures illustrate the vehicles' positions as a function of time with respect to that of the microphone (red circle), (b): spectrogram in dB SPL (analysis window length: 80 ms, temporal overlap: 60 ms (75%), spectral resolution: 12.5 Hz, apodization window: Hamming).

A commonly accepted approximation consists in saying that the mechanical noise, respectively the tyre/road noise predominates for vehicle running below 50 km/h, respectively upper 50 km/h. But in modern cars, the tyre/road also dominates at low speed for constant speed driving [111]. Thus, the major assumption of this thesis is that vehicles are not under acceleration during the observation (lasting between 1 to 4 seconds in general). Each observation is partitioned in short audio signal frames (30 ms long) within which the vehicle is considered as static.

Let us return to the spectro-temporal representation depicted in the chapter 1, reintroduced in Fig. 4.1. This signal was acquired at a sampling rate of 51.2 kHz and with a quantification of 24 bits. One can clearly dissociates the first pass-by at 18 seconds and a second one, weaker because the vehicle is further away, at nearly 28 seconds. The spectral contents of the background noise and pass-by noise are clearly distinguishable, in particular, the closer the vehicle, the richer the spectral content of the pass-by noise is. Most of its energy is almost uniform below 4 kHz. The energy in the band 5kHz-8kHz is 40 dB below that in the band 0kHz-3kHz. No strong components such as harmonics or




(a) Auto-correlation for vehicle A

(b) Auto-correlation for vehicle B

(c) Cross-correlation for vehicle A and B

(d) Cross-correlation for two successive frames of vehicle B

Figure 4.2: (a) auto-correlation for one audio frame extracted at the vehicle A pass-by, (b) idem as (a) but for the v Auto- and cross-correlation for two different road vehicles.

pure tones emerge so that in the remainder of this document, the pass-by noise will be modeled as wideband and stochastic with a 3.5 kHz bandwidth.

Another investigation regarding the correlation of the sound sources was carried out. Two audio frames of 41 ms each were extracted from an audio recording with two successive road vehicles passing by. Each frame exactly corresponds to the broadside position of one vehicle, designated by A and B. The auto-correlation for frame A (respectively B) is depicted in Fig. 4.2a (respectively 4.2b). The Dirac-shaped result confirms the broadband nature of the signal in these two frames. Fig. 4.2c shows the cross-correlation for frames A and B. Here, no peak appears, validating the assumption that two different vehicles, even being on the same section of road with quite similar speed, load, position etc., are uncorrelated. Therefore, this quick check justifies the use of cross-correlation-based methods to locate multiple vehicles simultaneously because the number of peaks in the cross-correlation is directly related to the number of sound sources. Finally, Fig. 4.2d depicts the cross-correlation between two successive temporal frames B and B' such that the frame B' corresponds to the next 41 ms. Here again, no peak appears, meaning that a vehicle pass-by can be modeled as a succession of uncorrelated and static point emitters placed one after the other.

These preliminary observations serve as the base for the developments that follow, especially regarding the target model and the observation function.

## 4.2 Signal Model

Because of the finite speed of sound and of the fact that the target is continuously moving, each DOA estimate actually corresponds to the DOA of the vehicle at an earlier, rather than at a current, time. This is the so-called *retardation effect* well known by people working on aircraft/ballistic tracking [112]. In this type of application, the speed of sound, nearly 1235 km/h, is comparable to the speed of the target, so that errors in position may reach several hundred of meters. This effect is not taken into account here because there is no need to know the exact position of the vehicle at each time but rather

its speed. It takes between 30 ms and 40 ms to get a new position update and during this time interval, a vehicle travels about one meter at 100 km/h. Therefore, acquired microphone signals can be simply modeled as an extension of (2.2)-(2.3) such that:

$$y_1(t) = \sum_{k=1}^{N} \alpha_k s_k(t - \delta_{1k}) + n_1(t), \tag{4.1}$$

$$y_2(t) = \sum_{k=1}^{N} \alpha_k s_k(t - \delta_{1k} - \tau_{12,k}) + n_2(t), \tag{4.2}$$

where $N$ is the number of sound sources.

## 4.3 Target model

The closer the model is to reality, the more robust the tracking is against noise. On the other hand, a highly precise model increases te risk of failure in case of model mismatch. In urban and peri-urban areas, vehicles are expected to run between 50 km/h and 100 km/h at a constant speed during the observation so the predominant noise is the tyre/road one [113, 111]. The tyre/road noise is a combination of several physical mechanisms [114]:

- vibratory phenomena caused by the irregularities of the road surface and by the deformation of the tyre on the contact zone, producing frequencies below 1000 Hz;
- resonance phenomena caused by the air confined in cavities between the tyre and the road surface, producing frequencies around 1000 Hz;
- amplification phenomena, the so-called *horn effect*, caused by the noise reflected between the surface of the tyre and the surface of the road at the front and rear parts of the tyre;
- screeching phenomena caused by the succession of adhesion and detachment of the tyre rubber, producing frequencies above 1000 Hz.

It appears that the tyre/road noise highly depends on the tyre type (rubber, tread patterns), the road surface (grain, porosity), and the vehicle speed, making its modeling quite difficult since the characteristics of the vehicles are of course not *a priori* known. Consequently, a simplistic but general model is, in our opinion, the only way to ensure a robust tracking in the real world.

As demonstrated with Fig. 2.7, generalized cross-correlation (GCC) displays both front and rear axles trajectories. As a consequence, a specific model dedicated to this high resolution observation is proposed. Instead of considering a unique point emitter as in the literature, for instance [12, 30, 115, 32, 20], a new model is introduced. It consists in considering a two-axle vehicle as the summation of two static monopoles radiating stochastic and identically distributed sounds separated by a wheelbase length $wb$ in the x-y plane. This is what the terms *bimodal sound source model* refer to in what

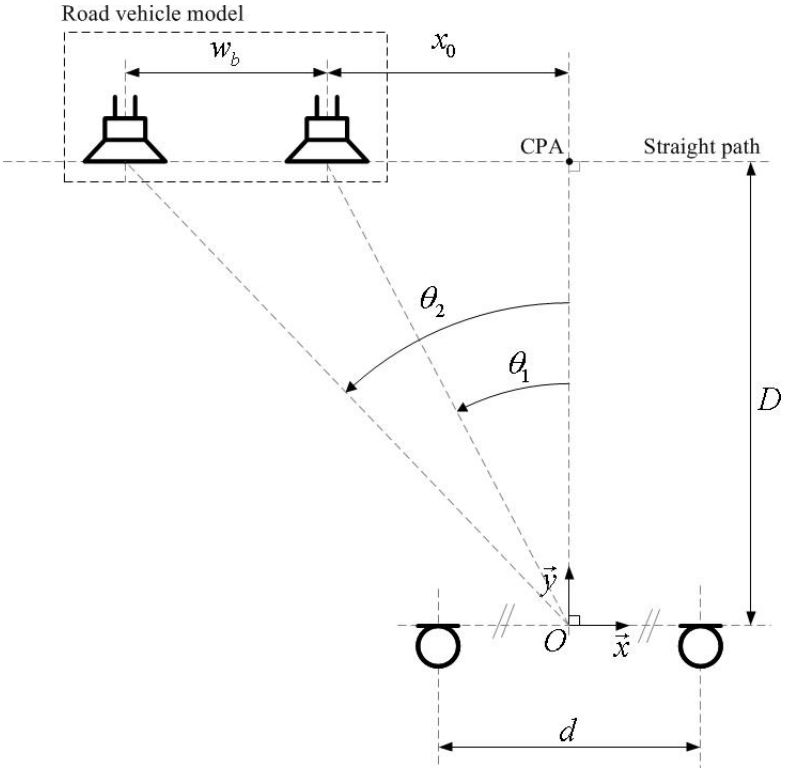



Figure 4.3: Bimodal sound source model of a two-axle road vehicle, wavefronts are acquired by a microphone array placed in parallel to the road lane. The vehicle is assumed to be static for each observation.

follows. This model is illustrated in Fig. 4.3: the wavefield is captured by a two-element microphone array with known spacing $d$, placed in parallel to the lane, at a distance $D$ to its CPA; $x_0$ denotes the distance between the front rear and the CPA, $\theta_j$ denotes the DOA of the j$^{th}$ axle, $j \in [1, 2]$. The vehicle speed is considered as a constant along the abscissa and close to zero along the ordinate.

Consequently, a new target state vector $\alpha_t$, initially expressed in Eq. (3.51), is proposed. It includes a fourth parameter $wb_t$ denoting the wheelbase length such that:

$$\alpha_t = [x_t, y_t, \dot{x}_t, wb_t]^T. \tag{4.3}$$

**Remark** We acknowledge that a much more realistic model could be considered, for instance such as those proposed by V. Cevher *et al.* [35, 77, 36]. These research works are to the best of our knowledge, the only antecedents focusing on wheelbase length estimation through acoustic sensing. In these papers, a wave-pattern-based recognition algorithm for joint speed and wheelbase estimation was suggested, using a one-channel pass-by recording acquired on the roadside. Engine, tyre, exhaust and air turbulence noises were meticulously modeled. Tyre/road noise directionality, interferences between tyres, microphone directionality and frequency response, were also taken into account. In

a totally opposite philosophy, we limit our model to the minimum *a-priori* knowledge of two-axles. This choice is mainly motivated by our experience of real world signals that may be strongly affected by interfering noises or other vehicles in the monitored area. In such cases, resorting to a too precise model may limit the practical applicability of the algorithm. Secondly, the simpler the model, the larger the potentiality to extend it for other applications is.

## 4.4 Dynamical model

Because of the short observation duration (between 1 and 3 seconds per vehicle), each target object is supposed to move at a nearly constant speed and to follow a nearly straight trajectory according to the state equation:

$$p(\alpha_t|\alpha_{t-1}) = \mathcal{N}\left(\mathbf{F}\alpha_{t-1}, \mathbf{V}\right), \tag{4.4}$$

with the prediction matrix $\mathbf{F}$ and the statistical noise covariance $\mathbf{V}$ given by:

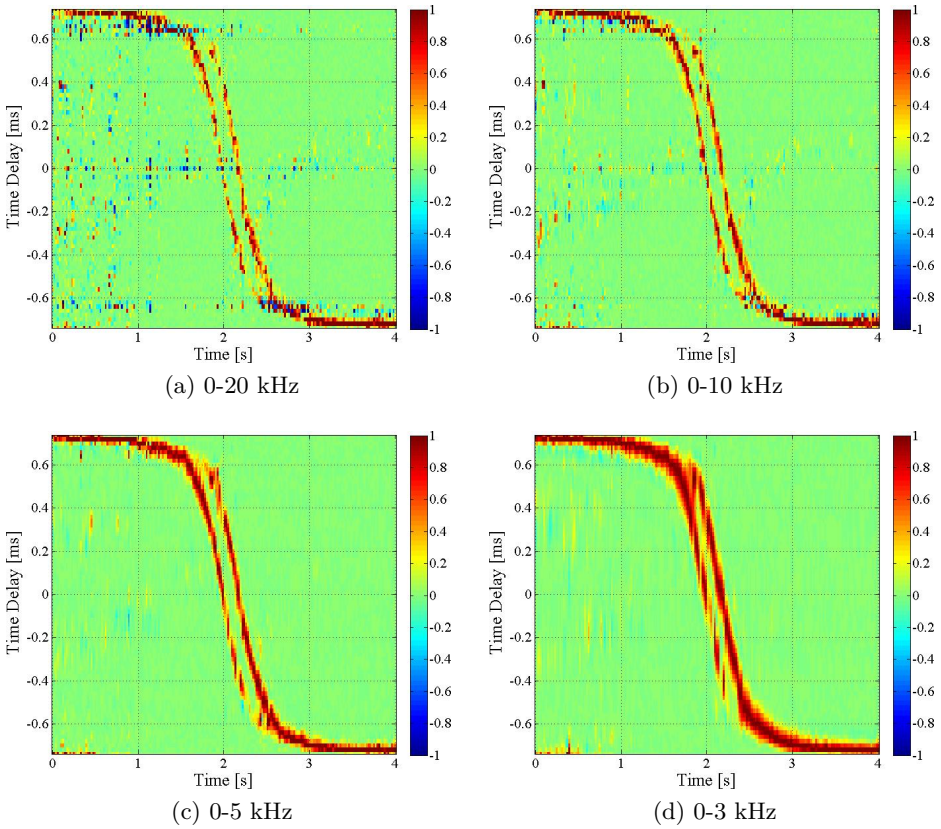
$$\mathbf{F} = \begin{pmatrix} 1 & 0 & \Delta T & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \mathbf{V} = \begin{pmatrix} \sigma_x^2 & 0 & 0 & 0 \\ 0 & \sigma_y^2 & 0 & 0 \\ 0 & 0 & \sigma_{\dot{x}}^2 & 0 \\ 0 & 0 & 0 & \sigma_{wb}^2 \end{pmatrix}$$

where $\sigma_x^2$ (respectively $\sigma_y^2$, $\sigma_{\dot{x}}^2$ and $\sigma_{wb}^2$) are the noise variances of $x$ (respectively $y$, speed $\dot{x}$ and wheelbase length $wb$).

The speed can be positive or negative, depending on the target direction. Both constant speed and straight trajectory assumptions translate in practice into low uncertainties on the speed and co-ordinate states, that is, low values of $\sigma_{\dot{x}}^2$ and $\sigma_y^2$. Note that sufficient knowledge of the vehicle abscissa is a strong requirement for vehicle positioning using bearing-only measurements [84].

## 4.5 Observation model

The PHAT processor may be seen as a cross-power spectrum whitening, meaning that the correlation in amplitude between signals is discarded. This approach can be justified when it comes to estimating only phase differences. In general, a much more accentuated peak than the classical cross-correlation one is achieved. The price to pay is that spurious peaks can appear, for instance, because of a spatially coherent noise at low frequencies and/or power too low at high frequencies. Whatever the signal to noise ratio, coherent noises are considered as other sources by the GCC-PHAT. Therefore, it is often of interest to work only on the spectral band in which most of the energy of the useful signal lies. This can be done using the *Bandpass-PHAT* (BPHAT) weighting. This processor was




(a) 0-20 kHz (b) 0-10 kHz

(c) 0-5 kHz (d) 0-3 kHz

Figure 4.4: Influence of the BPHAT processor bandwidth ($B_w$ and $f_c$) on the quality of observation exemplified on a real signal.

previously proposed for speaker localization by DiBiase in [116] p. 46 or for water pipes leak localization by Gao *et al* in [117, 118]. It is defined as:

$$\psi_{bphat}(f) = \begin{cases} \psi_{phat}(f) & \text{if } f_c - B_w/2 \leq |f| \leq f_c + B_w/2 \\ 0 & \text{otherwise.} \end{cases} \tag{4.5}$$

where $f_c$ and $B_w$ respectively denote the central frequency and the bandwidth on which the BPHAT transform is applied. To be effective, the spectral band on which the BPHAT is applied needs to be identical or within the bandwidth of the signal of interest. This point is illustrated in Fig. 4.4: a real pass-by measurement has been processed by four different BPHAT-CCTS which differ in the parameters $B_w$ and $f_c$. Taking a too large bandwidth (weighting bandwidth larger than signal bandwidth) causes the apparition of spurious peaks: Fig. 4.4a and Fig. 4.4b. Adapting the bandwidth properly greatly improves the contrast of the axles: Fig. 4.4c and Fig. 4.4d.

According to (4.5) and (2.16), one can demonstrate that the closed-form expression of

the GCC-BPHAT function for the single source case is (see Appendix A.2 for a proof):

$$R_{s_1 s_1}^{bphat}(\tau) = 2B_w \cos\left[2\pi f_c(\tau - \tau_{12})\right] \mathrm{sinc}\left[B_w(\tau - \tau_{12})\right]. \tag{4.6}$$

For the two-sound source case and under the assumption that each source delivers a zero-mean signal, uncorrelated with the other, one gets:

$$\begin{aligned} R_{s_1 s_2}^{bphat}(\tau) &= R_{s_1 s_1}^{bphat}(\tau) + R_{s_2 s_2}^{bphat}(\tau), & (4.7)\\ &= 2B_w \left(A_1 + A_2\right), & (4.8) \end{aligned}$$

with

$$A_k = \cos\left[2\pi f_c(\tau - \tau_{12,k})\right] \mathrm{sinc}\left[B_w(\tau - \tau_{12,k})\right], \ k \in [1, 2].$$

**Remark** It may be noted that, regarding the application targeted for these developments, the non-correlation of sources is in this case a debatable assumption since sources coming from the axles of a vehicle would somewhat be correlated (same speed and loading for instance). Consequently, cross-terms in the correlation measure should be considered but are non easily quantifiable. This is why they are neglected as a first approximation.

## 4.6 Likelihood model

The proposed likelihood model is an extension of the model (3.54) that now takes into account the wheelbase length state such that:

$$p(\beta_t|\alpha_t^{(n)}) = \frac{1}{2}\left(R_{s_1 s_2,t}^{bphat}(\tau_{12,1,t}^{(n)}) + R_{s_1 s_2,t}^{bphat}(\tau_{12,2,t}^{(n)})\right) \ \forall n \in [1, 2, ..., N_p], \tag{4.9}$$

where $\tau_{12,1,t}^{(n)}$ and $\tau_{12,1,t}^{(n)}$ denote the TDOA between microphones 1 and 2 inherent to the $n^{th}$ candidate positions for front and rear axles at time $t$ respectively. As the likelihood measure is updated at each time step, the time index $t$ is dropped in this paragraph for the sake of clarity in the notation. Both time-delays are given by:

$$\tau_{12,1}^{(n)} = \frac{\sqrt{\left(x^{(n)} - d/2\right)^2 + \left(y^{(n)}\right)^2} - \sqrt{\left(x^{(n)} + d/2\right)^2 + \left(y^{(n)}\right)^2}}{c} \tag{4.10}$$

$$\tau_{12,2}^{(n)} = \frac{\sqrt{\left(x^{(n)} - wb^{(n)} - d/2\right)^2 + \left(y^{(n)}\right)^2} - \sqrt{\left(x^{(n)} - wb^{(n)} + d/2\right)^2 + \left(y^{(n)}\right)^2}}{c} \tag{4.11}$$

An interpretation of (4.9) is that to each particle $n$ located at $(x^{(n)}, y^{(n)})$ corresponds a "particle-image" located at $(x^{(n)} - wb^{(n)}, y^{(n)})$, both belonging to the same state vector $\alpha_t^{(n)}$. Each of the two particles is projected onto the correlation measure $R_{s_1 s_2,t}^{bphat}$ using relations (4.10)-(4.11). This returns two likelihood measures (one per axle) which are summed to give the final likelihood of the candidate $\alpha_t^{(n)}$. This principle is illustrated

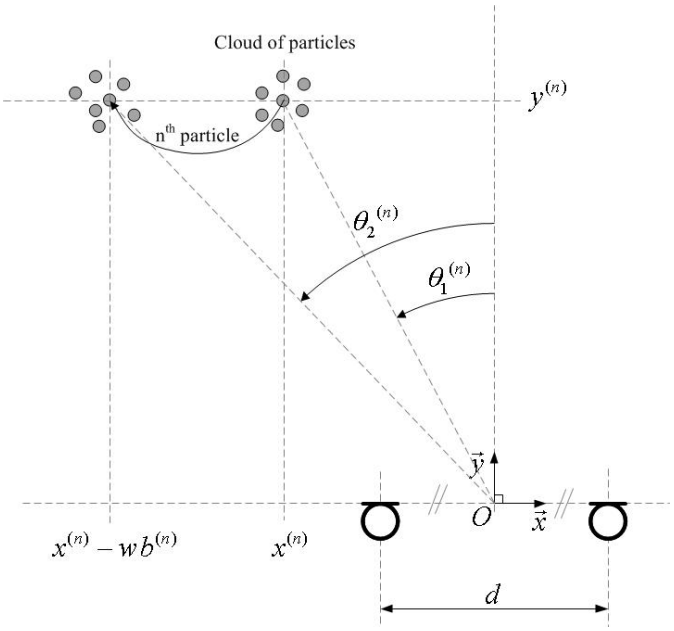



Figure 4.5: Basic bimodal likelihood model. Each particle $n$ of the front axle at coordinate $(x^{(n)}, y^{(n)})$ build a particle image for the rear axle through the $n^{th}$ candidate for wheelbase $wb^{(n)}$.

in Fig. 4.5. The term *bimodal* used before is justified here in the sense that the two observation modes corresponding to front and rear axles are jointly tracked. By extension, such a likelihood-based particle filtering is called bimodal particle filtering (BPF) in the following.

One limitation of model (4.9) is that both axles are considered with same importance at all times, that is, whatever the vehicle position. In reality, only the front (respectively rear) axle is observed when the vehicle is approaching (respectively leaving). When the vehicle is in front of the array (broadside), both axles are observed. As an improvement to model (4.9), positive weighting factors $\gamma_{1,t}$ and $\gamma_{2,t}$ are introduced into the likelihood model such that:

$$p(\beta_t|\alpha_t^{(n)}) = \gamma_{1,t} R_{s_1 s_2,t}^{bphat}(\tau_{12,1,t}^{(n)}) + \gamma_{2,t} R_{s_1 s_2,t}^{bphat}(\tau_{12,2,t}^{(n)}). \tag{4.12}$$

When the vehicle is approaching (respectively leaving) the algorithm ideally should give more weight to the front axle (respectively the rear axle), so that $\gamma_1$ should be larger (respectively smaller) than $\gamma_2$. One simple way to allocate the contribution of axles through these weights is presented in the following.

Let us introduce the quantity $\tau_{12,0}$ representing the TDOA relative to the vehicle center,

averaged over the $N_p$ candidates such that:

$$\begin{aligned} \tau_{12,0}^{(n)} &= \frac{\sqrt{\left(x^{(n)} - wb^{(n)}/2 - d/2\right)^2 + \left(y^{(n)}\right)^2} - \sqrt{\left(x^{(n)} - wb^{(n)}/2 + d/2\right)^2 + \left(y^{(n)}\right)^2}}{c}, \\ \tau_{12,0} &= \frac{1}{N_p}\sum_{n=1}^{N_p} \tau_{12,0}^{(n)}. \end{aligned} \tag{4.13}$$

By setting:

$$\gamma_1 = \frac{1}{2}\left(\frac{c\tau_{12,0}}{d} + 1\right), \tag{4.14}$$

$$\gamma_2 = 1 - \gamma_1, \tag{4.15}$$

we achieve the desired effect. Indeed, when the vehicle is approaching, *i.e.* $0 \leq \tau_{12,0} \leq d/c$ then $\gamma_1 \geq 0.5$ and $\gamma_2 \leq 0.5$, giving more importance to the front axle. When the vehicle is leaving, *i.e.* $-d/c \leq \tau_{12,0} \leq 0$ then $\gamma_1 \leq 0.5$ and $\gamma_2 \geq 0.5$, giving more importance to the rear axle. When the vehicle is in the broadside direction, $\tau_{12,0} = d/c$ yielding $\gamma_1 = \gamma_2 = 0.5$. Both axles are considered with equal importance. Such a strategy is illustrated in Fig. 4.6.

## 4.7 Initialisation and stopping conditions

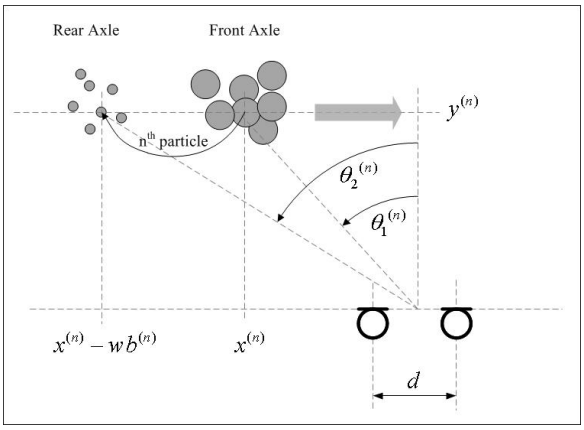
The tracking begins when a new approaching car is detected at a predefined abscissa. At this instant, the initial state vector is drawn from a Gaussian distribution and initial weights are all equal and normalized, such that, for $n \in [1, 2, ..., N_p]$:

$$\alpha_0^n \sim \mathcal{N}\left(\begin{bmatrix} \mu_{x,0} \\ \mu_{y,0} \\ \mu_{\dot{x},0} \\ \mu_{wb,0} \end{bmatrix}, \begin{bmatrix} \sigma_{x,0}^2 & 0 & 0 & 0 \\ 0 & \sigma_{y,0}^2 & 0 & 0 \\ 0 & 0 & \sigma_{\dot{x},0}^2 & 0 \\ 0 & 0 & 0 & \sigma_{wb,0}^2 \end{bmatrix}\right), \tag{4.16}$$

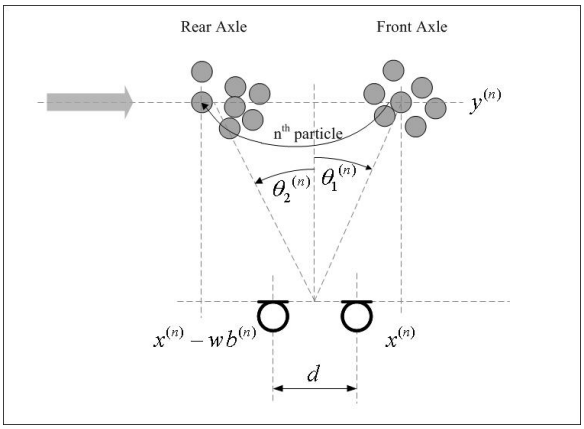
$$w_0^{(n)} = \frac{1}{N_p}, \tag{4.17}$$

where the means $\mu_{.,0}$ denotes the *a priori* knowledge of the target state vector, and the noise variances $\sigma_{.,0}^2$ denotes the uncertainty in this knowledge.

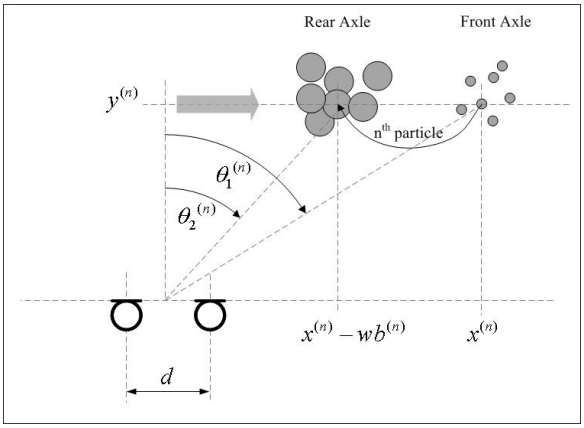
The tracking is stopped after a predefined duration.




(a) Vehicle is approaching. More importance is given to the particles of the front axle




(b) Vehicle is in the broadside DOA. Front and rear axles are considered with equal importance




(c) Vehicle is leaving. More importance is given to the particles of the rear axle

Figure 4.6: Improved bimodal likelihood model. Distribution of the particle weights as a function of the vehicle DOA.

## 4.8 Simulation

One typical BPF result is detailed using an *in-silico* experiment, and depicted in Fig. 4.7. The BPHAT-CCTS, Fig. 4.7a, is built from the closed-form expression of the GCC-BPHAT function of Eq. (4.8), in which the primary correlations are weighed according to the likelihood weighting model (4.14)-(4.15). Geometrical, acoustical and statistical parameters of the simulated scenario are summarized in Table 4.1.

In this example, the observation does not correspond to the whole CCTS but only to the part delimited by the two black lines in Fig. 4.7. The observation is considered to start at $t = 0$ and to finish at $t = T$ seconds independently of the true time axis.

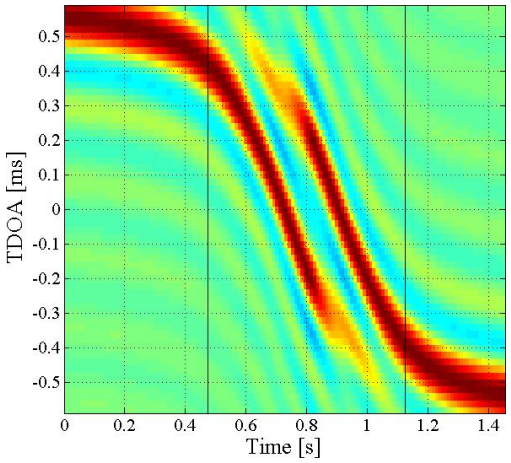
Fig. 4.7b and Fig. 4.7c depict the distributions of the particles as a function of time, respectively, for speed and wheelbase states. At $t = 0$ (first black line), speed and wheelbase states are drawn from the Gaussian distribution $\mathcal{N}(\mu_{\dot{x},0}, \sigma_{\dot{x},0})$ and $\mathcal{N}(\mu_{wb,0}, \sigma_{wb,0})$ respectively. For demonstration purpose, the *a priori* $\mu_{\dot{x},0}$ and $\mu_{wb,0}$, denoted by blue crosses A, are clearly below the actual values, denoted by red dashed lines. One can also note that for the wheelbase state case, Fig. 4.7c, no candidates correspond to the actual value of 2.5 m; all the particles are contained between 1 m and 2.25 m.

After a few iterations, particles converge properly towards their respective target values. One possible way to build an estimate therefore simply consists in computing the mean of the particle distribution at the end of the tracking: let us call them $\mu_{\dot{x},T}$ for speed, and $\mu_{wb,T}$ for the wheelbase length. These values are depicted by the blue crosses B.

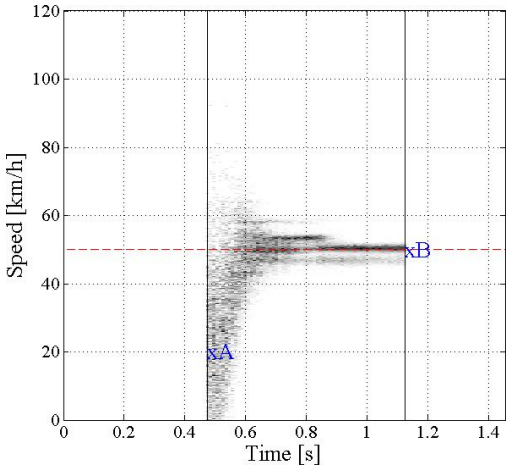
Table 4.2 summarizes the average performance of this scenario over $N_{test} = 100$ runs. The performance of the estimator of the $j^{th}$ coordinate of the state vector, $j \in [1, 2, 3, 4]$, are characterized by the global error $\Sigma_{\epsilon,j}$, the global percentage error $\Sigma^{\%}_{\epsilon,j}$, the global standard deviation $\Sigma_{\sigma,j}$ and the relative total standard deviation $\Sigma^{\%}_{\sigma,j}$, all defined in Appendix A.3.

In this example, the performance is convincing regarding the global error for speed and wheelbase length (-1.1 km/h and -17 cm respectively) knowing that the *a priori* values were quite far from the actual ones (-30 km/h and -100 cm respectively). The repeatability of the speed estimate is very good (1.7 km/h of standard deviation only). The relative standard deviation achieved by the wheelbase length estimator is larger but stays below 10%.

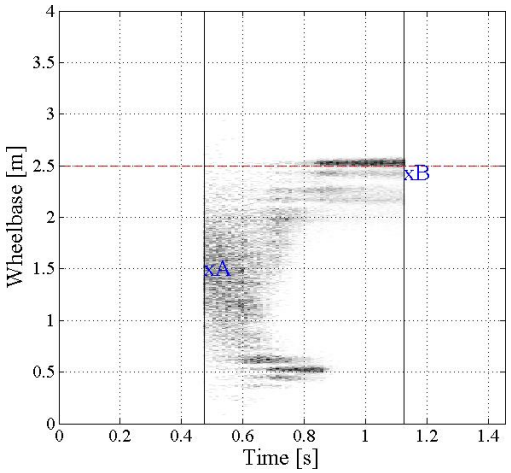
Looking again at Fig. 4.7a, one can remark that in reality, the wheelbase information is strongly expressed only when the vehicle is close to its CPA, namely between 0.7 seconds and 0.9 seconds approximatively in this example. This is a rather short time interval for the particles to converge. On the other hand, the information on speed is always present during the observation. This explains in part why the performance for speed is better than that for wheelbase length, and also why particles for speed, Fig. 4.7b, converge quicker than particles for wheelbase, Fig. 4.7c.




(a) observation




(b) state *speed*




(c) state *wheelbase*

Figure 4.7: Typical example of a tracking result applied to speed estimation. The observation likelihood function is delimited by the two vertical black lines on the CCTS (a). (b) represents the evolution of the speed state histogram with a false *a priori* starting.

| | | **Actual states** | ***A priori* states** | **Initial STD** | **Noise STD** |
|---|---|---|---|---|---|
| Particle Filter | States | $x_0$ = -3 m | $\mu_{x,0}$ = -3 m | $\sigma_{x,0}$ = 0.1 m | $\sigma_x = \sigma_{x,0}/\lambda$ |
| | | $y_0$ = 3.5 m | $\mu_{y,0}$ = 3.5 m | $\sigma_{y,0}$ = 0.1 m | $\sigma_y = \sigma_{y,0}/\lambda$ |
| | | $\dot{x}$ = 50 km/h | $\mu_{\dot{x},0}$ = 20 km/h | $\sigma_{\dot{x},0}$ = 20 km/h | $\sigma_{\dot{x}} = \sigma_{\dot{x},0}/\lambda$ |
| | | $wb$ = 2.5 m | $\mu_{wb,0}$ = 1.5 m | $\sigma_{wb,0}$ = 0.4 m | $\sigma_{wb} = \sigma_{wb,0}/(2\lambda)$ |
| | noise parameter | | | $\lambda$ = 200 | |
| | number of particles | | | $N_p$ = 10000 | |
| | start tracking condition | | | $x_0$ = -3.5 m | |
| | end tracking condition | | | $x_T$ = 3 m | |
| Observation | speed of sound | | | c = 343 m/s | |
| | inter-sensor distance | | | d = 0.2 m | |
| | length of window analysis | | | $N_s$ = 2048 samples | |
| | percentage overlap | | | 75 % | |
| | BPHAT bandwidth | | | $B_w$ = 4500 Hz | |
| | BPHAT central frequency | | | $f_c$ = 2500 Hz | |
| | sampling frequency | | | $f_s$ = 50 kHz | |

Table 4.1: Default parameters of the bimodal particle filtering and observation function used in the test.

| | Actual | $\Sigma_\mu$ | $\Sigma_\epsilon$ | $\Sigma_\epsilon^{\%}$ | $\Sigma_\sigma$ | $\Sigma_\sigma^{\%}$ |
|---|---|---|---|---|---|---|
| speed | 50 km/h | 48.9 km/h | -1.1 km/h | -2.2% | 1.7 km/h | 3.4% |
| wheelbase | 2.5 m | 2.32 m | - 0.17 m | -6.8 % | 0.2 m | 7.3 % |

Table 4.2: Performance analysis of the bimodal particle filtering for the parameters of Table 4.1.

## 4.9 Influence of the BPF internal parameters and CCTS observation quality

The performance of any tracking algorithm increases with the quality of the observation and is also ruled by the internal parameters of the algorithm. In the present case, the observation determines the weight of the particles at each iteration, and thus, the particle resampling. Similarly, the internal parameters (initial states, initial noise, dynamic noise, number of particles) govern how to explore the observation, and thus, the particle convergence. As highlighted by Lichtenauer *et al.* [119] and Abbott *et al.* [120], research works focusing on how the observation quality or internal parameters affect the tracking performance are rare. Inspired by these two pioneering papers, some *in-silico* tests were

carried out in order to assess the influence of the parameters involved in the BPF.

In this testing campaign, the bimodal particle filtering was applied to theoretical observations, built from the closed-form expression of the GCC-BPHAT function, Eq. (4.8), and the likelihood weighting model (4.14)-(4.15). Default parameters both for the observation function (CCTS-BPHAT) and filtering are summarized in Table 4.1.

**Remark** It is important to note that the high number of parameters to adjust (from Table 4.1: 13 (15) for the unimodal (bimodal) particle filter, 6 for the observation function) make the optimal algorithm difficult to define and the practitioner's experience is often crucial in the application of such methods. In the literature, the values of parameters are also rarely explained. Moreover, an optimum choice is always related to a specific observation. In the present case, another observation (vehicle) will require another set of values. Finally, the inter-dependencies between parameters that come into play make the search for this optimum even more complex. In this section, we focus our attention on the most important parameters. Each one is studied separately in order to assess its influence on the tracking performance.

### 4.9.1 Influence of the number of particles

It is known that the estimation accuracy of the posterior increases [121] and the risk of loss of tracking decreases [122] as the number of particles ($N_p$) increases. On the other hand the complexity of the algorithm, and thus the computation time, increases linearly with $N_p$ [123], so that the practitioner should properly adjust $N_p$ by considering both the execution time and tracking performance in the light of the available CPU ressources.

In this section, the influence of the number of particles ($N_p$) is evaluated through the particle filtering performance on speed estimation. Simulations are carried out using the same parameters as in Table 4.1 except that $N_p$ is now variable and ranges from 10 to 2500. In this test, the wheelbase length is supposed to be exactly known, namely the *a priori* value at initialisation $\mu_{wb,0}$ equals the actual value $wb$. The tracking is launched at an arbitrary but known vehicle position (before the broadside position) and is stopped when its actual abscissa ($x$) equals wb/2, *i.e.* when the vehicle is in the broadside DOA. Such a tracking zone is depicted by the two black lines in Fig. 4.8a. On this plot, the observation corresponds to a vehicle speed of 50 km/h. The experiment is conducted for two different vehicle speeds: 50 km/h and 100 km/h.

Three different *a priori* initial speed $\mu_{\dot{x},0}$ are tested. These *a priori* are linked to the actual speed $\dot{x}$ through a bias $\epsilon_{\%}^{(i)}$ such that:,

$$\mu_{\dot{x},0}^{(i)} = (1 + \epsilon_{\%}^{(i)}/100) \times \dot{x}. \tag{4.18}$$

In this experiment, $\epsilon_{\%}^{(1)}$, $\epsilon_{\%}^{(2)}$ and $\epsilon_{\%}^{(3)}$ were equal to $-50$, 0 and $+50$ respectively. This means that when the actual speed ($\dot{x}$) is equal to 50 km/h, respectively 100 km/h,

particles are launched at 25 km/h, 50 km/h and 75 km/h, respectively 50 km/h, 100 km/h and 150 km/h.

The global error $\Sigma_\epsilon$ and the total standard deviation $\Sigma_\sigma$, defined in Appendix A.3, are computed over $N_{test} = 200$ runs.

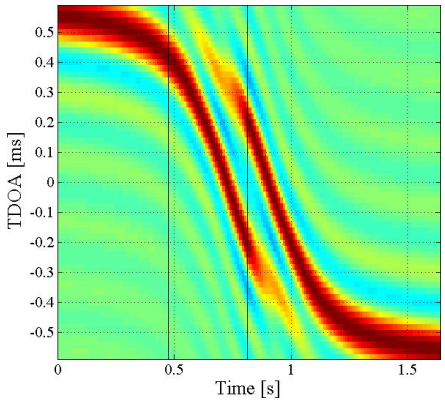
Results are depicted in Fig. 4.8. As expected by the theory, the execution time[1] evolves linearly with the number of particles, Fig. 4.8b. In parallel, mean errors and standard deviations of estimates follow an asymptotic behavior and remains constant as $N_p$ increases, Fig. 4.8c, Fig. 4.8d, Fig. 4.8e, Fig. 4.8f.

This asymptotic behavior is due to the dynamical noise injected at each iteration, which force particles to explore states around the mode even if this latter is very sharp. Other simulations, not detailed here, involving a noise parameter $\lambda$ of 400 instead of 200, *i.e.* a reduction of the dynamic noise by a factor 2 (see Table 4.1) , effectively showed lower error and standard deviation for high $N_p$. But in all cases, the asymptotic behavior remains true. One can conclude that above a certain threshold, increasing the number of particles is not determinant for the algorithm behavior. This is reminiscent of observations made by Burguera *et.al* in [124].

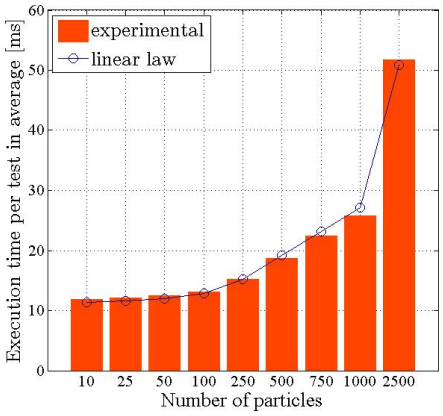
Comparing results depicted in Fig. 4.8c, Fig. 4.8d (target at 50 km/h) with those depicted in Fig. 4.8e, Fig. 4.8f (target at 100 km/h) demonstrates that with the same tracking parameters, the faster the vehicle, the poorer the results. This is due to the starting and stopping conditions, governing the observation duration (period between the vertical black lines), which are based here on a spatial criterion: the filtering begins at the same time in all situations, and stops when the vehicle is in the broadside direction. Consequently, the observation is shorter when the vehicle speed is larger, which explains in part why performance are better for the slowest vehicle. On the ground, high speeds also deteriorate the correlation measurement because of the relative Doppler shifts between sensors.

Lastly, it appears without surprise that the closer the *a priori* values to the actual, the better the performances (blue bars in the graphs). Thus, for applications in which the target speed is well known, even a low number of particles may provide satisfactory results. Comparing green and orange bars in Fig. 4.8c, Fig. 4.8d, that is, overestimated and underestimated initial speeds respectively, does not highlight any differences between the two cases. For a faster target, Fig. 4.8e, Fig. 4.8f, it seems better to underestimate the actual speed (orange bars) rather than to overestimate it (green bars) although the differences between the two cases are very small too.

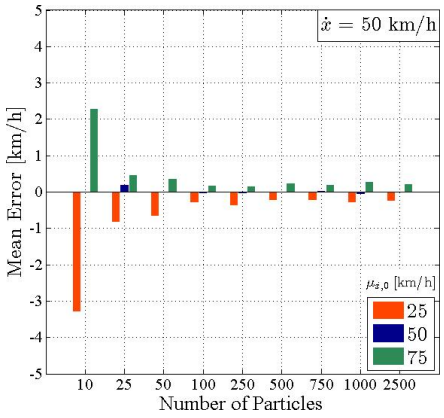
[1]Note that the presented execution times correspond to a non optimized Matlab implementation and may be drastically reduced using another programming language.



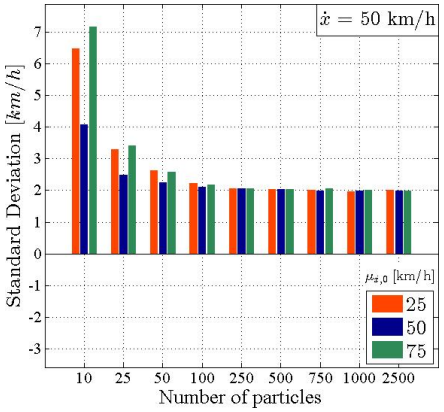

(a) Observation, $\dot{x} = 50$ km/h



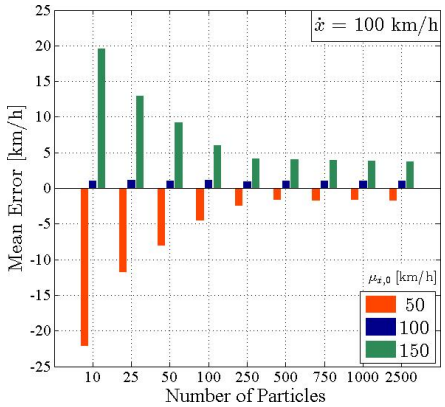

(b) Execution time averaged over 200 runs, $\dot{x} = 50$ km/h



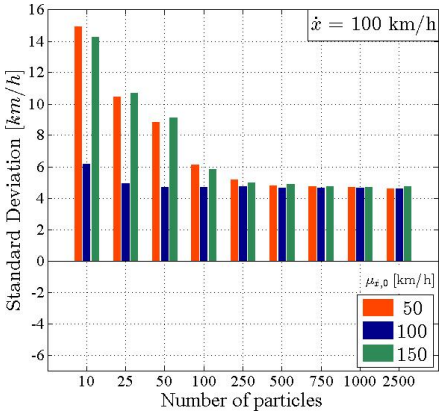

(c) Mean error (km/h) averaged over 200 runs, $\dot{x} = 50$ km/h




(d) Total standard deviation (km/h), $\dot{x} = 50$ km/h




(e) Mean error (km/h) averaged over 200 runs, $\dot{x} = 100$ km/h




(f) Total standard deviation (km/h), $\dot{x} = 100$ km/h

Figure 4.8: Influence of the number of particles on the bimodal particle filtering tracking performances.

### 4.9.2 Influence of the initial speed

This section assesses how the difference between actual and *a priori* speeds at initialisation influences the performance of the BPF. Simulations are carried out with the same parameters as in Table 4.1 except that the actual speed $\dot{x}$ and the *a priori* speed $\mu_{\dot{x},0}$ are now the variables. The former ranges from 50 km/h to 100 km/h, the latter is ruled by Eq.4.18 where the bias $\epsilon_{\%}$ ranges from -100 to +100. Results are averaged over $N_{test}$ = 200 runs. The number of particles is set to $N_p = 500$. This results in two matrices (actual speed in columns, initial bias in lines) whose elements are the global percentage error (in absolute value), Fig. 4.9a, and the total standard deviation, Fig. 4.9b, of speed estimates respectively.

As highlighted in section 4.9.1, these plots confirm that, for a given bias, the higher the actual speed, the poorest the performance is because of a shorter observation duration. One can also notice, as previously, that for fast targets, it is slightly better to underestimate the true speed than overestimate it.

The red dashed line in Fig. 4.9a depicts the boundaries within which the error is lower than 3% of the actual speed. This region is large for low target speeds and decreases as the target speed increases. However, one can observe that in this test, the algorithm is very robust to false *a priori* values.

**Remark** The reference error of 3% is the one claimed by the professional radar manufacturer ViaTraffic[2] for its *Viacount II* and for vehicle speed below 100 km/h. It is thus satisfactory to see that, in simulations at least, the proposed approach achieves comparable performance.

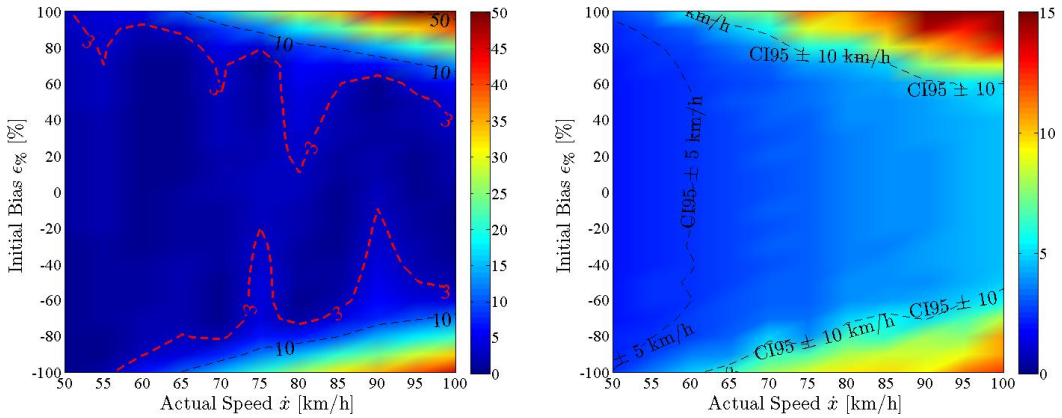
The repeatability of the BPF is depicted in Fig. 4.9b. If the difference between *a priori* and actual speeds is below 40 km/h, the CI95 of the estimates is within the ± 10 km/h limit. If the vehicle speed is below 60 km/h, the ± 5 km/h is almost guaranteed.

### 4.9.3 Influence of the initial position

Another practical question that arises concerns the road section of interest in which to track the road vehicle. In other words, given a tracking zone length, what is the optimal abscissa that introduces the smallest possible errors in the estimates. This is the topic of the following simulations.

The same parameters as in Table 4.1 are used except that the initial abscissa $x_0$ is now variable and extends from -8 m to 8 m; speed, ordinate and wheelbase are assumed to be exactly known. Theoretically, the speed estimate does not depend on $x_0$ if the latter is exactly known. In practice, it may be difficult for passive acoustic-based system to detect vehicle in a restricted section of space only, especially if the detection is done far away from the array. We therefore search to evaluate the particle filter with respect to

[2] http://www.viatraffic.de




(a) Absolute global percentage error (%) (b) Relative standard deviation (km/h),

Figure 4.9: Influence of biased *a-priori* speed values on bimodal particle filtering speed estimates in term of absolute global percentage error (a) and relative standard deviation (b) as a function of the actual target speed.

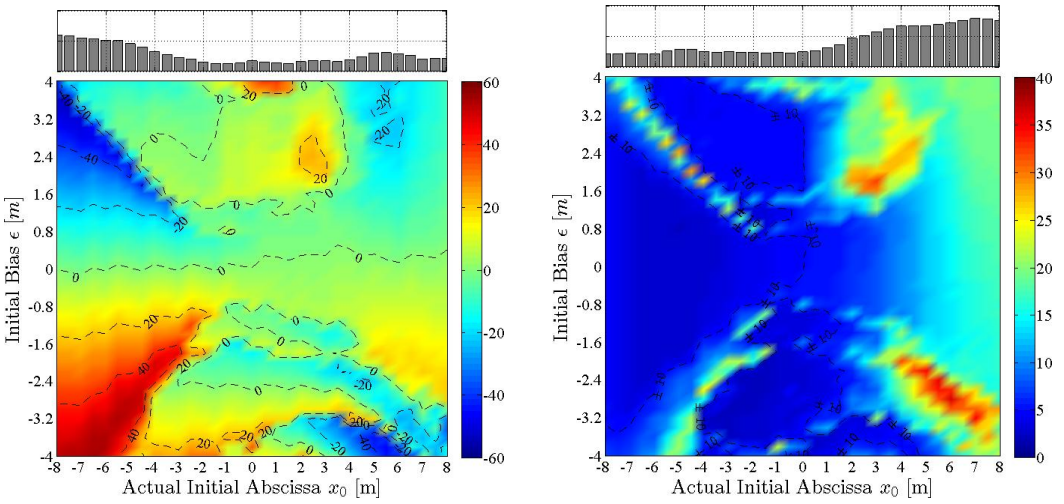
$x_0$ but also on bias introduced at initialisation such that:

$$\mu_{x,0} = x_0 + \epsilon. \tag{4.19}$$

In this test, $\epsilon$ varies from $-4$ m to $+4$ m with a step of 0.5 m. The condition to stop the tracking is that only 3 meters are observed each time, *i.e.* $x_T = x_0 + 3$. Here again $N_p = 200$ and $N_{test} = 200$. The results are stored into two matrices (actual initial abscissa $x_0$ in columns, initial bias $\epsilon$ in lines) whose elements are the global percentage error (in relative value), Fig. 4.10a, and the total standard deviation, Fig. 4.10b, of speed estimates respectively.

These plots reveal that knowledge of the initial abscissa is a very critical point. As expected, no problem occurs when the particle filter is initialized with the right initial abscissa ($\epsilon \approx 0$), whatever its initial value but the error quickly increases for an initial error of a few centimeters. An overestimation results from an underestimation of the initial distance and inversely. The summary of the absolute error and standard deviation are depicted for guidance at the top of each map. The preferred are those for which the values of the graph are minimal, namely, regions beginning between -2 m and +2m are those for which error on the *a priori* initial abscissa is the less penalizing.

It is interesting to observe that beginning the tracking too early induces accuracy problems (Fig. 4.10a when $x_0 < -2m$) and beginning the tracking too late induces precision problems (Fig. 4.10b when $x_0 > +2m$). Moreover, tracking a vehicle too early may compromise the assumption speed during the observation in practice. A mode detailed analysis of the distribution of the error shows that it is slightly better to overestimate the distance of the vehicle rather than the opposite.




(a) Error (km/h)

(b) Total Standard Deviation (km/h) and enhancing of ± 10 km/h interval confidence (CI95, km/h)

Figure 4.10: Influence of a false *a-priori* initial abscissa on the speed estimates when the vehicle is traveling at 75 km/h. Errors and standard deviations are expressed in km/h as a function of the actual vehicle position at $t$=0 (abscissa) and of the error on this position (ordinate). For both pictures, the absolute summation is depicted by a secondary axis on top. Specific CI95 at ± 10 km/h is enhancing in (b).

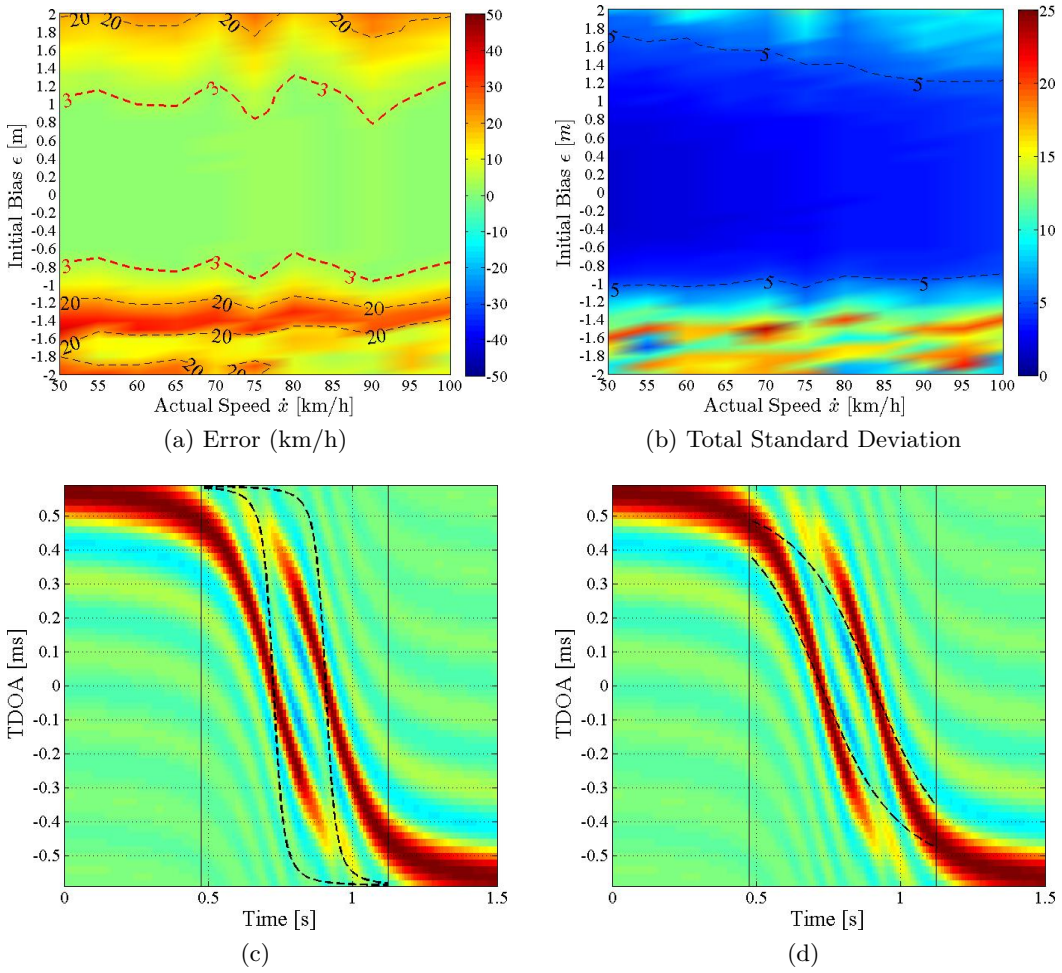
### 4.9.4 Influence of the *a priori* distance to the tyres

The distance between the microphone array and the nearest tyres is denoted as $D$ in the model depicted in Fig. 4.3. On the ground, $D$ can be roughly measured using a measuring tape or a laser range finder but this value actually varies from several tens of centimeters as the distance from the roadside is different for each motorist. In order to evaluate how this parameter influences the behavior of the BPF, a test is conducted using the same parameters as in Table 4.1 except that $\dot{x}$ and *a priori* target ordinate $\mu_{y,0}$ are now the variables. The former ranges from 50 km/h to 100 km/h, the latter is ruled by Eq.4.20 with the bias $\epsilon$ ranging from $-2$ m to $+2$ m.

$$\mu_{y,0} = D + \epsilon. \tag{4.20}$$

The actual distance to the road is set to $D = 2.5$ m, $N_p = 200$ and $N_{test} = 200$. Results are stored into two matrices (actual speed $\dot{x}$ in columns, initial bias $\epsilon$ in lines) whose elements are the global percentage error (in relative value), Fig. 4.10a, and the total standard deviation, Fig. 4.10b, of speed estimates respectively.

This is clearly demonstrated seeing both Fig. 4.11a and Fig. 4.11b that the accuracy and precision in speed estimates are not symmetrical in $\epsilon$. An underestimation of $D$ involves an estimation of lower quality rather than an overestimation. Fig. 4.11c and Fig. 4.11d help to understand this effect. In these plots, the observation of a vehicle




(a) Error (km/h) (b) Total Standard Deviation

(c) (d)

Figure 4.11: Top: Influence of the *a priori* initial ordinate $\mu_{y,0}$ on the speed estimates. Errors (a) and standard deviations (b) are expressed in km/h as a function of the actual vehicle speed $\dot{x}$ and of the initial bias $\epsilon$ [Eq. (4.20)]. Below: comparison between the observed CCTS and the model followed by the particles (black dashed lines) at initialisation when the initial ordinate $\mu_{y,0}$ is an underestimation (c) or overestimation (d) of the actual $D$. All other parameters: $\dot{x}$, $x_0$ and $wb$ are exactly known.

travelling at speed $\dot{x} = 50$ km/h at a distance $D = 2.5$ m from the microphone array and with a wheelbase $wb = 2.5$ is represented. The black dashed line corresponds to the initial model, that is, the trajectory that the particles should follow if the observation was not taken into account. In the initial model of Fig. 4.11c, $D$ is underestimated ($\mu_{y,0} = D - 2\text{m}$), and in Fig. 4.11d, $D$ is overestimated ($\mu_{y,0} = D + 2\text{m}$). In the first case, particles initially follow a horizontal line inducing a rapid loss of the observation. After some iterations, particle resampling is not ruled by the CCTS and the particles simply follow their initial model quite independently from the observation, which results in an overestimation of speed and large variance because of the stochastic nature of the process. In the second case, the model is incorrect again. But with respect to the observation,

the particles are more capable to focus on the observation as they quickly intersect the actual traces, giving a better result in both error and standard deviation in such a case.

### 4.9.5 Influence of interruptions of information

This last experiment explores how the BPF behaves when the observation is temporary unavailable. The duration of interruption is expressed in percentage of the total observation length. Two types of interruption are tested: the first one consists in replacing a part of the observation by a spatially consistent noise, see Fig. 4.12a, and the second one is a random noise, drawn from a zero-mean, unit variance Gaussian distribution, see Fig. 4.12b. Parameters are the same as in Table 4.1 except that the number of particles $N_p$ is set to 200. Results are depicted in Fig. 4.12.

In the case of a spatially consistent perturbation, Fig. 4.12c and Fig. 4.12d, performance begin to be undermined above 40% of missing observation. The estimates fall towards zero value in speed, which is, in a sense, correct because the noise which presents no TDOA evolution, acting like an immobile sound source.

In the case of an incoherent perturbation, Fig. 4.12e and Fig. 4.12f, the error increases when 60% of the observation is missing, and the CI95 is below 10% until 50% of observation at least. Above 80% of missing observation, the estimates fall towards the *a-priori* speed value , which is correct since the observation does not play a role anymore so that particles follow their initial model.

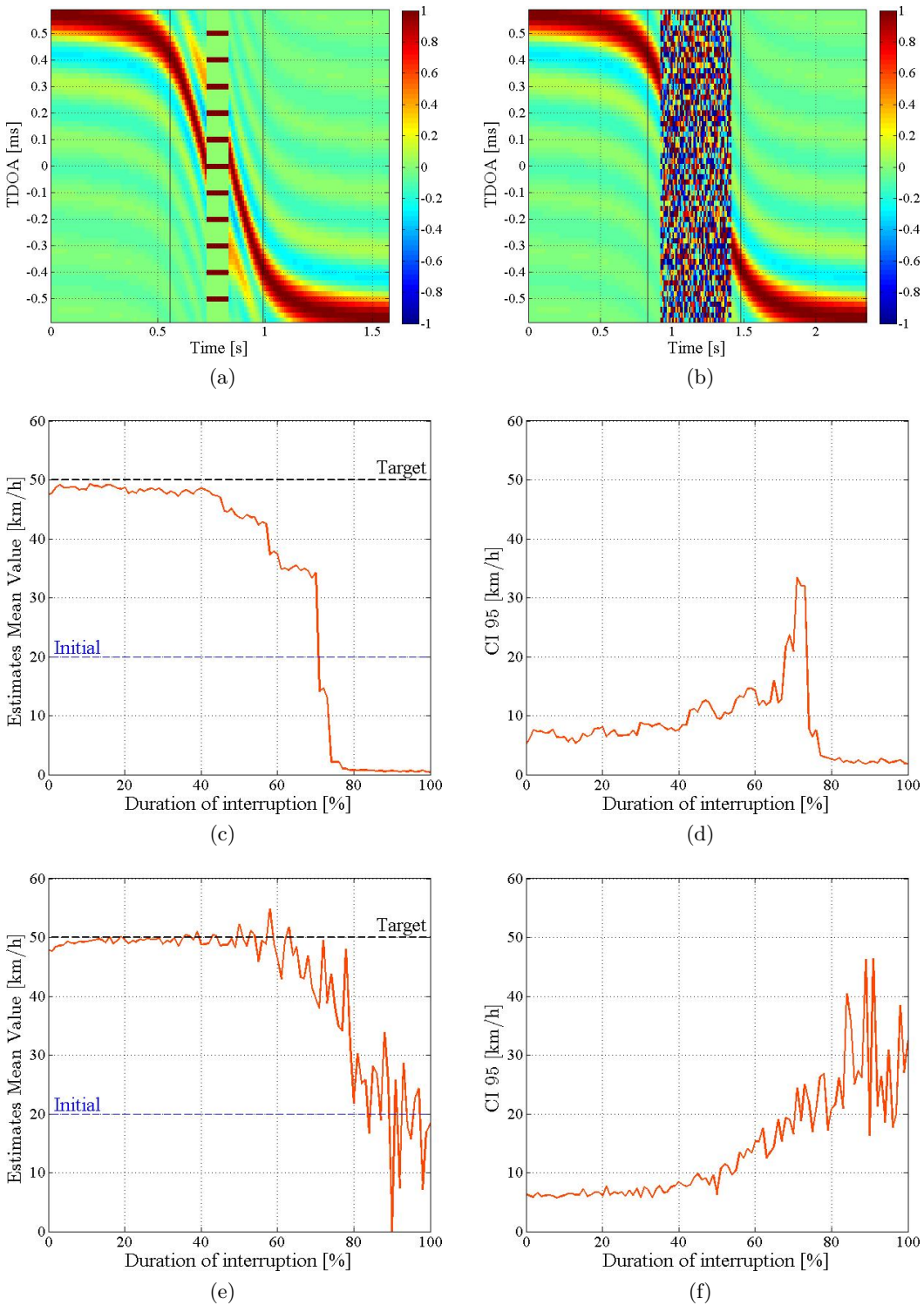



Figure 4.12: Top: examples of observations interrupted at (a) 25% by a spatially consistent noise and at (b) 75% by a random noise. Below: tracking performances as a function of the interruption length. (c) and (d): particles mean estimates and CI95 for spatially consistent noise, (e) and (f): particles mean estimates and CI95 for incoherent noise.

## 4.10 Conclusion

In this chapter, physical mechanisms at the origin of pass-by noise have been listed, confirming its broadband and stochastic nature. This latter point has also been experimentally verified with spectral and correlation analysis. It was also revealed that for sufficiently short observations, pass-by noise may be modeled as a succession of static and uncorrelated broadband sound sources. This validates the global strategy consisting in filtering over time successive location estimates in order to cluster coherently the measurements and dissociate vehicles. In the context of our application, pass-by noise is mainly due to the tyre/road interactions, suggesting the possibility of observing front and rear axles independently using a suitable correlation measure.

According to chapter 2, one of the most efficient correlation measure regarding the acoustical conditions is the phase-transform generalized cross correlation (GCC-PHAT). Looking more closely at the spectral content of interest, we proposed an optimal form of GCC-PHAT consisting in applying the PHAT processor onto a specific bandwidth, as large as possible, but allowing to discard coherent noises responsible for spurious peaks in the CCTS. We realized lately that such an optimization had already been proposed for speaker localization and water pipes leak localization. However, this chapter provides an analytical expression of this approach for the cases of one and two independent sources, allowing, for instance, to predict the observation shape as a function of the acoustical scenario and also assessing the tracking algorithm *in-silico*.

The generic particle filtering algorithm has also been improved by considering the acoustical and geometrical properties of our application. A new target model including the wheelbase length and taking into account of the change of axle acoustically dominant over the pass-by has been proposed. We called this new algorithm the *bimodal particle filtering* (BPF).

Finally, simulations were performed and permitted to evaluate the proposed $BPF$ with regard to some of the most important parameters such the number of particles, the uncertainty on the *a-priori* knowledge, and the quality of the observation (temporary interruption).

# 5 Specifications for the microphone array

## 5.1 Introduction

In the previous chapters, it has been shown that filtering successive generalized cross-correlation (GCC) estimate with a specific particle filter (PF) makes possible the joint estimation of speed and wheelbase length of vehicles as they pass by. This is due to the broadband nature of the predominant component of the pass-by noise which results from the tyre/road interactions.

In its simplest form, the proposed approach needs a pair of microphones placed on the roadside, in parallel to the road lane. But until now, nothing has been said on the optimal inter-sensor distance $d$, the influence of the number of sensors, as well as that of the array geometry, on the estimation procedure. This is the purpose of the present chapter to answer to these questions regarding the practical constraints, the acoustical conditions, and the objectives of the applied context.

To the best of our knowledge, no former studies focused on microphone array design with objective to estimate wheelbase length. Most of the time, vehicles are considered as a point emitter and the array design is limited to one or two sensors placed on the shoulder of the road, generally without any justification of the inter-sensor distance (if given). One and two-sensor arrays are extensively found in the literature, especially for classification and motion parameter estimation [15, 17, 125, 35, 126, 127, 128, 36, 129, 12, 16, 30, 31, 32, 130]. Detection and localization are handled using more complex, but always planar and compact arrays, like linear [20, 21, 23], circular [115] or crossed ones [59]. Other types of arrays that are beyond the scope of this thesis can be evoked. They are distributed arrays and/or large aperture arrays and/or arrays comprising a large number of sensors and/or arrays at a height of a few meters [18, 19, 131, 132, 22]. Authors mainly focus on localisation and extraction problems by investigating spatial filtering.

Mathematically speaking, it is a well-known result [133] that the optimal microphone

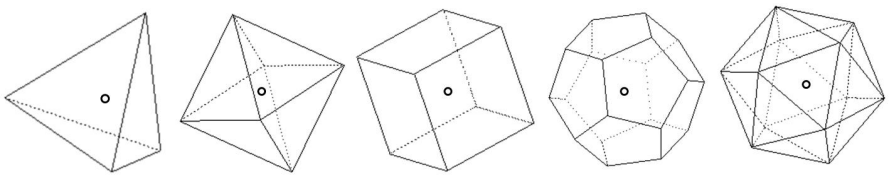


Figure 5.1: Platonic solids. From left to right: tetrahedron, octahedron, hexahedron (cube), dodecahedron and icosahedron.

arrangement for TDOA-based SSL consists in placing the sensors according to a *Platonic solid* with the target at the center, as depicted in Fig. 5.1. This is the geometry which enables the maximal reduction of variance in source position estimates. The main problem is that such a geometry is difficult, if not impossible, to achieve in an RTM context and also quite far from the primary objective being to develop a small, light and easily movable device, namely, what should be called a *compact* array[1]. Moreover, the platonic solid is the optimal arrangement for the single static source localization problem, but nothing is said about monitoring two moving sound sources at the same time. In the present case, we are looking for the optimal $d$ for which the two traces inherent to the rear and front axles in the CCTS are clearly depicted. The purpose of this chapter is to present a design methodology, first by specifying the optimal inter-sensor distance, then by discussing the required number of sensors.

## 5.2 Inter-sensor distance

According to Eq. (2.11), the reliability of a GCC-based time-delay estimator depends on the characteristics of the peaks (width, emergence and spacing) in the correlation measure. Let us recall that a closed-form expression of this correlation measure is given in Eq. (4.8). The characteristics of the peaks are dependent of the spectral properties ($B_w$, $f_c$) of the sources and the geometrical parameters ($x_0$, $D$, $w_b$, $d$) of the scene. In-situ, distance to the road $D$ and inter-sensor distance $d$ are the only modifiable parameters, except for normative measurements where $D$ is imposed, such as for instance in [135]. Thus, what we propose is a strategy aiming at optimizing $d$ in order to discriminate the two peaks related to the two axles as best as possible.

### 5.2.1 Cramer-Rao Lower Bound

The Cramer-Rao lower bound (CRLB) defines the best performance than can be achieved by an unbiased estimator. Its relatively simple form is used for many engineering problems in which a parameter vector must be estimated from observations depending

[1]The term "compact array" traditionally refers to an array with inter-sensor distances much smaller than the smallest acoustic wavelength that has to be processed [134]. In this work, wavelengths of interests vary from 7 cm to more than one meter. We will see later that such a definition of the compactness is not respected. However, the term compact is used here to mark the difference with distributed arrays, as Platonic ones, for which sensors are separated by much larger distances.

on another parameter. The CRLB is defined as the inverse of the Fisher information matrix. Compute this matrix and maximise it enables the parameter vector delivering the observation to be optimized in order to deliver the best estimates as possible. As a first attempt, this standard method has been derived and results are discussed in the following.

The parameter to estimate is $wb$ and the parameter to optimize is $d$. The available measurements are $\tau_{12,1}$ and $\tau_{12,2}$, respectively denoted $\tau_1$ and $\tau_2$ for clarity below. Let us consider the following relations between actual delays and their estimates:

$$\tau_1 = \hat{\tau}_1 + n_1, \tag{5.1}$$
$$\tau_2 = \hat{\tau}_2 + n_2, \tag{5.2}$$

where $\hat{\tau}_j$ is an estimate of $\tau_j$ and $n_j$ is a zero mean Gaussian noise with variance $\sigma_j^2$ denoting the uncertainty on the measurements, $j \in [1, 2]$. Thus, $\hat{\tau}_2$ can be expressed as a function of $\hat{\tau}_1$ and $wb$:

$$\hat{\tau}_2 = f(\hat{\tau}_1, wb). \tag{5.3}$$

According to the model in Fig. 4.3, one gets:

$$\tan\theta_1 = \frac{x_0}{D}, \tag{5.4}$$
$$\tan\theta_2 = \frac{x_0 - wb}{D}. \tag{5.5}$$

Substituting Eq. (5.4) into Eq. (5.5) gives:

$$\tan\theta_2 = \tan\theta_1 - \frac{wb}{D}, \tag{5.6}$$

where $\theta_k^{(n)}$, $k \in [1, 2]$, is expressed by :

$$\theta_k = \arcsin\left(\frac{c\tau_k}{d}\right). \tag{5.7}$$

Substituting Eq. (5.7) into Eq. (5.6) gives:

$$\tan\left(\arcsin\left(\frac{c\tau_2}{d}\right)\right) = \tan\left(\arcsin\left(\frac{c\tau_1}{d}\right)\right) - \frac{wb}{D}, \tag{5.8}$$

yielding:

$$f(\hat{\tau}_1, wb) = \frac{d}{c}\sin\left\{\arctan\left[\tan\left(\arcsin\left(\frac{c\hat{\tau}_1}{d}\right)\right) - \frac{wb}{D}\right]\right\}. \tag{5.9}$$

The Fisher information matrix is given by [136] page 47:

$$F = A' \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} A, \tag{5.10}$$

where

$$A = \begin{pmatrix} \partial\tau_1/\partial\hat{\tau}_1 & \partial\tau_1/\partial wb \\ \partial\tau_2/\partial\hat{\tau}_1 & \partial\tau_2/\partial wb \end{pmatrix}, \tag{5.11}$$

$$= \begin{pmatrix} 1 & 0 \\ \partial f/\partial\hat{\tau}_1 & \partial f/\partial wb \end{pmatrix}. \tag{5.12}$$

The optimal $d$ maximizes the determinant of $F$ (D-optimality criterion) [137]. The determinant of $F$ is given by:

$$|F| = |A|^2 \sigma_1^2 \sigma_2^2. \tag{5.13}$$

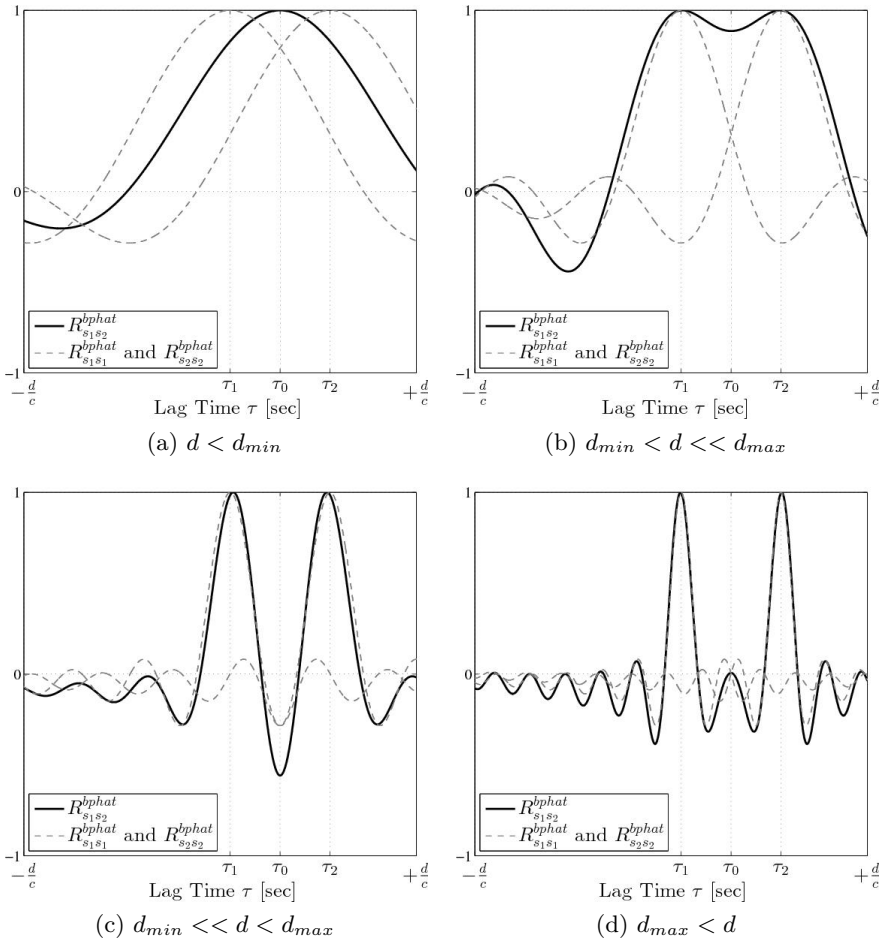
Maximizing (5.13) is the same as maximizing $|A|^2 = (\partial f/\partial wb)^2$ with respect to $d$. This quantity is expressed by:

$$\left(\frac{\partial f}{\partial wb}\right)^2 = \left(\frac{d}{cD}\right)^2 \underbrace{\left( \frac{\left(wb\sqrt{d^2 - c^2\hat{\tau}_1^2} - cD\hat{\tau}_1\right)^2}{D^2\left(d^2 - c^2\hat{\tau}_1^2\right)} + 1 \right)^{-3}}_{\xi}. \tag{5.14}$$

Since $d^2 \geq c^2\tau^2$, the term $\xi$ is positive whatever the value of $d$. It comes that the larger the value of $d$, the better the estimate of $wb$. But this result is not satisfactory for many practical reasons. The main one is that the larger the value of $d$, the lower the correlation between sensor signals is. Thus, what we propose is to assess how the parameter $d$ influences the shape of the observation model (4.8).

### 5.2.2 Minimal and maximal inter-sensor distance

Because of the additive effect, due to the sum operator in Eq. (4.8), axles cannot be distinguished for very small values of $d$ and phantom sources (spurious peaks) appear for very large values of $d$. Such an effect is depicted in Fig. 5.2. For all plots, the acoustic scenario is the same, $d$ being the only variable. The GCC-BPHAT function and the primary correlations are drawn in black and gray respectively. The actual TDOAs $\tau_1$ and $\tau_2$ and their average value $\tau_0$ are also represented. In Fig. 5.2a, $d$ is so small that it is impossible to predict the existence of the two sources. In Fig. 5.2b, both peaks begin to appear since $d$ has been increased. In Fig. 5.2c, $d$ has been increased again and both peaks are clearly distinct. In Fig. 5.2d $d$ has been increased again and both peaks are well distinguished but one spurious peak appears at $\tau_0$.




(a) $d < d_{min}$ (b) $d_{min} < d << d_{max}$

(c) $d_{min} << d < d_{max}$ (d) $d_{max} < d$

Figure 5.2: Illustration of the additive effect in (4.8) as a function of the inter-sensor distance $d$.

As spurious peaks do not have any physical meaning here, it is always better to avoid them because of possible misinterpretations, especially when it comes to estimating the number of axles during pass-by. Consequently, the inter-sensor distance should be limited to values between a minimal distance $d_{min}$, above which both axles are distinct, and a maximal distance $d_{max}$, below which no spurious peaks appear. Inspired by Fig. 5.2, the two peaks are distinct once $R^{bphat}_{s_1s_2}(\tau)$ is locally convex around $\tau_0$, yielding an implicit expression of $d_{min}$:

$$d_{min} = \arg\min_{d>0} \left(g_{\tau_0} > 0\right), \tag{5.15}$$

where

$$g_{\tau_0} = \left.\frac{\partial^2 R^{bphat}_{s_1s_2}(\tau)}{\partial \tau^2}\right|_{\tau_0}. \tag{5.16}$$

Similarly, the condition for avoiding a central spurious peak is that $R^{bphat}_{s_1s_2}(\tau)$ is not convex

anymore around $\tau_0$ for larger values of $d$. An implicit expression of $d_{max}$ is therefore:

$$d_{max} = \arg \min_{d > d_{min}} \left( g_{\tau_0} < 0 \right). \tag{5.17}$$

To conclude, the domain $[d_{min}, +\infty[$ defines what one can call a *range of bimodality detection*, that is, the set of inter-sensor distances for which the two peaks are observable. But in order to avoid central spurious peaks, one needs to restrict this range to $[d_{min}, d_{max}]$. We called this domain the *range of undistorted bimodality* (RUBI).

### 5.2.3 Range of undistorted bimodality

According to Eq. (5.15) and Eq. (5.17) and considering a given acoustic scenario (fixed value of $D$, $w_b$ and $x_0$), the sign of $g_{\tau_0}$ may be expressed as a function both of the spectral properties of the BPHAT transform $(B_w, f_c)$ and the inter-sensor distance $d$ thanks to Eq. (4.8), (5.15) and (5.17). This is what Fig. 5.3 illustrates. The vertical and horizontal axis have been specifically chosen for the sake of generalization so that spectral values are not necessarily acoustic values. This is the reason why $d$ is normalized by the halved central wavelength $\lambda_c = c/f_c$. This plot has been generated using arbitrary geometrical parameters: $w_b = 2.47$ m, $D = 6.3$ m and $x_0 = 0$ m. Grey zones (respectively white zones) correspond to a negative sign (respectively positive sign) of $g_{\tau_0}$. The six plots on the right of Fig. 5.3 show the GCC-BPHAT at different points of the abacus (A,B,C,D,E and F).

Consider a BPHAT transform between 250 Hz and 4750 Hz, i.e. $B_w/f_c = 1.8$. In zone I, the two peaks are undetectable (point A). They begin to appear at the boundary between zone I and zone II (point B). The two peaks are clearly distinct in middle of the zone II (point C). Then, in zone III, IV and upper, secondary lobes appear around $\tau_0$ (point D, E, F). So, in this example, the RUBI is delimited by B and D (between 12 cm and 34 cm) and the optimal distance $d_{opt}$ is somewhere within this range.

In Fig. 5.4, the same scenario as above is considered, except that the variable is now the DOA $\theta$ of the center of the vehicle (at coordinate $[x_0 + wb/2,\ D]$) instead of the ratio $B_w/f_c$, the latter is fixed here to 1.8 for the whole plot. By considering the zone II, one can see that the opening angle in which bimodality is observable is more or less wide depending on $d$. For instance, setting $d = 5\lambda_c/2$ allows a bimodal tracking on an angle range of about 90° (±45°) as depicted by points A, B, and C. Reducing $d$ to $3\lambda_c/2$ will reduce the observation area to nearly 70° (±35°) as depicted by points D, E and F.

### 5.2.4 Optimal inter-sensor distance

The objective of this section is to find, given a scenario ($B_w$, $f_c$, $D$), which value of $d$ within the RUBI $[d_{min}, d_{max}]$ enables the best wheelbase length estimation ? According to section 5.2.1, the CRLB-based method would answer $d_{max}$. The true answer is in practice a little more complex, mainly because of two points. Firstly the model (5.1)-(5.2)

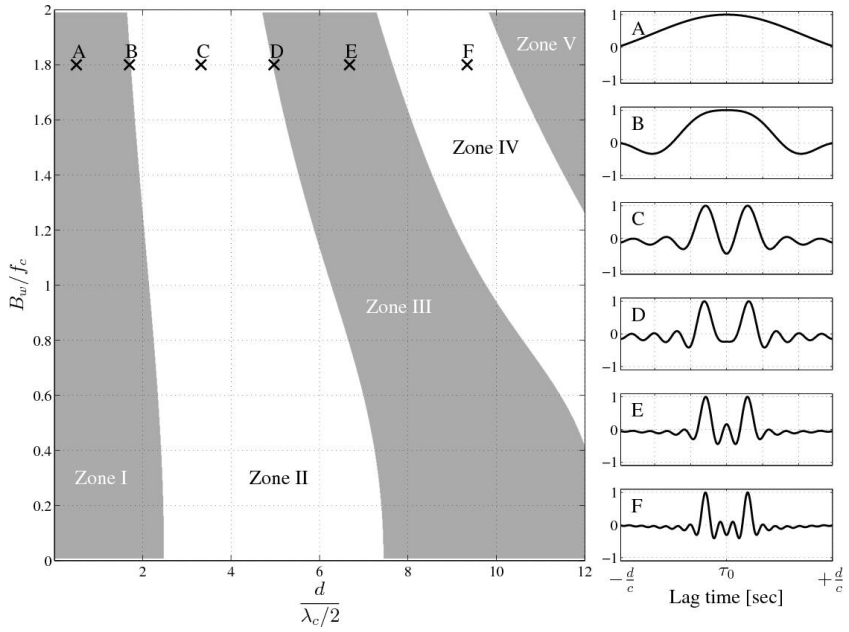



Figure 5.3: Sign of $g_{\tau_0}$ [Eq. (5.16)] as a function of the spectral properties of the BPHAT transform (*i.e.* $B_w$, $f_c$, $\lambda_c$) and the inter-sensor distance $d$. Grey (resp. white) areas correspond to a negative (resp. positive) sign.

supposes that the two TDOAs inherent to front and rear axle are observable. Actually, the additive effect described in section 5.2.2 induces a bias so that, except for specific values of $d$, the peaks never correspond to the actual TDOAs when they are more than one. Secondly, TDOAs are not estimated after a peak-picking procedure in the present case, but after a Monte-Carlo-based procedure. This allows to find the peaks without any thresholding steps. However, deriving a mathematical formalism linking the performance of the convergence to $d$ is rather tricky.

As an alternative to the conventional method, an ad-hoc optimization procedure in which the data processing algorithm itself (in this case, the particle filtering) is taken into account is proposed in the next paragraph.

Let us consider the illustrative example depicted in Fig. 5.5. Two different theoretical observations (i.e. GCC-BPHAT functions depicted in black) are considered. They differ by the value of $d$: on top, Fig. 5.5a or Fig. 5.5b, $d$ belongs to the RUBI, and below, Fig. 5.5c or Fig. 5.5d, $d > d_{max}$ so that a spurious peak appears at $\tau$=0. All other parameters are the same for both cases, they are $f_c = 2500$ Hz, $B_w = 1.8 f_c$, $w_b = 2.47$ m, $D = 6.3$ m and $x_0 = w_b/2$. Note that the value of $x_0$ implies that $\tau_2 = -\tau_1$ (vehicle in the broadside direction). That is why only the positive part of the observation is represented.

On the left side, a particle set at initialisation is depicted. The distribution is uniform over the range of possible delays, *i.e.* between 0 and $d/c$. On the right side, the same set of particles is depicted after one multinomial resampling. One can observe that in Fig. 5.5b, all particles coalesce around the target value, but in Fig. 5.5d, some particles

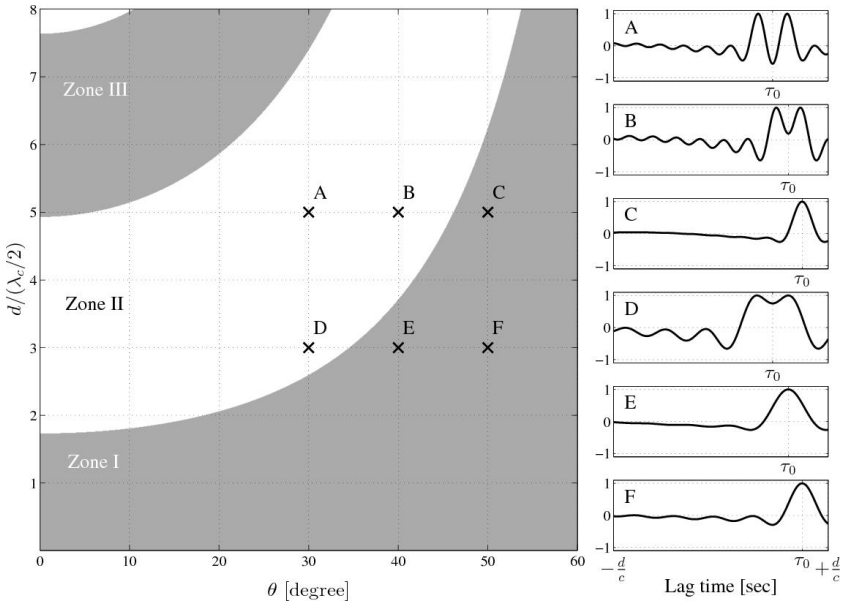



Figure 5.4: Sign of $g_{\tau_0}$ [Eq. (5.16)] as a function of the inter-sensor distance $d$ (normalized by the halved wavelength) and the vehicle direction of arrival $\theta$ in degree. The ratio $B_w/f_c$ is set to 1.8.

also coalesce around the central spurious peak. In the latter case, particles are separated into two groups (one around the correct peak, the other around the spurious peak). The convergence is not as satisfactory as in the first case, especially when an estimate is returned by taking the mean of the particles.

Reiterating the procedure many times (for instance 100 or 200 times) permits to derive some statistics (global coefficient of variation, global mean percentage error) and quantify the convergence of the particles for each tested $d$. This is the idea explained in more detail by the algorithm 2.

As previously demonstrated, zone I should not be considered because of the non-observability of the two peaks ($d < d_{min}$). Global mean percentage error and global coefficient of variation are logically high in this area. From the beginning of the zone II (RUBI), both the accuracy and repeatability of the estimator increase. As predicted by the Fisher information matrix, the general trend is that the larger the inter-sensor distance, the better the estimate. However, with the proposed approach, one local minimum appears within the RUBI suggesting that, in the present case, setting $d = 2\lambda_c < d_{max}$ provides a better estimator than setting $d = d_{max}$. Hence, by integrating both the analytical model of the correlation measure and the Monte-Carlo-based tracking process in the optimization procedure, a much more adapted design is obtained in comparison with deriving the CRLB.

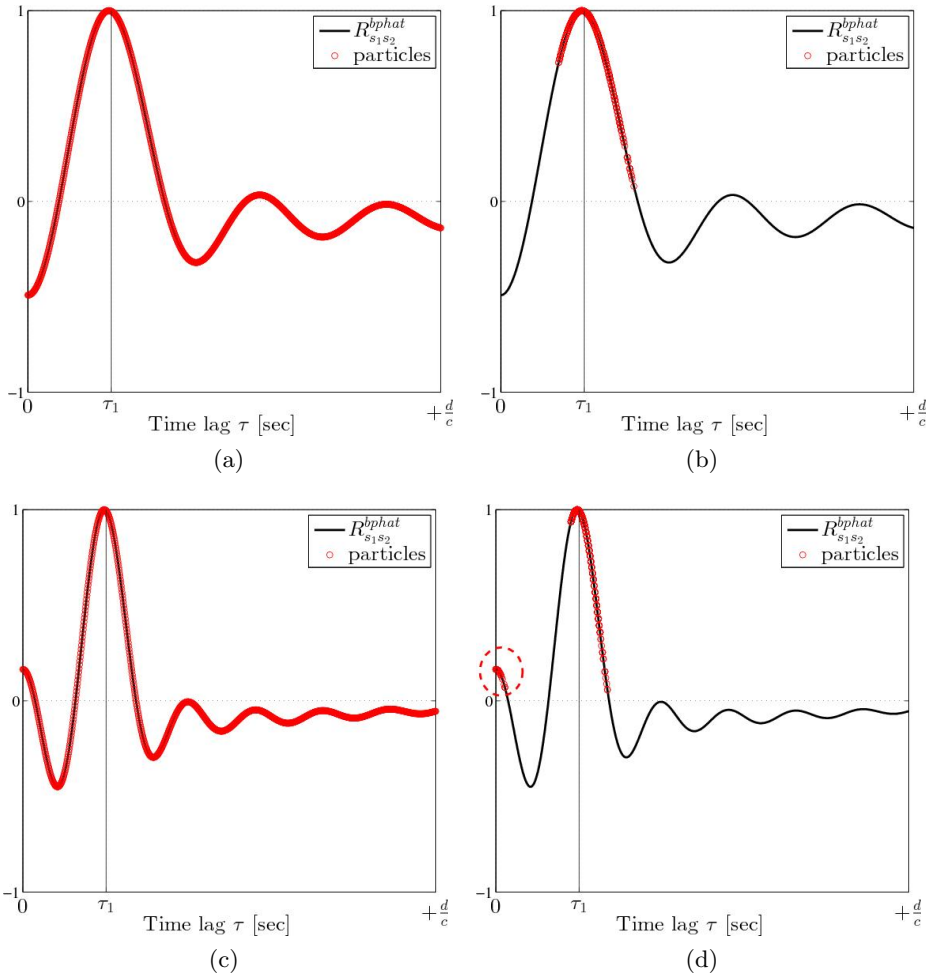



Figure 5.5: Effect of a spurious peak on the particles distribution.

### 5.2.5 Experimental measurement

A preliminary experiment has been carried out to confront the theoretical RUBI with one in-situ measurement. A car was equipped with two loudspeakers, each being fixed in front of a wheel of the left side, see Fig. 5.7a. Each loudspeaker emitted a white noise, independent with the other. The wheelbase of the car was of $w_b = 2.47$ m. A linear array was located on the roadside at a height of 80 cm and at a distance $D = 6.3$ m to the loudspeakers during pass-by. The array was composed of 7 microphones allowing different pairs from 7 cm to 50 cm, Fig. 5.7b. The sensors were 1/4" omnidirectional ICP microphones from *PCB Piezoelectronics*. The vehicle speed was nearly 60 km/h during the measurement. The recording was collected on November 2, 2012 on the EPFL Campus (Lat. 46°31'7.74"N, Long. 6°33'56.39"E). The location was free for reverberation but quite noisy because of a demolition site 150 meters away and a light wind (20 km/h in average). The sky was clear and the temperature was 17°C.

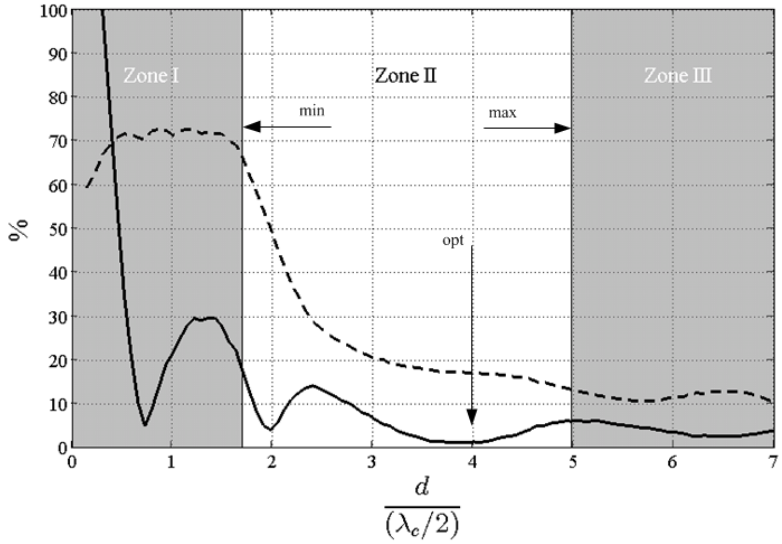



Figure 5.6: Mean percentage error (thick line) and mean coefficient of variation (dashed line) of TDOA estimation as a function of $d$, both expressed in %.

One BPHAT-CCTS per pair ($B_w/f_c = 1.8$, $f_c = 2500$ Hz) was computed. Some examples are depicted in Fig. 5.8 and Fig. 5.9. Red, respectively green, lines represent the period of time during which the vehicle is in the 60° opening angle ($-30° \leq \theta \leq +30°$), respectively 90° opening angle ($-45° \leq \theta \leq +45°$).

From Fig. 5.3, the minimal inter-sensor distance respects the equality $d/(\lambda_c/2) \approx 1.8$, *i.e.* $d_{min} \approx 12$ cm in the present case. In Fig. 5.8a and Fig. 5.8b, $d$ equals 9 cm and 10 cm respectively. As expected, front and rear axles are not dissociated at all. On Fig. 5.8c, $d$ equals 12 cm and one can perceive the very beginning of the separation of the two traces. This is confirmed by Fig. 5.8d and Fig. 5.8e in which $d$ equals 14 cm and 18 cm respectively.

From Fig. 5.4, the distance enabling the dissociation of axles over an opening angle of 60° respects the equality $d/(\lambda_c/2) \approx 2.8$, $d \approx 19$ cm . This is a rather good prediction regarding Fig. 5.8f and Fig. 5.9a in which $d$ is equal to 19 cm and 21 cm respectively: the traces are well separated from one red line to the other. Similarly, covering an opening angle of 90° requires $d$ to be 31 cm. One observe such an objective is actually achieved for a lower inter-sensor distance, for instance in Fig. 5.9c with $d$ equals tos 28 cm.

From Fig. 5.3, the maximal inter-sensor distance respects the equality $d/(\lambda_c/2) \approx 5$, *i.e.* $d_{max} \approx 34$ cm in the present case. This is clearly demonstrated by inspecting Fig. 5.9e for which $d = 33$ cm and Fig. 5.9f for which $d = 40$ cm that in the first case no spurious peak appears between both traces, in opposition to the second case in which a third "phantom axle" appears between the two actual ones.

Finally, from Fig. 5.6, the optimal inter-sensor distance respects the equality $d/(\lambda_c/2) \approx 4$, *i.e.* $d_{opt} \approx 27$ cm. Indeed, one can conclude that the best contrast is achieved for $d = 28$


(a) (b)

Figure 5.7: Experimental setup. (a) car equipped with two loudspeakers, (b) linear array.

cm in this test, as shown in Fig. 5.9c.

**Algorithm 2** Proposed assessment of a candidate distance $d$.

**For** $k = 1$ **to** $k = 200$
*Initialisation*

- Initialize the particles with a uniform distribution over the set of possible delays: $\alpha_0 \sim \mathcal{U}(0, d/c)$;
- Attribute the same weight to all particles: $\forall n \in [1, 2, ..., N_p]$, $w_0^{(n)} = 1/N_p$;

*Weighting*

- Weight the particles according to the observation: $\forall n \in [1, 2, ..., N_p]$, $\tilde{w}_0^{(n)} = w_0^{(n)} R_{s_1 s_2}^{bphat}(\alpha_0^{(n)})$;
- Normalize the weights: $\forall n \in [1, 2, ..., N_p]$, $w_0^{(n)} = 1/\sum_{n=1}^{N_p} \tilde{w}_0^{(n)}$;

*One resampling step*

- Resample the particles according to their weights using the multinomial resampling , this returns a new set of particles $\alpha_1$;

*Assessment of the $k^{th}$ run*

- Compute the mean percentage error: $mpe_{d,k} = \dfrac{1}{N_p} \sum_n \dfrac{\alpha_1^{(n)} - \tau_1}{\tau_1}$;
- Compute the coefficient of variation: $cv_{d,k} = \dfrac{\sqrt{\frac{1}{N_p} \sum_p \left( \alpha_1^{(p)} - \frac{1}{N_p} \sum_n \alpha_1^{(n)} \right)^2}}{\frac{1}{N_p} \sum_n \alpha_1^{(n)}}$;

**endfor**
*Output of the algorithm*

- Compute the averaged percentage error: $MPE_d = \dfrac{1}{200} \sum_{k=1}^{200} mpe_{d,k}$;
- Compute the averaged coefficient of variation: $CV_d = \dfrac{1}{200} \sum_{k=1}^{200} cv_{d,k}$.

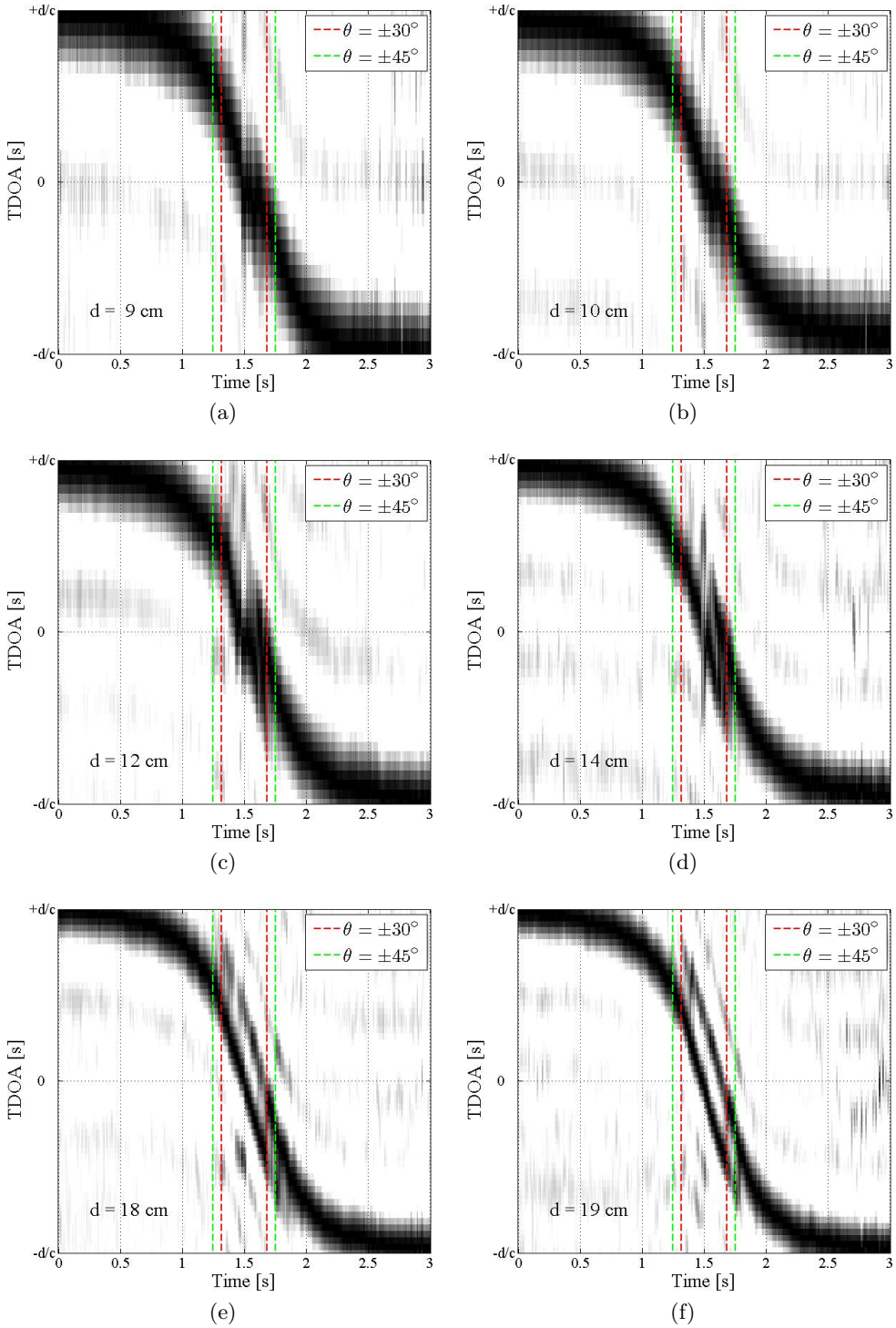



Figure 5.8: Real BPHAT-CCTS as a function of the inter-sensor distance (1/2).

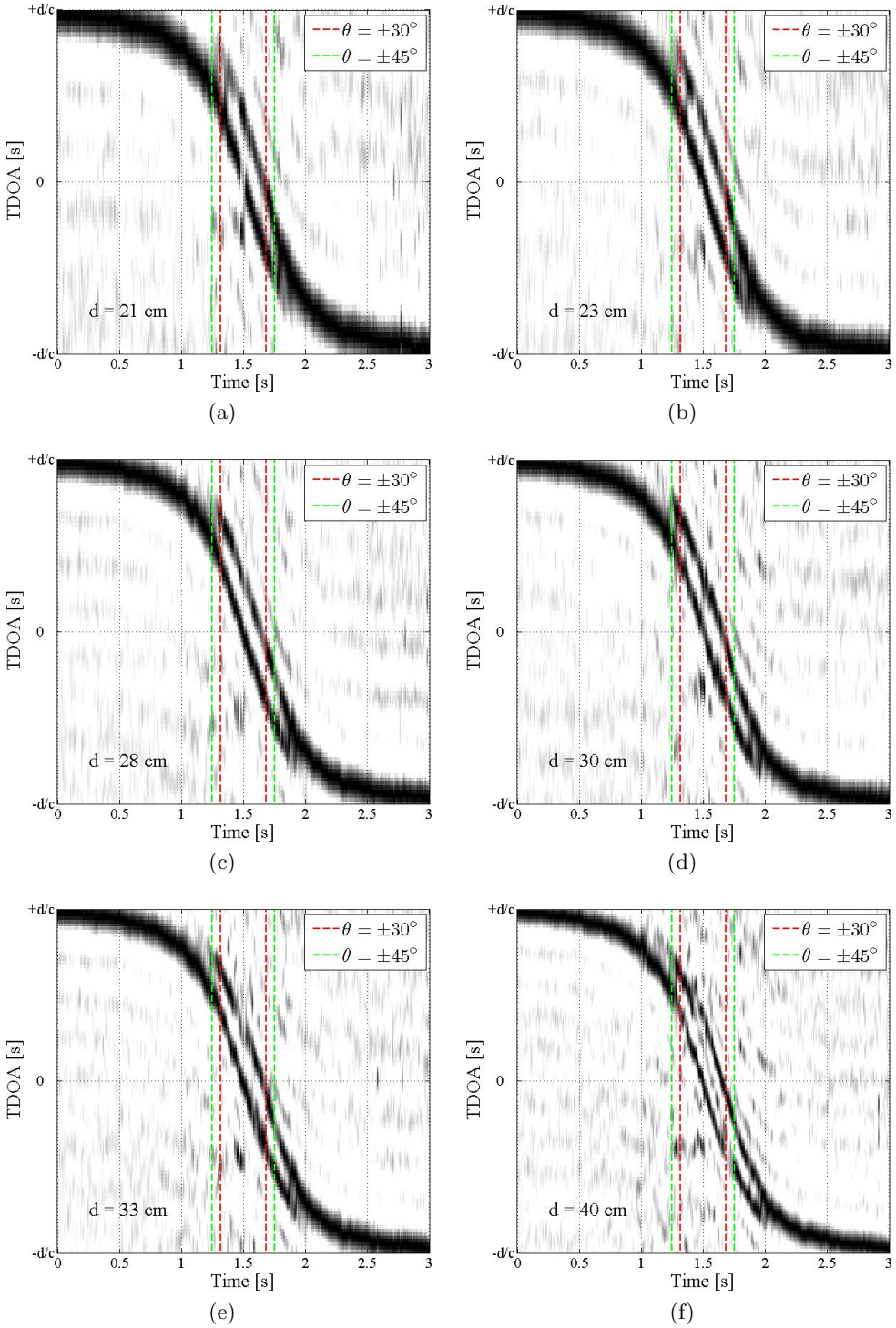
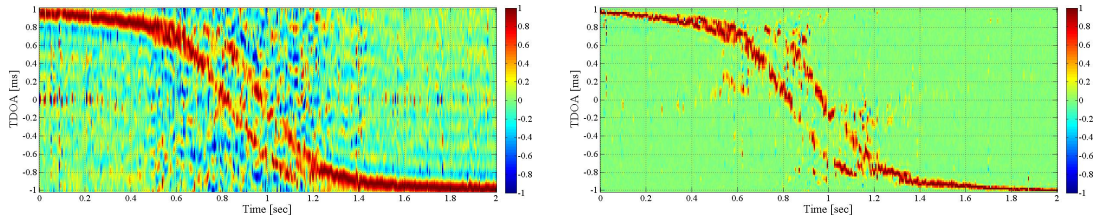



Figure 5.9: Real BPHAT-CCTS as a function of the inter-sensor distance (2/2).




(a) (b)

Figure 5.10: BPHAT-CCTS achieved using a single pair (a) and the three pairs (b) of an equilateral triangle shaped array.

## 5.3 Number of sensors

It is known that the TDOA-based localization in 2 (or $n$) dimensions needs at least 4 (or $n+2$) sensors for solving all the spatial ambiguities that might occurs [138]. Using prior knowledge on the source positions, these ambiguities may be reduced and thus, the number of required sensors. In the present context, the microphone array is placed on the shoulder of the road and vehicles are theoretically constrained by the road path. As a consequence, two microphones placed in parallel to the road lane are sufficient to estimate, without any ambiguity, the position (DOA) of a vehicle as it passes by.

However, the main risk when using a single sensor pair is that interfering noise sources coming from the rear of the array (agricultural machinery, animals, other road etc.) are mixed up with useful signal observations. Also, replacing $\theta$ by $\pi - \theta$ in Eq. (2.5), both DOA produces the same $\tau_{12}$. This ambiguity remains unsolved whatever the number of sensors if they are all aligned.

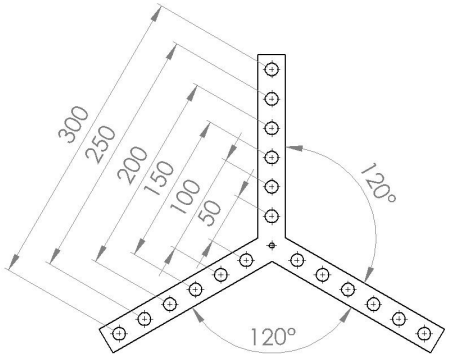
The solution that was retained to counteract this effect consists in adding a third microphone in the horizontal plane to form an equilateral triangle array with the two first ones.

This three-element array produces three CCTS (one by pair) which can be combined to improve the observation contrast by taking advantage of the measurement redundancy. In our experience, the combination of CCTS using the MULTI-PHAT technique provides impressive results in our context (see Appendix A.4 ).

This is what is illustrated in Fig. 5.10. Two BPHAT-CCTS of the same vehicle passage ($d = 35$ cm, $f_c = 2500$ Hz, $B_w = 4500$ Hz) are depicted using a single sensor pair in parallel to the road lane (Fig. 5.10a), and the MULTI-PHAT technique applied to the three pairs of an equilateral shaped array (Fig. 5.10b). It is clear on these example that taking advantage of the redundant information by adding supplementary sensors drastically improves the estimation.

The prototype developed at the end of this thesis is shown in Fig. 5.11a. It consists of a camera tripod and a home-made plexiglass holder. The camera tripod easily enables the array to be hung to urban furniture. The plexiglass holder is composed of multiple holes

(a)




(b)

Figure 5.11: (a) microphone array prototype, (b) microphone holder aperture (in mm)

to vary the array aperture with respect to the scenario, Fig. 5.11b.

**Remark** It is clear that the presented prototype is solely dedicated for research purposes. For instance, it needs to be connected to an independent data acquisition system and is not designed to resist harsh weather conditions. However, some efforts in this direction have been started and a brief review of techniques permitting to counteract the wind noise is proposed in the appendix A.7.

## 5.4 Conclusion

This chapter discusses the inter-sensor distance optimization for the observation of the front and rear tyre-asphalt interactions from cross-correlation measurements. According to the CRLB derivation of model (5.1), the larger the distance the better the estimation is. But this result cannot be blindly applied as the correlation of sensor signals has to be maintained, discarding the use of a too large aperture array. A heuristic methodology has therefore been proposed consisting in i) expressing the closed-form expression of the observation [done in chapter 4], ii) defining a range within which the inter-sensor distance must be contained, iii) filtering the modeled observation with a sequential Monte-Carlo method for each inter-sensor distance within this range and iv) looking at which candidates yield the most accurate and repeatable time-delay estimates.

In addition, we argued in favor of a third microphone, added to the theoretically sufficient 2-element array, and the use of the MULTI-PHAT technique, as a mean to exploit the information redundancy between sensor pairs, so as to improve the robustness of time-delay estimates. It has been demonstrated on a real example, and will be confirmed in the next chapter, that using three sensors instead of two effectively provides much less noisy observations. This is due to the exclusive effect of the MULTI-PHAT technique

which is discards most of the incoherent spatially interfering noises.

# 6 In-situ measurements: validation of the methods

## 6.1 Introduction

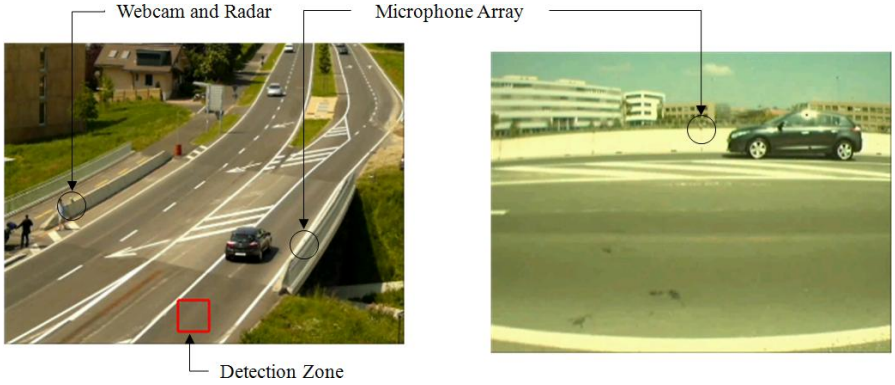
In the previous chapters, a new tracking algorithm dedicated to jointly estimate speed and wheelbase length of two-axle vehicles was presented, the bimodal particle filtering (BPF). As the BPF is fed by observations of a microphone array, a design strategy was proposed to provide the best pass-by noise measurements, under the contraint of a small, easily movable array. In this chapter, some experimental results are presented and discussed. They were post-processed on two databases collected near the EPFL campus (EPFL database) and in St-Maurice, Switzerland (St-Maurice database). Both are presented below.

#### EPFL database

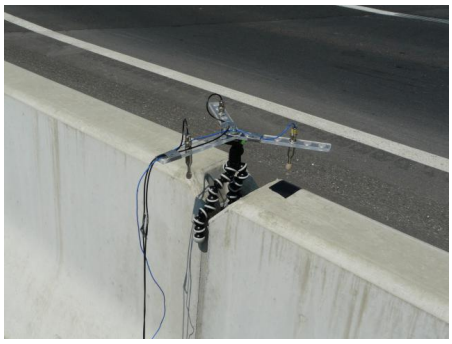
This database was collected on $25^{th}$ May 2012 at the Route Cantonale of Ecublens, near the EPFL campus, Switzerland (Lat. 46°31'0.28"N, Long. 6°33'50.41"E). The triangular microphone array was disposed on the roadside at a height of 84 cm, an average distance of $D = 2.5$ m to the vehicles closest wheels, and an inter-sensor distance $d$ of 20 cm. The three sensors used were 1/4" omnidirectional ICP microphones from *PCB Piezoelectronics*. The array was situated between a traffic roundabout (120 meters upstream) and a traffic light (345 meters downstream). Vehicle speed ranged between 50 km/h to 75 km/h. The speed limitation is officially 70 km/h, but it is not uncommon that users slow down seeing staff and equipment on the roadside. The location is free from reverberation, the nearest building being distant of 30 meters. The day was warm and windless, and the sky was generally clear. A view from the sky of the location provided by the Google Earth database is depicted in Fig. 6.1a.

The audio signals were collected using the NetdB acquisition device from *01dB-Metravib* (today *ACOEM*). The sampling rate was 51.2 kHz, the quantification was 24 bits and it was made sure that all tracks were acquired synchronously. A standardized radar

(a)




(b)



(c)

Figure 6.1: Experimental setup of the "EPFL-Database". (a): view from the sky of the setup emplacement (M: microphone array, R: radar + webcam, C: in height camera), (b): views of the two cameras (top and side), location of the microphone array and radar are highlighted by black circles, (c): zoom on the microphone array.

Doppler type Viacount II[1] was set up on the opposite shoulder. The Viacount II is a

[1] kindly lent by the Swiss society *ViaTraffic*

professional traffic counter device providing speed (in km/h), direction (sign of the speed) and length (in number of reflected pulses) of vehicles. The scene was continuously filmed by two cameras, one placed on the road side near the radar to get a view of the sides of all the vehicles and another placed on the balcony of a nearby building to get a more global view of the scene. Both devices produced video at 30 frames per second. Fig. 6.1b depicts the two views provided by cameras and the location of the microphone array and radar.

Only the right-hand traffic lane is considered in this experiment, namely the lane where a black vehicle is present on Fig. 6.1b. Audio and video signals were synchronized off-line thanks to a pre-measurement consisting in broadcasting the same radio FM program close to each device. An home-made detection algorithm was implemented to return the apparition time of each new vehicle in this lane through "successive image differences" considering pixels within the red square of Fig. 6.1b

Due to the quantity of data: one high definition camera, two webcams (only one is used here), one radar and twelve audio tracks (only three are used here) and also due to the battery limitation of the devices (video and audio acquisition ones), the exploitable recording duration was only 240 seconds. We acknowledge this is relatively short since only 24 vehicles were detected during this time. On the other hand, this allows us about each passage in more detail. The brand and model of each vehicle was identified manually using the movies so that their actual wheelbase length is also known in addition to their speed and time of apparition. All the vehicles' pass-by are depicted in Appendix A.6.

**St-Maurice database**

This database was collected on November 2010, $11^{th}$, at Route du Simplon, St Maurice, Switzerland (Lat. 46° 12'39.01", Long. 7° 0'21.93"). The road was rectilinear and composed of two opposite lanes, and located in a quite calm residential area. The triangle microphone array was set up at 8.8 m from the left-hand traffic lane lane and 5.1 m from the right-hand traffic lane at a height of about 80 cm, the inter-sensor distance was 25 cm. The sensors were the same as in the EPFL database. In this campaign, 139 vehicles were recorded in 14 minutes, 72 on the right-hand side, 67 on the left-hand side, Fig. 6.2a.

The scene was continuously filmed by a webcam placed near the array. Microphones were connected to the *NI USB-9233* acquisition card from *National Instruments*. The three channels were simultaneously cadenced at a sampling rate of 50 kHz. A Matlab interface was developed to start the recordings and to monitor in real time the smooth running of operations by displaying spectra, sound pressure levels and waveforms of the channels, Fig. 6.2b.

(a) (b)

Figure 6.2: Experimental set-up of the "St-Maurice-Database". (a): microphone array and (b): acquisition device and Matlab interface

## 6.2 Discrete formulation of signals

A recording of audio signals may be seen as a matrix of size N× M where $M$ is the number of microphones and $N$ is the number of samples. $N$ is related to the recording duration $T$ (sec) by:

$$N = \left\lfloor \frac{T}{f_s} \right\rfloor, \tag{6.1}$$

where $\lfloor . \rfloor$ denotes the floor function and $f_s$ denotes the sampling rate (in Hz). In practice, a recording is processed on successive short audio frames of length $N_s$ samples, each overlapping the previous by $N_o$ samples. A recording of $N$ samples therefore produces $N_w$ observations (frames) with:

$$N_w = \left\lfloor \frac{N - N_s}{N_s - N_o} \right\rfloor + 1. \tag{6.2}$$

This frame-by-frame processing methodology is exemplified in Fig. 6.3.

Let $\mathbf{y}_j^q$ be the $q^{th}$ audio frame of size $N_s \times 1$ of the $j^{th}$ channel. One can write:

$$\mathbf{y}_j^q = [y_j[m], y_j[m-1], ..., y_j[m-N_s+1], \tag{6.3}$$

with

$$m = (q-1)(N_s - N_o) + N_s. \tag{6.4}$$

Let $\mathbf{Y}_j^q$ be the discrete Fourier transform (DFT) of $\mathbf{y}_j^q$ of size $N_s \times 1$. The definition of

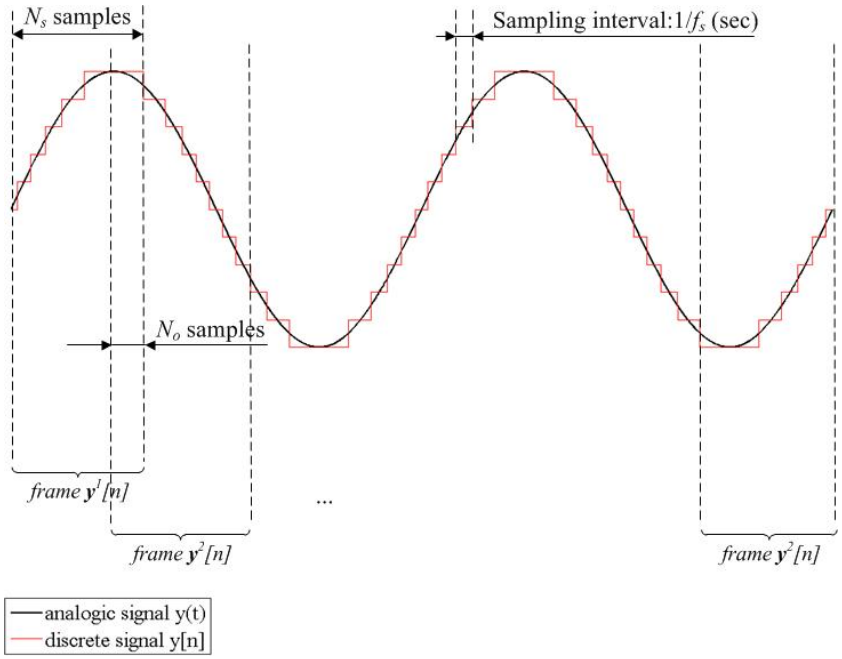



Figure 6.3: The frame-by-frame digital audio signal processing methodology.

$\mathbf{Y}_j^q$ is:

$$\mathbf{Y}_j^q[k] = DFT(\mathbf{y}_j^q) = \sum_{n=1}^{N_s} \mathbf{y}_j^q[n] e^{-2i\pi k \frac{n-1}{N_s}}, \tag{6.5}$$

$$\mathbf{y}_j^q[n] = IDFT(\mathbf{Y}_j^q) = \frac{1}{N_s} \sum_{k=1}^{N_s} \mathbf{Y}_j^q[k] e^{2i\pi k \frac{n-1}{N_s}}. \tag{6.6}$$

The discrete counterpart of the continuous GCC-BPHAT function computed using the $q^{th}$ pair of frames of sensors 1 and 2 is the vector $\mathbf{R}_{bphat}^q$ of size $N_s \times 1$ can be computed using:

$$\mathbf{R}_{bphat}^q = \begin{cases} \mathbf{Re}\left\{ IDFT\left( \frac{\mathbf{Y}_1^q[k]\mathbf{Y}_2^q[k]^*}{|\mathbf{Y}_1^q[k]\mathbf{Y}_2^q[k]^*|} \right) \right\} & \text{if } k \text{ is in the BPHAT bandwidth,} \\ 0 & \text{otherwise.} \end{cases} \tag{6.7}$$

Finally, a CCTS image $\mathbf{R}_{phat}$ consists in the concatenation of the $N_w$ discrete correlation measurements $\mathbf{R}_{phat}^q$, $q \in [|1, N_w|]$, such that:

$$\mathbf{R}_{phat} = [R_{phat}^1, R_{phat}^2, ..., R_{phat}^{N_w}]. \tag{6.8}$$

In practice, $N_s$ needs to be sufficiently large to get reliable measurements. Indeed, when using the cross-correlation function, the longer the signals, the smaller the variance of the time-delay estimates is [51]. On the other hand, $N_s$ need to be sufficiently small for

the static source assumption to hold. We empirically noticed that a window duration between 30 ms and 40 ms is a good trade-off. A vehicle traveling at 90 km/h moves by less than one meter during this time. In the following, recordings are processed using $N_s$ = 2048 samples and $N_o = 0.75N_s$.

## 6.3 Speed estimation

In this part, performance of the proposed approach with regard to speed estimation are investigated. The term *performance* refers here to the precision (related to the error between actual and estimated states) and the accuracy (related to the repeatability) of the method. Such an assessment is carried out on the EPFL-Database. Different strategies are compared.

### 6.3.1 Tracking strategies

Although three microphones have been used for the experiment, laid out on an equilateral triangle, two microphones placed in parallel to the road lane are theoretically sufficient to localize a vehicle without ambiguity, as explained in chapter 5. Similarly, a unimodal particle filter, as described in setion 3.6, is theoretically sufficient for the estimation of speed since wheelbase does not play any role on it. Therefore, four different strategies are assessed:

- observation with 2 microphones + tracking with unimodal particle filter: 2MUPF;
- observation with 2 microphones + tracking with bimodal particle filter: 2MBPF;
- observation with 3 microphones + tracking with unimodal particle filter: 3MUPF;
- observation with 3 microphones + tracking with bimodal particle filter: 3MBPF.

Each strategy is run 200 times for each of the 24 pass-by. At the end of each run, the mean and standard deviation of the $N_p$ particles for the speed state are computed. Then, at the end of the 200 runs, the global error and relative standard deviation are returned using Eq. (A.22) and Eq. (A.24).

As an example, the superposition of 200 runs launched on the $20^{th}$ pass-by is depicted in Fig. 6.4 using the 2MUPF strategy, Fig. 6.4a, and the 3MBPF strategy, Fig. 6.4b. On these plots, each point of each red line represents the mean value of the particle coordinates (in x and y) transduced in terms of time-delay. Regarding the observation of these examples (CCTS in black and white), it clearly appears that a much less noisy result is achieved using three microphones, Fig. 6.4b, than using only two microphones, Fig. 6.4a. Also, regarding the particles trajectories (in red), the unimodal model, Fig. 6.4a, make the particles switch abruptly from one axle to another at the beginning of the observation (nearly 0.15 sec), while this is not the case using the bimodal model, Fig. 6.4b. Pictures of vehicle 20 are depicted in Appendix A.6. The radar indicates that its speed was 79 km/h. The estimate returned by the 2MUPF strategy is 78 km/h, and the estimate returned by the 3MUPF strategy is 79 km/h. The underestimation

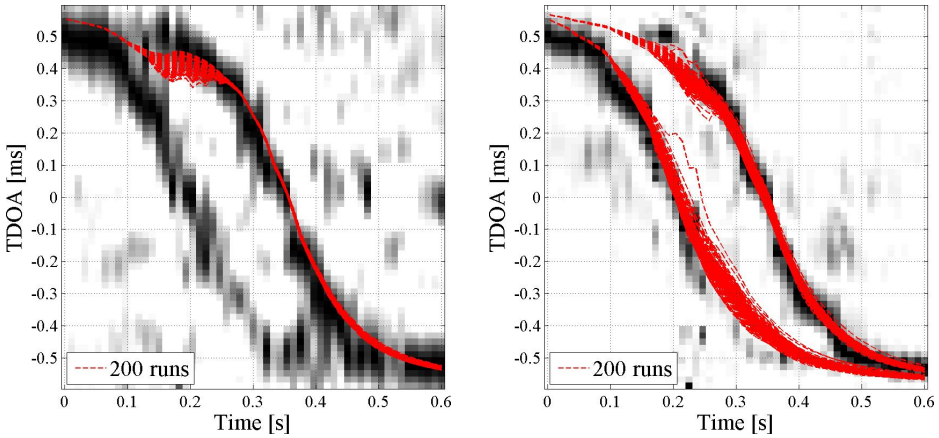



Figure 6.4: Superposition of observations and 200 particles trajectories launched with same initial conditions (examples). On (a) the observation is provided by two microphones only and the tracking is performed by the unimodal particle filter (2MUPF strategy), on (b), the observation is provided by the three microphones and the tracking is performed by the bimodal particle filter (3MBPF strategy).

| Abs. diff. radar / acoustic | 2MUPF | 3MUPF | 2MBPF | 3MBPF |
|---|---|---|---|---|
| 0 km/h - 3 km/h | 9 | 8 | 11 | 12 |
| 3 km/h - 5 km/h | 1 | 5 | 3 | 6 |
| 5 km/h - 10 km/h | 7 | 7 | 7 | 4 |
| $\geq$ 10 km/h | 7 | 4 | 3 | 2 |
| total | 24 | 24 | 24 | 24 |

Table 6.1: Number of vehicles (over 24) belonging to different margin of errors.

of the 2MUPF strategy is certainly due to this change in the tracked trace, even if the difference is small for both approaches in this example.

### 6.3.2 Results

Results on all pass-by are depicted in Fig. 6.5 and Fig. 6.6. For each strategy, the acoustic speed estimates (red crosses) and their CI95 (vertical red lines) are confronted to the radar Doppler estimates (black crosses) as a function of the vehicle ID. For clarity, actual speeds have been sorted in ascending order. The absolute difference between acoustic and radar estimates are depicted by a bar graph and compared to various thresholds: $\pm$ 3 km/h, $\pm$ 5 km/h and $\pm$ 10 km/h. The number of vehicles belonging to each of these error intervals is given in Table 6.1.

The poorest results are achieved using the 2MUPF strategy. More than half of the estimates (14 over 24) have an error greater than 5 km/h. This number is reduced to 11 when applying the BPF instead of the UPF, and to 10 when using three microphones instead of two.

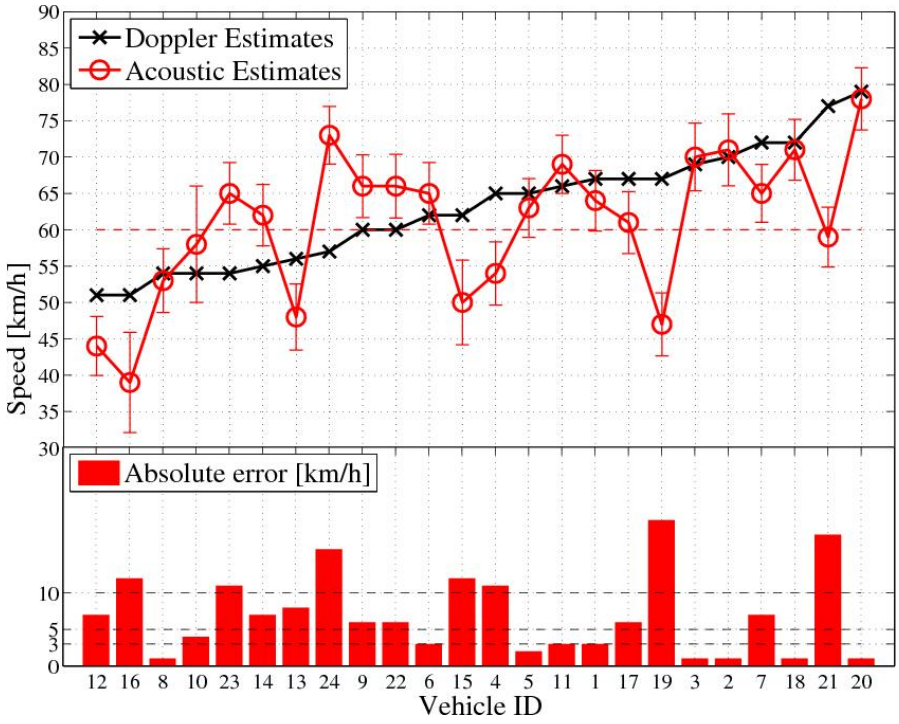
When two microphones are used, applying the BPF permits to estimate the speed of the vehicle 19 with only 2 km/h of error while it was of 24 km/h with the UPF as depicted in Fig. 6.5a and Fig. 6.5b. By looking more closely at the CCTS of vehicle 19, one can observe that the two traces are well visible but spurious peaks are present between them. Thus, particles easily lose the track of the front axle to suddenly track the rear axle, Fig. 6.7a. This results in a large underestimation of the actual speed. The bimodal particle filter is much less disturbed by these spurious peaks as particles are forced to converge towards a solution that takes into account a constant wheelbase length of nearly 2 or 3 meters.

A totally opposite effect may be noticed for vehicle 8: the speed estimation is much worse by using the 2MBPF strategy than using the 2MUPF strategy. In the first case, an error of 6 km/h is obtained, and of 1 km/h only in the second case. This is mainly due to a mismatch between model and observation. The acoustic energy radiated by this vehicle mainly comes from the front of the vehicle so that the rear axle is almost invisible on the trace. The UPF correctly tracks the trace, Fig. 6.7c, but the BPF has difficulty to converge because it constantly tries to stabilize itself by tracking a second axle which is badly observed, Fig. 6.7c.

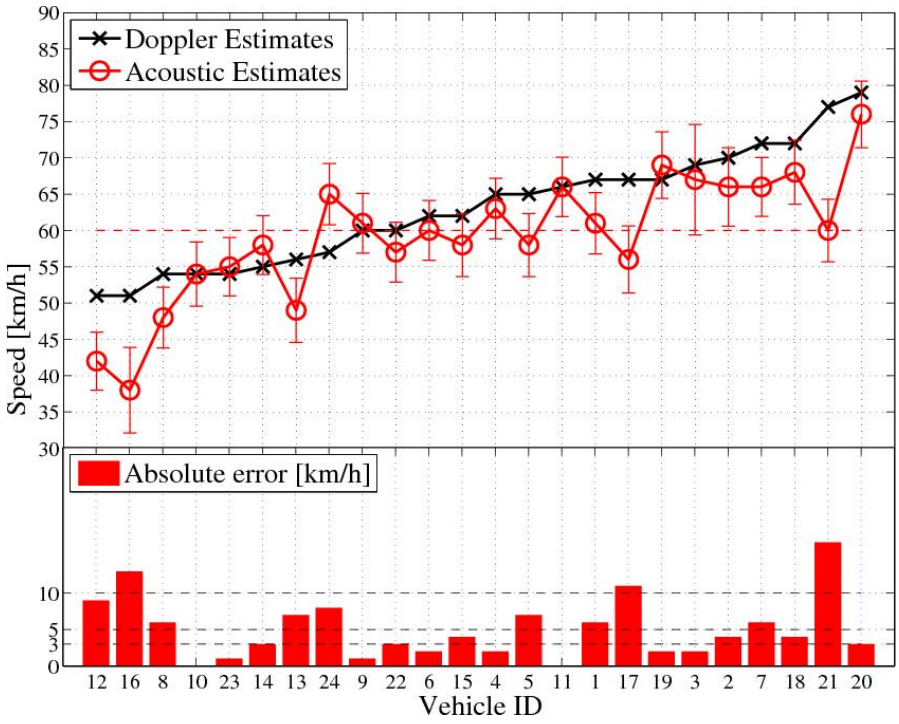
Again, vehicle 19 is the best example of the improvement brought by a third microphone even when the unimodal particle filter is used. The use of a third microphone permit to attenuate spurious peaks that are not consistent with a sound source coming from the road. As a consequence, a better contrast is obtained as depicted in Fig. 6.7e and Fig. 6.7f. For this example, the initial error of 20 km/h obtained with the 2MUPF strategy is reduced to 8 km/h with the 3MUPF strategy.

In some rare cases, it may happen that the MULTI-PHAT technique is so selective that the information of interest is partially missing. This is for example the case with vehicle 14. The trace of the front axle is well visible from the beginning to the end of the observation with two microphones, Fig. 6.7g, but not anymore in Fig. 6.7h where three microphones are used. Consequently, in the 3MUPF strategy, particles are temporarily lost and continue their trajectory based on the latest available observations before locking again as soon as the trace is visible. Thus, a less accurate estimate is obtained. In this case, an error of 10 km/h was observed when using three microphones, and of 7 km/h when using two microphones.

But, globally speaking, Table 6.1 shows that changing from the 2MUPF strategy to the 2MBPF or 3MBPF strategy brings an improvement especially regarding the $\pm$ 5 km/h error margin. Even better results are obtained with the 3MBPF strategy for which less than 10 km/h error is achieved in 92% of the cases.

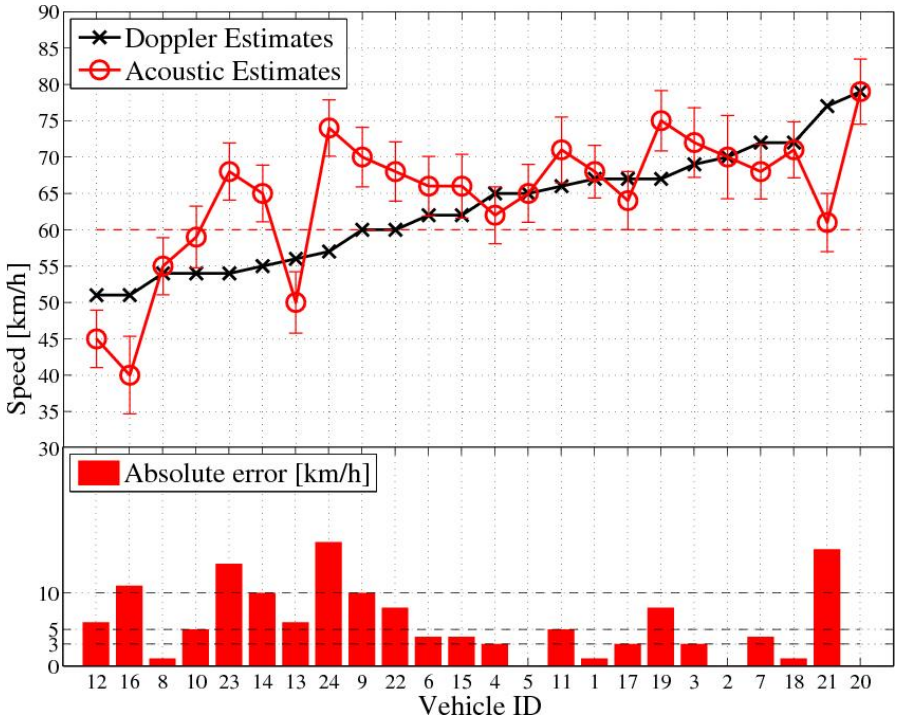



(a) 2MUPF

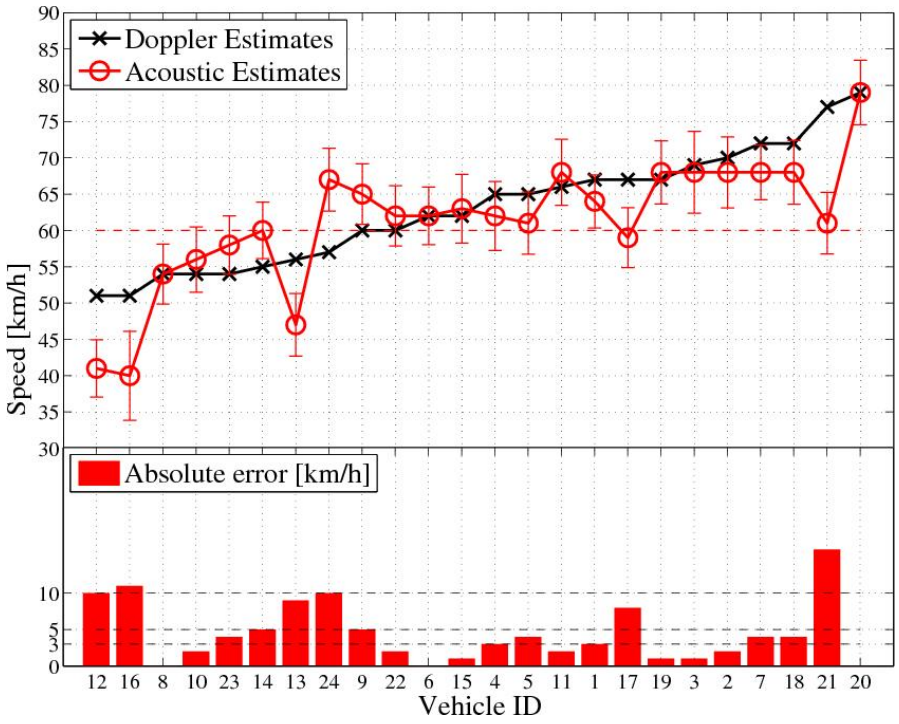



(b) 2MBPF

Figure 6.5: Comparison between Doppler and acoustic speed estimates as a function of the vehicle ID for the UPF-based strategies using (a) two and (b) three microphones. For clarity, actual speeds have been sorted in ascending order.

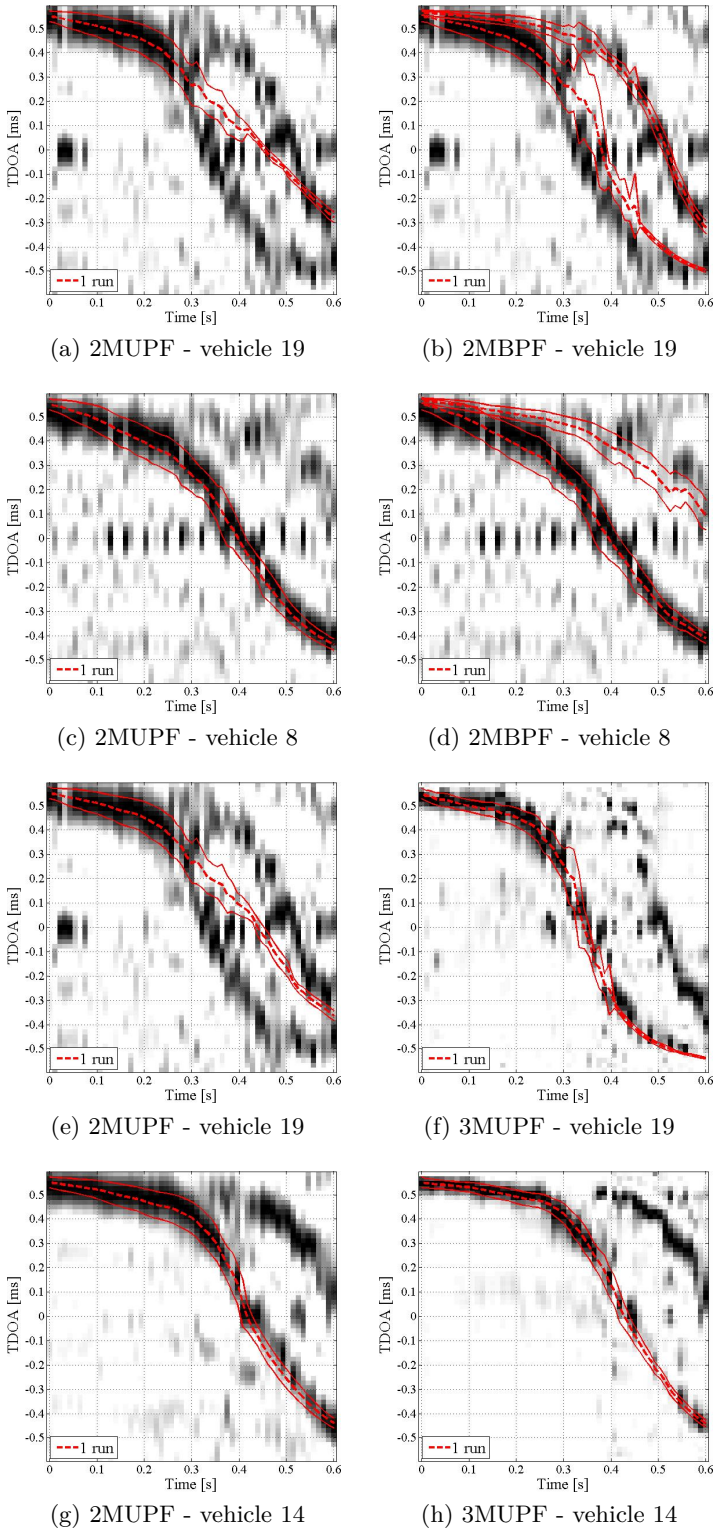



(a) 3MUPF




(b) 3MBPF

Figure 6.6: Comparison between Doppler and acoustic speed estimates as a function of the vehicle ID for the BPF-based strategies using (a) two and (b) three microphones. For clarity, actual speeds have been sorted in ascending order.




(a) 2MUPF - vehicle 19

(b) 2MBPF - vehicle 19

(c) 2MUPF - vehicle 8

(d) 2MBPF - vehicle 8

(e) 2MUPF - vehicle 19

(f) 3MUPF - vehicle 19

(g) 2MUPF - vehicle 14

(h) 3MUPF - vehicle 14

Figure 6.7: Comparison between observations and particles trajectories after one run (problematical cases).

### 6.3.3 Problematic (but still interesting) cases

This section focuses on cases in which the error is larger than 5 km/h, despite the use of the 3MBPF strategy.

These cases corresponds to vehicles 16 and 21 for which an error greater than 10 km/h was observed, and vehicles 12, 13, 17 and 24 for which an error between 5 km/h and 10 km/h was observed. Understanding why such errors have occurred may be instructive for the practitioner.

**Vehicles 16 and 21**

The largest differences between 3MBPF estimates and radar ones appear for vehicle 16 and 21, with values of 11 km/h and 16 km/h respectively. In reality, these two pass-by are due to the same motorbike which passed two times during the measurement, Fig. 6.8a.

The trace of this motorbike during its first passage is depicted in Fig. 6.8b. This clearly demonstrates that the hypothesis of bimodality does not match at all with this observation. Indeed, although the motorbike has two wheels, the rolling noise is totally masked by the engine noise which is preponderant for such vehicles. The front and rear wheels are not observed and the $BPF$-based methods fails.

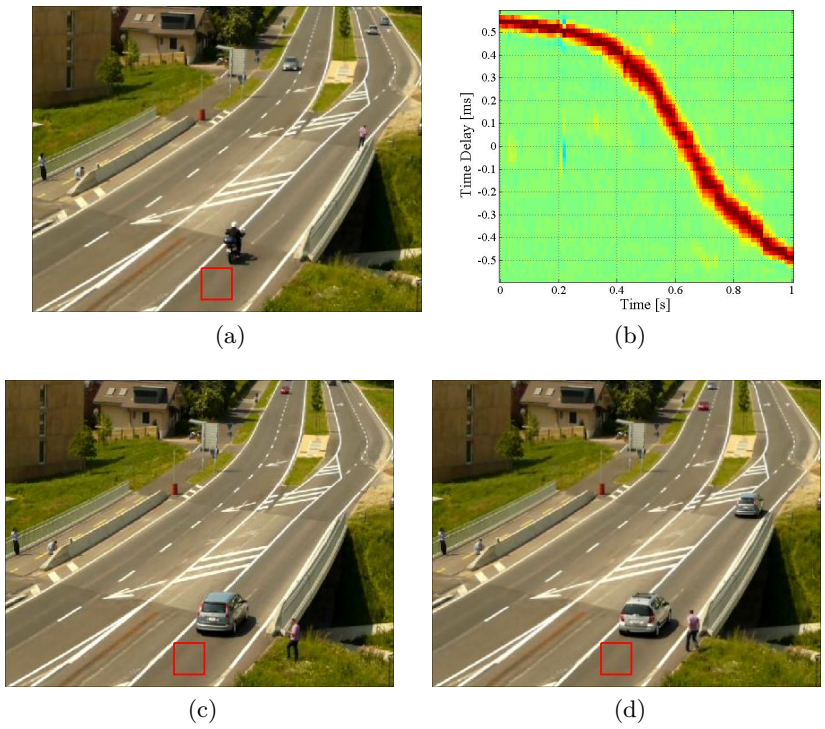
However, one can note from Fig. 6.5a and Fig. 6.6a that $UPF$-based strategies do not provide a better result. The error is always greater than 10 km/h whatever the number of sensors thus the problem lies elsewhere.

Fig. 6.8a, 6.8c and 6.8d depict the video frame corresponding to the time index returned by the detection algorithm, that is, the instant when the vehicle is supposed to be located at coordinate $[\mu_{x,0},\mu_{y,0}]$ (*a priori* initial position for the particles). By looking attentively at these pictures, one can note that the initial position of the moto is quite different from those of the cars, in terms of abscissa and ordinate. Cars are closer to the array than the moto at initialisation. Note that this difference is amplified on the $x$ axis since the sound radiated by the motorbike mainly comes from the exhaust system which is situated at the rear of the vehicle. This default in the detection comes from the opted strategy chosen, which consisted in waiting that the vehicle was completely out of the red square to launch the particles. Consequently, when the vehicle is smaller, respectively longer, than the majority of the cars, its actual position is further, respectively closer, to the *a priori* one. In the case of pass-by 16, respectively 21, this implies a significant underestimation of the actual speed of 11 km/h, respectively 23 km/h. Adjusting the initial *a-priori* coordinate of the moto to more realistic values reduced the error to 2 km/h for the pass-by 16 and 3 km/h for pass-by 21.

**Vehicle 12, 13, 17, 24**

The error for each of these vehicles are between 5 km/h and 10 km/h.

Like the motorbike 16 (or 21), the car 13 and the van 12 drive on the left of the road




(a) (b)

(c) (d)

Figure 6.8: (b) Trace of the $16^{th}$ pass-by (motorbike), (a), (c) and (d): initial positions of vehicles 16, 14 and 15 respectively.

when compared to other vehicles as depicted in Appendix in Fig. A.9g and Fig. A.9d. Adjusting $\mu_{y,0}$ to a more realistic value reduces the error of 9 km/h for vehicle 13 and of 10 km/h for vehicle 12 to 0 km/h and 5 km/h respectively. On the contrary, the vehicle 24 is moving on the right of the road, Fig. A.11j. Again, adjusting $\mu_{y,0}$ to the actual ordinate reduces the error from 10 km/h to 2 km/h. Finally, vehicle 17 is a van whose rear axle is difficult to observe by a correlation measure as depicted in Fig. A.10f. Consequently, the 3MUPF model provides a better estimate (3 km/h error) than the 3MBPF strategy (8 km/h error).

**Remark** These experiments highlight well one of the major challenges of target tracking, the so-called *measurement-origin uncertainty*, discussed in chapter 4 and also evoked in the literature, for instance in [139, 140, 141]. One way to handle this issue is to implement a reliable detection technique providing the exact initial position of cars.

### 6.3.4 Benefits of the bimodality in harsh conditions

Regarding the accuracy of the speed estimates, unimodal and bimodal models returned the same results in case of ideal observational conditions. However, bimodal observation model is preferable in harsh situations. This is the conclusion that can be reached from the exploitation of the EPFL Database and such a trend is confirmed by one recording

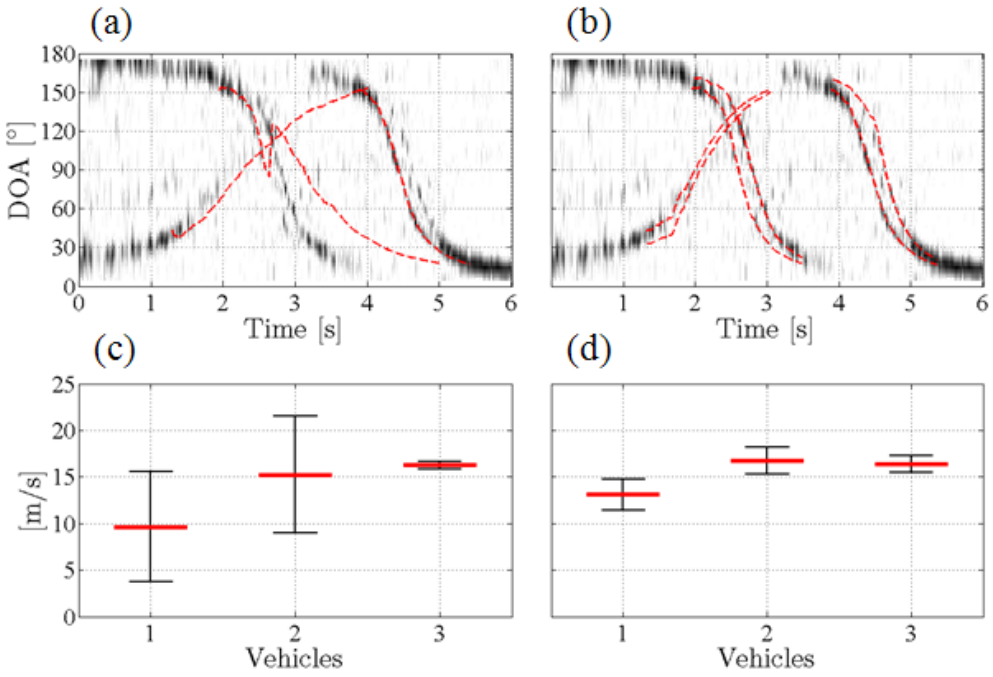



Figure 6.9: DOA as a function of time, and speed estimates of three vehicles in a real harsh situation. Raw observation and results are superimposed using (a) : a unimodal sound source PF-based model, (b) a bimodal sound source PF-based model. The same measurement is processed 200 times, mean and IC95 in speed estimates are represented in (c) and (d) for each vehicle.

coming from the St-Maurice Database, depicted in Fig. 6.9.

In this 6-second recording, two vehicles pass each other quickly followed by a third one. The observation is compared to the result returned by one run of $UPF$ in Fig. 6.9a and $BPF$ in Fig. 6.9b. This clearly demonstrates that the $UPF$ method makes the particles follow the most dominant of the two axles, and needs to overcome a large gap when the dominant axle is changing, which typically happens when the vehicle is in the broadside situation. Risks of failure during this gap are accentuated when another vehicle is tracked at the same time as is the case here. This risk is drastically reduced using the bimodal observation model where no gap is noticed anymore and a wheelbase estimate is provided.

Both methods have been applied 200 times to this measurement. Results on speed estimation for each case (mean and CI95) are depicted in Fig. 6.9c and Fig. 6.9d. The actual speeds are unknown but looking at the confidence interval, one can note that the CI95 of the estimates for vehicle 1 and 2 cover a very large zone of around 11 m/s. In contrast, these intervals are drastically reduced when using the $BPF$ method. Regarding the third vehicle, one can note that both approaches lead on the same speed estimation but the CI95 is lower for the standard technique compared to the bimodal one. Particles following a unimodal model track the front axle sound source, which happens to be the dominant and less noisy one, whereas particles following the bimodal model are also

driven by the rear axle which is observed with lower quality. Thus, the convergence is made more difficult for the particles, due to the wheelbase state having a larger variance.

## 6.4 Wheelbase length estimation

Experiments on wheelbase length estimation were carried out using the EPFL Database. The two $BPF$-based strategies (two or three microphones) are tested and depicted in Fig. 6.10. The acoustic estimates (in red) are compared to the actual ones (in black) and their absolute differences are represented by a bar chart. The *a priori* wheelbase length $\mu_{wb,0}$ here equals 2.25 m (arbitrary choice) and is represented by a red dashed line. For clarity, actual wheelbase lengths have been sorted in ascending order.

First of all, one can observe that, despite an *a priori* wheelbase length relatively distant from reality, the trend in estimates is pretty good for wheelbase lengths varying between 2.4 m and 2.8 m. When the wheelbase is difficultly observed, the final result tends to be close to the *a priori* value $\mu_{wb,0}$. This is obviously the case for motorbikes 16 and 21, but also for vehicles 2, 8, 12, 13, 17 and 19, as it is true that the observation are very noisy (see Appendix A.6). The situation is even worse for cars 2, 12 and 17 as their wheelbase is quite distant from the *a priori* value.

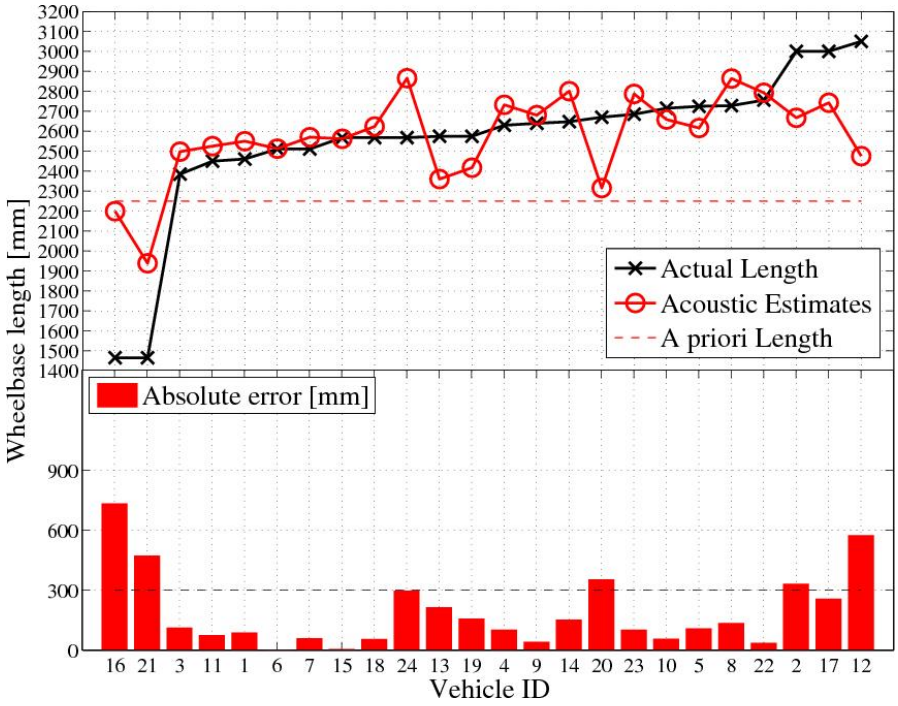
Anyway, if both pass-by of motorbikes 16 and 21 are excluded from the database, as being out of context, it appears that for 18 out of 22 cases , respectively 17 out of 22 cases, the error is less than 30 cm when using two microphones, respectively three microphones. Such an error is typically less than the diameter of a wheel.

**Remark** It appears that for wheelbase length estimation, the use of three microphones is not as relevant as for speed estimation since results seem better when using two microphones here. We suppose this is due to the MULTI-PHAT which is too selective and which deteriorates the sides of peaks. This phenomenon should be investigated in more detail using controlled moving sources and simulations. A strategy to investigate could be to rely on MULTI-PHAT for speed estimation, and on SRP-PHAT (non-destructive) for wheelbase length estimation.

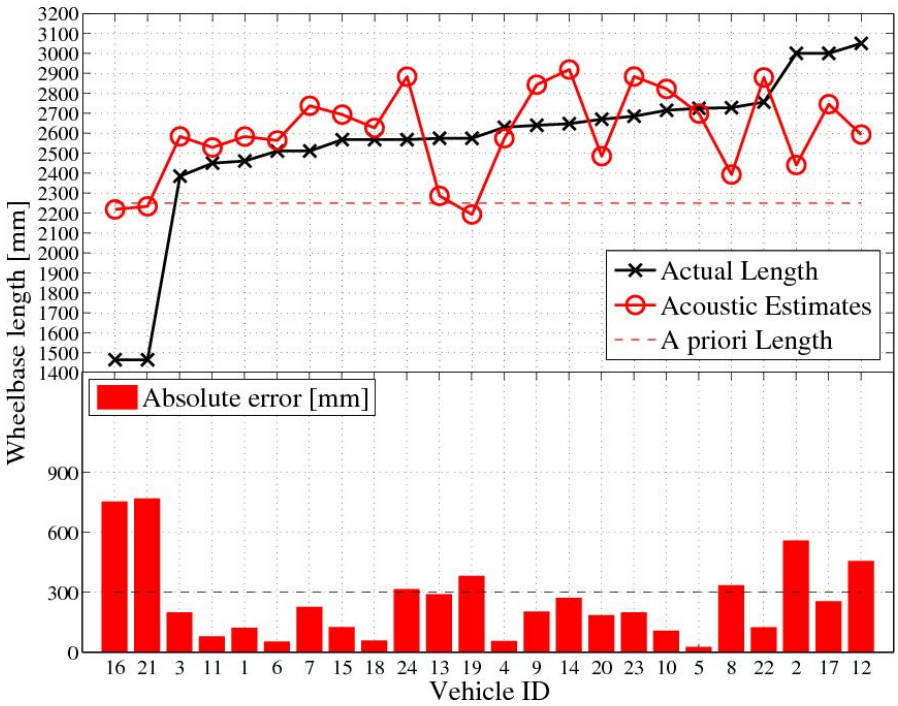
## 6.5 Detection

### 6.5.1 Broadside detection

The audio recordings of the *EPFL database* have been partitioned into short frames of length $N_s = 2048$ samples without overlap. A single microphone has been considered for all the experiments. Each frame which did not correspond to the presence of a vehicle in the detection zone (at least in part) was labeled as 0, otherwise as 1. This procedure of categorization was automatically performed after having developed an ad-hoc video-based algorithm using signals of the in-height camera. This resulted in a database composed of $N_w^0 = 12643$ frames of type 0 and $N_w^1 = 1633$ frames of type 1. Audio features




(a) 2MBPF




(b) 3MBPF

Figure 6.10: Confrontation between actual and acoustic wheelbase estimates as a function of the vehicle ID when using two (a) and three (b) microphones. For clarity, actual wheelbase lengths have been sorted in ascending order.

were extracted from each frame. Features that have been considered are classic ones in automatic music classification and are described in Appendix A.5.

To illustrate the methodology, let us consider an example: the zero crossing rate (ZCR). ZCR is a measure of how many times a signal crosses the zero axis. Its definition is:

$$ZCR[q] = \frac{1}{N_s - 1} \sum_{n=1}^{N_s-1} |sign(\mathbf{y}_1^q[N_s - n + 1]) - sign(\mathbf{y}_1^q[N_s - n])|, \tag{6.9}$$

where

$$sign(\mathbf{y}_1^q[n]) = \begin{cases} 1 & \text{if } \mathbf{y}_1^q[n] \geq 0, \\ -1 & \text{if } \mathbf{y}_1^q[n] < 0, \end{cases}$$

where $q$ is the frame number. The ZCR is traditionally used to distinguish clean or periodic signals (low ZCR) from more noisy ones (high ZCR).

We investigated the performance of ZCR, *i.e.* true positive rate (TPR) and false positive rate (FPR), both defined in section 3.7.1, as a function of the spectral band of the frame considered. That is, without pre-filtering, corresponding to the raw definition of ZCR, *e.g.* Eq. (6.9), and with a pre-filtering according to the standard octave band decomposition, from the 63-Hz band to the 16-kHz band (9 bands).

Fig. 6.11a depicts each of the 10 ZCR obtained as a function of time for one audio recording in which a pass-by occurs. The raw ZCR is symbolized by a thick black line, the other ones by fine-colored lines. At first glance, the raw ZCR is not a discriminant feature because no noticeable difference appears between frames of class 1 (inside the red dashed box) and frames of class 0 (other frames). This is also true for octave-band decomposition up to 8 kHz. However, the 16 kHz octave band brings a more important contrast. This observation is confirmed in Fig. 6.11b, where each ZCR is normalized between 0 and 1. The 16 kHz octave band is the one for which the contrast between detection zone and other zones is the most important. Considering the whole EPFL database, one can plot the distribution of raw and 16 kHz ZCR for both classes. As expected, no distinction between classes is possible when ZCR is applied to the raw signal: Fig. 6.11c, but this changes when it is applied to a pre-filtered signal at 16 kHz octave band 6.11d. A ROC analysis has been derived to automatically find the best confusion matrix according to the range of the ZCR. The optimal threshold is depicted by a dashed black line in both cases. It clearly appears on this example that the detection performance is greatly improved.

The performance of explored features are summarized in Table 6.2. For almost all the features, their discriminative ability is drastically improved when they are applied to the appropriate frequency band. Note that when the optimal band is not specified, it means that the performance of the corresponding feature is better when the raw signal is considered. One of the most spectacular cases of improvement is the ZCR. Without optimization, this feature classifies each frame well nearly half of the time, but when it is

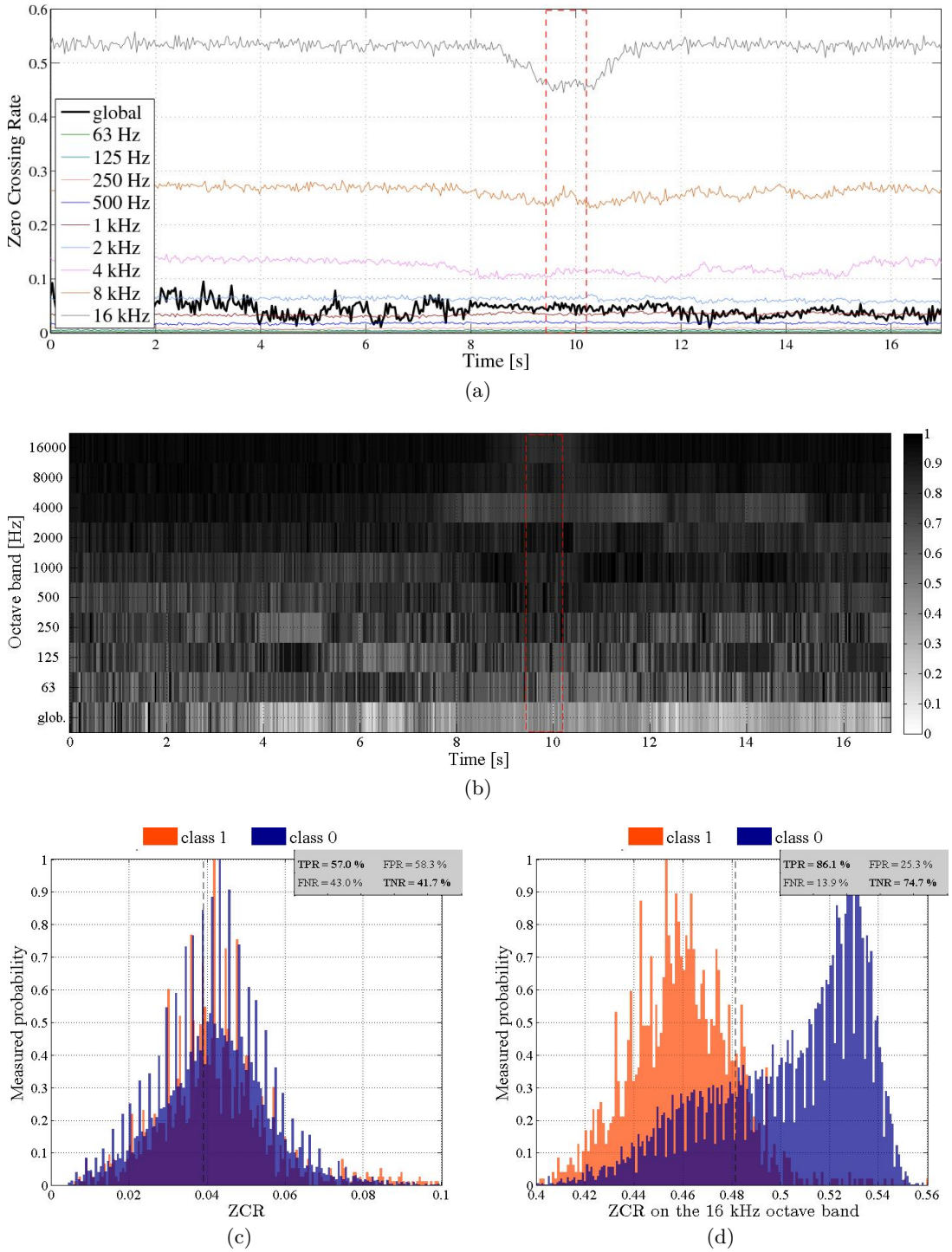



Figure 6.11: Example of a feature that has been optimized. (a): raw zero crossing rate (ZCR) (black line) and per octave band, (b): normalized raw ZCR (first line from the bottom) and per octava band, the best contrast is achieved for the 16 kHz band (first line from top), (c): histogram and confusion matrix for the raw ZCR between the two classes, (d): histogram and confusion matrix for the 16 kHz ZCR between the two classes.

| | on the whole signal | | on the optimal band | | |
|---|---|---|---|---|---|
| Feature | TPR | FPR | TPR | FPR | Optimal octave band |
| MAC | 79.9% | 17.1% | 84.9% | 16.4% | 500 Hz |
| KRT | 59.2% | 49.1% | 87.7% | 21.4% | 16 kHz |
| SGC | 56% | 72.9% | 87.6% | 21.4% | 16 kHz |
| SRF ($\gamma_{srf}$=0.85) | 65.5% | 50.9% | 85.4% | 19.7% | 16 kHz |
| SPL | 79.9% | 17.1% | 82.4% | 17.5% | 8 kHz |
| ZCR | 57% | 58.3% | 86.1% | 25.3% | 16 kHz |
| SKW | 55.5% | 46% | 76% | 30.4% | 16 kHz |
| SBW | 67% | 37.1% | | | |
| SRF ($\gamma_{srf}$=0.99) | 54% | 29.5% | | | |
| SRF ($\gamma_{srf}$=0.95) | 70.5% | 37.5% | | | |

Table 6.2: Performance of raw and optimized features for broadside detection.

applied to the 16 kHz octave band, the prediction is correct more than 8 times out of 10. Among the most performant features are the spectral gravity center (SGC), spectral roll-off point (SRF), maximum of the auto-correlation (MAC) and ZCR. Actually, these descriptors reflect almost the same information, namely the enrichment of the spectral content in the high frequencies when a vehicle passes just in front of the network. There is no doubt that this physical property should be a solid basis of a broadside detection algorithm.

One needs to keep in mind that such an analysis of each feature taken independently is only the first step of a classifier design procedure. Each feature vector should be normalized and correlation between features should be studied in order to select the most representative ones. Such a selection is traditionally done using statistical techniques like a principal component analysis (PCA) aiming at maximizing the variance between features. Recently, Rabaoui *et. al* proposed not to select the best features, but the best combinaison of features using the support vector machine approach [142].

**Remark** The octave band decomposition was inspired here by traditional acoustic measurement of room acoustics or music classification. Depending on the application, one may choose another decomposition like Mel bands [109], commonly used by the automatic speech recognition community, Bark bands, convenient for modeling the human auditory system [143] or even non-standard decompositions like in [144]. During the supervision of two master thesis, one about applause sound detection and the other about owl cries detection, one optimal band per feature was determined by looking for which part of the spectrum (parameterized by $f_{min}$ and $f_{max}$) maximized the Kullback-Leibler divergence between distributions of classes 0 and 1 based on a training database.

### 6.5.2 Endfire detection

This section provides experimental results on the endfire detection strategy described in section 3.7.2. The audio recordings of the *St-Maurice database* were considered. We

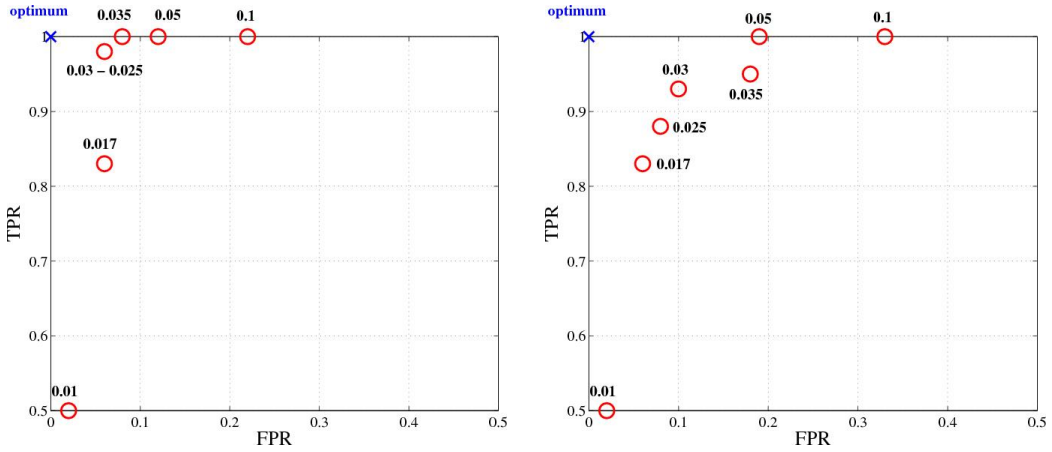



Figure 6.12: ROC curve for the threshold $\Lambda$ through real measurements. (a): detection of vehicles coming from left, (b): detection of vehicles coming from right.

remind that for this campaign, 139 vehicles had been recorded, 72 came from left, 67 came from right. The measured cross-correlation time series $CCTS_{meas}$ is the concatenation of the $K$ last cross-correlation measurements, where $K$ is defined by: the length of the detection zone $L$ (in m), the expected speed of the vehicle $\mu_{\dot{x},0}$ (m/s), the length of the window analysis $N_s$ and of the overlap $N_o$ (samples), the sampling frequency $f_s$ such that:

$$K = \left\lfloor \frac{\lfloor N/(\mu_{\dot{x},0} f_s) \rfloor - N_s}{N_s - N_o} \right\rfloor + 1 \tag{6.10}$$

Again, we decided to take advantage of the information redundancy brought by the three pairs in the array and we relied on three different measured and theoretic $CCTS$, one couple theory-measurement per pair. Consequently, three 2D Pearson coefficients were returned $r^{(1)}$, $r^{(2)}$ and $r^{(3)}$. A final classifier can be:

$$\mathcal{D}[q] = \prod_{p=1}^{3} r^{(p)}, \tag{6.11}$$

where $q$ is the current audio frame. The score of the classifier (6.11) has been compared with several thresholds $\Lambda$ and for each one, it has been decided manually, by replaying the images of the webcam, if the detection was correct or not. The processing was performed on frames of 1024 samples with 75% overlap. The corresponding ROC curves are plotted in Fig. 6.12.

It appears that the method works better for vehicle detection on the nearest lane (left to right direction of circulation) than on the farthest lane (right to left). Indeed, it clearly appears that red circles are closer to the optimum point in Fig. 6.12a than in Fig. 6.12b. According to our observations, this is mainly due to the fact than vehicles coming

from right may be masked by those leaving on the nearest lane, making impossible the detection of the most distant vehicle. The masking effect is a difficult (if not impossible) problem to solve using a single compact microphone array. No attempt toward this direction has been developed during this work but a possible solution could be to use several arrays so that an array can detect vehicles that have been missed by another array. According to Fig. 6.12, the optimal threshold $\Lambda$ is 0.03 for both flow directions. Applying this value to the whole St-Maurice database yields a TPR of 94% and a FPR of 3% for the detection of vehicles coming from left and a TPR of 90% and FPR of 6% for the detection of vehicles coming from right.

## 6.6 Conclusion

In this chapter we discussed two *in-situ* test campaigns, one mainly assessing the proposed tracking method and the broadside detection strategy (EPFL Database), the other mainly assessing the endfire detection strategy (St-Maurice Database).

For tracking, four strategies were proposed, from the more basic (two microphones and unimodal particle filter) to the more evolved (three microphones and bimodal particle filter). Methods have been assessed on the basis of 24 pass-by recorded with audio and video devices. The study confirmed the usefulness of a third microphone compared to the two theoretically sufficient ones. Indeed, switching from two to three sensors expands the percentage of vehicles for which the error in speed estimates is below 10 km/h from 70% to 83% when using the unimodal particle filter and from 87.5% to 92% when using the bimodal particle filter. The proposed strategy consists in applying the bimodal particle filter on observations provided by three sensors using the MULTI-PHAT technique (3MBPF strategy). This is what provides the best results in speed estimation with 75% of vehicles having an error below 5 km/h. Promising results have also been achieved regarding the wheelbase length estimation problem since 91% of the two-axle vehicles monitored returned an error below 30 cm using the bimodal particle filter, with an observation provided by two microphones. These are excellent first results since we reach a spatial accuracy comparable to the radius of a wheel.

The detection problem has been assessed separately. Two strategies have been considered: the broadside detection strategy (for detecting vehicles in front of the array) and the endfire detection strategy (for detecting vehicles upstream the array). For the first one, different audio features classically found in the music classification literature have been tested. Our work consisted in optimizing each of them by searching which octave band dissociate at best the two situations: vehicle in front of the array, no vehicle in front of the array. Kurtosis or spectral gravity center, once having been optimized, present a good potential of detection, having a true positive rate above 87% and a false negative rate below 20%. The endfire detection have been investigated in more detail. A new method has been proposed, consisting in establishing a score between two theoretical and observed cross-correlation time series of same size. Applying this procedure to real

measurements yields a true positive rate of 94% and a false positive rate of 3% for approaching vehicles on the right-hand lane, and true positive rate of 90% and false positive rate of 6% for approaching vehicles on the left-hand lane.

This study globally revealed that both precise detection, defining the initial conditions, and axles number estimation, defining the appropriate target model, are crucial to ensure good tracking performance. The former point has been evoked in this chapter and firsts results are promising. The latter point is investigated in the next chapter by relying on subspace-based theory.

# 7 Potential improvement of the method

## 7.1 Introduction

Experiments of the previous chapter revealed encouraging results for two-axle vehicles tracking using the proposed bimodal particle filter. The BPF has also proved to work better than the classical UPF even for harsh conditions. But these experiments have also revealed the unapplicability of the BPF for vehicles, like motorbikes, whose predominant exhaust noise does not allow the observation of both axles. Acoustically speaking, even having one front and one rear wheel, a motorbike can be considered as "one-axle vehicle", *i.e.* a point emitter. For this case, UPF is often better.

Consequently, a good way to improve the proposed methodology might be to automatically estimate the number of observed axles before applying the PF. It could therefore permit to choose which method to apply, such as, unimodal, bimodal or $n$-modal $PF$ depending if one, two or $n$ axles are observed. This is commonly called a *source number estimation* problem, well known from in source separation, clustering, or multiple-target tracking for instance. Support Vector Machine [145], Information Theoretic Criteria [146, 147, 148], Minimum Eigenvalue Varied Rate Criteria [149], Beam Eigenvalue Approaches [150] are existing techniques to solve this problem. The idea behind these methods consists in studying the rank of the covariance matrix of the observations. This is called the *subspace-based theory*, for instance described in [44]. This theory requires a number of sensors $M$ far larger than the number of sources $N$ ($M >> N$).

This chapter initializes a study based on the subspace-based theory, but for cases when the number of sensors is equal to the maximal possible number of sources ($M \geq N$). For now on, the only case of two pure tone sound sources is discussed, but it permits to reveal very interesting aspects of the relationship between the rank of the correlation matrix of the observations and the microphone array geometry.

## 7.2 The subspace approach

Consider an array of $M$ omnidirectional sensors with the same impulse response at locations $\mathbf{r}_j^m \in \mathbb{R}^2$, $j \in [1, 2, ..., M]$ and let $N_{max}$ be the number of maximal mutually independent and isotropic active sound sources in the medium. Each source is characterized by its location $\mathbf{r}_k^s \in \mathbb{R}^2$, its wavelength $\lambda_k$ and its amplitude $\beta_k$, $k \in [1,...,N_{max}]$. Assume that $\mathbf{r}_k^s$ and $\lambda_k$ are known for all $k$. Let $N$ be the number of active radiating sound sources during the observation, that is, sources having an amplitude different from zero, $0 \leq N \leq N_{max}$.

A simplistic but often sufficient model of sensor array signal processing considers the observations $\mathbf{X} \in \mathbb{C}^{M\times 1}$ as linear combinations of complex source signals $S \in \mathbb{C}^{N\times 1}$ attenuated and delayed in time through a complex mixing matrix $A \in \mathbb{C}^{M\times N}$ summed with an independent and identically distributed zero mean gaussian noise $W \in \mathbb{C}^{M\times 1}$:

$$\mathbf{X} = \mathbf{AS} + \mathbf{W}. \tag{7.1}$$

An interpretation of (7.1) is that each line $\mathbf{x}_j$ of $\mathbf{X}$ is a linear combination of each complex source signal $\mathbf{s}_k$ through the complex coefficient $a_{jk}$ of $\mathbf{A}$ [44]. Therefore, in presence of $N$ sources located so as to avoid type I ambiguity (*i.e.* spatial ambiguity, such as two sources placed symmetrically about the axe of a linear array), the rank of the correlation matrix $\mathbf{R}$ is equal to $N$ with $\mathbf{R}$ defined as:

$$\mathbf{R} = \mathbb{E}\left\{\mathbf{X}\mathbf{X}^H\right\}, \tag{7.2}$$

and $(.)^H$ is the hermitian operator. It appears that estimating the number of active sound sources is equivalent to estimating the rank of $\mathbf{R}$. This may be achieved by studying the eigenstructure of $\mathbf{R}$. Using the definition of the mathematical expectation the expression (7.2) may be expanded as below:

$$\mathbf{R} = \mathbf{A}\psi\mathbf{A}^H + \sigma^2 I_{N\times N}, \tag{7.3}$$

where $\psi$ is the signal correlation matrix and $\sigma^2 I_{N\times N}$ is the noise correlation matrix. The $M$ eigenvalues $\Lambda_j$ of $\mathbf{R}$ obey the following relations [45, 151]:

$$\begin{array}{rcll} \Lambda_j & = & \mu_j + \sigma^2 & \forall j \in [1, 2, \ldots, N] \text{ and } \mu_j \in \mathbb{R}^+, \\ \Lambda_j & = & \sigma^2 & \forall j \in [N+1, \ldots, M]. \end{array} \tag{7.4}$$

Hence, if $M > N$, the eigenvectors $V_j$ associated to the eigenvalues $\Lambda_j$ can be separated in two groups:

$E_S$ = $[V_1, V_2, \ldots, V_N]$ the signal subspace associated to the $N$ largest eigenvalues,
$E_N$ = $[V_{N+1}, V_{N+2}, \ldots, V_M]$ the noise subspace associated to the $M - N$ smallest eigenvalues.

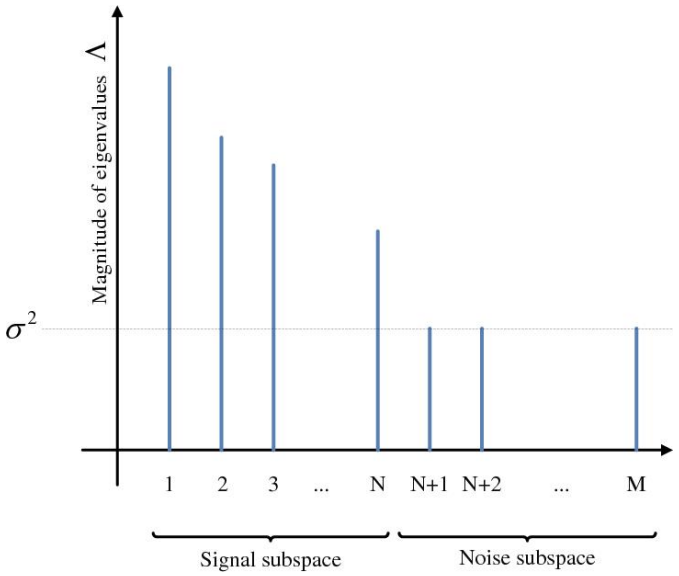



Figure 7.1: Typical theoretic distribution of eigenvalues of the covariance matrix in presence of $N$ sources and $M>N$ microphones: the $M-N$ smallest eigenvalue are equal.

The rank of $\mathbf{R}$ can so be deduced from the multiplicity of its smallest eigenvalues as illustrated in Fig. 7.1.

If the theory seems very attractive because of its simplicity, in practice, the smallest eigenvalues are never perfectly equal because of the finite size of the observations [146] making the signal and noise eigenspaces difficult to distinguish. This is why several methods have been proposed. One of the most popular relies on the Information Theoretic Criterion: the idea is to test a family of $P$ hypothesis, where the hypothesis $p$ reflects the equality between the $M-p$ smallest eigenvalues, and see which hypothesis best fits the data (*i.e.* which hypothesis has the maximum likelihood). As maximum likelihood estimators are generally biased, penalty functions are introduced to correct the bias. The most well-known of them are the AIC (Akaike Information Criterion) [152], MDL (Maximum Description Length) [148], EDC (Efficient Detection Criterion) [147], MDL-BSS [153] to list a few. But all above-mentioned are effective if and only if the number of sensors is larger than the number of sources $M{=}N_{max}$.

In what follows, the case $M \geq N_{max}$ is investigated for $N_{max} = 2$.

## 7.3 Array geometry vs. rank of the correlation matrix

From Eq. (7.1), all the information about i) sensors locations in relation to ii) the sources locations and iii) the sources wavelength is contained in $\mathbf{A}$. Consider the case where the number of sensors $M = N_{max} = 2$. In such a situation, the mixing matrix $\mathbf{A}$ has the following form:

$$\mathbf{A} = \begin{pmatrix} \gamma_{11}e^{-i2\pi a} & \gamma_{12}e^{-i2\pi b} \\ \gamma_{21}e^{-i2\pi c} & \gamma_{22}e^{-i2\pi d} \end{pmatrix}, \tag{7.5}$$

where

$$a = \frac{||\mathbf{r}_1^m - \mathbf{r}_1^s||}{\lambda_1}, \quad b = \frac{||\mathbf{r}_1^m - \mathbf{r}_2^s||}{\lambda_2}, \quad c = \frac{||\mathbf{r}_2^m - \mathbf{r}_1^s||}{\lambda_1}, \quad d = \frac{||\mathbf{r}_2^m - \mathbf{r}_2^s||}{\lambda_2}, \tag{7.6}$$

and

$$\gamma_{jk} = \frac{\beta_k}{4\pi \left|\left|\mathbf{r}_j^m - \mathbf{r}_k^s\right|\right|^2}. \tag{7.7}$$

In the context of a compact and far-field sensor array, the distances between sensors is low compared to distances between microphones and sources. Hence the above model can be simplified by letting $\gamma_{jk} = \alpha_k$ where $\alpha_k$ is a positive constant which represents the initial intensity level of the source $j$.

Under the assumption of mutually uncorrelated sources and independent and identically distributed (i.i.d) noise, $rank\{\mathbf{R}\} = rank\{\mathbb{E}\left\{\mathbf{A}\mathbf{A}^H\right\}\}$. Let us explore what the expression of an eigenvalue $\Lambda$ of $\mathbf{A}\mathbf{A}^H$ is. An eigenvalue $\Lambda$ obeys $P(\Lambda) = 0$ with:

$$\begin{aligned} P(\Lambda) &= det\left(AA^H - \Lambda I_{M\times M}\right), & (7.8)\\ &= \Lambda^2 - 2(\alpha_1^2 + \alpha_2^2)\Lambda + 4\alpha_1^2\alpha_2^2 sin^2(\pi(a - b + c - d)). & (7.9) \end{aligned}$$

This yields two solutions:

$$\Lambda_1 = \frac{2(\alpha_1^2 + \alpha_2^2) + \sqrt{\Delta}}{2}, \Lambda_2 = \frac{2(\alpha_1^2 + \alpha_2^2) - \sqrt{\Delta}}{2}, \tag{7.10}$$

with $\Delta$ equals to:

$$\Delta = 4(\alpha_1^2 + \alpha_2^2)^2 - 16\alpha_1^2\alpha_2^2 sin^2(\pi(a - b + c - d)). \tag{7.11}$$

Now that expressions for $\Lambda_1$ and $\Lambda_2$ have been established as a function of the position of sensors, and positions, amplitude and wavelength of sources, let us explore what the optimal position of a microphone is - given the position of the other one and all the other sources-related parameters - in a source separation context.

### 7.3.1 Optimal array for source separation

When the number of sensors is the same as the number of sources, a perfect source separation can be achieved under the condition that the observations are independent (in a second order sense). Given $\mathbf{r}_1^m$, $\mathbf{r}_1^s$, $\mathbf{r}_2^s$ $\lambda_1$ and $\lambda_2$, one has to find the position $\mathbf{r}_2^m$ for which the geometric multiplicity of $\mathbf{A}\mathbf{A}^H$ equals $N_{max}$, *i.e*:

$$\text{find } \mathbf{r}_2^m \quad \text{such that} \quad dim[Ker(\mathbf{A}\mathbf{A}^H - \lambda\mathbf{I})] \quad = N_{max}. \tag{7.12}$$

**Reminder** The algebraic and geometric multiplicities are two distinct measures of the number of eigenvectors belonging to an eigenvalue. The algebraic multiplicity of an eigenvalue is defined as the multiplicity of the corresponding root of the characteristic polynomial (7.8). The geometric multiplicity of an eigenvalue is defined as the dimension of the associated eigenspace, namely, the number of linearly independent eigenvectors with that eigenvalue.

Since $AA^H$ is an Hermitian matrix, algebraic and geometric multiplicities are the same. Thus, a sufficient condition to verify (7.12) is to make the eigenvalues equal. From (7.10) and (7.11), one gets:

$$\Lambda_1 = \Lambda_2 \Leftrightarrow \Delta = 0, \tag{7.13}$$

which leads to:

$$a - b + c - d = \pm\frac{1}{\pi} Arcsin\left(\frac{\alpha_1^2 + \alpha_2^2}{2\alpha_1\alpha_2}\right). \tag{7.14}$$

Because the parameters $a$, $b$, $c$ and $d$ are real, the initial intensities $\alpha_1$ and $\alpha_2$ of the sources have to respect the following contraint to give a physical solution:

$$\left|\frac{\alpha_1^2 + \alpha_2^2}{2\alpha_1\alpha_2}\right| \leq 1. \tag{7.15}$$

Without loss of generality, setting $\alpha_2 = x\alpha_1$ with $x \in \mathbb{R}^+$ yields:

$$1 + x^2 \leq 2x. \tag{7.16}$$

The solution $x = 1$ is the only physical one. The optimal position $\mathbf{r}_2^m$ can be found only when both sources have the same initial radiating intensity. In other cases, only suboptimal separation can be achieved with two microphones and more evolved methods have to be used (such as spatial filtering using many more sensors). If $x = 1$, Eq. (7.14) yields the final equality constraint $h$ that $\mathbf{r}_2^m$ has to verify with respect to $\mathbf{r}_1^m$, $\mathbf{r}_1^s$ and $\mathbf{r}_2^s$:

$$h(\mathbf{r}_2^m) = \frac{1}{\lambda_1}\left(||\mathbf{r}_1^s - \mathbf{r}_2^m|| - ||\mathbf{r}_1^s - \mathbf{r}_1^m||\right) + \frac{1}{\lambda_2}\left(||\mathbf{r}_2^s - \mathbf{r}_1^m|| - ||\mathbf{r}_2^s - \mathbf{r}_2^m||\right) \pm \frac{1}{2} = 0. \tag{7.17}$$

Let us now hold the same reasoning for source number estimation.

### 7.3.2 Optimal array for source number estimation

With $N_{max} = 2$ in the present case, the objective is to discriminate between three cases:

- case a): both sources radiate;
- case b): one source radiates;
- case c): no sources radiate.

Because of the initial intensity of the sources is not known, the eigenvalues cannot be predicted. Assuming both sources radiate with equal intensity, one should use the ratio:

$$r = \frac{\Lambda_2}{\Lambda_1}. \tag{7.18}$$

In the case where $\mathbf{r}_2^m$ respects the constraint (7.17), cases a) and c) cannot be dissociated. Indeed, in case a) $r$ is equal to 1 because of the independence of the two signals for this geometry. But by definition of an i.i.d noise $\mathbf{W}$, $r$ is also equal to 1 in the case c) because both eigenvalues are equal to $\sigma^2$.

**Remark** The previous point revealed that in the case of two sources radiating with equal intensity, the array geometry for their optimal separation is different from that for the estimation of their number. This is rather bad news since most of the source separation algorithms require the exact number of sources.

Let us continue with the source number estimation problem. According to the above-mentioned results, another optimal $\mathbf{r}_2^m$ has to be found for this specific purpose.

The first condition that $r$ has to respect is to not be equal to 1 or 0 in the case a) in order to avoid ambiguity with case c) and b) respectively, *i.e.*:

$$\text{choose } \mathbf{r}_2^m \quad \text{such that} \quad a - b + c - d \neq \begin{cases} \pm\frac{1}{2} & \text{if } \alpha_1 = \alpha_2, \\ \mathbb{Z} & \text{otherwise.} \end{cases} \tag{7.19}$$

For instance, if one wants $r$ to be equal to 0.5 in case a), the following constraint has to be respected:

$$h(\mathbf{r}_2^m) = \frac{1}{\lambda_1}\left(||\mathbf{r}_1^s - \mathbf{r}_2^m|| - ||\mathbf{r}_1^s - \mathbf{r}_1^m||\right) + \frac{1}{\lambda_2}\left(||\mathbf{r}_2^s - \mathbf{r}_1^m|| - ||\mathbf{r}_2^s - \mathbf{r}_2^m||\right) \pm \frac{1}{\pi}\arccos\left(\frac{1}{3}\right) = 0. \tag{7.20}$$

### 7.3.3 Optimization procedure

This part gives more details on how to find the optimal position $\mathbf{r}_2^m$ automatically, given an acoustic scenario.

From Eq. (7.17) and (7.20), the optimal position of $\mathbf{r}_2^m$ can be found using a standard optimization method formulated as:

$$\min_{\mathbf{r}_2^m \in \mathbb{R}^2} \quad f(\mathbf{r}_2^m), \tag{7.21}$$

$$\text{subject to} \quad h(\mathbf{r}_2^m) = 0, \tag{7.22}$$

where the function to minimize is the distance between microphones, from the primary

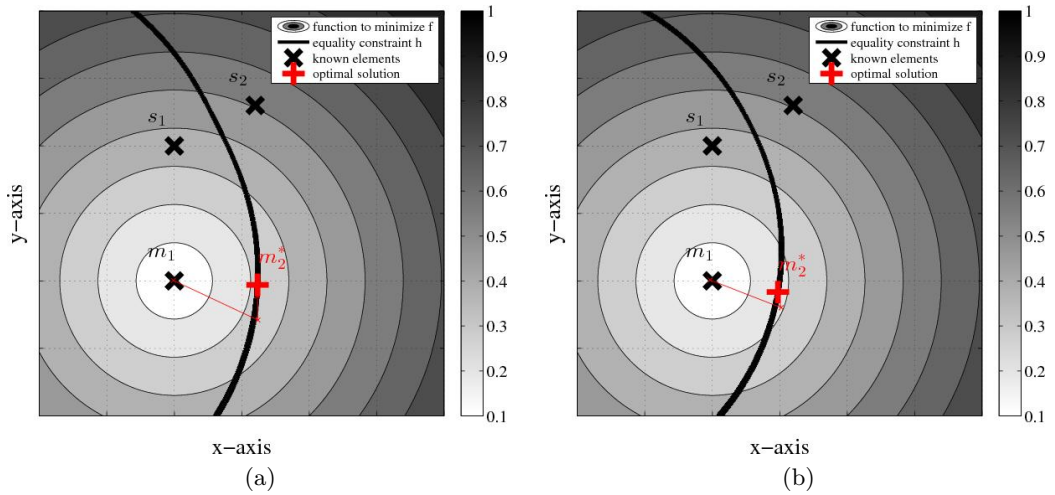



Figure 7.2: Optimal location of the second microphone $m_2$ given the location of the first one $m_1$ for two different contexts: sources separation (a) and sources detection (b). The acoustical scenario is the same for both cases.

objective of having the smallest possible array:

$$f(\mathbf{r}_2^m) = ||\mathbf{r}_2^m - \mathbf{r}_1^m|| \,. \tag{7.23}$$

A standard method to solve such a non linear convex optimization problem is the Local Sequential Quadratic Programming method (Local-SQP). For a complete description of Local-SQP, see for example [154] page 465.

As an example, let us consider a simulation with the following acoustic scenario:

- position of the sources: $\mathbf{r}_1^s = [0, 0]^T$ m, $\mathbf{r}_2^s = [0.6, 0.3]^T$ m;
- frequencies of the sources: $f_1 = 600$ Hz, $f_2 = 500$ Hz;
- intensity of the sources: $\alpha_1 = \alpha_2 = 1$;
- position of the first microphone: $\mathbf{r}_1^m = [0, -1]^T$ m;
- speed of sound: $c = 343$ m/s.

Solutions of the optimal $\mathbf{r}_2^m$ are depicted in Fig. 7.2a, respectively Fig. 7.2b, for the source separation, respectively source number estimation. On these plots, the two sources and the first microphone are symbolized by black crosses, the function to minimize is symbolized by the concentric circles (the darker, the farthest of $\mathbf{r}_1^m$), the equality constraint function is the black curve and the solution is the red cross. For the source separation, the optimal $\mathbf{r}_2^m$ is $[0.62, -1.03]^T$ m and for the source number estimation, the optimal $\mathbf{r}_2^m$ is $[0.48, -1.08]^T$ m.

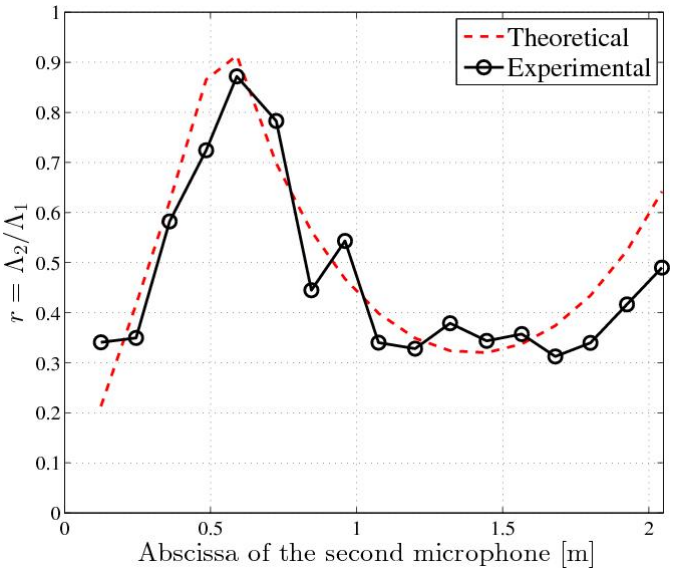



Figure 7.3: Theoretical and experimental values of $r$ [Eq. (7.18)] as a function of the position of the second sensor [Table 7.1].

## 7.4 Experimental measurements in anechoic conditions

An experimental measurement was carried out to validate the presented theoretical results. Two sound sources radiated a pure tone with respective frequencies $f_1 = 2000$ Hz and $f_2 = 3000$ Hz, and the same intensity. The two loudspeakers were placed in an anechoic room at coordinates $\mathbf{r}_1^s = [0, 0]^T$ m, $\mathbf{r}_2^s = [0.5, 0]^T$ m. The first microphone was placed at $\mathbf{r}_1^m = [0, -4]^T$ m. The second microphone was placed at the same ordinate of -4 m and acquisition was performed for 17 different abscissas. The tested abscissas are given in Table 7.1.

| Position Number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Abscissa of $m_2$ [m] | 0.125 | 0.245 | 0.36 | 0.48 | 0.59 | 0.725 | 0.845 | 0.96 | 1.075 |
| Position Number | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | |
| Abscissa of $m_2$ [m] | 1.2 | 1.32 | 1.445 | 1.565 | 1.68 | 1.8 | 1.925 | 2.045 | |

Table 7.1: Successive tested abscissas of microphone $m_2$.

For each coordinate $\mathbf{r}_2^m$, the ratio $r$ of Eq. (7.18) was computed. One can remark on Fig. 7.3 that theoretical and experimental values of $r$ match pretty well for small distances between $m_1$ and $m_2$ (below 1.6 m). The fifth position of $m_2$ is the one which maximizes the independence between recordings. It can be said this position is the optimal one for source separation with regard to all tested positions (or positioning constraints).

A standard result in statistical signal processing theory is that an efficient[1] estimator of

[1] a finite-sample estimator is said efficient if it is unbiased and if it attains the Cramer-Rao Lower Bound

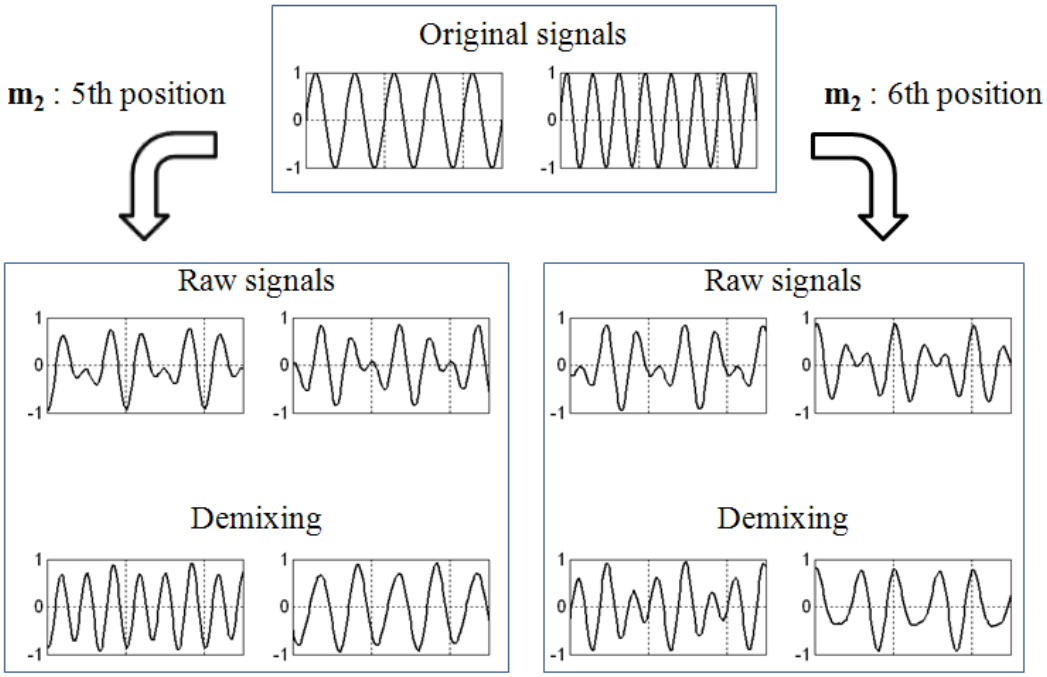



Figure 7.4: Experimental result of sound source separation using Eq. (7.24) when the second microphone is placed at the optimal position (the fifth one) and another one 13.6 cm side (the sixth one).

**S** is [136]:

$$\hat{\mathbf{S}} = \left(\mathbf{A}^H\mathbf{A}\right)^{-1}\mathbf{A}^H\mathbf{X}. \tag{7.24}$$

Computation of the estimate was performed for two positions: the fifth (optimal : , i.e $m_2 = [0.59, -4]^T$) and the sixth (i.e $m_2 = [0.725, -4]^T$). Both results are represented in Fig. 7.4. As expected the estimate of the original sources is conclusive when $\mathbf{r}_2^m$ is the optimal position and much less when $\mathbf{r}_2^m$ is at a few centimeters from the optimum. This proves the influence of the microphone array geometry on the performance of a source separation application.

Similarly, the smallest microphone array making $r$ equals to 0.5 is obtained for $\mathbf{r}_2^m$ between the second and third position. A recording was carried out with $\mathbf{r}_2^m = [0.31, 0]m$ and with $s_1$ and $s_2$ radiating randomly (Fig. 7.5-a). The three different cases : no signal, one signal and two signals, are clearly distinguishable and conform to the theory as illustrated in Fig. 7.5-b. Using the fifth position would have not permitted to differentiate the "no signal" case from the "two signals" case, confirming that an optimal microphone array for source separation is not necessary optimal for source number estimation and vice-versa.

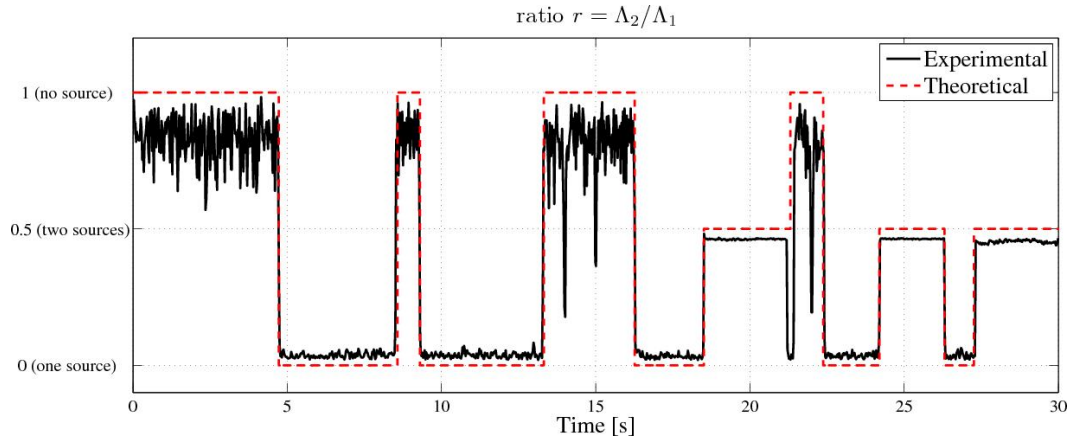



Figure 7.5: Theoretical and experimental result of source number estimation using Eq. (7.18) .

## 7.5 Conclusion

This chapter discussed the subspace-based theory as a potential framework to estimate the number of axles as vehicles pass by. The objective is to adapt the tracking model with respect to the very first observations. To be efficient, the subspace-based approach requires a high number of sensors, larger than the number of sources. It also requires a high number of snapshots (observations). Neither of these conditions is achieved in the context of this thesis. We therefore investigated the case when the number of sensors is equal to two when the number of sources can be equal to zero, one or two. The idea is to control the rank of the correlation matrix by acting on the microphone array geometry. For the case of two independent, tonal sound sources acquired by two microphones, the relationship between these eigenvalues and the acoustical scenario has been derived. The analysis of these expressions led us to design an optimization procedure aiming at finding the best array for source separation and source number estimation. It has been proved that the two estimation problems do not admit the same solution. A new technique for source number estimation has been proposed. It consists in studying the ratio of the eigenvalues of the correlation matrix. Depending on the array, this ratio can be equal to three different values corresponding to the three cases, namely no source, one source and two sources. This makes possible the source number estimation in the case where the noise-subspace can potentially be unavailable. All the presented developments have been validated through experimental results. A logical extension to this work could be to investigate the case of broadband and independent sound sources.

# 8 Conclusions and Perspectives

This thesis introduced a novel acoustic road traffic monitoring technique through passive acoustic sensing. With the objective to facilitate traffic data analysis, currently a difficult task due to data heterogeneity, we proposed audio processing strategies involving small, light and easily movable microphone arrays used both as sound level meter and also as all-in-one traffic analysis stations.

We considered the context of an unknown number of moving wideband, sound sources, in a non-reverberant and non-dispersive medium monitored by a small number of sensors placed on the roadside. The efforts were mainly focused on the observation, detection and estimation of motion and geometrical parameters of passing-by vehicles.

The first part of this work relates to the observation of vehicles through their pass-by noise. Inspired by current standardized sound-level-meter-based measurements, we aimed at developing audio processing algorithms applicable to compact microphone arrays (the smallest possible number of microphones, the smallest possible size) placed on the roadside. After a review of airborne sound source localization techniques and time-delay estimators, we oriented our efforts towards the phase-transform generalized cross-correlation function (GCC-PHAT), which is one of the most relevant tools for the extraction of the vehicle trajectories. But it is also relevant for multiple axle trajectories, observed during pass-by, using only two microphones placed in parallel to the road lane.

The first original contribution of this thesis was to improve GCC-PHAT processing in the light of the acoustical properties of the pass-by noise in order to improve the axle observation quality. This gave the *GCC-BPHAT* function, whose analytical expression has also been derived. It was tested in simulations and for parameter optimization in the case of one or $N$ uncorrelated stochastic sound sources.

In the second part, we developed a procedure enabling the joint automatic and joint estimation of speed and wheelbase length for two-axle vehicles, through a set of GCC-BPHAT-based observations. For this purpose, classical Bayesian-based tracking algorithms were reviewed. Due to the non-linearity and non-Gaussianity of the problem at hand, Kalman filters were discarded in favor of the particle filtering technique.

Two main contributions of the thesis are related to this aspect. The first contribution was to bridge passive acoustics monitoring and Bayesian statistics, linking each probability function to the available acoustic-based measurements and the *a-priori* knowledge of the target motion. An experiment has been carried out to validate this approach, and also for pedagogical purposes. The second contribution was to establish a target model, specifically dedicated to two-axle vehicles, including their geometrical properties, and the variable contribution of each tyre/road interaction as a function of the vehicle direction of arrival. We called this new model the *bimodal sound source model*. Combination of this model and a particle filter led the *bimodal particle filter* making possible for the time, wheelbase length estimation with a two-microphone array, and potentially returning an estimate in real time.

These developments on observation and tracking have then inspired an innovative strategy for microphone array design, constituting the fourth contribution of this thesis. This strategy take into account both data processing and measurement techniques for the optimization of the inter-sensor distance between microphones. We also argued in favor of a third microphone, added to the (theoretically sufficient) two-element array, and the use of the MULTI-PHAT technique, as a mean to exploit the redundant information between sensor pairs.

Finally, three detection strategies have been proposed. To the best of our knowledge, two of them have never been proposed before and can be considered as the fifth and sixth contributions of this thesis. Among the two, the *endfire detection strategy* consists in continuously monitoring a zone upstream the array by looking at the evolution of time-delay of arrivals and comparing it to a model through the 2D Pearson correlation coefficient, the other one consists in looking at the ratio of the eigenvalues of the observation correlation matrix using a specific array design based on the knowledge of the source position and wavelength.

All the proposed theoretic developments have also been assessed through *in-situ* measurements that we designed.

Speed estimates were compared with those obtained using standardized radar ones. When the 3MBPF strategy is used (triangular array associated to a bimodal particle filter), the error is below 5 km/h for 75% of vehicles. Moreover, BPF-based strategies enable wheelbase length estimation, in addition to speed, of two-axle vehicles during pass-by. Best results for wheelbase length estimation are achieved with two microphones (2MBPF), the estimation error is below 30 cm for 91% of the two-axle detected vehicles, that is, a spatial accuracy smaller than a wheel diameter.

The proposed acoustic-based strategy that consists in detecting vehicles in the endfire direction provides good results, since 94% of the vehicles have been correctly detected for the left-to-right direction (closest lane) and 90% for those of the opposite lane (farthest lane).

**Perspectives**

Many questions are still open and pave the way to future works. They are mainly related to observation, detection and tracking.

Given that the tracking step requires a precise knowledge of target positions at initialisation, it is essential to design a precise and reliable detection algorithm. We showed in the experimental part of this thesis that answering the question "is there any vehicle?" is relatively simple using acoustics, but the question "is there any vehicle and if yes, where exactly ?" is a much more tricky problem whether resorting on video or acoustic-based techniques, especially when the microphone array is compact and the detection zone is far upstream it. One solution could be to install two arrays, spaced a few meters apart, the first one returning an alert as soon as a vehicle is in front of it, the second dedicated to the tracking. Another solution could be to resort to multimodal detection, for instance, using video and audio signals like in [115, 155]. Indeed, it is more and more confirmed that *multimodal* detectors allow better results *in-situ* than with unimodal ones [115, 156].

Another possibility of improvement concerns the automatic adaptability of the likelihood model with respect to the observation. In this thesis only conventional two-axle vehicles, like cars, have been modeled, but many other kinds of vehicles exist so that an "ideal" filter would require to store different models in memory (a model for trucks, a model for motorbikes etc...) and switch between models according to the very first measurements. Similarly, one target could obey several dynamical models during one observation (for instance, constant speed at the beginning, then a stop followed by an acceleration). Recent filters have been developed to allow model switches, namely Interacting Multiple Model (IMM) filters [157, 158, 86]. This should constitute the object of further investigations as a potential improvement of the presented work.

The main drawback of particle filtering is the number of parameters to adjust, making its use dependent on the practitioner's experience. An improvement could be to automatically set each parameter to its optimal values after a quick pre-processing of the very first measurements.

Finally, only 2D and compact arrays are addressed in this thesis. A way to improve the observation could be to resort to 3D and/or distributed arrays like [24, 25, 156].

# A Appendix

## A.1 Hyperbolic localization in 2D: some analytical solutions

This section presents a short mathematical development on a turnkey solution to locate a source in the Cartesian 2D plan when two centered sensor pairs of same size are available. Note that the extensions to the 3D case and arbitrary sensor distribution are the object of many theoretical studies that are outside the context of this thesis. We advise the reader interested in such more complex cases to refer to the papers of Chan *et. al* [159, 160] and Spiesberger [161, 138, 162]. The following calculations were developed by ourselves.

### A.1.1 Hyperbola equation

Let $\mathcal{R}_{0;\vec{i},\vec{j}}$ be a direct orthonormal basis and $S$ a source with two dimensional coordinates $\mathbf{r}^s$ in $\mathcal{R}$ and two microphones $m_1$ and $m_2$ with respective coordinate $\mathbf{r}_1^m = [d/2, 0]^T$ and $\mathbf{r}_2^m = [-d/2, 0]^T$ in $\mathcal{R}$. The sound speed $c$ is assumed constant and the medium homogeneous. Let $\tau_{12}$ be the TDOA of the wave between $m_1$ and $m_2$. $\tau_{12}$ is related to the source and sensors positions through the relation:

$$\tau_{12} = \frac{||\mathbf{r}^s - \mathbf{r}_2^m|| - ||\mathbf{r}^s - \mathbf{r}_1^m||}{c}, \tag{A.1}$$

Note that the numerator of (A.1) describes an half-hyperbola $H_1$ with focal points $\mathbf{r}_1^m$ and $\mathbf{r}_2^m$ with equation:

$$H_1 : \frac{x^2}{a_1^2} - \frac{y^2}{b_1^2} = 1, \tag{A.2}$$

where $x$ and $y$ are the variable of the 2D orthonormal basis and $a_1$ and $b_1$ are scalars

defined by:

$$\begin{aligned} a_1 &= \frac{-c\tau_{12}}{2}, \\ b_1 &= \sqrt{\left(\frac{d}{2}\right)^2 - a_1^2}. \end{aligned} \tag{A.3}$$

### A.1.2 Intersection of two hyperbola

Let $H_i$ be an hyperbola of $\mathcal{R}$ . After a rotation of angle $\theta_i$ and a translation from O towards the point $\mathrm{M}_i(\mathrm{x}_{0i},\mathrm{y}_{0i})$, any point of $H_i$ with coordinates $(\mathrm{x}_i\text{'},\mathrm{y}_i\text{'})$ satisfies the relation:

$$\begin{pmatrix} x_i' \\ y_i' \end{pmatrix} = \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} x_{0i} \\ y_{0i} \end{pmatrix}. \tag{A.4}$$

Hence,

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos\theta_i & \sin\theta_i \\ -\sin\theta_i & \cos\theta_i \end{pmatrix} \begin{pmatrix} x_i' - x_{0i} \\ y_i' - y_{0i} \end{pmatrix}. \tag{A.5}$$

Replacing $x$ and $y$ by their expression in (A.2) gives:

$$H_i : A_1 x_i'^2 + B_1 x_i' y_i' + C_1 y_i'^2 + D_1 x_i' + E_1 y_i' + F_1 = 1, \tag{A.6}$$

with

$$\begin{aligned} A_i &= \frac{\cos\theta_i{}^2}{a_i^2} - \frac{\sin\theta_i{}^2}{b_i^2}, \\ B_i &= 2\gamma_i\tau_i, \\ C_i &= \frac{\sin\theta_i{}^2}{a_i^2} - \frac{\cos\theta_i{}^2}{b_i^2}, \\ D_i &= -2A_i x_{0i} - 2\gamma_1 y_{0i}\tau_{12,i}, \\ E_i &= -2C_i y_{0i} - 2\gamma_1 x_{0i}\tau_{12,i}, \\ F_i &= A_i x_{0i}^2 + 2\gamma_i\tau_i x_{0i} y_{0i} + C_i y_{0i}^2. \end{aligned} \tag{A.7}$$

where $\gamma_i = \cos\theta_i \sin\theta_i$ et $\tau_i = \frac{1}{a_i^2} + \frac{1}{b_i^2}$. Finding the intersection point of coordinate $(\mathrm{x}_s,\mathrm{y}_s)$ between two hyperbolas $\mathrm{H}_1$ and $\mathrm{H}_2$ is thus equivalent to solving the following system:

$$\begin{cases} A_1 x_s^2 + B_1 x_s y_s + C_1 y_s^2 + D_1 x_s + E_1 y_s + F_1 = 1 \\ A_2 x_s^2 + B_2 x_s y_s + C_2 y_s^2 + D_2 x_s + E_2 y_s + F_2 = 1 \end{cases} \tag{A.8}$$

In the case where the two pairs are orthonormal and centered, then $D_i = E_i = F_i =$

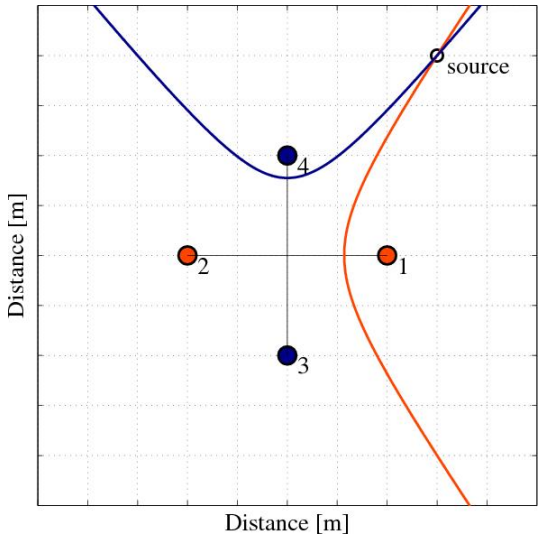



Figure A.1: Hyperbolic-based sound source localization using two centered and orthogonal pairs.

$B_i = 0$, $\forall$ i $\in$ {1,2}. The system to solve becomes:

$$\begin{cases} A_1 x_s^2 + C_1 y_s^2 = 1 \\ A_2 x_s^2 + C_2 y_s^2 = 1 \end{cases} \tag{A.9}$$

Mathematically speaking, four solutions are possible: $(x_a, y_a), (-x_a, y_a), (x_a, -y_a), (-x_a, -y_a)$ with:

$$\begin{aligned} x_a &= \sqrt{\frac{C_2 - C_1}{A_1 C_2 - A_2 C_1}}, \\ y_a &= \sqrt{\frac{A_1 - A_2}{A_1 C_2 - A_2 C_1}}. \end{aligned} \tag{A.10}$$

Physically speaking, only one of these solutions is possible:

$$\begin{aligned} x_a &= -sign(\tau_{12}) \sqrt{\frac{C_2 - C_1}{A_1 C_2 - A_2 C_1}}, \\ y_a &= -sign(\tau_{34}) \sqrt{\frac{A_1 - A_2}{A_1 C_2 - A_2 C_1}}, \end{aligned} \tag{A.11}$$

## A.2 Closed-form expression of the GCC-BPHAT function in the single source case

Without noise and under free-field conditions, the signal acquired by one sensor is a delayed version of the signal acquired by the other sensor, such that:

$$y_2(t) = y_1(t + \tau_{12}). \tag{A.12}$$

Eq. (A.12) may be translated to the frequency domain by:

$$Y_2(f) = Y_1(f)e^{+2j\pi f\tau_{12}}, \tag{A.13}$$

where $Y_i$(f) and $y_i$(t) are related by the Fourier and inverse Fourier transforms according to the conventions:

$$Y_i(f) = \int_{-\infty}^{+\infty} y_i(t)e^{-2j\pi ft}dt, \tag{A.14}$$

$$y_i(t) = \int_{-\infty}^{+\infty} Y_i(f)e^{+2j\pi ft}df. \tag{A.15}$$

Substituting (A.13) into the expression of the generalized cross-correlation (2.16) with $\psi_g(f)$ the BPHAT weighting (4.5) gives:

$$R^{bphat}(\tau) = \int_{-\infty}^{+\infty} \frac{Y_1Y_1^*}{|Y_1Y_1^*|}e^{2j\pi f(\tau-\tau_{12})}, \tag{A.16}$$

$$= \int_{-f^+}^{-f^-} e^{2j\pi f(\tau-\tau_{12})} + \int_{f^-}^{f^+} e^{2j\pi f(\tau-\tau_{12})}, \tag{A.17}$$

$$= 2\mathbf{Re}\left[\int_{f^-}^{f^+} e^{2j\pi f(\tau-\tau_{12})}df\right], \tag{A.18}$$

where **Re**[.] is the real part. Futhermore:

$$\int_{f^-}^{f^+} e^{2j\pi f(\tau-\tau_{12})}df = \frac{e^{2j\pi f^+(\tau-\tau_{12})} - e^{2j\pi f^-(\tau-\tau_{12})}}{2j\pi(\tau-\tau_{12})}, \tag{A.19}$$

$$= \frac{e^{j\pi(f^++f^-)(\tau-\tau_{12})}sin(\pi(f^+ + f^-)(\tau-\tau_{12}))}{\pi(\tau-\tau_{12})}. \tag{A.20}$$

Replacing $f^+ + f^-$ by $2f_c$ and $f^+ - f^-$ by $B_w$ yields expression (4.6).

## A.3 Global percentage error and relative standard deviation

This section establishes some mathematical metrics to assess the performance of a particle filtering algorithm.

A test scenario is defined by the following geometrical, acoustical, and statistical parameters:

- the distance to the road $D$;
- the inter-sensor distance $d$;
- the speed of sound $c$;
- the bandwidth of the signal of interest: $B_w$ and $f_c$;
- the actual target state values: $x$,$y$,$\dot{x}$,$wb$;
- the *a priori* target state values at initialisation: $\mu_{x,0}$,$\mu y, 0$,$\mu_{\dot{x},0}$,$\mu_{wb,0}$;
- the *a priori* target state values at initialisation: $\mu_{x,0}$,$\mu_{y,0}$,$\mu_{\dot{x},0}$,$\mu_{wb,0}$;
- the uncertainties on the *a priori*: $\sigma^2_{x,0}$,$\sigma^2_{y,0}$,$\sigma^2_{\dot{x},0}$,$\sigma^2_{wb,0}$;
- the dynamical noise variances: $\sigma^2_x$,$\sigma^2_y$,$\sigma^2_{\dot{x}}$,$\sigma^2_{wb}$;
- the number of particles $N_p$.

Due to the stochastic nature of the Monte-Carlo-based process, the performance of the bimodal particle filter are averaged over a high number of runs for each tested scenario. For each run $k$, $k \in [1, 2, ..., N_{test}]$, the mean and standard deviations $\mu^{(k)}(\alpha_j, T)$ and $\sigma^{(k)}(\alpha_j, T)$ of the $j^{th}$ particle distribution is returned at the end of the tracking. From the state vector expression (4.3), $j = 1$ corresponds to the abscissa state $x$, $j = 2$ is for the ordinate state $y$, $j = 3$ is for the speed state $\dot{x}$ and $j = 4$ is for the wheelbase state $wb$.

After the $N_{test}$ runs, the $j^{th}$ *global error* is computed. The global error is defined as the relative difference between the actual value $\alpha_j$ and the quantity $\Sigma_{\mu,j}$:

$$\Sigma_{\epsilon,j} = \Sigma_{\mu,j} - \alpha_j, \tag{A.21}$$

where $\Sigma_{\mu,j}$ is the mean of all the $N_{test}$ means:

$$\Sigma_{\mu,j} = \frac{1}{N_{test}} \sum_{k=1}^{N_{test}} \mu^{(k)}_{\alpha_j,T}. \tag{A.22}$$

Note that for graphical reasons, the global error can be expressed in percentage, giving the *global percentage error* $\Sigma^{\%}_{\epsilon,j}$ expressed by:

$$\Sigma^{\%}_{\epsilon,j} = 100 \times \Sigma_{\epsilon,j}/\alpha_j. \tag{A.23}$$

Following the same idea, the *global standard deviation* $\Sigma_{\sigma,j}$ is obtained by computing the

within-group sum-of-square (SSW), that is, the mean of the $N_{test}$ variances $\sigma^{(k)}(\alpha_j, T)$ and the between-group sum-of-square (SSB), that is, the variance of the $N_{test}$ means $\mu^{(k)}(\alpha_j, T)$, such that:

$$\Sigma_{\sigma,j} = \sqrt{SSW_j + SSB_j}, \tag{A.24}$$

where

$$SSW_j = \frac{1}{N_{test}} \sum_{k=1}^{N_{test}} \left(\sigma^{(k)}_{\alpha_j,T}\right)^2, \tag{A.25}$$

$$SSB_j = \frac{1}{N_{test}} \sum_{k=1}^{N_{test}} \left(\mu^{(k)}_{\alpha_j,T} - \Sigma_{\mu,j}\right)^2. \tag{A.26}$$

To clarify, the SSW is a measure of the variation of particles within each run and the SSB is a measure of the differences between estimates on each test. Again one can express the *relative total standard deviation*:

$$\Sigma^{\%}_{\sigma,j} = 100 \times \Sigma_{\sigma,j}/\Sigma_{\epsilon,j}. \tag{A.27}$$

## A.4 The SRP-PHAT and MULTI-PHAT Techniques

The SRP-PHAT [116], or Global Coherence Field (GCF) [163], is a popular and powerful tool to compute acoustic maps, also called spatial likelihood function (SLF). In its most traditional form, it relies on a filter-and-sum beamformer, whence the term *SRP* for "Steered Response Power", and GCC-PHAT functions, whence the term *PHAT*. A description of this technique is proposed below.

Let $\mathbf{r}^s$ be the actual coordinate of a sound source monitored by an array of $M$ sensors with known geometry. Such an array is constituted of $P$ sensor pairs such that:

$$P = \frac{M(M-1)}{2}. \tag{A.28}$$

Let $\mathbf{r}^m_{p,1}$ and $\mathbf{r}^m_{p,2}$ be the sensors position forming the $p^{th}$ pair of the array, $p \in [1, 2, ..., P]$, and let $\mathbf{r}^{(n)}$ be the $n^{th}$ candidate position among a set of $N$ ones. The hypothetical delay $\tau_p^{(n)}$ between the sensors of the $p^{th}$ pair and inherent to $\mathbf{r}^{(n)}$ equals:

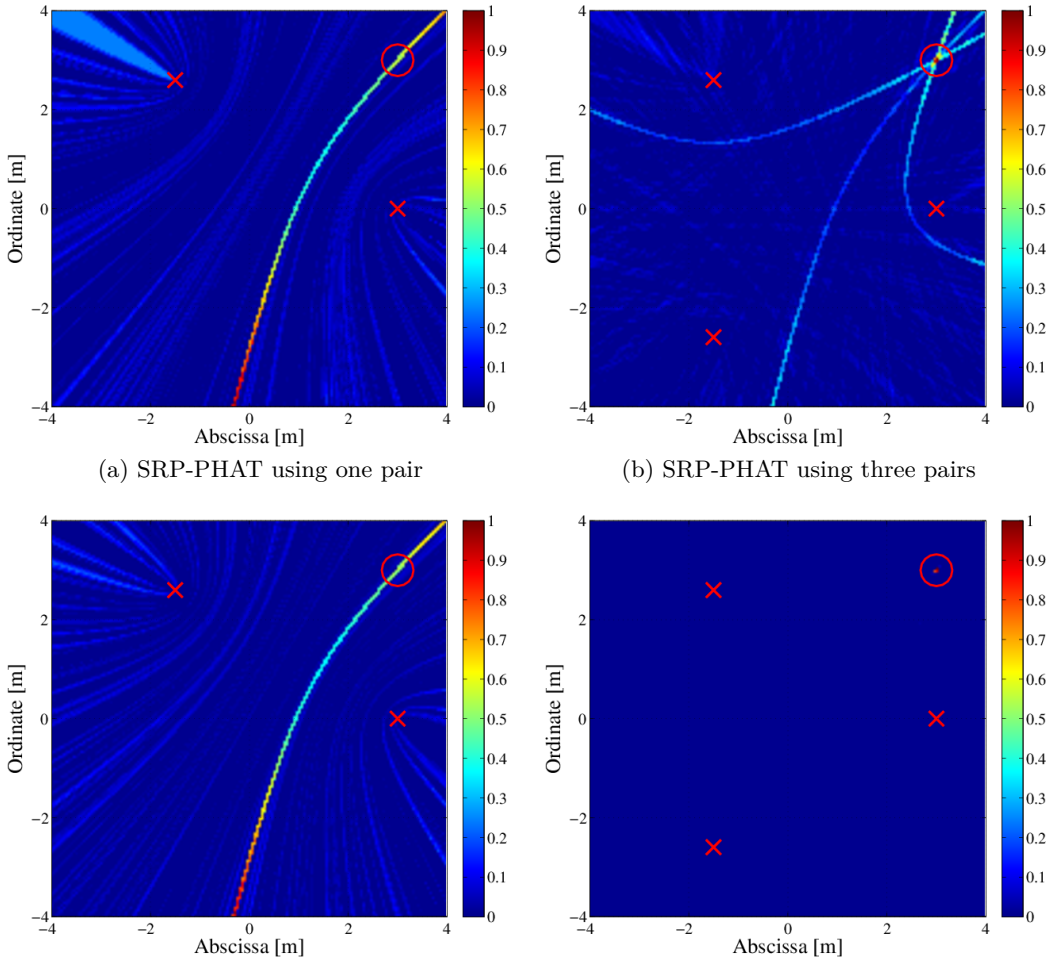
$$\tau_p^{(n)} = \frac{\left|\left|\mathbf{r}^{(n)} - \mathbf{r}^m_{p,1}\right|\right| - \left|\left|\mathbf{r}^{(n)} - \mathbf{r}^m_{p,2}\right|\right|}{c}. \tag{A.29}$$

The key idea of SRP-PHAT is to consider the $p^{th}$ correlation measure $R_p^{phat}(\tau_p^{(n)})$ as a kind of likelihood of the candidate position $\mathbf{r}^{(n)}$. This is based on the fact that if $\mathbf{r}^{(n)} = \mathbf{r}^s$, then $R_p^{phat}(\tau_p^{(n)})$ is high, and if $\mathbf{r}^{(n)} \neq \mathbf{r}^s$, then $R_p^{phat}(\tau_p^{(n)})$ is low. The SRP-PHAT function $\Lambda$ is defined as follows [164]:

$$\Lambda(\mathbf{r}^{(n)}) = \frac{1}{P}\sum_{p=1}^{P} R_p^{phat}(\tau_p^{(n)}). \tag{A.30}$$

$\Lambda(\mathbf{r}^{(n)})$ gives the likelihood of the candidate $\mathbf{r}^{(n)}$ given all the correlation measurements $R_p^{phat}$, $p \in [1, 2, ..., P]$. Computing this quantity for each candidate and normalize all the results between 0 and 1 produces an acoustic map. Two illustrative examples are depicted in Fig.A.2a and Fig.A.2b. For both plots, the actual sound source position is $\mathbf{r}^s = [3, 3]$ (red circle), the sensors are symbolized by red crosses, and the search area is a square of 8x8 m divided in small 5x5 cm square candidates. In Fig.A.2a, a single pair is considered. We retrieve the hyperbola defined by Eq.A.29, with $\mathbf{r}^{(n)}$ the variable. In Fig.A.2b, three pairs are used. The three hyperbola intersect at the actual source position.

In practice, $\tau_{p,\mathbf{x}}$ is approximated by taking the closest integer delay. Many modern versions of the original SRP-PHAT algorithm may be found [165, 166, 163], in which source directivity, microphone directivity, and source-microphone distances are taken into account. Modern approaches consist in replacing the acoustic signals delivered by the microphones by the principal components of the correlation matrix [167]. The objective




(a) SRP-PHAT using one pair

(b) SRP-PHAT using three pairs

(c) MULTI-PHAT using one pair

(d) MULTI-PHAT using three pairs

Figure A.2: Simulated SLF using SRP-PHAT and MULTI-PHAT techniques on one or three pairs of sensors.

of these enhancement is always to counteract the effect of reverberant conditions. This is out of the scope here because no reverberation occurs in the present work. However, one can note that the main drawback of the SRP-PHAT algorithm is that all high values in each pair are present in the resulting SLF, because the sum represents a union of values. For instance, all three hyperbola are well visible in Fig.A.2b, in which regions of high likelihood create so-called ghosts positions. Another approach therefore consists in using the product operator instead of the sum. This is called the MULTI-PHAT technique, and Eq (A.30) is then replaced by [168, 169]:

$$\Lambda(\mathbf{r}^{(n)}) = \frac{1}{P} \prod_{p=1}^{P} R_p^{phat}(\tau_p^{(n)}). \tag{A.31}$$

MULTI-PHAT is applied to the same scenario as above, and results are depicted in

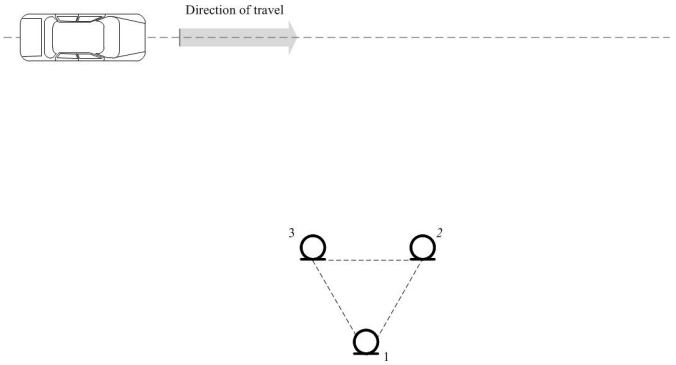
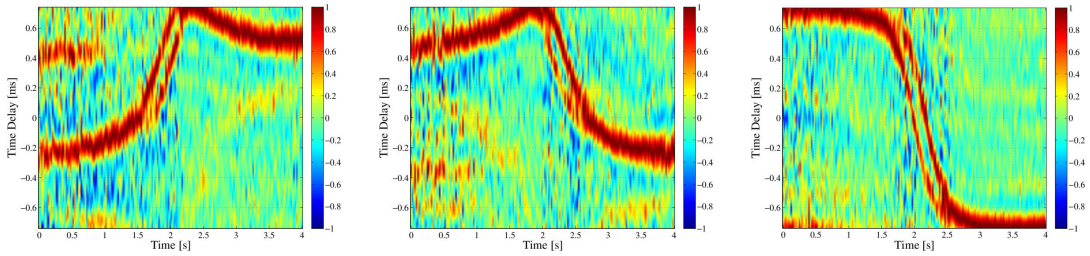



Figure A.3: Microphone array laid out in an equilateral triangle with size $d$ on the roadside and with one pair (2-3 here) parallel to the road lane.




(a) (b) (c)

Figure A.4: PHAT-CCTS using the pair of sensors (a) 1-2, (b) 1-3 and (c) 2-3.

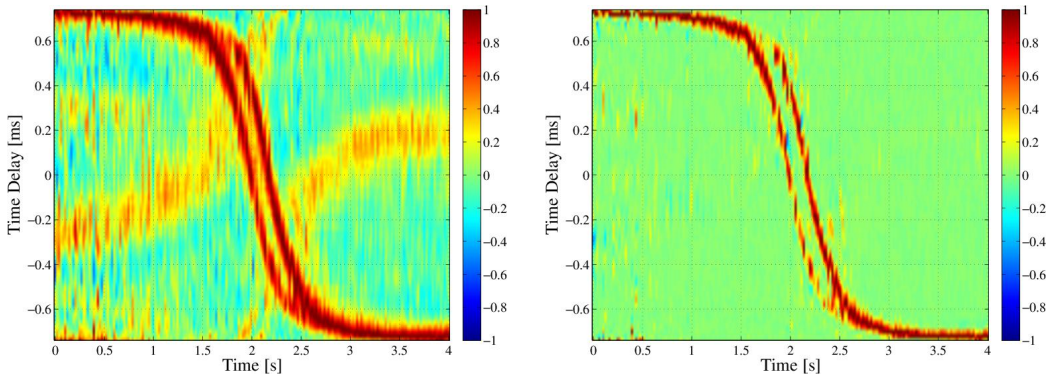
Fig.A.2c and Fig.A.2d. A drastic improvement is achieved in both cases, especially when using the three pairs, Fig.A.2d, where a single and correct mode is visible without any ghost in the remaining of the search area.

### Application to pass-by noise

MULTI-PHAT and SRP-PHAT techniques are now compared on a real pass-by noise measurement. Three microphones laid out in an equilateral triangle were placed on the roadside. The position of each sensor (1, 2 and 3) is schematically shown in Fig.A.3. The pass-by noise coming from an unknown two-axle road vehicle was recorded during 4 seconds. The cross-correlation time series of each pair is depicted in Fig.A.4a (pair 1-2), Fig.A.4b (pair 1-3) and Fig. A.4c (pair 2-3).

Combining the three observations according to the SRP-PHAT and MULTI-PHAT techniques gives the final CCTS depicted in Fig.A.5a and Fig. A.5b respectively.

It clearly appears that the SRP-PHAT-based combination is not appropriate because of the ghosts brought by the pairs 1-2 and 1-3. Considering the three pairs is even worst than considering only the pair 2-3 (parallel to the road lane). However, the MULTI-PHAT-based combination give a much better contrast by considering the three pairs rather than using only the pair 2-3. No ghost appears, and both axles are discriminated




(a) (b)

Figure A.5: Combination of multiple CCTS using the (a) SRP-PHAT procedure and (b) MULTI-PHAT procedure.

more precisely, making the MULTI-PHAT the ad-hoc technique we used by default when multiple pairs are available.

**Other examples of applications**

The LEMA actively works on room acoustics. One objective is to design more versatile rooms, for instance by developing (semi-) active modal control systems [170, 171], and measuring the room characteristics required by the acoustician for describing the sound field as reverberation time, clarity and spatial decay. Usually, the evaluation of rooms is based on the impulse response between one source and one microphone at different measurement points depending of the reflections studied. In this context, the LEMA has developed a 8-element microphone array [47], Fig.A.6a. An active array can return the direction of arrival of each early reflection in a room, which is a useful information for the room acoustician investigating which area of the ceil, floor or wall is responsible of a given reflection [172]. For pedagogic purposes, this tool is sometimes turned into a speaker localization device by applying the MULTI-PHAT technique using all the centered pairs of the array. The returned acoustic map is compared to video images provided by a camera laid out at the center of the array. A typical result is depicted in Fig A.6b.

Another example is Unmanned Aerial Vehicles (UAV). A big problem when operating multiple aircrafts is the increased risk of mid-air collisions. Sensor technology to detect aircrafts in order to prevent collisions currently receives a lot of attention in the research community, due to an increased use of military UAVs and the desire to operate in civilian airspace. In LEMA, we assessed the feasibility of acoustic embedded sensors with the goal to design an autonomous anti-collision system. The developed algorithm suppresses all the harmonics due to the propeller noise and correlates measurements between sensors to locate remaining sound sources around. Fig.A.6c illustrates a prototype of embedded tetrahedral microphone array and Fig.A.6d is a typical result of the localization function delivered by the SRP-PHAT algorithm when another UAV is in front of the array. More

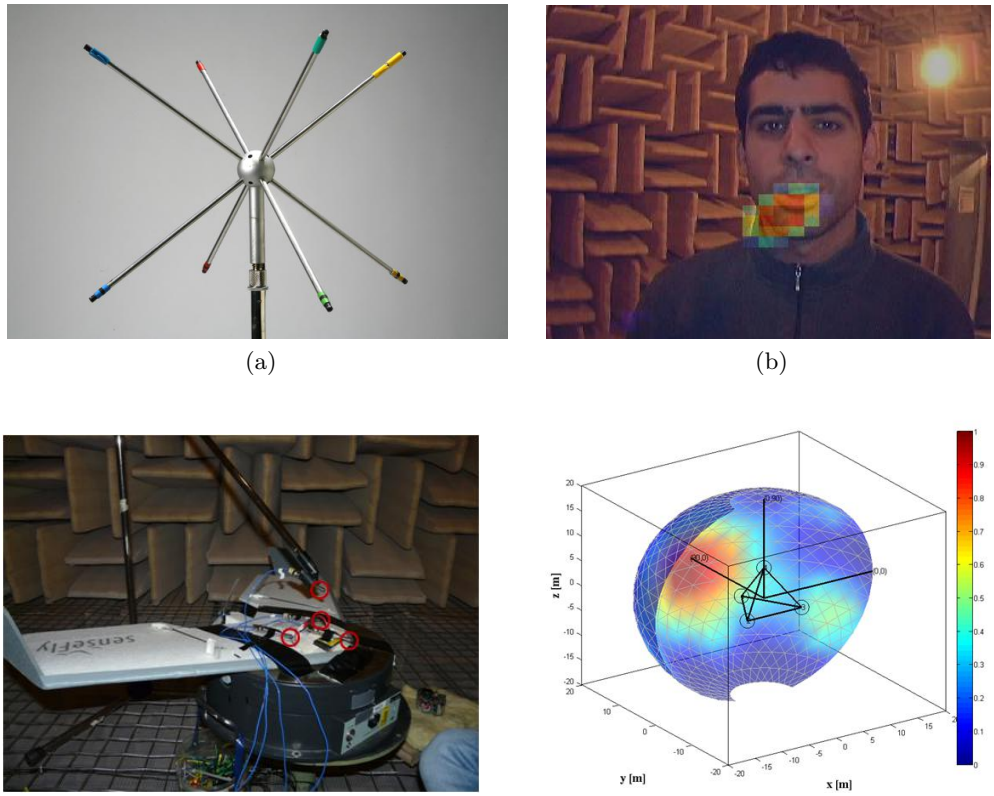



Figure A.6: Speaker localization: (a) 8-elements microphone array (cubic), (b) acoustic map (in azimuth and elevation) compared to the video signal from webcam at the center of the array. UAV localization: (c) protoype of an embedded tetrahedral microphone array, (d): acoustic map delivered by the SRP-PHAT algorithm when another UAV is in front of the array.

details can be found in [173].

## A.5 Audio features

This appendix presents some audio features commonly used in automatic music classification.

**sound pressure level (SPL) (dB)**

This is the logarithmic measure of the effective sound pressure of all surroundings sounds relative to the reference sound pressure level $p_{ref}$ in the air:

$$SPL[q] = 20log10\left(\frac{\sqrt{\sum_{n=1}^{N_s}\mathbf{y}_1^q[Ns-n+1]^2}}{\eta p_{ref}}\right), \tag{A.32}$$

where $\eta$ is the microphone sensitivity (in V/Pa) and $p_{ref} = 20\mu Pa$ (considered as the threshold of human hearing).

**spectral gravity center (SGC) (Hz)**

This is the frequency which splits the power spectral density into two parts of equal energy:

$$SGC[q] = \frac{\sum_{k\geq 0}^{k_s/2} k|\mathbf{Y}_1^q[k]|^2}{\sum_k |\mathbf{Y}_1^q[k]|^2}, \tag{A.33}$$

where $k_s/2$ is the Nyquist frequency bin. Perceptually speaking, the spectral centroid is strongly correlated with the brightness of a sound. The higher the centroid, the brighter the sound is [174].

**spectral roll-off point (SRF) (Hz)**

This is the frequency below which a given percentage $\gamma_{srf}$ of the signal energy is contained:

$$\sum_{k\geq 0}^{SRF[q]} |\mathbf{Y}_1^q[k]|^2 = \gamma_{srf}\sum_{k>0}|\mathbf{Y}_1^q[k]|^2. \tag{A.34}$$

The value of $\gamma_{srf}$ varies with authors: $\gamma_{srf}$ equals 0.95 in [143], 0.93 in [142], 0.92 in [175] or 0.85 in [174]. The SRF is higher for signals with strong energy components at high frequencies, so it is traditionally used to distinguish noisy from harmonic signals [143].

**zero crossing rate (ZCR) [%]**

This is a measure of the number of times the signal crosses the zero axis. It is defined by:

$$ZCR[q] = \frac{1}{N_s - 1} \sum_{n=1}^{N_s-1} |sign(\mathbf{y}_1^q[N_s - n + 1]) - sign(\mathbf{y}_1^q[N_s - n])|, \tag{A.35}$$

where

$$sign\mathbf{y}_1^q[n]) = \begin{cases} 1 & \text{if } \mathbf{y}_1^q[n] \geq 0 \\ -1 & \text{if } \mathbf{y}_1^q[n] < 0 \end{cases}$$

ZCR is traditionally used for distinguishing clean or periodic signals (low ZCR) from more noisy ones (high ZCR). It is highly correlated with the spectral gravity center since it is an indirect measure of the signal spectral content.

**maximum of the auto-correlation (MAC)**

The higher the periodicity in the signal, the higher this feature:

$$MAC[q] = \max\left(IDFT\left(\mathbf{Y}_1^q[k]\mathbf{Y}_1^q[k]^*\right)\right). \tag{A.36}$$

**spectral kurtosis (KRT) and spectral skewness (SKW)**

Spectral kurtosis and spectral skewness are measures of the spectrum shape.

The kurtosis measures the spectrum sharpness. It is equal to 3 if the spectral distribution is Gaussian, less for a flatter and more for a sharper one. It is given by:

$$KRT[q] = \mathbb{E}\left[\left(\frac{|\mathbf{Y}_1^q| - \mu_{\mathbf{Y}_1^q}}{\sigma_{\mathbf{Y}_1^q}}\right)^4\right], \tag{A.37}$$

where $\mu_{\mathbf{Y}_1^q}$ and $\sigma_{\mathbf{Y}_1^q}$ are the mean and standard deviation of $|\mathbf{Y}_1^q|$ respectively.

The skewness measures the symmetry of the spectrum around its mean value. It is positive if the distribution tail spreads to the right, and negative otherwise. The skewness is null for the Gaussian and any other symmetrical distribution. Multiple definitions of skewness exist but we chose the following one:

$$SKW[q] = \mathbb{E}\left[\left(\frac{|\mathbf{Y}_1^q| - \mu_{\mathbf{Y}_1^q}}{\sigma_{\mathbf{Y}_1^q}}\right)^3\right]. \tag{A.38}$$

**spectral bandwidth (SBW)**

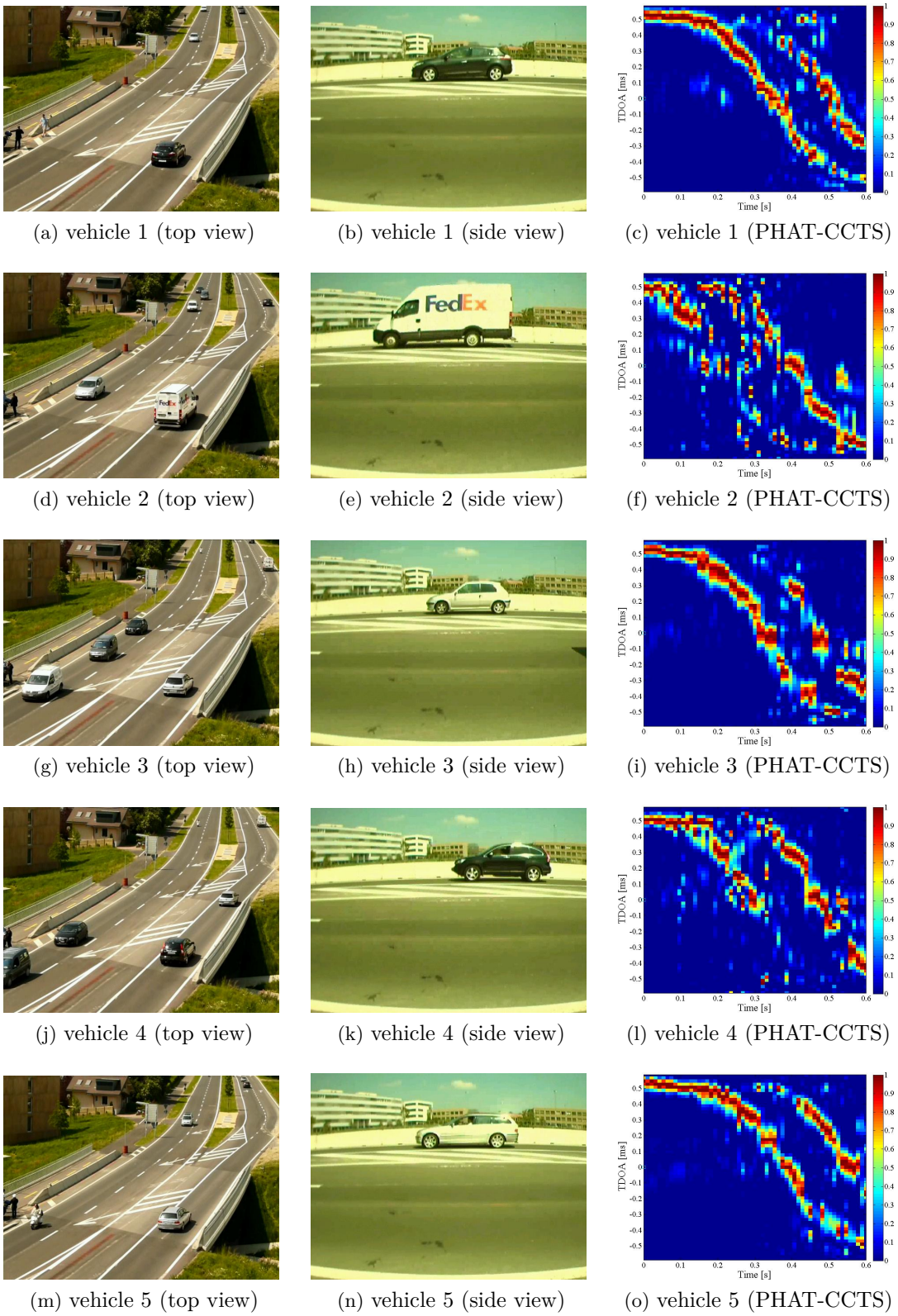
The spectral bandwidth is quite close to a spectral standard deviation except that the reference point is not the spectral mean but the spectral gravity center (SGC) defined

above. It is expressed by:

$$SBW[q] = \sqrt{\frac{\sum_k (k - \sqrt{SGC[q]})^2 \times |\mathbf{Y}_1^q[k]|}{\sum_k k}}. \tag{A.39}$$

**Remark** Other popular features in music classification and not explored here are the wavelets and mel frequency cepstral coefficients (MFCC). MFCC is a very popular feature in speech and music recognition. However, according to [109] and [176], MFCC are relevant for structured sounds, such as speech and music, but their performance degrades in the presence of noise. Moreover, MFCC is not effective for analyzing sounds with a broad flat spectrum as is the case for rain or vehicle noise. Wavelet transform is an interesting tool because it overcomes the classical tradeoff of time vs frequency resolution of the STFT. Wavelet coefficients have been used as features for vehicle detection in [126].

## A.6 EPFL Database




(a) vehicle 1 (top view) (b) vehicle 1 (side view) (c) vehicle 1 (PHAT-CCTS)

(d) vehicle 2 (top view) (e) vehicle 2 (side view) (f) vehicle 2 (PHAT-CCTS)

(g) vehicle 3 (top view) (h) vehicle 3 (side view) (i) vehicle 3 (PHAT-CCTS)

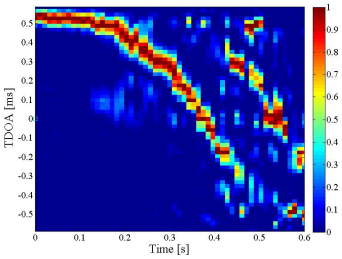
(j) vehicle 4 (top view) (k) vehicle 4 (side view) (l) vehicle 4 (PHAT-CCTS)

(m) vehicle 5 (top view) (n) vehicle 5 (side view) (o) vehicle 5 (PHAT-CCTS)

Figure A.7: Top view, side view and observation (BPHAT-CCTS) of vehicles 1 to 5


(a) vehicle 6 (top view)

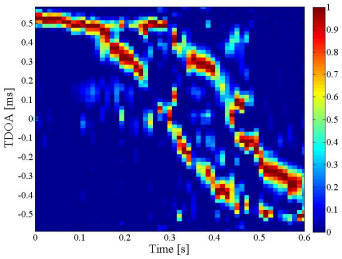

(b) vehicle 6 (side view)



(c) vehicle 6 (PHAT-CCTS)


(d) vehicle 7 (top view)

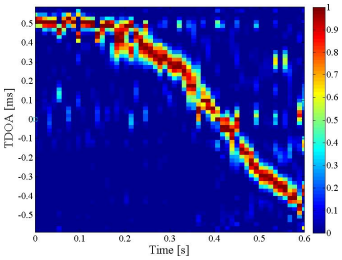

(e) vehicle 7 (side view)



(f) vehicle 7 (PHAT-CCTS)


(g) vehicle 8 (top view)

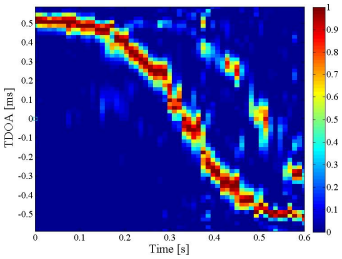

(h) vehicle 8 (side view)



(i) vehicle 8 (PHAT-CCTS)


(j) vehicle 9 (top view)

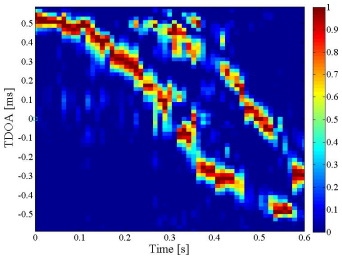

(k) vehicle 9 (side view)



(l) vehicle 9 (PHAT-CCTS)


(m) vehicle 10 (top view)


(n) vehicle 10 (side view)

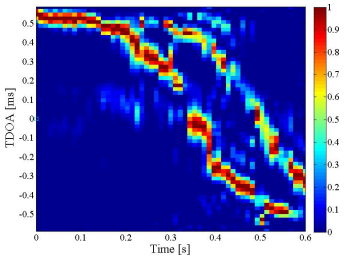


(o) vehicle 10 (PHAT-CCTS)

Figure A.8: Top view, side view and observation (BPHAT-CCTS) of vehicles 6 to 10


(a) vehicle 11 (top view)

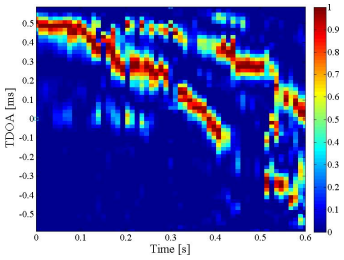

(b) vehicle 11 (side view)



(c) vehicle 11 (PHAT-CCTS)


(d) vehicle 12 (top view)

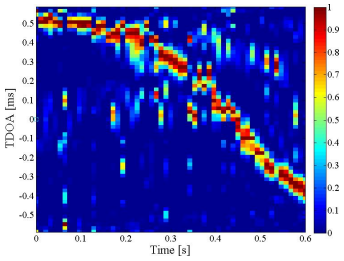

(e) vehicle 12 (side view)



(f) vehicle 12 (PHAT-CCTS)


(g) vehicle 13 (top view)

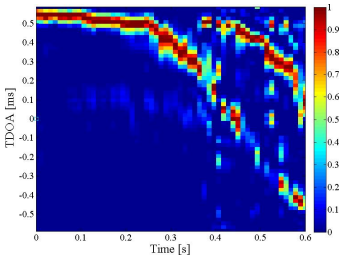

(h) vehicle 13 (side view)



(i) vehicle 13 (PHAT-CCTS)


(j) vehicle 14 (top view)

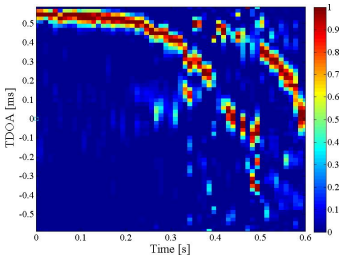

(k) vehicle 14 (side view)



(l) vehicle 14 (PHAT-CCTS)


(m) vehicle 15 (top view)


(n) vehicle 15 (side view)

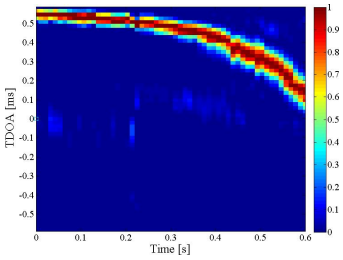


(o) vehicle 15 (PHAT-CCTS)

Figure A.9: Top view, side view and observation (BPHAT-CCTS) of vehicles 11 to 15


(a) vehicle 16 (top view)

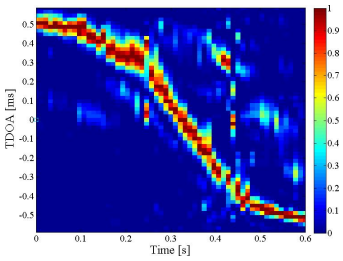

(b) vehicle 16 (side view)



(c) vehicle 16 (PHAT-CCTS)


(d) vehicle 17 (top view)

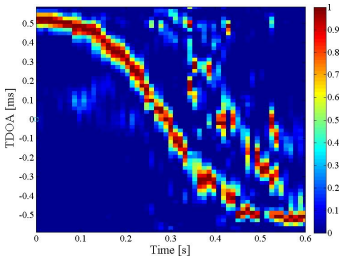

(e) vehicle 17 (side view)



(f) vehicle 17 (PHAT-CCTS)


(g) vehicle 18 (top view)

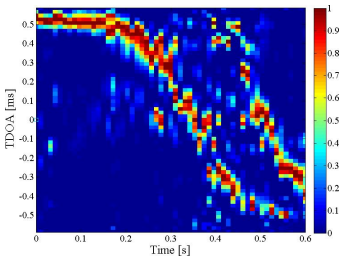

(h) vehicle 18 (side view)



(i) vehicle 18 (PHAT-CCTS)


(j) vehicle 19 (top view)

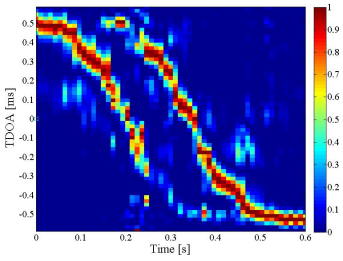

(k) vehicle 19 (side view)



(l) vehicle 19 (PHAT-CCTS)


(m) vehicle 20 (top view)


(n) vehicle 20 (side view)

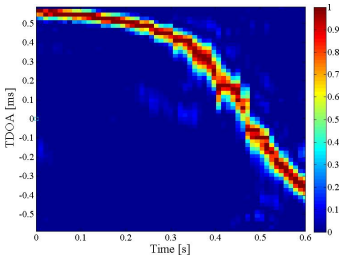


(o) vehicle 20 (PHAT-CCTS)

Figure A.10: Top view, side view and observation (BPHAT-CCTS) of vehicles 16 to 20


(a) vehicle 21 (top view)

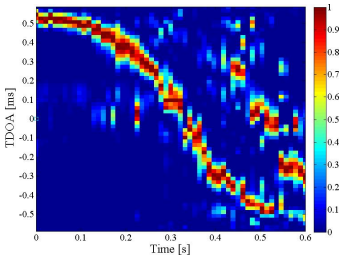

(b) vehicle 21 (side view)



(c) vehicle 21 (PHAT-CCTS)


(d) vehicle 22 (top view)

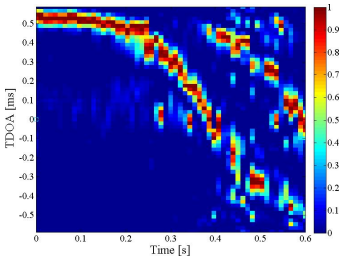

(e) vehicle 22 (side view)



(f) vehicle 22 (PHAT-CCTS)


(g) vehicle 23 (top view)

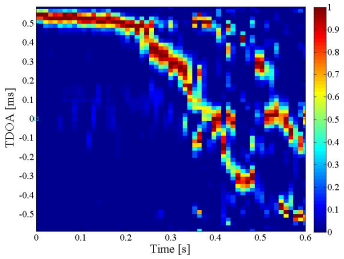

(h) vehicle 23 (side view)



(i) vehicle 23 (PHAT-CCTS)


(j) vehicle 24 (top view)


(k) vehicle 24 (side view)



(l) vehicle 24 (PHAT-CCTS)

Figure A.11: Top view, side view and observation (BPHAT-CCTS) of vehicles 21 to 24

## A.7 Counteracting the wind noise: a state of the art

According to standards, normative measurements of pass-by noise level should not be performed if the wind speed of wind is higher than 5 m/s [135]. In case of long term monitoring (several days), one has to take precautions against wind. A short state of the art on the different ways to reduce the influence of wind in measurements is proposed in this appendix.

The noise measured by a microphone within an airflow is caused by two distinct phenomena: the pressure fluctuations induced at the microphone diaphragm due to the turbulence in the flow (determined by the atmospheric conditions and terrain properties), and those induced by turbulent wake of the microphone (determined by the microphone shape and local wind speed) [177]. This causes the microphone to measure a *pseudo-noise* which is not due to an incoming acoustic wave. The pseudo-noise affects sound pressure level measurements and should ideally be at least 10 dB below sound level being measured [178]. The purpose of a windscreen is to reduce the effect of the pseudo-noise while allowing the acoustic signal to propagate to the microphone diaphragm with minimal attenuation. In the same range of ideas, the purpose of a denoising algorithm is to suppress the frequency components of the pseudo-noise. Both approaches are discussed.

### A.7.1 Types of windscreens

Various microphone coverings can be used to reduce pseudo-noise. The four most frequent types of windscreen are listed and illustrated below:

- basket-style: Fig.A.12a;
- solid foam: Fig.A.12b;
- hollow foam: Fig.A.12c;
- nose cone: Fig.A.12d.

(a) basket-style (b) solid foam (c) hollow foam (d) nose cone

Figure A.12: The four most common types of windscreen: (a) basket-style windscreen, (b) foam windscreen, (c) hollow windscreen and (d) nose cone windscreen.

The choice of a protection depends on the type of microphone used: pressure-gradient microphones or pressure microphones [179]. The basket-style windscreen and the hollow foam windscreen are adapted for pressure-gradient microphones only. Because they are

less sensitive to wind noise, pressure microphones should be preferred to pressure-gradient microphones in the RTM context.

Pressure microphones may be equipped of nose-cone or solid-foam windscreens. With solid-foam windscreens, the sound field is not distorted except at high frequencies. Moreover these windscreens are light and can be aerodynamically shaped. On the other hand, nose-cones are designed to reduce the aerodynamical noise present when the microphone is exposed to high wind speeds in a known direction. A highly polished surface gives the least possible resistance to air flow and thereby reduces the noise produced by the microphone itself. In practice, the wind direction is varying so that a solid foam windscreen should be the retained solution instead.

According to [180] and [179], the larger the windscreen, the more effective it will be. But, not surprisingly, highly effective windscreens are found to have the worst effect on the sound. Some balance has to be found regarding the sonic deterioration versus the minimization of the wind-induced noise. In [178], different types of microphone, associated to different windscreens, are compared within a wind of 28 m/s. The best result (lowest pseudo-noise measurement) was obtained by associating a 1/2" microphone equipped with a sharp nose cone windscreen. The signal is distorted above 4 kHz with a solid foam windscreen.

### A.7.2 Microphones

Experimental studies [179], [178] and [177] show that the pseudo-noise level is quite similar for 1" and 1/2" microphone diameters in case of low wind speed (6-10 m/s), but is more prevalent for small diameter microphones in case of high speed (>28 m/s). The greater problems for small diameter microphones result from air turbulence causing a higher instantaneous total pressure on the surface of a small area microphone than of on a larger area microphone [181]. The maximum interference is obtained when the microphone is oriented towards the wind source [179]. To limit such an effect, the membranes should face the road in practice.

In [182], theoretical and experimental studies showed that the turbulent noise signal can be reduced considerably by means of a probe microphone, *i.e.* a microphone placed at the end of a cylindrical tube with an axial slit and covered with cloth. A probe microphone is a kind of microphone especially designed for difficult measurement situations in harsh environments, *e.g.* to measure dynamic pressure in high-temperature airflows at the exhaust of a turbine. Usually it is of very small size, low weight and its right-angle design makes the probe microphone particularly well suited for such measurements. This solution has not been tested during this thesis but one should check if such microphones have no effect on the time delay measurements.

### A.7.3 Signal processing to attenuate wind noise

Because of its non-stationary nature, the wind noise cannot be handled by conventional noise reduction algorithms, such as spectral subtraction or statistics-based estimators like in [183], [184], [185] and [186]. But, as the definition of stationarity is relative to the observation duration, these simple methods may be sufficient when processing short signal frames. If the wind noise fluctuates a lot, non-stationary spectral subtraction methods should be considered, such as the so-called *noise tracking technique* [187].

As is mentioned in [188], many methods for separating non-stationary broad band signals are based on a priori source modeling (using Gaussian mixture models [189], vector quantization [190], Linear Predictive Coding (LPC) analysis [191] or non-negative sparse coding [192]). This is a good approach when the processing focuses only on a priori sounds (vehicles pass-by noise). According to [193], the Non-Negative Matrix factorization algorithm of [188, 192] provides the best wind noise reduction, but the computational complexity of this method is high and must be discarded for embedded processing purposes.

Recently, Franz and Bitzer proposed in [194] an algorithm for wind reduction dedicated to hearing aids, through the combination of a single-channel low-frequency reduction algorithm and a correlation detector between two channels. This procedure works in real time but is based on the strong assumption that the pseudo noises at each ear is highly uncorrelated. This is approximately true because of the presence of the head, but this assumption is not so evident for “empty” microphone arrays. No investigation were assessed on this point for our side.

# Bibliography


[1] O. Hänninen and A. Knol, "European perspectives on environmental burden of disease - estimates for nine stressors in six european countries," tech. rep., National Institute for Health and Welfare, 2011.

[2] G. Bolte, *Environmental health inequalities in Europe*, ch. Chapter 4. Environment-related inequalities, pp. 86–113. World Health Organization, 2012.

[3] W. Babisch, "Transportation noise and cardiovascular risk - review and synthesis of epidemiological studies - dose-effect curve and risk estimation," tech. rep., Federal Environmental Agency, 2006.

[4] "Ordonnance de la protection contre le bruit (opb), etat le $1^{er}$ août 2010," tech. rep., Le Conseil fédéral suisse, 1986.

[5] B. Miranda, O. Jacquat, and D. Zürcher, "Plan directeur de recherche environnement pour les années 2013-2016, axes, domaines et thèmes de recherche prioritaires.," tech. rep., Officice Fédéral de l'Environnement (OFEV), Berne. Connaissance de l'environnement n° 1206, 2012.

[6] "Practitioner handbook for local noise action plans - recommendations from the silence project," tech. rep., Sixth Framework Programme of the European Commission, 2008.

[7] E. Minge, J. Kotzenmacher, and S. Peterson, "Evaluation of non-intrusive technologies for traffic detection," tech. rep., Minnesota Department of Tranportation - Research Services - Office of Policy Analyis, Research and Innovation, sep 2010.

[8] S. L. Skszek, ""state-of-the-art" report on non-traditional traffic counting methods," tech. rep., Arizona Department of Transportation, oct. 2001.

[9] L. Klein, M. Mills, and D. Gibson, "Traffic detector handbook: Third edition - volume i," tech. rep., Federal Highway Administration, oct. 2006.

[10] L. E. Y. Mimbela, L. A. Klein, P. Kent, J. L. Hamrick, K. M. Luces, and S. Herrera, "A summary of vehicle detection and surveillance technologies used in intelligent

transportation systems," tech. rep., Funded by the Federal Highway Administration's Intelligent Transportation Systems Joint Program Office, produced by The Vehicle Detector Clearinghouse, 2000.

[11] M. Hallenbeck and H. Weinblatt, "Equipment for collecting traffic load data," tech. rep., Transportation Research Board of the National Academies, 2004.

[12] S. Chen, Z. Sun, and B. Bridge, "Automatic traffic monitoring by intelligent sound detection," in *Proceedings of the IEEE Conference on Intelligent Transportation System (ITSC)*, pp. 171–176, dec. 1997.

[13] J. Perez-Lorenzo, R. Viciana-Abad, P. Reche-Lopez, F. Rivas, and J. Escolano, "Evaluation of generalized cross-correlation methods for direction of arrival estimation using two microphones in real environments," *Applied Acoustics*, vol. 73, no. 8, pp. 698–712, 2012.

[14] P. J. Yauch and et al., *Traffic signal control equipment: state of the art.* Washington, D:C: Transportation Research Borad, National Research Council, 1990.

[15] H. C. Choe, R. E. Karlsen, G. R. Gerhart, and T. J. Meitzler, "Wavelet-based ground vehicle recognition using acoustic signals," *Proceedings of SPIE*, vol. 2762, pp. 434–445, 1996.

[16] J. F. Forren and D. Jaarsma, "Traffic monitoring by tire noise," in *Proceedings of the IEEE Conference on Intelligent Transportation Systems (ITSC)*, pp. 177–182, 1997.

[17] E. Brockmann, B. Kwan, and L. Tung, "Audio detection of moving vehicles," in *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics.*, vol. 4, pp. 3817–3821, oct. 1997.

[18] J. P. Kuhn, B. C. Bui, and G. J. Pieper, "Acoustic sensor system for vehicle detection and multi-lane highway monitoring," aug 1998.

[19] S. Chen, Z. Sun, and B. Bridge, "Traffic monitoring using digital sound field mapping," *IEEE Transactions on Vehicular Technology*, vol. 50, pp. 1582–1589, nov. 2001.

[20] K. Kodera, A. Itai, and H. Yasukawa, "Sound localization of approaching vehicles using uniform microphone array," in *Proceedings of IEEE Conference on Intelligent Transportation Systems Conference (ITSC)*, pp. 1054–1058, oct. 2007.

[21] K. Kodera, A. Itai, and H. Yasukawa, "Approaching vehicle detection using linear microphone array," in *Proceedings of International Symposium on Information Theory and Its Applications (ISITA)*, pp. 1–6, dec. 2008.

[22] C. Kwak, M. Kim, K. Kim, S. Hong, and K. Kim, "Robust in-situ vehicle detection algorithm with acoustic transition bandpass filter," feb. 2009.

[23] N. Shimada, A. Itai, and H. Yasukawa, "A study on using linear microphone array-based acoustic sensing to detect approaching vehicles," in *Proceedings of International Symposium on Communications and Information Technologies (ISCIT 2010)*, pp. 182–186, oct. 2010.

[24] B. Barbagli, I. Magrini, G. Manes, A. Manes, G. Langer, and M. Bacchi, "A distributed sensor network for real-time acoustic traffic monitoring and early queue detection," in *Proceedings of the Fourth International Conference on Sensor Technologies and Applications (SENSORCOMM)*, pp. 173–178, jul. 2010.

[25] B. Barbagli, L. Bencini, I. Magrini, G. Manes, and A. Manes, "A real-time traffic monitoring based on wireless sensor network technologies," *Proceedings of the $7^{th}$ International Wireless Communications and Mobile Computing Conference (IWCMC)*, pp. 820–825, jul. 2011.

[26] V. Tyagi, S. Kalyanaraman, and R. Krishnapuram, "Vehicular traffic density state estimation based on cumulative road acoustics," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, pp. 1156–1166, sep. 2012.

[27] J. C. Hassab, B. W. Guimond, and S. C. Nardone, "Estimation of location and motion parameters of a moving source observed from a linear array," *The Journal of Acoustical Society of America*, vol. 70, no. 4, pp. 1054–1061, 1981.

[28] J. Towers and Y. Chan, "Passive localization of an emitting source by parametric means," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, pp. 2791–2794, apr. 1990.

[29] C. Couvreur and Y. Bresler, "Doppler-based motion estimation for wide-band sources from single passive sensor measurements," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, pp. 3537–3540, apr. 1997.

[30] F. Pérez-González, R. López-Valcarce, and C. Mosquera, "Road vehicle speed estimation from a two-microphone array," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, apr. 2002.

[31] R. López-Valcarce, "Broadband analysis of a microphone array based road traffic speed estimator," in *Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2004*, pp. 533–537, jul. 2004.

[32] R. López-Valcarce, C. Mosquera, and F. Pérez-González, "Estimation of road vehicles speed using two omnidirectional microphones: a maximum likelihood approach," *EURASIP Journal on Applied Signal Processing*, vol. 8, pp. 1059–1077, 2004.

[33] O. Duffner, N. O'Connor, N. Murphy, A. Smeanton, and S. Marlow, "Road traffic monitoring using a two-microphone array," in *Audio Engineering Society, Convention 118*, p. 6355, may 2005.

[34] V. Cevher, R. Chellappa, A. Gurbuz, F. Shah, and J. McClellan, "Vehicle fingerprinting using drive-by-sounds," tech. rep., Maryland University, College Park. Center for Automation Research, nov. 2006.

[35] V. Cevher, F. Guo, A. Sankaranarayanan, and R. Chellappa, "Joint acoustic-video fingerprinting of vehicles, part 2," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 749–752, apr. 2007.

[36] V. Cevher, R. Chellappa, and J. McClellan, "Vehicle speed estimation using acoustic wave patterns," *IEEE Transactions on Signal Processing*, vol. 57, pp. 30–47, jan. 2009.

[37] A. Can, L. Dekoninck, M. Rademaker, T. V. Renterghem, B. D. Baets, and D. Botteldooren, "Noise measurements as proxies for traffic parameters in monitoring networks," *Science of The Total Environment*, vol. 410-411, pp. 198–204, 2011.

[38] F. Samaran, O. Adam, J.-F. Motsch, Y. Cansi, G. Ruzié, and C. Guinet, "Acoustic localization of two distinct blue whale (balaenoptera musculus) subspecies in the south-west indian ocean," *Journal of the Acoustical Society of America*, vol. 123, pp. 3774–3774, may 2008.

[39] S. Shivappa, M. Trivedi, and B. Rao, "Audiovisual information fusion in human-computer interfaces and intelligent environment: a survey," *Proceedings of the IEEE*, vol. 98, pp. 1692–1715, oct. 2010.

[40] A. Brutti, M. Omologo, and P. Svaizer, "Localization of multiple speakers based on a two step acoustic map analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4349–4352, apr. 2008.

[41] D. Welker, J. Greenberg, J. Desloge, and P. Zurek, "Microphone-array hearing aids with binaural output - part 2. a two-microphone adaptive system," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 543–551, nov. 1997.

[42] R. O. Nielsen, *Sonar Signal Processing.* Artech House, 1991.

[43] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, pp. 1408–1418, aug. 1969.

[44] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, pp. 276–280, mar. 1986.

[45] G. Bienvenu and L. Kopp, "Optimality of high resolution array processing using the eigensystem approach," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 31, pp. 1235–1248, oct. 1983.

[46] R. Roy, A. Paulraj, and T. Kailath, "Esprit - a subspace rotation approach to estimation of parameters of cisoids in noise," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, pp. 1340–1342, oct. 1986.

[47] E. Van Lancker, *Acoustic goniometry : a spatio-temporal approach.* PhD thesis, Ecole polytechnique fédérale de Lausanne, 2001.

[48] Y. H. Hu and D. Li, "Energy based collaborative source localization using acoustic microsensor array," in *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, pp. 371–375, dec 2002.

[49] D. Li and Y. H. Hu, "Least square solutions of energy based acoustic source localization problems," in *Proceedings of the International Conference on Parallel Processing Workshops (ICPP)*, pp. 443 – 446, aug. 2004.

[50] K. Ho and M. Sun, "An accurate algebraic closed-form solution for energy-based source localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2542–2550, nov. 2007.

[51] H. L. van Trees, *Detection, Estimation, and Modulation Theory, Part I.* Wiley-Interscience, 2001.

[52] J. Stuller, "Maximum-likelihood estimation of time-varying delay - part 2," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, pp. 300–313, mar. 1987.

[53] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, pp. 320–327, aug 1976.

[54] J. Chen, Y. Huang, and J. Benesty, "Time delay estimation," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems* (Y. Huang and J. Benesty, eds.), pp. 197–227, Springer US, 2004.

[55] C. Zhang, D. Florencio, and Z. Zhang, "Why does phat work well in lownoise, reverberative environments?," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2565–2568, mar. 2008.

[56] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 288–292, 1997.

[57] T. Gustafsson, B. Rao, and M. Trivedi, "Source localization in reverberant environments: modeling and statistical analysis," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 791–803, nov. 2003.

[58] A. Löytynoja and P. Pertilä, "A real-time talker localization implementation using multi-phat and particle filter.," in *Proceedings of the 17$^{th}$ European Signal Processing Conference (EUSIPCO)*, (Glasgow, Scotland, UK), pp. 1418–1422, 2009.

[59] B. Fazenda, H. Atmoko, F. Gu, L. Guan, and A. Ball, "Acoustic based safety emergency vehicle detection for intelligent transport systems," in *Proceedings of the joint Conference ICCAS-SICE*, pp. 4250–4255, 2009.

[60] E. J. Hannan and P. J. Thomson, "The estimation of coherence and group delay," *Biometrika*, vol. 58, pp. 469–481, dec. 1971.

[61] P. R. Roth, "Effective measurements using digital signal analysis," *IEEE Spectrum*, vol. 8, pp. 62–70, apr. 1971.

[62] G. Carter, A. Nuttall, and P. Cable, "The smoothed coherence transform," *Proceedings of the IEEE*, vol. 61, pp. 1497–1498, oct. 1973.

[63] H. R. Madala and A. G. Ivakhnenko, *Inductive Learning Algorithms for Complex System Modeling (Chapter 2)*. CRC Press, 1994.

[64] A. Hero and S. Schwartz, "A new generalized cross correlator," *IEEE Transactions on Acoustics, Speech and Signal Processing.*, vol. 33, pp. 38–45, feb. 1985.

[65] C. Eckart, "Optimal rectifier systems for the detection of steady signals," tech. rep., UC San Diego: Scripps Institution of Oceanography, 1952.

[66] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 273–276, 1994.

[67] J. Hassab and R. Boucher, "Optimum estimation of time delay by a generalized correlator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, pp. 373–380, aug 1979.

[68] F. Reed, P. Feintuch, and N. Bershad, "Time delay estimation using the lms adaptive filter - static behavior," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, pp. 561–571, jun. 1981.

[69] P. Marmaroli, X. Falourd, and H. Lissek, "A comparative study of time delay estimation techniques for road vehicle tracking," in *11$^{ème}$ Congrés Français d'Acoustique (CFA) - 2012 Institute of Acoustics (IOA) Annual Meeting*, 2012.

[70] K. Sasaki, T. Sato, and Y. Nakamura, "Holographic passive sonar," *IEEE Transactions on Sonics and Ultrasonics*, vol. 24, pp. 193–200, may 1977.

[71] S. Björklund, *A survey and comparison of time-delay estimation methods in linear systems*. PhD thesis, Linköpings Universitet, 2003.

[72] B. Quinn, "Doppler speed and range estimation using frequency and amplitude estimates," *The Journal of Acoustical Society of America*, vol. 98, pp. 2560–2566, nov. 1995.

[73] B. Ferguson, "A ground-based narrow-band passive acoustic technique for estimating the altitude and speed of a propeller-driven aircraft," *The Journal of Acoustical Society of America*, vol. 3, pp. 1403–1407, sep 1992.

[74] B. Ferguson and B. Quinn, "Application of the short-time fourier transform and the wigner-ville distribution to the acoustic localization of aircraft," *The Journal of Acoustical Society of America*, vol. 2, pp. 821–827, 1994.

[75] M. G. i Francitorra, *Sound source detection and noise measurement methods for aircraft noise monitoring in presence of background noise*. PhD thesis, Universitat Politècnica de Catalunya, 2008.

[76] R. Sen, P. Siriah, and B. Raman, "Roadsoundsense: acoustic sensing based road congestion monitoring in developing regions," in *Proceedings of the $8^{th}$ Annual IEEE Communication Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks*, 2011.

[77] V. Cevher, R. Chellappa, and J. McClellan, "Joint acoustic-video fingerprinting of vehicles, part 1," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, pp. 745–748, apr. 2007.

[78] J. L. Spiesberger, "Linking auto- and cross-correlation functions with correlation equations: Application to estimating the relative travel times and amplitudes of multipath," *Journal of the Acoustical Society of America*, vol. 104, no. 1, pp. 300–312, 1998.

[79] E. A. Lehmann, D. B. Ward, and WIlliamson, "Experimental comparison of particle filtering algorithms for acoustic source localization in a reverberant room," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 177–180, 2003.

[80] Z. Liang, X. Ma, and X. Dai, "Robust tracking of moving sound source using scaled unscented particle filter," *Applied Acoustics*, vol. 69, no. 8, pp. 673–680, 2008.

[81] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, pp. 174–188, 2002.

[82] N. Gordon, D. Salmond, and A. Smith, "Novel approach to nonlinear/non-gaussian bayesian state estimation," *Proceedings of IEEE Radar and Signal Processing*, vol. 140, pp. 107–113, apr. 1993.

[83] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME - Journal of Basic Engineering*, pp. 33–45, 1960.

[84] A. Farina, "Target tracking with bearings only measurements," *Signal Processing*, vol. 78, pp. 61–78, oct. 1999.

[85] X. Lin, T. Kirubarajan, Y. Bar-Shalom, and S. Maskell, "Comparison of ekf, pseudomeasurement and particle filters for a bearing-only target tracking problem," in *Proceedings of SPIE*, 2002.

[86] C. Kreucher and B. Shapo, "Multitarget detection and tracking using multisensor passive acoustic data," *IEEE Journal of Oceanic Engineering*, vol. 36, pp. 205–218, apr. 2011.

[87] B. Saulson and K. Chang, "Comparison of nonlinear estimation for ballistic missile tracking," in *Proceedings of SPIE*, vol. 5096, pp. 13–24, 2003.

[88] E. Chatzi and A. Smyth, "The unscented kalman filter and particle filter methods for nonlinear structural system identification with non-collocated heterogeneous sensing," *Struct. Control Health Monit.*, vol. 16, pp. 99–123, 2009.

[89] S. J. Julier and J. K. Uhlmann, "A new extension of the kalman filter to nonlinear systems," in *Proceedings of the SPIE*, vol. 3068, pp. 182–193, 1997.

[90] M. Wu and S. A.W., "Application of the unscented kalman filter for real-time nonlinear structural system identification," *Structural Control and Health Monitoring*, vol. 14, pp. 971–990, 2007.

[91] R. van der Merwe, A. Doucet, N. de Freitas, and E. Wan, "The unscented particle filter," tech. rep., Cambridge University, 2000.

[92] a. Kong, J. S. Liu, and W. H. Wong, "Sequential imputations and bayesian missing data problems," *Journal of the American Statistical Association*, vol. 89, pp. 278–288, mar. 1994.

[93] A. Doucet, S. Godsill, and C. Andrieu, "On sequential monte carlo sampling methods for bayesian filtering," *Statistics and Computing*, vol. 10, pp. 197–208, 2000.

[94] R. Douc, O. Cappé, and E. Mou, "Comparison of resampling schemes for particle filtering," in *Proceedings of the $4^{th}$ International Symposium on Image and Signal Processing and Analysis (ISPA)*, 2005.

[95] P. Djuric, J. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. Bugallo, and J. Miguez, "Particle filtering," *IEEE Signal Processing Magazine*, vol. 20, pp. 19–38, sep 2003.

[96] J. Candy, "Bootstrap particle filtering," *IEEE Signal Processing Magazine*, vol. 24, pp. 73–85, jul. 2007.

[97] A. Doucet and A. M. Johansen, "A tutorial on particle filtering and smoothing: fifteen years later," 2008.

[98] F. Legland, "Filtrage particulaire," in *Proceedings of 19$^{ème}$ colloque GRETSI sur le traitement du signal et des images*, vol. I, pp. 1–8, 2003.

[99] C. Hue, *Méthodes séquentielles de Monte-Carlo pour le filtrage non linéaire multi-objets dans un environnement bruité. Applications au pistage multi-cibles et à la trajectographie d'entités dans des séquences d'images 2D.* PhD thesis, Université de Rennes I, 2003.

[100] A. N. Ndjeng, *Localisation robuste multi-capteurs et multi-modèles.* PhD thesis, Université d'Evry Val d'Essonne, 2009.

[101] B. Balakumar, A. Sinha, T. Kirubarajan, and J. Reilly, "Phd filtering for tracking an unknown number of sources using an array of sensors," in *IEEE/SP 13$^{th}$ Workshop on Statistical Signal Processing, 2005*, 2005.

[102] W. Ng, J. Li, S. Godsill, and J. Vermaak, "Tracking variable number of targets using sequential monte carlo methods," in *Proceedings of the IEEE Statistical Signal Processing Workshop*, pp. 1286–1291, 2005.

[103] W. Ng, J. Li, S. Godsill, and J. Vermaak, "A hybrid approach for online joint detection and tracking for multiple targets," in *Proceedings of the IEEE Aerospace Conference*, pp. 2126–2141, mar. 2005.

[104] M. E. Hohil, J. R. Heberley, J. Chang, and A. Rotolo, "Vehicle counting and classification algorithms for unattended acoustic sensors," in *Proceedings of the SPIE*, vol. 5090, pp. 99–110, 2003.

[105] S. le Cessie and J. van Houwelingen, "Ridge estimators in logistic regression," *Applied Statistics*, vol. 41, no. 1, pp. 191–201, 1992.

[106] S. L. Salzberg, "C4.5: Programs for machine learning by j. ross quinlan," *Machine Learning*, vol. 16, pp. 235–240, 1994. 10.1007/BF00993309.

[107] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Min. Knowl. Discov.*, vol. 2, pp. 121–167, jun. 1998.

[108] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, jun. 2006.

[109] J. Lee and A. Rakotonirainy, "Acoustic hazard detection for pedestrians with obscured hearing," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, pp. 1640–1649, dec. 2011.

[110] J. Lelong, "Vehicle noise emission: evaluation of tyre-road and motor noise contributions," in *Proceedings of the 1999 International Congress on Noise Control Engineering (Internoise)*, 1999.

[111] U. Sandberg, "Tyre/road noise - myths and realities," in *Proceedings of the 2011 International Congress and Exhibition on Noise Control Engineering*, 2001.

[112] F. M. Dommermuth, "A simple procedure for tracking fast maneuvering aircraft using spatially distributed acoustic sensors," *The Journal of Acoustical Society of America*, vol. 82, no. 4, pp. 1418–1424, 1987.

[113] J.-F. Hamet, "Les mécanismes de génération du bruit de roulement et l'influence des caractéristiques de chaussée," *Acoustique & Techniques de l'Ingénieur*, vol. 32, pp. 2–10, 2003.

[114] G. Priolet, F. Anfosso-Lédée, Y. Pichaud, L. Ségaud, L. Toussaint, R. Durang, and P. Dunez, "Mesure en continu du bruit de contact pneumatique - chaussée. méthode d'essai des lpc n°63, version 2.0," tech. rep., France, Ministère de l'Ecologie, de l'Energie, du Développement Durable et de l'Aménagement du Territoire. Laboratoire Central des Ponts et Chaussées, 2008. Coll. Techniques et Méthodes des Laboratoires des Ponts et Chausées, 63.

[115] R. Chellappa, G. Qian, and Q. Zheng, "Vehicle detection and tracking using acoustic and video sensors," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, pp. 793–796, 2004.

[116] J. H. DiBiase, *A high-accuracy, low latency technique for talker localization in reverberant environments using microphone arrays*. PhD thesis, Brown University, may 2000.

[117] Y. Gao, M. Brennan, P. Joseph, J. Muggleton, and O. Hunaidi, "A model of the correlation function of leak noise in buried plastic pipes," *Journal of Sound and Vibration*, vol. 277, Issues 1-2, pp. 133–148, oct. 2004.

[118] Y. Gao, M. Brennan, and P. Joseph, "A comparison of time delay estimators for the detection of leak noise signals in plastic water distribution pipes," *Journal of Sound and Vibration*, vol. 292, Issues 3-5, pp. 552–570, 2006.

[119] J. Lichtenauer, M. Reinders, and E. Hendriks, "Influence of the observation likelihood function on particle filtering performance in tracking applications," in *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004*, pp. 767–772, may 2004.

[120] J. T. Abbott and T. L. Griffiths, "Exploring the influence of particle filter parameters on order effects in causal learning," in *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, 2011.

[121] V. Cevher and J. McClellan, "Fast initialization of particle filters using a modified metropolis-hastings algorithm: mode-hungry approach," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 129–132, may 2004.

[122] M. Bolic, S. Hong, and P. M. Djuric, "Performance and complexity analysis of adaptive particle filtering for tracking applications," in *Proceedings of the 36th Asilomar Conference on Signals Systems and Computers*, vol. 1, pp. 853–857, IEEE, 2002.

[123] F. Gustafsson, "Particle filter theory and practice with positioning applications," *IEEE Magazine in Aerospace and Electronic Systems*, vol. 25, pp. 53–82, jul. 2010.

[124] A. Burguera, Y. González, and G. Oliver, *Advances in sonar technology*, ch. Mobile robot localization using particle filters and sonar sensors, pp. 213–232. Sergio Rui Silva, 2009.

[125] J. Ding, S.-Y. Cheung, C.-W. Tan, and P. Varaiya, "Signal processing of sensor node data for vehicle detection," in *Proceedings of the 7th International IEEE Conference on Intelligent Transportation Systems, 2004.*, pp. 70–75, oct. 2004.

[126] A. Averbuch, V. A. Zheludev, N. Rabin, and A. Schclar, "Wavelet-based acoustic detection of moving vehicles," in *Multidimensional Systems and Signal Processing*, vol. 20, pp. 55–80, Springer Netherlands, may 2008. 10.1007/s11045-008-0058-z.

[127] B. Anami and V. Pagi, "An acoustic signature based neural network model for type recognition of two-wheelers," in *Procedings of the International Multimedia, Signal Processing and Communication Technologies Conference (IMPACT)*, pp. 28–31, mar. 2009.

[128] A. Starzacher and B. Rinner, "Single sensor acoustic feature extraction for embedded realtime vehicle classification," in *Parallel and Distributed Computing, Applications and Technologies, 2009 International Conference on*, pp. 378–383, dec. 2009.

[129] G. Padmavathi, D. Shanmugapriya, and M. Kalaivani, "Acoustic signal based feature extraction for vehicular classification," in *Proceedings of the 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE)*, vol. 2, pp. 11–14, aug 2010.

[130] S. Erb, "Classification of vehicles based on acoustic features," Master's thesis, TU Graz, 2007.

[131] M. L. Moran, R. J. Greenfield, and D. K. Wilson, "Acoustic array tracking performance under moderately complex environmental conditions," *Applied Acoustics*, vol. 68, no. 10, pp. 1241–1262, 2007.

[132] M. Azimi-Sadjadi, A. Pezeshki, and N. Roseveare, "Wideband doa estimation algorithms for multiple moving sources using unattended acoustic sensors," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 44, pp. 1585–1599, oct. 2008.

[133] B. Yang and J. Scheuing, "Cramer-rao bound and optimum sensor array for source localization from time differences of arrival," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, pp. 961–964, 2005.

[134] M. Buck, T. Wolff, T. Haulick, and G. Schmidt, "A compact microphone array system with spatial post-filtering for automotive applications," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 221–224, apr. 2009.

[135] "Iso 11819-1 : Acoustics - method for measuring the influence of road surfaces on traffic noise - part 1: Statistical pass-by method," 1997.

[136] S. M.Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall PTR, 2010.

[137] R. Mehra, "Optimal input signals for parameter estimation in dynamic systems - survey and new results," *IEEE Transactions on Automatic Control*, vol. 19, pp. 753–768, dec. 1974.

[138] J. L. Spiesberger, "Hyperbolic location errors due to insufficient numbers of receivers," *Journal of the Acoustical Society of America*, vol. 110, no. 5, pp. 2666–2666, 2001.

[139] X. Rong Li and V. Jilkov, "Survey of maneuvering target tracking. part i. dynamic models," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, pp. 1333–1364, oct. 2003.

[140] H. Chen, X. R. Li, and Y. Bar-Shalom, "On joint track initiation and parameter estimation under measurement origin uncertainty," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 40, no. 2, pp. 675–694, 2004.

[141] H. Ma and B.-H. Ng, "Distributive target tracking in wireless sensor networks under measurement origin uncertainty," in *Proceedings of the 3rd International Conference on Intelligent Sensors, Sensor Networks and Information.*, pp. 299–304, dec. 2007.

[142] A. Rabaoui, M. Davy, S. Rossignol, Z. Lachiri, and N. Ellouze, "Sélection de descripteurs audio pour la classification des sons environnementaux avec des svms mono-classe," tech. rep., Unité de recherche Signal, Image et Reconnaissance des formes (ENIT), Laboratoire d'Automatique, de Génie Informatique et Signal (INRIA), 2008.

[143] G. Peeters, *A Large Set Of Audio Features For Sound Description (similarity and classification) in the CUIDADO project*, 2004.

[144] G. Tzanetakis, "Song-specific bootstraping of singing voice structure," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, vol. 3, pp. 2027–2023, jun. 2004.

[145] K. Yamamoto, F. Asano, W. van Rooijen, E. Ling, T. Yamada, and N. Kitawaki, "Estimation of the number of sound sources using support vector machines and its application to sound source separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, pp. 485–488, apr. 2003.

[146] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, pp. 387–392, apr. 1985.

[147] L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai, "On detection of the number of signals in presence of white noise," *Journal of Multivariate Analysis*, vol. 20, pp. 1–25, oct. 1986.

[148] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 5, pp. 465–471, 1978.

[149] Y. Liu, J. Soraghan, and T. Durrani, "Detection of number of harmonics by maximum eigenvalue varied rate criteria," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, pp. 2543–2546, apr. 1990.

[150] J. Lei, C. Ping, and Y. Juan, "The source number estimation based on the beam eigenvalue method," in *Industrial Electronics and Applications, 2007. ICIEA 2007. 2nd IEEE Conference on*, pp. 2727–2731, May 2007.

[151] J.-J. Fuchs, "Estimating the number of sinusoids in additive white noise," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, pp. 1846–1853, dec. 1988.

[152] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, dec. 1974.

[153] R. Balan, "Estimator for number of sources using minimum description length criterion for blind sparse source mixtures," in *Proceedings of the 7th international conference on Independent component analysis and signal separation*, ICA'07, (Berlin, Heidelberg), pp. 333–340, Springer-Verlag, 2007.

[154] M. Bierlaire, *Introduction à l'optimisation différentiable*. Presses Polytechniques et Universitaires Romandes, 2006.

[155] M. Pucher, D. Schabus, P. Schallauer, Y. Lypetskyy, F. Graf, H. Rainer, M. Stadtschnitzer, S. Sternig, J. Birchbauer, W. Schneider, and B. Schalko, "Multimodal highway monitoring for robust incident detection," in *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pp. 837 –842, sept. 2010.

[156] T. Liu, Y. Liu, X. Cui, G. Xu, and D. Qian, "Molts: Mobile object localization and tracking system based on wireless sensor networks," in *Networking, Architecture and Storage (NAS), 2012 IEEE 7th International Conference on*, pp. 245 –251, june 2012.

[157] C. Kreucher, A. Hero, and K. Kastella, "Multiple model particle filtering for multitarget tracking," in *Proceedings of the twelfth Annual Conference on Adaptive Sensor Array Processing (ASAP)*, 2004.

[158] M. Morelande, C. Kreucher, and K. Kastella, "A bayesian approach to multiple target detection and tracking," *IEEE Transactions on Signal Processing*, vol. 55, pp. 1589–1604, may 2007.

[159] Y. T. Chan and K. C. Ho, "A simple and efficient estimator for hyperbolic location," *IEEE Transactions on Signal Processing*, vol. 42, pp. 1905–1915, aug 1994.

[160] Y. Chan and K. Ho, "An efficient closed-form localization solution from time difference of arrival measurements," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 393–396, apr. 1994.

[161] J. L. Spiesberger, "Locating animals from their sounds and tomography of the atmosphere: Experimental demonstration," *Journal of the Acoustical Society of America*, vol. 106, no. 2, pp. 837–846, 1999.

[162] J. L. Spiesberger, "Geometry of locating sounds from differences in travel time: Isodiachrons," *Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3168–3177, 2004.

[163] A. Brutti, *Distributed Microphone Networks for Sound Source Localization in Smart Rooms*. PhD thesis, International Doctorate School in Information and Communcation Technologies, 2007.

[164] A. Brutti, M. Omologo, and P. Svaizer, "Comparison between different sound source localization techniques based on a real data collection," in *Hands-Free Speech Communication and Microphone Arrays (HSCMA)*, pp. 69–72, may. 2008.

[165] B. Mungamuru and P. Aarabi, "Enhanced sound localization," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, pp. 1526–1540, 2004.

[166] C. Zhang, Z. Zhang, and D. Florencio, "Maximum likelihood sound source localization for multiple directional microphones," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, pp. 125–128, apr. 2007.

[167] X. Wan and Z. Wu, "Improved steered response power method for sound source localization based on principal eigenvector," *Applied Acoustics*, vol. 71, no. 12, pp. 1126–1131, 2010.

[168] E. A. Lehmann, *Particle Filtering Methods for Acoustic Source Localisation and Tracking*. PhD thesis, Research School of Information Sciences and Engineering, Department of Telecommunications Engineering, The Australian National University, Canberra, ACT, Australia, jul. 2004.

[169] P. Pertilä, *Acoustic source localization in a room environment and at moderate distances*. PhD thesis, Tampere University of Technology, 2009.

[170] R. Boulandet, *Tunable electroacoustic resonators through active impedance control of loudspeakers.* PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2012.

[171] P.-J. René, *Contributions aux études sur le couplage électroacoustique dans les espaces clos en vue du contrôle actif*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2006.

[172] X. Falourd, L. Rohr, M. Rossi, and H. Lissek, "Spatial echogram analysis of a small auditorium with observations on the dispersion of early reflections," in *Proceedings of the $39^{th}$ International Congress and Exposition on Noise control Engineering (Internoise), Lisbon, Portugal*, 2010.

[173] P. Marmaroli, X. Falourd, and H. Lissek, "A uav motor denoising technique to improve localization of surroundings noisy aircrafts: Proof of concept for anti-collision systems," in *$11^{ème}$ Congrés Français d'Acoustique (CFA) - 2012 Institute of Acoustics (IOA) Annual Meeting*, 2012.

[174] H.-G. Kim, N. Moreau, and T. Sikora, *MPEG-7 Audio and Beyond*. Wiley, 2005.

[175] D. Li, ishwar K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Image/Video Indexing and Retrieval*, vol. 22, pp. 533–544, 2001.

[176] S. Chu, S. Narayanan, and C.-C. Kuo, "Environmental sound recognition with time frequency audio features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, pp. 1142–158, aug 2009.

[177] D. Leclercq, J. Cooper, and M. Stead, "The use of microphone windshieds for outdoors noise measurements," in *Acoustics 2008, Geelong, Victorian, Australia 24 to 26 nov. 2008*, 2008.

[178] J. R. Pearse and M. J. Kingan, "Measurement of sound in airflow," in *Proceedings of the 13$^{th}$ International Congress on Sound and Vibration (ICSV)*, 2006.

[179] J. Wuttke, "Microphones and wind," in *Audio Engineering Society Convention 91*, 10 1991.

[180] G. Hessler, D. Hessler, P. Brandstatt, and K. Bay, "Experimental study to determine wind-induced noise and windscreen attenuation effects on microphone response for environmental wind turbine and other applications," *Noise Control Engineering Journal*, vol. 56, no. 4, pp. 300–309, 2008.

[181] G. W. Plice, "Wind and breath noise protector for microphones," 1989.

[182] W. Neise, "Theoretical and experimental investigations of microphone probes for sound measurements in turbulent flow," *Journal of Sound and Vibration*, vol. 39, no. 3, pp. 371–400, 1975.

[183] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, pp. 113–120, Apr. 1979.

[184] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 137–145, apr. 1980.

[185] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, pp. 1109–1121, dec. 1984.

[186] P. Wolfe and S. Godsill, "Simple alternatives to the ephraim and malah suppression rule for speech enhancement," in *Proceedings of the 11$^{th}$ IEEE Signal Processing Workshop on Statistical Signal Processing*, pp. 496–499, 2001.

[187] R. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using dft domain subspace decompositions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 541–553, mar. 2008.

[188] M. Schmidt, J. Larsen, and F.-T. Hsiao, "Wind noise reduction using non-negative sparse coding," in *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing*, pp. 431–436, aug. 2007.

[189] S. T. Roweis, "One microphone source separation," in *In Advances in Neural Information Processing Systems 13*, pp. 793–799, MIT Press, 2000.

[190] D. P. W. Ellis and R. J. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.*, 2006.

[191] E. Nemer and W. Leblanc, "Single-microphone wind noise reduction by adaptive postfiltering," in *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA '09. IEEE Workshop on*, pp. 177–180, oct. 2009.

[192] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH)*, 2006.

[193] K. T. Andersen, *Wind noise reduction in single channel speech signals.* PhD thesis, Technical University of Denmark, Department of Informatics and Mathematical Modeling, Intelligent Signal Processing, 2008.

[194] S. Franz and J. Bitzer, "Multi-channel algorithm for wind noise reduction and signal compensation in binaural hearing aids," in *International Workshop on Acoustic Echo and Noise Control*, 2010.

## Curriculum Vitae

Patrick Marmaroli was born in Saint-Julien-en-Genevois, France, in 1984. He received a M.Sc. degree in signal processing and trajectography from Sud-Toulon Var University and in electronics from the Institut Supérieur de l'Electronique et du Numérique de Toulon, France, in 2008. He carried out his Master thesis on *statistic and perceptive characterization of percussive sounds* at the Laboratoire de Mécanique et d'Acoustique (LMA-CNRS), Marseille, France. Since October 2008, he enrolled as a PhD student at the Laboratoire of Electromagnetics and Acoustics (LEMA) of the Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. He is actively involved in projects dealing with speech extraction, noise speech reduction, sound sources localization and tracking, sound sources detection and classification, microphone and loudspeaker array signal processing. He oversees the work of semester and master students as a Research and Teaching Assistant. His current research interests include acoustic array processing for denoising, localization and multi target tracking.

## List of representative publications

P. Marmaroli, M. Carmona, J.M. Odobez, X. Falourd and H. Lissek. *Observation of vehicle axles through pass-by noise: a strategy of microphone array design*, submitted to IEEE Transactions on Intelligent Transportation Systems.

P. Marmaroli, J-M. Odobez, X. Falourd and H. Lissek. *A bimodal sound source model for vehicle tracking in traffic monitoring.* in Proceedings of the $19^{th}$ European Signal Processing Conference (EUSIPCO), Barcelona, Spain, 2011.

P. Marmaroli, X. Falourd and H. Lissek. *Sensor array optimization for sources separation and detection in the at-worst determined case.* in Proceedings of the $18^{th}$ International Congress of Sound and Vibration (ICSV), Rio de Janeiro, Brasil, 2011.

P. Marmaroli, X. Falourd and H. Lissek. *Study of an octahedral antenna for both sound pressure level estimation and 3D localization of multiple sources.* in Proceedings of the $39^{th}$ International Congress and Exposition on Noise Control Engineering (INTER-NOISE), Lisbon, Portugal, 2010.

H. Lissek, P. Martin, J. Carmona, M. Imhasly, I. Millar, X. Falourd and P. Marmaroli. *Device and method for capturing and processing voice.* Patent number WO2011067292-A1, 2009.