Débruitage multicapteur appliqué à la téléphonie mains-libres en automobile

Charles Fox

▶ **To cite this version:**

Charles Fox. Débruitage multicapteur appliqué à la téléphonie mains-libres en automobile. Traitement du signal et de l'image [eess.SP]. Télécom ParisTech, 2013. Français. NNT : 2013ENST0074 . tel-01156542

HAL Id: tel-01156542
https://pastel.hal.science/tel-01156542

Submitted on 27 May 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

Parrot®

TELECOM ParisTech

2013-ENST-0074

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

TELECOM ParisTech

Spécialité « Signal et Image »

présentée et soutenue publiquement par

Charles Fox

le 5 décembre 2013

Débruitage multicapteur appliqué à la téléphonie mains-libres en automobile

Jury

M. Pascal SCALART, Professeur, ENSSAT Lannion	Rapporteur
M. Emmanuel VINCENT, Chargé de recherche, INRIA Nancy	Rapporteur
M. Karim ABED-MERAIM, Professeur, Polytech Orléans	Examinateur
M. Dirk SLOCK, Professeur, EURECOM	Examinateur
M. Roland BADEAU, Maître de conférences, Télécom Paristech	Directeur de thèse
M. Bertrand DAVID, Maître de conférences, Télécom Paristech	Directeur de thèse
M. Maurice CHARBIT, Professeur émérite, Télécom Paristech	Co-encadrant
M. Guillaume VITTE, Responsable traitement de la parole, Parrot S.A	Co-encadrant

TELECOM ParisTech
école de l'Institut Mines-Télécom - membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

THÈSE

Remerciements

Les travaux présentés dans cet ouvrage ont été effectués à l'aide d'un encadrement pléthorique. En effet, pas moins de 5 personnes se sont relayées durant ces trois ans pour diriger cette thèse. Cette spécificité, due aux départs en retraite en cours de route, a rendu parfois le travail d'encadrement particulièrement ardu.

J'aimerais donc remercier Maurice Charbit, Jacques Prado et Guillaume Vitte, pour m'avoir permis de monter ce projet, né en très grande partie grâce à leur enthousiasme et leur soutien. Un grand merci également à Roland Badeau et Bertrand David, qui ont récupéré le projet en cours de route pour faire face aux départs en retraite de Jacques et Maurice, et se sont assurés qu'il finisse dans les meilleures conditions possibles. Leurs conseils, ainsi que leur soutien amical, m'ont été précieux tout au long de cette période.

Merci également aux membres du jury de cette thèse, Pascal Scalart, Emmanuel Vincent, Karim Abed-Meraim et Dirk Slock pour leurs remarques et leurs analyses, qui m'ont permis de mieux mettre en perspective les travaux réalisés.

Le soutien de mes collègues de Parrot a souvent été d'une grande aide au cours de ces trois ans, qu'il soit scientifique ou festif. Bravo TDS ! Et merci à Pierre, Gaël, Louis, Nicolas, Fadi, Romain, Alex, Fabien, Etienne, Julie, Guillaume, Julien, Thomas, Thomas, Gaspard, Laure, Adrien, Vu, Phong, Benoit, Claire, Gilles, François, Michael, Mathieu, Dominique, ainsi qu'à tout ceux que j'ai eu l'occasion de croiser. Et bien sûr, bonne chance à Éric pour finir sa thèse, la deuxième CIFRE de Parrot.

Mes passages à mon bureau côté Télécom ont été (trop) rares, mais les autres thésards m'ont beaucoup apporté, notamment à travers le BDT. Merci particulièrement à Émilie, Émilie, Sylvain, Sylvain, Olivier, Yao, Cristina, Fabrice... pour les bons moments à Dareau et sur la Butte-aux-Cailles.

Des grandes bises aux amis hors thèse également (même si beaucoup sont ou seront docteurs), particulièrement Alice, Étienne, mon cousin Alex, Luc, Kader, Ricky, Popol, Marguerite, P'tit Sam, Moyen Sam, Pedro, Max', Karim, Clémish, Jérôme, Béné, Ugo...

Une pensée aussi à HHQQP, pour leur soutien régulier avant et après la fin de la Happy Hour.

Bien sûr, toute mon affection à ma famille. Merci à mes soeurs Mathilde et Anne-Sophie, ainsi qu'à Seb, pour m'avoir supporté si longtemps et être toujours là. Merci bien sûr à mes parents Catherine et Didier pour tout ce qu'ils ont fait pour moi. Merci à mes grands parents Roger, Solange, Yvette et son mari Gérard, pour être au top à chaque fois que l'on se voit.

Enfin, merci à la Coloc' de l'Amour pour cette belle idylle de trois ans ensemble. Les poulets dominicaux n'auront plus la même saveur sans vous.

Merci en particulier à Yvonne, qui a été une source d'inspiration tout au long de cette thèse.

Résumé

Les kits pour téléphonie mains-libres en voiture sont un équipement qui devient de plus en plus standard dans les véhicules actuels. Ces accessoires répondent à un besoin de communiquer tout en conduisant, pour des raisons professionnelles ou personnelles.

Or, la voiture s'avère un environnement acoustique particulièrement difficile. En effet, un habitacle de voiture présente une forte réverbération, du fait de la présence de nombreuses surfaces vitrées, et il est aussi très bruyant. Ce bruit vient de multiples sources, comme le moteur, le roulement du pneu sur la route, le vent, la circulation environnante... et ces sources varient fortement d'une condition de conduite à l'autre. La réduction de bruit ambiant au niveau de la prise de son dans l'habitacle constitue donc un élément majeur dans le confort des utilisateurs de ce type d'équipement.

L'objectif des travaux effectués dans le cadre de cette thèse est de fournir une solution efficace de réduction de bruit pour cette application, en utilisant plusieurs microphones.

Nous nous intéressons ici principalement à la situation d'autoroute. Une grande campagne de mesures dans un véritable habitacle automobile en roulant a permis d'observer des caractéristiques spatiales et spectrales du champ de bruit ambiant présent dans cette situation. Ces mesures nous ont permis de mettre en évidence de fortes différences en termes de Rapport Signal à Bruit d'entrée (RSB) et de cohérence spatiale du bruit capté selon la fréquence considérée.

Ces observations nous amènent à concevoir des systèmes hybrides : nous cherchons à appliquer des traitements différents en basses et hautes fréquences.

Nous avons développé une implémentation adaptative du beamforming Minimum Variance Distortionless Response, qui est efficace dans des conditions de fort RSB d'entrée, et quand la cohérence inter-capteur du bruit est faible. Nous avons également étudié le placement des capteurs pour cette approche de façon à maximiser ses performances, qui seront bonnes en hautes fréquences.

Pour les basses fréquences, nous avons étudié deux systèmes :

— L'un est basé sur de l'Annulation de Bruit Adaptative, exploitant un bruit fortement cohérent d'un capteur sur l'autre

— L'autre est un dérivé du Filtre de Wiener Multicanal

A chaque fois, une étude des performances en fonction de l'antenne de capteurs utilisée a été menée, pour utiliser la stratégie acoustique la plus appropriée. Les systèmes hybrides ainsi conçus ont été évalués de façon subjective, en faisant passer un test d'écoute à un panel d'individus. Ce test montre que le système hybride utilisant le filtrage de Wiener permet de réduire de façon significative la gêne liée au bruit ambiant, sans montrer de contrepartie sur la qualité de la parole transmise.

Mots-clés: Téléphonie Mains-libres, Automobile, Réduction de bruit, Multi-capteurs

Abstract

Hands-free car kits have gained a lot of popularity among drivers over the last years. These equipments fill the need for the users to communicate while driving their car, whether it is for professional or personal use.

Having a phone conversation in a car is really challenging, as the inside of an automobile is a strongly adverse acoustic environment. Indeed, this compartment is strongly reverberant, because of the presence of important glass surface (such as the windbreaker and windows), and it is also very noisy.

The noise comes from various sources, such as the engine, the contact of the tire on the road, the wind..., and those sources show different characteristics from one situation to another. Hence, the noise reduction for in-car voice pickup is a major element for the user's comfort.

The main objective of the work reported in this thesis is to build an efficient noise reduction solution for in-car telephony, using a plurality of microphones. We are in this work mostly interested in the freeway situation. Hence, a database of measurements has been recorded in a real car interior, to understand what are the spectral and spatial characteristics of the noise field in this situation. These measurements showed that the noise field has different characteristics in high and low frequencies. Indeed, the noise has more energy, and is more spatially coherent in the low frequency range.

This leads us to propose hybrid subband systems, in order to use different algorithmic approaches in high and low frequencies.

We made an adaptive implementation of the well-known Minimum Variance Distortionless Response beamforming, which is efficient when the input Signal-to-Noise Ratio is high, and the noise field shows a low spatial coherence. We also conducted an analysis on the impact of sensors' positions, in order to build a microphone array which will allow this method to be efficient in the high frequency range.

We also considered two different processings for the low frequency range :

— one is based on Adaptive Noise Cancellation, which uses the high spatial coherence of recorded noises,

— the other is based on Multichannel Wiener Filter.

For both methods, an analysis of the impact of sensors' positions has been made, in order to build an efficient microphone array. To assess the performance of these hybrid systems, a subjective evaluation has been conducted, through a listening test. This evaluation shows that the hybrid system using a Multichannel Wiener Filter in the low frequency range supresses a significant amount of noise, while keeping the voice distortion to a minimum, perceptually unnoticed.

Keywords: Noise Reduction, Automotive, Microphone array, Hands-free

Sommaire

Abréviations et acronymes

AEC Annulation d'écho acoustique (Acoustic Echo Cancellation)

ANC Annulation de bruit adaptative (Adaptive Noise Cancellation)

APA Affine Projection Algorithm

CR-ANC Annulation de bruit adaptative résistante aux fuites de Parole (Crosstalk Resistant Adaptive Noise Cancellation)

DSP Densité Spectrale de Puissance

ECM Electret Condenser Microphone

EQM Erreur Quadratique Moyenne

FBF Beamforming fixe (Fixed Beamforming)

FFT Transformée de Fourier Rapide (Fast Fourier Transform)

GSC Generalized Sidelobe Canceller

GSVD Generalized Singular Value Decomposition

HMM Modèle de Markov caché (Hidden Markov Model)

ITU-T International Telecommunication Union - Telecommunication Standardization Sector

KLT Transformée de Karhunen-Loeve (Karhunen-Loeve Transform)

LMS Least-Mean Square

LSA Log-Spectral Amplitude

MOS Mean Opinion Score

MSC Mean-Squared Coherence

MVDR Minimum Variance Distortionless Response

MWF Filtre de Wiener multicanal (Multichannel Wiener Filter)

NLMS Normalized LMS

NMF Nonnegative Matrix Factorization

NR Réduction de bruit (Noise Reduction)

OLA Overlap-Add

OM-LSA Optimally Modified Log-Spectral Amplitude

PESQ Perceptual Evaluation of Speech Quality

PCM Pulse Code Modulation

RIF Réponse Impulsionnelle Finie

RSB Rapport Signal-à-Bruit

SDW-MWF Speech Distortion Weighted Multichannel Wiener Filter

SIMO Single Input Multiple Output

SPP Probabilité de présence de parole (Speech Presence Probability)

SSL Stationnaire au second ordre au Sens Large

TFD Transformée de Fourier Discrète

TFDi Transformée de Fourier Discrète inverse

TFCT Transformée de Fourier à Court Terme

VAD Détecteur d'activité vocale (Voice Activity Detector)

Notations

$E[]$	Espérance mathématique
^{T}	Transposée
*	Conjuguée complexe
^{H}	Transconjuguée complexe
$x(t)$	Processus temporel scalaire
$\mathbf{x}(t)$	Processus temporel vectoriel
$X(f)$	Grandeur fréquentielle scalaire - Transformée de Fourier à temps discret de $x(t)$
$\mathbf{X}(f)$	Grandeur fréquentielle vectorielle - Transformée de Fourier à temps discret de $\mathbf{x}(t)$
a_k	k-ième composante du vecteur $\mathbf{a}$
$\varphi_x(f)$	Densité Spectrale de Puissance de $x(t)$
$\varphi_{xy}(f)$	Interspectre de $x(t)$ et $y(t)$
$R_x(\tau)$	Autocovariance du processus $x(t)$
$\mathbf{R}_x(\tau)$	Matrice de covariance du processus $\mathbf{x}(t)$
$\Sigma_x(f)$	Matrice spectrale du processus $\mathbf{x}(t)$
$\hat{x}$	Estimée de la grandeur x
$\otimes$	Produit de convolution
$\delta(t)$	Impulsion de Dirac discrète, qui vaut 1 quand $t = 0$, et 0 sinon
c_s	Célérité du son, que l'on considère ici égale à 340m.s^{-1}
$\arg x$	Phase de la valeur complexe x
$\lfloor x \rfloor$	Entier inférieur à x le plus proche de x, x étant réel
$\mathbf{I}_{k\times k}$	Matrice identité de taille $k \times k$
$\mathbf{0}_{k\times l}$	Matrice nulle de taille $k \times l$

Table des figures

Liste des tableaux

Introduction

Téléphonie mains-libres en automobile

Les kits mains libres pour la téléphonie en automobile sont des équipements de plus en plus présents dans les véhicules, que ces systèmes soient installés en option sur les véhicules vendus neufs, ou qu'ils soient installés a posteriori. Ces dispositifs permettent aux automobilistes de communiquer lorsqu'ils conduisent leur véhicule. Le marché des kits mains-libres représente un enjeu commercial important. Notamment, on peut observer sur la Figure 1 l'évolution prévue du taux de voitures neuves équipées d'un système mains-libres Bluetooth.

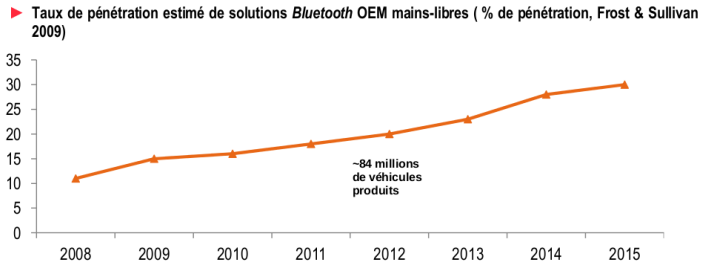


Figure 1 – Taux d'équipement estimé des véhicules neufs en systèmes mains-libres [Frost and Sullivan, 2009]

On voit que le nombre de voitures déjà équipées en sortie d'usine est très important. A ces véhicules se rajoutent les automobiles qui seront équipées plus tard. Ces accessoires répondent à un besoin de sécurité et à un besoin juridique, car dans de nombreux pays, la loi interdit de manipuler un téléphone portable tout en conduisant. Pour répondre à ce besoin de sécurité, les kits mains libres doivent fournir une qualité de communication assez bonne pour ne pas perturber le conducteur, qui doit rester concentré sur sa conduite.

Le confort en communication apporté par ces kits est donc un enjeu majeur. Les deux éléments principaux susceptibles d'altérer la qualité de la communication sont :

Le bruit ambiant qui provient de multiples sources, telles que le moteur, les turbulences environnantes, le roulement du pneu sur la route ;

L'écho acoustique provoqué par le fait que le système d'acquisition acoustique (un ou plusieurs microphones) capte, en plus de la voix du conducteur, la voix du locuteur distant ainsi que les réflexions de celle-ci sur les parois de la voiture

Ces effets sont schématisés dans la Figure 2.

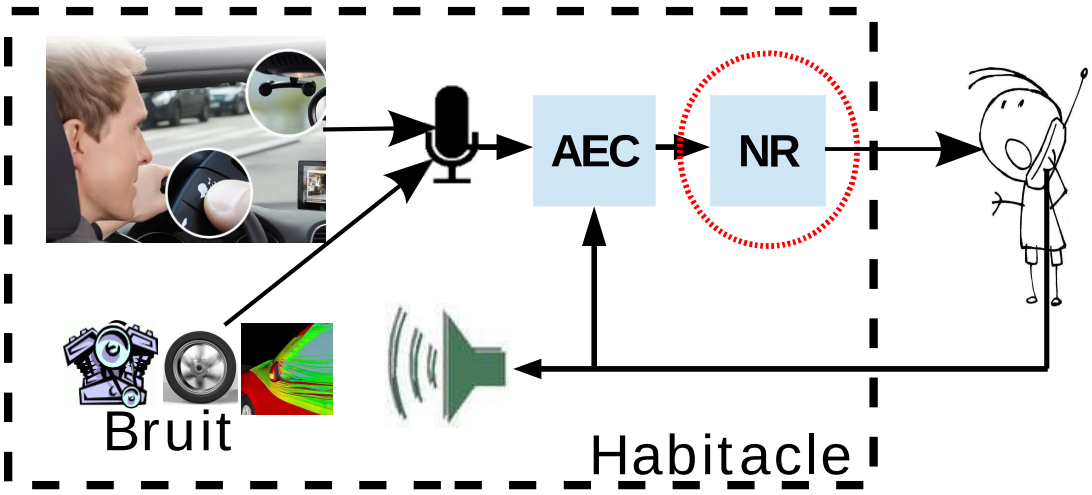


Figure 2 – Communication téléphonique en automobile. Les algorithmes d'amélioration de la parole sont l'Annulation d'écho acoustique (Acoustic Echo Cancellation) (AEC) et la Réduction de bruit (Noise Reduction) (NR).

On remarque que ces effets perturbent surtout le locuteur distant, et pas directement le conducteur. Mais une meilleure compréhension par le locuteur distant permet une conversation plus fluide, et plus confortable pour les deux utilisateurs. Dans les travaux présentés dans cette thèse, nous ne nous intéressons qu'à la perturbation liée au bruit ambiant, les effets de la présence d'écho de voix de la personne distante ne sont pas pris en compte.

L'automobile est un environnement particulièrement bruyant, ce qui affecte la communication. Le but de la réduction de bruit est donc, en premier lieu, d'améliorer l'intelligibilité de la conversation pour que celle-ci soit fluide, mais aussi de conserver une voix naturelle, pour le confort du locuteur.

Le débruitage en automobile est également rendu très difficile par le fait qu'il présente une grande variabilité. En effet, celui-ci va changer d'une voiture à l'autre (certaines voitures sont plus bruyantes que d'autres selon la conception de l'habitacle, de la motorisation...), mais aussi d'une situation à l'autre. En effet, le bruit à l'intérieur de l'habitacle sera très différent entre une situation de ville et d'autoroute, en fonction de la vitesse, de la météorologie...

Dans le cas d'une discussion face-à-face, le cerveau humain est capable de faire abstraction de toutes ces perturbations pour se concentrer sur la voix du locuteur. Le cerveau s'aide notamment des informations apportées par l'écoute binaurale, mais aussi des informations visuelles pour être robuste à toutes sortes de situations défavorables.

Dans le cas de la communication distante, ces informations ne sont pas présentes. C'est donc au traitement embarqué dans le système mains-libres d'améliorer le signal de parole capté pour assister le locuteur distant.

Systèmes de réduction de bruit

Dans cette optique, de nombreux systèmes ont été développés. Les systèmes utilisant un seul capteur exploitent les caractéristiques temps-fréquence des signaux captés. Il s'agit alors d'estimer des caractéristiques du bruit ambiant (principalement la Densité Spectrale de Puissance (DSP)), pour en déduire un gain fréquentiel qui permettra d'atténuer les points temps-fréquence du signal contenant surtout du bruit. L'efficacité d'un tel système repose essentielle-

ment sur la qualité d'estimation des caractéristiques du bruit, qui dépend des hypothèses faites sur le bruit ambiant (stationnarité, distribution des coefficients temps-fréquence...), et ces traitements n'exploitent que l'amplitude spectrale du signal capté [Ephraim and Malah, 1984]. La phase du signal observé n'est donc pas exploitée.

Des systèmes multi-capteurs sont également apparus. Ceux-ci tiennent également compte des informations spatiales sur le champ acoustique présent dans l'habitacle, celui-ci étant mesuré en plusieurs points de l'espace. Il est alors possible d'utiliser les mêmes méthodes que pour les systèmes monocapteur, en ajoutant dans le calcul du gain d'atténuation des informations spatiales sur le bruit et la parole [Guérin et al., 2003]. Il est également possible de faire "pointer" l'acquisition vers une direction particulière (où le locuteur est présent). C'est-à-dire que les ondes venant d'autres directions seront atténuées par le système, car elles ne contiennent que du bruit. Ces méthodes, dites de formation de voie ou beamforming, peuvent être exploitées avec une connaissance a priori de la position du locuteur, ou alors celle-ci peut être apprise par le traitement [Ito et al., 2012]. L'utilisation de ces méthodes est rendue difficile dans un environnement automobile, du fait de la présence d'une forte réverbération du signal de parole sur les parois de l'habitacle [Spriet et al., 2004]. Il est à noter que ces méthodes multi-capteurs permettent d'améliorer non seulement l'estimation de l'amplitude spectrale du signal utile, mais également l'estimation de sa phase [Trawicki and Johnson, 2009].

Environnement

Pour toutes ces méthodes de réduction de bruit, il est important de bien connaitre l'environnement acoustique dans lequel les traitements s'appliquent. Ceci permet, dans tous les cas (mono et multi-capteur(s)), d'établir des hypothèses sur les propriétés des signaux de bruit et de parole cohérentes, de manière à renforcer l'efficacité dans le milieu qui nous intéresse.

Dans le cas multi-capteurs, la connaissance de l'environnement permet de développer conjointement un traitement efficace, et une antenne de microphones adaptée au traitement proposé et à l'environnement [Kodrasi et al., 2011].

En effet, le placement des microphones peut avoir une grande influence dans les performances atteintes par un tel système, il est donc important que le placement des capteurs soit défini en même temps que les stratégies algorithmiques, et ceci en accord avec l'environnement acoustique considéré.

Évaluation des systèmes de réduction de bruit

Une des difficultés des traitements de débruitage de parole est l'évaluation des performances atteintes par un système. En effet, la qualité est avant tout subjective : c'est le confort de l'utilisateur qui est visé.

De nombreuses mesures de performances objectives, que l'on peut évaluer automatiquement, ont été développées [Hu and Loizou, 2008], mais il reste toutefois très difficile de rendre compte de tous les aspects de la perception de la parole par le cerveau humain. Ces mesures sont donc souvent de bons indicateurs, mais il faut être prudent avec les performances estimées par ces méthodes, qui ne rendent pas forcément compte de l'agrément ressenti par un utilisateur final.

Il est donc souvent incontournable d'évaluer ces performances par des méthodes subjectives. Il s'agit alors de faire écouter des signaux traités à un panel d'individus, qui notent les traitements

proposés selon un ou plusieurs critères [ITU-T, 2003].

Ce type de test est en pratique long à réaliser, car il faut de nombreux sujets pour obtenir un résultat significatif.

Il est donc souvent utile d'utiliser des mesures objectives au cours du développement d'un système de réduction de bruit, afin d'obtenir des indicateurs de performance rapidement, et de valider un algorithme en fin de développement par des tests d'écoute.

Apports de la thèse

Les travaux réalisés ont une base expérimentale. De nombreux enregistrements en situation réelle de voiture roulant dans différentes conditions ont été faits, afin d'étudier l'environnement acoustique propre à l'automobile

C'est à partir de ces observations que des méthodes de débruitage multi-capteurs ont été proposées. Celles-ci sont adaptatives et hybrides.

Considérer des méthodes adaptatives permet de prendre en compte le fait que l'environnement est variable au cours du temps, selon le changement d'allure, de type de route... Il est donc intéressant de concevoir des traitements permettant de s'adapter aux changements de conditions.

Les systèmes considérés sont hybrides, pour pouvoir prendre en compte différentes bandes de fréquences. En effet, les observations faites in situ montrent que les caractéristiques spectrales et spatiales du bruit considéré ne sont pas les mêmes en basses et hautes fréquences. Les traitements proposés permettent de prendre en compte cette différence pour appliquer un débruitage adapté pour chaque fréquence.

Nous avons donc développé un traitement adapté aux hautes fréquences, en implémentant un beamforming Minimum Variance Distortionless Response (MVDR) [Fox, 2012], auquel nous avons adjoint un système pour les basses fréquences. Pour le système basses fréquences, deux pistes ont été suivies.

— L'une est basée sur de l'Annulation de bruit adaptative (Adaptive Noise Cancellation) (ANC), pour prendre en compte les caractéristiques spatiales du bruit dans cette bande de fréquences ;

— L'autre est un dérivé du Filtre de Wiener multicanal (Multichannel Wiener Filter) (MWF).

Pour toutes ces méthodes, nous avons également évalué les performances atteintes en fonction du placement des capteurs. Ceci nous a permis, pour chaque système hybride, de concevoir une antenne de microphones adaptée au traitement considéré.

Le système hybride utilisant le MVDR en hautes fréquences et le dérivé du MWF en basses fréquences [Fox et al., 2012, Fox et al., 2013] permet de dépasser les performances atteintes par l'algorithme multi-capteur actuel Parrot dans la situation d'utilisation la plus courante (l'autoroute). Ces performances ont été évaluées subjectivement, en faisant passer un test d'écoute à un panel de sujets.

Organisation du manuscrit

Ce manuscrit est organisé en 6 chapitres. Dans le Chapitre 1, nous présentons de nombreuses méthodes de débruitage mono et multi-capteurs couramment utilisées, avant d'examiner plus en détail les méthodes pertinentes pour notre étude, au vu du contexte des travaux réalisés.

Le Chapitre 2 présente une grande partie du travail expérimental effectué. Nous montrons notamment quels sont les capteurs que l'on utilise, avec les caractéristiques de chacun. Nous étudions les sources acoustiques présentes dans un habitacle automobile, avec leurs caractéristiques spectrales et spatiales. Ce sont ces observations qui mettent en évidence des différences entre les hautes fréquences et les basses fréquences, et qui nous amènent à considérer des systèmes hybrides en sous-bandes.

Les méthodes considérées pour les hautes et basses fréquences font l'objet des trois chapitres suivants.

Le Chapitre 3 présente une implémentation adaptative du beamforming MVDR. Une étude est également menée pour montrer que cette approche sera efficace en hautes fréquences, mais présentera des performances limitées en basses fréquences. Nous verrons également une étude sur l'influence du placement et du type de capteurs utilisés sur les performances, afin de concevoir une antenne de microphones adaptée à ce traitement.

Le Chapitre 4 est consacré à l'étude d'un système basé sur l'ANC. Au vu des observations faites sur le bruit, cette méthode semble adaptée aux basses fréquences dans notre cas. Nous conduisons une approche expérimentale pour définir la meilleure implémentation pour ce système, à la fois acoustique et algorithmique.

Le Chapitre 5 détaille un traitement de type Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF). Nous examinons ce système dans sa globalité, et nous étudions également ses performances en fonction de la stratégie acoustique choisie. Ceci nous permet de concevoir une antenne de microphones adaptée à cet algorithme, pour les basses fréquences.

Le Chapitre 6 aborde conjointement les systèmes hybrides considérés et les résultats subjectifs obtenus, en comparaison avec l'algorithme actuel Parrot. Pour chaque système considéré, nous justifions l'approche en sous-bandes proposée, et nous présentons l'implémentation faite de ces systèmes. La méthodologie du test subjectif mené, ainsi que les résultats obtenus, sont ensuite présentés.

Chapitre 1

État de l'art

Sommaire

Nous allons dans ce Chapitre présenter les différentes applications du débruitage de parole, ainsi que les méthodes couramment utilisées. Nous présenterons ensuite le contexte de l'étude : la réduction de bruit dans la prise de parole en automobile, pour la téléphonie mains-libres. Après avoir présenté les spécificités de l'environnement automobile, nous définirons la stratégie choisie dans le cadre des travaux effectués lors de cette thèse, ainsi que les différentes pistes algorithmiques envisagées.

1.1 Généralités sur le débruitage de parole

1.1.1 Enjeux

La communication orale est un mode de transmission d'information incontournable, que ce soit dans sa forme la plus simple, comme la discussion face-à-face, ou dans des formes plus complexes, comme les interactions homme-machine rendues possibles par le développement de systèmes de reconnaissance vocale, ou encore la téléphonie.

Or, dans de nombreux cas, le son reçu par le capteur acoustique (que ce soit l'oreille du locuteur ou un système microphonique) est un signal de parole dégradé par la présence de bruit. Cela apparaît dès que le locuteur et ce capteur acoustique sont placés dans un environnement bruyant, ce qui est très souvent le cas, notamment dans le cadre de la téléphonie mobile.

C'est pourquoi des traitements de débruitage ont été développés, pour de nombreux environnements et de nombreuses applications, dont quelques unes sont données ci-dessous :

Aides auditives Dans le cas de personnes malentendantes, il est important d'améliorer le signal de parole reçu au niveau de l'oreille, afin d'améliorer l'intelligibilité d'une conversation. C'est pourquoi des appareils d'aide auditive embarquent un système de débruitage de parole dans une prothèse implantée dans l'oreille [Montgomery and Edge, 1988].

Télécommunications L'essor de la téléphonie, et particulièrement la téléphonie mobile, a permis de téléphoner depuis presque partout. Il est devenu commun d'appeler depuis un lieu bruyant, que ce soit dans la rue, au milieu d'une foule...

Il est donc important de pouvoir réduire le bruit au niveau de la prise de son, que ce soit directement pour un terminal téléphonique [Jeub et al., 2012] ou un système dit "mains libres", par exemple une oreillette Bluetooth [Tashev et al., 2005]. On peut aussi voir des systèmes pour des applications plus spécifiques, comme la téléphonie mains-libres en voiture [Goulding and Bird, 1990], ou les systèmes pour téléconférence [Diethorn, 1997].

Ces systèmes ne se limitent pas à la téléphonie. De fait, des systèmes de débruitage pour des situations très spécifiques comme la communication à partir du cockpit d'un avion, par exemple [Harrison et al., 1986], ont été développés.

Interfaces homme-machine De nombreux appareils, dont les smartphones, embarquent des moteurs de reconnaissance vocale pour pouvoir être commandés directement à la voix. Pour rendre la reconnaissance fiable, il est souvent utile de débruiter le signal de parole capté, pour que l'algorithme de reconnaissance ne soit pas gêné par le bruit parasite [Hirsch and Ehrlicher, 1995].

Une autre application est l'identification par empreinte vocale : il s'agit pour la machine de reconnaître la personne qui parle. Là aussi un étage de réduction de bruit est souvent utile pour la fiabilité de l'identification [Ortega-García and González-Rodríguez, 1996].

Analyse d'enregistrements Dans d'autres cas il est intéressant de pouvoir débruiter un enregistrement vocal *a posteriori*, comme par exemple dans le cas d'analyse d'enregistrements de surveillance, ou de restauration d'enregistrements dégradés par du bruit. Du fait de l'absence de contraintes liées à un traitement temps-réel, il est alors possible de développer des traitements spécifiques [Cohen, 2004].

1.1.2 Un aperçu des méthodes de débruitage de parole mono-capteur

De nombreux systèmes utilisent un unique microphone pour capter le signal de parole dégradé par le bruit. Les traitements associés utilisent donc un seul signal pour assurer le débruitage. On cite ici quelques unes des ces méthodes :

— les méthodes temporelles ;
— les méthodes sous-espace ;
— les méthodes fréquentielles ;
— les méthodes par factorisation de matrice.

Méthodes temporelles

Ces traitements prennent en compte directement le signal capté et échantillonné, sans transformation préalable.

Filtrage de Wiener [Wiener, 1950] C'est un estimateur du signal de parole au sens des moindres carrés. Cela consiste à appliquer un filtre linéaire sur les échantillons captés pour obtenir une estimation du signal de parole non bruité. Ce filtre est estimé en fonction des statistiques du signal observé, et la covariance entre signal de parole et signal observé. Du fait de la disponibilité de périodes de bruit seul (lorsque le locuteur est silencieux), il peut être plus avantageux d'estimer le bruit par une méthode de Wiener, puis de le soustraire au signal observé [Chen et al., 2006].

Filtrage de Wiener pondéré [Florencio and Malvar, 2001] La fonction de coût associée au filtrage de Wiener peut être séparée en, d'une part, une composante de distorsion sur le signal de parole et, d'autre part, une composante de bruit résiduel. En connaissant séparément les statistiques du signal de parole et du signal de voix, on peut accorder plus ou moins d'importance à l'une de ces composantes, selon qu'il soit préférable d'avoir un filtre favorisant un signal moins distordu mais plus bruité ou l'inverse [Chen et al., 2011].

Filtre de Kalman [Gannot et al., 1998] On suppose que le signal de parole et l'observation forment un processus bivarié Modèle de Markov caché (Hidden Markov Model) (HMM), linéaire et gaussien. Le filtre de Kalman effectue un estimation du signal de parole conditionnellement aux observations présentes et passées de façon récursive, en se basant sur un modèle de voix et de bruit (par exemple un modèle autorégressif). Il est possible d'introduire un modèle de parole plus complet, prenant en compte de la parole voisée et non-voisée [Goh et al., 1999].

Méthodes sous-espace

Ces méthodes s'appuient sur des modèles qui séparent l'espace d'observations en un sous-espace composé de parole et de bruit, et un sous-espace supplémentaire composé de bruit seul. Le principe est résumé dans la Figure 1.1.

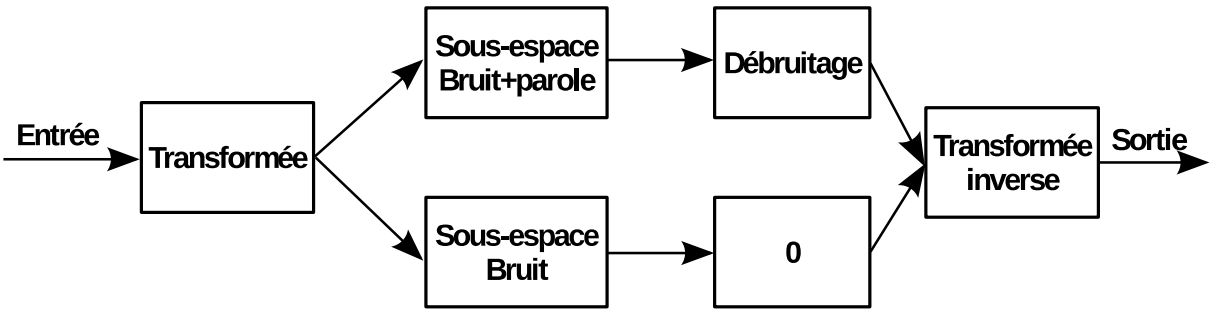


Figure 1.1 – Méthode sous-espace

On présente deux transformations utilisées couramment : la Transformée de Karhunen-Loeve (Karhunen-Loeve Transform) (KLT) et la transformée en ondelettes.

KLT [Ephraim and Van Trees, 1995] La KLT consiste à prendre comme base les vecteurs propres de la matrice de covariance du signal observé. En supposant un bruit additif blanc (éventuellement après blanchiment), les espaces propres présentant la plus grande énergie correspondent au sous-espace de voix bruitée, les autres à du bruit seul. On peut supprimer le sous-espace de bruit seul et estimer le signal de parole à partir du supplémentaire [Hermus and Wambacq, 2004]. Un traitement plus général pour bruits colorés est présenté dans [Hu and Loizou, 2003], utilisant une approche par Generalized Singular Value Decomposition (GSVD) et la connaissance de la matrice de covariance du bruit.

Méthodes par transformée en ondelettes [Seok and Bae, 1997] Le principe est de transformer le signal en ondelettes, et de seuiller les coefficients associés, en mettant à zéro les coefficients faibles (correspondant à du bruit), puis retransformer dans le domaine temporel. Différentes méthodes existent pour déterminer le seuil [Donoho, 1995, Bahoura and Rouat, 2001]. De plus, il est possible d'appliquer uns estimation de Wiener sur les coefficients correspondant à de la parole bruitée [Ambikairajah et al., 1998].

Méthodes fréquentielles

Ces méthodes exploitent une représentation fréquentielle des signaux considérés, en utilisant la transformée de Fourier. Cette représentation permet d'exploiter des propriétés spectrales, en appliquant un traitement sur chaque fréquence séparément.

Extension des méthodes temporelles La plupart des méthodes temporelles s'étendent naturellement au domaine fréquentiel : le filtrage de Wiener [Kirchauer et al., 1995], le filtrage de Wiener pondéré [Doclo et al., 2007], le filtrage de Kalman [Fujimoto and Ariki, 2000]. Il s'agit simplement d'adapter le modèle et les statistiques estimées au domaine fréquentiel.

Soustraction spectrale [Boll, 1979] Dans le domaine fréquentiel, en supposant le bruit et la voix non corrélés, la Densité Spectrale de Puissance (DSP) du signal à chaque fréquence est la somme de la DSP du bruit et de la DSP de la parole. En estimant la DSP du bruit, on débruite en retranchant cette estimation au spectre observé.

Estimation d'amplitude spectrale [Ephraim and Malah, 1984] Il s'agit d'estimer le module du signal de parole pour chaque fréquence au sens des moindres carrés. Cela correspond à l'application d'un gain à chaque fréquence, dépendant des statistiques des signaux de bruit et de parole, visant à atténuer les fréquences comportant surtout du bruit.

Estimation de la log-amplitude spectrale [Ephraim and Malah, 1985] Le principe est le même, mais l'estimation porte alors sur le logarithme du module, auquel l'oreille est plus sensible [Porter and Boll, 1984].

Méthodes par factorisation de matrices

En utilisant la Transformée de Fourier à Court Terme (TFCT), on obtient une représentation temps-fréquence du signal : le spectrogramme, qui donne une représentation fréquentielle du signal à un instant donné. Celui-ci est approximativement la somme du spectrogramme de parole et du spectrogramme du bruit. Les méthodes présentées ici ont pour but de séparer le spectrogramme observé en deux composantes : bruit et parole. Ces méthodes se démarquent des méthodes fréquentielles présentées avant par le fait qu'elles prennent en compte toutes les fréquences en même temps pour le débruitage, alors que les précédentes n'agissent que sur une fréquence à la fois. La séparation se faisant sur les modules, on utilise des méthodes de Non-negative Matrix Factorization (NMF). On utilise alors un dictionnaire de parole, composé d'un certain nombre de trames fréquentielles, généralement apprises sur des séquences de parole seule, et un dictionnaire de bruit, appris sur des séquences de bruit seul. Il s'agit alors d'estimer des coefficients d'activation qui vont déterminer quelle combinaison de ces dictionnaires représente le spectre traité. En séparant les composantes voix et bruit, on peut obtenir les spectrogrammes séparés, comme illustré sur la Figure 1.2.

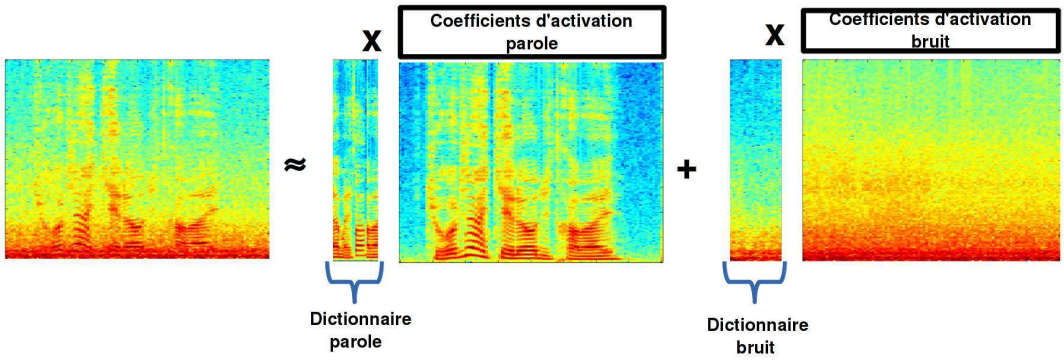


Figure 1.2 – Factorisation de matrice

Cette séparation peut se faire en intégrant des connaissances a priori sur les sources de voix et bruit [Wilson et al., 2008] ou de parcimonie sur les coefficients d'activation [Virtanen, 2007]. Il est également possible d'intégrer un modèle d'état caché dynamique (comme pour le filtrage de Kalman) pour cette situation [Févotte et al., 2013]. Il est possible d'utiliser des versions semi-

supervisées de cette séparation, dans le cas où l'on ne connait pas tous les dictionnaires à l'avance [Joder et al., 2012]. Une fois estimés par cette méthodes les modules du spectre de la parole et du bruit, on applique généralement un traitement spectral semblable au filtrage de Wiener. Ces méthodes sont surtout efficaces lorsque le bruit est redondant, et peut donc être expliqué par un dictionnaire restreint [Schmidt et al., 2007].

1.1.3 Un aperçu des méthodes de débruitage de parole multi-capteurs

Des systèmes d'acquisition utilisent une antenne de microphones afin d'ajouter des informations spatiales aidant le débruitage de la parole captée. Certaines méthodes utilisant plusieurs microphones sont des extensions naturelles de méthodes mono-capteur : nous les présenterons brièvement, avant de nous intéresser aux méthodes spécifiquement multi-capteurs.

Extension des méthodes mono-capteur

Filtrage de Wiener Certaines méthodes temporelles ont leur pendant multi-capteur, en combinant les échantillons enregistrés par les différents microphones. Les statistiques utilisées contiennent alors non seulement de l'information temporelle, mais aussi spatiale. Les dérivés du filtre de Wiener peuvent ainsi être utilisés en mono comme en multi-capteurs [Benesty et al., 2011, Spriet et al., 2004]. Les versions fréquentielles associées également [Doclo et al., 2007].

Méthodes de sous-espace La même extension peut être faite pour certaines méthodes de sous-espace, comme celles basées sur une GSVD des matrices de covariance du signal observé et du bruit [Doclo and Moonen, 2002].

Factorisation des méthodes fréquentielles [Balan and Rosca, 2002] Sous certaines hypothèses [Hendriks et al., 2009], les méthodes basées sur l'estimation du spectre (filtre de Wiener), de l'amplitude spectrale et la log-amplitude spectrale peuvent se factoriser dans le cas multi-capteurs en un traitement spatial suivi du même estimateur en mono-capteur. Le traitement spatial est un **beamforming Minimum Variance Distortionless Response** (MVDR) ([Capon, 1969, Habets et al., 2010a]), dépendant des caractéristiques spatiales du bruit et du signal de parole. Le traitement mono est appelé post-filtre, et dépend du type d'estimateur utilisé. Ainsi les estimateurs fréquentiels de Wiener, d'amplitude spectrale et de log-amplitude spectrale [Gannot and Cohen, 2004] se factorisent comme décrit dans la Figure 1.3.

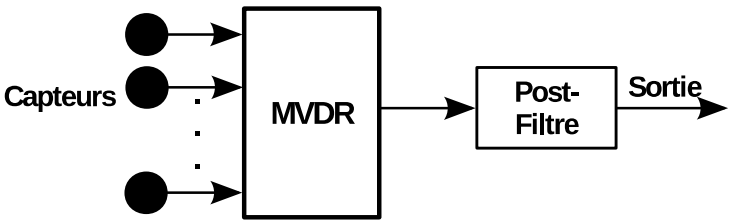


Figure 1.3 – Factorisation des méthodes fréquentielles

Les méthodes de soustraction spectrales peuvent être adaptées au cas multi-capteurs, en

se basant non plus sur le module spectral d'un signal unique, mais sur l'interspectre des signaux captés [Azirani et al., 1997].

Méthodes purement multi-capteurs

Beamforming Outre le MVDR, il existe de nombreuses façons de combiner les signaux reçus par les différents capteurs pour améliorer la qualité du signal. Citons par exemple le **matched beamforming**, qui consiste à créer des interférences constructives sur le signal de parole, en se basant uniquement sur la propagation du signal de voix. Le **Generalized Sidelobe Canceller** (GSC), illustré dans la Figure 1.4, [Griffiths and Jim, 1982] est une implémentation adaptative du MVDR qui est assez largement utilisée dans les systèmes de réduction de bruit. [Lepauloux, 2010].

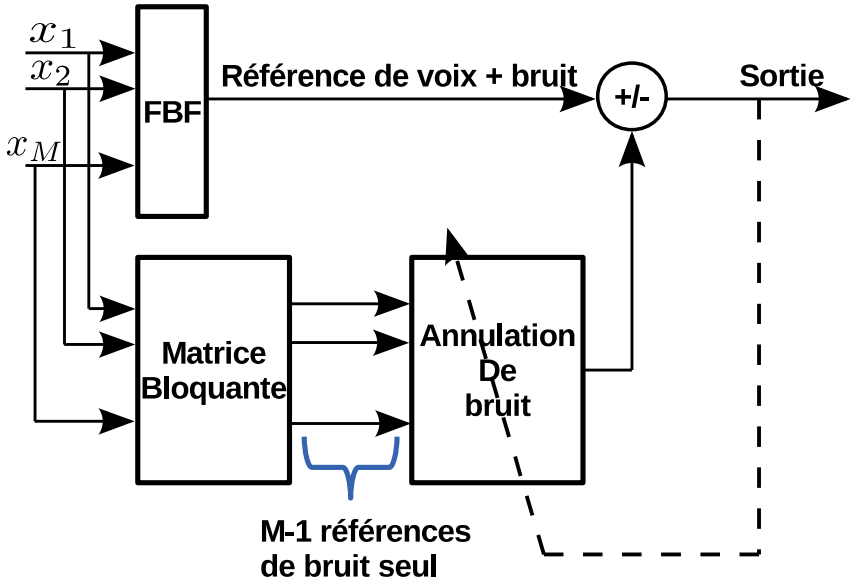


Figure 1.4 – GSC - Schéma global

Le Beamforming fixe (**Fixed Beamforming**) (FBF) est associé à une contrainte de non-distorsion, et la matrice de blocage permet d'annuler le signal de parole pour créer des références de bruit. L'annulation de bruit permet de minimiser l'énergie du bruit en sortie du traitement.

Méthodes d'Annulation de bruit adaptative [Widrow et al., 1975] Dans le cas particulier où l'on dispose un microphone captant la parole et du bruit, et d'un microphone captant uniquement le bruit (référence de bruit seul), il est possible, sous certaines hypothèses, d'annuler le bruit sur le premier capteur par filtrage adaptatif, à partir de la référence de bruit, comme illustré dans la Figure 1.5.

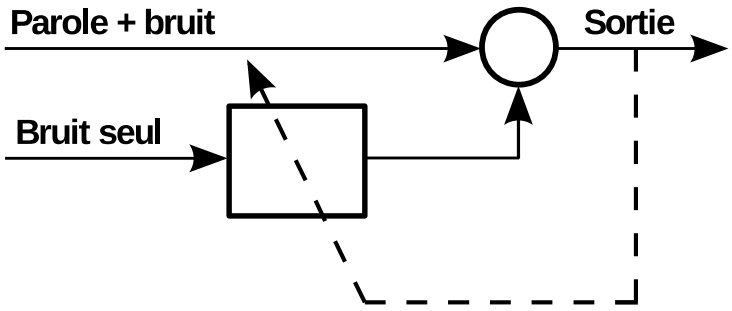


Figure 1.5 – Annulation de bruit adaptative (Adaptive Noise Cancellation) (ANC)

Or, dans de nombreux cas, il n'est pas possible d'obtenir une telle référence de bruit seul. S'il y a une fuite de parole sur cette référence de bruit, ce système provoque de la distorsion en sortie, ainsi que du bruit résiduel. C'est pourquoi des méthodes ont été développées pour le cas où la référence de bruit contient aussi de la parole [Gardner and Agee, 1980, Madhavan and De Bruin, 1990, Lepauloux, 2010]. Ces méthodes correspondent alors à une séparation de sources en aveugle [Djendi et al., 2013a], l'une des sources étant la parole, et l'autre le bruit. De telles méthodes peuvent également être utilisées avec des antennes comportant plus de deux capteurs [Zeng and Abdulla, 2006].

Une étude des performances atteignables et des cas d'utilisation de ces différentes méthodes multi-capteurs peut être trouvée dans [Bitzer et al., 1998].

1.1.4 Considérations sur l'estimation des statistiques nécessaires au débruitage

On a jusqu'ici présenté les méthodes de débruitage de façon très générale, en considérant comme connues les statistiques nécessaires à leur fonctionnement. Nous abordons brièvement cette problématique dans cette partie.

Cas mono-capteur

On a vu que la plupart des méthodes de débruitage présentées nécessitent la connaissance de statistiques sur le bruit, la parole, ou le signal observé. Ces statistiques peuvent être temporelles ou fréquentielles. On utilise de manière générale des estimateurs de l'espérance, qui sont une moyenne empirique d'une grandeur sur une certaine période temporelle. La principale difficulté est d'estimer les statistiques soit de la parole seule, soit du bruit seul. Il est alors courant d'utiliser les périodes de bruit seul (le locuteur ne parle pas) pour estimer ces dernières, et d'en déduire les statistiques de la parole en utilisant la décorrélation du bruit et de la parole, comme on l'illustre dans la Figure 1.6.

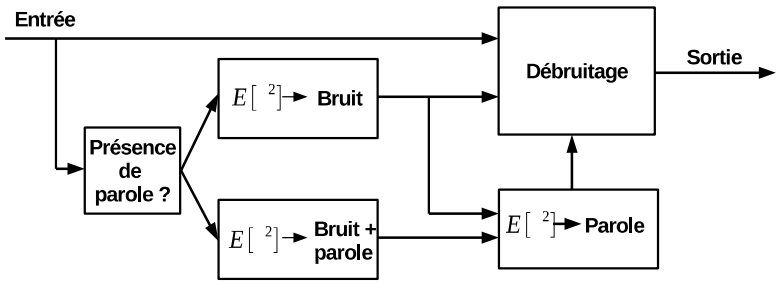


Figure 1.6 – Estimation de statistiques

Dans les traitements fréquentiels, la présence de parole peut être binaire [Hirsch and Ehrlicher, 1995], ou être caractérisée par une Probabilité de présence de parole (Speech Presence Probability) (SPP) [Cohen and Berdugo, 2002, Cohen, 2003, Cohen and Berdugo, 2001]. Notons à ce sujet que l'utilisation d'une probabilité de présence de parole peut aussi permettre de modifier les estimateurs utilisés dans le débruitage, en combinant un estimateur pour le cas où la parole est absente [Cappé, 1994] en plus d'un estimateur classique pour le cas où la parole est présente.

Cas multi-capteur

Dans le cas où l'on utilise une antenne de capteurs, il est nécessaire d'obtenir non seulement des statistiques temporelles, notamment pour le calcul du post-filtre vu plus haut, mais aussi des statistiques spatiales pour le traitement spatial. Il est possible d'exploiter la prise de son multi-microphones pour estimer les statistiques de la parole et du bruit. On utilise alors des informations spatiales mesurées sur ces capteurs, comme la covariance, en utilisant un modèle pour le champ de bruit. Citons par exemple le cas d'un bruit spatialement blanc [Zelinski, 1988], diffus [McCowan and Bourlard, 2003], ou des bruits dont la matrice de covariance est contenue dans un espace connu [Ito, 2012]. Dans le cas du GSC, il est possible d'égaliser la variance du bruit présent sur les références de bruits créées pour estimer la variance du bruit en sortie [Choi and Kang, 2011].

Les statistiques spatiales sur le bruit peuvent être estimées de la même façon qu'en mono-capteur, durant les périodes d'absence de parole. Il est également possible d'utiliser d'autres méthodes lorsqu'on dispose déjà d'informations sur la propagation du signal de parole (par exemple si l'on connait la géométrie de l'antenne de capteurs et la position du locuteur) [Hendriks and Gerkmann, 2011].

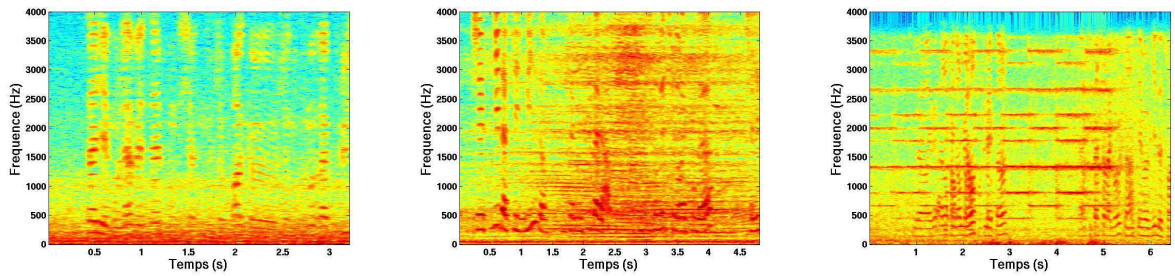
En ce qui concerne les informations relatives à la propagation de la parole, si le bruit est spatialement blanc, on peut utiliser une méthode de sous-espace sur les matrices de covariance spatiale observées [Affes and Grenier, 1997]. Ces méthodes peuvent être généralisées à d'autres types de champs de bruit [Abed-Meraim et al., 1997, Markovich et al., 2009]. Des méthodes basées sur la stationnarité du bruit et la non-stationnarité de la parole existent également [Gannot et al., 2001].

1.2 Contexte de l'étude : téléphonie mains-libres en automobile

1.2.1 Conditions de bruit

Les bruits présents en automobile proviennent de nombreuses sources : moteur, roulement des pneumatiques sur la route, bruits aérodynamiques, circulation environnante...

On a donc une forte variabilité des conditions de bruit. En effet, celles-ci dépendent de la vitesse du véhicule, du régime moteur, du type de voiture, du type de route (autoroute, ville...). On peut notamment avoir du bruit stationnaire lorsque l'on est sur autoroute à vitesse constante sans circulation autour, et du bruit fortement instationnaire en ville, en raison de la circulation (deux-roues qui doublent, sirènes...) et des variations de vitesse. Quelques exemples de spectrogrammes pour certaines situations sont présentés dans la Figure 1.7.



(a) (b) (c)

Figure 1.7 – Voix dégradée par divers bruits : (a) bruit d'autoroute, (b) passage d'un deux-roues, (c) sirène de pompiers

On voit que les bruits liés au moteur, vent et roulement (présents dans le cas du bruit d'autoroute) présentent aussi une puissance spectrale qui varie en fonction de la fréquence : ces bruits sont majoritairement concentrés dans les fréquences basses, en dessous de 1 kHz environ [Guérin et al., 2003], comme on le voit dans la Figure 1.7.

1.2.2 Propagation de la voix

L'onde sonore émise par un locuteur dans une voiture traverse un canal acoustique, constitué du chemin direct entre la bouche et le capteur, mais aussi de chemins passant par des réflexions sur les surfaces intérieures de la voiture, notamment les vitres. Un exemple de réponse impulsionnelle, modélisant ce chemin acoustique, est représenté dans la Figure 1.8.

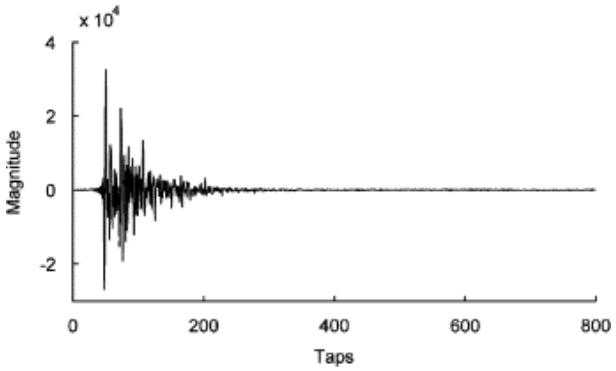


Figure 1.8 – Réponse impulsionnelle de voiture, à une fréquence d'échantillonnage de 16 kHz [Benallal and Benkrid, 2007]

On voit que c'est une réponse courte, mais qui comporte de fortes composantes liées aux chemins indirects. Il n'est donc *a priori* pas possible de modéliser ce canal acoustique uniquement par un chemin direct.

De plus, ce chemin acoustique, de fait des composantes liées à des réflexions sur des surfaces proches, sera fortement modifié s'il y a déplacement, même léger, du locuteur ou du capteur [Motojima et al., 2009].

Si l'on souhaite exploiter ce canal acoustique, il sera donc nécessaire de l'apprendre tout au long du traitement réalisé.

1.2.3 Contraintes de placement de microphones

Dans un habitacle automobile, il n'est possible de placer les microphones qu'à certains endroits, pour ne pas gêner le conducteur, et pour pouvoir fixer les capteurs à l'habitacle de la voiture.

Dans les systèmes mains-libres, les microphones sont généralement placés en hauteur, au niveau du pare-soleil ou du plafonnier central, comme représenté dans la Figure 1.9.

Figure 1.9 – Intérieur d'une voiture. La zone hachurée représente les emplacements possibles pour les microphones.

La prise de son est donc distante, car on ne peut pas placer de capteur à proximité immédiate

de la bouche du conducteur.

1.2.4 Méthodes employées dans ce contexte

Les méthodes fréquentielles ont été largement exploitées dans le cas de la réduction de bruit en automobile. Dans le cas mono-capteur, les méthodes basées sur l'estimation d'amplitude spectrale et de log-amplitude spectrale ont montré de bonnes performances pour améliorer la qualité du signal de parole transmis [Shozakai et al., 1997, Scalart et al., 1996]. Dans le cas multi-capteurs, des méthodes exploitant l'information spatiale pour l'estimation des statistiques du bruit et de la parole ont été implémentées avec succès [Li and Akagi, 2006]. Ces méthodes peuvent être couplées à un beamforming adaptatif [Fuchs et al., 2004]. Ces informations spatiales peuvent également être exploitées pour estimer une SPP [Vitte et al., 2012].

Outre ces méthodes de combinaison et d'estimation de statistiques, des traitements exploitant la différence de cohérence spatiale entre le signal de parole, très cohérent sur toute la bande de fréquence, et le bruit, qui perd en cohérence en hautes fréquences, ont été introduites dans le cas automobile, sous forme de soustraction spectrale à deux capteurs [Guérin et al., 2003].

1.2.5 Problématique de l'étude

Le traitement de réduction de bruit utilisant un seul microphone actuellement exploité par Parrot est basé sur l'estimation de la log-amplitude spectrale du signal utile [Cohen, 2002]. L'approche choisie pour cette thèse est le développement d'un étage multi-capteurs qui viendrait en amont du traitement mono-capteur déjà en œuvre, comme illusré dans la Figure 1.10.

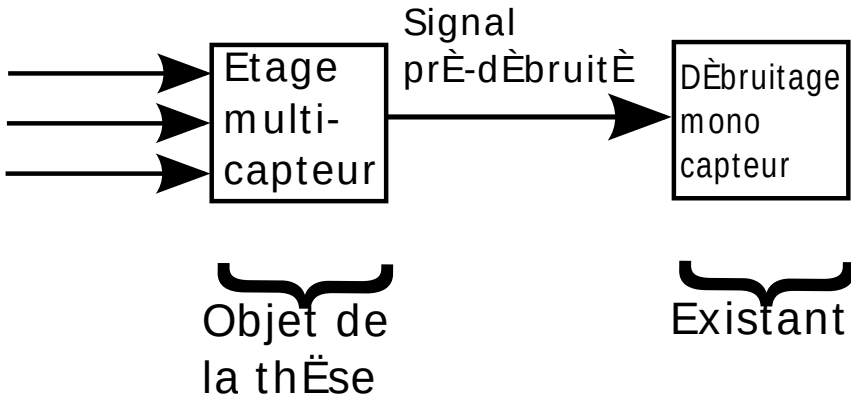


Figure 1.10 – Schéma de la problématique de l'étude

Il s'agit donc, à l'aide de la pluralité de capteurs, de fournir un signal mono-capteur déjà débruité. Celui-ci viendrait ensuite en entrée du traitement mono, et doit être le plus débruité possible pour permettre d'améliorer la sortie finale de ce post-filtre.

On s'intéresse ici au cas où cet étage multi-capteur est un filtrage linéaire, c'est-à-dire que si l'on écrit le signal sur l'entrée m $x_m(t)$, pour $m \in [1 \square M]$, on cherche un signal débruité $\hat{s}(t)$ tel que :

$$\hat{s}(t) = \{g^T \otimes x\}(t) \tag{1.1}$$

où $x(t) = [x_1(t) \; \square\square\square x_M(t)]^T$ et $g(t) = [g_1(t) \; \square\square\square g_M(t)]^T$ est le filtre recherché. En particulier, on s'intéresse ici à trois méthodes :

— le MVDR, dans le domaine fréquentiel,

— le Filtre de Wiener multicanal (Multichannel Wiener Filter) (MWF), dans le domaine temporel,

— l'ANC, dans le domaine temporel.

Nous allons à présent décrire plus en détail le système mono-capteur existant ainsi que son extension multi-capteur, avant de nous intéresser aux pistes explorées pour la création d'un étage multi-capteurs adapté à l'environnement automobile.

1.3 Débruitage fréquentiel

1.3.1 Principe général

On enregistre de la voix dans un environnement bruité à l'aide d'un ou plusieurs capteurs. Le signal capté est donc, pour chaque capteur, composé d'un signal de parole $s(t)$(le signal utile) et d'une composante de bruit $b(t)$, que l'on souhaite éliminer. Le signal observé échantillonné s'écrit alors, pour l'instant discret t :

$$x(t) = s(t) + b(t) \tag{1.2}$$

Il s'agit de passer dans le domaine temps-fréquence, d'appliquer un traitement au spectre du signal observé, qui vise à réduire le bruit, puis de repasser dans le domaine temporel.

Pour passer dans le domaine temps-fréquence, on utilise généralement la TFCT. La méthode consiste à segmenter le signal en trames successives recouvrantes et à appliquer une fenêtre d'analyse sur chaque trame.

Après Transformée de Fourier Discrète (TFD), on obtient ainsi une représentation fréquentielle locale pour chaque trame, sur laquelle on peut appliquer une transformation pour réduire le bruit. Ayant obtenu un signal débruité dans le domaine fréquentiel, le passage au domaine temporel se fait par Transformée de Fourier Discrète inverse (TFDi) pour obtenir la trame temporelle débruitée, puis l'on reconstruit le signal global par Overlap-Add (OLA), qui consiste à additionner les échantillons des trames se chevauchant. Un schéma global du traitement est présenté dans la Figure 1.11.

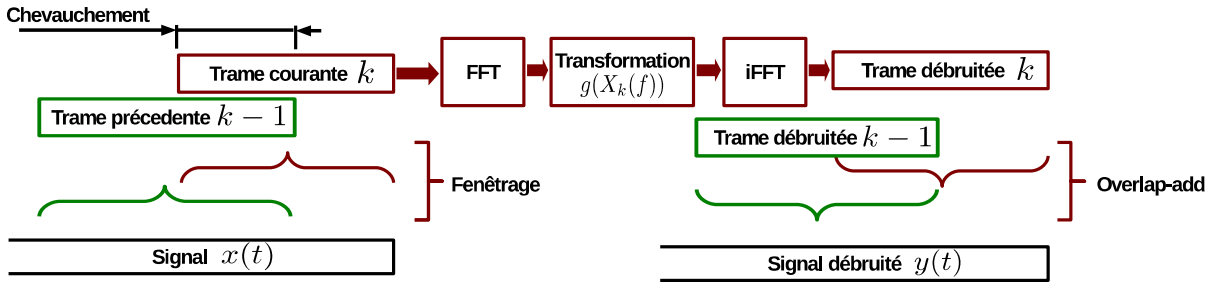


Figure 1.11 – Fenêtrage et Transformée de Fourier

On passe donc d'une représentation du signal purement temporelle (forme d'onde) à une représentation temps-fréquence, qui offre une information spectrale sur le contenu du signal

observé.

En écrivant le signal observé dans le domaine fréquentiel, on obtient, pour la fréquence discrète f et pour chaque trame n :

$$X_n(f) = S_n(f) + B_n(f)$$

Dans la suite, nous ne préciserons plus l'indice de trame n pour plus de lisibilité.

Dans la suite, $S(f)$ et $B(f)$ sont supposés indépendants, circulaires gaussiens, centrés et de variances respectives $\varphi_s(f)$ et $\varphi_b(f)$.

1.3.2 Débruitage mono-capteur

Estimation séparée de la phase et de l'amplitude spectrale

Le traitement mono-capteur existant consiste en une estimation séparée de l'amplitude et de la phase du signal utile. On utilise la notation polaire pour $S(f)$:

$$S(f) = A(f)e^{j\psi(f)} \quad \text{où } A(f) > 0 \quad \text{et} \quad \psi(f) \in [0\ 2\pi[\tag{1.3}$$

Et on estime séparément $A(f)$ et $\psi(f)$, et le signal non bruité estimé s'écrit alors :

$$\hat{S}(f) = \hat{A}(f)e^{j\hat{\psi}(f)} \tag{1.4}$$

Estimateur de phase Il a été démontré dans [Ephraim and Malah, 1984] que la phase estimée du signal utile au sens des moindres carrés est égale à la phase du signal observé, soit $\hat{\psi}(f) = \arg X(f)$.

Estimateur basé sur la log-amplitude spectrale Pour les applications de parole, il est intéressant de considérer une fonction de coût basée sur le logarithme de la puissance du spectre [Gray et al., 1980]. On s'intéresse donc à un estimateur d'amplitude minimisant l'Erreur Quadratique Moyenne (EQM) entre le logarithme de la puissance estimée et le logarithme de la puissance réelle. On cherche donc à minimiser la fonction de coût suivante, où $E[]$ désigne l'espérance mathématique :

$$E\left[|\ln \hat{A}(f) - \ln A(f)|^2 | X(f)\right] \tag{1.5}$$

Cette minimisation se fait en ayant la connaissance uniquement du signal observé $X(f)$. L'estimateur associé s'écrit alors [Ephraim and Malah, 1985] :

$$\hat{A}(f) = e^{E[\ln A(f) | X(f)]} \tag{1.6}$$

On définit les Rapport Signal-à-Bruit (RSB) *a priori*, et *a posteriori* selon

[McAulay and Malpass, 1980] :

$$RSB_{prio}(f) = \frac{\varphi_s(f)}{\varphi_b(f)} \tag{1.7}$$

$$RSB_{post}(f) = \frac{|X(f)|^2}{\varphi_b(f)} \tag{1.8}$$

On définit alors $v(f)$ de la façon suivante :

$$v(f) = \frac{RSB_{prio}(f)}{1 + RSB_{prio}(f)} RSB_{post}(f) \tag{1.9}$$

L'amplitude estimée s'écrit alors :

$$\hat{A}(f) = \underbrace{\frac{RSB_{prio}(f)}{1 + RSB_{prio}(f)} \exp\left(\frac{1}{2}\int_{v(f)}^{\infty} \frac{e^{-t}}{t} dt\right)}_{G_{LSA}} |X(f)| \tag{1.10}$$

Il s'agit d'un gain dépendant des RSB **a priori** et **a posteriori**, que l'on appelle Gain **Log-Spectral Amplitude** (LSA).

Le signal débruité est donné par :

$$\hat{S}(f) = G_{LSA}(f) X(f) \tag{1.11}$$

Remarque 1: On a jusqu'ici supposé que $\varphi_s(f)$ et $\varphi_b(f)$ étaient connus. En pratique, il est nécessaire de les estimer. On peut estimer $\varphi_b(f)$, en supposant le bruit ambiant stationnaire, sur les périodes où le locuteur ne parle pas (seul le bruit est présent). $\varphi_s(f)$ est alors estimé en exploitant la non-corrélation entre la parole et le bruit :

$$\hat{\varphi}_s(f) = \hat{\varphi}_x(f) - \hat{\varphi}_b(f) \tag{1.12}$$

$\hat{\varphi}_x(f)$ étant l'estimée de la DSP du signal observé.

1.3.3 Extension aux systèmes multicanaux

Modèle

On suppose que l'on dispose de M microphones, numérotés m de 1 à M. Chacun correspond à un signal d'entrée $x_m(t)$.

Le modèle pour ces signaux est :

$$x_m(t) = \{h_m \otimes s\}(t) + b_m(t) \tag{1.13}$$

$h_m(t)$ est la réponse impulsionnelle correspondant au canal acoustique entre la source de signal utile et le microphone m, $s(t)$ est le signal utile (la parole) et $b_m(t)$ est le bruit additionnel présent sur l'entrée m.

Ce mélange est présenté dans la Figure 1.12.

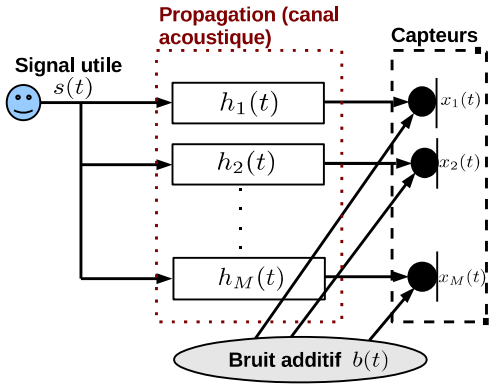


Figure 1.12 – Modèle d'acquisition

Cela peut se réécrire fréquentiellement :

$$X_m(f) = H_m(f)S(f) + B_m(f) \tag{1.14}$$

Pour traiter l'ensemble des entrées, nous utiliserons la notation vectorielle compacte suivante :

$$X(f) = H(f)S(f) + B(f) \tag{1.15}$$

où $S(f)$ et $B(f)$ sont supposés circulaires gaussiens, centrés et non-corrélés. On note $\varphi_s(f)$ la variance de $S(f)$ et $\Sigma_b(f)$ la matrice spectrale de $b(t)$. $H(f)$ est un vecteur déterministe de dimension M.

Estimation séparée de phase et d'amplitude spectrale

On s'intéresse de nouveau à l'estimation séparée des termes d'amplitude et de phase dans l'écriture $S(f) = A(f)e^{j\psi(f)}$.

Estimateur de phase L'estimateur optimal de la phase s'écrit ici :

$$\hat{\psi}(f) = E[\psi(f)|\, X(f)] \tag{1.16}$$

Or, il a été montré dans [Balan and Rosca, 2002] que cette espérance conditionnelle se réécrit de la façon suivante :

$$E[\psi(f)|\, X(f)] = E[\psi(f)|\, T(X(f))] \tag{1.17}$$

avec :

$$T(X(f)) = \frac{H(f)^H \Sigma_b(f)^{-1} X(f)}{H(f)^H \Sigma_b(f)^{-1} H(f)} \tag{1.18}$$

L'estimation de la phase se généralise donc ainsi, dans le cas multi-capteurs :

$$\hat{\psi}(f) = \arg \quad T(X(f)) \tag{1.19}$$

La phase optimale en ce sens est donc la phase du signal $T(X(f))$.

Estimateur d'amplitude spectrale Le problème pour le débruitage basé sur la log-amplitude spectrale s'écrit dans le cas multi-capteurs :

$$\hat{A}(f) = e^{E[\ln A(f) | X(f)]} \tag{1.20}$$

De même que pour la phase, cela se réécrit de la façon suivante :

$$e^{E[\ln A(f) | X(f)]} = e^{E[\ln A(f) | T(X(f))]} \tag{1.21}$$

Le signal débruité est donné par :

$$\hat{S}(f) = \underbrace{\frac{RSB_{prio}(f)}{1 + RSB_{prio}(f)} \exp\left(\frac{1}{2} \int_{v(f)}^{\infty} \frac{e^{-t}}{t} dt\right)}_{\text{Post-Filtre: } G_{LSA}} \frac{H(f)^H \Sigma_b(f)^{-1} X(f)}{H(f)^H \Sigma_b(f)^{-1} H(f)} \tag{1.22}$$

avec cette fois-ci :

$$RSB_{prio}(f) = \varphi_s(f) H(f)^H \Sigma_b(f)^{-1} H(f) \tag{1.23}$$

$$RSB_{post}(f) = |T(X(f))|^2 H(f)^H \Sigma_b(f)^{-1} H(f) \tag{1.24}$$

$$v(f) = \frac{RSB_{prio}(f)}{1 + RSB_{prio}(f)} RSB_{post}(f) \tag{1.25}$$

car la variance du bruit sur le signal $T(X(f))$ vaut $\left[H(f)^H \Sigma_b(f)^{-1} H(f)\right]^{-1}$.

Remarque 2: Considérons le problème de minimisation de l'énergie du bruit résiduel par un filtre linéaire, sous la contrainte de ne pas distordre le signal utile. En reprenant les notations de l'Équation (1.13), on écrit :

$$\hat{W}(f) = \underset{G}{\operatorname{argmin}} E\left[|G^H X(f)|^2\right] \quad \text{sous la contrainte} \quad G^H H(f) = 1 \tag{1.26}$$

La solution est $\hat{W}(f) = \frac{\Sigma_b(f)^{-1} H(f)}{H(f)^H \Sigma_b(f)^{-1} H(f)}$,

et le signal en sortie de ce filtrage est $T(X(f))$. Ce filtrage s'appelle le **beamforming** MVDR [Capon, 1969, Habets et al., 2010a, Habets et al., 2010b]. L'estimateur multi-capteur consiste donc en un **beamforming** MVDR, suivi d'un traitement mono-capteur.

1.3.4 Incertitude sur la présence de parole

On a jusqu'ici considéré que les points temps-fréquence auxquels on s'intéresse contiennent toujours de la parole. Or ce n'est pas exact : il y a des trames durant lesquelles le locuteur ne parle pas (ce sont alors des trames de bruit seul), mais même lorsque une trame contient de la parole, celle-ci n'est pas présente à toutes les fréquences, comme illustré sur la Figure 1.13.

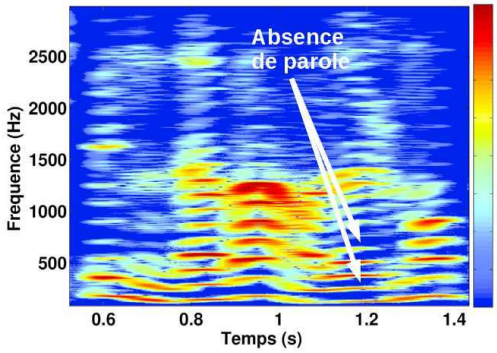


Figure 1.13 – Occupation du plan temps-fréquence par le signal de parole - Les parties bleues sont des zones où la parole n'est pas présente

On introduit alors de l'incertitude sur la présence du signal utile, et on définit les hypothèses de présence ou d'absence de signal utile [Ephraim and Malah, 1984, Cohen, 2001]

$$\begin{aligned} H_1^f &: X(f) = S(f) + B(f) \\ H_0^f &: X(f) = B(f) \end{aligned} \tag{1.27}$$

H_1^f est l'hypothèse "le signal utile est présent à la fréquence f " et H_0^f est l'hypothèse "le signal utile est absent à la fréquence f ".

Probabilité conditionnelle de présence de parole

On s'intéresse alors à la probabilité conditionnelle de présence de parole $p(H_1^f|X(f))$.

En utilisant la règle de Bayes, et sous hypothèses gaussiennes pour le bruit et le signal utile, cette probabilité s'écrit [Cohen, 2001] :

$$p(H_1^f|X(f)) = \left[1 + \frac{p(H_0^f)}{1-p(H_0^f)}(1+RSB_{prio}(f))e^{-RSB_{post}(f)\frac{RSB_{prio}(f)}{1+RSB_{prio}(f)}}\right]^{-1} \tag{1.28}$$

où $p(H_0^f)$ désigne la probabilité d'absence de parole a priori que l'on peut estimer par différentes approches [Cohen, 2001, Malah et al., 1999, Cohen and Berdugo, 2003].

Modification de l'estimateur de log-amplitude spectrale

On peut, pour toute fonction ρ de $S(f)$, écrire un estimateur de cette fonction en prenant en compte l'incertitude sur la présence de parole. Cela s'écrit :

$$E[\rho(S(f))|X(f)] = p(H_0^f)E\left[\rho(S(f))|X(f), H_0^f\right] + p(H_1^f)E\left[\rho(S(f))|X(f), H_1^f\right] \tag{1.29}$$

Pour l'estimateur basé sur la log-amplitude spectrale, cela s'écrit :

$$e^{E[\ln|S(f)||X(f)]} = e^{p(H_0^f)E\left[\ln|S(f)||X(f),H_0^f\right]}\, e^{p(H_1^f)E\left[\ln|S(f)||X(f),H_1^f\right]} \tag{1.30}$$

Or, lorsque l'on suppose une absence de parole (seul le bruit est présent), on peut penser intuitivement que le signal débruité doit être nul, pour supprimer tout le bruit. En pratique, il est plus judicieux d'appliquer un gain G_{min} sur le signal observé, pour conserver un côté naturel du bruit et minimiser l'apparition d'artefacts [Cappé, 1994], que l'on appelle bruit musical. Ainsi, la partie $e^{E[\ln S(f)|X(f),H_0^f]}$ s'écrit :

$$e^{E[\ln |S(f)||X(f),H_0^f]} = G_{min}|X(f)| \tag{1.31}$$

Et l'estimateur complet donne [Cohen, 2002] :

$$\begin{aligned} e^{E[\ln |S(f)||X(f)]} &= e^{E[\ln |S(f)||X(f),H_1^f]^p} G_{min}^{1-p} \\ &= G_{LSA}(f)^p G_{min}^{1-p}|X(f)| \end{aligned} \tag{1.32}$$

$G_{LSA}(f)$ étant le gain LSA et $p = p(H_1^f|X(f))$. On appelle ce gain le Gain Optimally Modified Log-Spectral Amplitude (OM-LSA). Ce gain peut également s'appliquer dans le cas multi-capteurs [Gannot and Cohen, 2004].

1.4 Étage multi-capteurs

Les méthodes présentées dans cette section ont été historiquement développées dans le cadre du beamforming GSC, basé sur l'obtention de références de bruit en annulant le signal de parole. En pratique, des références parfaites de bruit sont difficiles à obtenir, ce qui gêne le débruitage final par ANC (voir Figure 1.4, page 13). C'est pourquoi des méthodes plus robustes ont été développées. Nous allons présenter brièvement les défauts du GSC, avant de nous intéresser à deux méthodes d'annulation de bruit multi-capteurs ayant permis de rendre cette structure plus robuste.

Limitations du GSC

Dans le cas du GSC, les références de bruit s'obtiennent à partir de l'estimation de la propagation du signal de parole (H en reprenant les notations de Section 1.3.3). En pratique, il est difficile d'obtenir une estimation parfaite des transferts modélisant la propagation de la parole, même si des implémentations adaptatives permettent de l'apprendre [Choi et al., 2007].

Si l'on a une estimation présentant une erreur, cela provoque une présence de parole sur les références de bruit : on a des "fuites" de parole. Cela vient perturber l'annulation de bruit, et provoque une distorsion de la parole en sortie du traitement. Ces défauts seront d'autant plus marqués que le RSB sur la référence de bruit est grand [Widrow et al., 1975].

Pour compenser ces défauts, il est possible de remplacer l'ANC par une Annulation de bruit adaptative résistante aux fuites de Parole (Crosstalk Resistant Adaptive Noise Cancellation) (CR-ANC) [Lepauloux, 2010], ou d'utiliser un dérivé du MWF [Spriet et al., 2004]. Nous allons présenter ces deux systèmes d'annulation de bruit.

1.4.1 Annulation de bruit résistante aux fuites de parole

Les hypothèses pour l'annulation de bruit sur les signaux d'entrée sont :

— L'une des entrées est composée du signal utile à débruiter et d'une composante de bruit.

— Les autres entrées sont composées d'une version filtrée de cette composante de bruit : les bruits captés sont totalement cohérents d'un capteur à l'autre.

Cela s'écrit, dans le cas où l'on dispose de deux entrées :

$$\begin{aligned} x_1(t) &= s(t) + b(t) \\ x_2(t) &= \{h_b \otimes b\}(t) \end{aligned} \tag{1.33}$$

$h_b(t)$ étant un transfert modélisant la propagation du bruit.

Ici, on relâche l'hypothèse d'absence de signal de parole sur la deuxième entrée :

$$\begin{aligned} x_1(t) &= s(t) + b(t) \\ x_2(t) &= \{h_s \otimes s\}(t) + \{h_b \otimes b\}(t) \end{aligned} \tag{1.34}$$

$h_s(t)$ étant un transfert modélisant la propagation de la parole.

Durant les périodes d'absence de parole, cela devient :

$$\begin{aligned} x_1(t) &= b(t) \\ x_2(t) &= \{h_b \otimes b\}(t) \end{aligned} \tag{1.35}$$

En adaptant le filtrage adaptatif de la Figure 1.5 (de la page 14) uniquement sur les phases de bruit seul, on parvient donc à converger vers une bonne estimation de h_b [Toner and Campbell, 1993, Harrison et al., 1986]. On peut pour cela utiliser un Détecteur d'activité vocale (Voice Activity Detector) (VAD) [Ramirez et al., 2005].

En utilisant cette estimation dans les périodes de parole (en supposant le champ de bruit stable, c'est à dire que h_b est constant), la sortie du traitement s'écrit :

$$\hat{S}(f) = X_1(f) - \frac{X_2(f)}{H_b(f)} = S(f)\left(1 - \frac{H_s(f)}{H_b(f)}\right) \tag{1.36}$$

On obtient donc un signal exempt de bruit, mais comportant un signal utile distordu dans le cas où la deuxième entrée contient de la parole [Gardner and Agee, 1980].

Pour compenser cette distorsion, on utilise un second filtrage adaptatif, qui n'est actif que durant les périodes où la parole est présente [Madhavan and De Bruin, 1990]. Le schéma est présenté dans la Figure 1.14.

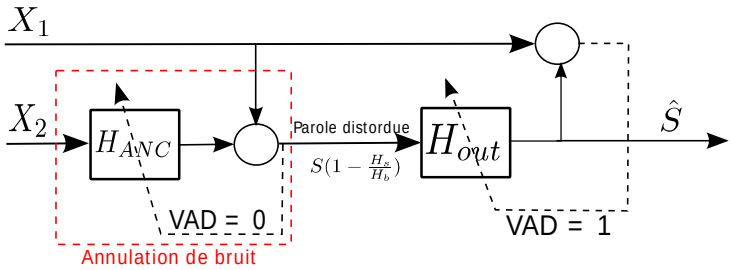


Figure 1.14 – CR-ANC. VAD est une variable contrôlant l'adaptation des filtres. Elle vaut 1 en présence de parole, et 0 lors des périodes de bruit seul.

L'adaptation du second filtre H_{out} converge vers :

$$\hat{H_{out}}(f) = \frac{E\left[[S(f)+B(f)]S(f)\left[1-\frac{H_s(f)}{H_b(f)}\right]\right]}{E\left[|S(f)\left[1-\frac{H_s(f)}{H_b(f)}\right]|^2\right]} = \left[1-\frac{H_s(f)}{H_b(f)}\right]^{-1} \tag{1.37}$$

Et la sortie de ce filtrage donne bien $S(f)$, le signal utile.

Notons que l'on peut également utiliser ce principe d'annulation de bruit et compensation de distorsion avec une antenne comprenant plus de 2 capteurs [Zeng and Abdulla, 2006].

1.4.2 SDW-MWF

On reprend ici le modèle donné dans l'Équation (1.13), puisque l'on suppose que l'on a de la parole sur toutes les entrées :

$$x_m(t) = \{h_m \otimes s\}(t) + b_m(t) \tag{1.38}$$

et l'on note $s_m(t) = \{h_m \otimes s\}(t)$ la parole utile présente sur le capteur m.

On cherche à identifier le bruit sur l'entrée 1, en appliquant un filtre de taille L sur chacune des entrées.

Cela revient à estimer un filtre w de taille ML, de la façon suivante :

$$\hat{w} = \underset{g}{\operatorname{argmin}}\, E\left[|b_1(t) - g^H x(t)|^2\right] \tag{1.39}$$

Avec les notations :

$$x(t) = [x_1(t)^H \quad x_2(t)^H \quad \cdots \quad x_M(t)^H]^H \tag{1.40}$$

avec :

$$x_m(t) = [x_m(t-L+1) \quad \cdots \quad x_m(t)]^H \tag{1.41}$$

Les vecteurs $s_m(t)$, $b_m(t)$, $s(t)$ et $b(t)$ sont définis de la même façon.

A noter que chaque x_m est de taille $L \times 1$, alors que $x(t)$ est de taille $ML \times 1$.

Le signal estimé s'écrit alors : $\hat{s}(t) = x_1(t) - \hat{w}^H x(t)$. Cela revient à prédire le bruit présent sur l'entrée 1 et à le retrancher : la formulation est la même que pour l'ANC.

L'écart quadratique moyen de l'Équation (1.39) peut s'écrire comme la somme de deux termes, en s'appuyant sur la décorrélation entre bruit et signal utile :

$$\hat{w} = \underset{g}{\operatorname{argmin}} \underbrace{E\left[|b_1(t) - g^H b(t)|^2\right]}_{e_b} + \underbrace{E\left[|g^H s(t)|^2\right]}_{e_s}; \tag{1.42}$$

e_s est l'apport de distorsion sur le signal utile, et e_b est le bruit résiduel. On peut pondérer ces erreurs selon que l'on préfère avoir plus de distorsion ou plus de bruit résiduel. Le problème devient [Spriet et al., 2004, Florencio and Malvar, 2001] :

$$\hat{w_\rho} = \underset{g}{\operatorname{argmin}} \left[E\left[|b_1(t) - g^H b(t)|^2\right] + \rho E\left[|g^H s(t)|^2\right]\right] \tag{1.43}$$

La solution est :

$$\hat{w}_\rho = \left[\rho E\left[s(t)s(t)^H\right] + E\left[b(t)b(t)^H\right]\right]^{-1} E\left[b(t)b_1^*(t)\right] \quad (1.44)$$

ρ étant un terme de pondération positif. L'interprétation est assez simple, en terme de performances :

— Plus ρ est petit, plus on accorde d'importance au bruit résiduel, au prix d'une distorsion plus forte sur le signal utile. Si ρ vaut zéro, le filtre w vaut $\delta(t)$: la sortie du filtre vaut alors $x_1(t)$, et le signal estimé est nul.

On peut estimer les matrices $E\left[s(t)s(t)^H\right]$ et $E\left[b(t)b(t)^H\right]$ en exploitant les périodes de signal où seul le bruit est présent, et en utilisant la décorrélation entre parole et bruit [Spriet et al., 2004].

1.5 Synthèse

Nous avons présenté dans ce Chapitre un panel de méthodes de réduction de bruit mono et multi-capteurs couramment utilisées dans le cadre du débruitage de parole. La présentation du contexte de l'étude effectuée ensuite, conjointement avec la présentation de l'environnement automobile, nous a permis de mettre en évidence la stratégie mise en œuvre dans la suite : il s'agit de développer un étage de filtrage multi-capteur, qui permettra d'obtenir un signal de parole mono-capteur déjà débruité. Nous appliquerons alors sur ce signal l'algorithme de débruitage déjà utilisé par Parrot, basé sur un gain OM-LSA.

Pour l'étage multi-capteurs, trois pistes ont été évoquées. Ce sont trois filtrages linéaires :

— un beamforming MVDR, qui étend naturellement le gain OM-LSA au cas multi-capteurs,

— un système basé sur l'ANC, robuste à la présence de parole sur tous les capteurs,

— un dérivé du MWF permettant de pondérer le compromis entre bruit résiduel et distorsion, le SDW-MWF.

Chapitre 2

Environnement acoustique

Sommaire

Ce Chapitre présente les spécificités acoustiques de l'environnement automobile. Dans le cadre de l'étude menée, une grande campagne de mesures acoustiques a été effectuée. Ce Chapitre présente la façon dont a été faite cette campagne, notamment les capteurs utilisés. Nous présentons ensuite les caractéristiques spectrales et spatiales que nous avons mises en évidence à l'aide de ces mesures, sur le bruit ambiant en automobile et sur le signal de parole.

Nous cherchons à définir quelles sont les propriétés acoustiques du bruit et du signal utile que nous pouvons exploiter afin de définir conjointement :

— des méthodes de débruitage multi-canal adaptées à l'environnement,
— une architecture acoustique cohérente avec ces méthodes, notamment pour le placement des capteurs, afin de créer une antenne de microphones optimale.

Des campagnes de mesure dans l'habitacle automobile ont donc été réalisées, pour caractériser certaines propriétés temporelles et spatiales du bruit et de la voix. Nous allons tout d'abord présenter les systèmes d'acquisition utilisés pour réaliser ces mesures, les bases audio ayant servi au développement et à l'évaluation des algorithmes, ainsi que les différents microphones utilisés. Nous nous intéresserons ensuite aux diverses sources acoustiques présentes dans l'habitacle automobile, bruit et parole. Des propriétés du bruit ambiant automobile seront alors exposées. Enfin, nous montrerons des propriétés du signal de parole, avant de conclure sur les propriétés exploitables du bruit et de la parole.

2.1 Système d'acquisition

Les enregistrements des diverses campagnes de mesure ont été faits dans un habitacle de voiture (Peugeot 207), dans des conditions réelles. C'est-à-dire que le système d'enregistrement était embarqué dans l'automobile pour capter des signaux correspondant aux conditions réelles de bruit sur route, pendant que l'un des expérimentateurs conduisait. Une grande partie des mesures a été faite sur autoroute, car cela correspond à une condition d'intérêt, du fait de son importance pratique (c'est la condition d'utilisation la plus courante), et de par son intérêt acoustique (c'est la condition qui présente le plus de bruit).

2.1.1 Chaîne globale d'acquisition

Nous présentons ici le système d'enregistrement multi-canal utilisé lors des campagnes d'enregistrement, dans sa globalité.

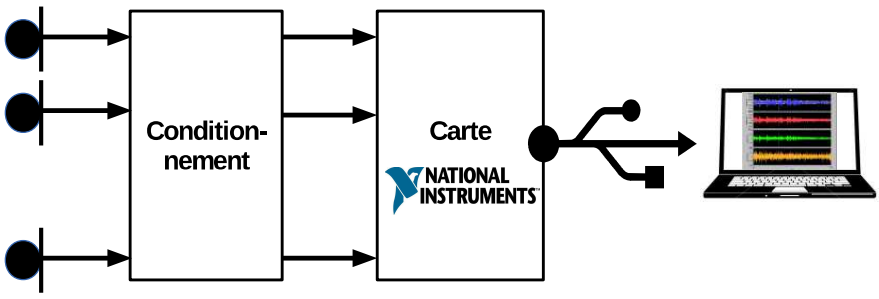


Figure 2.1 – Chaîne d'acquisition globale

La chaîne d'acquisition est composée de plusieurs étages, représentés dans la Figure 2.1. Nous allons décrire le rôle de chaque étage, de gauche à droite :

Capteurs Ce sont les microphones qui transforment un champ acoustique en tension électrique.

Conditionnement Ce composant a une double utilité : alimenter les transistors présents dans les microphones (des **Electret Condenser Microphone** (ECM)), et éliminer les composantes continues des signaux. Après cet étage, les signaux sont toujours analogiques.

Carte National Instruments Cette interface permet d'échantillonner les signaux, et de les transmettre vers un port USB après conversion analogique/numérique. L'échantillonnage se fait à 51.2 kHz, et la quantification se fait sur 16 bits, en **Pulse Code Modulation** (PCM) linéaire.

Ordinateur L'ordinateur pilote la carte National Instruments via un logiciel développé en Labview [1], et permet l'enregistrement des signaux au format wav PCM.

Notons que cette chaîne d'acquisition induit un bruit électrique. Celui-ci est un bruit quasi-blanc, avec une Densité Spectrale de Puissance (DSP) de -75 dB sur toute la bande. Tous les niveaux en dB sont donnés par rapport à la dynamique de la carte d'acquisition : celle-ci renvoie des échantillons entre -1 et 1, correspondant à une tension d'entrée allant de -100 mV à 100 mV. Cette dynamique a été choisie de façon à ne pas saturer l'entrée dans nos conditions d'enregistrement, tout en gardant une bonne dynamique de quantification.

Nous allons étudier plus en détail les capteurs que nous utilisons lors de ces acquisitions. Ce système d'acquisition permet d'enregistrer de façon synchrone jusqu'à 8 capteurs simultanément. On peut donc construire des antennes de microphones comportant jusqu'à 8 capteurs. Dans les antennes acoustiques considérées, on peut utiliser soit des capteurs acoustiques omnidirectionnels (captant de manière égale les ondes sonores, quelle que soit leur direction d'arrivée), ou alors des capteurs directionnels, qui vont atténuer les ondes venant de certaines directions par rapport à d'autres. Nous avons au cours de notre étude utilisé principalement deux modèles de capteur, l'un omnidirectionnel et l'autre cardioïde.

L'intérêt des capteurs cardioïdes est de pouvoir privilégier une direction d'arrivée d'onde, en atténuant les ondes venant d'autres directions. En connaissant la position du locuteur, on peut donc faire pointer un capteur unidirectionnel vers cette position : le signal de parole sera donc bien restitué, alors que les bruits venant d'autres directions seront atténués. Il est donc important de caractériser la directivité des capteurs pour savoir dans quelle mesure les bruits alentours seront déjà atténués au niveau du capteur.

2.1.2 Capteur omnidirectionnel

Un microphone omnidirectionnel produit une tension électrique indépendante de la direction d'arrivée de l'onde sonore. Pour vérifier cette propriété, des mesures de directivité sont faites sur ces capteurs : une source sonore est placée à côté du capteur en condition sde réverbération faible (chambre pseudo-anéchoïque), et l'on fait tourner la source autour du capteur pour échantillonner les directions d'arrivée de l'onde sonore, comme illustré dans la Figure 2.2. L'énergie du signal capté dans ces directions donne le diagramme de directivité.

1. http://www.ni.com/labview/f/

Figure 2.2 – Mesure de directivité. Les mesures se font ici tous les 10°

De plus, le signal émis pour la mesure est une rampe en fréquence, de façon à pouvoir tracer ce diagramme à différentes fréquences.

Le capteur omnidirectionnel utilisé est un Veco 6022B-9C403S-7AT2. Ses spécifications sont données dans la Table 2.1.

Type	ECM
Sensibilité	-40 dBV/Pa
Bande passante	10 Hz - 20 kHz

Table 2.1 – Spécifications du Veco 6022B-9C403S-7AT2 - Données constructeur

Son diagramme de directivité est donné dans la Figure 2.3.

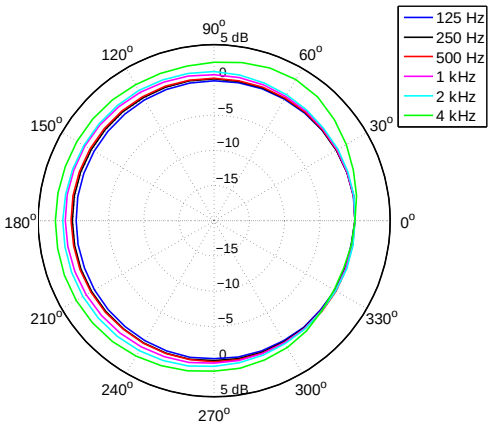


Figure 2.3 – Diagramme de directivité (en dB) du Veco 6022B-9C403S-7AT2, mesuré chez Parrot. Les niveaux sont donnés en dB, et sont normalisés par rapport à la réponse à 0°.

On constate que pour toutes les fréquences, ce diagramme est bien quasi-circulaire, ce qui est attendu pour un capteur omnidirectionnel.

Une autre caractéristique importante pour un capteur est sa réponse en fréquence. Cela indique si le capteur a tendance à atténuer certaines fréquences plutôt que d'autres. Pour la mesurer, une source émet un signal à différentes fréquences, en conservant la même puissance (on peut là aussi se servir d'un signal de type rampe fréquentielle). Il est à noter que la source est émise par un haut-parleur, ayant lui même sa propre réponse en fréquence. Celle-ci est mesurée à l'aide d'un capteur de mesure calibré (ici un GRAS AE46) dont la réponse en fréquence est connue. On peut ensuite compenser la réponse du haut-parleur utilisé. L'énergie du signal capté par le microphone à chaque fréquence donne alors sa réponse en fréquence.

La réponse en fréquence pour le microphone omnidirectionnel utilisé est donnée dans la Figure 2.4.

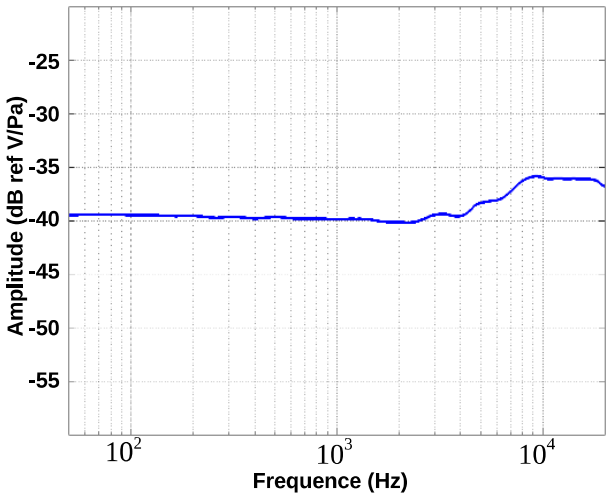


Figure 2.4 – Réponse en fréquence du Veco 6022B-9C403S-7AT2, mesurée chez Parrot

On voit que la réponse en fréquence de ce capteur est plate sur une large bande de fréquences, ce qui est désirable pour ne pas altérer le signal capté. En effet, un capteur ayant une réponse en fréquence trop variable pourra, dans le cas de la parole, affecter le timbre de la voix et la rendre moins naturelle. Par exemple, amplifier fortement les aigus (fréquences hautes) par rapport aux graves rend la voix plus nasillarde.

2.1.3 Capteur cardioïde

Un autre type de capteur est couramment utilisé dans des systèmes d'acquisition : le capteur cardioïde. Ce type de microphone mesure des différences de pression acoustique entre deux points pour créer une direction privilégiée d'acquisition : le but est d'atténuer les ondes acoustiques venant d'autres directions que celle choisie. Ce type de microphone est utile notamment dans le cas où l'on sait de quelle direction vient le signal d'intérêt : on peut améliorer la restitution de celui-ci en le plaçant dans la direction privilégiée, et les ondes parasites venant d'autres directions seront atténuées.

En particulier, dans un champ de bruit diffus (le bruit vient uniformément de toutes les directions), l'atténuation de bruit obtenue par rapport à un microphone omnidirectionnel est de 4.8 dB pour un capteur cardioïde parfait [Kahrs and Brandenburg, 1998].

Nous utilisons un capteur cardioïde, dont la directivité théorique est donnée dans la Figure 2.5.

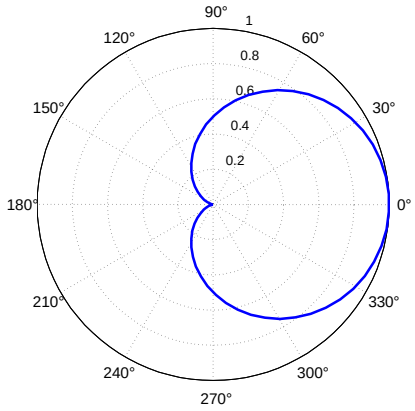


Figure 2.5 – Directivité théorique d'un capteur cardioïde, pour toutes les fréquences, sur une échelle linéaire.

Nous utilisons un capteur Merry 3100 EMC158-010-01, dont les spécifications sont données dans la Table 2.2.

Type	ECM
Sensibilité	-35 dBV/Pa
Bande passante	500 Hz - 20 kHz

Table 2.2 – Spécifications du Merry 3100 EMC158-010-01 - Données constructeur

Le diagramme de directivité mesuré est donné dans la Figure 2.6.

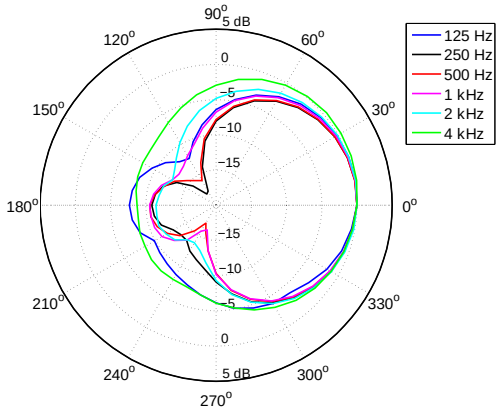


Figure 2.6 – Diagramme de directivité du Merry 3100 EMC158-010-01, mesuré chez Parrot. Les niveaux sont donnés en dB, et sont normalisés par rapport à la réponse à 0°.

On constate que ce diagramme varie fortement en fonction de la fréquence, et que l'on perd en directivité en basses (~ 125 Hz) et hautes fréquences (~ 4 kHz).

La réponse en fréquence de ce microphone, mesurée dans l'axe de directivité (0°), est donnée dans la Figure 2.7.

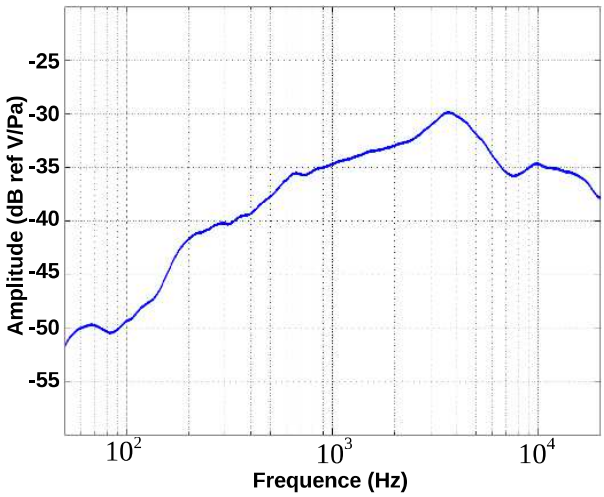


Figure 2.7 – Réponse en fréquence du Merry 3100 EMC158-010-01, mesurée chez Parrot

La réponse en fréquence de ce capteur est moins plate que celle mesurée sur l'omnidirectionnel. Notamment, il y a une forte chute de cette réponse en basses fréquences, ce qui explique les différences de spécification en terme de bande passante entre les deux capteurs.

Ce microphone présente donc une directivité plus avantageuse, mais sera moins neutre dans la mesure acoustique, notamment dans les basses fréquences, où sa réponse en fréquence chute fortement.

2.1.4 Comparaison entre les deux types de capteurs

On a présenté deux types de capteurs acoustiques :

Omnidirectionnel Ce microphone présente une réponse en fréquence plate, mais n'a pas de directivité : on ne peut exploiter la connaissance de la position du locuteur en utilisant un de ces capteurs.

Unidirectionnel Celui-ci présente une bonne directivité, mais dans une bande de fréquence limitée. De plus, sa réponse en fréquence n'est pas plate, et présente notamment une chute en basses fréquences (<500 Hz).

Le choix des capteurs est donc un facteur important dans la prise de son, et il faut bien tenir compte du compromis entre directivité (qui permet de gagner en Rapport Signal-à-Bruit (RSB) d'entrée) et réponse en fréquence (que l'on souhaiterait la plus plate possible). Pour se rendre compte du gain en RSB obtenu par l'utilisation d'un capteur cardioïde, nous avons placé un capteur omnidirectionnel et un cardioïde dans l'habitacle de la voiture, au niveau du plafonnier (voir Figure 2.9). Remarquons que dans cette manipulation, l'environnement n'est plus anéchoïque. Le niveau de bruit enregistré en utilisant un seul de ces capteurs a été déterminé en enregistrant du bruit seul, lorsque la voiture roulait sur autoroute, et le niveau de signal utile a été déterminé en plaçant un haut-parleur diffusant une source large bande (un bruit blanc) à la place du conducteur, alors que la voiture était dans un parking silencieux. On détermine alors l'énergie du bruit et du signal utile en par une méthode de Welch [Welch, 1967], sur des fenêtres

de 512 échantillons (soit 32 ms), recouvrantes à 50%. L'estimation se fait sur des signaux de 5 s. A partir de ces niveaux de bruit et de signal utile, on en déduit le RSB sur les capteurs, que l'on montre dans la Figure 2.8.

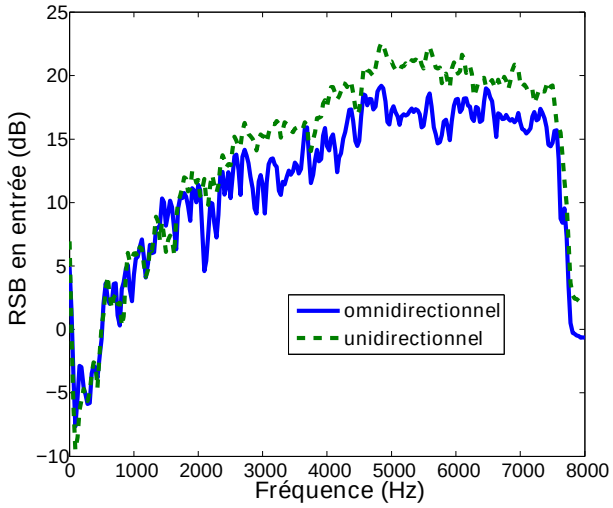


Figure 2.8 – RSB d'entrée pour un microphone unidirectionnel et un microphone omnidirectionnel

Grâce à sa directivité, on constate que le capteur cardioïde améliore grandement le RSB d'entrée, surtout à partir de 1 kHz. En dessous, le diagramme de directivité est moins favorable, et cet avantage s'estompe. Il est donc intéressant d'utiliser des capteurs cardioïdes lorsque l'on s'intéresse aux hautes fréquences.

2.2 Sources acoustiques en environnement automobile

Le signal de parole est produit par une unique source (la bouche du conducteur), alors que le bruit est constitué de plusieurs composantes. Parmi celles-ci, on trouve notamment :

— le moteur,
— le bruit de roulement (contact entre la route et les pneumatiques),
— le bruit aérodynamique (vent produit par la vitesse),
— la circulation environnante (véhicules à proximité),
— d'autres sources, comme des coups de klaxon ou des sirènes.

Nous allons présenter brièvement quelques-unes de ces sources. On s'intéresse notamment au contenu spectral de ces sources au cours du temps (le spectrogramme), mais aussi à une caractéristique spatiale, la Mean-Squared Coherence (MSC) [Cook et al., 1955], définie pour deux processus $x(t)$ et $y(t)$ stationnaires et d'intercorrélation stationnaire par :

$$\mathrm{MSC}_{xy}(f) = \frac{|\varphi_{xy}(f)|^2}{\varphi_x(f)\varphi_y(f)} \tag{2.1}$$

si $\varphi_x(f) \neq 0$ et $\varphi_y(f) \neq 0$. $\varphi_{ab}(f)$ désigne la transformée de Fourier de la fonction d'intercovariance des signaux $a(t)$ et $b(t)$, appelée interspectre :

$$\varphi_{ab}(f) = \int_{-\infty}^{+\infty} E\left[a(t)b(t-\tau)\right] e^{-2i\pi f \tau} d\tau \tag{2.2}$$

et $\varphi_x(f)$ désigne la DSP de $x(t)$. La MSC est une fonction réelle à valeurs dans $[0;1]$. La MSC représente quelle proportion du signal x est liée linéairement à y [Goulding and Bird, 1990]. En effet, si l'on écrit :

$$x(t) = \{h \otimes y\}(t) + v(t) \tag{2.3}$$

où $v(t)$ et $y(t)$ sont deux processus non-corrélés, alors :

$$\mathrm{MSC}_{xy}(f) = \frac{|H(f)|^2 |\varphi_y(f)|^2}{(|H(f)|^2 \varphi_y(f) + \varphi_v(f))\,\varphi_y(f)}. \tag{2.4}$$

On peut notamment dire que cette MSC vaut 1 sur toute la bande de fréquence si et seulement si $x(t) = \{h \otimes y\}(t)$. Des éléments sur la relation entre la MSC des processus $x(t)$ et $y(t)$, et la matrice spectrale du processus $\begin{pmatrix} x(t) \\ y(t) \end{pmatrix}$ sont présentés dans l'Annexe A.

Ici, nous nous intéressons à cette fonction lorsque x et y sont les signaux enregistrés par deux capteurs différents.

Estimation de la MSC

On utilise pour calculer la MSC un estimateur de la DSP des deux signaux considérés et un estimateur de l'interspectre, basés sur la méthode de Welch [Welch, 1967]. Cette méthode consiste à moyenner le périodogramme calculé sur des trames recouvrantes successives.

Ainsi, si l'on dispose de deux signaux $a(t)$ et $b(t)$ stationnaires et d'intercovariance stationnaire, séparés en N fenêtres recouvrantes, et si l'on note les Transformée de Fourier Discrète (TFD) de la trame n des signaux a et b à la fréquence discrète f, $A_n(f)$ et $B_n(f)$, respectivement, l'interspectre de l'équation (2.2) est estimé de la façon suivante :

$$\hat{\varphi}_{ab}(f) = \frac{1}{N} \sum_{n=0}^{N-1} A_n(f) B_n^*(f). \tag{2.5}$$

Lorsque les signaux sont stationnaires, cet estimateur est asymptotiquement non-biaisé, et sa variance est inversement proportionnelle au nombre de trames utilisées. Toutefois, la longueur des trames utilisées détermine la résolution fréquentielle : le choix des trames est donc un compromis entre résolution fréquentielle et précision de l'estimation. La MSC estimée est ensuite donnée par :

$$\hat{\mathrm{MSC}}(f) = \frac{\left|\sum_{n=0}^{N-1} A_n(f) B_n^*(f)\right|^2}{\sum_{n=0}^{N-1} A_n(f) A_n^*(f) \sum_{n=0}^{N-1} B_n(f) B_n^*(f)}. \tag{2.6}$$

On remarque que lorsque l'un des signaux a une faible énergie ($\varphi_x(f)$ ou $\varphi_y(f) \sim 0$), l'estimation de cette fonction aura une forte variance. Pour les MSC présentées ici, on a choisi des trames

de 512 échantillons (soit 32 ms, léchantillonnage se faisant à 16 kHz) pondérées par une fenêtre de Hann avec un recouvrement de 50%.

Conditions d'enregistrement

Les enregistrements des sources que l'on présente ici ont été faits dans un véritable habitacle de voiture, avec les capteurs placés comme montré dans la Figure 2.9.

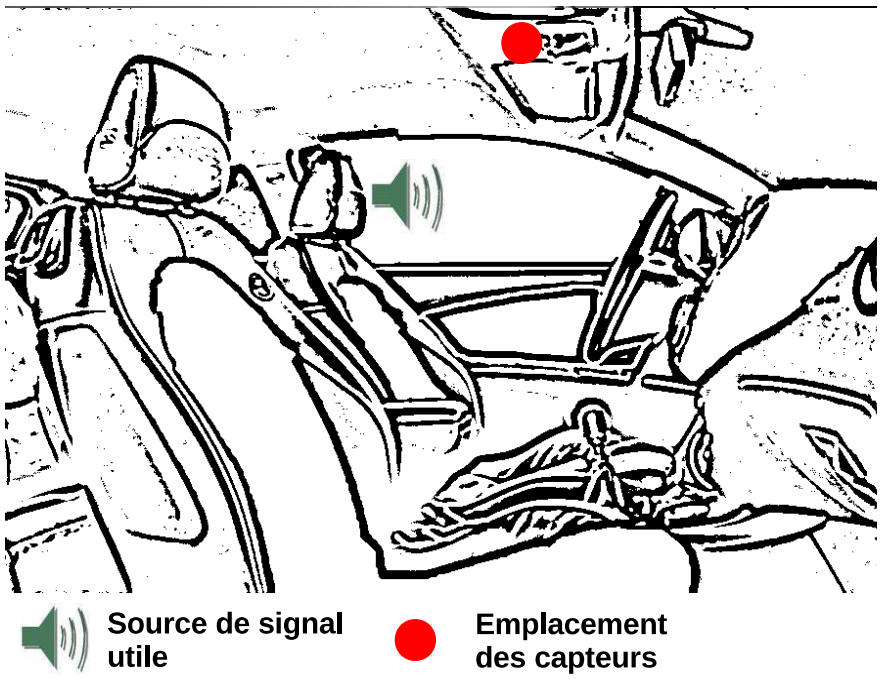


Figure 2.9 – Placement du locuteur et des capteurs dans l'habitacle

Lorsque l'on présente des sources de bruit, la source de signal utile n'est pas active : on est en situation de bruit seul, en roulant avec la voiture. Dans le cas de la parole seule, la source de signal utile est la bouche du conducteur, et la voiture ne roule pas et est placée dans un environnement silencieux (parking calme).

2.2.1 Signal de parole

Le signal de parole est produit par le conducteur de la voiture, placé à environ 50 cm de l'antenne de capteurs (qui est fixée sur le plafonnier, voir Figure 2.9). Cette antenne est composée de deux capteurs omnidirectionnels placés à 4 cm l'un de l'autre.

On présente dans la Figure 2.10 un signal de parole seule, enregistré dans l'automobile à l'arrêt, dans un parking calme, moteur éteint.

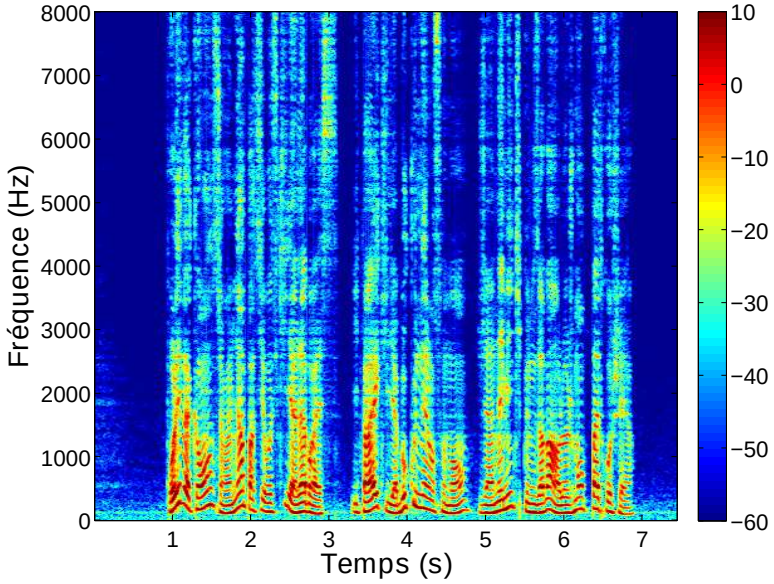


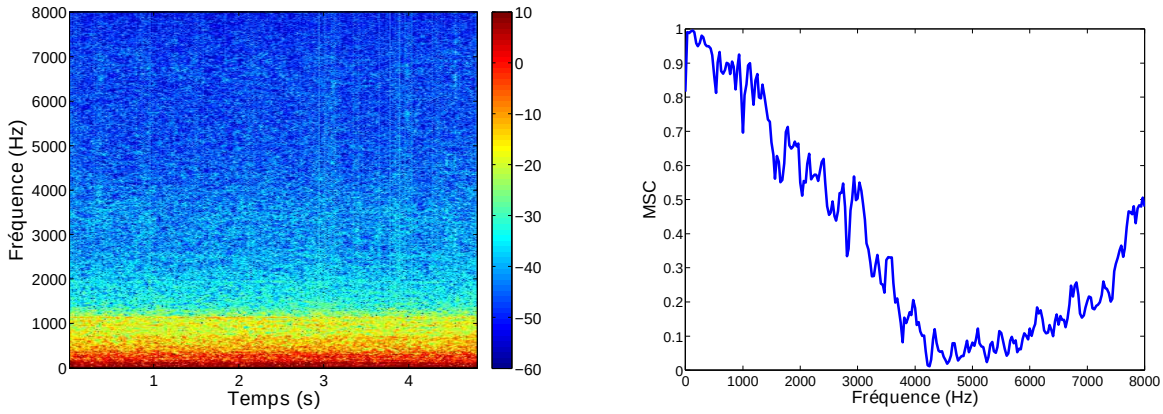
Figure 2.10 – Spectrogramme d'un signal de parole

On remarque que le signal de parole couvre une bande de fréquence assez large, mais perd en énergie en hautes fréquences. Nous reviendrons plus tard sur ses caractéristiques spatiales.

2.2.2 Sources de bruit stationnaires et permanentes

La superposition des bruits de moteur, de roulement et de vent peut souvent être considérée comme stationnaire (notamment lorsque le véhicule roule à vitesse constante). De plus, ces sources sont quasiment tout le temps présentes, alors que les bruits produits par les véhicules environnants (scooter qui double...) ou des sirènes ne sont présents qu'à certains moments bien précis.

On observe la composante de bruit liée au moteur, au roulement et au vent dans une voiture roulant à une vitesse constante de 130km/h. Son contenu spectral et sa MSC, estimée sur un signal de 1s, (pour des capteurs omnidirectionnels placés à 4 cm l'un de l'autre) sont présentés dans la Figure 2.11.



Spectrogramme de bruit d'autoroute (en dB) MSC de bruit d'autoroute entre les deux capteurs

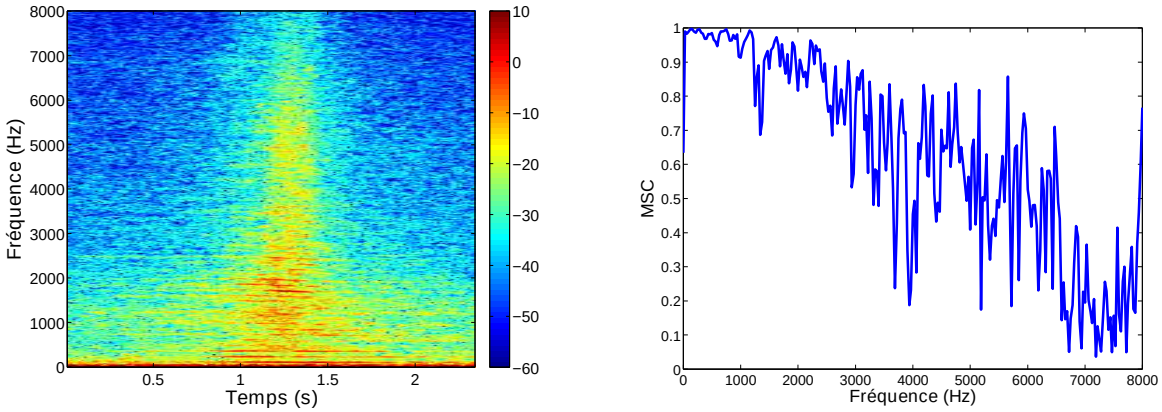
Figure 2.11 – Bruit stationnaire typique d'autoroute

Notons que l'énergie mesurée en hautes fréquences descend jusqu'à environ -50 dB, soit 25 dB au dessus du bruit électrique dû au système d'acquisition. On constate que le contenu spectral de ce bruit varie peu au cours du temps. De plus, le bruit est surtout concentré en basses fréquences, là où la MSC est la plus forte. Nous reviendrons plus tard sur le champ de bruit lié à ces sources, présentant un intérêt particulier puisque ces sources sont tout le temps présentes, et sont prépondérantes sur autoroute, qui est la situation la plus courante.

En comparant avec l'énergie du signal de parole, on peut remarquer que le RSB sur les signaux captés (parole + bruit) varie fortement selon la fréquence : il peut être très faible en basses fréquences (de l'ordre de -3 dB en dessous de 1 kHz), alors que cela s'améliore en plus hautes fréquences (entre 10 et 20 dB).

2.2.3 Autres sources de bruit

On s'intéresse ici aux sources de bruit intermittentes, qui apparaissent sur une période courte par rapport aux temps de convergence des algorithmes. Les exemples pris ici sont le dépassement de la voiture par un scooter, et un coup de klaxon à proximité de la voiture. Les exemples présentés ont été enregistrés en ville, avec les fenêtres ouvertes. Le spectrogramme du dépassement par un scooter, ainsi que la MSC des bruits (captés par deux omnidirectionnels placés à 4 cm l'un de l'autre) liée à cette source sont présentés dans la Figure 2.12. La MSC est estimée sur la période allant de 1s à 1.5s sur le spectrogramme.

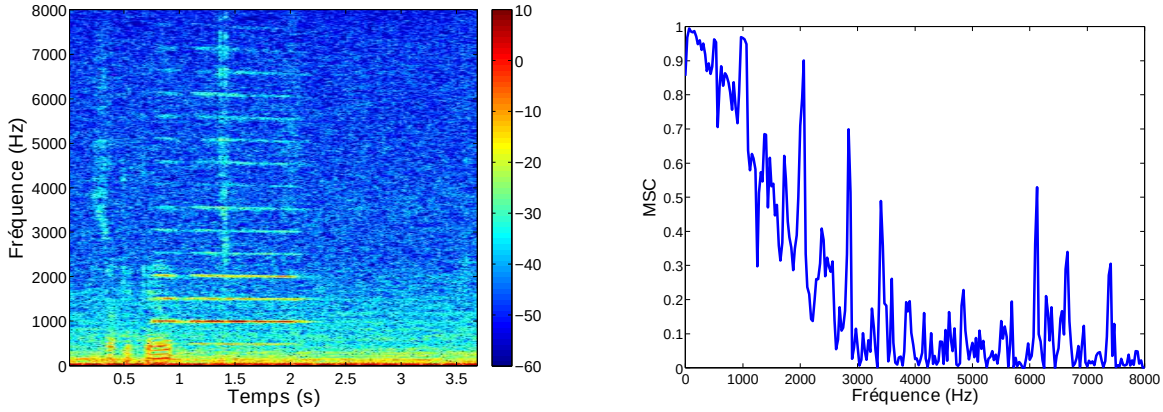


Spectrogramme de bruit de scooter (en dB) MSC de bruit de scooter entre les deux capteurs

Figure 2.12 – Bruit produit par un scooter à proximité

Ce bruit a un contenu spectral qui varie fortement au cours du temps, et son énergie est surtout concentrée en basses fréquences. De plus, ce bruit est très cohérent sur toute la bande de fréquences, comme le montre la forte MSC observée.

Le cas du coup de klaxon est présenté Figure 2.13. Celui-ci vient d'une source qui est plus éloignée que le scooter du cas précédent. Sa MSC est estimée sur la période allant de 1s à 2s sur le spectrogramme.



Spectrogramme de bruit de klaxon (en dB) MSC de bruit de klaxon entre les deux capteurs

Figure 2.13 – Bruit produit par un coup de klaxon à proximité

Le bruit de klaxon est harmonique sur toute la bande de fréquences, et présente des pics de MSC sur les harmoniques, qui diminuent en hautes fréquences. Ceci est dû à la distance plus forte entre l'habitacle où étaient placés les microphones et le véhicule d'où provient ce bruit.

On constate que ces bruit sont instationnaires, cohérents sur une large bande de fréquence, et dans certains cas harmoniques : il est difficile de les distinguer d'un signal de parole et leur débruitage est très difficile. De plus, leur durée est très faible : on ne peut les éliminer qu'à condition de s'autoriser un long délai de traitement, ce qui est impossible pour une application de téléphonie, où le débruitage se fait en temps réel. Nous nous concentrerons dans cette étude sur les bruits stationnaires présents dans la situation la plus courante : le bruit d'autoroute.

2.2.4 Difficultés liées à la téléphonie Wideband

Jusqu'ici, la norme de compression de la voix sur le réseau de téléphonie mobile (norme G.711 de l'International Telecommunication Union - Telecommunication Standardization Sector (ITU-T)) n'autorisait une bande passante n'allant que de 300 Hz à 3400 Hz : la partie hautes fréquences et très basses fréquences du spectre n'étaient pas transmises. La téléphonie Wideband, échantillonnée à 16 kHz au lieu de 8 kHz, tend à se démocratiser. La norme associée (G.722 de l'ITU-T) spécifie une bande passante allant de 50 Hz à 7 kHz, ce qui permet de restituer le spectre en haute et très basses fréquences, ce qui améliore grandement l'intelligibilité [Martin et al., 2008].

Dans le cas de la téléphonie mains-libres en automobile, la bande de fréquences en dessous de 300 Hz contient en général beaucoup de bruit, et la bande au-dessus de 4 kHz contient moins de voix : ce sont des zones où le rapport signal-à-bruit est défavorable, et leur débruitage est donc délicat. De plus, l'intelligibilité globale étant grandement améliorée, le bruit ambiant apporte une gêne moins perceptible, puisqu'il n'empêche plus la compréhension : les interlocuteurs auront plus tendance à désirer une meilleure qualité de voix plutôt qu'une plus grande réduction de bruit. Cette nouvelle norme permet donc une amélioration globale de qualité, mais la prise en compte de l'extension de bande passante amène de nouvelles contraintes quant à l'application des algorithmes de débruitage. Il faut donc prendre en compte cet aspect large bande dans cette étude.

De plus, le bruit d'autoroute étant faiblement cohérent au delà d'une certaine fréquence (voir

Figure 2.11), il est possible lorsque l'on se limite à la bande 300 - 3500 Hz de se mettre dans une situation où l'on peut supposer le bruit faiblement cohérent [Guérin et al., 2003], ce qui n'est pas possible dans le cas du Wideband.

2.3 Champ de bruit

On s'intéresse ici aux propriétés temporelles et spatiales du bruit présent dans un habitacle de voiture, en prenant en compte les capteurs utilisés. En ce qui concerne les propriétés spatiales, on s'intéresse à la cohérence présente entre les bruits enregistrés sur des capteurs différents. Nous allons en particulier présenter le modèle de bruit spatialement diffus couramment utilisé en environnement automobile [Meyer and Simmer, 1997, Guérin et al., 2003] et le comparer aux mesures réalisées. Pour les propriétés temporelles, nous allons nous intéresser aux problématiques de coloration spectrale du bruit. Ces modèles nous permettront de concevoir des simulations pertinentes lorsque nous nous intéresserons à la géométrie des antennes acoustiques considérées, notamment.

2.3.1 Bruit diffus

Nous allons présenter le modèle de bruit diffus sphérique, lorsque l'on utilise des capteurs omnidirectionnels ou cardioïdes. Ce modèle nous permettra de construire des simulations pour évaluer les performances des méthodes de réduction de bruit proposées.

Un champ de bruit diffus sphérique correspond à une situation où le bruit est produit par une infinité de sources indépendantes et de même puissance, placées dans toutes les directions. C'est un champ de bruit que l'on rencontre notamment dans les environnements réverbérants [Cook et al., 1955].

Capteurs omnidirectionnels

Dans le cas de capteurs omnidirectionnels, le signal enregistré par les microphones est simplement la pression acoustique à l'endroit où sont placés les capteurs. Dans ce cas, la MSC théorique pour un champ de bruit diffus sphérique est fonction uniquement de la distance entre les microphones, et s'écrit [Talham, 1981] :

$$\mathrm{MSC}(f) = \left[\mathrm{sinc}\left(\frac{2df}{c_s}\right)\right]^2 \tag{2.7}$$

où d la distance entre les microphones, c_s la célérité du son, f la fréquence et sinc est la fonction sinus cardinal, définie par :

$$\mathrm{sinc}(x) = \frac{\sin(\pi x)}{\pi x} \tag{2.8}$$

Un exemple de MSC pour ce type d'antenne est présenté dans la Figure 2.14.

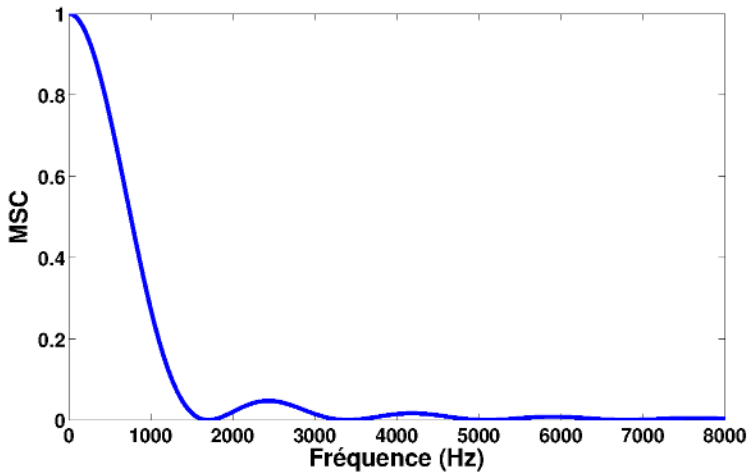


FIGURE 2.14 – Courbe de MSC obtenue pour deux microphones omnidirectionnels placés à 10 cm l'un de l'autre

On remarque que cette MSC est faible en hautes fréquences, et forte en basses fréquences : En particulier, elle atteint 1 en très basses fréquences. En très basses fréquences, on peut donc considérer que le bruit capté par un microphone est le filtré de celui capté par un autre microphone proche.

Capteurs cardioïdes

Pour des capteurs unidirectionnels cardioïdes, on peut calculer cette cohérence en prenant un repère pour situer les capteurs et leur axe de directivité (voir Figure 2.15).

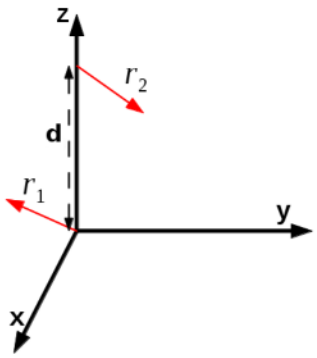


FIGURE 2.15 – Repère utilisé pour le calcul de la cohérence. Les microphones sont placés le long de l'axe z. Les flèches représentent les axes de directivité.

Les vecteurs r_1 et r_2 correspondent aux axes de directivité, et ont pour coordonnées $\begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix}$ et $\begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix}$, respectivement. On note $\delta = \frac{2\pi f d}{c_s}$. Dans ce repère, la MSC dans un cas de bruit diffus

vaut alors [Goulding and Bird, 1990] :

$$MSC(\delta) = \left[\frac{3}{4} \frac{\sin(\delta)}{\delta} + (x_1x_2 + y_1y_2) \left[\frac{\sin(\delta)}{(\delta)^3} - \frac{\cos(\delta)}{(\delta)^2} \right] + z_1z_2 \left[\frac{\sin(\delta)}{(\delta)} + \frac{2\cos(\delta)}{(\delta)^2} - \frac{2\sin(\delta)}{(\delta)^3} \right] + i(z_1 + z_2) \left[\frac{\cos(\delta)}{(\delta)} - \frac{\sin(\delta)}{(\delta)^2} \right] \right]^2 \quad (2.9)$$

La Figure 2.17 présente des exemples de MSC obtenues pour différents placements et orientations, schématisés dans la Figure 2.16.

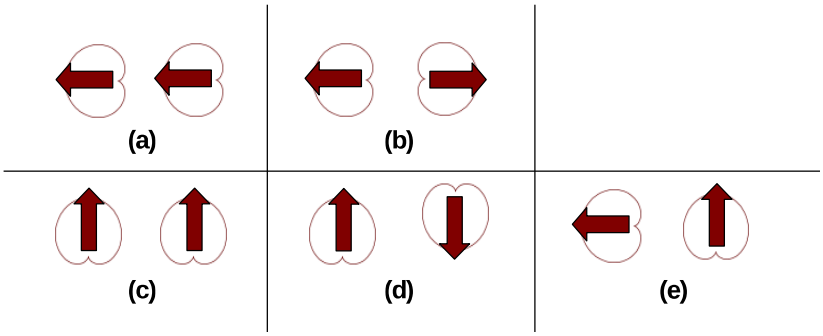


Figure 2.16 – Différents placements envisagés pour les antennes à deux microphones cardioïdes. La flèche représente l'axe de directivité.

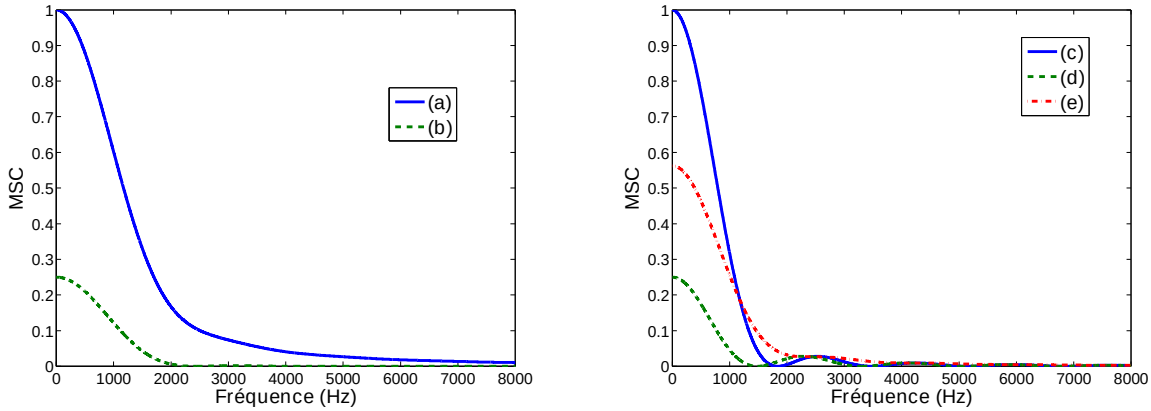


Figure 2.17 – Courbes de MSC théoriques pour des microphones cardioïdes placés à 10 cm de distance, dans les situation décrites dans la Figure 2.16

On remarque que la limite de MSC lorsque l'on s'approche de la fréquence zéro dépend grandement de l'angle entre les axes de directivité des cardioïdes utilisés : plus cet angle est grand, plus la MSC en basses fréquences est basse, le minimum étant atteint lorsque les microphones "pointent" dans des directions opposées.

Antenne mixte omnidirectionnel - cardioïde

Dans le cas d'une antenne composée d'un capteur omnidirectionnel et d'un capteur cardioïde, la cohérence dépend de l'angle entre l'axe de directivité du cardioïde et l'axe sur lequel sont placés les capteurs. La MSC vaut alors [Brandstein and Ward, 2001] :

$$MSC(\delta) = 3\left[\frac{1}{2}\frac{\sin(\delta)}{\delta} + i\frac{\cos(\theta)}{\delta^2}(\delta\cos(\delta) - \sin(\delta))\right]^2 \quad (2.10)$$

où θ est l'angle entre l'axe de placement des microphones et l'axe de directivité du cardioïde. Des cas extrêmes de placement sont illustrés dans la Figure 2.18.

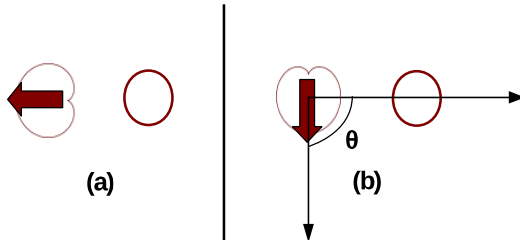


Figure 2.18 – Placements envisagés pour les antennes mixtes. Les cercles sont des capteurs omnidirectionnels.

La Figure 2.19 présente un exemple de MSC obtenue dans ce cas.

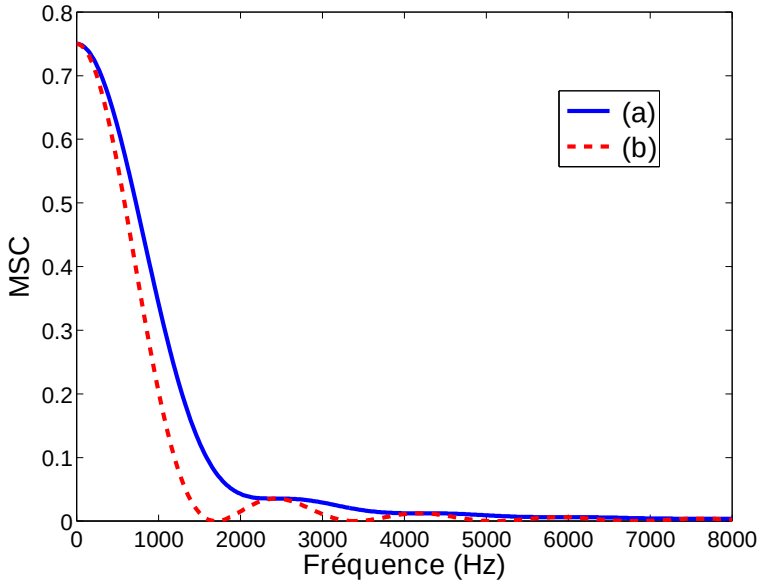


Figure 2.19 – Courbes de MSC obtenues avec un microphone cardioïde et un microphone omnidirectionnel placés à 10 cm l'un de l'autre, pour les orientations données dans la Figure 2.18

On remarque que cette MSC diminue lorsque la fréquence augmente. En très basse fréquencs, elle n'atteint pas 1 : on ne peut pas modéliser les bruits comme complètement cohérents en basses fréquences avec ce type d'antenne.

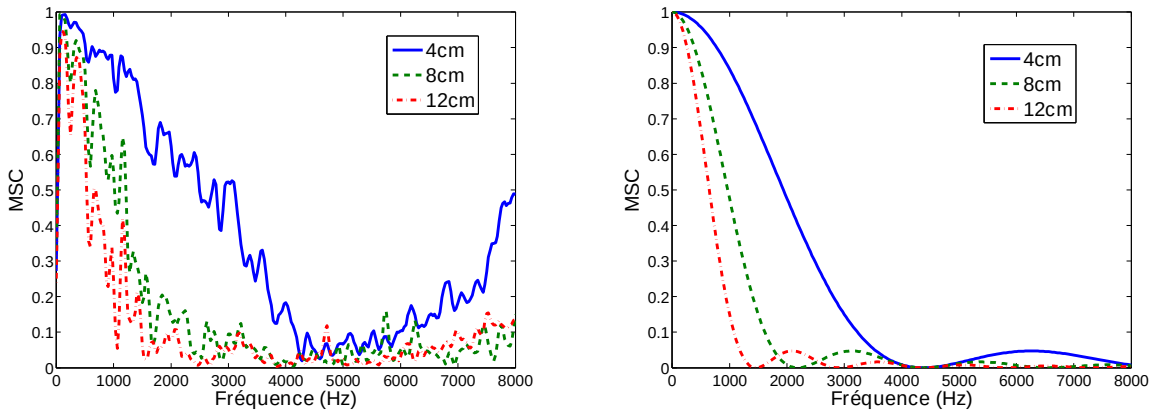
2.3.2 Mesures de cohérence

La MSC entre deux capteurs a été mesurée dans des cas de bruit seul, dans une voiture roulant à 130 km/h sur autoroute. L'enregistrement se fait à une fréquence d'échantillonnage de 16 kHz. Les spectres de puissance et les interspectres ont été estimés par la méthode de Welch [Welch, 1967] sur des fenêtres de 512 échantillons (32 ms), avec un recouvrement de 50%. L'estimation se fait sur un signal d'une durée d'environ 10 s.

On compare les courbes mesurées à celles prédites par le modèle de bruit diffus pour plusieurs types d'antenne :

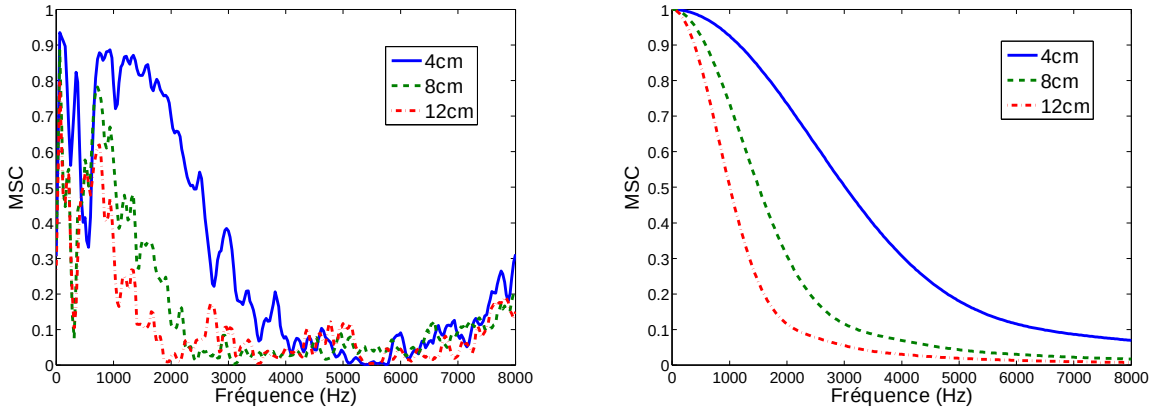
— deux microphones omnidirectionnels (Figure 2.20),
— deux microphones cardioïdes (Figure 2.21),
— un microphone omnidirectionnel et un cardioïde (Figure 2.22).

On utilise pour les courbes théoriques une célérité du son dans l'air égale à $c_s = 340$ m.s^{-1}.



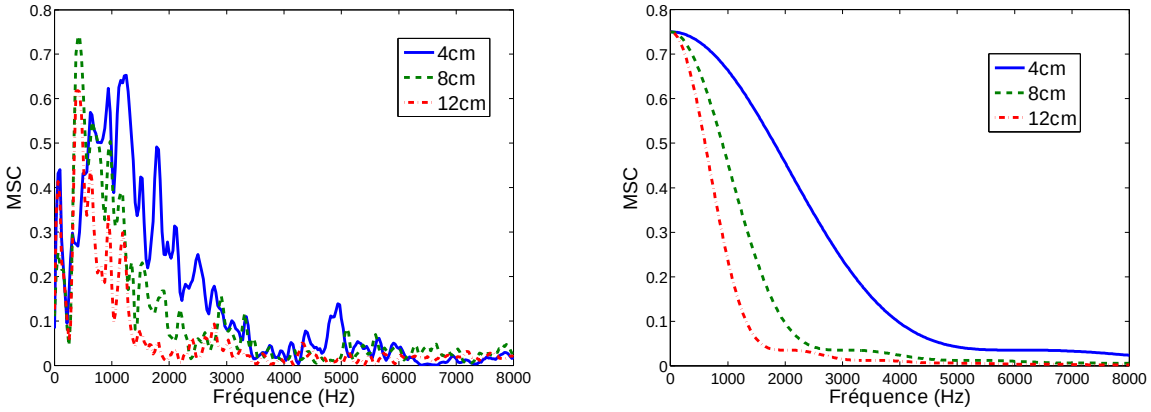
(a) (b)

Figure 2.20 – Courbes de MSC mesurées (a) et théoriques (b) pour des microphones omnidirectionnels, pour plusieurs distances



(a) (b)

Figure 2.21 – Courbes de MSC mesurées (a) et théoriques (b) pour des microphones cardioïdes placés dans la configuration (a) de la Figure 2.16, pour plusieurs distances



(a) (b)

Figure 2.22 – Courbes de MSC mesurées (a) et théoriques (b) pour un microphone cardioïde et un omnidirectionnel placés dans la configuration (a) de la Figure 2.18, pour plusieurs distances

Le modèle est plutôt bien respecté dans le cas des capteurs omnidirectionnels. Dans les autres cas, les mesures présentent une chute de cohérence en dessous de 1 kHz. Ceci peut être lié aux caractéristiques mesurées des capteurs cardioïdes. En effet, lorsque les signaux considérés ont peu d'énergie dans certaines fréquences, l'estimateur de MSC aura une forte variance. Or, comme on l'a souligné dans la Section 2.1.3, les capteurs cardioïdes ont une chute de réponse en fréquence dans le bas du spectre : dans ces fréquences, les signaux captés par ces microphones auront une faible énergie. De plus, la directivité pour ces fréquences n'étant pas spécifiée par les constructeurs de microphones, il peut y avoir un grande variabilité d'un capteur à l'autre.

Dans tous les cas, on observe que l'on a un maximum de cohérence en basses fréquences, alors que cette cohérence chute en hautes fréquences. Ce maximum de cohérence dépend du type de capteurs utilisé, ainsi que de leur placement.

En utilisant une antenne appropriée (2 omnidirectionnels, ou 2 unidirectionnels pointant dans la même direction), on peut donc avoir une MSC qui atteint 1 en basses fréquences, et qui reste très haute jusqu'à une fréquence limite dépendant de la distance entre les capteurs. On peut donc, dans cette bande, estimer une fonction de transfert entre deux capteurs, qui modélise une propagation relative du bruit entre ces microphones.

Cas du bruit très cohérent

On se place dans une situation où la MSC des bruits observés est très proche de 1, en basses fréquences. Dans cette situation, on s'intéresse à la propagation relative du bruit, en estimant un transfert linéaire entre les bruits captés.

Les transferts sont estimés dans la bande [0, 1 kHz], au sens des moindres carrés : on peut déterminer le filtre linéaire entre un signal de référence et une version filtrée en résolvant le problème, étant donné les deux signaux x_{ref} et x_1 de longueur N :

$$h = \underset{g}{\operatorname{argmin}} |x_1^H - g^H X_{ref}|^2 \tag{2.11}$$

où x_1 est le vecteur $[x_1(0) \square\square x_1(N-L)]^H$, X_{ref} est une matrice de Toeplitz d'échantillons

du signal x_{ref} de taille $L \times N - L + 1$ qui vaut :

$$X_{ref} = \begin{bmatrix} x_{ref}(0) & \cdots & x_{ref}(N-L) \\ \vdots & & \vdots \\ x_{ref}(L-1) & \cdots & x_{ref}(N-1) \end{bmatrix} \tag{2.12}$$

et L est la taille du filtre estimé. La solution est :

$$h = \left(X_{ref} X_{ref}^H \right)^{-1} X_{ref} x_1 \tag{2.13}$$

Dans tous les cas, ce transfert est assez stable au cours du temps. Un exemple est présenté dans la Figure 2.23.

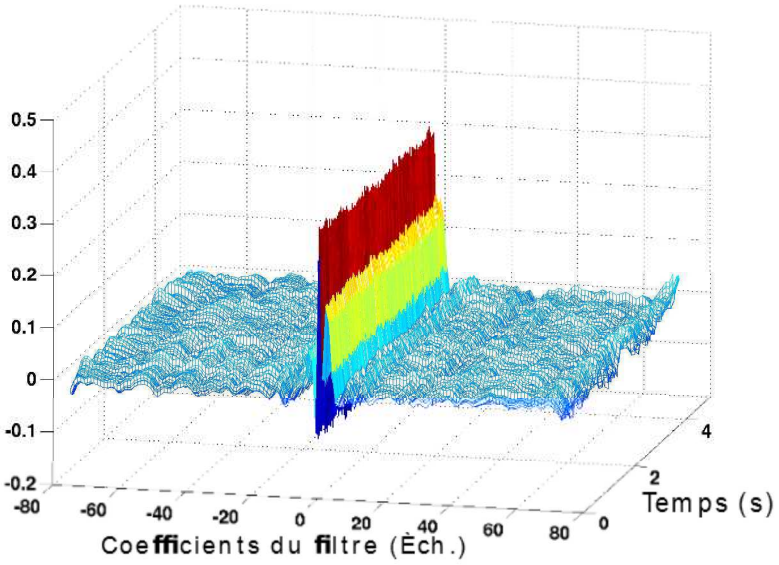


Figure 2.23 – Évolution de la propagation relative du bruit au cours du temps pour 2 microphones unidirectionnels placés à 12 cm l'un de l'autre.

Outre la faible variation de ce transfert (non-causal) au cours du temps, on remarque que les coefficients importants sont très proches de l'origine, et que ceux correspondant à un décalage supérieur à une dizaine d'échantillons sont très faibles : c'est un filtre court.

On s'intéresse ici aux cas où la MSC atteint la valeur 1 en basses fréquences : les antennes à deux omnidirectionnels, et les antennes à deux microphones unidirectionnels placés comme dans la situation (a) de la Figure 2.16.

Les transferts estimés sont représentés Figure 2.24.

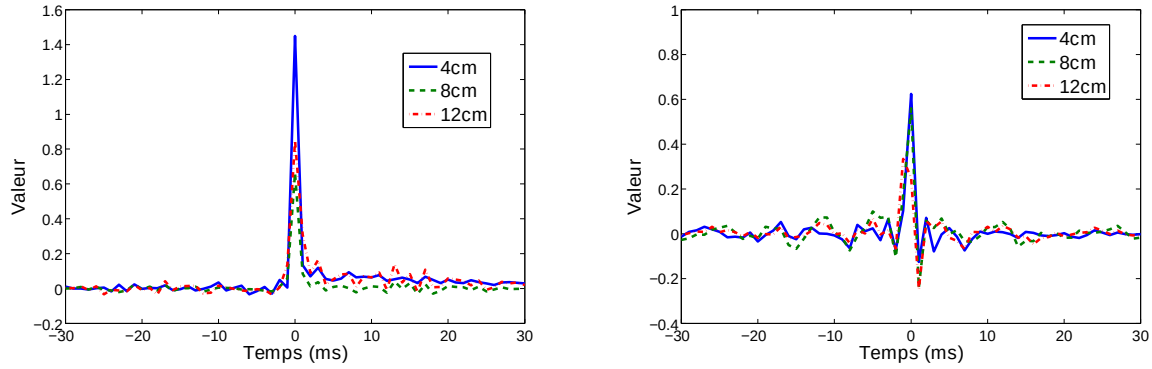


Figure 2.24 – Transferts entre bruits basses fréquences pour des antennes avec des microphones omnidirectionnels (a) et unidirectionnels (b)

On remarque que ces transferts ont toujours leur pic principal en zéro : il n'y a pas de retard d'un capteur sur l'autre sur les signaux de bruit. Ceci est dû au fait que le bruit vient de toutes les directions : n'ayant pas de direction d'arrivée privilégiée, les retards d'arrivée des ondes incidentes du bruit sur les capteurs s'annulent.

2.3.3 Caractéristiques spectrales du bruit

On observe ici la répartition de l'énergie du bruit d'autoroute sur l'horizon fréquentiel. La DSP du bruit d'autoroute a été estimée par la méthode de Welch, sur des fenêtres de 32 ms, recouvrantes à 50%. L'estimation se fait sur une durée totale de signal de 2 s. Un exemple est présenté dans la Figure 2.25.

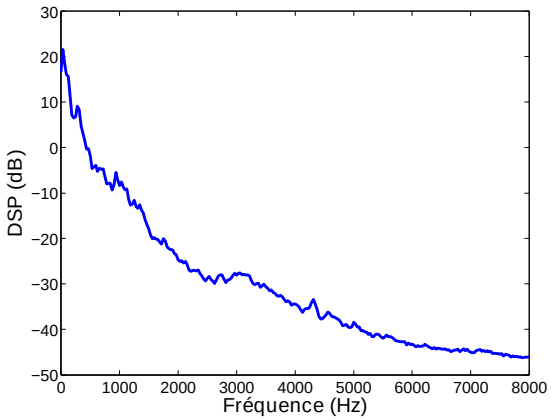


Figure 2.25 – DSP d'un bruit d'autoroute capté par un microphone omnidirectionnel

On constate que l'énergie du bruit décroît lorsque la fréquence augmente. Ainsi, les basses fréquences seront plus affectées par le bruit ambiant que les hautes fréquences. Ceci est également illustré sur le spectrogramme de bruit de la Figure 2.11 (page 39).

2.3.4 Synthèse

On retiendra particulièrement que le bruit stationnaire, principalement produit par le moteur, le roulement du pneu sur la chaussée et le vent, peut être modélisé spatialement comme un bruit diffus sphérique. Pour certaines antennes, il présente une forte cohérence spatiale en basses fréquences, ce qui permet dans cette bande de modéliser deux bruits enregistrés par deux microphones proches comme filtrés l'un de l'autre. On peut alors estimer le filtre correspondant à une propagation relative du bruit : ce filtre varie peu dans le temps, est court, et a un fort pic au temps zéro, quelle que soit la distance entre les microphones. C'est aussi dans la bande des basses fréquences que beaucoup de l'énergie du bruit est concentrée : il est donc judicieux d'exploiter la cohérence dans cette partie du spectre pour réduire fortement l'énergie du bruit.

2.4 Propagation de la voix

Contrairement au bruit que l'on a présenté, le signal de parole est produit par une unique source : le conducteur. C'est une source bien localisée dans l'espace, non diffuse. Dans ce cas, les microphones reçoivent le même signal de parole, au canal acoustique (la propagation dans l'habitacle) près.

Il est donc légitime de supposer que le signal de parole seul présentera une forte cohérence inter-capteurs, sur toute la bande de fréquence [Freudenberger et al., 2009].

Nous allons présenter un exemple de propagation entre le locuteur et un capteur, avant de vérifier si l'hypothèse de cohérence totale sur la parole est vérifiée. Nous présenterons ensuite la propagation relative de la parole entre les capteurs.

2.4.1 Propagation entre locuteur et capteur

On mesure le canal acoustique entre la position de la bouche du locuteur et un capteur en mettant un haut-parleur diffusant un signal large bande (ici, un bruit blanc) à la place du locuteur. On utilise le même système que celui présenté dans la Figure 2.9, en plaçant un microphone de mesure dont on connait parfaitement la réponse en fréquence (ici, un GRAS AE46) devant le haut parleur servant de source de signal utile. Ce microphone de mesure permet de compenser les effets de la réponse en fréquence du haut-parleur utilisé, pour obtenir exactement la source acoustique de référence. On peut ensuite déterminer le filtre linéaire entre cette référence et le capteur en résolvant le problème de l'Équation (2.11). Cette identification est faite sur un signal d'une durée de 10 s.

La réponse impulsionnelle mesurée est donnée dans la Figure 2.26.

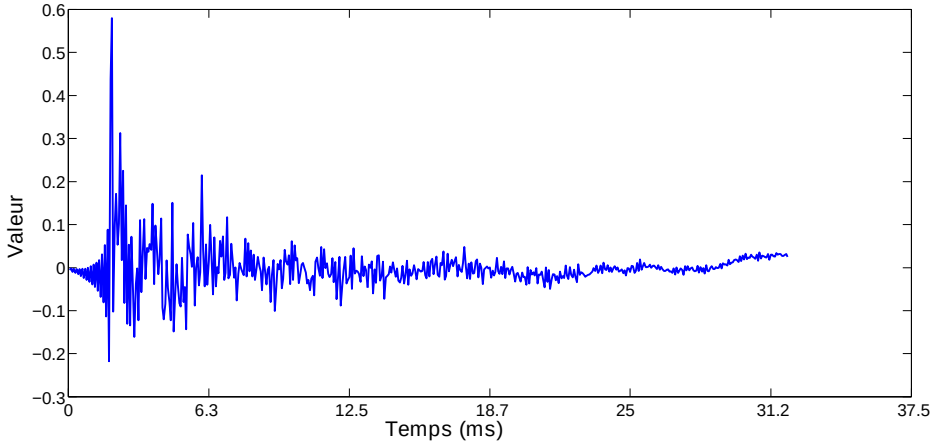


Figure 2.26 – Réponse impulsionnelle du canal acoustique entre la position du locuteur et un capteur omnidirectionnel, à 16 kHz. La source de signal est placée à environ 40 cm du microphone.

On constate que cette réponse est courte, et présente de fortes composantes de réflexion : l'habitacle a un impact important sur la propagation des sons produits à l'intérieur de la voiture, notamment du fait de la présence de nombreuses surfaces vitrées. Il faut donc que les traitements utilisés soient robustes à la présence de forts échos. On s'intéresse à présent à la réponse en fréquence de ce même chemin acoustique, présentée dans la Figure 2.27.

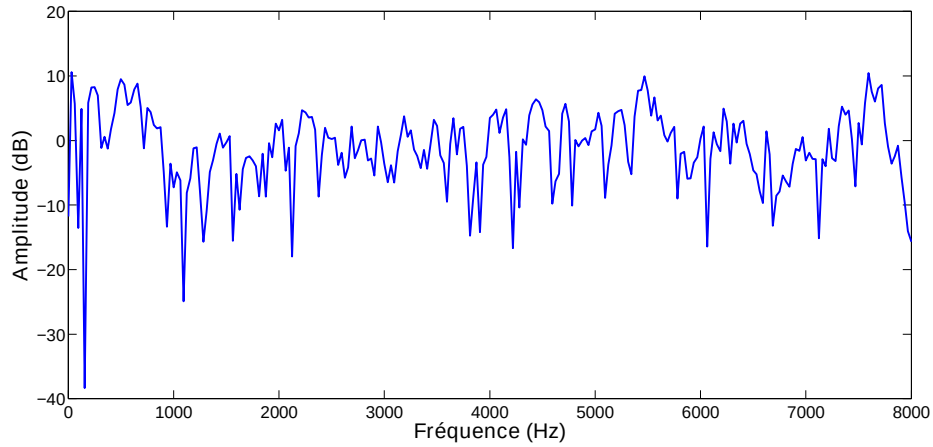


Figure 2.27 – Réponse en fréquence du canal acoustique entre la position du locuteur et un capteur omnidirectionnel

On remarque la présence de nombreux creux ponctuels : l'habitacle de la voiture absorbe certaines fréquences, du fait de la propagation ou d'interférences destructives.

2.4.2 Cohérence

On s'intéresse à présent au cas où l'on utilise plusieurs capteurs. On utilise le même système que précédemment, en mettant cette fois-ci une antenne de capteurs au niveau du plafonnier, et la source de signal utile est toujours un bruit blanc, et non de la parole. Ceci permet de s'affranchir de la non-stationnarité du signal de parole, ainsi que de sa répartition spectrale, qui

ne remplit pas toute la bande de fréquences. Les cardioïdes pointent vers la source ponctuelle, et les capteurs sont en position Endfire, comme montré dans la Figure 2.28.

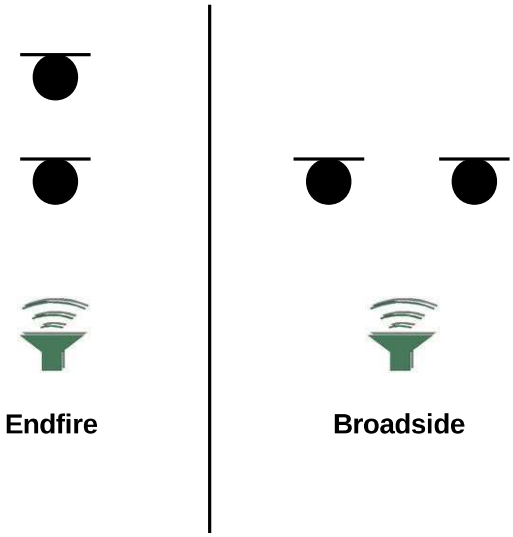


Figure 2.28 – Deux alignements de capteurs possibles : Endfire et Broadside

On s'intéresse à la MSC entre les signaux captés par deux microphones distincts, dans plusieurs situations. Ceci permet de vérifier si le signal utile présentera bien une forte cohérence entre les capteurs sur toutes les fréquences. Les MSC sont estimées sur des périodes de 10s de signal, par la même méthode que dans la Section 2.2, en utilisant des fenêtres de 32 ms recouvrantes à 50%. On présente les courbes de cohérence mesurées pour plusieurs types d'antenne :

— deux microphones omnidirectionnels (Figure 2.29),
— deux microphones cardioïdes (Figure 2.30),
— un microphone omnidirectionnel et un cardioïde (Figure 2.31).

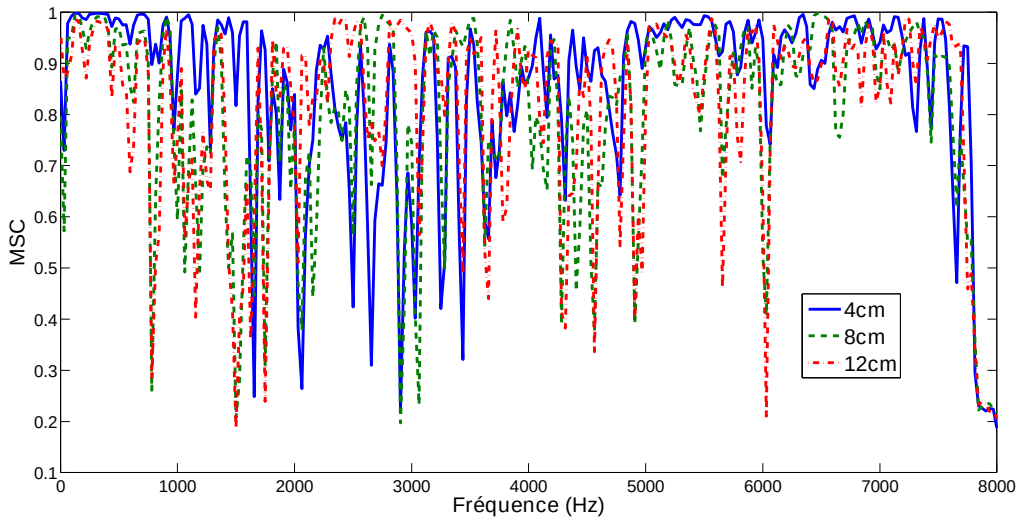


Figure 2.29 – MSC mesurée pour le signal ponctuel, pour des capteurs omnidirectionnels séparés de 4, 8 et 12 cm

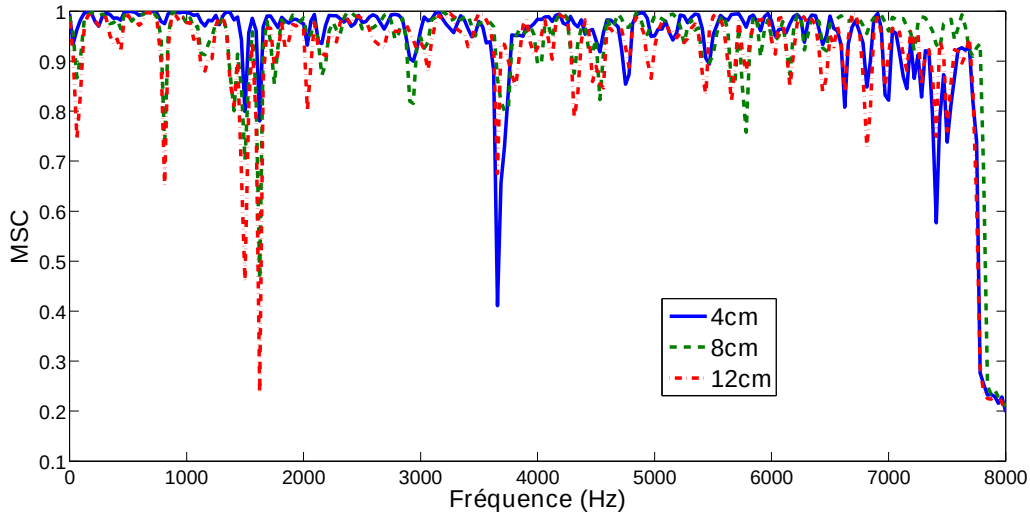


Figure 2.30 – MSC mesurée pour le signal ponctuel, pour des capteurs cardioïdes séparés de 4, 8 et 12 cm

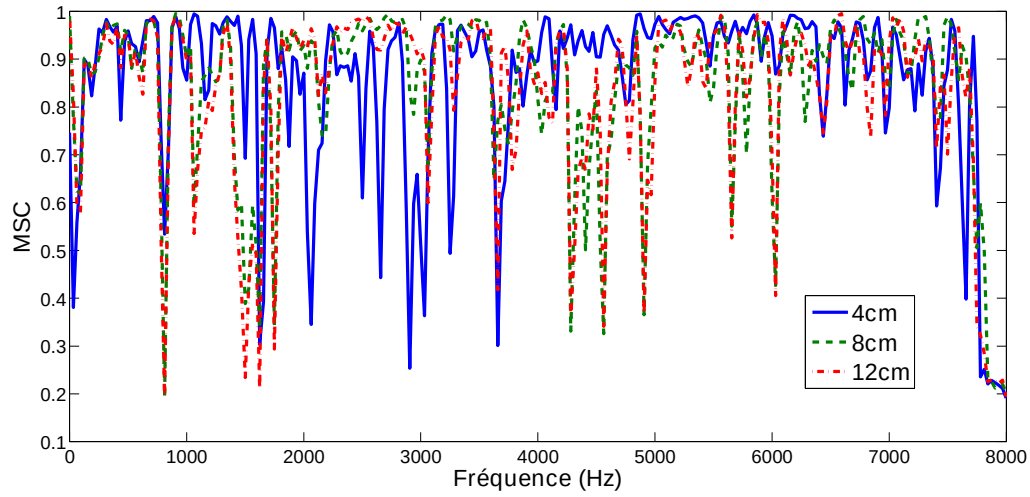


Figure 2.31 – MSC mesurée pour le signal ponctuel, pour un capteur cardioïde et un capteur omnidirectionnel séparés de 4, 8 et 12 cm

On constate que dans tous les cas, la cohérence entre les signaux captés est très forte sur toute la bande de fréquence, à part certains creux ponctuels. Ceux-ci sont à mettre en lien avec la propagation de l'habitacle de voiture (comme suggéré par la réponse en fréquence de la Figure 2.27), qui peut annuler des fréquences selon le placement du capteur. On a alors un problème de variance sur l'estimation de la MSC, comme vu dans la Section 2.2 (page 36).

On illustre ce phénomène en comparant la DSP des deux signaux captés et leur MSC dans le cas de deux microphones omnidirectionnels placés à 12 cm l'un de l'autre. On cherche alors à comparer ces grandeurs à la variance de l'estimée de la MSC. Pour estimer cette variance, on suppose que l'on dispose de N fenêtres recouvrantes pour chacun des signaux (voir Équation (2.5), page 37), avec $N = K \times L$. On a alors K regroupements de L fenêtres chacun. On estime la MSC des signaux sur chacun de ces regroupements $k \in [0\ K - 1]$, en reprenant les

notations de la Section 2.2 :

$$\hat{M}SC_k(f) = \frac{\left| \sum_{n=kL}^{(k+1)L-1} A_n(f) B_n^*(f) \right|^2}{\sum_{n=kL}^{(k+1)L-1} A_n(f) A_n^*(f) \sum_{n=kL}^{(k+1)L-1} B_n(f) B_n^*(f)} \tag{2.14}$$

La variance de la MSC estimée peut être estimée par :

$$\sigma^2_{\hat{M}SC}(f) = \frac{1}{K-1} \sum_{k=0}^{K-1} [\hat{M}SC_k(f) - \hat{M}SC(f)]^2 \tag{2.15}$$

où $\hat{M}SC(f)$ est la MSC estimée globalement sur les N fenêtres. La déviation standard est alors égale à $\sqrt{\sigma^2_{\hat{M}SC}(f)}$.

On choisit ici $N = 600$ (soit environ 10 s), $L = 15$ et $K = 40$.

La DSP des signaux enregistrés par ces microphones et leur MSC sont tracées dans la Figure 2.32, ainsi que la déviation standard de l'estimation de la MSC.

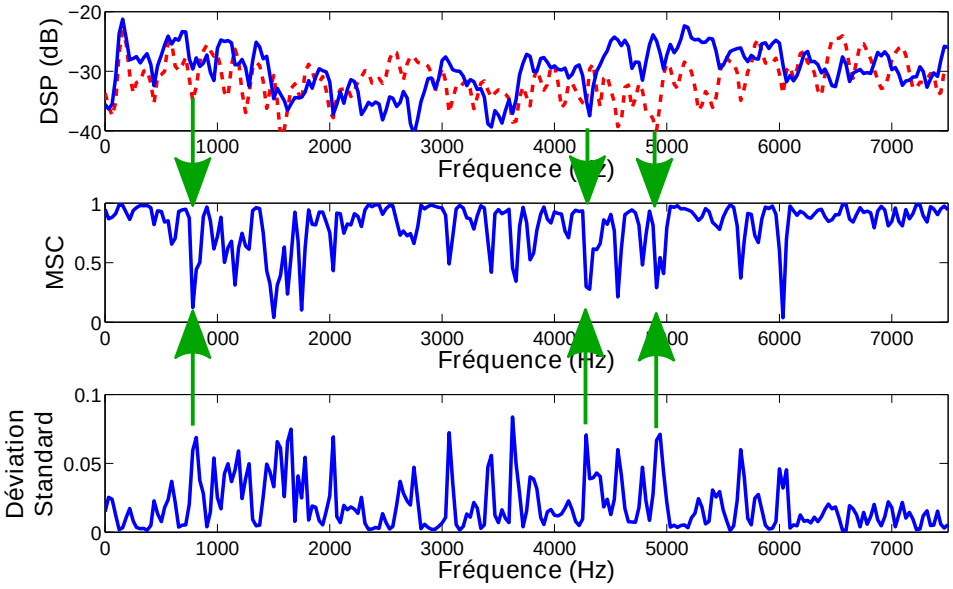


Figure 2.32 – DSP des signaux enregistrés par deux microphones omnidirectionnels placés à 12 cm l'un de l'autre (en haut), la MSC mesurée sur tout le signal, et la variance de l'estimation de MSC. Les flèches montrent des correspondances entre les creux dans une DSP, les creux dans la MSC et les pics dans la déviation standard.

Les creux observés dans les MSCs des Figures 2.29, 2.30 et 2.31 sont donc dus à la propagation de la parole dans l'habitacle. Comme on l'observe sur la Figure 2.27, certaines fréquences sont absorbées par l'habitacle, et la MSC est donc mal estimée sur ces fréquences, car la DSP d'au moins un des signaux considérés est trop faible pour avoir une estimation précise, comme on l'avait remarqué dans la Section 2.2.

La MSC restant proche de 1 sur toute la bande, on peut donc estimer un transfert linéaire entre les voix enregistrées par deux microphones différents, qui correspondent à la propagation relative de la parole entre ces microphones.

2.4.3 Propagation relative de la parole

A l'aide du même montage que pour les mesures de MSC, on estime des transferts entre les signaux captés lorsque l'on enregistre une source ponctuelle large bande, qui correspondent à la propagation relative du signal entre les capteurs.

Ces transferts sont plus longs que ceux estimés dans le cas du bruit cohérent, mais aussi moins stables au cours du temps : on présente un exemple d'évolution d'un de ces transferts dans la Figure 2.33. Pour voir la variation, ces transferts ont été estimés alors qu'un locuteur parlait dans la voiture à l'arrêt dans un parking calme, et le locuteur bougeait normalement sur le siège conducteur.

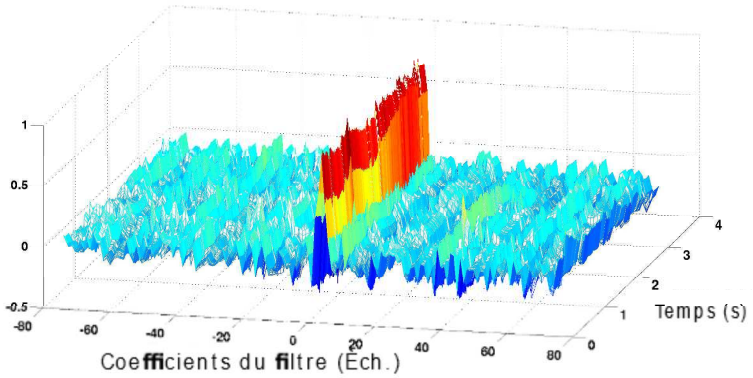


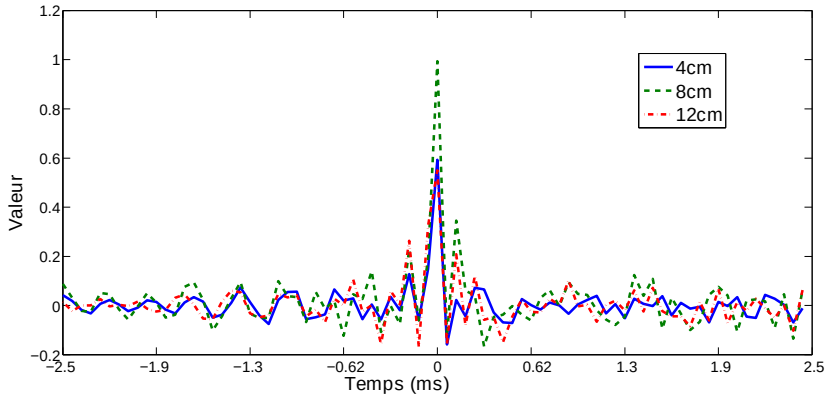
Figure 2.33 – Évolution de la propagation relative de la parole au cours du temps pour 2 microphones omnidirectionnels placés en Broadside à 8 cm l'un de l'autre. Le pic principal se situe en 0 : ce pic n'indique pas de retard.

Cette variabilité est due aux réverbérations proches sur les vitres et le pare-brise : de petits mouvements du locuteur peuvent grandement modifier la propagation [Motojima et al., 2009].

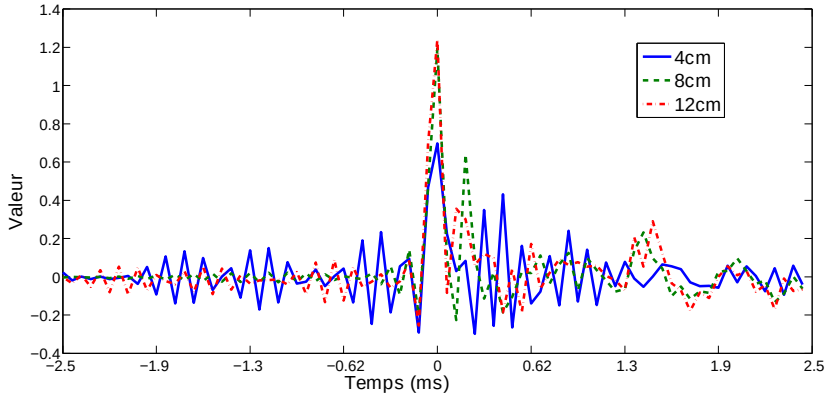
Pour se rendre compte des propriétés des différentes antennes, on utilise le même montage que dans la Section 2.4.1, avec un haut-parleur fixe, pour s'affranchir des mouvements du locuteur. On s'intéresse, pour différentes antennes, aux deux types de placement montrés Figure 2.28 : Broadside et Endfire.

Cas Broadside

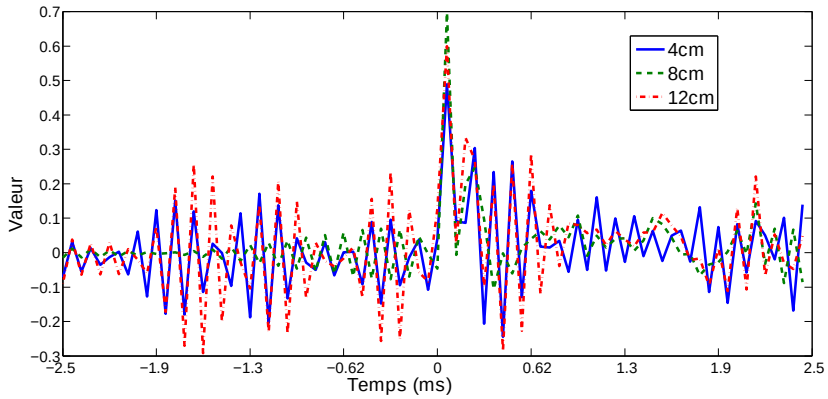
On s'intéresse ici au cas Broadside : les capteurs sont placés sur un axe perpendiculaire à la direction d'arrivée du chemin direct de la propagation de la parole. Ils sont donc à la même distance du locuteur. Les transferts estimés pour différentes antennes sont présentés dans la Figure 2.34.



Antenne de 2 microphones omnidirectionnels



Antenne de 2 microphones unidirectionnels



Antenne mixte de microphones unidirectionnel/omnidirectionnel

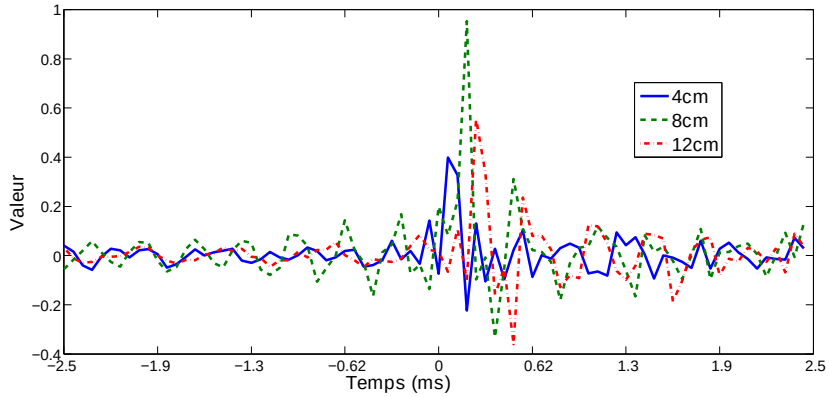
Figure 2.34 – Propagation relative de la parole pour plusieurs antennes dans le cas **Broadside**, à 16 kHz

On remarque que le pic principal, correspondant au chemin direct, est toujours au temps zéro quelle que soit l'antenne, ce qui illustre que les capteurs sont bien à la même distance du locuteur. Des chemins secondaires importants apparaissent, très marqués lorsque l'on utilise une antenne mixte unidirectionnel/omnidirectionnel.

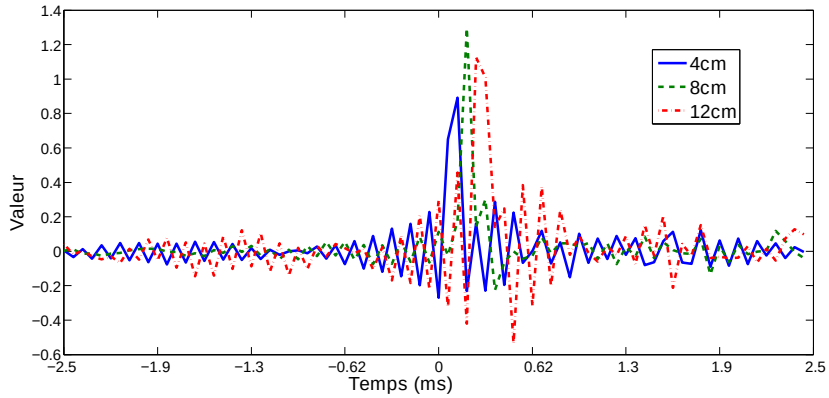
Cas Endfire

On présente les mêmes estimations, mais dans le cas **Endfire** cette fois-ci. Les microphones sont donc dans l'axe de la direction d'arrivée du chemin direct, et ne sont plus à la même distance

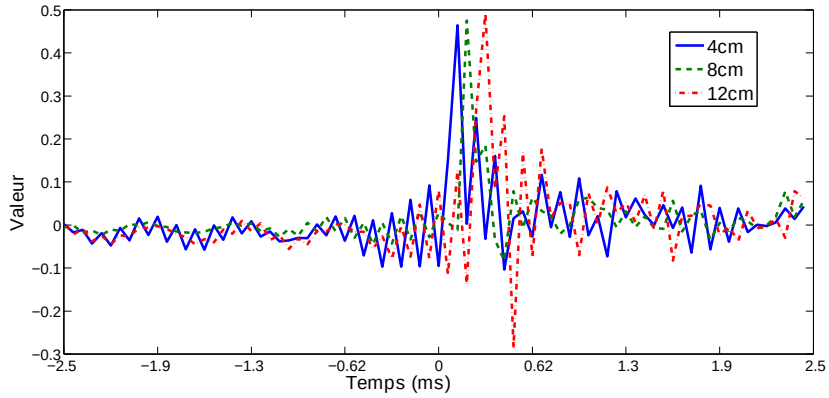
du locuteur. Les estimations sont présentées Figure 2.35.



Antenne de 2 microphones omnidirectionnels



Antenne de 2 microphones unidirectionnels



Antenne mixte de microphones unidirectionnel/omnidirectionnel

Figure 2.35 – Propagation relative de la parole pour plusieurs antennes dans le cas Endfire, à 16 kHz

Le pic principal se décale à mesure que la distance entre les capteurs augmente, ce qui illustre le retard d'arrivée de chemin direct entre les capteurs. Les chemins secondaires sont aussi présents quel que soit le type d'antenne.

En comparant ces estimations avec celles faites dans le cas du bruit (Figure 2.24, page 49), on déduit que dans la situation Endfire, le pic principal de la propagation relative de la parole se décale par rapport à celui correspondant au bruit très cohérent basses fréquences, à mesure que la distance entre les capteurs augmente. C'est donc un moyen de différencier la propagation

relative du bruit et de la parole, ce qui est intéressant si l'on souhaite faire de l'Annulation de bruit adaptative (Adaptive Noise Cancellation) (ANC) : on peut annuler le bruit en limitant l'impact sur la parole. On ne peut toutefois éloigner beaucoup les capteurs, puisque l'on perd alors en cohérence de bruit, et la ressemblance linéaire des composantes de bruit n'existe plus.

2.5 Synthèse

Nous avons mis en évidence un certain nombre de propriétés sur les sources de bruit et de voix, que nous allons résumer ici.

Sur le bruit Il y a deux types de bruits : un bruit stationnaire (moteur, vent, roulement) et d'autres, plus ponctuels (circulation environnante, klaxon...). On s'intéresse tout particulièrement au premier.

Celui-ci est stationnaire, a beaucoup d'énergie en basses fréquences. Son intensité diminue lorsque la fréquence augmente.

Spatialement, on peut le modéliser comme un bruit diffus sphérique : selon les antennes de capteurs utilisées, il peut être très cohérent spatialement en basses fréquences, mais cette cohérence chute dans tous les cas lorsque la fréquence augmente. Dans le cas où l'antenne permet d'avoir une grande cohérence en basses fréquences (lorsque l'on utilise des microphones omnidirectionnels, ou des unidirectionnels pointant dans la même direction), on peut considérer que jusqu'à une certaine fréquence, les bruits captés par deux microphones sont les mêmes à un filtre linéaire près. Ce filtre est alors assez court, et a un pic principal en zéro, il n'y a pas de décalage qui apparait sur la propagation relative du bruit, quelle que soit la position de l'antenne de capteurs.

Sur la parole La parole est un signal instationnaire, qui perd en énergie dans les hautes fréquences (>4 kHz). Des enregistrements de ce signal, faits en deux points de l'habitacle (spatialement proches), restent cohérents sur toute la bande de fréquences, étant produits par une source ponctuelle dans l'habitacle de la voiture. La propagation de ce signal varie beaucoup avec les mouvements, même légers, du conducteur.

La propagation montre l'importance de la réverbération de l'habitacle, qui ne permet pas de considérer la propagation relative de la parole entre les capteurs comme de simples retards. Toutefois, le chemin direct est assez important pour que l'on voie apparaitre un retard dépendant de la distance entre les capteurs, dans le cas où les capteurs sont alignés avec la source de parole.

Conclusion Voici quelques conclusions générales que l'on peut tirer de ces observations :

- On peut supposer que le bruit diffus est peu cohérent spatialement en hautes fréquences, pour toutes les antennes de capteurs.
- Selon l'antenne utilisée, on peut supposer le bruit très cohérent spatialement en basses fréquences, et même supposer que les bruits reçus par deux capteurs diffèrent uniquement d'un transfert linéaire.
- La parole captée en deux points donne des signaux qui sont très cohérents sur toute la bande de fréquences.

— Quel que soit le type de capteur utilisé, le RSB d'entrée sera très faible en basses fréquences.
— En utilisant des capteurs cardioïdes, on peut obtenir un RSB d'entrée assez fort en hautes fréquences.

Pour exploiter les caractéristiques spatiales du bruit et de la parole, il faut donc prévoir un système prenant en compte un bruit faiblement cohérent, que nous pourrons exploiter en hautes fréquences, ainsi qu'un système utilisant une forte cohérence de bruit (pour les basses fréquences), voire des bruits complètement cohérents, si l'on utilise des antennes permettant d'atteindre une MSC de 1 en basses fréquences.

Chapitre 3

Beamforming Minimum Variance Distortionless Response (MVDR) adaptatif

Sommaire

Nous allons dans ce Chapitre proposer une implémentation adaptative du beamforming MVDR. Celle-ci permet de mieux prendre en compte la réverbération présente dans un habitacle automobile, mais souffre de défauts. C'est pourquoi nous étudions également les performances de cette méthode en fonction de l'environnement, pour pouvoir définir dans quelles conditions elle sera le plus efficace, étant donné les caractéristiques de l'environnement automobile. Cette étude permet de définir quelle bande de fréquences et quelle antenne de capteurs sera la plus appropriée pour ce système.

Nous avons vu dans le Chapitre 1 que les méthodes de débruitage fréquentielles usuelles se décomposent en un beamforming MVDR suivi d'un traitement mono-capteur.

Nous cherchons donc à implémenter un tel beamforming dans le cas d'une prise de son dans un habitacle automobile. Ce filtrage multi-capteurs nécessite l'estimation, d'une part, de la propagation du signal de parole entre la bouche du locuteur et les microphones et d'autre part, des matrices spectrales des bruits captés.

Nous allons tout d'abord illustrer les difficultés posées par l'habitacle de la voiture pour l'estimation de la propagation de la parole. Nous présenterons ensuite une solution pour cette estimation, dont nous montrerons les limitations afin de comprendre dans quelles conditions elle est efficace. Nous présenterons ensuite une solution pour l'estimation des matrices spectrales des bruits. Enfin, nous analyserons l'impact de la géométrie de placement des capteurs sur les performances de la méthode proposée, afin de définir la meilleure stratégie acoustique pour l'exploitation de cette méthode.

3.1 Rappels sur le modèle et les notations

Le système considéré est une antenne de M microphones captant un unique signal utile. Ce modèle est généralement désigné dans la littérature anglo-saxonne par Single Input Multiple Output (SIMO). On suppose que le signal utile est affecté uniquement par la propagation acoustique (en tenant compte de la réverbération), modélisée par des filtres à Réponse Impulsionnelle Finie (RIF), et du bruit additif. Ainsi, le signal échantillonné capté sur le microphone numéroté m s'écrit, à l'échantillon t :

$$x_m(t) = \{h_m \otimes s\}(t) + b_m(t) \tag{3.1}$$

où $s(t)$ est le signal utile discret, $h_m(t)$ la réponse impulsionnelle correspondant à la propagation acoustique vers le microphone et $b_m(t)$ le bruit additif.

Dans l'équation (3.1), s et h_m sont à la fois inconnus, par conséquent $h_{m\,m\in[1\,M]}$ ne peuvent donc être identifiés qu'à une fonction de transfert près.

On peut écrire l'Équation (3.1) sous forme vectorielle compacte :

$$x(t) = \{h \otimes s\}(t) + b(t) \tag{3.2}$$

avec $x(t) = [x_1(t)\ \cdots\ x_M(t)]$, $h(t)$ et $b(t)$ étant définis de la même façon.

L'approche considérée dans ce chapitre est fréquentielle. Les données sont tronquées en fenêtres recouvrantes (avec un taux de recouvrement $r = \frac{1}{2}$, soit 50%) sur lesquelles nous appliquons une Transformée de Fourier à Court Terme (TFCT).

La TFCT du signal enregistré sur le capteur m s'écrit à la fréquence f_k et pour la trame n :

$$X_{m\,n}(f_k) = \frac{1}{\sqrt{N}} \sum_{t=\lfloor nN \rfloor + 1}^{\lfloor nN \rfloor + N} x_m(t) e^{-2i\pi f_k t} \tag{3.3}$$

où N est la taille de la fenêtre considéré (en échantillons), $\alpha = 1 - r$ avec $r \in]0,1[$, et :

$$f_k = \frac{k}{N}, k = (0, \ldots, N-1) \tag{3.4}$$

est une fréquence réduite qui correspond à une fréquence réelle de $(f_k\ F_s)$ Hz, F_s étant la fréquence d'échantillonnage.

Le modèle de mélange dans le domaine fréquentiel pour la fenêtre n et la fréquence f s'écrit :

$$X_{m,n}(f_k) = H_m(f_k)S_n(f_k) + B_{m,n}(f_k), \tag{3.5}$$

Cette écriture est une approximation, en supposant la longueur de h_m faible devant la longueur de la fenêtre sur laquelle on effectue la Transformée de Fourier Discrète (TFD). Si cette condition n'est pas remplie, l'écriture en fréquence correspond dans le domaine temporel à une convolution circulaire, et non linéaire. En écrivant les M entrées en une notation vectorielle compacte, on obtient

$$\mathbf{X}_n(f_k) = \mathbf{H}(f_k)S_n(f_k) + \mathbf{B}_n(f_k) \tag{3.6}$$

où les caractères en gras sont des vecteurs. Par exemple, $\mathbf{X}_n(f_k) = [X_{1,n}(f_k), \ldots, X_{M,n}(f_k)]^T$ contient pour la trame n les composantes de chaque capteur.

On suppose que $b(t)$ est un processus Stationnaire au second ordre au Sens Large (SSL), et on note $R_b(\tau)$ sa matrice de covariance au décalage τ, définie par :

$$R_b(\tau) = E\left[b(t+\tau)b(t)^H\right], \tag{3.7}$$

La matrice spectrale des bruits est alors définie comme la transformée de Fourier des matrices de covariance :

$$\Sigma_b(f) = \sum_{\tau \in \mathbb{Z}} R_b(\tau)e^{-2i\pi f\tau} \tag{3.8}$$

De la même façon, on définit la Densité Spectrale de Puissance (DSP) du signal utile $s(t)$ comme la transformée de son autocovariance :

$$\begin{aligned} R_s(\tau) &= E[s(t+\tau)s(t)^*] \\ \varphi_s(f) &= \sum_{\tau \in \mathbb{Z}} R_s(\tau)e^{-2i\pi f\tau} \end{aligned} \tag{3.9}$$

De façon générale, on note dans la suite $\varphi_a(f)$ la DSP d'un processus $a(t)$, et $\varphi_{ab}(f)$ l'interspectre de deux processus $a(t)$ et $b(t)$, qui est la transformée de Fourier de leur intercovariance, supposée stationnaire.

Nous utilisons un estimateur d'amplitude spectrale à court terme minimisant l'Erreur Quadratique Moyenne (EQM) sur la log-amplitude spectrale [Ephraim and Malah, 1985]. Remarquons que l'estimateur Log-Spectral Amplitude (LSA) multicanaux est décomposé en un beamforming MVDR suivi d'un estimateur de type Ephraim et Malah [Hendriks et al., 2009], comme on l'a vu dans la Section 1.3.3 (page 21). Ainsi, le signal débruité est donné par :

$$\hat{S}_n(f_k) = G_{OM-LSA}(f_k)T(\mathbf{X}_n(f_k)) \tag{3.10}$$

où $T(X_n(f_k))$ est la sortie du beamforming MVDR, donnée par :

$$T(X_n(f_k)) = \frac{H^H(f_k)\Sigma_b(f_k)^{-1}X_n(f_k)}{H^H(f_k)\Sigma_b(f_k)^{-1}H(f_k)} \tag{3.11}$$

Pour s'affranchir de l'indétermination dans l'équation (3.1), on suppose $H_1(f_k) = 1$. Cela revient à prendre le signal $\{h_1 \otimes s\}(t)$ comme référence [Habets et al., 2010a]. Pour pouvoir calculer le signal $T(X(f_k))$, il nous faut une estimation des transferts acoustiques $H(f_k)$ et des matrices spectrales des bruits $\Sigma_b(f_k)$. L'estimation de la propagation de la parole doit tenir compte d'une forte réverbération de l'environnement, comme on l'a vu dans la Section 2.4 (page 50), et doit également être adaptative, puisque cette propagation varie en fonction des mouvements, même faibles, du locuteur.

Nous allons dans un premier temps illustrer les difficultés posées par cet environnement, et la solution envisagée pour estimer de façon adaptative cette propagation, ainsi que les conditions pour que cette méthode soit efficace. Nous aborderons ensuite l'estimation des matrices spectrales de bruit Σ_b, avant de nous intéresser à l'influence du placement des microphones sur le comportement du beamforming proposé.

3.2 Estimation de la propagation relative de la parole

On s'intéresse ici à l'estimation de la propagation relative de la parole, notée $H(f)$ précédemment. Nous allons tout d'abord nous intéresser à une méthode basée uniquement sur la connaissance de la géométrie de l'antenne de microphones utilisée, dont nous montrerons qu'elle n'est pas efficace dans l'environnement automobile. Nous présenterons ensuite une alternative simple, basée sur la prédiction du signal de parole d'un capteur sur l'autre. Cette méthode est plus adaptée aux conditions automobiles, mais souffre également de défauts. L'étude de ces défauts nous permettra de déterminer dans quelle bande de fréquences cette méthode est efficace.

3.2.1 Problème posé par l'écho présent dans l'habitacle

Il est courant dans de nombreuses applications de considérer que la propagation relative de la parole ne dépend que du chemin direct entre la source de parole et les capteurs utilisés. Cette approche est valable lorsque la réverbération de l'environnement est négligeable, et que la source de signal utile est ponctuelle et isotrope. Cela permet de déterminer cette propagation relative uniquement à partir de la position du locuteur (la source de signal utile), et la géométrie de l'antenne, comme on l'illustre dans la Figure 3.1.

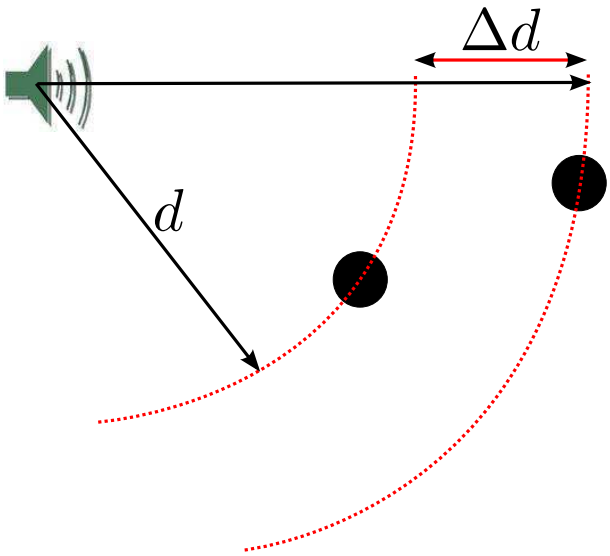


Figure 3.1 – Cas du seul chemin direct, ici pour une source isotrope en champ libre.

Cette propagation relative est alors simplement une atténuation, dépendante de la distance entre la source et les différents capteurs, et un retard, dépendant de la différence de marche entre les rayons acoustiques.

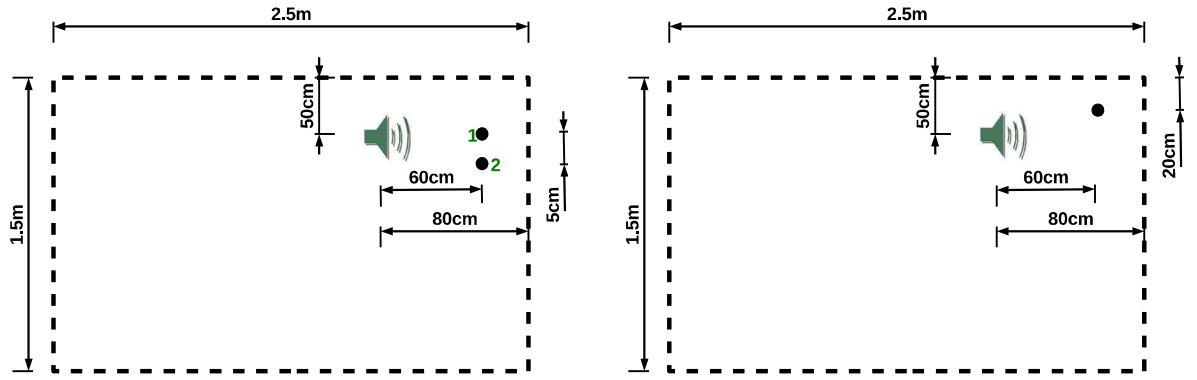
On peut alors calculer $\hat{H}(f_k)$ à partir de cette géométrie, pour $f \in [0;1]$:

$$\hat{H}(f_k) = \begin{bmatrix} 1 \\ \alpha e^{-2i\pi f_k F_s \frac{\Delta d}{c_s}} \end{bmatrix} \tag{3.12}$$

où $\alpha = \dfrac{d^2}{(d+\Delta d)^2}$ est l'atténuation du signal utile due à la différence de distance à la source Δd et c_s, la célérité du son.

Lorsque la partie réverbérée du signal capté devient trop énergétique, l'utilisation du chemin direct seul pour estimer $H(f)$ n'est plus efficace, car la propagation réelle du signal utile n'est pas bien représentée. Cela introduit une baisse de performance en terme de RSB, l'apparition de distorsion sur le signal utile. Or, comme on l'a vu dans le Chapitre 1, l'habitacle d'une voiture présente de fortes composantes de réverbération. Pour illustrer ceci, nous avons simulé une antenne de deux capteurs placée dans un environnement réverbérant et bruyant. La salle est simulée à l'aide d'une méthode d'images [Allen and Berkley, 1979] pour obtenir la réponse impulsionnelle entre la source et les capteurs, en construisant les sources images, et en tenant compte des coefficients de réflexion sur les murs[2]. La méthode est expliquée brièvement dans l'Annexe C (page 147). La situation simulée est présentée dans la Figure 3.2.

2. Nous avons utilisé le logiciel Room Impulse Response Generator de E.H.P Habets, disponible à l'adresse http://home.tiscali.nl/ehabets/rir_generator.html



Salle simulée, vue de haut Salle simulée, vue de côté

Figure 3.2 – Salle simulée, et position de la source de signal utile et des capteurs.

Ceci nous permet de faire varier les coefficients d'absorption des surfaces autour de la salle, et donc l'énergie relative des chemins secondaires (réverbérés) par rapport au chemin direct. Cette énergie est donnée par :

$$\varphi_{reverb} = \frac{\sum_{t=-\infty}^{+\infty} |h_{salle}(t) - h_{direct}(t)|^2}{\sum_{t=-\infty}^{+\infty} |h_{direct}(t)|^2} \tag{3.13}$$

où h_{salle} est la réponse impulsionnelle complète de la salle entre la source et le capteur, et h_{direct} est la réponse pour le chemin direct seul, calculée en considérant une absence de réflexion sur les murs de la salle par l'équation (3.12).

Nous avons utilisé comme source de signal utile $s(t)$ un bruit blanc, et nous avons ajouté sur chaque capteur un bruit blanc $b(t)$, indépendants d'un capteur à l'autre, pour plusieurs RSBs d'entrée. Nous avons alors estimé le gain en RSB dû au MVDR (par rapport aux signaux d'entrée), ainsi que la distorsion apportée sur le signal utile, en utilisant un modèle de chemin direct seul (on suppose que la propagation est composée uniquement d'un retard et d'une atténuation). On connaît ici les matrices spectrales des bruits (notées Σ_b), le problème de leur estimation ne se pose pas. Pour estimer ces critères de performance, nous avons appliqué le MVDR séparément sur les signaux de bruit et de parole (respectivement $b(t)$ et $\{h \otimes s\}(t)$). Le bruit et la parole en sortie s'écrivent en fréquence :

$$\hat{B}(f) = \frac{\tilde{H}(f)^H \Sigma_b(f)^{-1} B(f)}{\tilde{H}(f)^H \Sigma_b(f)^{-1} \tilde{H}(f)} \tag{3.14}$$

$$\hat{S}(f) = \frac{\tilde{H}(f)^H \Sigma_b(f)^{-1} H(f) S(f)}{\tilde{H}(f)^H \Sigma_b(f)^{-1} \tilde{H}(f)} \tag{3.15}$$

et dans ce cas, $\tilde{H}(f)$ dépend uniquement de la géométrie de l'antenne et de la position de la source de parole, et son expression est donnée dans l'équation (3.12). En revanche, $H(f)$ représente la vraie propagation, dépendante de la salle simulée. De plus, $\Sigma_b(f)$ est ici diagonale,

et vaut :

$$\Sigma_b(f) = \begin{bmatrix} \varphi_b(f) & 0 \\ 0 & \varphi_b(f) \end{bmatrix} \tag{3.16}$$

φ_b étant la DSP du bruit sur les 2 capteurs($\varphi_{b_1}(f) = \varphi_{b_2}(f) = \varphi_b(f)$). Ceci vient du fait que l'on suppose des bruits non-cohérents d'un capteur sur l'autre.

Le RSB en sortie vaut alors (on suppose des signaux centrés) :

$$RSB_s = \frac{E\left[|\hat{s}(t)|^2\right]}{E\left[|\hat{b}(t)|^2\right]} \tag{3.17}$$

$\hat{s}(t)$ et $\hat{b}(t)$ étant les processus engendrés par les filtres de réponses en fréquence définies par les équations (3.15) et (3.14), respectivement.

Le RSB d'entrée vaut :

$$RSB_e = \frac{E\left[|s(t)|^2\right]}{E\left[|b_1(t)|^2\right]} \tag{3.18}$$

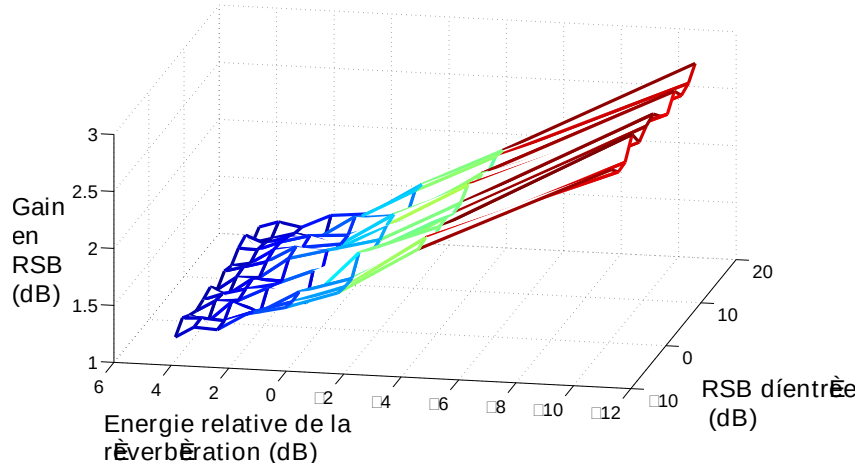
Le gain en RSB vaut donc :

$$\Delta_{RSB} = \frac{RSB_s}{RSB_e} \tag{3.19}$$

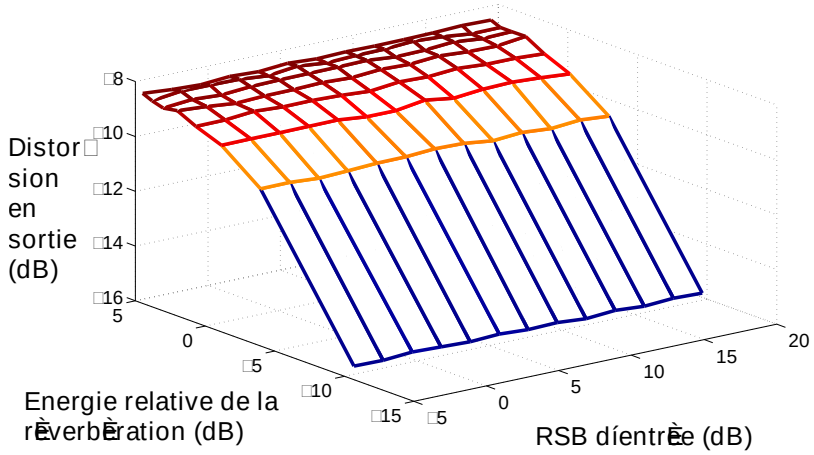
et la distorsion (normalisée par l'énergie du signal de parole) vaut :

$$D = \frac{E\left[|\hat{s}(t) - s(t)|^2\right]}{E\left[|s(t)|^2\right]} \tag{3.20}$$

Le gain en RSB par rapport aux entrées et la distorsion sont représentés dans la Figure 3.3, en fonction de l'énergie relative de la réverbération et du RSB en entrée.



Gain en RSB



Distorsion apportée par le MVDR

Figure 3.3 – Résultats de simulation en supposant un unique chemin direct

On remarque que ces résultats sont quasi-indépendants du RSB en entrée, mais que la présence de réverbération diminue nettement les performances de ce système. Or, comme on l'a vu dans le Chapitre 1, l'habitacle d'une voiture présente de fortes composantes de réverbération, du fait notamment de la présence de surfaces vitrées.

On cherche donc une alternative simple pour prendre en compte la présence de réverbération forte, et qui permette de suivre les changements de propagation dus aux mouvements du locuteur.

3.2.2 Estimation adaptative de la propagation de la parole

Si l'on suppose que les bruits captés par différents microphones sont incohérents, le seul signal cohérent entre les capteurs est le signal utile. On suppose que le canal acoustique peut varier dans le temps, et on note $H_n(f)$ la réponse en fréquence de ce canal dans la fenêtre n. Le mélange devient, en omettant la dépendance en fréquence :

$$X_n(f) = H_n(f)S_n(f) + B_n(f) \quad (3.21)$$

où on a supposé $H_{1n}(f) = 1$, car l'on utilise le capteur 1 comme référence pour le signal utile.

Ainsi, les transferts acoustiques $H_n(f)$ peuvent être estimés par rapport au capteur 1.

$H_n(f)$ est alors estimé par un algorithme Least-Mean Square (LMS) par bloc dans le domaine fréquentiel [Prado and Moulines, 1994] (voir Annexe B, page 143), selon le schéma donné dans la Figure 3.4. Un délai de Δ échantillons est appliqué sur le capteur de référence afin de prendre en compte les retards dus au placement des microphones et à la réverbération, dont on ne sait pas s'il seront positifs ou négatifs.

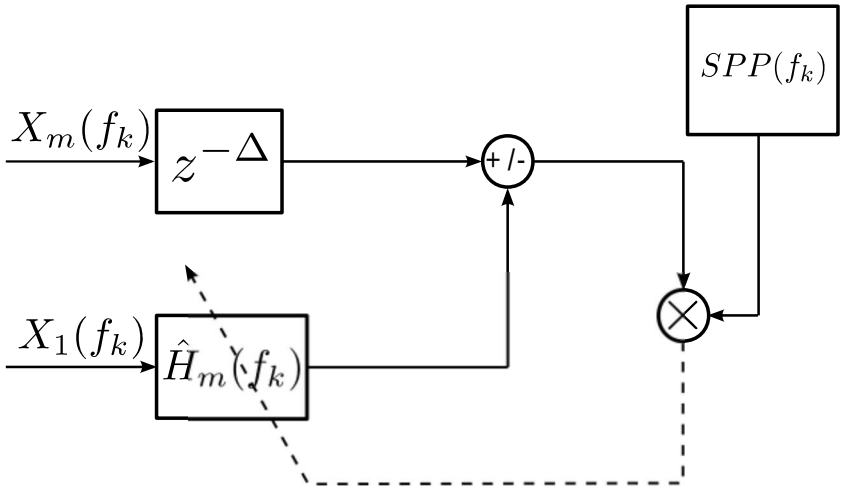


Figure 3.4 – Estimation de la propagation relative par un algorithme Bloc-LMS. $m \in [2 \square M]$ désigne ici l'indice du capteur pour lequel on estime cette propagation relative.

L'équation de mise à jour pour le canal m, à la trame n et à la fréquence f_k s'écrit (voir l'équation C.1, page 147) [Choi et al., 2007] [Fox, 2012] :

$$\hat{H}_{m\,n}(f_k) = \hat{H}_{m\,n-1}(f_k) + \mu(f_k \square n) X^*_{m\,n}(f_k) \left[X_{m\,n}(f_k) - \hat{H}_{m\,n-1}(f_k)\, X_{1\,n}(f_k) \right] \tag{3.22}$$

où

$$\mu(f_k \square n) = \mu_0 \frac{SPP_n(f_k)}{\hat{\varphi}_{x_1}(f_k \square n)} \square \tag{3.23}$$

$SPP_n(f_k)$ est la Probabilité de présence de parole (Speech Presence Probability) (SPP), μ_0 est un pas d'adaptation constant choisi expérimentalement. $\hat{\varphi}_{x_1}(f_k \square n)$ est une estimation de la DSP de x_1 à la trame n, obtenue en lissant le périodogramme de x_1 sur une fenêtre exponentielle. Ainsi, on peut mettre à jour cette estimation récursivement de la façon suivante :

$$\hat{\varphi}_{x_1}(f_k \square n) = \beta \hat{\varphi}_{x_1}(f_k \square n - 1) + (1 - \beta) |X_{1\,n}(f_k)|^2 \tag{3.24}$$

où β est un facteur d'oubli choisi expérimentalement. La SPP est une grandeur déjà utilisée dans l'algorithme Parrot. Elle est estimée en exploitant la non-stationnarité de la parole [Cohen and Berdugo, 2002], de la prédiction linéaire sur la parole [Pinto, 2011], ainsi que des

informations spatiales [Vitte et al., 2012].

Ce pas variable en fonction de la SPP permet d'adapter l'estimation de la propagation de la parole uniquement sur les points temps-fréquence où la parole est présente, et de ne pas mettre à jour H sur les phases de bruit seul, qui n'apportent pas d'information sur cette propagation [Choi et al., 2007] [Fox, 2012].

3.2.3 Limitations de l'approche utilisée

On cherche à mieux comprendre l'influence de l'environnement sur les performances du beamforming adaptatif proposé, et notamment sur l'influence de la cohérence entre les bruits captés par deux microphones différents. On considère ici un système à 2 microphones, et l'analyse se fait dans le domaine fréquentiel (on considère ici f continue dans $[0;1]$). On suppose que les bruits captés sur les différents microphones ont la même DSP $\varphi_{b_1}(f) = \varphi_{b_2}(f) = \varphi_b(f)$ et une cohérence $\frac{\varphi_{b_1 b_2}(f)}{\varphi_b(f)} = c(f)$, avec $|c(f)| < 1$. La matrice spectrale des bruits, que l'on suppose ici connue, s'écrit alors :

$$\Sigma_b(f) = \varphi_b(f) \begin{bmatrix} 1 & c(f) \\ c(f)^* & 1 \end{bmatrix} \tag{3.25}$$

Par la méthode de prédiction du signal utile présentée dans la Section 3.2.2 , on a une estimation de la propagation relative $H(f)$, utilisant le capteur numéroté 1 comme référence. Les deux signaux s'écrivent alors, en fréquence :

$$\begin{aligned} X_1(f) &= H_1(f)S(f) + B_1(f) = S(f) + B_1(f) \\ X_2(f) &= H_2(f)S(f) + B_2(f) = H(f)S(f) + B_2(f) \end{aligned}$$

en notant $H(f) = H_2(f)$ pour simplifier les notations. Si l'on suppose que le filtrage adaptatif présenté dans la Figure 3.4 converge vers la solution de Wiener, la propagation relative sur le capteur 2 est estimée par :

$$\begin{aligned} \hat{H}(f) &= \frac{E[X_2(f)X_1(f)^*]}{E[|X_1(f)|^2]} = \frac{H(f)\varphi_s(f) + E[B_2(f)B_1(f)^*]}{\varphi_s(f) + \varphi_{b_1}(f)} \\ &= \frac{H(f)}{1 + \frac{1}{RSB(f)}} + \frac{c(f)^*}{1 + RSB(f)} \end{aligned} \tag{3.26}$$

et l'estimation de $H(f)$ est donnée par :

$$\hat{H}(f) = \left[\frac{H(f)}{1 + \frac{1}{RSB(f)}} + \frac{c(f)^*}{1 + RSB(f)} \right]^1 \tag{3.27}$$

où :

$$RSB(f) = \frac{\varphi_s(f)}{\varphi_b(f)} \tag{3.28}$$

est le RSB d'entrée.

On constate que cette estimation diffère de la vraie valeur $H(f)$: il y a un biais dépendant

de la cohérence des bruits $c(f)$ et du RSB d'entrée.

Le filtre MVDR adaptatif estimé vaut alors :

$$G_{MVDR}(f) = \frac{\Sigma_b(f)^{-1}\hat{H}(f)}{\hat{H}(f)^H \Sigma_b(f)^{-1}\hat{H}(f)} \tag{3.29}$$

En remplaçant $\hat{H}(f)$ et $\Sigma_b(f)$ par leur valeur, on obtient :

$$G_{MVDR} = \frac{\left(1 - \frac{H(f)c(f)}{1+\frac{1}{RSB(f)}} - \frac{|c(f)|^2}{1+RSB(f)}\right)\frac{H(f)-c(f)^*}{1+\frac{1}{RSB(f)}}}{1 + \left(\frac{RSB(f)}{1+RSB(f)}\right)^2|H(f)-c(f)^*|^2 - |c(f)|^2} \tag{3.30}$$

Nous allons à présent étudier l'influence de l'environnement sur les performances obtenues en utilisant ce filtrage spatial. Cette étude va se faire en deux temps.

Nous allons d'abord présenter les expressions théoriques des critères de performance en fonction de la cohérence des bruits captés, de la propagation relative de la parole et du RSB d'entrée. Ces critères sont : la distorsion, le niveau de bruit résiduel, et le RSB en sortie du beamforming. Nous nous intéresserons ensuite aux tendances que l'on peut dégager sur ces critères en fonction de l'environnement, en mettant en regard les expressions théoriques présentées et les paramètres environnementaux pertinents.

Distorsion

On s'intéresse ici à la distorsion apportée par le traitement.

Pour un MVDR utilisant la méthode proposée pour l'estimation de la propagation relative du signal utile, cette distorsion vaut :

$$Disto_{MVDR}(f) = E\left[|S(f) - G_{MVDR}(f)^H H(f)S(f)|^2\right] = \varphi_s(f)E\left[\left|1 - G_{MVDR}(f)^H \frac{1}{H(f)}\right|^2\right] \tag{3.31}$$

En substituant $G_{MVDR}(f)$ par la valeur donnée dans l'Equation (3.30), on obtient :

$$Disto_{MVDR}(f) = \varphi_s(f)\left(\frac{1}{\frac{(1-|c(f)|^2)(1+RSB(f))^2}{|H(f)-c(f)^*|^2 RSB(f)} + RSB(f)}\right)^2 \tag{3.32}$$

où $RSB(f)$ est donné par l'équation (3.28).

On constate, comme l'on pouvait le déduire de l'équation (3.27), que la distorsion va être minimisée lorsque la cohérence sera faible. Elle sera aussi d'autant plus faible que $|H(f)-c(f)^*|^2$ est faible, et un RSB d'entrée fort permettra également de réduire cette distorsion.

On note alors :

$$R(f) = \frac{|H(f) - c(f)^*|^2}{1 - |c(f)|^2} \tag{3.33}$$

Nous donnerons des éléments d'analyse sur le comportement de cette quantité en fonction

de l'environnement dans la suite.

On peut alors exprimer la distorsion uniquement en fonction de $R(f)$ et du RSB d'entrée. La distorsion relative (normalisée par $\varphi_s(f)$) est donnée par :

$$\text{Disto}_r(f) = \frac{\text{Disto}_{MVDR}(f)}{\varphi_s(f)} = \frac{1}{\left(\frac{(1+\text{RSB}(f))^2}{R(f)\text{RSB}(f)} + \text{RSB}(f)\right)^2} \tag{3.34}$$

On trace donc la distorsion relative en fonction de ces deux grandeurs dans la Figure 3.5.

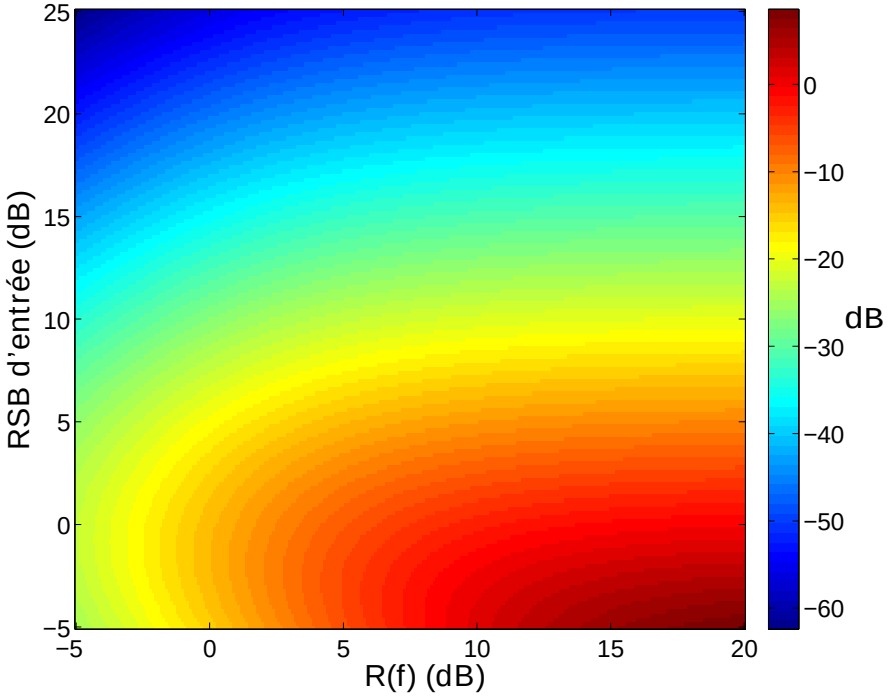


Figure 3.5 – Distorsion en sortie en fonction du RSB d'entrée et de $R(f)$

On constate notamment que cette distorsion est minimisée lorsque le RSB d'entrée est fort. Des éléments sur le comportement de $R(f)$ permettront de mieux appréhender l'influence de ce critère par la suite.

Bruit résiduel

Pour le MVDR adaptatif, on obtient :

$$\text{Bruit}_{MVDR}(f) = E\left[|G_{MVDR}(f)^H B(f)|^2\right] = \frac{1}{\|h(f)^H \Sigma_b(f)^{-1} h(f)\|} \tag{3.35}$$

Ce qui s'écrit :

$$\text{Bruit}_{MVDR}(f) = \varphi_b(f) \frac{1}{1 + \frac{\text{RSB}(f)^2 |H(f) - c(f)^*|^2}{(1+\text{RSB}(f))^2(1-|c(f)|^2)}} \tag{3.36}$$

Les conditions pour minimiser le bruit résiduel sont les inverses de celles permettant de

minimiser la distorsion. Il faut ici une forte cohérence pour obtenir un faible bruit en sortie. Il faut donc faire un compromis entre distorsion et bruit résiduel. On peut là aussi exprimer cette grandeur uniquement en fonction de $R(f)$ et du RSB d'entrée. On considère l'atténuation du bruit par rapport aux entrées, donnée par :

$$Bruit_{att}(f) = \frac{\varphi_b(f)}{Bruit_{MVDR}(f)} = 1 + \frac{RSB(f)^2 R(f)}{(1+RSB(f))^2} \tag{3.37}$$

et on trace cette grandeur en fonction de $R(f)$ et du RSB d'entrée dans la Figure 3.6.

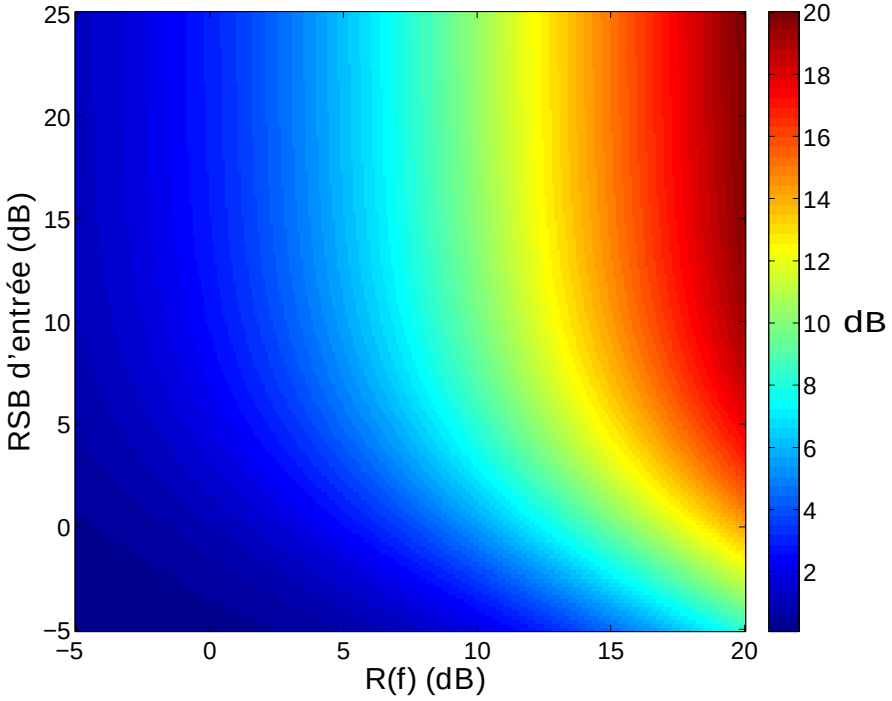


Figure 3.6 – Atténuation du bruit en fonction du RSB d'entrée et de $R(f)$.

RSB en sortie

On calcule le RSB de sortie, qui est le rapport entre l'énergie du signal utile filtré et l'énergie du bruit résiduel.

Pour le MVDR, on calcule le RSB de sortie de la façon suivante :

$$RSB_{MVDR}(f) = \frac{E\left[|G_{MVDR}(f)^H H(f) S(f)|^2\right]}{E\left[|G_{MVDR}(f)^H B(f)|^2\right]} \tag{3.38}$$

Cela vaut :

$$RSB_{MVDR}(f) = \left[\vec{H}(f)^H \Sigma_b(f)^{-1} \vec{H}(f) \varphi_s(f)\right] \left[\frac{1 + \frac{RSB(f)}{1+RSB(f)} |H - c(f)^*|^2 - |c(f)|^2}{\vec{H}(f)^H \Sigma_b(f)^{-1} \vec{H}(f)}\right]^2 \tag{3.39}$$

soit :

$$RSB_{MVDR}(f) = \frac{RSB(f)}{1-|c(f)|^2} \frac{1+\frac{RSB(f)}{1+RSB(f)}\frac{|H(f)-c(f)^*|^2}{1-|c(f)|^2}}{1+\frac{RSB(f)^2}{(1+RSB(f))^2}\frac{|H(f)-c(f)^*|^2}{1-|c(f)|^2}} \tag{3.40}$$

Remarquons que $\frac{RSB(f)}{1+RSB(f)} > \frac{RSB(f)^2}{(1+RSB(f))^2}$ car $\frac{RSB(f)}{1+RSB(f)} \in [0;1]$, et

$$\begin{aligned} 1+\frac{RSB(f)}{1+RSB(f)}\frac{|H(f)-c(f)^*|^2}{1-|c(f)|^2} &> 1+\frac{RSB(f)^2}{(1+RSB(f))^2}\frac{|H(f)-c(f)^*|^2}{1-|c(f)|^2} \\ &> 1+\underbrace{\frac{RSB(f)^2}{(1+RSB(f))^2}}_{>0}\frac{|H(f)-c(f)^*|^2}{1-|c(f)|^2} \end{aligned} \tag{3.41}$$

On peut en conclure que le RSB de sortie, dans le cas de ce MVDR adaptatif, est toujours supérieur au RSB d'entrée.

Lien avec l'acoustique

On remarque que dans les expressions des critères de performances présentés, la quantité $R(f)$, donnée dans l'équation (3.33), joue un rôle important. Pour mieux comprendre son comportement, on considère une situation de champ libre, sans réverbération. Dans ce cas, la propagation relative de la parole $H(f)$ correspond à un retard seul (on néglige les effets de l'atténuation), dépendant de la fréquence f, et de la distance parcourue par l'onde du signal utile entre les deux capteurs Δd :

$$H(f) = e^{-2i\pi f F_s \frac{\Delta d}{c_s}} \tag{3.42}$$

Dans ce cas, $R(f)$ ne dépend que de la phase de $H(f)$ et de la cohérence inter-capteurs du bruit, que l'on considère ici réelle (ce qui est le cas lorsque l'on est en présence d'un bruit diffus, en utilisant des capteurs omnidirectionnels par exemple).

On peut alors interpréter $f F_s \frac{\Delta d}{c_s}$ selon le placement des capteurs :

- Si les microphones sont placés en Broadside, Δd est nul sur toute la bande de fréquence : il n'y a pas de retard sur le signal utile entre les deux capteurs.
- Dans un autre cas, $f F_s \frac{\Delta d}{c_s}$ varie fortement. $f F_s \frac{\Delta d}{c_s}$ augmente le plus vite lorsque Δd est maximum : c'est lorsque les microphones sont en Endfire, et que la distance entre eux est grande.

Concernant la cohérence, on peut considérer que celle-ci sera proche de 1 en basses fréquences, et proche de 0 en hautes fréquences, comme on l'a vu lors de la présentation du champ de bruit diffus sphérique dans la Section 2.3 (page 42).

On présente les valeurs prises par $R(f)$, en fonction de $f F_s \frac{\Delta d}{c_s}$ et de la cohérence inter-capteurs des bruits $c(f)$, dans la Figure 3.7.

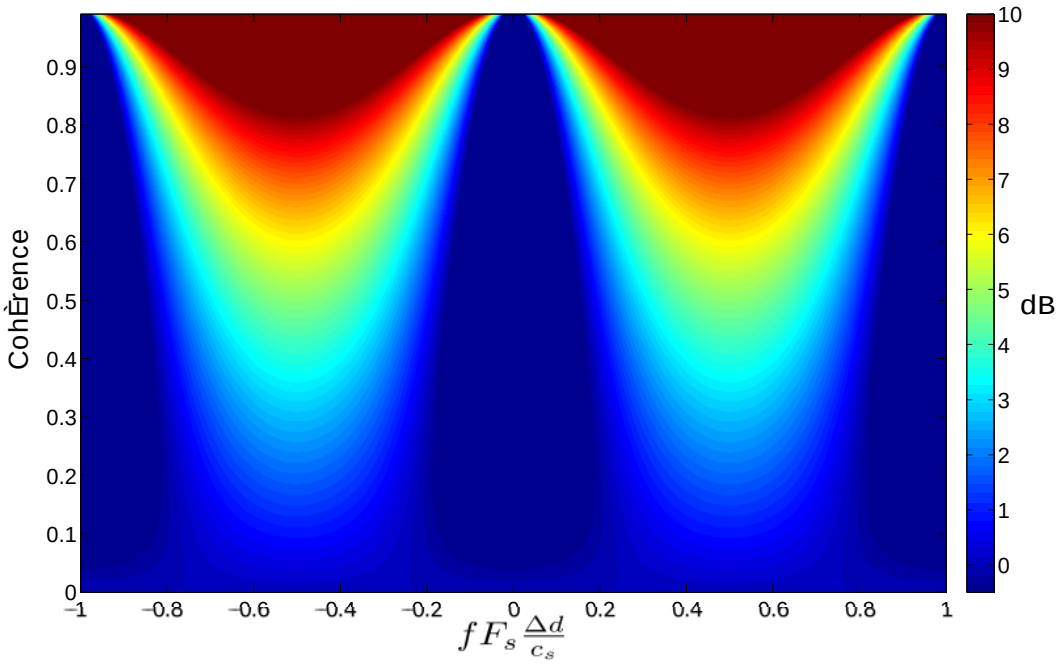


Figure 3.7 – Évolution de $R(f)$ en fonction de $f\,F_s\frac{\Delta d}{c_s}$ et de la cohérence inter-capteurs du bruit

On voit que lorsque la cohérence est forte, le dénominateur $1 - |c(f)|^2$ permet à $R(f)$ d'atteindre de grandes valeurs. En liant cette quantité aux interprétations acoustiques faites sur la phase de $H(f)$ et sur la Mean-Squared Coherence (MSC) inter-capteurs du bruit, on peut en déduire le comportement général de $R(f)$ en fonction de la fréquence et du placement des capteurs, que l'on résume dans le tableau 3.1.

Placement \ Bande de fréquence	Basses fréquences	Hautes fréquences
Broadside	Faible	Faible
Endfire	Très fort	Assez faible (varie avec la fréquence)

Table 3.1 – Comportement général des valeurs $R(f)$, selon le placement des capteurs et la bande de fréquences considérés

On peut à présent s'intéresser aux performances que l'on peut attendre, en fonction de $R(f)$ et du RSB d'entrée. Notamment, la distorsion relative (normalisée par la puissance du signal utile en entrée) et l'atténuation de bruit ne dépendent que de $R(f)$ et du RSB d'entrée.

En mettant les Figures 3.5 et 3.6 en regard de ce que l'on a vu sur les valeurs prises par $R(f)$ dans la Figure 3.7, et avec les observations sur le RSB d'entrée faites dans le chapitre 2, on peut comprendre dans quelles conditions l'on va se trouver, selon les bandes de fréquences :

— En basses fréquences, le RSB d'entrée est faible (de l'ordre de -3~0 dB), et l'on peut obtenir $R(f)$ plus ou moins grand selon le placement des capteurs. C'est une zone où cette méthode introduit une forte de distorsion, et ne permet pas une forte atténuation de bruit.

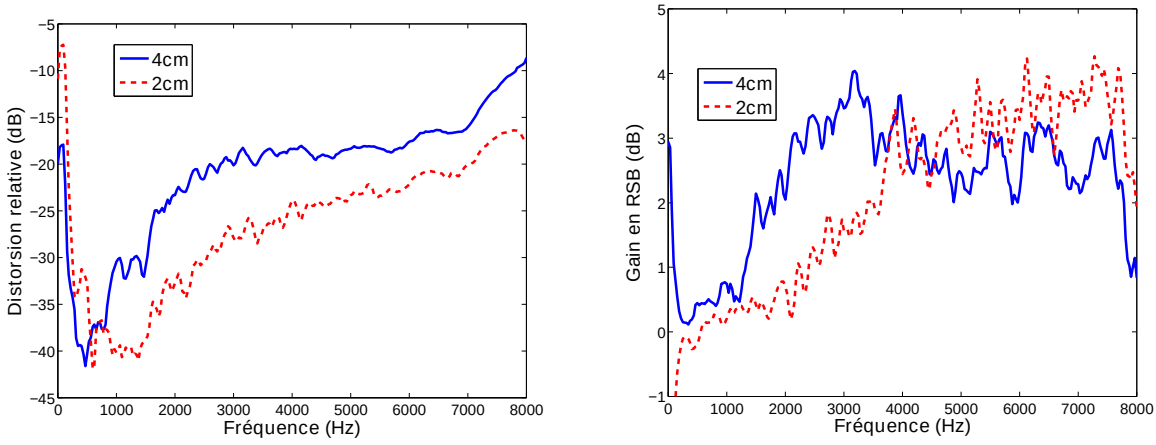
— En hautes fréquences, le RSB d'entrée est fort, mais $R(f)$ reste faible. On peut obtenir une atténuation de bruit de l'ordre de 3 dB (le maximum théorique atteignable avec

deux capteurs en utilisant un MVDR sur du bruit non-cohérent d'un capteur sur l'autre [Bitzer et al., 1998]), et une distorsion très faible.

Pour mieux comprendre ceci, nous avons simulé deux capteurs unidirectionnels (pour maximiser le RSB d'entrée) en champ libre, dans un champ de bruit diffus. Nous avons généré des bruits ayant la même DSP qu'un bruit d'autoroute enregistré en voiture avec un capteur cardioïde, et ayant une cohérence correspondant à un bruit diffus sphérique en utilisant la méthode présentée dans [Habets et al., 2008]. Nous avons simulé une source de signal utile blanche, placée à 40 cm des capteurs (dans une position Broadside).

En appliquant la méthode d'estimation de la propagation relative du signal utile proposée pour calculer un MVDR sur les signaux enregistrés par ces deux capteurs, séparément sur le bruit et le signal utile, nous avons pu estimer la distorsion relative introduite par le MVDR, ainsi que le gain en RSB que l'on obtient, par rapport aux signaux d'entrée, et ce pour chaque fréquence.

Les grandeurs estimées, pour des capteurs placés à 2 et 4 cm l'un de l'autre, sont présentées dans la Figure 3.8.



Distorsion relative (dB) Gain en RSB (dB)

Figure 3.8 – Performances obtenues pour le MVDR en bruit diffus simulant un bruit d'autoroute, en fonction de la fréquence, pour plusieurs distances.

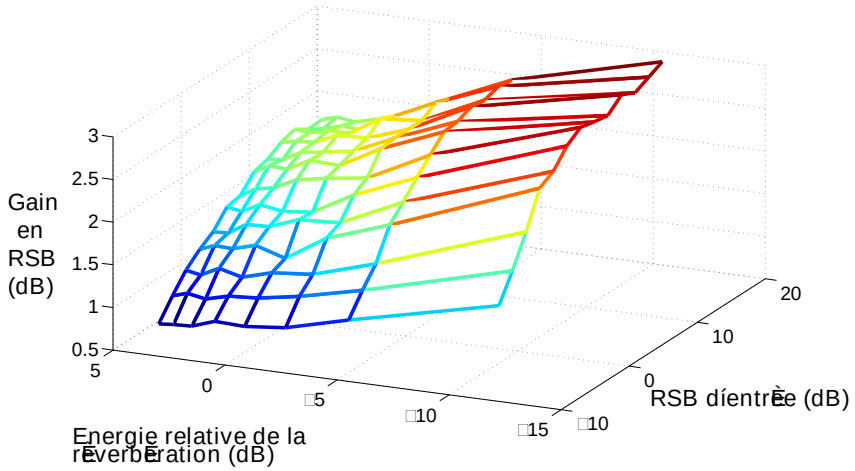
On remarque que l'on ne gagne en RSB qu'à partir de 1 kHz (pour l'antenne de capteurs à 4 cm l'un de l'autre). En dessous de cette fréquence, on n'a pas d'amélioration du RSB par rapport aux entrées, et l'intérêt d'utiliser cette méthode en dessous de cette fréquence est faible, même si c'est la bande de fréquences où la distorsion est la plus faible. Sur les courbes de gain en RSB, on constate également qu'à partir de 4 kHz (qui correspond à la fréquence de Nyquist pour l'échantillonnage spatial à 4 cm), l'antenne de microphones séparés de 2 cm présente un plus fort RSB que celle de microphones séparés de 4 cm : il est donc judicieux d'utiliser différentes antennes dans le cas de la téléphonie Wideband (où l'échantillonnage se fait à 16 kHz), utilisées dans différentes bandes de fréquence. Dans le cas présenté ici, il est judicieux d'utiliser l'antenne de 4 cm entre 1 kHz et 4 kHz, et l'antenne de 2 cm au-delà de 4 kHz.

Conclusion

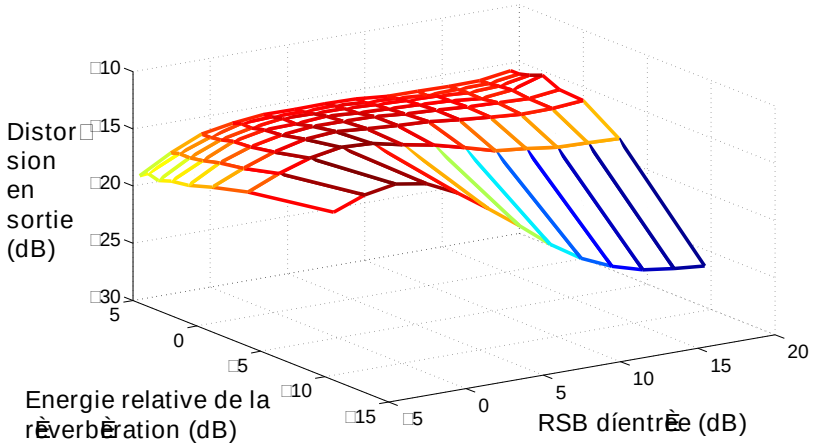
On peut donc voir que la méthode proposée va permettre au MVDR d'être Distortionless uniquement si le bruit présente une faible cohérence spatiale (de façon à avoir R(f) faible), et si le RSB d'entrée est favorable. D'un autre côté, si cette cohérence est forte, le bruit résiduel sera plus faible.

Cette méthode serait donc efficace lorsque les bruits sont faiblement cohérents, et lorsque le RSB d'entrée est favorable, comme le montrent les Figures 3.5 et 3.6.

Pour illustrer l'intérêt de cette méthode dans un cas réverbérant, on montre la même simulation que dans la Section 3.2.1, mais en utilisant cette fois la méthode proposée pour l'estimation de la propagation relative de la parole. Les bruits sont toujours supposés non-cohérents entre les capteurs, et les résultats sont montrés dans la Figure 3.9.



Gain en RSB



Distorsion apportée par le MVDR

Figure 3.9 – Résultats de simulation en utilisant la méthode proposée

On constate que les performances ne sont plus indépendantes du RSB d'entrée : ceci confirme que le RSB d'entrée doit être favorable pour espérer de bonnes performances. En contrepartie, la dépendance à l'énergie de la réverbération est nettement estompée, et les performances sont bien plus satisfaisantes lorsque l'on est en présence d'une forte réverbération, notamment en

termes de distorsion.

3.3 Estimation des matrices spectrales de bruit

On s'intéresse ici à l'estimation des matrices spectrales du bruit $b(t)$ à la fréquence f_k et à la trame temporelle n, et l'on note :

$$B_{m,n}(f_k) = \frac{1}{\sqrt{N}} \sum_{t=\lfloor nN \rfloor + 1}^{\lfloor nN \rfloor + N} b_m(t) e^{-2i\pi f_k t} \tag{3.43}$$

et $B_n(f_k) = [B_{1,n}(f_k) \dots B_{M,n}(f_k)]^T$.

On suppose que $b(t)$ est un processus réel, stationnaire au sens large et gaussien. Dans ce cas, si la taille N de la fenêtre est suffisamment grande, on sait que :

$$E\left[B_n(f_k) B_n(f_{k'})^H\right] \sim \Sigma_b(f_k)\delta_{k,k'} \tag{3.44}$$

et on peut construire un estimateur de $\Sigma_b(f_k)$ basé sur le périodogramme moyenné en temps. On utilise alors une moyenne sur les trames sur une fenêtre exponentielle glissante, ce qui revient à utiliser un estimateur récursif. Ainsi, l'estimateur à la fréquence f_k et à la trame n s'écrit :

$$\hat{\Sigma}_b(f_k, n) = \alpha \hat{\Sigma}_b(f_k, n-1) + (1-\alpha) B_n(f_k) B_n(f_k)^H \tag{3.45}$$

$\alpha \in]0, 1[$ étant un facteur d'oubli exponentiel. Des éléments sur l'estimation de matrices spectrales peuvent sont présentés dans l'Annexe A.

Une difficulté de cette méthode est que l'on ne dispose pas des signaux de bruit seuls. On choisit donc de faire cette adaptation sur les phases de bruit seul, sur lesquelles le locuteur ne parle pas.

Les matrices spectrales $\Sigma_b(f_k, n)$ sont estimées en utilisant la SPP introduite dans la Section 3.2.2 dans l'équation de mise à jour suivante :

$$\hat{\Sigma}_b(f_k, n) = \alpha_n(f_k) \hat{\Sigma}_b(f_k, n-1) + (1-\alpha_n(f_k)) X_n(f_k) X_n(f_k)^H \tag{3.46}$$

$$\alpha_n(f_k) = \alpha_0 + (1-\alpha_0) SPP_n(f_k) \tag{3.47}$$

où α_0 est un facteur d'oubli choisi expérimentalement. Cette méthode permet d'adapter la matrice spectrale des bruits plus rapidement si la composante à la fréquence considérée ne contient que du bruit, et de bloquer l'adaptation s'il contient de la parole. Toutefois, l'estimation de la SPP n'est pas instantanée : il y a un délai entre le début d'une période de parole et la montée de cette probabilité. Ainsi, on a une apparition de "fuites" de voix dans les matrices spectrales du bruit, ce qui semble provoquer une distorsion sur la voix, qualifiée de "reverb" ou de "métallisation" lors d'écoutes informelles. Cette fuite est plus gênante lorsque le RSB d'entrée est favorable, car l'énergie de la fuite est plus importante que le bruit estimé.

Pour contrer ce défaut, un test sur le RSB à posteriori a été introduit. Le RSB a posteriori est défini par $RSB_{post}(f_k, n) = \frac{|X_{1,n}(f_k)|^2}{\varphi_{b_1}(f_k)}$.

L'avantage de cet indicateur est qu'il dépend directement de l'observation : il augmente

tout de suite dès que l'énergie de l'observation augmente, ce qui est le cas lorsqu'une phase de parole commence. Les algorithmes Parrot utilisent déjà une estimation de $\varphi_{b_1}(f_k)$. Cette quantité $RSB_{post}(f_k, n)$ est donc disponible.

Il est à noter que l'estimation de $\varphi_{b_1}(f_k)$ étant particulièrement basée sur une hypothèse de stationnarité du bruit, cette quantité augmente aussi lors de présence de bruit instationnaire.

La nouvelle estimation des matrices spectrales du bruit devient donc :

Algorithme 1 Estimation de $\Sigma_b(f_k, n)$

```
if RSB_post(f_k, n) < Ω then
    α_n(f_k) = α_0 + (1 − α_0)SPP(f_k, n)
    Σ̂_b(f_k, n) = α_n(f_k)Σ̂_b(f_k, n − 1) + (1 − α_n(f_k))X_n(f_k)X_n^H(f_k)
else
    Σ̂_b(f_k, n) = Σ̂_b(f_k, n − 1)
end if
```

Dans l'Algorithme 1, Ω est un seuil défini expérimentalement. Il est choisi assez faible, pour être sûr de bloquer l'adaptation lors de présence de parole, quitte à ne pas mettre à jour sur certaines périodes de bruit.

Pour se rendre compte de l'efficacité de cette méthode, un test d'écoute a été mené. Il s'agit d'un test par paires, c'est à dire que l'on fait écouter deux sons aux sujets : le même signal d'entrée traité par deux algorithmes différents. Le sujet indique alors quel son a, selon lui, la meilleure qualité, tout en pouvant écouter chaque son autant de fois qu'il veut. Les sons sont désignés "A" et "B", et les différents traitements sont mélangés : le sujet ne sait jamais quel algorithme correspond aux sons qu'il entend. Lors de ce test, chacun des 20 sujets a écouté et noté 8 paires : il y avait 8 signaux d'entrée différents, et 2 algorithmes testés.

Les 2 algorithmes sont, avec leur désignation :

UAB1 Un algorithme utilisant uniquement l'Équation (3.46) pour l'estimation de Σ_b

UAB2 Le même algorithme, mais utilisant aussi le test de RSB présenté dans l'Algorithme 1.

Les algorithmes sont ensuite notés les uns par rapport aux autres, en regardant pour chaque paire d'algorithmes dans quelle mesure l'un a été préféré à l'autre. Les résultats pour les deux systèmes considérés sont présentés dans la Table 3.2.

UAB1	UAB2
77%	23%

Table 3.2 – Pourcentage de préférence pour chaque méthode d'estimation des matrices spectrales de bruit.

On remarque une préférence marquée pour le système utilisant le test sur le RSB *a posteriori* en plus de la SPP pour l'estimation des matrices spectrales du bruit. Ce test permet donc de limiter les fuites de paroles dans ces matrices, et d'atténuer de façon audible les effets néfastes que cela engendre.

3.4 Considérations sur le placement des capteurs

Dans le cas d'un *beamforming* MVDR, il est généralement avantageux d'utiliser un placement *Endfire* pour les capteurs [Brandstein and Ward, 2001, Yu and Hansen, 2010]. Or, en utilisant

la méthode proposée, notamment pour l'estimation de la propagation de la parole, ce n'est pas forcément le cas. Cette section a pour but de mieux comprendre comment le placement influe sur les performances des algorithmes utilisés, selon les critères présentés dans la section 3.2.3 (page 70).

3.4.1 Simulation de placement

On suppose une situation où l'on dispose de deux capteurs, qui sont unidirectionnels ou omnidirectionnels. L'un des microphones (le capteur de référence) est placé à l'origine d'un repère, avec la source de signal utile placée à une distance d de ce capteur (voir Figure 3.10).

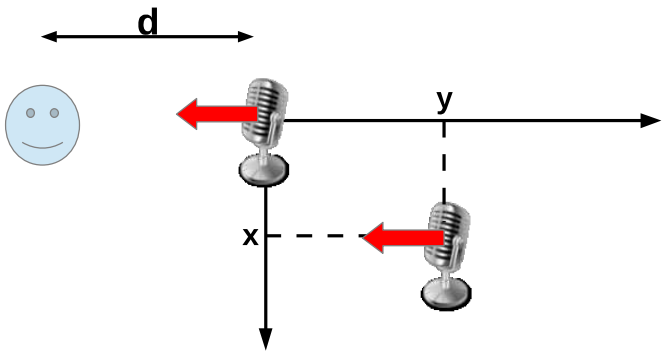


Figure 3.10 – Situation envisagée pour les simulations. Les flèches rouges représentent l'axe de directivité lorsque l'on utilise des microphones cardioïdes.

On suppose une condition de champ libre. Le transfert relatif pour le signal utile entre les deux capteurs correspond alors à un retard pur. En fréquence, cela correspond, à $H(f) = \alpha e^{-2i\pi f F_s}$, où τ est le retard relatif et vaut $\tau = \frac{\Delta d}{c_s} = \frac{\sqrt{x^2+(d+y)^2} - d}{c_s}$, et où c_s est la célérité du son. α correspond à l'atténuation due à la différence de distance à la source de signal utile R_{en} et à la directivité des microphones g, lorsqu'il y en a une (lorsque l'on utilise des capteurs cardioïdes).

$$\begin{aligned} \alpha &= R_{en}\, g \\ R_{en} &= \frac{\sqrt{(d+y)^2+x^2}}{d} \\ g &= \begin{cases} \frac{1}{2}\left(1+\frac{1}{1+(\frac{x}{y+d})^2}\right) & \text{Si unidirectionnel} \\ 1 & \text{Si omnidirectionnel} \end{cases} \end{aligned} \tag{3.48}$$

On suppose également un champ de bruit diffus : ainsi, la cohérence des bruits présents sur les capteurs peut être donnée en fonction du type de capteur utilisé et du placement des capteurs, grâce au modèle présenté dans la Section 2.3 (page 42).

La DSP du bruit sur chaque microphone dépend du type de capteur. Celle-ci correspond à la DSP du bruit mesuré dans une voiture roulant sur autoroute, et sa forme est donnée, pour des microphones omnidirectionnels et unidirectionnels, dans la Figure 3.11.

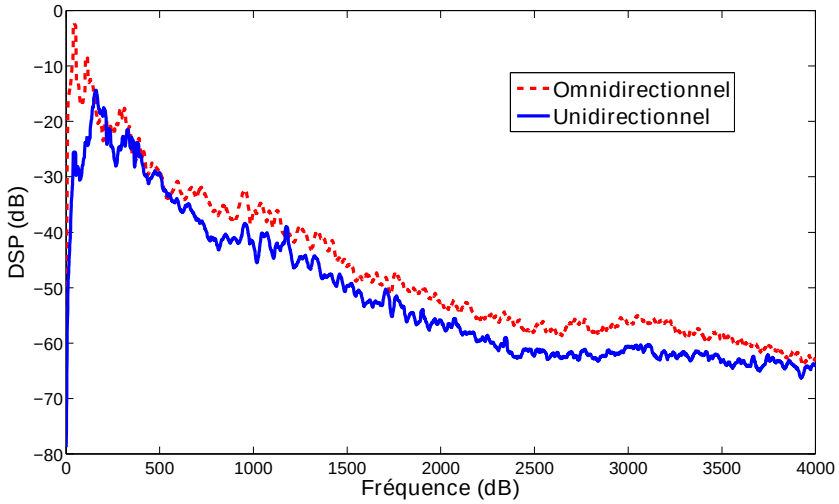


Figure 3.11 – DSP utilisées pour les bruits, pour les deux types de capteurs envisagés

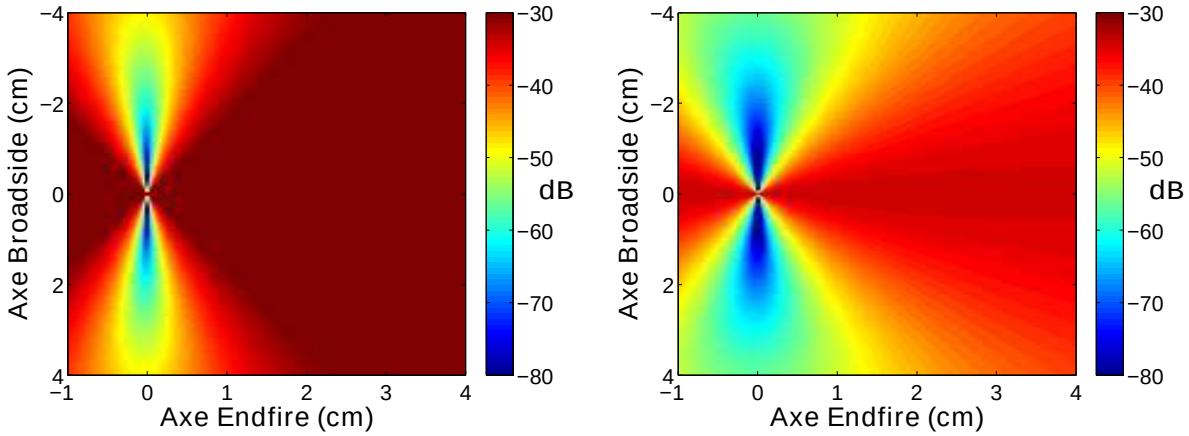
On va ici comparer les critères de performance présentés dans la Section 3.2.3, dans le cas où les capteurs utilisés sont deux microphones omnidirectionnels ou deux microphones unidirectionnels. On intègre les critères de performances sur la bande de fréquence [1 kHz 4 kHz]. Cette bande de fréquences a été choisie car on n'obtient pas de gain significatif avec cette méthode en dessous de 1 kHz (voir Figure 3.8), et l'algorithme Parrot pour la téléphonie Wideband sépare le signal en deux bandes : une bande entre 0 et 4 kHz, et une seconde entre 4 et 8 kHz, notamment pour des raisons de coût de calcul. On peut donc construire une antenne de capteurs pour chacune de ces bandes.

On restreint l'écartement des capteurs en dessous de 4 cm, qui correspond à la fréquence de Nyquist pour l'échantillonnage spatial à 4 kHz. Au delà de cette distance, les performances se détériorent, comme montré dans la Figure 3.8.

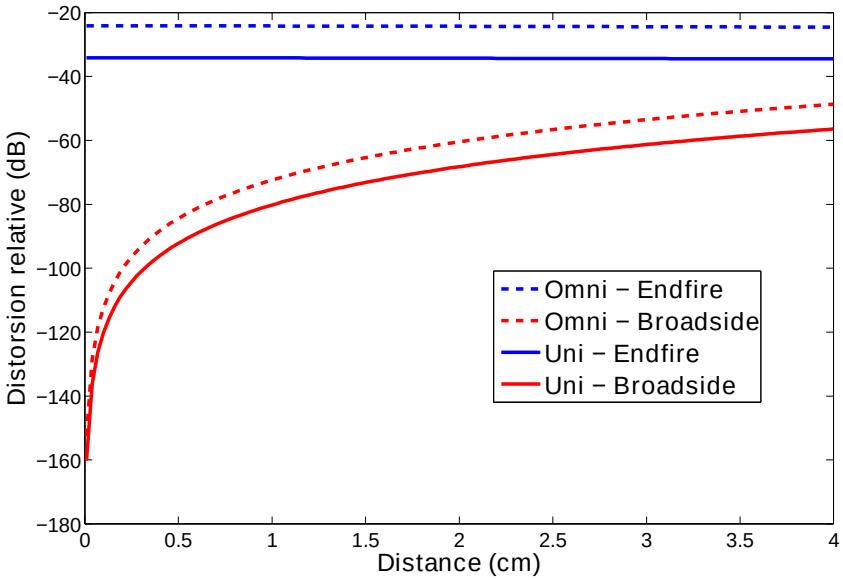
Pour chaque critère, on présente les performances que l'on peut attendre dans le plan (le second microphone est placé dans la zone $x \in [-4\text{cm} \quad 4\text{ cm}]$, $y \in [-1\text{ cm} \quad 4\text{ cm}]$), et les coupes sur les 2 axes $x = 0$ et $y = 0$ qui correspondent aux situations Endfire et Broadside , respectivement.

3.4.2 Distorsion

La distorsion est normalisée par l'énergie du signal utile, et son tracé est visible dans la Figure 3.12. On cherche à la rendre la plus petite possible, car une distorsion de $-\infty$ dB signifie que le signal utile reste intact.



Capteurs omnidirectionnels Capteurs unidirectionnels



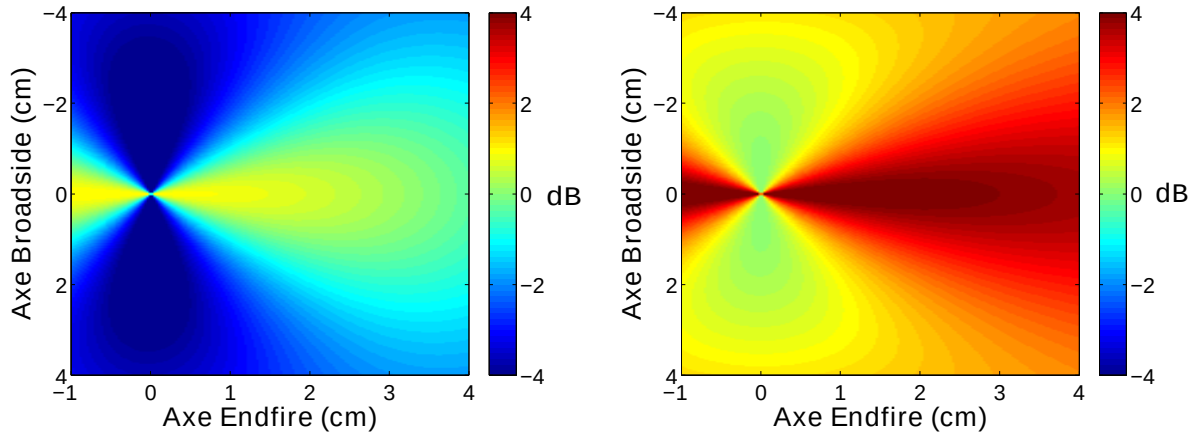
Coupes sur les 2 axes : **Endfire** et **Broadside**

Figure 3.12 – Distorsion pour le MVDR

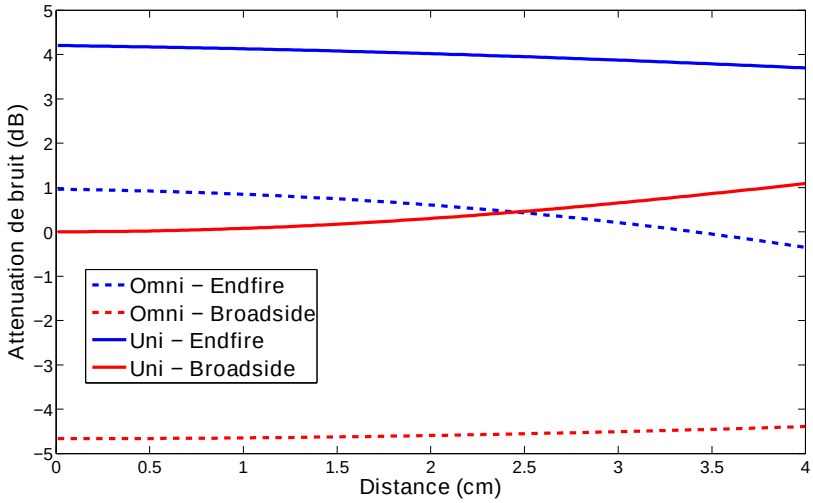
On remarque que le placement **Broadside** est largement préférable si l'on souhaite minimiser la distorsion en sortie, et ce quel que soit les capteurs utilisés. Il faut mettre ceci en balance avec les gains que l'on peut espérer sur la réduction de bruit et l'amélioration de RSB.

3.4.3 Bruit résiduel

On présente ici l'atténuation de bruit par rapport à l'énergie du bruit présent sur un capteur unidirectionnel (qui est le système d'acquisition de référence). C'est un critère que l'on cherche à maximiser. L'atténuation de bruit en fonction du placement pour un traitement MVDR est présenté dans la Figure 3.13.



Capteurs omnidirectionnels Capteurs unidirectionnels



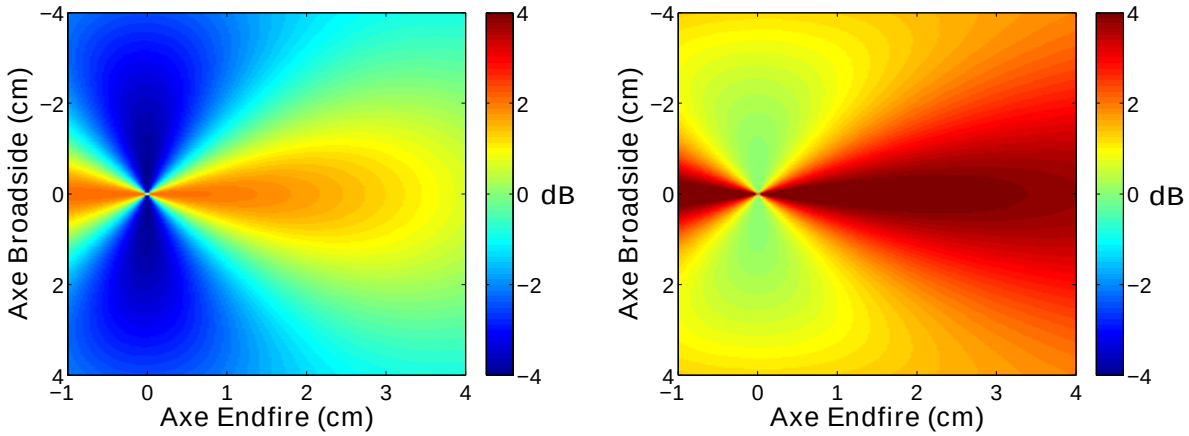
Coupes sur les 2 axes : **Endfire** et **Broadside**

Figure 3.13 – Atténuation de bruit pour le MVDR

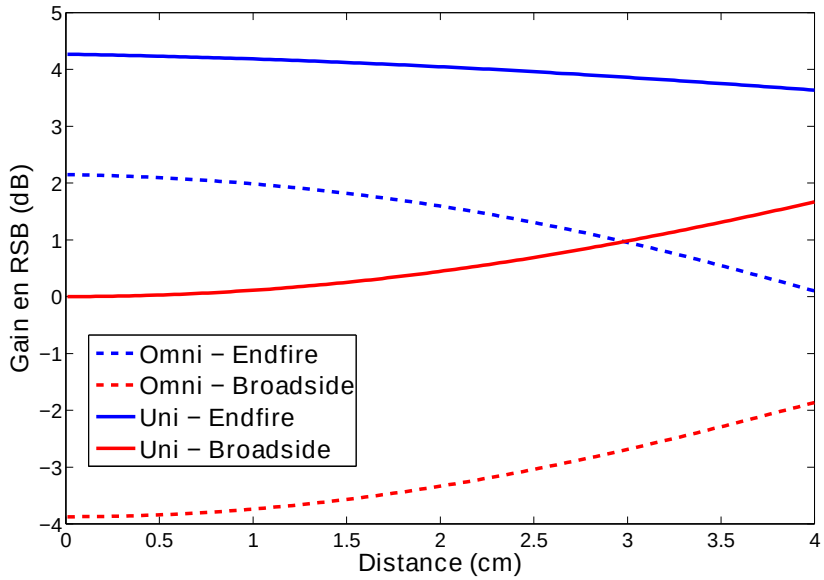
La différence de RSB d'entrée entre capteurs omnidirectionnel et cardioïde joue ici grandement : les omnidirectionnels ne permettent pas toujours de réduire le bruit par rapport à un unique capteur cardioïde, étant désavantagés dès le départ par leur absence de directivité. Dans tous les cas, le placement **Endfire** permet de mieux minimiser le bruit, mais c'est au prix d'une plus forte distorsion.

3.4.4 RSB en sortie

On présente ici le gain en RSB par rapport au RSB en entrée sur un capteur unidirectionnel, et ce critère est présenté en fonction du placement dans la Figure 3.14. On cherche à maximiser ce gain en RSB.



Capteurs omnidirectionnels Capteurs unidirectionnels



Coupes sur les 2 axes : Endfire et Broadside

Figure 3.14 – RSB de sortie pour le MVDR

On voit que là aussi, pour de faibles distances, le placement Endfire est préférable. Toutefois, en augmentant la distance entre capteurs, on obtient des performances intéressantes pour le placement Broadside . Dans le cas de capteurs omnidirectionnels, on constate que l'écart de RSB d'entrée avec des capteurs cardioïdes est très pénalisant : on n'obtient d'amélioration de RSB par rapport à un unique cardioïde que dans certaines situations, alors que la combinaison utilisant des cardioïdes améliore toujours le RSB par rapport à un unique microphone, comme on pouvait le prédire à partir de l'Équation (3.40) (page 74).

3.4.5 Conclusion

Du fait de leur directivité, les capteurs cardioïdes ont un avantage en terme de RSB d'entrée lorsque l'on est dans des fréquences au dessus de 1 kHz, comme on l'a vu dans le Chapitre 2. Or c'est dans cette bande de fréquences que ces simulations ont été réalisées, et l'avantage de ces capteurs apparaît clairement dans les performances atteignables : il est dans tous les cas judicieux d'utiliser des capteurs cardioïdes.

En ce qui concerne le placement, c'est le placement Endfire qui permet la meilleure réduction de bruit, et le plus fort gain en RSB. Mais c'est aussi sur cet axe que l'on a la plus forte distorsion.

Pour un placement Broadside , les performances en termes de réduction de bruit et de gain en RSB s'améliorent à mesure que l'on éloigne les capteurs, et la distorsion reste très faible, par rapport au placement Endfire (à 4cm, on gagne 20 dB de distorsion par rapport à un placement Endfire).

On choisit donc d'utiliser des capteurs cardioïdes en position Broadside , avec une distance la plus grande permise. À 4 kHz, la demi-longueur d'onde est de 4 cm : c'est le maximum que l'on puisse se permettre pour respecter les conditions d'échantillonnage spatial.

3.5 Synthèse

On a vu que la réverbération dans l'habitacle automobile ne permettait pas d'utiliser des méthodes d'estimation de la propagation relative de la parole basée uniquement sur la position de la source de parole et de la géométrie de l'antenne. Nous avons présenté une solution adaptative simple pour estimer cette propagation, qui est plus robuste à la réverbération, et qui permet de suivre les variations de cette propagation.

Toutefois, cette méthode a ses limites :

— Lorsque l'on a du bruit cohérent d'un capteur sur l'autre, cela provoque l'apparition de distorsion sur le signal débruité.

— Cette méthode est peu robuste aux RSBs d'entrée faibles.

C'est pourquoi nous envisageons de l'utiliser dans la bande de fréquences au-delà de 1 kHz : ainsi, on limite l'impact de la cohérence du bruit, et on se place dans des conditions de RSB d'entrée plus favorables, surtout lorsque l'on utilise des capteurs cardioïdes.

Le placement des capteurs a également une forte influence sur les performances : une étude dans des conditions simples nous amène à choisir un placement Broadside, avec des capteurs éloignés de 4 cm dans la bande [1 kHz, 4 kHz].

De plus, nous avons présenté une méthode adaptative pour estimer les matrices spectrales du bruit nécessaires au MVDR, robuste aux attaques de parole, qui est plus efficace qu'une moyenne récursive basée sur une probabilité de présence de parole.

Ces méthodes combinées nous permettent d'implémenter un beamforming MVDR adaptatif pour les fréquences supérieures à 1 kHz, mais il faut trouver d'autres méthodes pour les basses fréquences, robustes à de faibles RSBs d'entrée et à des bruits fortement cohérents d'un capteur sur l'autre.

Chapitre 4

Annulation de bruit adaptative résistante aux fuites de parole (CR-ANC)

Sommaire

Ce Chapitre est dédié à l'étude d'un système d'Annulation de bruit adaptative (Adaptive Noise Cancellation) (ANC), qui prend en compte le fait que la parole est présente sur tous les capteurs utilisés. Une étude sur les performances atteintes est conduite, afin de déterminer la meilleure implémentation algorithmique et acoustique pour un tel système, dans le cadre de la téléphonie en automobile. Cette étude est conduite expérimentalement, sur des signaux enregistrés en voiture.

Nous présentons ici en détail le système de Annulation de bruit adaptative résistante aux fuites de Parole (Crosstalk Resistant Adaptive Noise Cancellation) (CR-ANC), que nous avons évoqué dans la Section 1.4.1 (page 25). Nous commençons par rappeler le principe général de cette méthode, qui se compose d'un étage d'ANC (qui permet de réduire le niveau de bruit, mais apporte une forte distorsion), suivi d'un étage de compensation de distorsion. On s'intéresse plus précisément à un système à trois microphones :

— un capteur de référence,

— deux capteurs permettant de réaliser l'ANC.

Nous nous intéressons ensuite aux performances atteintes en sortie de l'étage de compensation de distorsion. Plus précisément, nous étudions le lien entre ces performances et les caractéristiques du signal en sortie de l'étage d'ANC. Une étude de l'étage d'ANC nous permet alors de comprendre comment obtenir un signal en sortie de cet étage permettant de maximiser les performances globales du système, en fonction de l'environnement.

Ces performances dépendent grandement de l'implémentation faite de ce système, ainsi que de la configuration acoustique choisie. C'est pourquoi nous présentons dans la suite une étude expérimentale, qui permet de choisir conjointement les paramètres d'implémentation pertinents et une configuration acoustique cohérente. Cette étude est faite sur l'étage d'ANC, puis sur l'étage de compensation de distorsion.

4.1 Principe général

On rappelle ici le principe général de l'ANC dans le cas où le signal de parole est présent sur tous les signaux captés. Un scéma global, pour un système à 3 microphones, est présenté dans la Figure 4.1. On considère 3 entrées $x_m(t)$, $m \in [1 \square 3]$, qui s'écrivent à l'instant discret t :

$$x_m(t) = \{h_m^s \otimes s\}(t) + b_m(t) \tag{4.1}$$

où $h_m^s(t)$ représente la propagation du signal de parole entre la source et le capteur m, $s(t)$ est le signal de parole et $b_m(t)$ est le bruit additif sur le capteur m. Comme dans le Chapitre 3, on suppose $h_1^s(t) = \delta(t)$, δ étant l'impulsion de Dirac, pour des raisons d'identifiabilité du canal acoustique.

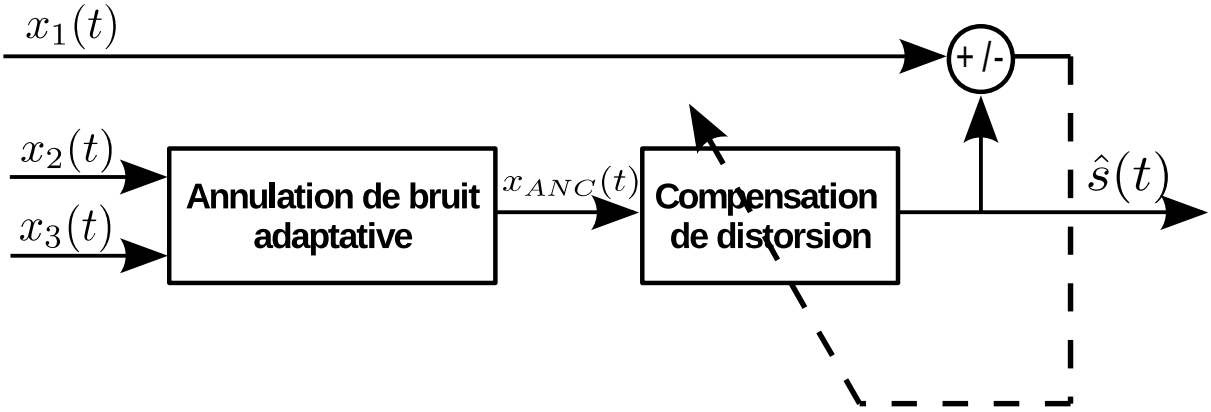


Figure 4.1 – Système d'annulation de bruit à 3 microphones [Zeng and Abdulla, 2006].

Ce système est composé de deux étages : un premier d'ANC, et un de compensation de distorsion.

Ce dernier est indispensable, car l'ANC, en présence de parole sur toutes les entrées (comme c'est le cas dans notre application) provoque une forte distorsion du signal utile sur sa sortie, notée $x_{ANC}(t)$ [Widrow et al., 1975]. Ce signal $x_{ANC}(t)$ est donc composé d'un signal de parole filtré, ainsi que d'un bruit résiduel :

$$x_{ANC}(t) = \underbrace{\{h_d \otimes s\}(t)}_{\text{parole filtrée}} + \underbrace{b_{ANC}(t)}_{\text{bruit résiduel}} \tag{4.2}$$

On constate que les caractéristiques du signal débruité $\hat{s}(t)$ vont dépendre des propriétés du signal de sortie de l'étage d'annulation de bruit ($x_{ANC}(t)$), d'une part, et du signal capté de référence $x_1(t)$, d'autre part. Nous allons étudier le comportement de la compensation de distorsion en fonction des caractéristiques du signal de sortie de l'ANC, avant de nous intéresser aux propriétés de ce signal en fonction de l'environnement. Ensuite, nous confronterons ces propriétés à l'environnement automobile pour comprendre comment se mettre dans la meilleure situation pour que ce système présente les meilleures performances possibles. Enfin, nous présenterons les performances atteintes par le système global.

4.2 Compensation de distorsion

On s'intéresse ici au deuxième étage adaptatif de la Figure 4.1, qui dépend de la sortie de l'étage d'annulation de bruit (x_{ANC}) et du capteur de référence, numéroté 1.

Ces deux signaux sont supposés ainsi :

$x_1(t)$ contient le signal utile (que l'on souhaite estimer) et un bruit additif ;

$x_{ANC}(t)$ contient le signal utile filtré (du fait d'une distorsion linéaire), et un bruit additif.

On remarque que l'on se place ici dans un cadre général où l'on suppose que le bruit résiduel sur la sortie de l'ANC peut être cohérent avec le bruit présent sur le signal de référence.

On écrit ces entrées :

$$\begin{aligned} x_1(t) &= s(t) + b(t) + u(t) \\ x_{ANC}(t) &= \{h_d \otimes s\}(t) + \{h_b \otimes b\}(t) \end{aligned} \tag{4.3}$$

où $s(t)$ est le signal utile, $h_d(t)$ est le filtre correspondant à la distorsion du signal utile sur x_{ANC}. $b(t) + u(t)$ est le bruit additif présent sur $x_1(t)$, séparé en une composante $b(t)$, cohérente avec le bruit présent sur $x_{ANC}(t)$, $\{h_b \otimes b\}(t)$, et l'autre non-cohérente avec le bruit résiduel sur $x_{ANC}(t)$, $u(t)$.

Fréquentiellement, ces signaux s'écrivent :

$$\begin{aligned} X_1(f) &= S(f) + B(f) + \Upsilon(f) \\ X_{ANC}(f) &= H_d(f)S(f) + H_b(f)B(f) \end{aligned} \tag{4.4}$$

où $X_1(f)$, $X_{ANC}(f)$, $S(f)$, $B(f)$ et $\Upsilon(f)$ sont les transformées de Fourier de $x_1(t)$, $x_{ANC}(t)$, $s(t)$, $b(t)$ et $u(t)$, respectivement. $H_d(f)$ et $H_b(f)$ sont les transformées de Fourier de $h_d(t)$ et

$h_b(t)$, respectivement.

La compensation de distorsion consiste en un filtrage adaptatif entre l'entrée $x_{ANC}(t)$ et $x_1(t)$, comme illustré sur la Figure 4.2.

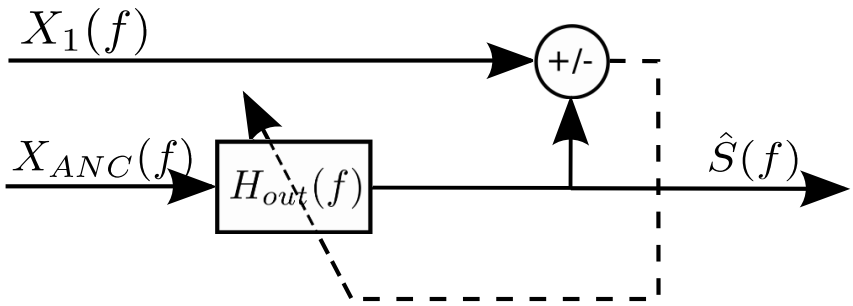


Figure 4.2 – Compensation de distorsion

Si on suppose des signaux $S(f)$, $B(f)$ et $\Upsilon(f)$ stationnaires, et que la convergence du filtrage adaptatif est atteinte, ce filtre adaptatif converge vers la solution de Wiener [3]

$$H_{out}(f) = \frac{H_d(f)\varphi_s(f) + H_b(f)\varphi_b(f)}{|H_d(f)|^2\varphi_s(f) + |H_b(f)|^2\varphi_b(f)} \tag{4.5}$$

où $\varphi_s(f)$ et $\varphi_b(f)$ sont les Densités Spectrales de Puissance (DSPs) de $s(t)$ et $b(t)$, respectivement.

Le signal débruité est donné par la sortie de ce filtre, et vaut $\hat{S}(f) = H_{out}(f)^* X_{ANC}(f)$

On peut, à partir de l'expression de ce filtre, s'intéresser aux performances que l'on obtient sur le signal débruité, en termes de distorsion, de bruit résiduel, et de RSB.

4.2.1 Distorsion

On s'intéresse ici à la distorsion linéaire présente sur le signal de sortie, en fonction des entrées de l'étage de compensation de distorsion.

On écrit le signal de sortie $\hat{S}(f)$:

$$\hat{S}(f) = H_{out}(f)^* X_{ANC}(f) = \left[\frac{H_d(f)\varphi_s(f) + H_b(f)\varphi_b(f)}{|H_d(f)|^2\varphi_s(f) + |H_b(f)|^2\varphi_b(f)}\right]^* (H_d(f)S(f) + H_b(f)B(f)) \tag{4.6}$$

et on s'intéresse aux composantes de signal utile et de bruit résiduel présents sur ce signal, séparément.

La composante utile a pour expression :

$$\hat{S}^{sig}(f) = H_{out}(f)^* H_d(f)S(f) = \left[\frac{H_d(f)\varphi_s(f) + H_b(f)\varphi_b(f)}{|H_d(f)|^2\varphi_s(f) + |H_b(f)|^2\varphi_b(f)}\right]^* H_d(f)S(f) \tag{4.7}$$

On introduit la valeur du RSB sur le signal $X_{ANC}(f)$, que l'on appelle $RSB_{ANC}(f)$, défini

3. Ceci est un cas idéal, supposant que l'on s'autorise un filtre non-causal de longueur infinie.

par :

$$RSB_{ANC}(f) = \frac{|H_d(f)|^2 \varphi_s(f)}{|H_b(f)|^2 \varphi_b(f)} \quad (4.8)$$

On peut écrire la composante de signal utile sur le signal débruité :

$$\hat{S}^{sig}(f) = \left[\frac{1 + \frac{H_b(f)}{H_d(f) RSB_{ANC}(f)}}{1 + \frac{1}{RSB_{ANC}(f)}} \right] S(f) \quad (4.9)$$

On définit la distorsion normalisée présente sur la sortie par :

$$Disto(f) = \frac{E\left[|S(f) - \hat{S}^{sig}(f)|^2 \right]}{\varphi_s(f)} \quad (4.10)$$

En utilisant l'équation (4.9), on obtient :

$$Disto(f) = \left[\frac{1 - \frac{H_b(f)}{H_d(f)}}{1 + RSB_{ANC}(f)} \right]^2 \quad (4.11)$$

On remarque que cette distorsion est fonction du RSB sur le signal de sortie de l'ANC x_{ANC}. Il faut donc maximiser ce RSB pour avoir une faible distorsion.

4.2.2 RSB en sortie

Le signal débruité étant un filtrage du signal de sortie de l'ANC, le RSB en sortie sera le même que celui sur le signal de sortie de l'ANC.

Les performances du système sont donc largement dépendantes du RSB sur le signal de sortie de l'ANC. Il faut parvenir à ce que ce RSB soit haut, même si le signal utile présent sur x_{ANC} est déformé.

4.3 Annulation de bruit adaptative

On a donc vu que pour avoir les meilleures performances possibles, il faut obtenir en sortie d'ANC un signal ayant un fort RSB.

On s'intéresse donc à un système d'ANC à deux microphones, dont les entrées s'écrivent :

$$\begin{aligned} x_2(t) &= s(t) + b_2(t) \\ x_3(t) &= \{h_s \otimes s\}(t) + b_3(t) \end{aligned} \quad (4.12)$$

où $b_2(t)$ et $b_3(t)$ sont les bruits additifs, $s(t)$ le signal utile et $h_s(t)$ est la propagation relative du signal utile.

En fréquences, cela s'écrit :

$$X_2(f) = S(f) + B_2(f)$$
$$X_3(f) = H_s(f)S(f) + B_3(f) \tag{4.13}$$

où $X_2(f)$, $X_3(f)$, $S(f)$, $B(f)$ et $H_s(f)$ sont les transformées de Fourier de $x_2(t)$, $x_3(t)$, $s(t)$ et $b(t)$ et $h_s(t)$, respectivement.

Lorsque le signal utile est absent, durant les périodes où le locuteur ne parle pas, les entrées s'écrivent :

$$X_2(f) = B_2(f)$$
$$X_3(f) = B_3(f) \tag{4.14}$$

L'annulation de bruit adaptative consiste à estimer un filtre entre ces bruits seuls par un algorithme de filtrage adaptatif [Harrison et al., 1986], comme illustré dans la Figure 4.3.

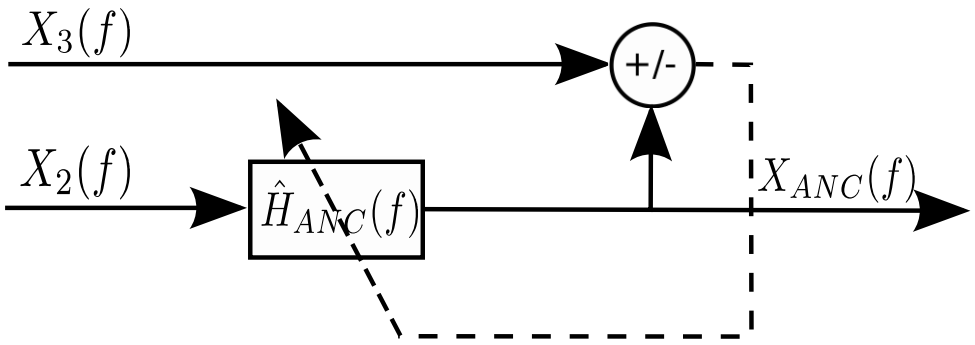


Figure 4.3 – Annulation de bruit adaptative

Si $B_2(f)$ et $B_3(f)$ sont stationnaires, et si l'on utilise un algorithme adapté, ce filtre converge vers la solution de Wiener :

$$\hat{H}_{ANC}(f) = \frac{\varphi_{b_2 b_3}(f)}{\varphi_{b_2}(f)} \tag{4.15}$$

4.3.1 Atténuation du bruit

On s'intéresse ici à la capacité d'un tel système à réduire l'énergie du bruit présent sur les capteurs, sans s'intéresser aux conséquences sur le signal de parole.

Le bruit en sortie est donné par :

$$X^b_{ANC}(f) = B_3(f) - \hat{H}_{ANC}(f)B_2(f) \tag{4.16}$$

Sa DSP est donnée par [Goulding and Bird, 1990] :

$$\varphi^b_{ANC}(f) = \varphi_{b_3}(f) - |\hat{H}_{ANC}(f)|^2 \varphi_{b_2}(f) \tag{4.17}$$

où $\varphi_{b_3}(f)$ et $\varphi_{b_2}(f)$ sont les DSPs de b_2 et b_3, respectivement.

On s'intéresse alors à l'atténuation du bruit par rapport à $B_3(f)$, définie par :

$$\Delta_b(f) = \frac{\varphi_{b_3}(f)}{\varphi^b_{ANC}(f)} \quad (4.18)$$

On trouve alors :

$$\begin{aligned}\Delta_b(f) &= \frac{\varphi_{b_3}(f)}{\varphi_{b_3}(f) - |\hat{H}_{ANC}(f)|^2\varphi_{b_2}(f)} \\ &= \frac{1}{1 - \dfrac{|\varphi_{b_2 b_3}(f)|^2}{\varphi_{b_2}(f)\varphi_{b_3}(f)}}\end{aligned} \quad (4.19)$$

Or, $\dfrac{|\varphi_{b_2 b_3}(f)|^2}{\varphi_{b_2}(f)\varphi_{b_3}(f)}$ est la Mean-Squared Coherence (MSC) entre les bruits $B_2(f)$ et $B_3(f)$, que l'on note $MSC_{23}(f)$. L'atténuation de bruit dépend donc uniquement de la MSC entre les bruits présents sur les capteurs [Goulding and Bird, 1990] :

$$\Delta_b(f) = \frac{1}{1 - MSC_{23}(f)} \quad (4.20)$$

Ainsi, dans ce cas idéal, l'atténuation de bruit est donnée, en fonction de la MSC entre les bruits captés, dans la Figure 4.4.

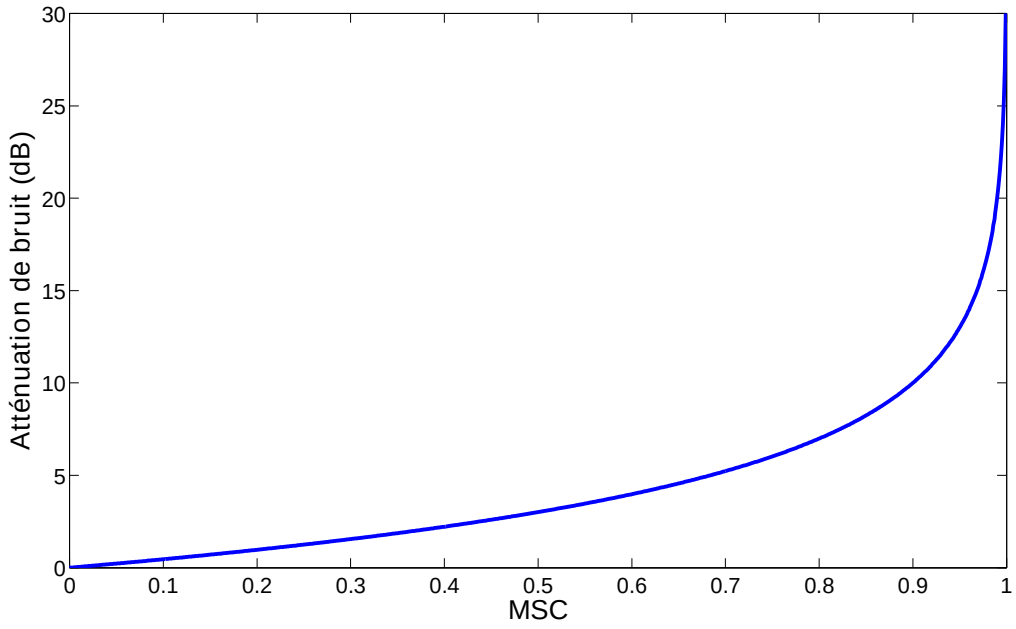


Figure 4.4 – Atténuation théorique de bruit pour un système d'ANC, en fonction de la MSC entre les bruits captés.

Ce système est donc très efficace en terme de réduction de bruit lorsque la MSC entre les bruits captés est grande. L'atténuation de bruit peut être théoriquement infinie (l'énergie du bruit en sortie est nulle) si cette MSC est égale à 1, mais cette expression ne prend pas en compte les erreurs d'estimation du filtre h_{ANC}, ni les limitations sur la longueur de ce filtre.

La MSC pour les bruits captés en automobile étant très forte en basses fréquences pour

certaines antennes (deux microphones omnidirectionnels ou deux microphones unidirectionnels pointant dans la même direction, voir Section 2.3, page 42), il est judicieux d'utiliser ce système dans les basses fréquences, avec ce type d'antenne de capteurs.

En supposant la stationnarité des bruits, cette expression pour l'atténuation de bruit reste valable pour les périodes où le signal utile est présent, puisque l'on n'estime h_{ANC} que sur les périodes de bruit seul, et que l'on fige l'adaptation du filtre sur les périodes où le signal utile est présent.

4.3.2 Atténuation de la parole

L'ANC a aussi un impact sur le signal utile, et peut notamment réduire son énergie en même temps que l'énergie du bruit. Or, pour maximiser le RSB en sortie de l'ANC, il faut réduire l'énergie du bruit tout en gardant un haut niveau de signal utile. Nous allons étudier ici l'impact de l'ANC sur l'énergie du signal utile.

Le signal utile en sortie de l'ANC s'écrit (durant les phases où le locuteur parle) :

$$X_{ANC}^{sig}(f) = H_s(f)S(f) - \hat{H}_{ANC}(f)^*S(f) = S(f)[H_s(f) - \hat{H}_{ANC}(f)] \tag{4.21}$$

Sa DSP est donnée par :

$$\varphi_{ANC}^{sig}(f) = \varphi_s(f)|H_s(f) - \hat{H}_{ANC}(f)|^2 \tag{4.22}$$

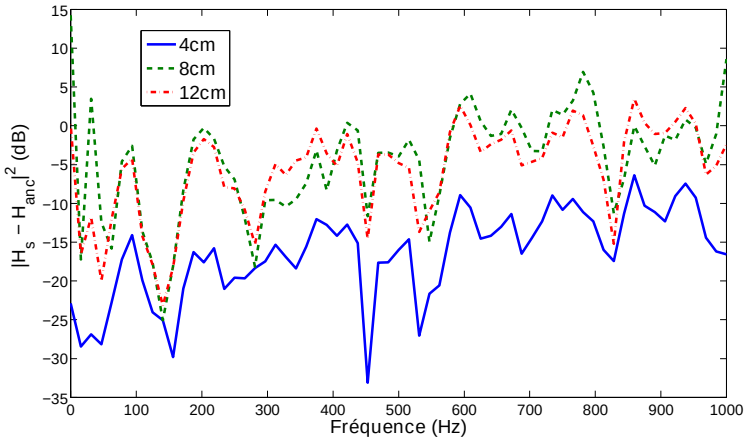
Ainsi, l'atténuation du signal utile est donnée par :

$$\Delta_{sig}(f) = \frac{\varphi_s(f)}{\varphi_{ANC}^{sig}(f)} = \frac{1}{|H_s(f) - \hat{H}_{ANC}(f)|^2} \square \tag{4.23}$$

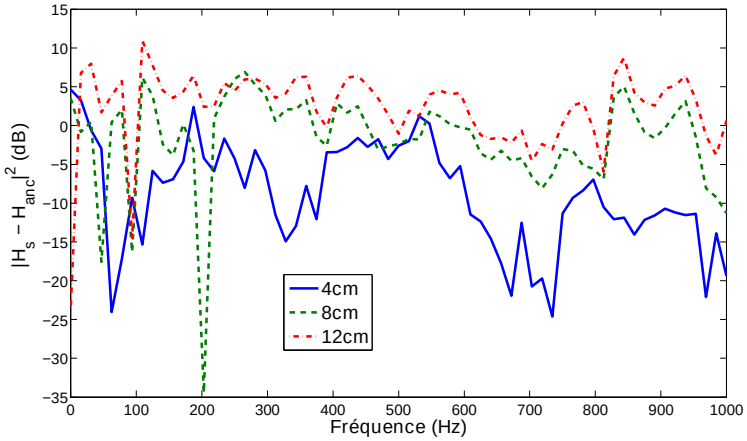
Celle-ci dépend de la différence entre la propagation relative du signal utile H_s et le filtre d'annulation de bruit estimé sur les périodes de bruit seul $\hat{H}_{ANC}$.

Nous avons mesuré cette quantité pour différentes configurations acoustiques. Pour cela, nous utilisons pour estimer la propagation relative de la parole le même système que dans la Section 2.4 (page 50), en limitant le signal à la bande [0, 1 kHz]. Pour estimer $\hat{H}_{ANC}$, nous utilisons le même système que celui présenté dans la Section 2.3 (page 42), pour le cas des bruits basses fréquences, très cohérents. On limite également les signaux de bruit à la bande [0, 1 kHz]. On considère les antennes qui vont permettre de maximiser la MSC entre les bruits enregistrés (pour permettre une bonne annulation de bruit) : 2 microphones omnidirectionnels ou 2 microphones unidirectionnels pointant dans la même direction (vers le conducteur). De plus, nous nous intéressons au placement Endfire, qui permet de mieux différencier $\hat{H}_{ANC}$ de H_s, comme vu dans la Section 2.3. On peut alors calculer $|H_s(f) - \hat{H}_{ANC}(f)|^2$, dans la bande de fréquence [0, 1 kHz]. Ces grandeurs sont représentées, pour les différentes antennes, dans la Figure 4.5.

On remarque que dans tous les cas, éloigner les microphones permet d'augmenter cette quantité, et donc de limiter l'impact de l'ANC sur le signal utile. En ce qui concerne la distance entre les capteurs, il faut donc un compromis entre une forte atténuation de bruit (qui suppose une forte MSC, et donc des capteurs proches) et une faible atténuation de parole, qui suppose des capteurs plus éloignés.



Antenne de capteurs omnidirectionnels



Antenne de capteurs unidirectionnels

Figure 4.5 – Différence entre propagation relative de la parole et filtre d'annulation de bruit, pour deux microphones placés en Endfire, pour plusieurs distances.

4.3.3 RSB en sortie

À partir des observations faites précédemment, on peut estimer le gain en RSB sur la sortie de l'ANC, par rapport aux entrées. Celui-ci s'écrit :

$$\Delta \mathrm{RSB}(f) = \frac{\varphi_{b_2}(f)}{\varphi_s(f)} \frac{\varphi^{sig}_{ANC}(f)}{\varphi^{b}_{ANC}(f)} = \frac{|H_s(f) - \hat{H}_{ANC}(f)|^2}{1 - \mathrm{MSC}_{23}(f)} \tag{4.24}$$

Ce gain en RSB est illustré dans la Figure 4.6.

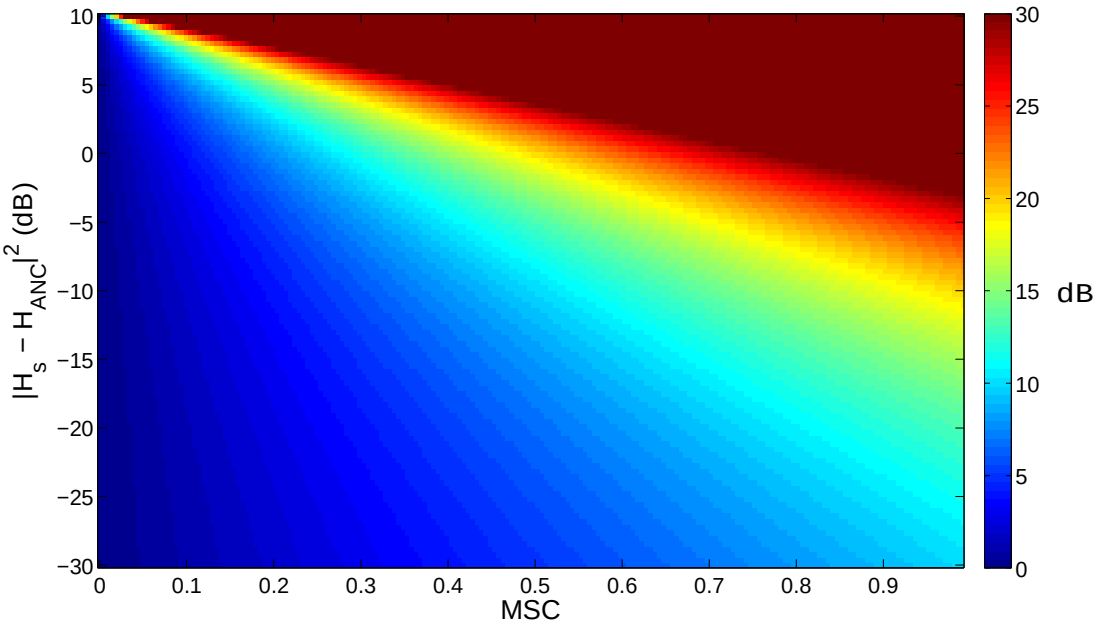


Figure 4.6 – Gain en RSB en fonction de la MSC entre les bruits captés et de $|H_s(f) - \hat{H}_{ANC}(f)|^2$.

Il faut donc une solution acoustique qui permette une grande MSC des bruits captés dans la bande de fréquences qui nous intéresse, et qui vérifie certaines propriétés sur la propagation relative de la parole pour permettre de maintenir $|H_s(f) - \hat{H}_{ANC}(f)|^2$ haut.

4.4 Implémentation et stratégie acoustique

Les performances théoriques de l'ANC discutées jusqu'ici considéraient un cas idéal. On ne tenait pas compte de la longueur limitée des filtres, ni des erreurs d'estimation faites par le filtrage adaptatif utilisé. Or ces approximations ont une grande influence sur les performances obtenues.

4.4.1 Influence des erreurs d'estimation

On propose d'illustrer l'importance de ces approximations dans le cas où la MSC des bruits captés vaut 1. Dans ce cas, on peut s'attendre, d'après la Figure 4.6, à obtenir un gain en RSB infini sur la sortie de l'ANC. Or ce n'est pas le cas en pratique, car en utilisant un algorithme de filtrage adaptatif de type Least-Mean Square (LMS), même après convergence, le filtrage adaptatif n'est pas exactement égal à la solution de Wiener. De plus, la longueur limitée du filtre utilisé perturbe également l'estimation : on n'obtient pas d'annulation parfaite de bruit. Enfin, le fait de stopper l'adaptation durant les périodes où le locuteur parle suppose de pouvoir détecter ces périodes de façon exacte, par un Détecteur d'activité vocale (Voice Activity Detector) (VAD) : selon les performances du VAD, il est possible que le filtre soit en partie estimé sur des phases où la parole est présente, ce qui induit un biais dans l'estimation [Lepauloux, 2010].

En effet, si l'on considère que le bruit sur un des capteurs est le filtré de celui présent sur

l'autre capteur (ce qui correspond à une MSC de 1), les bruits s'écrivent, en fréquence :

$$\begin{aligned} B_2(f) &= B(f) \\ B_3(f) &= H_b(f)B(f) \end{aligned} \tag{4.25}$$

On considère que le filtre ANC converge vers la solution de Wiener, à une erreur $\Delta h(f)$ près :

$$\hat{H}_{ANC}(f) = H_b(f) + \Delta h(f) \tag{4.26}$$

Le signal en sortie s'écrit :

$$X_{ANC}(f) = (H_s(f) - [H_b(f) + \Delta h(f)])\,S(f) - \Delta h(f)B(f) \tag{4.27}$$

Le gain en RSB par rapport aux entrées s'écrit :

$$\Delta_{RSB}(f) = \frac{|H_s(f) - [H_b(f) + \Delta H_{ANC}(f)]|^2}{|\Delta h(f)|^2} \leq \left(1 + \frac{|H_s(f) - H_b(f)|}{|\Delta h(f)|}\right)^2 \tag{4.28}$$

Ainsi, le gain en RSB maximum que l'on peut espérer est de $\left(1 + \frac{|H_s(f) - H_b(f)|}{|\Delta h(f)|}\right)^2$.

Nous avons reporté dans la Figure 4.7 la valeur maximum de $\Delta_{RSB}(f)$ en fonction de $|\Delta h(f)|^2$ et $|H_s(f) - H_b(f)|^2$.

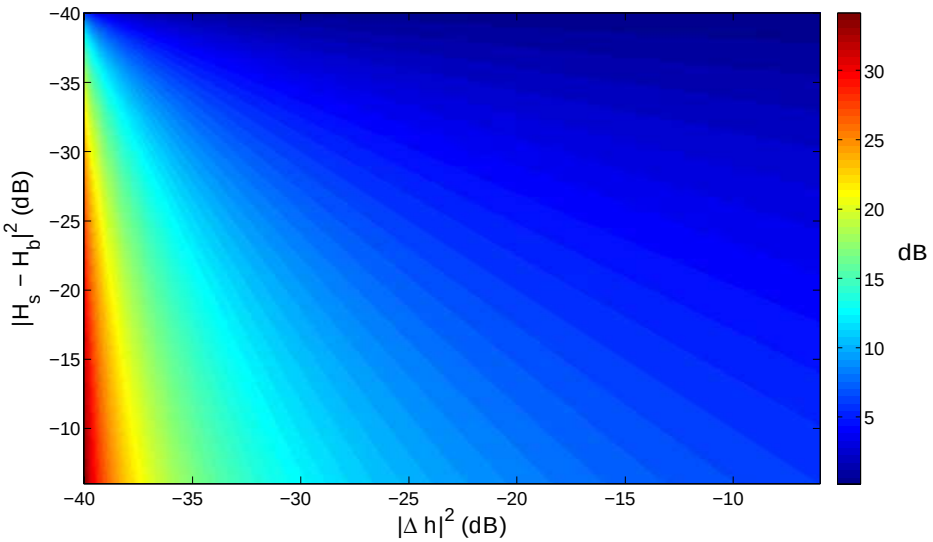


Figure 4.7 – Gain en RSB maximum en fonction de l'erreur d'estimation $|\Delta h|^2$ et de la différence de propagation bruit et parole $|H_s(f) - H_b(f)|^2$, dans le cas du bruit très cohérent.

Par rapport à ce que l'on a sur la Figure 4.6 sur l'axe MSC = 1, les performances atteignables sont nettement dégradées lorsque l'erreur faite sur l'estimation du filtre ANC $|\Delta h|^2$ grandit.

L'implémentation va donc jouer un rôle majeur dans les performances que l'on pourra atteindre grâce à cette méthode. C'est pourquoi nous proposons une analyse expérimentale des performances atteintes, qui prend en compte non seulement l'antenne de microphones utilisée,

mais aussi l'implémentation des filtrages adaptatifs dans les étages d'ANC et de compensation de distorsion.

4.4.2 Implémentation

Pour prendre en compte ces éléments, on se propose d'étudier les performances que l'on peut obtenir dans des cas proches de la réalité : on enregistre séparément bruit d'autoroute et parole dans un habitacle automobile, avec différentes antennes de capteurs. Le montage est le même que celui utilisé dans la Section 2.2 (page 36). On peut ainsi faire une ANC en connaissant les composantes de bruit et de parole sur la sortie, pour estimer des critères de performance, notamment le RSB, puisque c'est celui-ci qui détermine la performance du système global (comme on l'a vu dans la Section 4.2).

On utilise un algorithme Normalized LMS (NLMS) pour l'adaptation du filtre ANC.

A chaque instant t, on définit le vecteur $x_2(t) = [x_2(t - L + 1)\, x_2(t - L + 2)\, \square\square\square x_2(t)]^T$, où L est la longueur du filtre estimé, et l'on minimise la fonction de coût suivante :

$$J(h) = E\left[|x_3(t - \Delta) - h^H x_2(t)|^2\right] \tag{4.29}$$

Le retard Δ permet de prendre en compte les retards entre $x_2(t)$ et $x_3(t)$, dont on ne sait pas si ils seront positifs ou négatifs. On utilise ici $\Delta = \lfloor \frac{L}{2} \rfloor$, $\lfloor z \rfloor$ désignant l'entier inférieur à z le plus proche de z.

A chaque instant où la parole n'est pas présente, le filtre d'ANC h_{ANC} est mis à jour selon l'équation :

$$h_{ANC}(t+1) = h_{ANC}(t) + \frac{\mu_0}{x_2(t)^H x_2(t) + \varrho} \left[x_3(t - \Delta) - h_{ANC}(t)^H x_2(t) \right] x_2(t) \tag{4.30}$$

où μ_0 est le pas d'adaptation, dont la valeur sera discutée plus loin, et ϱ est une faible constante permettant de ne pas diverger lorsque $x_2(t)^H x_2(t) \sim 0$.

Lorsque la parole est présente, l'adaptation est bloquée et l'on a simplement :

$$h_{ANC}(t+1) = h_{ANC}(t) \tag{4.31}$$

On utilise alors des signaux comportant une période de bruit seul, pour permettre au filtre ANC de converger, suivie d'une période de parole bruitée, comme illustré dans la Figure 4.8.

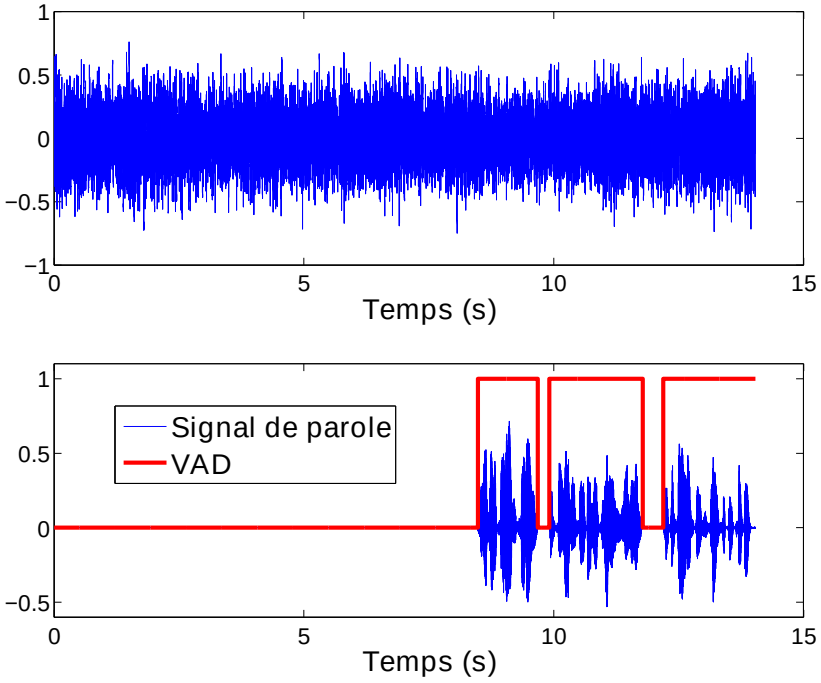


Figure 4.8 – Exemple de formes d'onde de signaux utilisés pour les expériences. En haut, le signal de bruit, en bas le signal de parole. VAD est la variable contrôlant l'adaptation : elle vaut 0 lorsque la parole est absente (période d'adaptation), et 1 sur les phases de parole (adaptation bloquée).

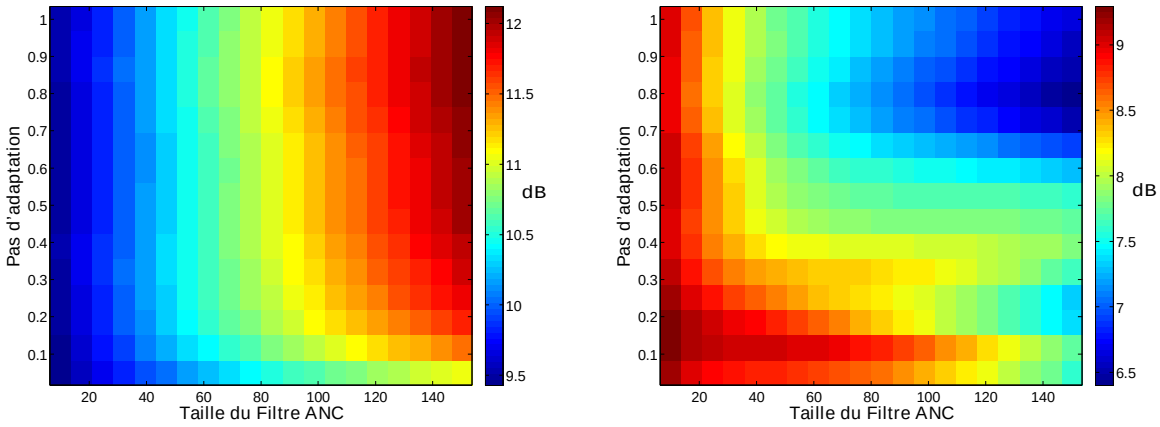
Ici, on dispose de signaux dont les périodes d'activité vocale sont parfaitement connues, de façon à mettre en évidence les performances atteintes en évitant d'éventuelles erreurs dues au VAD. Ce sont donc les performances atteignables lorsque ce détecteur est parfait.

On peut alors, ayant les signaux de bruit et de parole séparés, estimer :

— l'énergie du bruit sur la sortie de l'ANC lorsque l'on adapte le filtre (période de bruit seul),
— l'énergie du bruit sur la sortie de l'ANC sur les périodes où le filtre n'est pas adapté (période de parole bruitée),
— l'énergie du signal de parole sur la sortie de l'ANC sur les périodes où le filtre ne s'adapte pas (période de parole bruitée).

Dans tous les cas, nous présenterons ces quantités normalisées par les mêmes signaux de bruit et parole, captés par un microphone unidirectionnel, et on limite les signaux à la bande [0, 1 kHz], afin de garder une forte cohérence entre les bruit enregistrés par les différents capteurs.

On s'intéresse d'abord aux différences d'énergie de bruit résiduel entre les périodes de bruit seul et les périodes de parole bruitée, en fonction de la longueur du filtre utilisée L et du pas d'adaptation μ_0. Un exemple de ces quantités est présenté, dans le cas d'une antenne composée de deux microphones omnidirectionnels placés à 8 cm l'un de l'autre en Endfire, dans la Figure 4.9.



Atténuation sur les périodes de bruit seul — Atténuation sur les périodes de parole bruitée

Figure 4.9 – Atténuation de bruit obtenue par ANC par rapport à un capteur unidirectionnel, en fonction du pas d'adaptation et de la taille du filtre. A gauche, lorsque le filtre s'adapte (période de bruit seul) et à droite, lorsque le filtre ne s'adapte pas (période de parole bruitée).

On remarque une différence très forte entre les périodes de bruit seul et les périodes de parole bruitée. Lorsque le filtre s'adapte, l'atténuation de bruit devient plus forte lorsque la taille du filtre augmente. Lorsque le filtre ne s'adapte pas, l'atténuation est plus forte lorsque la taille du filtre est petite, et lorsque le pas d'adaptation est faible. Ceci s'explique par le fait qu'un pas d'adaptation plus petit permet au filtre estimé de moins varier autour de la solution optimale [Haykin, 2005], et une taille de filtre faible permet de prendre en compte moins de coefficients de filtres, mal estimés lorsque le filtre ne s'adapte pas.

La taille du filtre utilisé et le pas d'adaptation doivent donc être choisis soigneusement pour pouvoir obtenir une bonne atténuation de bruit dans les cas où le filtre ne s'adapte pas, afin de pouvoir avoir le meilleur RSB possible en sortie de l'ANC dans les périodes durant lesquelles le locuteur parle.

4.4.3 Antenne acoustique

En utilisant le même système que dans la Section 4.4.2, on peut estimer le RSB en sortie de l'ANC, sur les phases de parole. C'est cette valeur qui déterminera l'efficacité du système global, comme vu dans la Section 4.2.

Il faut donc prendre en compte non seulement l'atténuation du bruit, mais aussi l'atténuation de la parole.

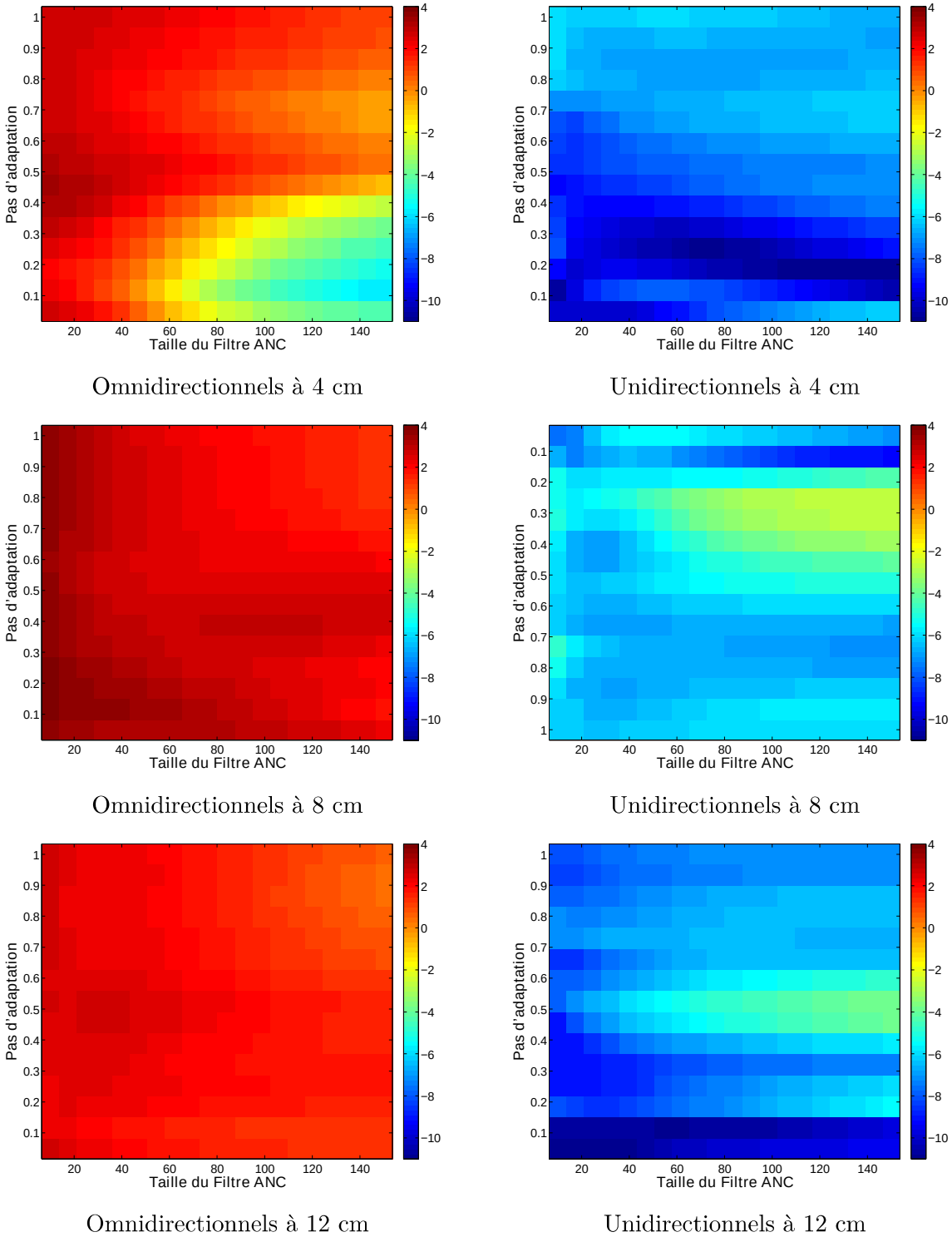
On s'intéresse ici au gain en RSB, par rapport au signal capté par un microphone unidirectionnel, et toujours dans la bande de fréquences [0, 1 kHz], et on s'intéresse particulièrement au comportement de ce gain en RSB en fonction de l'antenne de capteurs utilisée.

On se restreint aux antennes de deux capteurs omnidirectionnels et aux antennes de deux capteurs unidirectionnels, placés en Endfire, afin d'avoir une MSC proche de 1 sur les bruits captés, de manière à atténuer fortement le bruit en sortie, tout en limitant l'atténuation de la parole.

On a donc utilisé les mêmes simulations que celles présentées dans la Figure 4.8 avec ces antennes de microphones, et en considérant diverses valeurs pour la taille du filtre d'ANC, ainsi

que pour le pas d'adaptation du NLMS.

Les valeurs obtenues pour le gain en RSB sont présentées dans la Figure 4.10.



Omnidirectionnels à 4 cm | Unidirectionnels à 4 cm

Omnidirectionnels à 8 cm | Unidirectionnels à 8 cm

Omnidirectionnels à 12 cm | Unidirectionnels à 12 cm

Figure 4.10 – Gain en RSB en sortie de l'ANC, par rapport à un capteur unidirectionnel, en dB. Les capteurs sont placés en position Endfire.

Les différences entre les deux types d'antennes (capteurs omnidirectionnels ou unidirectionnels) sont flagrantes : en utilisant des capteurs unidirectionnels, il n'y a pas d'amélioration de RSB. On obtient au contraire une détérioration : l'atténuation de la parole n'est pas compensée

par l'atténuation de bruit. Ceci est dû aux chutes de MSC entre les bruits captés à certaines fréquences, que l'on avait remarquées dans la Figure 2.21 (page 46). De plus, dans la bande de fréquences [0, 1 kHz], un capteur unidirectionnel ne permet pas d'améliorer le RSB d'entrée par rapport à un capteur omnidirectionnel, comme on l'a constaté sur la Figure 2.8 (page 36) : l'intérêt d'utiliser ce type de capteur disparaît en basses fréquences.

Sur les antennes de capteurs omnidirectionnels, on constate que la plus grande amélioration a lieu lorsque l'on utilise un filtre de longueur faible. Sur l'antenne de capteurs omnidirectionnels placés à 4 cm l'un de l'autre, on remarque que dans certains cas, le RSB est dégradé. En mettant cette tendance en regard des courbes de la Figure 4.5, on en déduit que cela est dû à une faible différence entre la propagation relative de la parole et le filtre d'ANC.

L'antenne de deux capteurs omnidirectionnels placés à 8 cm l'un de l'autre semble être le meilleur compromis entre une forte MSC des bruits captés (microphones proches), et une différentiation de la propagation relative de la parole par rapport au filtre ANC (microphones éloignés). C'est cette antenne qui permet le meilleur gain en RSB, quels que soient les paramètres utilisés pour le NLMS (pas d'adaptation et taille de filtre).

On choisit d'utiliser cette antenne, avec une longueur de filtre courte (30 coefficients), et un faible pas d'adaptation (0.1).

4.5 Performances globales

On a donc pu définir une implémentation et une stratégie optimale pour l'ANC. Nous cherchons maintenant à définir une implémentation pour le second filtrage adaptatif, permettant de compenser la distorsion présente sur la sortie de l'ANC.

Celui-ci est fait par un algorithme NLMS, que l'on adapte lorsque la parole est présente. Dans ces périodes, la mise à jour de ce filtre est similaire à celle présentée dans l'équation (4.30). En utilisant les mêmes simulations que celles présentées dans la Figure 4.8, on cherche à estimer la distorsion relative en sortie de ce filtre, ainsi que le gain en RSB obtenu par cette méthode.

La distorsion relative est donnée par :

$$\mathrm{Disto} = \frac{\sum_{t\in P} \left[s(t) - \hat{s}^{sig}(t) \right]^2}{\sum_{t\in P} \left(s(t) \right)^2} \tag{4.32}$$

où P est l'ensemble des instants où la parole est présente, et $\hat{s}^{sig}(t)$ est la composante de parole en sortie du système global.

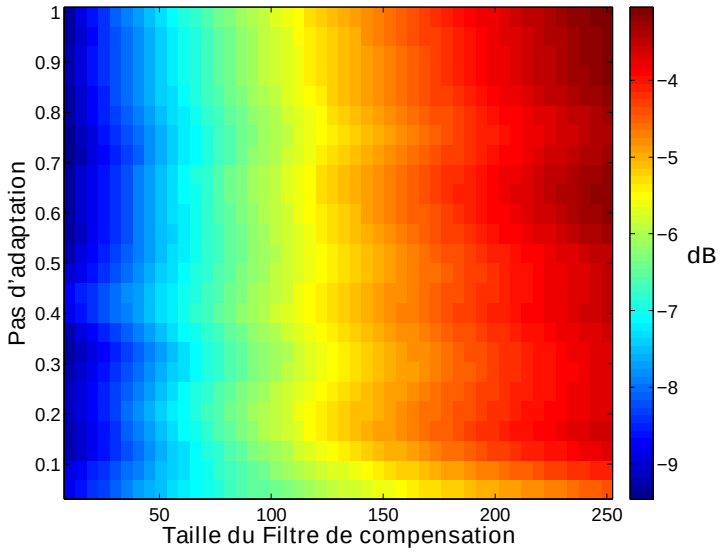
Le gain est RSB est donné par :

$$\Delta\mathrm{RSB} = \frac{\sum_{t\in P} \left[\hat{s}^{sig}(t) \right]^2}{\sum_{t\in P} \left(\hat{s}^{b}(t) \right)^2}, \quad \frac{\sum_{t\in P} \left(s(t) \right)^2}{\sum_{t\in P} \left(b_1(t) \right)^2} \tag{4.33}$$

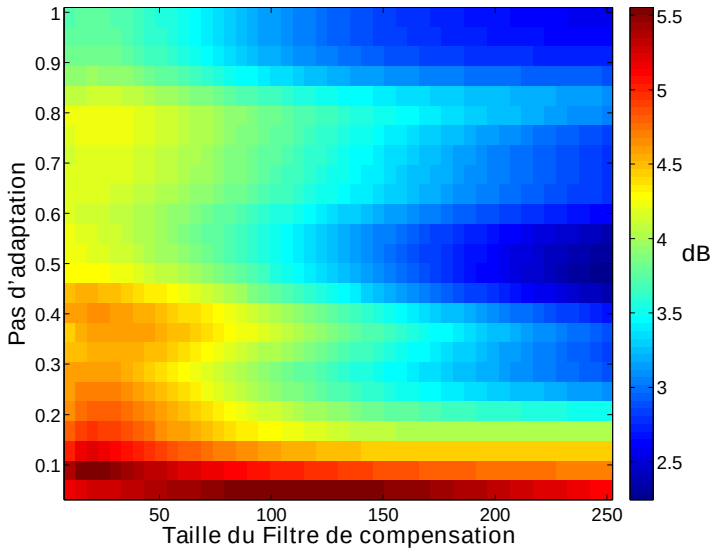
où $\hat{s}^{b}(t)$ est la composante de bruit en sortie du système global.

Le signal x_1 est enregistré par un capteur unidirectionnel placé à proximité de l'antenne constituée de 2 capteurs omnidirectionnels utilisés pour l'annulation de bruit. On a choisi un capteur unidirectionnel pour pouvoir ensuite combiner ce système (en basses fréquences) avec le système présenté dans le Chapitre 3 (en hautes fréquences), en ayant le même capteur de référence pour les deux systèmes. Les performances en fonction des paramètres du NLMS utilisé

(pas d'adaptation et longueur de filtre) sont présentées dans la Figure 4.11.



Distorsion relative en sortie



Gain en RSB en sortie

Figure 4.11 – Performances atteintes par le système global présenté Figure 4.1, page 88

On remarque que le gain en RSB est toujours positif : on améliore le RSB dans tous les cas. La distorsion relative est minimisée lorsque la taille du filtre est faible, alors que le RSB est maximisé lorsque la taille du filtre et le pas d'adaptation sont faibles.

On choisit donc d'utiliser un filtre à 50 coefficients), et un pas d'adaptation de 0.1 pour l'adaptation de ce filtre, qui permet une compensation de distorsion par rapport à la sortie de l'ANC.

4.6 Synthèse

Nous avons présenté dans ce Chapitre un système d'CR-ANC, permettant de réduire le bruit ambiant lorsque l'on a une grande cohérence inter-capteurs sur les bruits enregistrés. Cette condition sur la cohérence des bruits captés nous a permis de d'établir que c'est en basses fréquences que ce système sera efficace (dans la bande [0, 1 kHz]).

Cette méthode est basée sur l'utilisation de deux filtrage adaptatifs :

— un étage d'ANC, dont la sortie présente un bruit réduit, mais un signal de parole fortement distordu,

— un étage de compensation de distorsion, qui permet de réduire la distorsion sur la voix lorsque le RSB sur la sortie de l'ANC est très fort.

Toutefois, cette méthode apporte une grande distorsion sur la voix lorsqu'on n'est pas dans un cas idéal (en terme de cohérence des bruits, et de convergence de filtrage adaptatif). Nous avons donc présenté une étude expérimentale sur les performances obtenues en fonction de l'implémentation des filtrages adaptatifs (des NLMS), et de l'antenne de capteurs utilisée. Nous avons donc choisi des filtres courts (30 échantillons pour l'ANC et 50 échantillons pour la compensation de distorsion), qui s'adaptent lentement.

Pour la configuration acoustique, nous avons choisi deux microphones omnidirectionnels placés en Endfire, à 8 cm l'un de l'autre, pour l'étage d'ANC, et un capteur unidirectionnel placé à proximité pour la compensation de distorsion.

Chapitre 5

Speech Distortion Weighted Multichannel Wiener Filter (SDW-MWF)

Sommaire

Ce Chapitre a pour objet l'étude d'un dérivé du Filtre de Wiener multicanal (Multichannel Wiener Filter) (MWF). Ce système permet de pondérer le bruit en sortie du système et la distorsion apportée sur la voix, selon que l'on souhaite privilégier l'un ou l'autre de ces critères. Les performances de ce système sont ensuite étudiées, en fonction de l'environnement, afin de définir la bande de fréquences dans laquelle ce système fonctionnera le mieux, ainsi que l'antenne de capteurs la mieux adaptée.

On s'intéresse ici à un système dont l'expression est proche de l'Annulation de bruit adaptative (Adaptive Noise Cancellation) (ANC). On cherche à identifier le bruit présent sur un capteur de référence. En retranchant cette estimation au signal capté, on peut ainsi obtenir une estimation du signal de parole sur ce capteur.

On cherche alors à estimer ce bruit au sens de Wiener, en minimisant l'écart quadratique moyen entre le bruit présent sur le capteur de référence (numéroté 1) et le bruit estimé, qui est un filtrage des signaux enregistrés sur les entrées.

Cet écart quadratique est la somme de deux contributions :

— l'énergie du bruit résiduel,

— l'énergie de la distorsion apportée sur le signal utile.

On peut, en connaissant les statistiques du second ordre du signal de parole et du bruit, pondérer ces deux contributions, de façon à effectuer un compromis entre réduction de bruit et distorsion.

Nous allons commencer par présenter ce système dans son ensemble, ainsi que son implémentation adaptative. Nous allons ensuite nous intéresser aux performances que l'on peut en attendre, en fonction de l'environnement et de la fréquence de fonctionnement. Enfin, nous verrons comment le placement des capteurs influence ces performances, dans la bande de fréquences qui nous intéresse.

5.1 Principe et modèle

On présente ici l'estimation du bruit par filtrage de Wiener pondéré, ainsi que son implémentation adaptative.

Nous reprenons les notations du Chapitre 3 (page 61), pour chaque capteur noté m, entre 1 et M :

$$x_m(t) = s_m(t) + b_m(t) \quad (5.1)$$

où $s_m(t) = \{h_m \otimes s\}(t)$ est la parole présente sur le capteur m. On souhaite estimer l'une des versions retardées du bruit présente sur un capteur de référence, numéroté 1, pour le retrancher. Cela revient à estimer un filtre w de taille ML, de la façon suivante :

$$\hat{w} = \underset{g}{\operatorname{argmin}} E\left[|b_1(t-\Delta) - g^H x(t)|^2\right] \quad (5.2)$$

Avec les notations :

— $x_m(t)$ est le vecteur $[x_m(t-L+1) \quad \cdots \quad x_m(t)]^T$,

— $x(t) = [x_1(t)^T \quad x_2(t)^T \quad \cdots \quad x_M(t)^T]^T$.

Les vecteurs $s_m(t)$, $b_m(t)$, $s(t)$ et $b(t)$ sont définis de la même façon.

A noter que chaque x_m est de taille $L \times 1$, alors que $x(t)$ est de taille $ML \times 1$.

Le retard Δ permet de prendre en compte les retards entre le signal de référence et les autres entrées, dont on ne sait pas si ils seront positifs ou négatifs.

Le signal estimé s'écrit alors : $\hat{s}_1(t-\Delta) = x_1(t-\Delta) - \hat{w}^H x(t)$. Cela revient à prédire le bruit présent sur le capteur 1 et à le retrancher : la formulation est la même que pour faire de l'ANC. Ce principe est résumé dans le schéma présenté Figure 5.1.

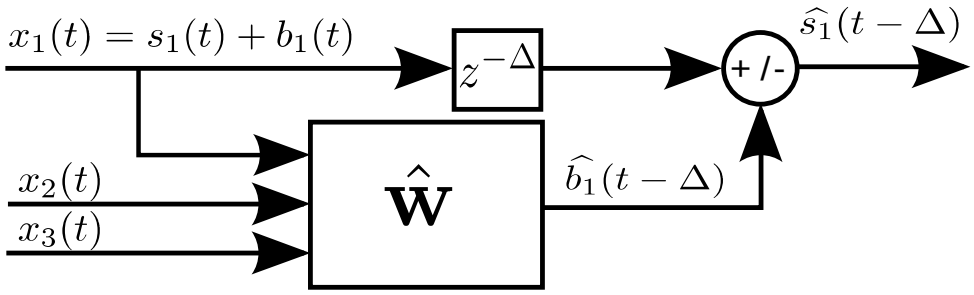


Figure 5.1 – Schéma de la prédiction de bruit par MWF

La solution est donnée par le filtre de Wiener :

$$\hat{w} = \left[E\left[x(t)x(t)^T \right] \right]^{-1} E\left[x(t)b_1(t-\Delta)\right] \quad (5.3)$$

5.1.1 Weighted Wiener Filter

Dans le cas du filtre de Wiener $\hat{w}$, l'Erreur Quadratique Moyenne (EQM) présentée dans l'équation (5.2) peut s'écrire comme la somme de deux termes (en invoquant la décorrélation entre bruit et signal utile) :

$$\hat{w} = \underset{g}{\operatorname{argmin}} \left[E\left[|b_1(t-\Delta) - g^H b(t)|^2 \right] \right] + \left[E\left[|g^H s(t)|^2 \right] \right]$$
$$\underbrace{\qquad\qquad}_{e_b} \qquad\qquad \underbrace{\qquad\qquad}_{e_s}$$

e_s est l'apport de distorsion sur le signal utile, et e_b est le bruit résiduel. On peut pondérer ces erreurs selon que l'on préfère avoir plus de distorsion ou plus de bruit résiduel. Le problème devient [Spriet et al., 2005] :

$$\hat{w}_\rho = \underset{g}{\operatorname{argmin}} \left[E\left[|b_1(t-\Delta) - g^H b(t)|^2 \right] \right] + \left[\rho E\left[|g^H s(t)|^2 \right] \right] \quad (5.4)$$

L'indice ρ indique que l'on régularise la fonction de coût pour pondérer la distorsion d'un facteur ρ. La solution est :

$$\hat{w}_\rho = (\rho R_s(t) + R_b(t))^{-1} E\left[b(t)b_1^*(t-\Delta)\right]$$

où $R_s(t) = E\left[s(t)s(t)^H \right]$ et $R_b(t) = E\left[b(t)b(t)^H \right]$.

L'interprétation est assez simple, en terme de performances : plus ρ est petit, plus l'on accorde de l'importance au bruit résiduel, au prix d'une distorsion plus forte sur le signal utile. Notamment, si $\rho = 0$, on a :

$$\hat{w}_\rho = R_b(t)^{-1} R_b(t)^{(:,L-\Delta)} = [0 \cdots \underbrace{1}_{(L-\Delta)\text{-ième position}} \cdots 0]^H \quad (5.5)$$

où $R_b(t)^{(:,L-\Delta)}$ est la $(L-\Delta)$ -ième colonne de la matrice $R_b(t)$.

Le signal de sortie vaut donc $x_1(t-\Delta) - x_1(t-\Delta)$ et est nul : le bruit a été entièrement supprimé, au prix d'une forte distorsion.

On utilise cet algorithme de manière adaptative, par un algorithme de descente de gradient [Spriet et al., 2005]. Un résumé de l'algorithme est présenté dans la Figure 5.2.

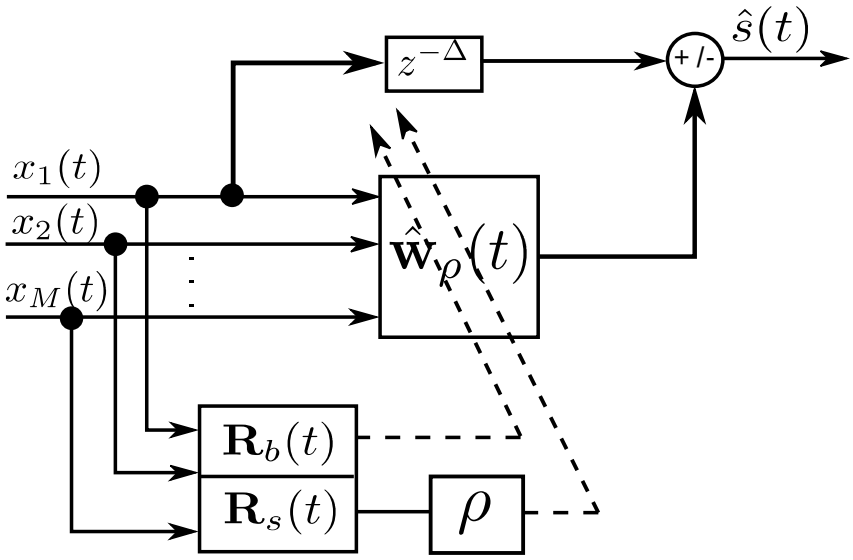


Figure 5.2 – Schéma de l'implémentation adaptative du SDW-MWF

5.1.2 Implémentation adaptative

On a donc besoin d'estimer les matrices $\mathbf{R}_s(t) = E\left[\mathbf{s}(t)\mathbf{s}(t)^H\right]$, $\mathbf{R}_b(t) = E\left[\mathbf{b}(t)\mathbf{b}(t)^H\right]$, le vecteur $E[\mathbf{b}(t)b_1^*(t-\Delta)]$ ainsi que le paramètre ρ (qui règle le compromis entre réduction de bruit et distorsion). Si l'on suppose que l'on dispose d'un Détecteur d'activité vocale (Voice Activity Detector) (VAD) et que le bruit $\mathbf{b}(t)$ est stationnaire, on peut estimer $\mathbf{R}_b(t)$ durant les phases de silence, où seul le bruit est présent sur les entrées. Sur ces phases de silence, on estime alors cette matrice récursivement :

$$\mathbf{R}_b(t) = \begin{cases} \lambda_b \mathbf{R}_b(t-1) + (1-\lambda_b)\mathbf{x}(t)\mathbf{x}(t)^H & \text{s'il n'y a pas de parole} \\ \mathbf{R}_b(t-1) & \text{sinon} \end{cases}$$

$\lambda_b \in]0\,1[$ étant un facteur d'oubli. On peut estimer $E[\mathbf{b}(t)b_1^*(t-\Delta)]$, en remarquant que c'est la $(L-\Delta)$ -ième colonne de la matrice $\mathbf{R}_b(t)$ (si $\Delta < L$). Pour estimer $\mathbf{R}_{\mathbf{s}}$, on invoque la décorrélation du bruit et du signal utile. Si l'on note $\mathbf{R}_x(t) = E\left[\mathbf{x}(t)\mathbf{x}(t)^H\right]$, on peut alors écrire :

$$\mathbf{R}_x(t) = \mathbf{R}_s(t) + \mathbf{R}_b(t) \tag{5.6}$$

On peut estimer $\mathbf{R}_x(t)$ de la même façon que $\mathbf{R}_b(t)$, sans condition sur la présence de parole :

$$\mathbf{R}_x(t) = \lambda_x \mathbf{R}_x(t-1) + (1-\lambda_x)\mathbf{x}(t)\mathbf{x}(t)^H \tag{5.7}$$

$\lambda_x \in]0\,1[$ étant également un facteur d'oubli. On choisit généralement $\lambda_x \leq \lambda_b$, car on considère le bruit de fond plus stationnaire que le signal de parole utile [Guérin et al., 2003].

On en déduit $\mathbf{R}_s(t) = \mathbf{R}_x(t) - \mathbf{R}_b(t)$. En ce qui concerne la longueur du filtre L, cela doit correspondre à une réalité spatiale et temporelle : il faut assez de coefficients pour prédire temporellement le bruit (cohérence temporelle du bruit) et spatialement (transfert spatial entre les microphones). Le coefficient ρ est réglé expérimentalement pour obtenir un compromis satisfaisant entre distorsion et bruit résiduel. La valeur 1 correspondant au filtrage de Wiener non pondéré apportant une distorsion trop grande (constatée subjectivement), on augmente ρ jus-

qu'à obtenir une qualité de voix satisfaisante. On utilise ces estimateurs pour faire une descente de gradient sur la fonction de coût suivante :

$$J_\rho(w) = \left[E\left[|b_1(t-\Delta) - w^H b(t)|^2 \right] + \rho \left[E\left[|w^H s(t)|^2 \right] \right] \right] \quad (5.8)$$

Son gradient vaut :

$$\delta J_\rho(w) = 2\left(\rho R_s(t) + R_b(t)\right) w - 2E\left[b(t) b_1^*(t-\Delta)\right] \quad (5.9)$$

L'équation de mise à jour est alors :

$$w(t) = w(t-1) - \mu \delta J_\rho(w) \quad (5.10)$$

où μ est un pas d'adaptation positif, proportionnel à $\frac{1}{x(t)^H x(t) + \varrho}$:

$$\mu = \frac{\mu_0}{x(t)^H x(t) + \varrho} \quad (5.11)$$

ϱ étant une constante positive faible permettant de pas diverger lorsque $x(t)^H x(t) \sim 0$.

5.1.3 Équivalence SDW-MWF - MVDR

On montre dans cette section que dans l'approche de type SDW-MWF dans le domaine fréquentiel, le filtrage purement spatial correspond à un beamforming Minimum Variance Distortionless Response (MVDR).

On s'intéresse ici au cas où l'on cherche d'abord à prédire le bruit pour le soustraire ensuite. On reprend les notations de l'Équation (3.6) :

$$X_n(f) = H(f) S_n(f) + B_n(f) \quad (5.12)$$

$X_n(f)$ est le vecteur des M entrées à la trame n et à la fréquence f, $H(f)$ est la propagation du signal de parole, $S_n(f)$ est le signal de parole et $B_n(f)$ est le vecteur des bruits captés.

On note alors $\Sigma_b(f)$ la matrice spectrale des bruits captés, et $\varphi_s(f)$ la Densité Spectrale de Puissance (DSP) du signal de parole.

On s'intéresse donc à la minimisation de $E\left[|B_1(f) - W(f)^H B(f)|^2 \right] + \rho E\left[|W(f)^H H(f) S(f)|^2 \right]$.

La solution s'écrit alors :

$$\hat{W}(f) = \left[\rho \varphi_s(f) H(f) H(f)^H + \Sigma_b(f) \right]^{-1} \Sigma_b(f)^{(:1)} \quad (5.13)$$

où $\Sigma_b(f)^{(:1)}$ désigne la première colonne de $\Sigma_b(f)$.

En utilisant le lemme d'inversion matricielle et le fait que $\Sigma_b(f)^{-1} \Sigma_b(f)^{(:1)} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$, l'expression devient :

$$\hat{W}(f) = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} - \underbrace{\frac{\rho\varphi_s(f)}{\rho\varphi_s(f) + \frac{1}{H(f)^H \Sigma_b(f)^{-1} H(f)}}}_{\text{Post-filtre}} \underbrace{\frac{\Sigma_b(f)^{-1} H(f)^H}{H(f)^H \Sigma_b(f)^{-1} H(f)}}_{\text{MVDR}} \tag{5.14}$$

On estime alors le signal débruité par $\hat{S}(f) = X_1(f) - \hat{W}(f)^H X(f)$.

Or, $X_1(f) = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}^H X(f)$

Donc, la sortie s'écrit :

$$\begin{aligned} \hat{S}(f) &= \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}^H - \hat{W}(f)^H X(f) \\ &= \underbrace{\frac{\rho\varphi_s(f)}{\rho\varphi_s(f) + \frac{1}{H(f)^H \Sigma_b(f)^{-1} H(f)}}}_{\text{Post-filtre}} \underbrace{\frac{H(f)\Sigma_b(f)^{-1}}{H(f)^H \Sigma_b(f)^{-1} H(f)}}_{\text{MVDR}} X(f) \end{aligned} \tag{5.15}$$

La partie spatiale du filtrage global correspond donc à un filtrage MVDR. On remarque notamment que lorsque ρ devient grand (on cherche à avoir le moins de distorsion possible, au prix d'un bruit résiduel plus fort), ce système est un beamforming MVDR, le post-filtre tendant vers 1.

5.2 Performances

Comme dans le Chapitre 3, on analyse les performances attendues de ce traitement en fonction de l'environnement, dans le domaine fréquentiel. On considère là aussi un système à $M = 2$ capteurs.

On reprend les notations de l'Équation (3.6) (page 63) :

$$\begin{aligned} X_1(f) &= H_1(f)S(f) + B_1(f) = S(f) + B_1(f) \\ X_2(f) &= H_2(f)S(f) + B_2(f) = H(f)S(f) + B_2(f) \end{aligned}$$

soit

$$X_n(f) = H(f)S_n(f) + B_n(f) \tag{5.16}$$

où $X_n(f)$ est le vecteur des 2 entrées à la trame n et à la fréquence f, $H(f)$ est la propagation relative du signal de parole, $S_n(f)$ est le signal de parole et $B_n(f)$ est le vecteur des bruits captés.

On note alors $\Sigma_b(f)$ la matrice spectrale des bruits captés, et l'on suppose :

$$\Sigma_b(f) = \varphi_b(f) \begin{bmatrix} 1 & c(f) \\ c(f)^* & 1 \end{bmatrix} \tag{5.17}$$

où $\varphi_b(f)$ est la DSP du bruit sur les capteurs 1 et 2 (que l'on suppose de même puissance spectrale $\varphi_{b_1}(f) = \varphi_{b_2}(f) = \varphi_b(f)$), et $c(f)$ est la cohérence entre le bruit sur le capteur 1 et celui sur le capteur 2 :

$$c(f) = \frac{\varphi_{b_1 b_2}(f)}{\varphi_b(f)} \tag{5.18}$$

On note $\varphi_s(f)$ la DSP du signal de parole, et $H(f)$ s'écrit :

$$H(f) = \begin{bmatrix} 1 \\ H(f) \end{bmatrix} \tag{5.19}$$

La fonction de coût pour le SDW-MWF [Spriet et al., 2005] s'écrit pour la fréquence f et pour un filtre G :

$$\begin{aligned} J_\rho^f(G) = & \varphi_{b_1}(f) - G^H \begin{bmatrix} \varphi_{b_1}(f) \\ \varphi_{b_1 b_2}(f)^* \end{bmatrix} - \begin{bmatrix} \varphi_{b_1}(f) & \varphi_{b_1 b_2}(f) \end{bmatrix} G \\ & + G^H \left[\rho \varphi_s(f) H(f) H(f)^H + \Sigma_b(f) \right] G \end{aligned} \tag{5.20}$$

où ρ est le paramètre réglant le compromis entre bruit résiduel et distorsion. En notant : $C_b(f) = \begin{bmatrix} \varphi_{b_1}(f) \\ \varphi_{b_1 b_2}(f)^* \end{bmatrix}$, on introduit la valeur du filtre optimal, minimisant cette fonction de coût :

$$W_{MWF}(f) = \left[\rho \varphi_s(f) H(f) H(f)^H + \Sigma_b(f) \right]^{-1} C_b(f) \tag{5.21}$$

En remplaçant $H(f)$, $\Sigma_b(f)$ et $C_b(f)$ par leur valeur, on obtient :

$$W_{MWF}(f) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \frac{\frac{\rho \varphi_s(f)}{\varphi_b(f)(1-|c(f)|^2)} \begin{bmatrix} 1 - c(f) H(f) \\ H(f) - c(f)^* \end{bmatrix}}{1 + \frac{\rho \varphi_s(f)}{\varphi_b(f)} 1 + \frac{|H(f) - c(f)^*|^2}{1-|c(f)|^2}} \tag{5.22}$$

A partir de cette expression, on cherche à déterminer quelles sont les conditions les plus favorables pour obtenir de bonnes performances à l'aide de ce traitement. Nous ré-introduisons la grandeur $R(f) = \frac{|H(f) - c(f)^*|^2}{1 - |c(f)|^2}$, dont nous avions décrit le rôle dans la Section 3.2.3 (page 74), afin de pouvoir estimer les performances en fonction de $R(f)$ et du RSB d'entrée $\frac{\varphi_s(f)}{\varphi_b(f)}$. Les performances vont être données par la distorsion, l'énergie du bruit en sortie et le gain en RSB par rapport aux entrées.

5.2.1 Distorsion

La distorsion en sortie du SDW-MWF s'écrit :

$$\mathrm{Disto}_{MWF}(f) = E\left[\left|W_{MWF}(f)^H H(f) S(f)\right|^2\right] = \varphi_s(f) E\left[\left|W_{MWF}(f)^H \frac{1}{H(f)}\right|^2\right] \tag{5.23}$$

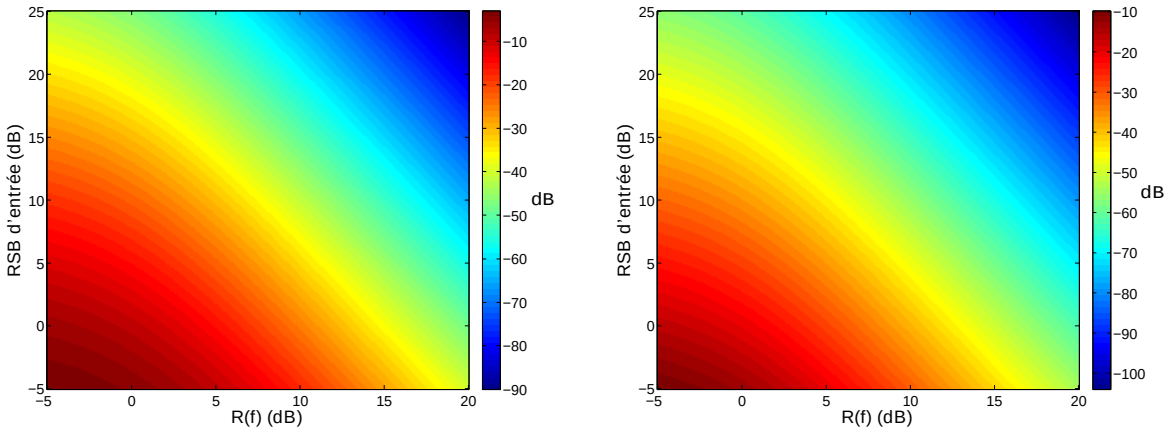
En remplaçant le filtre $W_{MWF}(f)$ par sa valeur, cela s'écrit :

$$\mathrm{Disto}_{MWF}(f) = \varphi_s(f) \left[\frac{1}{1+\rho \mathrm{RSB}(f)} \frac{1}{1+\frac{|H(f)-c(f)^*|^2}{1-|c(f)|^2}}\right]^2 \tag{5.24}$$

On peut alors exprimer la distorsion uniquement en fonction de $R(f) = \frac{|H(f) - c(f)^*|^2}{1 - |c(f)|^2}$ et du RSB d'entrée. La distorsion relative (normalisée par $\varphi_s(f)$) est donnée par :

$$\mathrm{Disto}_r(f) = \left[\frac{1}{1+\rho \mathrm{RSB}(f)(1+R(f))}\right]^2 \tag{5.25}$$

On trace cette distorsion normalisée en fonction de $R(f)$ et du RSB d'entrée, pour deux valeurs de ρ, dans la Figure 5.3.



Distorsion pour $\rho = 1$ Distorsion pour $\rho = 5$

Figure 5.3 – Distorsion en sortie en fonction de $R(f)$ et du RSB d'entrée, pour un facteur de pondération ρ de 1 et 5.

On constate qu'ntre les deux valeurs de ρ, on a surtout un changement d'échelle : augmenter ρ permet bien d'obtenir une plus faible distorsion quelle que soit la situation par ailleurs.

On remarque que la distorsion est une fonction décroissante du RSB d'entrée, mais aussi une fonction décroissante de $R(f)$. Pour minimiser la distorsion, on a donc intérêt à se placer dans une situation où $R(f)$ et le RSB d'entrée sont forts.

5.2.2 Bruit résiduel

Le calcul de la variance du bruit résiduel en sortie du filtrage s'écrit pour le SDW-MWF :

$$\mathrm{Bruit}_{MWF}(f) = E\left[|B_1(f) - W_{MWF}(f)^H B(f)|^2\right] \quad (5.26)$$

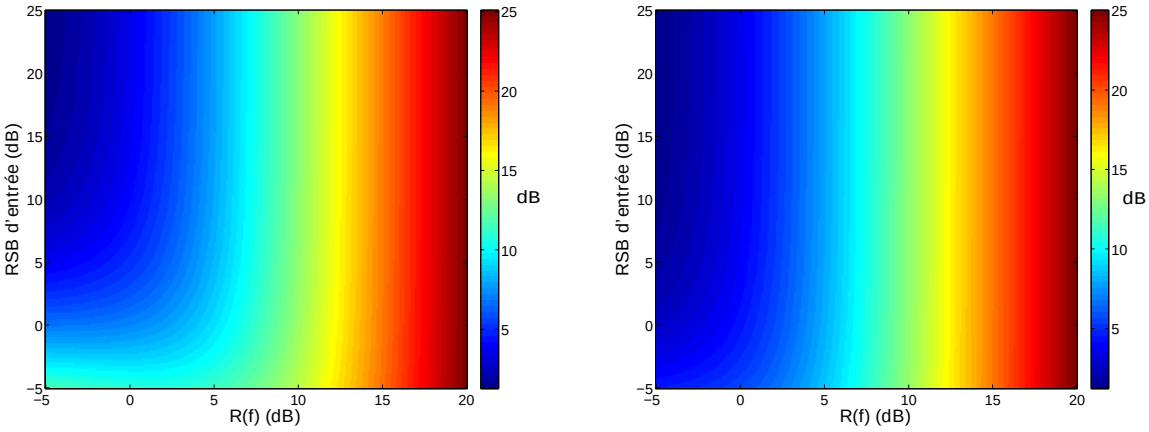
Le calcul donne :

$$\mathrm{Bruit}_{MWF}(f) = \varphi_b(f) \left(\underbrace{\frac{\rho \mathrm{RSB}(f)}{1+\rho \mathrm{RSB}(f)}}_{A} \underbrace{\frac{1}{1+\frac{|H(f)-c(f)^*|^2}{1-|c(f)|^2}}}_{} \right)^2 \left(\underbrace{1 + \frac{|H(f) - c(f)^*|^2}{1 - |c(f)|^2}}_{B} \right) \quad (5.27)$$

On peut là aussi exprimer cette grandeur uniquement en fonction de $R(f)$ et du RSB d'entrée. On considère l'atténuation du bruit par rapport aux entrées, donnée par :

$$\mathrm{Bruit}_{att}(f) = \frac{\varphi_b(f)}{\mathrm{Bruit}_{MWF}(f)} = \left[\frac{1+\rho \mathrm{RSB}(f)\,(1+R(f))}{\rho \mathrm{RSB}(f)}\right]^2 \frac{1}{(1+R(f))} \quad (5.28)$$

et on trace cette grandeur en fonction de $R(f)$ et du RSB d'entrée, pour deux valeurs de ρ dans la Figure 5.4.



Atténuation de bruit pour $\rho = 1$ Atténuation de bruit pour $\rho = 5$

Figure 5.4 – Atténuation de bruit en sortie en fonction de $R(f)$ et du RSB d'entrée, pour un facteur de pondération ρ de 1 et 5.

On remarque ici que augmenter ρ diminue bien l'atténuation du bruit : ce paramètre règle bien le compromis distorsion/atténuation du bruit.

On remarque que l'atténuation du bruit dépend surtout de $R(f)$. Pour maximiser cette atténuation, il faut donc se placer dans une situation où $R(f)$ est grand.

5.2.3 RSB en sortie

Pour le SDW-MWF, le RSB de sortie s'écrit [Lawin-Ore and Doclo, 2012] :

$$\mathrm{RSB}_{MWF}(f) = \frac{E\left[|S(f) - W_{MWF}(f)^H H(f) S(f)|^2\right]}{E\left[|B_1(f) - W_{MWF}(f)^H B(f)|^2\right]} \quad (5.29)$$

ou encore :

$$\mathrm{RSB}_{\mathrm{MWF}}(f) = \mathrm{RSB}(f) \left(1 + \frac{|H(f) - c(f)^*|^2}{1 - |c(f)|^2}\right) \quad (5.30)$$

On constate que le RSB en sortie est toujours supérieur au RSB en entrée, comme cela avait déjà été démontré dans [Doclo and Moonen, 2005]. De plus, ce RSB en sortie ne dépend pas du paramètre ρ, qui règle simplement le compromis distorsion/atténuation du bruit, mais n'affecte pas le RSB.

Le gain en RSB s'exprime en fonction de $R(f)$ uniquement, et vaut :

$$\Delta\mathrm{RSB}(f) = \frac{\mathrm{RSB}_{\mathrm{MWF}}(f)}{\mathrm{RSB}(f)} = 1 + R(f) \quad (5.31)$$

Et on trace sa valeur dans la Figure 5.5.

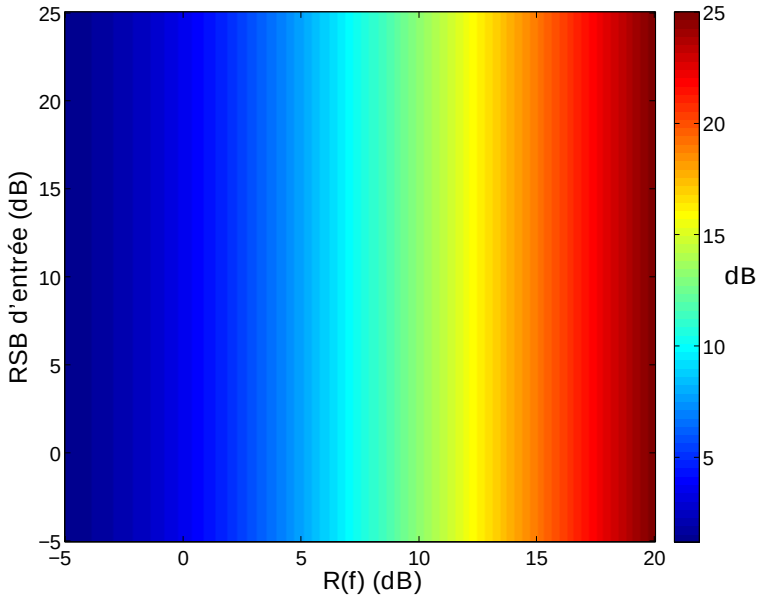


Figure 5.5 – Gain en RSB en fonction de $R(f)$ et du RSB d'entrée.

Il faut donc là aussi avoir $R(f)$ grand pour pouvoir maximiser le gain en RSB.

5.2.4 Conclusion

On a donc vu que pour optimiser tous ces critères de performance, il faut se placer dans une situation où $R(f)$ est grand. Or, d'après ce que l'on a vu dans la Section 3.2.3, c'est en basses fréquences que l'on peut obtenir $R(f)$ le plus grand, en profitant de la forte cohérence entre les bruits reçus par les différents capteurs.

On choisit donc d'utiliser ce système dans la bande de fréquences [0, 1 kHz], où notamment la solution présentée dans le Chapitre 3 n'est pas efficace.

Nous allons à présent nous intéresser à l'optimisation du placement des capteurs, afin d'obtenir les meilleures performances dans cette bande de fréquences.

5.3 Considérations sur le placement des capteurs

On simule ici la même situation que dans la Section 3.4.1 (page 80), en utilisant deux capteurs.

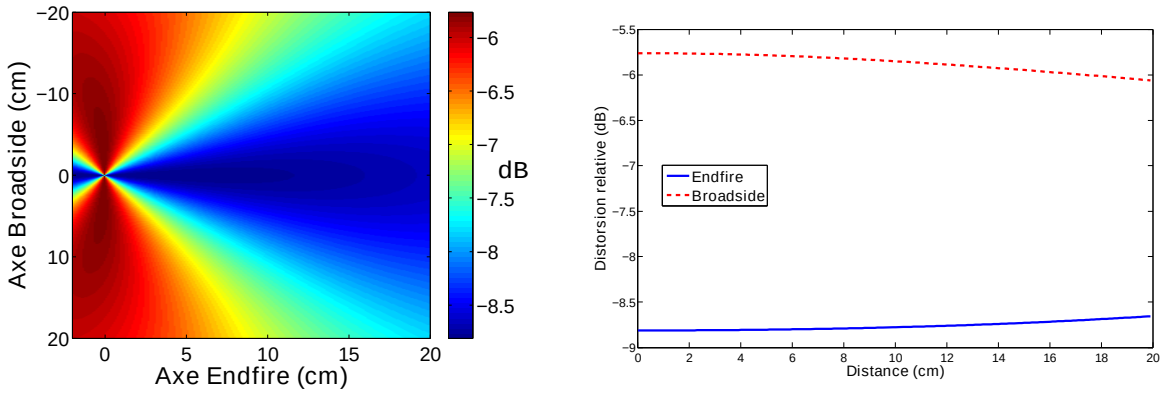
Toutefois, on ne s'intéresse ici qu'au cas où les microphones sont omnidirectionnels. En effet, on va intégrer les critères de performances présentés dans la Section 5.2 sur la bande de fréquence [0, 1 kHz]. Or, dans ces fréquences, les capteurs unidirectionnels perdent leur directivité, et donc leur intérêt, comme on l'a vu dans la Section 2.1 (page 30). On s'intéresse donc uniquement aux antennes composées de deux microphones omnidirectionnels.

La DSP du bruit simulé est donc celle présentée pour des capteurs omnidirectionnels dans la Figure 3.11 (page 81), et les simulations sont faites en choisissant un facteur de pondération $\rho = 5$.

Pour chaque critère, on présente les performances que l'on peut attendre dans le plan (le second microphone est placé dans la zone $x \in [-20\text{ cm} \quad 20\text{ cm}]$, $y \in [-2\text{ cm} \quad 20\text{ cm}]$), et les coupes sur les 2 axes $x = 0$ et $y = 0$ qui correspondent aux situations Endfire et Broadside, respectivement.

5.3.1 Distorsion

On présente ici la distorsion normalisée par l'énergie du signal utile, et son comportement est illustré sur la Figure 5.6. On souhaite ici rendre ce critère le plus petit possible, car une distorsion de $-\infty$ dB signifie que le signal utile reste intact.



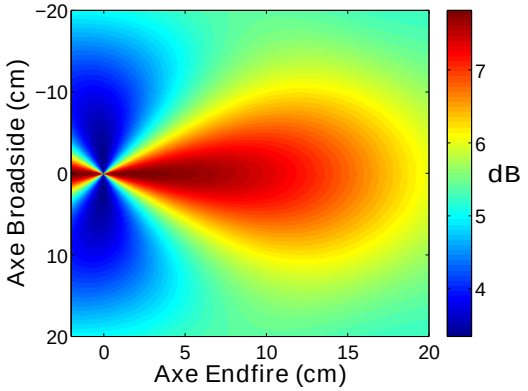
Distorsion relative dans le plan — Coupes sur les 2 axes : Endfire et Broadside

Figure 5.6 – Distorsion relative en sortie pour le SDW-MWF, avec $\rho = 5$, pour des capteurs omnidirectionnels.

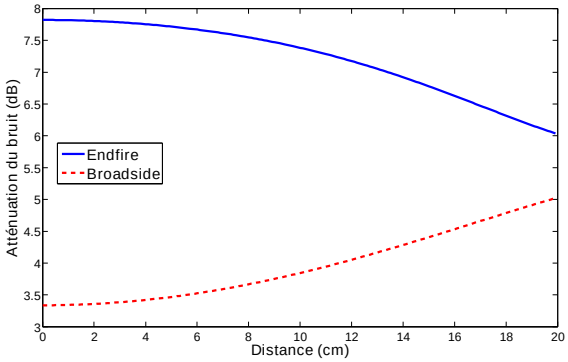
On remarque qu'ici, le placement Endfire est préférable pour minimiser la distorsion en sortie. De plus, ce critère varie peu avec la distance entre les capteurs utilisés.

5.3.2 Bruit résiduel

On présente ici l'atténuation du bruit obtenue, par rapport à l'énergie du bruit sur les entrées. On souhaite maximiser ce critère, pour avoir le moins de bruit résiduel possible. Ce critère est présenté dans la Figure 5.7.



Atténuation du bruit dans le plan



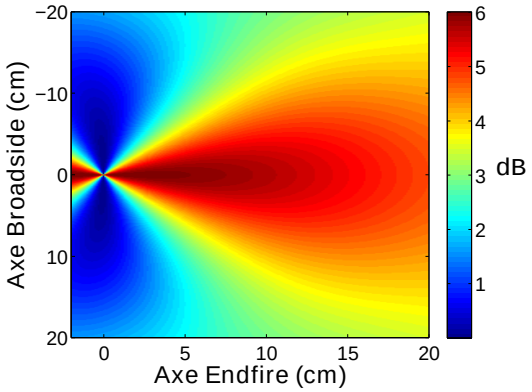
Coupes sur les 2 axes : Endfire et Broadside

Figure 5.7 – Atténuation de bruit pour le SDW-MWF, avec $\rho = 5$, pour des capteurs omnidirectionnels.

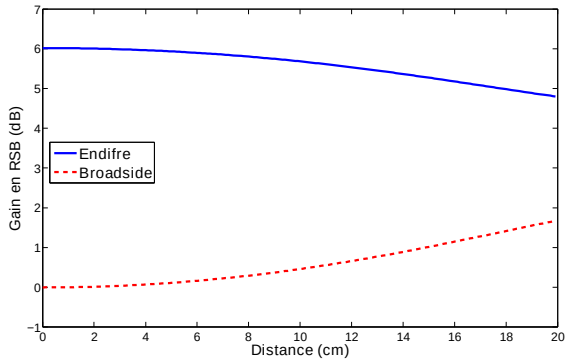
Là aussi, un placement Endfire est préférable, et utiliser des microphones proches l'un de l'autre permet d'obtenir une meilleure réduction de bruit.

5.3.3 RSB en sortie

On présente ici le gain en RSB par rapport au RSB en entrée. C'est un critère que l'on cherche à maximiser, pour avoir un signal de sortie dont l'énergie du bruit est réduite par rapport à l'énergie de la parole. Son comportement en fonction du placement des capteurs (l'un par rapport à l'autre) est présenté dans la Figure 5.8.



Gain en RSB dans le plan



Coupes sur les 2 axes : Endfire et Broadside

Figure 5.8 – Gain en RSB pour le SDW-MWF, avec $\rho = 5$, pour des capteurs omnidirectionnels.

Ici aussi, un placement Endfire avec des microphones proches l'un de l'autre est la meilleure solution pour améliorer le RSB du signal de sortie.

5.3.4 Conclusion

Ici, il n'y a pas d'ambiguïté : pour tous les critères, c'est un placement Endfire qui permet d'obtenir les meilleures performances avec ce système. On choisit donc pour ce traitement basses fréquences une antenne de capteurs omnidirectionnels placée en Endfire, avec des capteurs proches l'un de l'autre (4 cm).

5.4 Synthèse

Nous avons présenté dans ce Chapitre le SDW-MWF, qui permet d'estimer le bruit sur un capteur de référence afin de le soustraire, ce qui donne le signal débruité de sortie.

Une étude des performances atteignables a également permis de définir que ce système fonctionne mieux en basses fréquences (dans la bande [0, 1 kHz]) qu'en hautes fréquences. Toutefois, cette étude n'a pas démontré que ce système était inefficace au dessus de 1 kHz.

Nous avons également étudié les performances obtenues en fonction du placement des capteurs. Nous avons ainsi défini que la meilleure stratégie acoustique est d'utiliser une antenne de capteurs omnidirectionnels placés en Endfire, proches l'un de l'autre. En effet, c'est ce placement qui permet d'obtenir en même temps une faible distorsion, une meilleure atténuation du bruit et la meilleure amélioration du RSB.

Chapitre 6

Systèmes hybrides et résultats subjectifs

Sommaire

Nous allons présenter ici les systèmes hybrides que l'on souhaite utiliser. Ceux-ci sont des combinaisons des méthodes présentées dans les Chapitre 3, 4 et 5. En prenant en compte les performances obtenues pour chacune de ces méthodes en fonction de l'environnement, et les mesures acoustiques faites dans le Chapitre 2, nous combinons ces méthodes de façon à ce que chacune soit utilisée dans la bande de fréquences qui permet d'optimiser son fonctionnement.

Nous présentons ensuite le test d'écoute que nous avons mené pour évaluer les systèmes hybrides ainsi conçus, ainsi que les résultats obtenus.

Nous avons présenté plusieurs systèmes de débruitage multi-capteurs, qui ont des performances différentes selon l'environnement. Ainsi, en mettant en regard les performances de méthodes présentées dans les Chapitres 3, 4 et 5 et les caractéristiques de l'environnement automobile présentées dans le Chapitre 2, on peut voir que les méthodes présentées ont un domaine de fonctionnement différent :

MVDR adaptatif - Chapitre 3 Fonctionne en hautes fréquences, où les bruits sont faiblement cohérents d'un capteur sur l'autre, et où le Rapport Signal-à-Bruit (RSB) d'entrée est plus fort.

CR-ANC - Chapitre 4 Fonctionne uniquement en basses fréquences, où la cohérence entre les bruits captés est forte. Ceci permet une plus forte atténuation de bruit.

SDW-MWF - Chapitre 5 Fonctionne mieux en basses fréquences, où les bruits captés sont plus cohérents d'un microphone sur l'autre.

Ces différences de comportement nous amènent à proposer des systèmes hybrides en sous-bande pour exploiter au mieux les méthodes envisagées [Fox et al., 2012, Fox et al., 2013].

Nous avons donc étudié deux systèmes hybrides : l'un exploitant un MVDR adaptatif en hautes fréquences, et une annulation de bruit adaptative en basses fréquences, et l'autre exploitant toujours un MVDR adaptatif en hautes fréquences, et un SDW-MWF en basses fréquences. Dans tous les cas, la séparation des bandes se fait à 1 kHz.

De plus, l'algorithme Parrot pour le **Wideband** sépare le signal en deux bandes : une de 0 à 4 kHz, et l'autre de 4 kHz à 8 kHz. D'après ce que l'on a constaté, il faudrait appliquer un MVDR en utilisant deux microphones unidirectionnels, placés en **Broadside** proches (2 cm d'écartement), dans la bande [4 kHz, 8 kHz], comme on l'a vu sur la Figure 3.8 (page 76). Or, des tests d'écoute informels ont suggéré que cela n'apportait pas de meilleure qualité par rapport à un traitement mono-capteur. C'est pourquoi nous présenterons ici les stratégies hybrides uniquement dans la bande de fréquences [0, 4 kHz].

Les systèmes **Wideband** considérés sont donc illustrés dans la Figure 6.1.

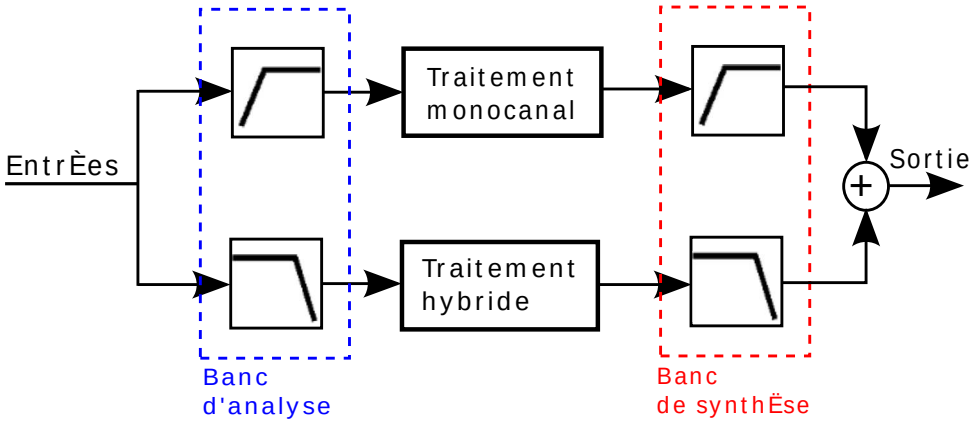


Figure 6.1 – Architecture pour la téléphonie **Wideband**. Les signaux sont séparés en deux bandes : de 0 à 4 kHz et de 4 kHz à 8 kHz.

Pour chaque sous-bande, les traitements se font sur des trames de 256 échantillons (soit 32ms), recouvrantes à 50%.

Ces méthodes ont été évaluées de manière subjective, en faisant passer un test d'écoute à un panel d'individus, afin de mieux évaluer l'impact de ces systèmes sur un utilisateur final. Pour

cette évaluation, nous avons fabriqué une antenne comportant tous les capteurs nécessaires aux deux approches hybrides considérées. Cette antenne est présentée dans la Figure 6.2.

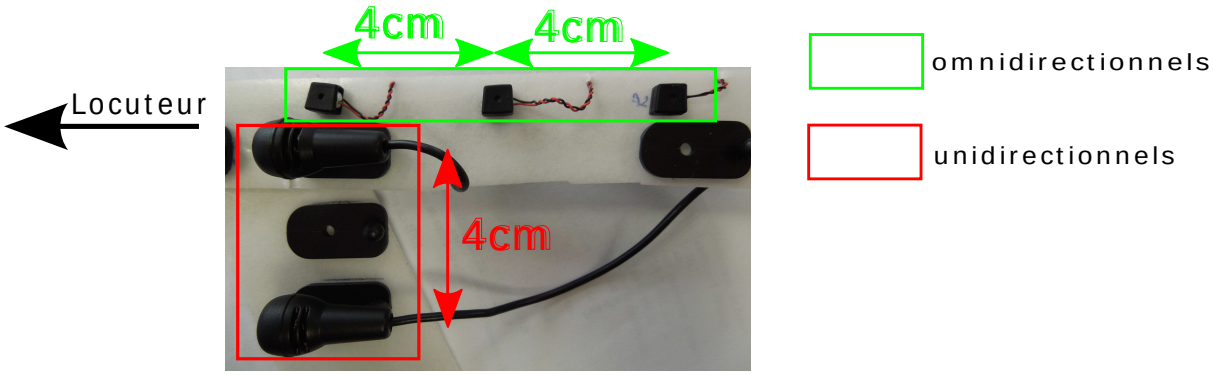


Figure 6.2 – Antenne conçue pour l'enregistrement de la base de test. Le locuteur se situe sur la gauche.

L'antenne est faite de façon à ce que les microphones unidirectionnels soient placés en Broadside, et les microphones omnidirectionnels en Endfire.

Nous allons commencer par présenter ces méthodes hybrides, puis nous détaillerons la méthodologie des tests d'écoute utilisés, et les résultats obtenus.

6.1 Système MVDR + CR-ANC

6.1.1 Principe général

Conformément à la stratégie que l'on a présentée dans la Section 1.2.5 (page 18), on construit un étage multi-capteurs, dont la sortie sera traitée par l'algorithme de réduction de bruit mono-capteur déjà utilisé par Parrot.

Cet étage multi-capteurs est donc hybride, il est construit comme suit :

un banc de filtre qui permet de séparer la partie basses fréquences, en dessous de 1 kHz.

un traitement différencié Annulation de bruit adaptative résistante aux fuites de Parole (Crosstalk Resistant Adaptive Noise Cancellation) (CR-ANC) en basses fréquences, et MVDR en hautes fréquences.

une reconstruction fréquentielle Les sorties de chaque traitement sont alors fréquentielles, et l'on reconstruit le spectre, qui sert d'entrée à l'étage mono-capteur (basé sur un gain Optimally Modified Log-Spectral Amplitude (OM-LSA), présenté dans la Section 1.3.4, page 23).

Un schéma global de ce système est présenté dans la Figure 6.3.

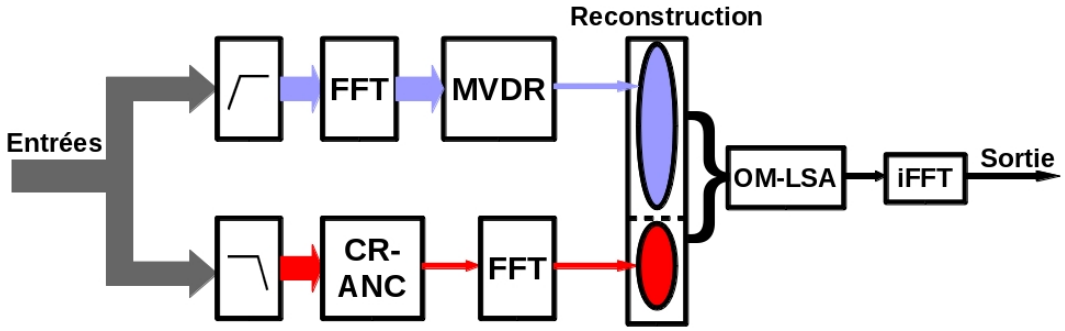


Figure 6.3 – Schéma global du système hybride MVDR + CR-ANC. Les flèches épaisses représentent des signaux multicanaux.

6.1.2 Justification de l'approche

On a présenté les approches MVDR adaptatif et CR-ANC dans les Chapitres 3 et 4, respectivement.

On a notamment vu dans la Section 3.2.3 (page 70) que l'approche MVDR adaptatif fonctionne surtout en hautes fréquences. En effet, en hautes fréquences, le RSB d'entrée est plus favorable, et les bruits captés sont moins cohérents d'un microphone sur l'autre. Ceci permet à l'estimation adaptative de la propagation relative de la parole présentée Section 3.2.2 (page 68) d'être efficace, alors que ses performances diminuent en basses fréquences. En basses fréquences, une telle approche induit notamment plus de distorsion, comme on l'a vu Section 3.2.3 (page 70).

En ce qui concerne l'approche par CR-ANC, la Figure 4.6 (page 96) montre que le gain de performance est d'autant plus important que la Mean-Squared Coherence (MSC) entre les bruits captés est proche de 1. En effet, lorsque la MSC est faible, l'atténuation de la parole par l'étage d'Annulation de bruit adaptative (Adaptive Noise Cancellation) (ANC) présentée Section 4.3.2 (page 94) n'est plus compensée par l'atténuation du bruit. Ainsi, le gain en RSB sur la sortie de l'annulation de bruit n'est plus suffisant pour permettre au système global d'atteindre de bonnes performances. Or, d'après la Section 2.3.2 (page 46), c'est en basses fréquences, et en utilisant une antenne appropriée (composée de deux omnidirectionnels, par exemple), que l'on maximise cette MSC. Ce système apportera donc de bonnes performances uniquement en basses fréquences.

C'est pourquoi il est judicieux de considérer un système tel que celui présenté Figure 6.3.

6.1.3 Implémentation

Antenne de capteurs

Pour choisir l'antenne de capteurs associée à chaque traitement, il est important de choisir :

— une antenne permettant d'obtenir les meilleures performances possibles pour chaque système,

— un capteur de référence commun pour les deux systèmes

Ainsi, conformément aux observations faites dans la Section 3.4.5 (page 84), on choisit pour le MVDR adaptatif sur la bande de fréquences [1 kHz 4 kHz] une antenne de deux capteurs cardioïdes placés en Broadside à 4 cm l'un de l'autre.

Pour la partie CR-ANC sur la bande [0, 1 kHz], on choisit :

— deux capteurs omnidirectionnels placés en Endfire à 8 cm l'un de l'autre, conformément aux mesures faites dans la Figure 4.10, en entrée de l'étage d'ANC,
— un capteur unidirectionnel, commun avec l'un de ceux utilisés pour le MVDR, comme capteur de référence pour la compensation de distorsion.

Cette stratégie acoustique est représentée sur le schéma de la Figure 6.4.

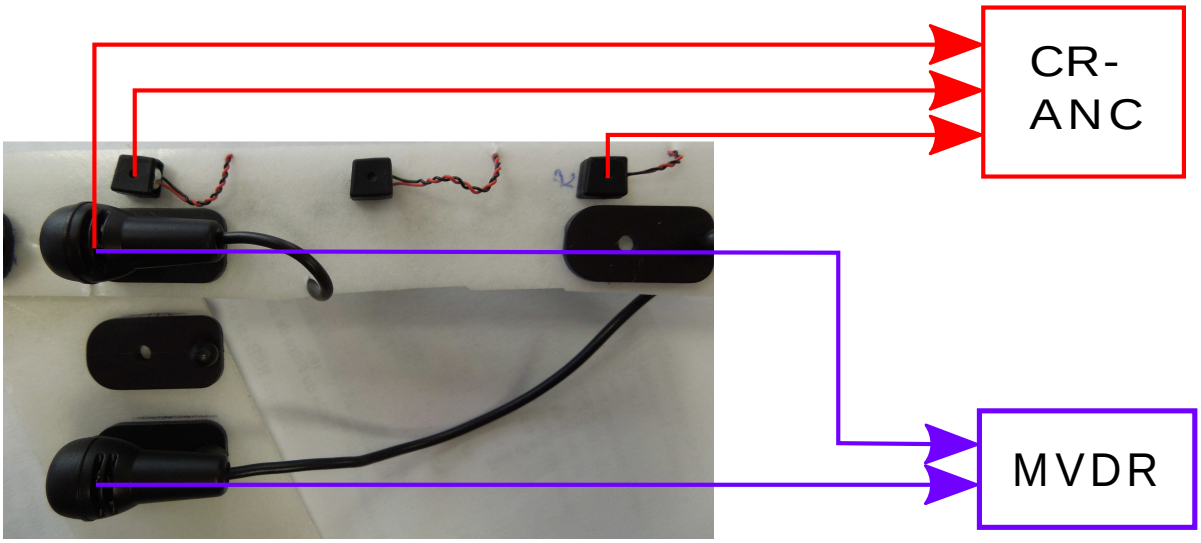


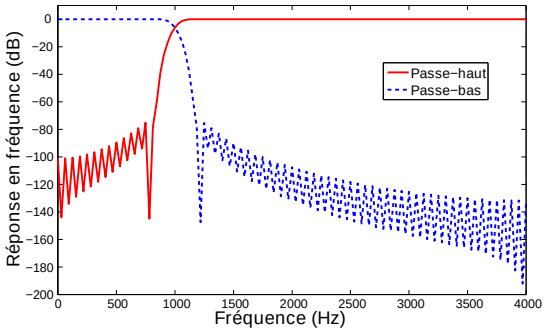
Figure 6.4 – Photo de l'antenne utilisée pour le système hybride CR-ANC + MVDR, avec le schéma des capteurs utilisés pour chaque méthode.

Banc de filtres

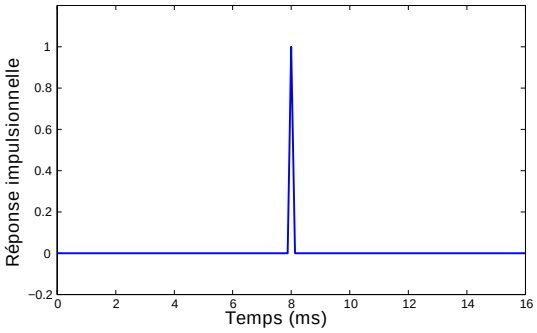
Pour séparer la bande basse de la bande haute, nous avons utilisé un banc de filtre asymétrique sans sous-échantillonnage. Pour cela, nous avons créé un filtre passe-bas $h_{bf}(t)$ et un filtre passe-haut $h_{hf}(t)$ à partir d'une fenêtre de Blackman, les deux de longueur $L = 129$, en contraignant la somme de leur réponse impulsionnelle de la façon suivante :

$$h_{bf}(t) + h_{hf}(t) = \delta\left(t - \frac{L+1}{2}\right) \tag{6.1}$$

Cette contrainte permet de se passer de filtre de synthèse, la reconstruction globale étant juste la somme des deux signaux en sortie des filtres. On obtient alors le signal d'origine, retardé de $\frac{L+1}{2}$ échantillons. La réponse en fréquence des filtres considérés et la somme de leurs réponses impulsionnelles sont présentées dans la Figure 6.5.



Réponse en fréquence des deux filtres considérés



Réponse impulsionnelle globale du banc de filtres

Figure 6.5 – Banc de filtres non symétrique construit pour le système hybride CR-ANC + MVDR

CR-ANC

Les paramètres d'implémentation ont été étudiés dans le Chapitre 4. On utilise donc, pour l'étage d'ANC, un algorithme **Normalized LMS** (NLMS) permettant d'adapter un filtre de longueur 30, avec un pas d'adaptation μ_0 de 0.1 (voir Équation (4.30), page 98).

Pour l'étage de compensation de distorsion, on utilise aussi un NLMS, qui adapte un filtre de longueur 50 avec un pas d'adaptation μ_0 de 0.1.

Il faut également un Détecteur d'activité vocale (**Voice Activity Detector**) (VAD) pour pouvoir contrôler l'adaptation de ces filtres, en discriminant les périodes où la parole est présente des périodes de bruit seul. Pour ce VAD, il est important que toutes les périodes de parole soient bien détectées (quitte à avoir certaines fausses détections) : si l'ANC s'adapte sur des périodes de parole, l'impact sur la qualité finale est plus important que si cet étage ne s'adapte pas sur certaines périodes de bruit seul. On a vu dans la Section 3.3 (page 78) que l'on utilisait un seuil sur le RSB **a posteriori** pour détecter les phases de bruit seul. Ces grandeurs seront donc calculées dans la bande des hautes fréquences, où l'on utilise un MVDR adaptatif. On choisit donc d'utiliser la moyenne de ces valeurs sur les hautes fréquences comme indicateur de présence de parole, et le VAD est résumé dans l'algorithme suivant :

Algorithme 2 Détection d'Activité Vocale.

```
if (1/N) Σ_{k=0}^{N-1} RSB_post(f_k, n) < Ω_CR-ANC then
    VAD = 0
else
    VAD = 1
end if
```

où $\Omega_{\text{CR-ANC}}$ est un seuil déterminé expérimentalement et on utilise ici $\Omega_{\text{CR-ANC}} = 4$.

MVDR

On s'intéresse ici aux paramètres utilisés pour le fonctionnement du MVDR adaptatif. On a besoin des paramètres pour :

— l'estimation de la propagation relative de la parole (Section 3.2.2, page 68),

— l'estimation des matrices spectrales du bruit (Section 3.3, page 78).

En ce qui concerne l'adaptation pour l'estimation de la propagation relative de la parole, on utilise un pas d'adaptation pour le NLMS fréquentiel de $\mu_0 = 0.5$, choisi expérimentalement (voir Équation (3.22)). De plus, le signal de référence pour l'adaptation est retardé de $\Delta = 16$ échantillons pour permettre une estimation non-causale. Pour estimer la puissance du signal à chaque fréquence, on utilise un lissage récursif sur le périodogramme avec un facteur de lissage $\beta = 0.8$ (voir Équation (3.24), page 69).

Pour l'estimation des matrices spectrales de bruit, on utilise un lissage récursif. La constante de lissage utilisée est $\alpha_0 = 0.75$, qui est ensuite modulée par la Probabilité de présence de parole (Speech Presence Probability) (SPP) (voir Équation (3.47), page 78). Le seuil sur le RSB a posteriori est choisi expérimentalement et vaut $\Omega = 1.5$, soit 1.8 dB.

6.2 Système MVDR + SDW-MWF

6.2.1 Principe général

La stratégie hybride est la même que celle présentée dans la Section 6.1.1, à ceci près que l'approche utilisée en basses fréquences n'est plus celle présentée dans le Chapitre 4, mais le SDW-MWF, présenté dans le Chapitre 5. Ce système fonctionne mieux en basses fréquences, comme on l'a vu dans la Section 5.2, est il est donc judicieux de l'utiliser ici.

Un schéma global de ce système hybride est présenté dans la Figure 6.6.

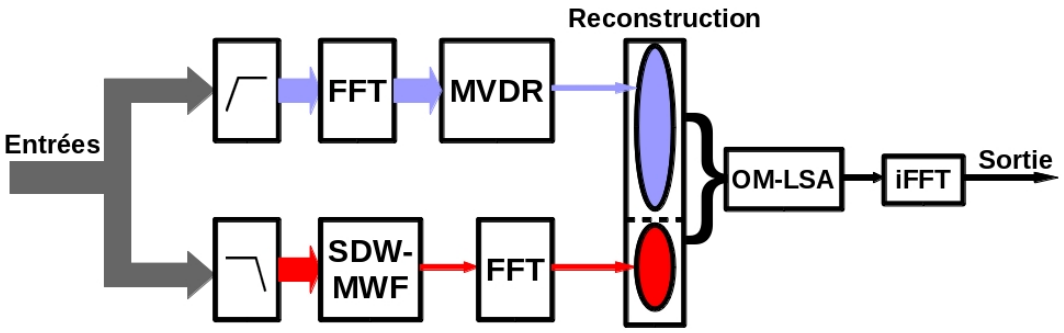


Figure 6.6 – Schéma global du système hybride MVDR + SDW-MWF. Les flèches épaisses représentent des signaux multicanaux.

6.2.2 Justification de l'approche

On a vu dans la Section 3.2.3 que le MVDR adaptatif n'allait donner de bonnes performances qu'en hautes fréquences, mais il n'a pas été montré que le SDW-MWF ne présenterait pas de bonnes performances en hautes fréquences : il a simplement été montré qu'il fonctionnerait mieux en basses fréquences qu'en hautes fréquences.

Il est donc possible qu'une approche SDW-MWF sur toute la bande soit plus efficace que la stratégie hybride proposée ici. Nous allons donc justifier cette approche hybride de façon expérimentale [Fox et al., 2012].

Pour cela, on se propose de comparer 3 systèmes :

MVDR La méthode présentée dans le Chapitre 3, appliquée sur toute la bande de fréquences.

SDW-MWF La méthode présentée dans le Chapitre 5, appliquée sur toute la bande de fréquences.

Hybride La stratégie hybride, avec le SDW-MWF en bande basse et le MVDR en bande haute, la coupure se faisant à 1 kHz.

On utilise une base de signaux enregistrés en voiture. Le bruit et la parole sont enregistrés séparément, de façon à avoir connaissance du signal de parole pur sur chacun des capteurs. Ceci nous permet d'utiliser deux critères de performance objective : le **Perceptual Evaluation of Speech Quality** (PESQ) [ITU-T, 2000] et le RSB segmental [Hansen and Pellom, 1998]. Ces systèmes sont évalués dans un contexte **Narrowband** : les signaux sont échantillonnés à 8 kHz, et on ne s'intéresse qu'à la bande de fréquence de 0 à 4 kHz. Notamment, l'architecture présentée dans la Figure 6.1 n'est pas prise en compte ici.

Les traitements sont faits sur des trames de 256 échantillons (soit 32ms), recouvrantes à 50%.

Le RSB segmental en sortie est calculé de la façon suivante :

$$\mathrm{RSB}_{\mathrm{seg}} = \frac{10}{M} \sum_{m=0}^{M-1} \log_{10} \left(\frac{\sum_{t=Nm}^{Nm+N-1} s(t)^2}{\sum_{t=Nm}^{Nm+N-1} (s(t) - \hat{s}(t))^2} \right) \tag{6.2}$$

M étant le nombre de fenêtres considérées, correspondant ici à 1min de signal, N la taille des fenêtres, correspondant à 128ms, $s(t)$ le signal de parole propre sur le capteur de référence (numéroté 1) et $\hat{s}(t)$ le signal en sortie du système. Comme il est d'usage de le faire, les fenêtres correspondant à un RSB inférieur à -10 dB ou supérieur à 30 dB ne sont pas prises en compte.

Le PESQ est présenté sur une échelle **Mean Opinion Score** (MOS), c'est à dire qu'il est modifié pour apparaître comme une note donnée par des utilisateurs, sur une échelle de 0 à 5 [Rix, 2003].

Les enregistrements sont faits à l'aide de deux capteurs unidirectionnels placés à 4 cm l'un de l'autre en position **Broadside** placés au niveau du plafonnier, comme sur le montage présenté dans la Figure 2.9 (page 38).

Le mélange des signaux de parole et de bruit se fait de façon à avoir des RSB d'entrée allant de -5 dB à 15 dB. Pour chaque RSB d'entrée (-5 dB, 0 dB, 5 dB, 10 dB et 15 dB), les performances sont évaluées sur 1 minute de signal.

Le PESQ en sortie est présenté dans la Figure 6.7. La Figure 6.8 présente le gain obtenu en RSB segmental, par rapport au RSB segmental mesuré sur le signal bruité de référence, présent sur le capteur 1.

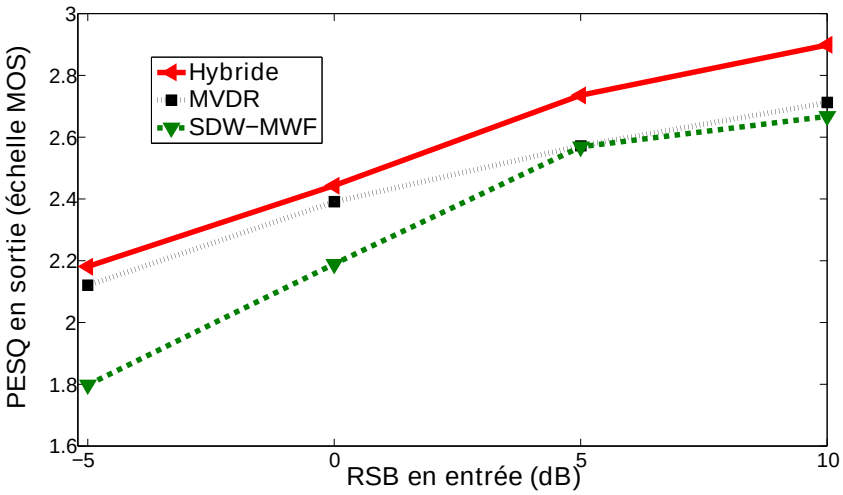


Figure 6.7 – PESQ mesuré pour chaque système (MVDR, SDW-MWF et hybride), à différents RSB d'entrée, sur une échelle allant de 0 à 5.

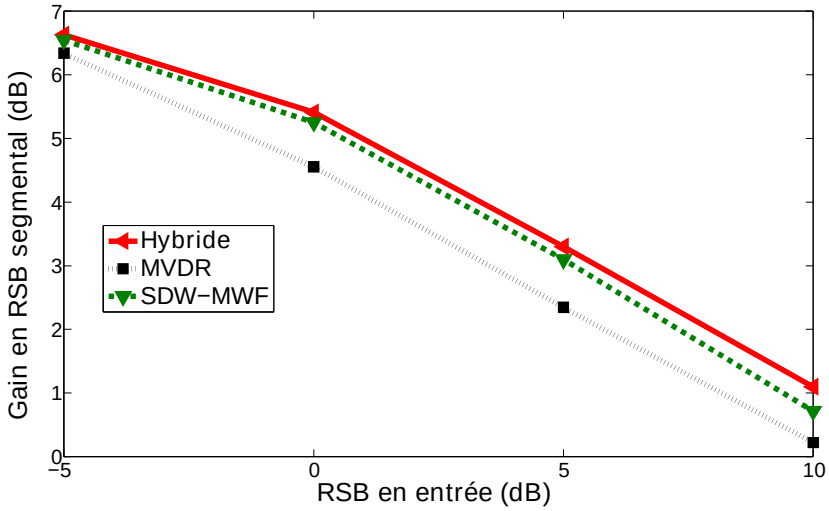


Figure 6.8 – RSB segmental mesuré pour chaque système (MVDR, SDW-MWF, et hybride) à différents RSB d'entrée.

On constate que la stratégie hybride obtient de meilleurs résultats que les méthodes pleine bande, et ce pour les deux critères présentés et pour toutes les conditions de RSB d'entrée. Cette stratégie semble donc judicieuse.

6.2.3 Implémentation

Antenne acoustique

On utilise pour la partie MVDR hautes fréquences la même antenne que dans la Section 6.1.3 : deux microphones unidirectionnels placés en Broadside, à 4 cm l'un de l'autre.

Pour le SDW-MWF en basses fréquences, on utilise un de ces capteurs comme référence (pour avoir la même référence sur toute la bande de fréquences), et un microphone omnidirectionnel placé en Endfire, proche (à 4 cm).