Amélioration de la compréhension de la parole et de l'écoute spatiale pour les malentendants appareillés

Adrien Llave

DOCTORAT BRETAGNE LOIRE / MATHSTIC

CentraleSupélec

THÈSE DE DOCTORAT DE

CentraleSupélec
COMUE UNIVERSITÉ BRETAGNE LOIRE

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Signal, Image, Vision*

Par

Adrien LLAVE

Amélioration de la compréhension de la parole et de l'écoute spatiale pour les malentendants appareillés

Thèse présentée et soutenue à CentraleSupélec, le 11 mars 2022
Unité de recherche : CentraleSupélec, institut d'électronique et des technologies du numérique, UMR 6164
Thèse N° : 2022CSUP0003

Rapporteurs avant soutenance :

Véronique Zimpfer	Chercheuse HDR	Institut franco-allemand de recherches de Saint-Louis, Saint-Louis
Alexandre Garcia	Professeur des Universités	Conservatoire national des arts et métiers, Paris

Composition du Jury :

Président :	Alexandre Garcia	Professeur des Universités	Conservatoire national des arts et métiers, Paris
Rapporteurs :	Véronique Zimpfer	Chercheuse HDR	Institut franco-allemand de recherches de Saint-Louis, Saint-Louis
	Alexandre Garcia	Professeur des Universités	Conservatoire national des arts et métiers, Paris
Examinateurs :	Mathieu Paquier	Professeur des Universités	Université de Bretagne Occidentale, Brest
	Romain Sérizel	Maître de conférences	Université de Lorraine, Nancy
Dir. de thèse :	Renaud Séguier	Professeur	CentraleSupélec, Cesson-Sévigné
Encadrant :	Simon Leglaive	Maître de conférences	CentraleSupélec, Cesson-Sévigné

"All knowledge degenerates into probability."
David Hume
A Treatise of Human Nature, 1739

Résumé

Les personnes affectées par une perte auditive légère ou modérée peuvent bénéficier de prothèses auditives acoustiques pour compenser celle-ci. Les algorithmes de compensation de niveau sonore permettent de ramener les sons faibles dans la plage audible de l'auditeur•rice [1] sans pour autant dépasser le niveau d'inconfort. Néanmoins, dans un environnement complexe, c'est-à-dire en présence de forte réverbération, de bruit ambiant ou de multiples locuteur•rice•s, ces algorithmes déforment le signal au point que celui-ci ne semble plus naturel à l'auditeur•rice. En particulier, ils distordent les indices acoustiques nécessaires à la localisation auditive. Celle-ci, bien que secondaire vis-à-vis de l'intelligibilité de la parole, reste centrale pour se situer par rapport à un danger, par exemple. De plus, l'audition humaine bénéficie aussi de mécanismes psychoacoustiques de démasquage basés sur ces mêmes indices permettant d'étendre la compréhension de la parole dans des situations très bruitées, ce phénomène est souvent appelé l'effet *cocktail-party*. Il est alors d'usage de placer un algorithme de débruitage en amont pour réduire l'effet néfaste du bruit sur l'intelligibilité de la parole. Cependant, les méthodes de débruitage multicanales, dites de *beamforming*, détruisent aussi les indices de localisation.

Depuis une quinzaine d'années, de nombreux travaux ont porté sur la préservation des indices de localisation au travers des étages de débruitage et de compensation de niveau sonore, indépendamment. Ces travaux ont été menés dans des communautés scientifiques relativement distinctes et peu de travaux les considèrent ensemble.

Dans cette thèse, nous proposons de considérer les fonctions de débruitage et de compensation de niveau sonore au sein d'un seul formalisme en matière de modélisation et d'algorithme de traitement du signal. Cette approche conduit à des formulations de problèmes plus complexes, et il est alors nécessaire d'introduire de nouvelles hypothèses pour aboutir à une solution réalisable dans le contexte des prothèses auditives.

Dans un premier temps, nous considérons une scène sonore composée d'un

[1] Nous adoptons dans cette thèse une forme d'écriture inclusive.

locuteur cible et d'un bruit ambiant. Une évaluation objective et perceptive permet de montrer que la méthode proposée préserve mieux les caractéristiques acoustiques de la scène sonore tout en obtenant les mêmes performances de compréhension de la parole que les méthodes état-de-l'art.

Dans un second temps, nous mettons en évidence que la méthode proposée ne permet pas de préserver les indices de localisation interauraux, hors de la direction cible. La prise en compte de cette information rend difficile l'obtention d'une solution analytique. Nous proposons alors trois méthodes visant à préserver les indices interauraux basées sur des approximations différentes. Deux d'entre elles offrent un compromis satisfaisant entre les performances originales de la méthode et la préservation des indices interauraux.

Enfin, nous considérons un scénario composé de plusieurs locuteurs et d'un faible bruit ambiant. Une méthode unifiant débruitage et compensation de niveau sonore a déjà été proposée pour ce scénario dans la littérature. Cependant, le problème d'optimisation sous-jacent se retrouve très contraint pour préserver à la fois les locuteur·rice·s de la scène sonore, les indices de localisation associés et réduire le niveau du bruit ambiant. De plus, les prothèses auditives sont des systèmes électroniques avec une capacité de calcul très réduite et une très forte contrainte en matière de latence. Nous proposons de considérer une hypothèse supplémentaire sur la parcimonie des signaux de parole dans le plan temps-fréquence de sorte à diminuer le coût calculatoire tout en remplissant au mieux les objectifs du problème d'optimisation.

Abstract

The hearing-impaired listeners experiencing a mild to moderate sensorineural hearing loss can benefit from hearing aids. The loudness compensation methods allow for bringing back the soft sound into the audible range of the listener without exceeding the pain threshold. However, in a complex auditory scene, *i.e.* with a lot of reverberation, ambient noise, or/and various speakers, these algorithms distort the acoustic signal in many ways, making it sound unnatural. Notably, it distorts the localization cues. Although secondary to speech intelligibility, auditory localization remains crucial for situating oneself in relation to a danger, for example. Moreover, human hearing takes advantage of psychoacoustic unmasking mechanisms based on the same acoustical cues, allowing us to keep good speech comprehension scores even in highly noisy situations. This phenomenon is called the *cocktail-party* effect, or the spatial release from masking. A denoising algorithm usually achieves this task before the loudness compensation. However, the multichannel denoising methods are known to remove these cues.

For about fifteen years, many works have focused on preserving localization cues through the denoising stage and the loudness compensation ones, independently. These methods have been developed in relatively different scientific communities, and few studies have considered the two stages as a whole.

In this thesis, we propose considering the denoising and loudness compensation tasks into the same formalism to overtake the limitations due to the local trade-off achieved at the level of each stage. This approach leads to a complex optimization problem which has to be simplified to get a method compatible with the constraints of the hearing aids application.

First, we consider an auditory scene composed of one speech target source and ambient noise. An objective and perceptual evaluation shows that the proposed method achieves a better global auditory scene preservation while getting the same speech intelligibility performance than the state-of-the-art methods.

Second, we show that the proposed method distorts the interaural localization cues outside the target direction. Taking these cues into account makes it

difficult to find a closed-form solution. We propose three methods to preserve them based on different approximations. Two of them can achieve a satisfying trade-off between the method original performance and the preservation of interaural cues.

Finally, we consider a multi-talker auditory scene with an ambient noise A method of unifying denoising and loudness compensation has already been proposed in the literature. However, the underlying optimization problem is highly constrained to preserve both the speakers and the associated location cues and to reduce the ambient noise level. Moreover, hearing aids are electronic systems with a very small computational capacity and a strong latency constraint. We propose to consider an additional assumption on the sparsity of speech signals in the time-frequency domain to decrease the computational cost while fulfilling the objectives of the optimization problem as well as possible.

Remerciements

En premier lieu, je souhaite remercier Renaud Séguier qui m'a fait confiance et m'a accompagné tout au long de ce travail comme directeur de thèse. Je tiens tout autant à remercier Simon Leglaive qui a rejoint l'encadrement de cette thèse en cours de route, dont j'ai appris beaucoup et qui a été d'une aide précieuse. Merci aussi à Stéphane Laurent pour les conversations enrichissantes que nous avons pu avoir autour des prothèses auditives.

Ce travail de thèse a été évalué par Alexandre Garcia, Véronique Zimpfer, Mathieu Paquier et Romain Sérizel que je remercie grandement d'avoir accepté de participer au jury et pour leur remarques. Un grand merci aussi à Rozenn Nicol et Olivier Macherey qui ont accepté de suivre mon travail de thèse au sein du comité de suivi du doctorant. Leur regard extérieur et leurs mots m'ont été précieux.

Merci aux quarante-deux oreilles, et moitié moins de cerveaux, qui ont accepté de participer à la campagne de test d'écoute que j'ai menée dans le cadre de cette thèse. Il va sans dire que leurs réponses ont grandement contribuées à mon travail. L'anonymat que je leur ai promis ne me permet pas de les nommer, mais je n'en oublie aucun•e !

Merci aussi à tous•tes les collègues de CentraleSupélec : Bastien, Corentin, Morgane, Lilian, Eloïse, Samir, Guénolé, pour tous les moments partagés, aussi bien scientifiques qu'amicaux. *Un ringraziamento speciale a* Diego pour le partage tout aussi scientifique que musical. Qui serais-je sans remercier Bastien, Léa et Emma pour tout leur soutien, et Loïc, meilleur colocataire tout simplement. Merci à toi, Pauline. Ce travail de thèse n'auraient pas pu aboutir non plus de cette manière sans tous•tes les ami•e•s avec qui j'ai partagé la vie pendant cette période. Qu'ielles ne m'en veuillent pas de ne pas les nommer exhaustivement, je suis certain qu'ielles se reconnaitront et j'ai hâte de les retrouver autour d'un verre, sur un parquet, devant ou sur une scène, sur les chemins... Et merci enfin à Hélène de m'avoir accompagné et soutenu ces dernières années.

Un merci particulier à Benoit Hauray pour son enseignement et la bienveillance dans ses cours qui m'ont permis de pas oublier que j'avais aussi un

corps à faire bouger !

Enfin, merci à Nathalie, ma mère a qui je dois tellement, ainsi qu'à Malou, qui a entretenu ma curiosité pour les sciences toute mon enfance.

Table des matières

Liste des figures

[2] Données aberrantes.

Liste des tables

Acronymes

AMA angle minimal audible

ATF fonction de transfert acoustique (de l'anglais *Acoustic Transfer Function*)

BTE derrière l'oreille (de l'anglais *Behind-The-Ear*)

CD compresseur de dynamique

CLCMV LCMV avec compression indépendante

CMMWF filtre de Wiener multicanal multicibles compressées

DOA direction d'arrivée (de l'anglais *Direction Of Arrival*)

DRR rapport signal direct à réverbéré (de l'anglais *Direct-to-Reverberation Ratio*)

ECR taux de compression effectif (de l'anglais *Effective Compression Ratio*)

EQM erreur quadratique moyenne

ERB largeur de bande rectangulaire équivalente (de l'anglais *Equivalent Rectangular Bandwidth*)

FAS filtrage et sommation (de l'anglais *Filter and Sum*)

FIR réponse impulsionnelle finie (de l'anglais *Finite Impulse Response*)

H-NF nullformer préservant les HRTFs

HLS simulation de perte auditive (de l'anglais *Hearing Loss Simulation*)

HRIR réponse impulsionnelle liée à la tête (de l'anglais *Head Related Impulse Response*)

HRTF fonction de transfert liée à la tête (de l'anglais *Head Related Transfer Function*)

IC cohérence interaurale

IC_{max} maximum de l'absolu de la cohérence interaurale

ID indice de directivité

IIR réponse impulsionnelle infinie (de l'anglais *Infinite Impulse Response*)

ILD différence de niveau interaural (de l'anglais *Interaural Level Difference*)

ISSD différence spectrale inter-sujet (de l'anglais *Inter-Subject Spectral Difference*)

ITC dans le canal (de l'anglais *In-The-Canal*)

ITD différence de temps interaural (de l'anglais *Interaural Time Difference*)

ITE dans l'oreille (de l'anglais *In-The-Ear*)

ITF fonction de transfert interaurale (de l'anglais *Interaural Transfer Function*)

ITF1-NF nullformer préservant les ITFs selon la définition probabiliste

ITF2-NF nullformer préservant les ITFs selon la définition déterministe

JLC-NF nullformer conjointement linéairement contraint (de l'anglais *jointly linearly constrained nullformer*)

jnd plus petite différence perceptible (de l'anglais *Just Noticeable Difference*)

LCMV linéairement contraint à variance minimale (de l'anglais *Linearly Constrained Minimum Variance*)

MaxID maximisant l'ID

MBSTOI STOI binaural modifié

MPDR sans distorsion à puissance minimale (de l'anglais *Minimum Power Distortionless Response*)

MSC carré de l'amplitude de la cohérence interaurale (de l'anglais *Magnitude Squared Coherence*)

MSNR à RSB maximal (de l'anglais *Maximum Signal-to-Noise Ratio*)

MUSHRA test multi stimuli avec référence cachée et une ancre (de l'anglais *MUlti Stimulus test with Hidden Reference and Anchor*)

MVDR sans distorsion à variance minimale (de l'anglais *Minimum Variance Distortionless Response*)

MVDR-ICP MVDR préservant l'IC post-filtré

MWF filtre de Wiener multicanal (de l'anglais *Multi-channel Wiener Filter*)

MWF-IC MWF avec préservation de l'IC

MWF-ITF MWF avec préservation de la fonction de transfert interaurale

MWF-N MWF avec estimation partielle du bruit

PPP probabilité de présence de la parole

PPPM probabilité de présence de la parole multicanale

RAU unité rationalisée par l'arcsinus (de l'anglais *Rationalized Arcsine Units*)

RSB rapport signal à bruit

rsbCD compresseur de dynamique piloté par le RSB

RTF fonction de transfert relative (de l'anglais *Relative Transfer Function*)

SAR rapport signal à artefact (de l'anglais *Signal to Artifact Ratio*)

SCs indices spectraux (de l'anglais *Spectral Cues*)

SDR rapport signal à distorsion (de l'anglais *Signal to Distortion Ratio*)

SDW-MWF MWF à distorsion de parole pondéré (de l'anglais *Speech-Distortion-Weighted-MWF*)

SIR rapport signal à interférence (de l'anglais *Signal to Interference Ratio*)

SRP puissance de la réponse visée (de l'anglais *Steered Response Power*)

SRT seuil de compréhension de la parole (de l'anglais *Speech Reception Threshold*)

STOI mesure objective de l'intelligibilité à court terme (de l'anglais *Short Term Objective Intelligibility*)

T-F temps-fréquence

TFCT transformée de Fourier à court terme

TFD transformation de Fourier discrète

VAD détecteur d'activité vocale (de l'anglais *Voice Activity Detector*)

WF filtre de Wiener (de l'anglais *Wiener Filter*)

Chapitre 1

Introduction

Dans ce chapitre, nous décrivons succinctement comment se caractérisent les pertes auditives que nous considérons dans la suite de ce travail ainsi que de la manière de caractériser la compréhension de la parole et les perceptions de l'espace sonore et la localisation auditive. Nous rapportons aussi les performances dans ces tâches, aussi bien typiques que pathologiques. Enfin, nous présentons les questions qui sont traitées dans les chapitres suivants, l'organisation du manuscrit et les publications associées à ce travail de thèse.

1.1 Pertes auditives

1.1.1 Caractérisation

Une perte auditive désigne une diminution partielle à totale d'une ou plusieurs capacités associées à l'audition (détection, compréhension, localisation, etc.). Ce phénomène est avant tout subjectif, consistant à faire l'expérience d'un décalage avec le reste de ses congénères. Néanmoins, celui-ci a des conséquences souvent très concrètes sur la qualité de vies des personnes touchées et est, en ce sens, objectivable. Ainsi, la psychologie expérimentale permet de quantifier le phénomène par des expériences évaluant la réponse des auditeur•rice•s à une tâche faisant intervenir différentes capacités attendues du système auditif.

L'outil de diagnostic privilégié est l'audiogramme. Il consiste à déterminer le seuil de niveau sonore à partir duquel l'auditeur•rice est capable d'entendre un son sinusoïdal accordé tour à tour sur les fréquences 0,5, 1, 2 et 4 kHz. Cet outil fait intervenir la capacité de détection du système auditif. Il présente l'avantage d'être très simple à mettre en œuvre et de limiter le nombre de paramètres de l'expérience.

Néanmoins, le système auditif a de nombreuses fonctions, plus ou moins complexes, et ne peut être résumé à la détection de la présence de signaux sinusoïdaux, dits sons purs. Par exemple, bien qu'en général un audiogramme montrant une perte auditive plus importante implique une compréhension de la parole plus faible, il existe une variabilité importante dans les performances dans cette tâche pour des sujets présentant le même audiogramme [Popelka et al., 2016]. Ceci peut être dû au fait qu'une perte auditive peut se traduire autrement que par une diminution de la sensibilité, comme par exemple par la diminution de la résolution temporelle, ou fréquentielle. Par ailleurs, des travaux récents ont montré que certaines personnes rapportent des difficultés auditives tout en présentant un audiogramme normal. Ce phénomène est appelé « pertes auditives cachées » [1] [Füllgrabe, 2015].

Dans le cadre de cette thèse nous nous intéressons particulièrement aux

[1] *hidden hearing losses*, en anglais.

pertes auditives neurosensorielles bilatérales[2] dues à l'âge, aussi appelées la presbyacousie. Ces pertes sont généralement liées à la détérioration de l'organe de Corti dans la cochlée [Gates and Mills, 2005]. Elles sont principalement caractérisées par une augmentation du seuil de perception, sans modification du niveau maximum toléré. Généralement, l'importance de la perte est corrélée avec la fréquence. Ce phénomène est appelé le recrutement de sonie,[3] il peut aussi bien être mis en évidence par des tests de perception du niveau sonore [Steinberg and Gardner, 1937] que par des mesures physiologiques, *e.g.* la tension électrique sur la membrane basilaire [Eggermont, 1977] comme illustré en Fig. 1.1 pour un son pur de fréquence 2 kHz. On observe sur celle-ci que pour un malentendant, l'excitation du nerf auditif est semblable à celle d'un normoentendant pour un son d'intensité élevé (supérieur à 70 dB au minimum, selon la perte) mais que cette courbe diverge en-deçà. Pour des niveaux inférieurs, la sensibilité est plus faible et donc la pente est plus forte. Les pertes auditives peuvent être classifiées en utilisant le pire point de l'audiogramme, *i.e.* à la pire oreille et pire fréquence, de la manière suivante :

— légère : 25-45 dB ;
— modérée : 45-65 dB ;
— sévère : 65-85 dB ;
— profonde : $\geq$ 85 dB.

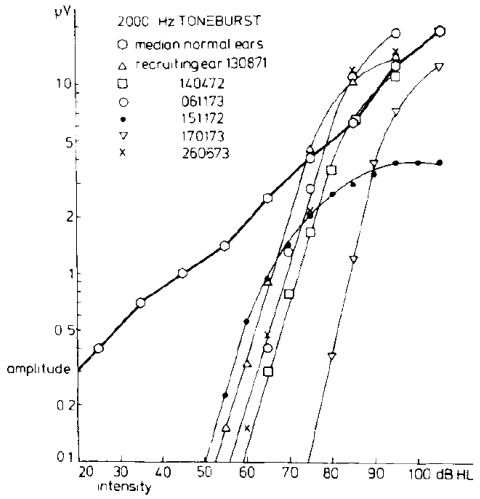
Pour une perte allant jusqu'à 60 dB, le principal moyen de rétablir en partie les fonctions du système auditif est l'appareillage avec une ou deux prothèses auditives acoustiques. Celles-ci sont équipées d'un ou plusieurs microphones et d'un haut-parleur transmettant le son traité jusqu'à l'entrée du canal auditif.

1.1.2 Performances de compréhension de la parole

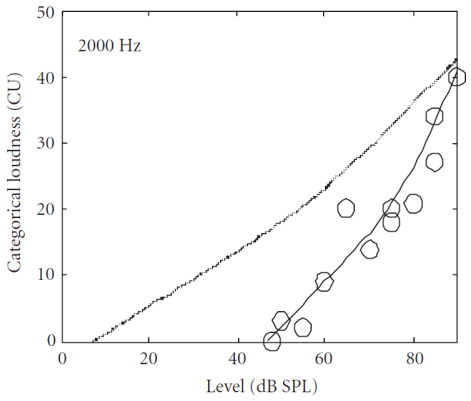
La compréhension de la parole est une des capacités du système auditif que l'on souhaite rétablir en priorité. En effet, elle est considérée comme la plus critique car de celle-ci dépend l'intégration sociale et l'autonomie de l'auditeur•rice. C'est pourquoi l'intelligibilité de la parole est un des critères les plus importants et qui a reçu le plus d'attention dans la conception des prothèses auditives.

Dans cette sous-section, nous décrivons dans les grandes lignes les caractéristiques du signal de parole puis nous présentons les différents critères utilisés pour évaluer sa compréhension. Enfin, nous présentons les performances pour ces critères des auditeur•rice•s normoentendant•e•s et malentendant•e•s avec et sans appareils, et ce, dans différents scenarii.

[2] qui touchent les deux oreilles.
[3] *loudness recruitment*, en anglais.



(a)



(b)

FIGURE 1.1 – (a) Excitation électrique du nerf en fonction du niveau de pression acoustique de la médiane des sujets normoentendants et pour six sujets malentendants (extraite de [Eggermont, 1977]). (b) Niveau sonore perçu en fonction du niveau de pression acoustique, pour des malentendants (cercles) et pour des normoentendants (ligne pointillée) [Hamacher et al., 2005]. Pour les deux graphiques, le stimulus employé est une sinusoïde accordée à 2 kHz.

Caractérisation de la parole La parole est un stimulus complexe pour lequel il est difficile d'associer directement une caractéristique physique à son intelligibilité. Sa dynamique, *i.e.* la différence entre le niveau le fort et le plus faible, est de l'ordre de 50 dB [Cox et al., 1988] et on considère que son niveau moyen lors d'une conversation est de l'ordre de 65 dB_{SPL}. La bande-passante nécessaire à sa compréhension s'étend de 100 à 8000 Hz, alternant des périodes harmoniques (sons voisés [4]) et inharmoniques (sons non-voisés) plutôt en haute fréquence [Stelmachowicz et al., 2001]. Ces deux caractéristiques physiques forment le socle minimal de description de la voix qui nous seront utiles dans la suite.

Critères d'évaluation On peut diviser les critères d'évaluation de perception de la parole en deux catégories : perceptifs et objectifs. Le critère considéré comme le plus écologique reste le pourcentage de mots reconnus et sa version corrigée : l'unité rationalisée par l'arcsinus (de l'anglais *Rationalized Arcsine Units*, RAU) [Studebaker, 1985, Sherbecoe and Studebaker, 2004]. Lorsque l'on évalue l'intelligibilité de la parole en présence de bruit, on obtient alors un

[4] Faisant intervenir les cordes vocales.

point de mesure par niveau de bruit. Cependant, on préfère en général avoir un critère synthétique permettant de classer différentes conditions d'expérimentales. Par ailleurs, cela implique un grand nombre de mesures, et sachant que le temps de test est limité dû à la fatigue des auditeur•rice•s, il contraint le nombre de conditions expérimentales testées. Pour cela, il a été proposé de mesurer le rapport signal à bruit (RSB) pour lequel un certain pourcentage (en général 50 %) de la parole est comprise, appelé le seuil de compréhension de la parole (de l'anglais *Speech Reception Threshold*, SRT) [Levitt, 1971, Nilsson et al., 1994]. Il faut noter que cette procédure est adaptative, elle réduit le temps de test mais le nombre de stimuli présentés pour mesurer un SRT est variable.

Différents corpus de parole existent pour réaliser ce genre de test. La tâche peut consister à reconnaitre des chiffres, des syllabes isolées ou des phrases entières avec plus de sens. Dans le dernier cas, il a par exemple été proposé de suivre la construction syntaxique suivante : *prénom, verbe, nombre, objet, couleur* [Hagerman, 1982]. Cela permet de pouvoir générer un très grand nombre de phrases syntaxiquement correctes à partir de listes de mots d'une taille raisonnable, en piochant aléatoirement dans celles-ci. De plus, ce genre de corpus doit tâcher d'être représentatif de la répartition des phonèmes dans la langue [Jansen et al., 2012].

Néanmoins, les tests subjectifs restent couteux et compliqués à mettre en œuvre. Il a donc été proposé des modèles de prédiction de l'intelligibilité de la parole par rapport à du bruit ou des artéfacts. En premier lieu, l'indice d'intelligibilité de la parole (*Speech Intelligibility Index*, SII) [French and Steinberg, 1947, ANSI, 1997] [5] donne une valeur entre 0 et 1 et n'est pertinent que pour les bruits stationnaires. Il est principalement basé sur une mesure du RSB pondéré par bande de fréquence. Un autre modèle, l'indice de transmission de la parole (*Speech Transmission Index*, STI) se base, quant à lui, sur une mesure de la modulation de l'enveloppe des signaux. Cela permet de considérer l'influence de distorsions non-linéaires sur l'intelligibilité [Goldsworthy and Greenberg, 2004]. Cependant, il prédit une amélioration de l'intelligibilité pour certains algorithmes de débruitage qui sont pourtant connus pour ne pas avoir d'effet sur celle-ci [Ludvigsen et al., 1993]. Par ailleurs, la corrélation entre les résultats de ces modèles et le pourcentage de mots reconnus est assez faible. Un modèle est venu répondre à ces problèmes, la mesure objective de l'intelligibilité à court terme (de l'anglais *Short Term Objective Intelligibility*, STOI) [Taal et al., 2010] se basant sur le RSB à court terme par bande de fréquence. Celui-ci est devenu le plus populaire aujourd'hui.

Comme nous le verrons dans la suite, le RSB n'est pas le seul facteur déter-

[5] Anciennement appelé l'indice d'articulation.

minant les performances de compréhension de la parole. D'autres mécanismes, propre à l'écoute binaurale, *i.e.* à deux oreilles, permettent de comprendre la parole dans des situations particulièrement défavorables. Les prédicteurs ont été amélioré en prenant en compte le modèle d'alignement et d'annulation [6] [Durlach, 1963]. Sans rentrer dans le détail, celui-ci fait l'hypothèse que le système auditif combine les signaux des oreilles droite et gauche en les soustrayant après leur avoir appliqué un gain et un délai réglé de sorte à réduire le RSB à court-terme. Ainsi, le SII a été étendu aux signaux binauraux [Lavandier et al., 2012] [7] et permet de prendre en compte le niveau de réverbération et le masquage de plusieurs locuteurs interférant. Plus récemment, le STOI binaural modifié (MBSTOI) [Andersen et al., 2018] a été proposé pour étendre le STOI d'une manière similaire.

Performances dans le bruit Le système auditif est remarquablement robuste pour comprendre la parole dans le bruit. Dans un scénario composé d'une source de parole et d'un bruit spatialement diffus, le SRT est de l'ordre de −6 dB pour les normoentendant•e•s en écoute monaurale [Jansen et al., 2012]. Par ailleurs, l'écoute binaurale permet d'améliorer le SRT de 3 dB comparé à une écoute monaurale [Arweiler and Buchholz, 2011], autant pour les normoentendant•e•s que pour les malentendants. Ce phénomène est appelé l'effet *cocktail-party* [Cherry, 1953, Bronkhorst and Plomp, 1989, Bronkhorst, 2015, Middlebrooks et al., 2017]. En particulier, cette tâche est plus facilement accomplie lorsque les sources cible et de bruit ne se situent pas dans la même direction. Dans un scénario où le bruit est composé de locuteurs interférant et bien localisés dans l'espace, le bénéfice apporté par l'effet cocktail-party sur le SRT peut atteindre jusqu'à 12 dB [Bronkhorst and Plomp, 1988, Allen et al., 2008] pour les normoentendant•e•s. Dans ce cas, deux mécanismes ont été identifiés pour expliquer l'amélioration de la compréhension de la parole par rapport à l'écoute monaurale : d'une part, l'écoute dite à la *meilleure-oreille* et d'autre part, le démasquage binaural.

Les résultats de [Brungart and Iyer, 2012] et [Edmonds and Culling, 2006] suggèrent que le système auditif exploite le son venant de l'oreille bénéficiant du meilleur RSB mais ne traite pas indépendamment chaque bande fréquentielle, c'est ce qu'on appelle l'écoute à la *meilleure-oreille*. Ce mécanisme repose sur le masquage acoustique effectué par la tête lorsqu'une source est située sur le coté (pour l'oreille du coté opposé, dite controlatérale). Ce masquage se manifeste par une différence de niveau entre les deux oreilles, appelée la différence de niveau interaural (de l'anglais *Interaural Level Difference*, ILD). Ainsi, si la

[6] *equalization-cancellation theory*, en anglais.
[7] Cette version prédit le SRT et non plus un score entre 0 et 1.

source cible se trouve en face de l'auditeur•rice et le bruit sur le coté, ce dernier arrive plus fort d'un coté que de l'autre et donc le RSB à l'oreille controlatérale est supérieur à celle du même coté, dite ipsilatérale. C'est cette différence qui est exploitée pour améliorer l'intelligibilité dans ce genre de scénario.

Par ailleurs, de sorte à créer un percept auditif unique et cohérent, les signaux droite et gauche sont combinés dans le cerveau moyen au niveau du complexe olivaire supérieur et du colliculus inférieur [Skottun et al., 2001]. Pour décrire cela, le modèle d'égalisation-annulation [Durlach, 1963] a été proposé et, bien que simple et non basé sur un modèle fonctionnel, il est suffisant pour expliquer le gain apporté par le démasquage binaural. Dans ce modèle, les signaux droite et gauche sont retardés et amplifiés de sorte à aligner au mieux le signal correspondant à la source masquante, c'est l'étape dite d'égalisation. En effet, lorsqu'une onde sonore arrive depuis le coté de l'auditeur•rice, celle-ci arrive retardée dans l'oreille controlatéral par rapport à l'oreille ipsilatérale dû au temps de propagation dans le champ de pression acoustique. Ce retard est appelé la différence de temps interaural (de l'anglais *Interaural Time Difference*, ITD). Ensuite, les deux signaux ainsi traités sont soustraits de sorte à minimiser le niveau du bruit dans la sortie. Ce modèle se rapproche fortement de la stratégie adoptée par certains algorithmes de débruitage multicanaux de type filtrage spatial [Elko and Anh-Tho Nguyen Pong, 1995, Benesty et al., 2016]. [Edmonds and Culling, 2005] montrent que l'ITD n'a pas besoin d'être cohérente d'une bande fréquentielle à une autre pour que le système auditif en fasse le bénéfice. Cela suggère que le traitement est fait indépendamment pour chaque bande de fréquence, contrairement à ce qui fait pour l'ILD.

Selon certains auteurs [Glyde et al., 2013a], l'ILD joue un rôle plus important que l'ITD dans le gain apporté par l'écoute binaural sur le SRT mais d'autres études [Ellinger et al., 2017] arrivent à des résultats contraires.

De manière général, avec des pertes auditives, les mécanismes permettant de bénéficier de la diversité spatiale de la scène sonore pour la compréhension de la parole en présence de sources masquantes fonctionnent moins bien [Glyde et al., 2013b]. [Arbogast et al., 2005] et [Marrone et al., 2008a] montrent que les malentendant•e•s perdent 5 à 6 dB sur le SRT par rapport aux normoentendant•e•s. Ces derniers montrent aussi un effet de diminution du bénéfice lié à l'âge pour les deux populations (normo et malentendantes) mais ce résultat n'a pas été confirmé dans d'autres expériences [Glyde et al., 2013b]. Par ailleurs, ils montrent que la réverbération réduit de 4 dB les bénéfices de la séparation spatiale sur le SRT. En effet, celle-ci peut aussi jouer un rôle néfaste sur cette tâche car elle limite l'accès à l'ITD. Dans une autre étude [Marrone et al., 2008b], les auteurs s'intéressent à l'effet du port des prothèses auditives. Ils montrent ainsi que les malentendant•e•s perdent 1 dB de bénéfice lié à la séparation spatiale avec un appareillage bilatéral et perdent encore 1 dB avec un

appareillage unilatéral. Cela montre l'intérêt d'être appareillé des deux cotés mais suggère que les prothèses auditives ont aussi un effet délétère sur les indices acoustiques permettant de bénéficier de la séparation spatiale de la cible et du bruit.

Nous avons vu que la compréhension de la parole était améliorée lorsque la source et celle de bruit sont spatialement séparées. Cet avantage est lié aux mêmes indices acoustiques qui permettent la localisation, l'ILD et l'ITD. Dans la sous-section suivante, nous allons décrire plus particulièrement les performances de localisation puis dans la section suivante nous décrirons en particulier comment est modélisée l'écoute spatiale et quelles sont les informations nécessaires au système auditif pour accomplir cette tâche.

1.1.3 Performances de localisation

Pour évaluer les performances de localisation, on s'appuie principalement sur deux critères : l'erreur angulaire d'une part et les erreurs de quadrant d'autre part [Middlebrooks, 1999b]. La première se définie comme l'écart-type des erreurs et permet d'évaluer la précision de la localisation tandis que la seconde est définie comme le taux de réponse dont l'erreur est supérieure à $\pm 90°$ et évalue les performances globales de localisation en matière de confusion entre quadrants (avant-arrière, haut-bas et éventuellement gauche-droite). Un critère supplémentaire peut être ajouté pour évaluer de manière encore plus fine la précision de la localisation, c'est la plus petite différence perceptible (de l'anglais *Just Noticeable Difference*, jnd) de l'angle, ou l'angle minimal audible (AMA). Celui-ci peut varier suivant la direction et le stimulus employé.

Pour une population normoentendante, le plus petit AMA mesuré est de 1° pour un signal sinusoïdal de 750 Hz présenté à la direction frontale [Mills, 1958] et peut monter jusqu'à 12° avec un bruit blanc [Häusler et al., 1983]. Pour une population malentendante, celui-ci est supérieur à 7° et peut monter jusqu'à 30° [Mills, 1958].

En ce qui concerne l'erreur angulaire, celle-ci est en moyenne de 4° pour les normoentendant•e•s tandis qu'elle est de l'ordre de 13° pour les malentendant•e•s [Van den Bogaert et al., 2006]. L'appareillage bilatéral n'améliore pas les performances et même les empirent avec 16° d'erreur angulaire en moyenne. L'ajout d'un algorithme de débruitage empire encore un peu plus les performances avec une moyenne de 18° d'erreur. Ces résultats ont été obtenus pour un signal d'alerte (sonnerie de téléphone) et en absence de bruit. Des résultats similaires ont été trouvé en utilisant de la parole comme stimulus [Best et al., 2010]. En présence de bruit (RSB de 0 dB), la localisation de la parole est très mauvaise pour les malentendants allant de 20-25° malgré un temps d'acclimatation de trois semaines [Keidser et al., 2009] et jusqu'à 30-35° même après 10 à 15

semaines d'acclimatation [Drennan et al., 2005].

Une fois appareillés, les résultats sont très variables allant de 22 à 45 % d'erreur [Best et al., 2010, Vaillancourt et al., 2011, Hassager et al., 2017b]. Enfin, il faut noter que le taux de confusion avant-arrière diminue fortement lorsque l'on rajoute des indices visuels dans le test.

Les métriques considérées jusqu'à maintenant permettent une évaluation bien contrôlée des performances de localisation. Cependant, elles nous informent peu sur les cas d'utilisation réel de la localisation auditive. Une étude [Brimijoin et al., 2014] s'est intéressée à une tâche plus complexe, consistant à se réorienté vers un nouvel interlocuteur qui s'adresse à nous. Plus écologique, cette tâche est néanmoins plus difficile à évaluer. Les auteurs ont choisi d'évaluer le temps d'orientation, la complexité du mouvement entre la nouvelle cible et la précédente, et enfin le taux de mauvaises orientations au début du mouvement. Pour toutes ces métriques, ils ont trouvé que l'ajout d'un algorithme de débruitage diminue les performances lorsque le nouvel interlocuteur est placé à plus de 120° sur le coté. En particulier, le taux de mauvaises orientations y est supérieur à 30 %. Cela est dû au fait que l'algorithme de débruitage employé limite sa direction de visée à la direction frontale et détruit les indices de localisation des sources hors-axe. Ce résultat nous pousse à considérer des algorithmes plus souples en matière de direction de visée.

Pour une revue extensive de ce sujet, la•e lecteur•rice pourra se référer à [Akeroyd, 2014] et à [Denk et al., 2019]. Dans la section suivante, nous allons voir comment est modélisée la localisation auditive humaine et quelles sont les informations présentes dans les signaux audio dont le système auditif se sert pour accomplir cette tâche.

1.2 Écoute spatiale

Le système auditif accomplit un certain nombre de tâches que l'on peut séparer en deux catégories : d'une part, *identifier* les sources sonores en présence ; d'autre part, les *localiser*. L'identification consiste aussi bien à attribuer une cause au signal perçu, *e.g.* un locuteur à un signal de parole ou une porte qui se referme à un claquement, qu'à attribuer un sens à celui-ci, *e.g.* interpréter une phrase dans une vocalisation humaine. La localisation auditive, quant à elle, consiste à situer l'événement dans l'espace par rapport à l'auditeur.

La localisation auditive est une tâche qui peut être vue comme la résolution du problème suivant : identifier la position des sources dans l'environnement à partir de deux points d'observation du champ de pression acoustique, nos tympans [Van Opstal, 2016]. Dans le cas général, une scène sonore complexe est composée d'un nombre Q de sources sonores. Le système auditif a alors accès

à deux observations, notées $x_{\mathrm{L}}(t)$ et $x_{\mathrm{R}}(t)$, modélisées comme des combinaisons linéaires des sources sonores :

$$\begin{cases} x_{\mathrm{L}}(t) = \sum_{q=1}^{Q} (h_{\mathrm{L},\theta_q,\phi_q} \star s_q)(t) \\ x_{\mathrm{R}}(t) = \sum_{q=1}^{Q} (h_{\mathrm{R},\theta_q,\phi_q} \star s_q)(t) \end{cases}, \tag{1.1}$$

où $s_q(t)$ est le signal de la $q^{\text{ème}}$ source dans le domaine temporel, t l'indice temporel, $h_{\mathrm{L},\theta_q,\phi_q}(t)$ et $h_{\mathrm{R},\theta_q,\phi_q}(t)$ les réponses impulsionnelles des canaux acoustiques entre la $q^{\text{ème}}$ source située dans la direction d'arrivée (de l'anglais *Direction Of Arrival*, DOA) d'angles azimut et élévation (θ_q, ϕ_q) et les oreilles gauche et droite, respectivement ; enfin, $\star$ réfère au produit de convolution. On fait l'hypothèse que trouver la DOA consiste à identifier le couple $\{h_{\mathrm{L},\theta_q,\phi_q}(t), h_{\mathrm{R},\theta_q,\phi_q}(t)\}$ correspondant dont le système auditif a appris dans le passé une certaine représentation [Zakarauskas and Cynader, 1993, Baumgartner et al., 2014]. On a alors un système à deux équations et Q inconnues, le plus souvent supérieur à 2. On dit alors que le problème est « *mal-posé* », dans le sens où il n'existe pas de solution unique pour les $\{h_{\mathrm{L},\theta_q,\phi_q}(t), h_{\mathrm{R},\theta_q,\phi_q}(t)\}$. Pour résoudre ce problème, le système auditif se sert d'*a priori* et de l'information apportée par d'autres sens, comme la vue.

Dans cette section, nous allons tout d'abord introduire les systèmes de coordonnées utilisés dans le cadre de l'écoute spatiale. Puis, nous détaillons les indices acoustiques, présents dans le son au niveau de nos tympans, utilisés par le système auditif pour accomplir la localisation. Enfin, nous décrivons les critères objectifs visant à comparer ces derniers. Cela nous sera utile pour évaluer leur préservation au sein d'une chaîne de traitement dans les prothèses auditives.

1.2.1 Système de coordonnées

La localisation auditive se fait par rapport à la tête de l'auditeur, *i.e.* dans un repère égocentré. Selon les auteurs, l'origine du repère est définie comme le centre de la tête [Algazi et al., 2001a] ou comme le milieu de l'axe interaural [Bahu, 2016], *i.e.* le segment entre l'entrée des deux canaux auditifs.

En général, le système de coordonnées sphériques est préféré au système cartésien car il correspond plus à notre perception. Il est constitué de trois dimension :

— l'angle d'azimut, noté $\theta \in [0; 360°]$, avec $0°$ la direction frontale ;
— l'angle d'élévation, noté $\phi \in [-90; 90°]$, avec $0°$ le plan horizontal ;
— la distance au centre.

Dans la suite, on se concentre sur la localisation de la direction d'arrivée d'un son et non sur sa distance. Sauf mention contraire, on considérera que l'on se place à 1,5 m de l'auditeur, *i.e.* la distance typique entre deux personnes dans une conversation.

Un autre système de coordonnées sphériques est parfois employé permettant de faire une identification plus direct entre chaque dimension du système et les indices de localisation dont nous parlerons dans la suite [Macpherson and Middlebrooks, 2002]. Contrairement au système de coordonnées sphériques présenté précédemment dont les pôles sont sur l'axe vertical, le système de coordonnées dit *latéral-polaire* ou *horizontal polaire* a ses pôles sur l'axe interaural.[8] Les dimensions sont alors :

— l'angle latéral, noté $\Phi \in [-90; 90°]$;
— l'angle polaire, noté $\Theta \in [-90; 270°]$ avec $-90°$ la direction visant le sol ;
— la distance au centre.

1.2.2 Indices de localisation

Le système auditif extrait de l'information des signaux observés au niveau des tympans de sorte à estimer la direction d'arrivée d'un son. Ces indices sont de plusieurs natures, on peut les classer de la sorte :

— ceux liés à la morphologie de l'auditeur (endogènes) :
  — binauraux : extrait à partir des signaux des deux oreilles ;
  — monauraux : extrait à partir du signal de chaque oreille indépendamment ;
— ceux liés aux *a priori*sur la scène sonore (exogènes) :
  — nature des sources sonores, *e.g.* niveau sonore, contenu spectral ;
  — caractéristique spatiale, *e.g.* réverbération.

Les indices exogènes informent principalement sur la distance de la source. Comme nous nous concentrons sur l'estimation de la direction d'arrivée, lorsque nous parlerons d'indices de localisations, nous référerons aux indices endogènes.

Les indices de localisation sont usuellement catégorisés selon trois types : l'ITD due à la différence de temps de propagation entre les oreilles ; l'ILD due principalement au masquage de la tête ; et les indices spectraux dus à la diffraction sur le pavillon et le torse.

Différence de temps interaurale (ITD) L'ITD est maximale pour une source placée à $\pm 90°$ d'azimut sur le plan horizontal, elle excède rarement les 700 μs. La jnd est d'environ 10-20 μs pour les faibles valeurs d'ITD (sur le plan sagittal) et peut aller jusqu'à 60 μs pour les valeurs les plus grandes [Akeroyd,

[8] La droite passant par l'entrée des deux canaux auditifs.

2014]. La jnd varie aussi en fonction du type de source (plus ou moins de transitoire) ainsi que de la présence de réverbération [Klockgether and van de Par, 2016]. En utilisant un modèle simpliste dans lequel les tympans sont modélisés comme deux points et en négligeant tout obstacle entre, on peut montrer que l'ensemble des points pour lesquels l'ITD est la même forme une surface hyperbolique symétrique par rotation par rapport à l'axe interaural [Blauert, 1969]. Cette hyperbole est plus ou moins resserrée en fonction de l'ITD. Si on se base uniquement sur ce critère pour localiser une source, il existe donc un très grand nombre de solutions possibles. Cette surface hyperbolique peut être approximée par un cône pour une source distante de plus d'1 m. Pour cette raison, l'ensemble des positions de source possibles compatible avec l'ITD observé est appelé le « cône de confusion ». De plus, à partir de 1,5 kHz, le temps de propagation est trop grand pour permettre une discrimination basée sur la phase (ITD) [Kuhn, 1977, Macpherson and Middlebrooks, 2002, Aaronson and Hartmann, 2014]. On comprend donc qu'il est nécessaire de s'appuyer sur d'autres critères pour lever l'ambiguïté et affiner l'estimation de DOA.

Différence d'intensité interaurale (ILD) Il se trouve que c'est à partir de 1,5 kHz que la tête devient assez grande par rapport à la longueur d'onde associée pour devenir un obstacle acoustique et entrainer ainsi une différence d'intensité entre les deux oreilles en fonction de la direction d'arrivée [Duda and Martens, 1998]. Celle-ci renforce la latéralisation mais n'est pas cruciale, la localisation latérale étant principalement pilotée par l'ITD. Néanmoins, pour des stimuli dont le contenu spectral se situe principalement au dessus de 1,5 kHz, l'ILD vient jouer un rôle majeur dans la localisation latérale étant donné qu'une discrimination basée sur l'ITD est impossible dans cette zone du spectre [Macpherson and Middlebrooks, 2002]. L'ILD due au masquage de la tête varie fortement avec la distance en-deçà d'un mètre. En conséquence, l'association de l'ITD et de l'ILD permet de passer d'un cône de confusion à un « tore de confusion » [Shinn-Cunningham et al., 2000] restreignant ainsi l'ensemble des positions de source possible permettant d'expliquer l'observation d'un couple ITD-ILD.

Indices spectraux (SC) Avant d'atteindre le tympan, une onde acoustique rencontre un certain nombre d'obstacles, comme le torse, la tête et le pavillon de l'oreille. A leur contact, elle est tantôt diffractée ou réfléchie. Arrivée au tympan, qui s'apparente à un point de mesure, l'onde sonore a vue son contenu spectral transformé par ces phénomènes. En particulier, les réflexions dans le pavillon créent des phénomènes d'interférences tantôt destructives, tantôt constructives qui se caractérisent dans la réponse en fréquence par des pics et des creux très saillants [Spagnol et al., 2010] comme on peut l'observer

en Fig. 1.2. Leur présence ainsi que leur emplacement sont très dépendants de la direction, c'est pourquoi le système auditif s'appuie sur eux pour lever l'ambiguïté sur le cône/tore de confusion. Les indices spectraux (de l'anglais *Spectral Cues*, SCs) sont présent à partir de 700 Hz pour ce qui est de la contribution du torse pour une direction ipsilatéral[9] et de la diffraction de la tête pour une direction contralatérale[10] [Algazi et al., 2002] et à partir de 3 kHz pour la contribution du pavillon [Batteau, 1967]. Les SCs situés au-delà de 3 kHz permettent principalement de résoudre les confusion avant-arrière tandis que ceux situés en dessous peuvent aider à résoudre les confusions haut-bas [Algazi et al., 2001b].

Il a été montré qu'il est important de considérer une bande-passante allant jusqu'à 16 kHz pour bénéficier pleinement des SCs dans la résolution des confusions avant-arrière [Langendijk and Bronkhorst, 2002, Denk et al., 2019]. De plus, une résolution fréquentielle logarithmique de l'ordre de la demie octave est suffisante pour préserver les performances de localisation [Langendijk and Bronkhorst, 2002].

Modélisation (HRTF) L'ensemble des indices de localisation endogènes peuvent être modélisés comme une paire de filtres linéaires entre la source de l'onde sonore et l'entrée du canal auditif [Møller, 1992]. Un filtre est appelé la fonction de transfert liée à la tête (de l'anglais *Head Related Transfer Function*, HRTF) et il en existe une paire différente pour chaque position dans l'espace par rapport à l'auditeur. Son pendant temporel s'appelle la réponse impulsionnelle liée à la tête (de l'anglais *Head Related Impulse Response*, HRIR), celle-ci a une durée de l'ordre de 2 ms comme il peut être observé en Fig. 1.2. En Fig. 1.3 est représentée l'amplitude des HRTFs sur le plan horizontal pour l'oreille gauche. On observe clairement le masquage dû à la tête pour les azimuts négatifs, *i.e.* pour une source provenant de la droite de l'auditeur.

Au-delà d'une distance située entre un et deux mètres, les indices de localisation endogènes ne permettent pas de rendre compte de la distance [Shinn-Cunningham et al., 2000]. On peut alors distinguer deux zones selon l'axe des distances : (i) la zone proximale pour laquelle les indices de localisation sont dépendants de la distance ; et (ii) la zone distante, pour laquelle ils sont indépendants de la distance. La frontière entre ces deux zones est parfois nommé l'*horizon auditif* [Blauert, 2013]. Pour estimer la distance au-delà, le système auditif se base sur les indices exogènes comme le niveau sonore *a priori* [Zahorik, 2005] ou le rapport entre signal direct et réverbéré [Zahorik, 2002].

[9] Du même coté que l'oreille.
[10] Du coté opposé à l'oreille.

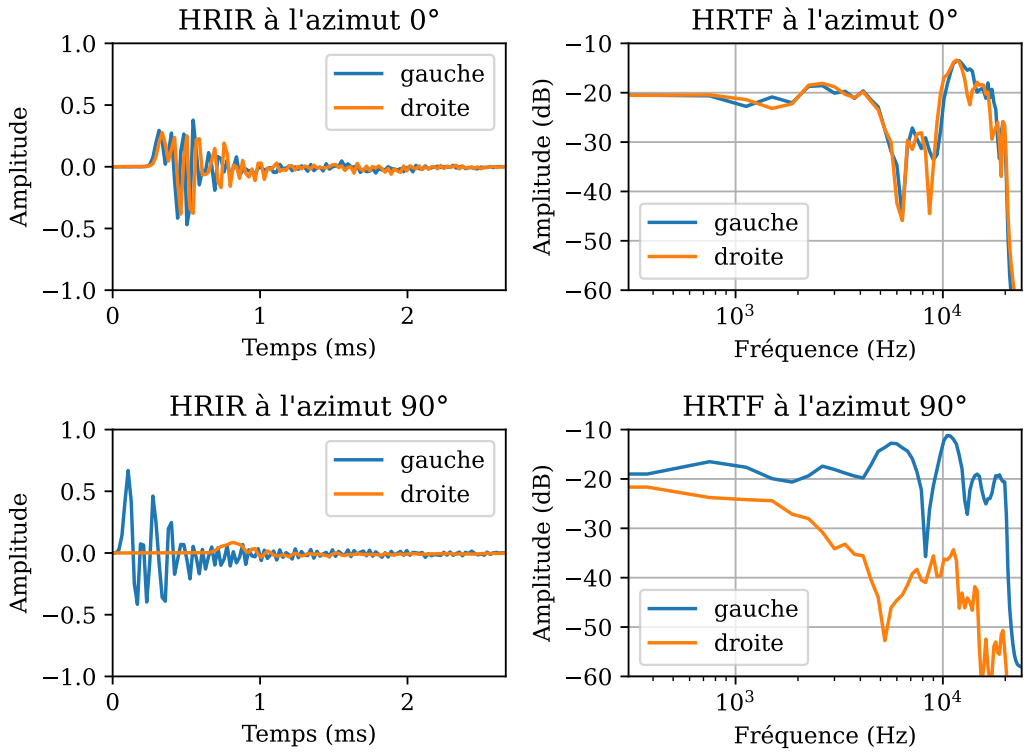


FIGURE 1.2 – Exemple d'HRIR et d'HRTF pour les azimut 0° et 90° à l'élévation 0°.

1.2.3 Externalisation

L'externalisation désigne la perception d'une source sonore comme provenant de l'extérieur de la tête de l'auditeur. A l'inverse, lorsqu'un son est perçu comme provenant de l'intérieur de la tête, on parle d'internalisation. Les indices interauraux permettent de créer une perception de latéralisation, *i.e.* de déplacement le long de l'axe interaural, mais ne permettent ni de résoudre l'ambiguïté sur le cône de confusion ni de créer un sentiment d'externalisation [Durlach et al., 1992].

Ce sont principalement les SCs qui permettent l'externalisation [Hartmann and Wittenberg, 1996]. La réduction de la résolution spectrale affecte l'externalisation pour un lissage du spectre par un banc de filtres d'une largeur supérieure à une largeur de bande rectangulaire équivalente (de l'anglais *Equivalent Rectangular Bandwidth*, ERB) en milieu faiblement réverbérant (temps de réverbération à 60 dB (TR60) de 0,16 s) [Hassager et al., 2016]. Il a aussi été montré que des indices de localisation dynamiques dus à un mouvement de la

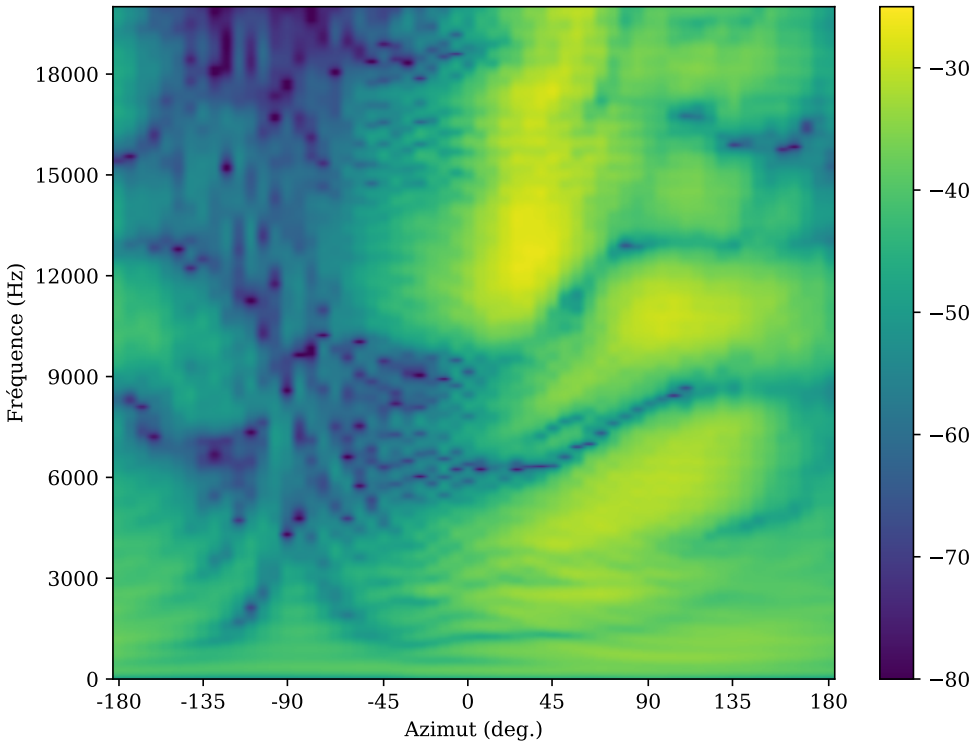


FIGURE 1.3 – Exemple d'amplitude d'HRTF (en dB) sur le plan horizontal pour l'oreille gauche.

tête augmentent la sensation d'externalisation, en particulier pour des sources situées sur le plan médian [Hendrickx et al., 2017].

[Boyd et al., 2012] ont montré que l'externalisation chez les normoentendant•e•s augmente grâce à une bande-passante supérieure à 6,5 kHz et les indices de localisation contenus dans l'HRTF. Pour les malentendants appareillés, l'ajout de bande-passante de 6,5 à 15 kHz n'améliore pas leur externalisation mais l'ajout des HRTFs l'améliore lorsque le microphone de la prothèse auditive est placé dans le pavillon de l'oreille (et ne l'améliore pas lorsque celui-ci est placé au-dessus de l'oreille) suggérant que les indices spectraux dus au pavillon sont important même s'ils n'accèdent pas à toute la bande-passante habituelle. Ils ont montré aussi que globalement, les malentendant•e•s perçoivent les sons toujours un peu au même niveau d'externalisation, jamais complètement internalisés ni complètement externalisés, à l'inverse des normoentendant•e•s.

1.2.4 Réapprentissage des indices de localisation

Le modèle qui fait consensus sur la manière dont le système auditif associe les indices de localisation acoustiques à une position dans l'espace consiste à le décrire comme un dispositif décisionnel se basant sur une carte de similarité entre les indices de localisation perçus et ceux appris [Hofman et al., 1998, Langendijk and Bronkhorst, 2002, Baumgartner et al., 2014]. Ce modèle nécessite donc une phase d'apprentissage pour associer les indices de localisation à une position dans l'espace basée sur des percepts multisensoriels (visuels, auditifs, proprioceptifs).

Une implication directe de cette hypothèse, presqu'un corollaire, est de penser que cette capacité d'apprentissage persiste dans le temps et que la cartographie mentale de l'espace sonore peut être réapprise si celle-ci venait à changer. En réalité, celle-ci est amenée à évoluer tout au long de la croissance du corps humain, pendant près de vingt ans donc. Elle est aussi capable de s'adapter à un changement brusque dans la morphologie de l'auditeur [Hofman et al., 1998] ou à l'arrivée progressive d'une perte auditive [Keating and King, 2013]. Après une modification de la forme du pavillon par exemple [Hofman et al., 1998], les performances de localisations redeviennent normales après quelques semaines, sans protocole de rééducation. Il est remarquable que la nouvelle carte auditive ne vient pas remplacer l'ancienne, mais s'ajouter. En effet, si le pavillon retrouve sa forme initiale, les performances de localisation sont équivalentes à celles d'origine.

Cette capacité est très importante car l'appareillage d'un patient revient à modifier les indices de localisation auxquels il aura accès [Majdak et al., 2013, Van den Bogaert et al., 2011, Denk et al., 2018]. Pour une large revue sur la plasticité cérébrale dans le réapprentissage des indices de localisation, la•e lecteur•rice intéressé•e pourra se référer à [Mendonça, 2014].

1.2.5 Critères d'évaluation objectifs

Nous avons vu qu'il existe une littérature importante traitant du rôle des indices acoustiques dans la localisation auditive. Par ailleurs, nous avons aussi vu que la préservation de ces indices est importante pour assurer une écoute de qualité aux malentendants appareillés mais que les prothèses en elles-même les dégradent. Nous verrons dans la suite que les traitements utilisés communément dans les prothèses auditives dégradent encore plus ces indices acoustiques. Pour cette raison, nous devons nous munir de critères objectifs permettant d'évaluer la préservation des indices de localisation de sorte à pouvoir concevoir des prothèses auditives qui les préservent tout au long de la chaîne de traitement.

Pour évaluer la préservation des indices interauraux, il suffit d'en calculer

la différence absolue [Marquardt, 2015]. En revanche, la quantification de la préservation des SCs ainsi que des indices de localisation globaux dans une scène sonore complexe sont plus compliqués.

La différence spectrale inter-sujet (ISSD)

La différence spectrale inter-sujet (de l'anglais *Inter-Subject Spectral Difference*, ISSD) vise à quantifier la distance entre deux réponses en fréquence indépendamment du niveau global [Middlebrooks, 1999a]. Elle est définie de la manière suivante :

$$\mathrm{ISSD} = \frac{1}{D}\sum_{d=1}^{D} \mathrm{var}_b\left(\tilde{H}_1(b,d) - \tilde{H}_2(b,d)\right), \tag{1.2}$$

où $\tilde{H}_1(b,d)$ et $\tilde{H}_2(b,d)$ désignent deux spectres d'amplitude pour lesquels l'axe fréquentiel a été rééchantillonné selon une échelle en $1/12^{\text{ème}}$ d'octave ; les indices d et b réfèrent à la direction et à la bande fréquentielle, respectivement ; et $\mathrm{var}_b(.)$ désigne la variance empirique selon l'axe fréquentiel.

La cohérence interaurale

La cohérence interaurale est définie comme la valeur absolue de la corrélation croisée normalisée entre les signaux des oreilles droite et gauche passés au travers d'un banc de filtre, $x_\mathrm{R}(k,n)$ et $x_\mathrm{L}(k,n)$:

$$\mathrm{IC}_{\max}(k) = \max_{\nu}\left|\frac{\sum_n x_\mathrm{L}(k,n+\nu)\; x_\mathrm{R}(k,n)^*}{\sqrt{\sum_n |x_\mathrm{L}(k,n)|^2}\sqrt{\sum_n |x_\mathrm{R}(k,n)|^2}}\right|. \tag{1.3}$$

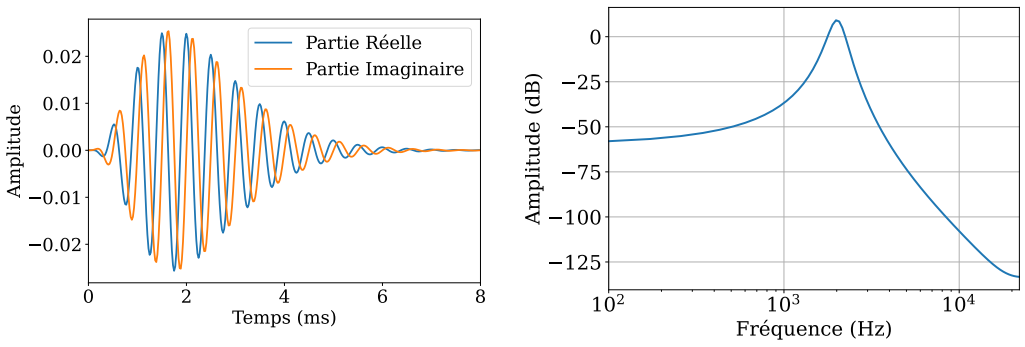
Le banc de filtre utilisé en pratique peut être composé de filtres passe-bande espacés d'un tiers d'octave [Hartmann et al., 2005] ou de filtres Gammatone complexe du $4^{\text{ème}}$ ordre [Hassager et al., 2017b] avec un espacement de une ERB, modélisant de manière plus fidèle la membrane basilaire [Glasberg and Moore, 1990] :

$$g_{f_c,\beta}(t) = \underbrace{t^{n-1}e^{-2\pi\beta t}}_{\text{enveloppe}} e^{j2\pi f_c t}, \tag{1.4}$$

avec f_c la fréquence centrale du filtre, n l'ordre du filtre et β le facteur de qualité [Hassager et al., 2016] :

$$\beta = 1,149 \;\times\; 24,7(4,37.10^{-3}\; f_c + 1). \tag{1.5}$$

La réponse impulsionnelle et l'amplitude de la fonction de transfert d'un tel filtre sont illustrées en Fig. 1.4.



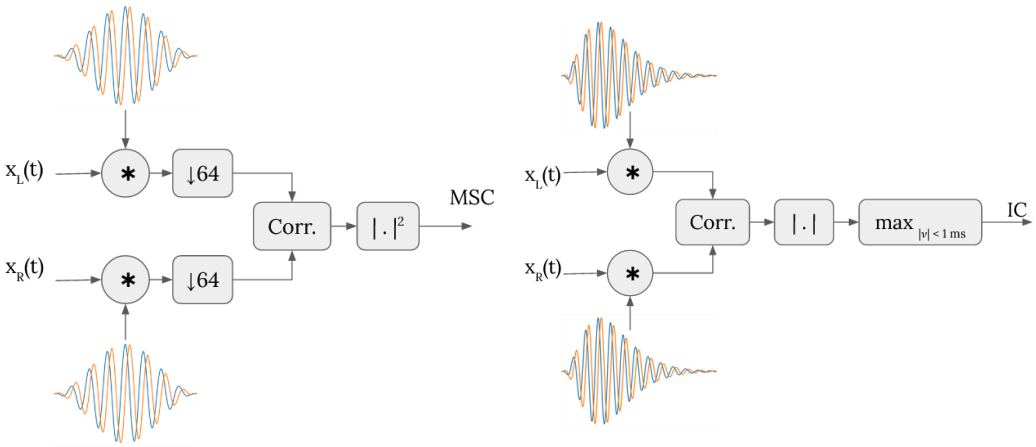
(a) (b)

FIGURE 1.4 – Filtre Gammatone complexe du 4ème ordre centré sur la fréquence 2 kHz dans le domaine temporel (a) et son amplitude en dB dans le domaine fréquentiel (b).

Cette métrique s'est montrée pertinente pour évaluer la préservation des indices de localisation dans un milieu réverbérant [Hassager et al., 2017b]. Pour cette raison, ce critère a été utilisé comme terme de coût dans la procédure d'optimisation permettant de concevoir les algorithmes de débruitage préservant les indices de localisation d'une scène sonore complexe [Marquardt, 2015, Itturriet and Costa, 2019]. Pour des raisons d'analyse et de facilité d'implémentation, la corrélation croisée est appliquée aux signaux dans le domaine de la transformée de Fourier à court terme (TFCT), ce qui équivaut à un banc de filtre. La TFCT et son utilisation dans les prothèses auditives sera définie plus en détail en section 1.3. Afin de faciliter la manipulation du terme relatif à la cohérence interaurale dans la procédure d'optimisation, il est préférable de considérer le carré de celle-ci, nommé le carré de l'amplitude de la cohérence interaurale (de l'anglais *Magnitude Squared Coherence*, MSC). En considérant que $x_{\mathrm{L}}(k,\ell)$ et $x_{\mathrm{R}}(k,\ell)$ sont deux variables aléatoires distribuées selon une loi Gausienne complexe centrée circulaire, on peut utiliser la définition statistique de la corrélation croisée :

$$\mathrm{MSC}(k) = \left| \frac{\mathbb{E}\left[x_{\mathrm{L}}(k,\ell) x_{\mathrm{R}}(k,\ell)^*\right]}{\sqrt{\mathbb{E}\left[|x_{\mathrm{L}}(k,\ell)|^2\right]}\ \sqrt{\mathbb{E}\left[|x_{\mathrm{R}}(k,\ell)|^2\right]}} \right|^2 . \tag{1.6}$$

Cela revient à considérer que le banc de filtre est constitué de filtres passe-bande, linéairement espacés, avec pour réponse impulsionnelle une exponentielle complexe multipliée par une fenêtre (de Hann, généralement) d'une longueur approximative de 10 ms. Le temps entre deux trames peut être interprété comme un sous-échantillonnage d'un rapport d'environ 64 effectué en sortie du



(a) (b)

FIGURE 1.5 – Schéma bloc comparant le calcul de la MSC (gauche) et l'IC_{max} (droite).

banc de filtre. En Fig. 1.5 sont représentés sous forme de schéma-bloc les calculs de l'IC_{max} et de la MSC de sorte à illustrer leur similarités et leur différences.

1.3 Chaîne de traitement du signal

Dans cette section, on décrit l'organisation générale des traitements du son au sein des prothèses auditives. Nous détaillerons dans les sections 2.1 et 2.2 le fonctionnement des algorithmes accomplissant la compensation des pertes auditives et la réduction de bruit.

Afin de bien comprendre la suite, nous rappelons que la conception des audioprothèses est fortement contrainte par des aspects techniques comme la miniaturisation, l'autonomie énergétique et la latence entrée-sortie [Gerlach et al., 2021]. Ces contraintes peuvent entrer en interaction, *e.g.* le choix de la taille de la batterie joue sur les contraintes à la fois de miniaturisation et d'autonomie énergétique.

Une des contraintes majeures dans la conception d'algorithmes pour les prothèses auditives est la latence tolérable par l'auditeur. Celle-ci est de maximum 9 ms, au-delà, la désynchronisation labiale[11] ainsi que le décalage entre la production et la réception de la propre voix de l'auditeur devient trop gênant [Stone and Moore, 2003]. D'autres travaux [Denk et al., 2019] suggèrent

[11] Le décalage temporel entre la parole perçue et le mouvement des lèvres d'un interlocuteur.

qu'un délai supérieur à 6 ms peut entrainer une diminution des performances de localisation sur le plan horizontal dû au mélange avec le signal direct par l'évent [12] dans les basses fréquences. En effet, il faut garder à l'esprit que l'auditeur perçoit toujours une partie du son direct et pas uniquement celui issu du haut-parleur de la prothèse auditive.

1.3.1 Acquisition des signaux

Les algorithmes de traitement du son employés dans les audioprothèses modernes nécessitent de numériser les signaux des microphones. Les circuits électroniques numériques consomment principalement lors des fronts d'horloge. On peut alors dire en première approximation que doubler la fréquence d'échantillonnage du signal double la consommation énergique de l'appareil. Or, la parole voisée [13] contient son énergie entre 100 et 4 kHz et la parole non-voisée jusqu'à 8 kHz environ. L'octave de 4 à 8 kHz est notamment importante pour différencier les phonèmes /f/ et /ʃ/ [Stelmachowicz et al., 2001]. En réalité, la parole contient de l'énergie au moins jusqu'à 16 kHz mais cette partie du signal n'est pas importante pour la compréhension du message. En revanche, il a été montré que des indices de localisation spectraux sont utilisés par le système auditif jusqu'à 16 kHz et que la perte de l'octave la plus haute (8 à 16 kHz) augmente, notamment, le nombre de confusion avant-arrière [Langendijk and Bronkhorst, 2002]. Le choix de la fréquence maximale à restituer par la prothèse auditive relève donc d'un compromis entre préservation de l'information contenue dans le signal audio et consommation d'énergie. En général, même dans les dispositifs modernes, l'octave 8-16 kHz est sacrifiée. De sorte à respecter la limite du théorème de Nyquist-Shannon, la fréquence d'échantillonnage est en général de 16 kHz pour une fréquence maximale restituée de 8 kHz.

1.3.2 Organisation des traitements

La plupart des traitements standards employés dans les prothèses auditives s'appliquent au signal dans le domaine fréquentiel. Pour cette raison, la transformation de Fourier discrète (TFD) est appliquée aux signaux des microphones juste après leur numérisation. Ceci équivaut à passer les signaux au travers d'un banc de filtre.

Les algorithmes de compensation des pertes auditives forment l'étage crucial des prothèses. Ceux-ci peuvent être de deux types : un compresseur de dynamique qui compense la perte de sensibilité aux sons de faible intensité et

[12] Au niveau du haut-parleur dans le canal auditif, conduit permettant la circulation de l'air.

[13] Faisant intervenir les cordes vocales.

le compresseur fréquentielle qui va faire en sorte de ramener dans la bande-passante audible par le malentendant une portion du spectre qui ne lui est plus accessible. Ces traitements permettent d'améliorer l'intelligibilité de la voix mais leur performance décroit en présence de bruit [Rhebergen et al., 2009]. Pour cela, un étage de réduction de bruit est généralement ajouté en amont.

Par ailleurs, le compresseur de dynamique peut appliquer un gain très important au signal (jusqu'à 60 dB en haute fréquence) et le haut-parleur est très proche des microphones (quelques centimètres). Ceci entraîne un risque d'effet Larsen très important, *i.e.* une instabilité dû à un bouclage du système et à un très fort gain dans la boucle de contre-réaction. Une manière de limiter ce risque et d'isoler au maximum les microphones du haut-parleur en bloquant au mieux le canal auditif avec un bouchon sur-mesure. Cependant, un bouchage hermétique du canal auditif n'est pas envisageable car les tissus doivent pouvoir respirer. Il existe donc une fuite de son entre l'extérieur et l'intérieur du canal auditif, notamment en basses fréquences. Afin de limiter ce risque, potentiellement très dangereux pour l'auditeur, un algorithme peut être ajouté au début de chaîne afin de détecter et interrompre un début d'effet Larsen. En Fig. 1.6 est illustré la chaîne de traitement décrite ci-dessus.

Dans la suite de ce document, on suppose que l'on se trouve dans un régime de fonctionnement stable, *i.e.* absence d'effet Larsen, on néglige alors la contribution de l'algorithme de réduction d'effet Larsen. De plus, on va considérer uniquement le compresseur de dynamique dans la suite. Celui-ci est le seul à être tout le temps présent, le compresseur fréquentiel étant réservé à des pertes particulières pour lesquelles une zone du spectre est totalement inaccessible.

Par ailleurs, les progrès récents permettent de faire communiquer sans-fil les prothèses auditives droite et gauche lorsque l'auditeur est appareillé des deux cotés. Cela permet de considérer un réseau de microphones deux fois plus important et surtout de bénéficier d'une diversité spatiale bien plus grande. En effet, l'espacement entre les microphones au sein d'une prothèses auditive est de l'ordre de quelques centimètres alors que l'espacement entre les deux prothèses est en moyenne de 16 cm. Cette diversité spatiale accrue permet aux algorithmes de réduction de bruit multicanaux d'être plus efficaces, notamment en basses fréquences. Pour ces raisons, nous considérons dans ce travail un système de prothèses auditives binaurales faisant l'hypothèse d'une parfaite communication (sans perte et sans délai) entre les oreilles droite et gauche, comme illustré en Fig. 1.7.

1.3.3 Chaîne de traitement temps-réel

On verra par la suite qu'il est commode de travailler dans le domaine temps-fréquence. Pour cela, il est nécessaire d'acquérir le signal sur une certaine durée

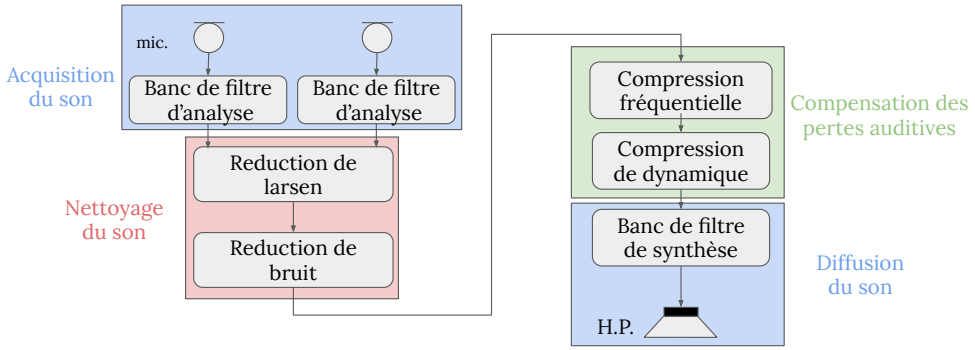


FIGURE 1.6 – Schéma bloc d'une chaîne de traitement du signal générique dans les prothèses auditives.

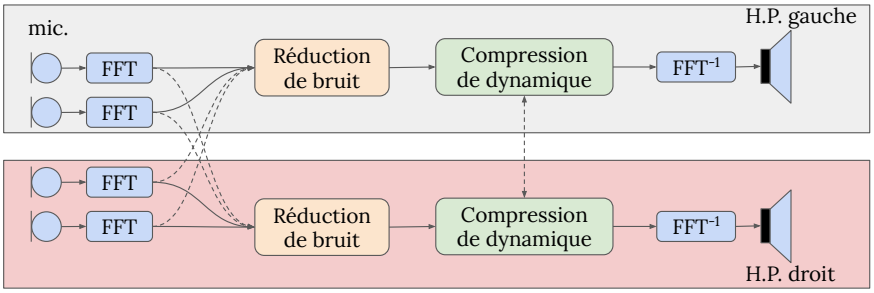


FIGURE 1.7 – Schéma bloc de principe d'une chaine de traitement de prothèses auditives binaurales (en gris l'appareil gauche, en rouge l'appareil droit). Les liens en pointillé désignent les signaux audio à transmettre d'un appareil à l'autre.

avant d'y appliquer la transformation de Fourier. Ce segment de son est appelée une trame. La durée de celle-ci doit être suffisamment longue de sorte à pouvoir acquérir une période entière de la fréquence la plus basse à représenter, *e.g.* de l'ordre de 100 Hz pour la voix, et suffisamment courte de sorte que la latence introduite ne soit pas rédhibitoire pour l'auditeur. Comme on l'a dit précédemment, la latence tolérable maximale est de l'ordre de 9 ms et une fréquence de 100 Hz correspond à une période de 10 ms. Il est important de rappeler que cette latence est incompressible et correspond au temps d'acquisition du son à traiter. Il n'inclue pas le temps de traitement de celui-ci.

Les traitements qui vont être appliqués à chaque trame sont effectués dans le domaine fréquentiel. Cela implique que rien n'assure la continuité du signal de sortie dans le domaine temporel. Pour cela, on introduit une période de recouvrement entre chaque trame ainsi qu'un fondu pour adoucir la discontinuité [Crochiere, 1980] comme illustré en Fig. 1.8. En général, chaque trame se

recouvre à moitié avec sa voisine précédente de sorte que l'on se retrouve toujours dans une phase de transition, on nomme cela un recouvrement à 50 %. Le fondu peut-être uniquement appliqué lors de la synthèse (après la transformation de Fourier inverse). Cependant, il peut être intéressant d'appliquer aussi un fondu lors de l'analyse (avant la transformation de Fourier). En effet, ne pas appliquer de fondu, ou fenêtrage, lors de l'analyse revient à convoluer le spectre par un sinus cardinal dont la taille du lobe principal est égale à l'inverse de la taille de la fenêtre. Cela conduit à un étalement spectral important. L'application d'une fenêtre bien choisie peut réduire cet étalement dans le domaine fréquentiel. En pratique, on choisit typiquement la fenêtre de Hann qui permet un recouvrement constant et donc une reconstruction parfaite (lorsqu'aucun traitement n'est appliqué dans la chaîne). Il est alors d'usage de prendre la racine carré de la fenêtre de Hann pour les fenêtres d'analyse et de synthèse de sorte à obtenir un recouvrement constant sur l'ensemble de la chaîne :

$$w(n) = \begin{cases} \sin\left(\frac{\pi n}{N}\right) & \forall n \in \{1, ..., N\} \\ 0 & \text{sinon,} \end{cases} \tag{1.7}$$

avec N la taille de la fenêtre. On applique alors la TFD à chaque trame acquise afin d'obtenir la représentation du signal dans le domaine temps-fréquence, notée $x(k, \ell)$:

$$x(k, \ell) = \frac{1}{N} \sum_{n=1}^{N} w(n)\ x\left(n + \ell \frac{N}{2}\right) e^{-j2\pi \frac{kn}{N}}, \tag{1.8}$$

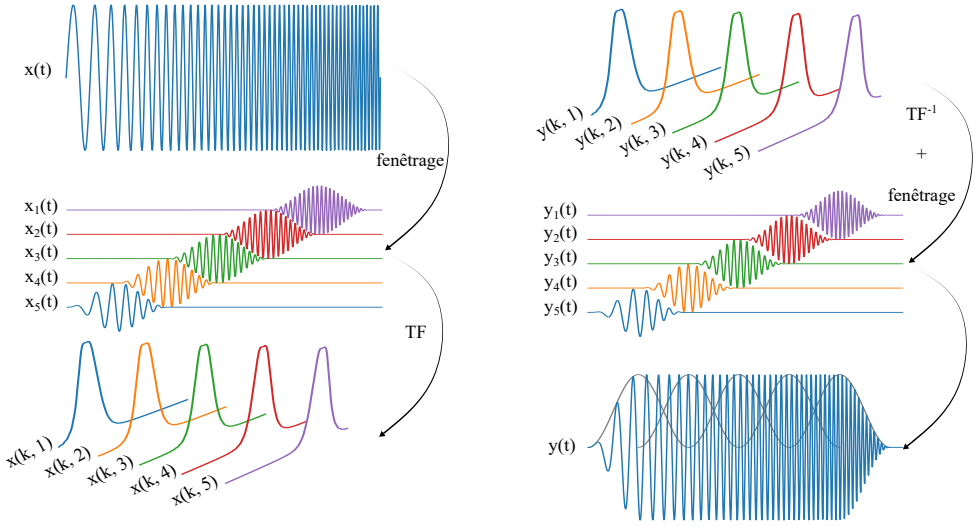
où k et ℓ réfèrent respectivement aux indices de fréquence et de trame temporelle et $x\left(n + \ell \frac{N}{2}\right)\ \forall n \in \{1, ..., N\}$ est la $\ell^{\text{ème}}$ trame temporelle. On peut alors retrouver le signal d'origine en appliquant la transformation de Fourier discrète inverse et en combinant les trames ℓ et $\ell - 1$ qui se recouvrent :

$$x\left(n + \ell \frac{N}{2}\right) = w(n) \sum_{k} x(k, \ell) e^{j2\pi \frac{kn}{N}} + w\left(n + \frac{N}{2}\right) \sum_{k} x(k, \ell - 1) e^{j2\pi \frac{kn}{N}}. \tag{1.9}$$

1.4 Structure du manuscrit et contributions

1.4.1 Plan et contributions

Dans le chapitre 2, nous présentons en détail les méthodes de débruitage et compensation du recrutement de la sonie proposées dans la littérature. En particulier, nous nous intéressons à celles visant à préserver les indices de localisation et plus généralement les caractéristiques acoustiques de la scène sonore.



(a) (b)

FIGURE 1.8 – Chaîne d'analyse (a) et de synthèse (b) de type « recouvrement constant » (COLA) appliquée à un signal sinusoïdal balayant le spectre.

Enfin, nous discutons de l'interaction entre ces deux étages de traitement du signal et des différentes manières de les associer.

En chapitre 3, nous proposons d'unifier les fonctions de débruitage et de compensation des pertes auditives au sein d'un même problème d'optimisation de sorte à proposer une association de ces deux fonctions basée sur une modélisation des signaux. La méthode proposée fait apparaitre une architecture parallèle où la source de parole cible et le reste de la scène sonore sont traitées séparément avant d'être recombinées.

Dans le chapitre 4, nous nous focalisons sur l'estimateur de bruit employé dans le chapitre précédent. Nous mettons en évidence qu'il distord fortement les indices interauraux de sources provenant de l'hémisphère frontale et nous proposons trois algorithmes permettant de répondre à ce problème.

Enfin, dans le chapitre 5, nous étendons notre modèle de scène sonore au cas multi-locuteur. Dans ce cas, il est très difficile de à la fois débruiter et préserver les sources de parole présentes dans la scène sonore. De plus, considérer plusieurs cibles fait augmenter significativement le coût calculatoire du filtre de débruitage. Nous proposons alors de prendre en compte les propriétés de parcimonie de la parole pour à la fois mieux débruiter sans trop dégrader les sources cibles et réduire le coût calculatoire du filtre.

Nous terminons ce manuscrit par le chapitre 6, fournissant une synthèse des travaux présentés et des pistes de réflexions inspirées par les résultats obtenus

dans cette thèse.

1.4.2 Publications associées à cette thèse

— Llave, A. and Séguier, R. (2019). Influence sur les indices de localisation du beamforming et de la compression de dynamique dans les audioprothèses. In *XXVIIème colloque GRETSI (GRETSI 2019)*, Lille, France

— Llave, A., Leglaive, S., and Seguier, R. (2020). Localization Cues Preservation in Hearing Aids by Combining Noise Reduction and Dynamic Range Compression. In *12th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, page 9, Auckland, New Zealand. IEEE

— Llave, A. and Leglaive, S. (2021b). On the Speech Sparsity for Computational Efficiency and Noise Reduction in Hearing Aids. In *13th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, page 5, Tokyo, Japan. IEEE

— Llave, A. and Leglaive, S. (2021a). Joint denoising and dynamic range compression in binaural hearing aids. *Soumis à The Journal of the Acoustical Society of America*, page 12 (**article soumis**)

Chapitre 2

État de l'art

Ce chapitre a pour but de présenter l'état des travaux sur la compensation des pertes auditives et du débruitage dans les prothèses auditives ainsi que la manière dont ces traitements ont été assemblés, voire co-conçus.

En section 2.1, nous présentons la manière classique de compenser les pertes auditives grâce à la compression de dynamique multi-bandes. Nous y discutons ses performances pour la compréhension de la parole et la localisation auditive dans différents scenarii défavorables. Nous introduisons alors les différentes tentatives d'amélioration répondant à ces situations.

Dans la section 2.2, nous présentons l'état de l'art des méthodes de débruitage dans la prothèses auditives. Tout d'abord, nous introduisons les méthodes classiques, ensuite nous détaillons celles spécifiques à notre application, permettant de préserver les indices de localisation de la cible et du bruit.

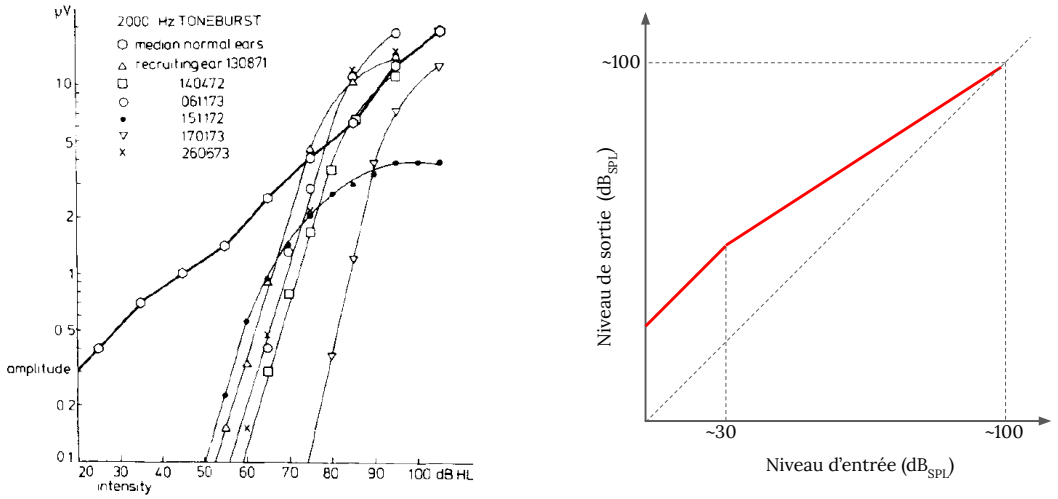
Les algorithmes de débruitage multicanaux nécessitent d'estimer la(les) réponse(s) du canal acoustique entre la(les) source(s) visée(s) et les microphones ainsi que les directions d'arrivées. Nous présentons donc en section 2.3 les méthodes d'estimation de ces premières en temps-réel et en amont de l'utilisation et en section 2.4, les méthodes d'estimation de ces dernières.

À partir de l'étude du compresseur de dynamique et des algorithmes de débruitage, nous discutons dans la section 2.5 leur interaction. Nous détaillons ensuite notre positionnement et les axes d'études que nous avons choisis et qui sont traités dans les chapitres suivants.

2.1 Compensation des pertes auditives

On cherche à compenser le recrutement de sonie présenté en section 1.1.1. Rappelons que ce phénomène consiste en une baisse du seuil d'audition sans diminution du seuil d'inconfort. La pente de la fonction entre niveau sonore perçu et niveau de pression acoustique devient alors plus forte (voir Fig. 2.1a). L'idée est alors de compenser l'excès de cette pente par un compresseur de dynamique (CD) comme illustré en Fig. 2.1b. Comme l'importance de la perte auditive dépend de la fréquence, un CD est appliqué indépendamment à chaque bande fréquentielle. Le nombre de bandes varie selon les auteurs et est généralement compris entre 2 et 16 [Wiggins and Seeber, 2011, Hassager et al., 2017b, Stone and Moore, 2008, Schwartz and Shinn-Cunningham, 2013] mais peut monter jusqu'à 30 dans les approches visant à modéliser plus précisément la membrane basilaire [Kortlang et al., 2016, Ernst et al., 2018].

Dans cette section, nous présentons tout d'abord le principe général du CD puis nous décrivons ses impacts sur les performances de localisation et de compréhension de la parole dans différents scenarii (avec ou sans réverbération, en présence de bruit ou dans un scénario multi-locuteur).



(a) (b)

FIGURE 2.1 – (a) Excitation électrique du nerf en fonction du niveau sonore de la médiane des sujets normoentendants et pour six sujets malentendants (extraite de [Eggermont, 1977]). Et (b) une courbe entrée-sortie typique d'un CD utilisé dans les prothèses auditives.

2.1.1 Principe général

Plaçons nous dans le domaine de la TFCT où k et ℓ sont respectivement les indices de fréquence et de trame temporelle. Le principe est d'appliquer un gain $g(k,\ell)$ au signal d'entrée $x(k,\ell)$:

$$y(k,\ell) = g(k,\ell)\ x(k,\ell). \tag{2.1}$$

Ce gain est calculé à partir de l'entrée $x(k,\ell)$. Pour ce faire, le spectre de l'entrée $x(k,\ell)$ est passé au travers d'un banc de filtre de B bandes et sa puissance instantanée, notée $P_b(\ell)$, est extraite :

$$P_b(\ell) = \frac{1}{K}\sum_{k=1}^{K} |w_b(k)\ x(k,\ell)|^2 \qquad \forall\ b\ \in \{1,...,B\}, \tag{2.2}$$

où $w_b(k) \in \mathbb{R}$ est la fonction de transfert de la $b^{\text{ème}}$ bande fréquentielle. L'enveloppe du signal d'entrée, notée $\tilde{P}_b(\ell)$, est calculée pour chaque bande comme étant la puissance instantanée lissée par un filtre passe-bas du premier ordre avec une constante de temps différente selon si le signal est dans une phase croissante ou décroissante (respectivement nommées attaque et relâche) :

$$\tilde{P}_b(\ell) = \alpha P_b(\ell) + (1-\alpha)\tilde{P}_b(\ell-1) \tag{2.3}$$

avec

$$\alpha = \begin{cases} \alpha_{\mathrm{A}} & \text{si } P_b(\ell) > \tilde{P}_b(\ell-1) \\ \alpha_{\mathrm{R}} & \text{sinon,} \end{cases} \tag{2.4}$$

où α_{A} et α_{R} sont les coefficients de lissage correspondant respectivement aux constantes de temps d'attaque et de relâche.

Le protocole de réglage de prothèses auditives [ANSI, 2003] définit la constante de temps comme la période de transition, en réponse à un échelon tension de 55 à 90 dB, entre le début de celui-ci et le moment où la sortie atteint un niveau 3 dB en dessous de la valeur cible. Cette définition varie sensiblement de celle communément admise en physique selon laquelle la constante de temps correspond à la durée pour atteindre 63 % de la valeur finale. Cependant, elle permet d'être plus globale dans la mesure où le filtre n'est pas toujours appliqué à la puissance de l'enveloppe, comme présenté ici, mais au gain de compression exprimé en dB par exemple [Ngo et al., 2012, Giannoulis et al., 2012]. Les caractéristiques de la parole varient rapidement, il est donc important d'avoir des temps d'attaque et de relâche assez courts pour suivre le niveau de la parole au niveau de chaque syllabe et, ainsi, en assurer l'audibilité. Il est préconisé d'avoir un temps de relâche inférieur à 60 ms mais il est important que ce temps soit supérieur à 30 ms de sorte à assurer correctement le lissage de l'enveloppe [Jerlvall and Lindblad, 1978, Kuk, 1996]. Le choix du temps d'attaque est moins critique et peut être de l'ordre de quelques millisecondes, 5 ms typiquement, afin d'être suffisamment réactif.

Ensuite, le gain de compression est calculé comme une fonction non-linéaire pour un niveau d'entrée dépassant le seuil T (affine dans le domaine des dB) et une amplification linéaire en-deçà :

$$G_b(\ell) = \begin{cases} K + \left(\tilde{P}_b^{\mathrm{dB}}(\ell) - T\right)\left(\frac{1}{R} - 1\right) & \text{si } \tilde{P}_b^{\mathrm{dB}}(\ell) > T \\ K & \text{sinon,} \end{cases} \tag{2.5}$$

où $\tilde{P}_b^{\mathrm{dB}}(\ell) = 10\,\log_{10}\tilde{P}_b(\ell)$ et R est l'inverse du coefficient de la droite (appelé ratio ou taux de compression), voir Fig. 2.1b pour un exemple. Le gain en dB est ensuite ramené dans le domaine linéaire $g_b(\ell) = 10^{\frac{G_b(\ell)}{20}}$ et enfin passé au travers du banc de filtre inverse de sorte à être exprimé en temps-fréquence pour être appliqué au signal d'entrée :

$$g(k,\ell) = \sum_b w_b(k)\; g_b(\ell). \tag{2.6}$$

Il faut noter que la manière d'agencer les différentes opérations permettant de calculer le gain de compression n'est pas unique et peu varier d'une application à une autre, ou d'un constructeur à un autre. En effet, c'est un

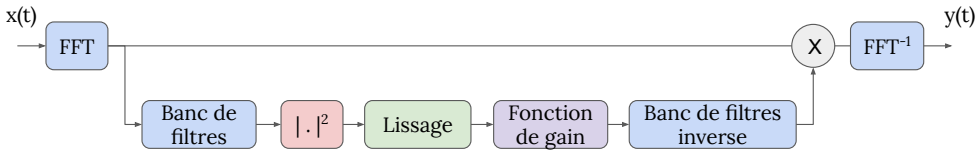


FIGURE 2.2 – Schéma général d'un CD multibande.

traitement qui ne répond pas à un problème bien posé mathématiquement et pour lequel une solution optimale et fondée pourrait se dégager. Il est possible par exemple de voir le lissage adaptatif (attaque et relâche) placé à différents endroits dans la chaîne, ou encore de se servir du signal de sortie pour calculer le gain à la place du signal d'entrée.[1] En Fig. 2.2 est représenté le schéma bloc du compresseur tel que décrit dans cette section et qui sera utilisé dans la suite en l'absence de précision. Pour une revue détaillée des CD utilisés dans les prothèses auditives, la•e lecteur•rice pourra se référer à [Kates, 2005] et pour une revue plus générale des CD, à [Giannoulis et al., 2012].

2.1.2 Compression de la parole en présence de bruit

Il a été bien montré que la compression de dynamique permet d'améliorer les performances de compréhension de la parole dans un environnement calme [Souza and Turner, 1998]. En revanche, en présence de bruit (stationnaire ou intermittent), celles-ci sont détériorées [Rhebergen et al., 2009]. Ceci peut s'expliquer par deux phénomènes : premièrement, le gain de compression est estimé à partir du signal d'entrée, *i.e.* le mélange entre le signal de parole et de bruit. Or, le RSB peut varier fortement, localement, dû à la dynamique naturelle de la parole. Cela a pour conséquence de sous-estimer le gain de compression lors des périodes de faible RSB et donc de moins amplifier les segments de parole de niveau faible en présence de bruit qu'en environnement calme [Souza et al., 2006, May et al., 2018]. Deuxièmement, le gain est appliqué au mélange de la parole et du bruit. Par conséquent, un bruit stationnaire peut voir son enveloppe être fortement modifiée par le gain appliqué au mélange et devenir corrélée au signal de parole [Stone and Moore, 2007] et inversement si le bruit est intermittent, *e.g.* une source de parole masquante. Il a été montré que cette comodulation entre les sources entraine des distorsions qui peuvent être délétères pour la compréhension de la parole [Naylor and Johannesson, 2009]. Un autre effet collatéral est la diminution du RSB à long terme en sortie, ceci a été montré empiriquement [Hagerman and Olofsson, 2004, Souza et al., 2006] et théoriquement pour n'importe quelle fonction de compression

[1] Ceci a pour conséquence de faire du compresseur un système bouclé, le rendant difficile à modéliser et possiblement instable.

concave [Corey, 2019, Chapitre 6]. Cet effet est dépendant des constantes de temps choisies pour le filtre de lissage d'enveloppe. En effet, plus les constantes de temps sont grandes, moins le compresseur a d'effet en général.

Pour répondre à ce problème, [May et al., 2018] ont proposé d'amplifier de façon moindre les périodes dominées par le bruit de sorte à améliorer le RSB en sortie et limiter la comodulation entre les deux composantes. Pour ce faire, ils proposent de prolonger le gain de compression calculé à partir du dernier segment identifié comme étant dominé par la parole en appliquant une constante de temps de relâche très grande sur les segments de signaux identifiés comme étant dominés par le bruit :

$$\alpha_{\mathrm{R}} = \begin{cases} \alpha_{\mathrm{R,\ fast}} & \text{si } \xi(k,\ell) > 1, \\ \alpha_{\mathrm{R,\ slow}} & \text{sinon,} \end{cases} \tag{2.7}$$

où $\xi(k,\ell)$ est le RSB à l'entrée du compresseur. Les valeurs de $\alpha_{\mathrm{R,\ fast}}$ et de $\alpha_{\mathrm{R,\ slow}}$ sont respectivement de 40 et 2000 ms. Cependant cette solution semble peu satisfaisante. En effet, ses bénéfices sur le RSB en sortie sont limités, notamment à faible RSB en entrée et le gain à appliquer aux périodes dominées par le bruit est hautement dépendant du contenu de la parole les précédant ainsi que de la fiabilité du classifieur permettant d'identifier si le segment est dominé ou pas par la parole. Selon ces deux facteurs, le gain appliqué à la période de bruit peut varier jusqu'à une dizaine de dB. Il est important de noter que la classification se fait indépendamment pour chaque bande fréquentielle, ceci a pour conséquence que d'une période de bruit à l'autre, le filtre appliqué peut-être très différent et donc apporter un changement de timbre important, sans aucune justification.

2.1.3 Influence sur les performances de localisation et de compréhension de la parole

Dans le cas d'un appareillage bilatéral, *i.e.* où les deux oreilles sont appareillées, un CD est appliqué indépendamment à chaque oreille. On définit alors $g_L(k,\ell)$ et $g_R(k,\ell)$, les gains des CD appliqués respectivement aux signaux destinés aux oreilles gauche et droite. Dans ce cas, rien n'assure que $g_L(k,\ell) = g_R(k,\ell)$, ce qui entraine une distorsion de l'ILD présenté à l'auditeur. De telles distorsions de l'ILD entrainent plus d'erreurs de localisation, d'internalisation, ainsi qu'une impression de dédoublement et/ou d'élargissement de l'image de la source sonore [Wiggins and Seeber, 2011, Wiggins and Seeber, 2012] en condition anéchoïque.

La manière intuitive de résoudre ce problème est de relier les deux CD, *i.e.*

d'appliquer le même gain $\bar{g}(k,\ell)$ aux signaux présentés aux deux oreilles :

$$y_L(k,\ell) = \bar{g}(k,\ell)\ x_L(k,\ell), \tag{2.8}$$
$$y_R(k,\ell) = \bar{g}(k,\ell)\ x_R(k,\ell), \tag{2.9}$$

où $x_L(k,\ell)$ et $x_R(k,\ell)$ sont les signaux d'entrée des CD respectivement gauche et droite. Afin de ne pas dépasser le niveau d'inconfort à l'une des deux oreilles, on prend le minimum des deux :

$$\bar{g}(k,\ell) = \min\{g_L(k,\ell), g_R(k,\ell)\}. \tag{2.10}$$

On parle alors de CD appairés (*linked compression*) et il devient nécessaire d'établir une communication entre les deux appareils, on parle alors de prothèses auditives binaurales.

L'efficacité de cette stratégie n'est pas clairement avérée. En condition anéchoïque, certains auteurs trouvent une amélioration des performances de localisation [Wiggins and Seeber, 2012] quand d'autres non [Korhonen et al., 2015]. L'influence sur la compréhension de la parole en environnement bruité n'est pas claire non plus avec certaines études montrant une amélioration [Wiggins and Seeber, 2013, Schwartz and Shinn-Cunningham, 2013] et d'autres non [Ibrahim et al., 2012]. En présence de réverbération, le CD appairé présente les même problèmes que le CD indépendant sur les performances de localisation [Hassager et al., 2017b] et n'améliore pas la compréhension de la parole [Ernst et al., 2018]. Une stratégie très similaire à celle de [May et al., 2018], présentée en Eq. (2.7), a été proposée par [Hassager et al., 2017a] de sorte à préserver le rapport signal direct à réverbéré (de l'anglais *Direct-to-Reverberation Ratio*, DRR) et l'IC, montrant des résultats convaincants dans un milieu réverbéré avec un temps de réverbération à 30 dB (TR_{30}) d'environ 500 ms, correspondant à un salon d'après les auteurs. Les caractéristiques d'une réverbération étant très complexes et non résumable à un temps de réverbération, ce résultat demande à être reproduit dans différents types d'environnements.

2.1.4 Métrique

Afin de quantifier le niveau de compression d'un signal, on peut utiliser l'ECR. Celui-ci consiste à mesurer le rapport entre la dynamique du signal en entrée et en sortie du compresseur. Il est d'usage de mesurer l'ECR par bande de fréquence puis de le moyenner selon l'axe fréquentielle :

$$\mathrm{ECR} = \frac{1}{B}\sum_{b=1}^{B}\frac{\mathcal{D}_x(b)}{\mathcal{D}_y(b)}, \tag{2.11}$$

où $\mathcal{D}_x(b)$ et $\mathcal{D}_y(b)$ sont respectivement les dynamiques des signaux en entrée et en sortie du compresseur, en dB, pour la $b^{\text{ème}}$ bande. La définition de la plage de dynamique varie selon les auteurs, tantôt définie comme l'écart-type de l'histogramme des niveaux en dB [Alexander and Rallapalli, 2017] ou comme la différence entre le $95^{\text{ème}}$ et le $5^{\text{ème}}$ centile [Souza et al., 2006, Corey and Singer, 2017, May et al., 2018] :

$$\mathcal{D}_x(b) = \sqrt{\mathbb{E}\left[|X(b,\ell) - \mathbb{E}\left[X(b,\ell)\right]|^2\right]}, \tag{2.12}$$

ou

$$\mathcal{D'}_x(b) = X^{(95)}(b) - X^{(5)}(b), \tag{2.13}$$

où $X^{(c)}(b)$ correspond à la valeur du $c^{\text{ème}}$ centile de l'histogramme des niveaux en dB du signal $x(b,\ell)$. La première définition fait l'hypothèse que la distribution des niveaux du signal est gaussienne, ce qui n'est pas le cas. Bien que la seconde ne fait pas ce genre d'hypothèse, les valeurs des centiles utilisés pour la mesure peuvent varier d'une étude à l'autre, rendant le critère paramétrique et donc la comparaison des scores difficile d'une étude à l'autre. De manière général, l'ECR est toujours inférieur au taux de compression R du fait du lissage de l'enveloppe du signal.

2.2 Réduction du bruit

Nous avons vu précédemment que les malentendants montrent des difficultés à comprendre la parole dans un environnement bruité et que les bénéfices de la compression de dynamique y sont aussi détériorés. Pour répondre à ce problème, la stratégie usuelle est d'ajouter un traitement de débruitage en amont de l'étage de compensation des pertes auditives.

Dans cette section, nous présentons les types d'algorithme de débruitage que l'on retrouve habituellement dans les audioprothèses. Ils peuvent fonctionner à partir d'un seul signal (monocanal) ou des signaux de plusieurs microphones (multicanal). Un algorithme monocanal repose sur l'hypothèse que le signal d'intérêt, *e.g.* parole ou musique, et le bruit ont des propriétés de stationnarités différentes tandis que les algorithmes multicanaux bénéficient de l'information spatiale apportée par la diversité des observations du champs de pression acoustique. Ces derniers ont pour conséquence d'amplifier différemment les sons selon leur direction d'arrivée, *i.e.* ils forment un faisceau en direction de la cible. Pour cette raison, on les appelle communément des algorithmes de *beamforming* ou beamformers. Ceux-ci parviennent à une importante amélioration du RSB en sortie et préservent les indices de localisation de la cible. Cependant, le bruit résiduel est aussi perçu comme venant de la direction cible. Cet

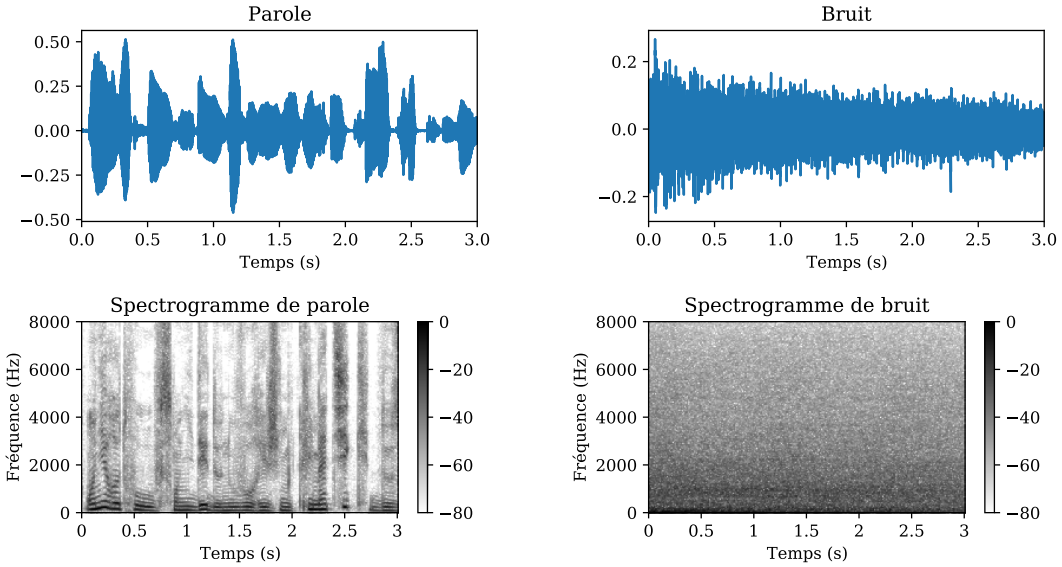


FIGURE 2.3 – Signaux de parole (gauche) et de bruit de rue (droite) dans le domaine temporel (haut) et sous forme de spectrogramme en dB (bas).

inconvénient a reçu beaucoup d'attention au cours de la dernière décennie et de nombreux algorithmes ont été présentés afin d'y répondre. Pour autant, ce sujet reste encore ouvert. La dernière partie de cette section décrit les beamformers préservant les indices de localisation du bruit et la Fig. 2.8 résume les liens entre les différents algorithmes de beamforming.

2.2.1 Algorithme monocanal

Dans cette sous-section, on présente la stratégie la plus courante pour réduire un bruit dit stationnaire. On entend par là un signal dont les caractéristiques évoluent peu au cours du temps. Cela exclue les bruits impulsifs, de type cliquetis ou craquement et considère plutôt le bruit de fond (bruit thermique, ventilation, etc.) ou la « rumeur » (circulation en ville, babillage, etc.). Pour différencier le signal d'intérêt et le bruit, l'algorithme se fonde sur l'hypothèse que le bruit est présent en tout point du plan temps-fréquence (T-F) alors que le signal de parole concentre son énergie sur un nombre limité de fréquence évoluant au cours du temps. Cette hypothèse est illustrée en Fig. 2.3.

Modèle des signaux

Le signal observé dans le domaine de la TFCT, noté $x(k,\ell)$, est modélisé comme étant la somme de la source de parole et du bruit :

$$x(k,\ell) = \ s(k,\ell) + n(k,\ell), \tag{2.14}$$

où k et ℓ sont, respectivement, les indices de fréquence et de trame temporelle ; $x(k,\ell)$, $s(k,\ell)$ et $n(k,\ell)$ sont des nombres complexes. On modélise $s(k,\ell)$ et $n(k,\ell)$ comme des variables aléatoires indépendantes suivant une distribution normale complexe isotropique centrée, notée $\mathcal{N}_{\mathbb{C}}(0,\phi)$, de variance ϕ :

$$s(k,\ell) \sim \mathcal{N}_{\mathbb{C}}(0, \phi_s(k,\ell)), \tag{2.15}$$

$$\text{et } n(k,\ell) \sim \mathcal{N}_{\mathbb{C}}(0, \phi_n(k,\ell)). \tag{2.16}$$

On fait aussi l'hypothèse que pour chaque point T-F, la parole est soit présente, soit absente et que le bruit est toujours présent. Cela peut être présenté comme deux hypothèses mutuellement exclusives, respectivement $\mathcal{H}_1$ et $\mathcal{H}_0$:

$$\mathcal{H}_0 : \quad x(k,\ell) = \ n(k,\ell), \tag{2.17}$$

$$\mathcal{H}_1 : \quad x(k,\ell) = \ s(k,\ell) + n(k,\ell). \tag{2.18}$$

Formulation du problème et solution

L'estimation de la source, notée $\hat{s}(k,\ell)$, est construite comme le produit du signal d'entrée et d'un gain $w(k,\ell)$ dans le domaine de la TFCT :

$$\hat{s}(k,\ell) = w(k,\ell)\ x(k,\ell). \tag{2.19}$$

Ce gain équivaut à un filtre adaptatif dans le domaine temporel. La détermination de $w(k,\ell)$ peut s'écrire comme un problème d'optimisation pour lequel on souhaite minimiser la puissance (ou variance, dans un cadre probabiliste) de l'erreur, *i.e.* la différence entre la source idéale et son estimation :

$$w_{\text{WF}}(k,\ell) = \underset{w}{\text{argmin}} \left\{ \mathbb{E}\left[|s(k,\ell) - w\ x(k,\ell)|^2 \right] \right\}. \tag{2.20}$$

La fonction de coût est quadratique positive et permet de dériver une solution analytique. Celle-ci est appelée le filtre de Wiener (de l'anglais *Wiener Filter*, WF), notée $w_{\text{WF}}(k,\ell)$, et peut s'écrire de la manière suivante :

$$w_{\text{WF}}(k,\ell) = \frac{\phi_s(k,\ell)}{\phi_s(k,\ell) + \phi_n(k,\ell)} = \frac{\xi(k,\ell)}{1+\xi(k,\ell)}, \tag{2.21}$$

où $\xi(k,\ell) = \frac{\phi_s(k,\ell)}{\phi_n(k,\ell)}$ est le RSB d'entrée. On remarque alors qu'il est nécessaire de connaitre les variances des sources de parole et de bruit. Par définition, ces paramètres sont inaccessibles en pratique car on souhaite justement séparer ces deux sources à partir du mélange.

Estimation des paramètres

Il existe de nombreuses manières d'estimer la variance du bruit [Loizou, 2013]. Celles-ci sont plus ou moins fondées théoriquement, certaines sont heuristiques et basées sur une approche assez intuitive du modèle des signaux sans être tout à fait fondée théoriquement. Dans ce travail, on considère un algorithme basé sur une approche bayésienne et utilisé de manière standard dans la littérature [Ephraim and Malah, 1984, Cohen, 2002].

Tout d'abord, en utilisant les Eq. (2.17) et (2.18), on peut développer l'expression de la variance du bruit comme suit :

$$\begin{aligned}\phi_n(k,\ell) &= \mathbb{E}\left[|n(k,\ell)|^2|x(k,\ell)\right] \\ &= \mathbb{E}\left[|n(k,\ell)|^2|\mathcal{H}_0\right] P(\mathcal{H}_0|x(k,\ell)) + \mathbb{E}\left[|n(k,\ell)|^2|\mathcal{H}_1\right] P(\mathcal{H}_1|x(k,\ell)).\end{aligned} \tag{2.22}$$

Il est alors courant d'estimer les espérances conditionnelles en utilisant un filtre récursif lors des phase d'absence de la parole et de bloquer l'estimation lors des phases de présence [Loizou, 2013, p. 407] :

$$\begin{aligned}\mathbb{E}\left[|n(k,\ell)|^2|\mathcal{H}_0\right] &\approx \alpha_n|x(k,\ell)|^2 + (1-\alpha_n)\phi_n(k,\ell-1) \\ \mathbb{E}\left[|n(k,\ell)|^2|\mathcal{H}_1\right] &\approx \phi_n(k,\ell-1),\end{aligned}$$

où $\alpha_n \in [0,1]$ est le facteur de lissage. On peut alors réécrire l'Eq. (2.22) de la manière suivante en posant $p(k,\ell) = P(\mathcal{H}_1|x(k,\ell))$:

$$\begin{aligned}\hat{\phi}_n(k,\ell) &= (\alpha_n|x(k,\ell)|^2 + (1-\alpha_n)\hat{\phi}_n(k,\ell-1))(1-p(k,\ell)) + \hat{\phi}_n(k,\ell-1)p(k,\ell) \\ &= \alpha_n(1-p(k,\ell))|x(k,\ell)|^2 + \left[1-\alpha_n(1-p(k,\ell))\right]\hat{\phi}_n(k,\ell-1).\end{aligned} \tag{2.23}$$

A cette étape, il reste donc à estimer $p(k,\ell)$, la probabilité de présence de la parole sachant l'observation du mélange. On peut alors utiliser le théorème de Bayes pour faire apparaitre la vraisemblance des données sachant chacune des hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$, notées respectivement $P(x(k,\ell)|\mathcal{H}_0)$ et $P(x(k,\ell)|\mathcal{H}_1)$ et la probabilité *a priori*de chacune des hypothèses, notées $P(\mathcal{H}_0)$ et $P(\mathcal{H}_1)$:

$$P(\mathcal{H}_1|x(k,\ell)) = \frac{P(x(k,\ell)|\mathcal{H}_1)P(\mathcal{H}_1)}{P(x(k,\ell)|\mathcal{H}_0)P(\mathcal{H}_0) + P(x(k,\ell)|\mathcal{H}_1)P(\mathcal{H}_1)}. \tag{2.24}$$

Les vraisemblances sont dérivées à partir du modèle présenté en Eq. (2.16) et les probabilités *a priori*peuvent être fixées ou estimées de manière adaptative [Cohen, 2002]. En intégrant les modèles de vraisemblance, on peut réécrire l'équation sous la forme :

$$P(\mathcal{H}_1|x(k,\ell)) = \left(1 + (1+\xi(k,\ell))\frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)}e^{\frac{-\gamma(k,\ell)\xi(k,\ell)}{1+\xi(k,\ell)}}\right)^{-1} \tag{2.25}$$

où $\xi(k,\ell) = \frac{\phi_s(k,\ell)}{\phi_n(k,\ell)}$ et $\gamma(k,\ell) = \frac{|x(k,\ell)|^2}{\phi_n(k,\ell)}$ sont appelés dans la littérature les RSBs *a priori* et *a posteriori*, respectivement [McAulay and Malpass, 1980].

Limites

L'estimation des paramètres du filtre de Wiener peut être difficile, notamment pour un RSB faible (inférieur à 0 dB) et mène à l'introduction d'artéfacts appelés bruit « musical » évoquant le bruit de l'eau qui coule. Par ailleurs, le choix d'une distribution gaussienne pour modéliser les coefficients TFD de la parole peut sembler trop simpliste. En effet, celle-ci est en réalité plus « pointue » qu'une gaussienne pour un signal de parole (et de bruit, dans une moindre mesure). De récents travaux [Fontaine et al., 2017, Leglaive et al., 2019] ont cherché à concevoir le même genre d'algorithme de débruitage en utilisant une distribution α-stable dont le kurtosis est paramétrable pour modéliser plus fidèlement la distribution des coefficients TFD. Cependant, il est beaucoup plus difficile de dériver des algorithmes avec ce modèle et de nouveaux paramètres à estimer sont particulièrement difficiles d'accès. Dans la suite nous considérons uniquement le modèle gaussien. Dans le contexte des prothèses auditives, il semble que ce genre d'algorithme ne permette pas d'améliorer la compréhension de la parole dans le bruit [Chong and Jenstad, 2018], contrairement aux algorithmes multicanaux [McCreery et al., 2012a, Völker et al., 2015]. Cependant, ils améliorent le confort d'écoute et réduisent la fatigue auditive [Chong and Jenstad, 2018, Wong et al., 2018].

2.2.2 Algorithmes de débruitage multicanaux (beamforming)

Le débruitage multicanal est utilisé de longue date dans les prothèses auditives [Luo et al., 2002]. L'algorithme historiquement employé est appelé le réseau de microphone différentiel du premier ordre [Benesty et al., 2016]. Celui-ci s'appuie sur un réseau composé de deux microphones alignés avec la ligne du regard de l'auditeur. Il a la particularité de former un diagramme de directivité invariant en fréquence [2] évitant une distorsion du timbre des sources hors-axe. Cependant, celui-ci souffre de nombreuses limitations. En effet, il amplifie le bruit de fond électronique dans les basses fréquences et le choix de la direction de visée est limité à l'avant ou l'arrière de la droite formée par les deux microphones. De plus, cette famille d'algorithmes se généralise mal à tout type de géométrie de réseau de microphone, notamment à la géométrie du réseau formé lorsque l'on fait communiquer les deux prothèses auditives. Pour ces raisons, les réseaux de microphones différentiels ont été délaissés dans le cadre des prothèses auditives au profit des réseaux de microphone additifs. Dans la suite on ne considèrera que ces derniers et on verra qu'ils sont plus flexibles

[2] Sur la bande-passante considérée.

Notation	$\in$	Definition
$\mathbb{E}[.]$		L'espérance mathématique
$\mathbf{x}(k,\ell)$	$\mathbb{C}^M$	Les signaux des microphones
$\mathbf{h}(k)$	$\mathbb{C}^M$	Les ATFs de la source
$\mathbf{H}(k)$	$\mathbb{C}^{M\times Q}$	Les ATFs des Q sources
$\mathbf{n}(k,\ell)$	$\mathbb{C}^M$	Le bruit capté par les microphones
$\xi(k,\ell) = \frac{\phi_s(k,\ell)}{\phi_n(k,\ell)}$	$\mathbb{R}^+$	Le RSB
$\mathbf{\Phi_x}(k,\ell) = \mathbb{E}\left[\mathbf{x}(k,\ell)\mathbf{x}(k,\ell)^H\right]$	$\mathbb{C}^{M\times M}$	La matrice de covariance des signaux d'entrée
$\mathbf{\Phi_n}(k,\ell) = \mathbb{E}\left[\mathbf{n}(k,\ell)\mathbf{n}(k,\ell)^H\right]$	$\mathbb{C}^{M\times M}$	La matrice de covariance du bruit
$\mathbf{\Gamma_n}(k,\ell)$	$\mathbb{C}^{M\times M}$	La matrice de cohérence du bruit ($\mathbf{\Phi_n}$ normalisée)
$\mathbf{\Gamma}_{\text{diff}}(k)$	$\mathbb{C}^{M\times M}$	La matrice de cohérence du champ diffus isotropique

TABLE 2.1 – Notations mathématiques.

pour concevoir des algorithmes aux objectifs divers et qu'ils s'adaptent bien aux configurations de microphones utilisés en pratique.

Dans cette sous-section, nous présentons de manière didactique et progressive les différents algorithmes de débruitage multicanaux que l'on retrouve dans les prothèses auditives. Cette famille d'algorithme comporte de nombreuses ramifications selon les choix faits dans la modélisation et la formulation du problème.

Toutes les notations mathématiques nécessaires à la compréhension de cette partie sont rassemblées dans la Tab. 2.1. De manière générale, les symboles en gras minuscule réfèrent à un vecteur colonne et ceux en gras majuscule à une matrice.

Modèles des signaux

Scène sonore Considérons une scène sonore composée de Q sources de parole, dont le signal de la $q^{\text{ème}}$ source est noté $s_q(t)$, arrivant sous forme d'onde plane au niveau des M microphones, dont le signal du $m^{\text{ème}}$ élément est noté $x_m(t)$, et d'un bruit ambiant noté $n_m(t)$, différent dans chaque microphone mais potentiellement corrélé. Chaque microphone observe une version altérée de chaque source de parole. Les distorsions sont principalement dues au retard de propagation entre l'émission et la réception ainsi qu'à la diffraction sur un obstacle se situant sur le trajet de l'onde, *e.g.* une tête. Celles-ci sont étroitement liées aux HRTFs définies en Section 1.2 (p.9) et sont donc modélisables par une convolution avec une réponse impulsionnelle notée $h_{m,q}(t)$. Ce scénario

est illustré schématiquement en Fig. 2.4. On peut alors écrire :

$$x_m(t) = \sum_{q=1}^{Q} (h_{m,q} \star s_q)(t) + n_m(t), \tag{2.26}$$

où $\star$ est le produit de convolution. Ce dernier rend la manipulation du modèle assez compliquée. Afin de s'en affranchir, on préfère se placer dans le domaine fréquentiel où il devient un produit simple. Cependant, comme nous sommes dans une application temps-réel, il est nécessaire de remplacer la transformée de Fourier sur le signal entier par une TFCT, comme décrit en Section 1.3 (p.19). Cette approximation est valide seulement si les $h_{m,q}(t)$ sont assez courtes par rapport à la fenêtre d'analyse de la TFCT [Avargel and Cohen, 2007]. Dans notre application, la durée de $h_{m,q}(t)$ est de l'ordre de 2 ms et celle de la fenêtre d'analyse est de 10 ms. On peut alors écrire :

$$x_m(k,\ell) = \sum_{q=1}^{Q} h_{m,q}(k) s_q(k,\ell) + n_m(k,\ell), \tag{2.27}$$

où $h_{m,q}(k)$ est appelée l'fonction de transfert acoustique (de l'anglais *Acoustic Transfer Function*, ATF) entre la $q^{\text{ème}}$ source et le $m^{\text{ème}}$ microphone. Nous verrons qu'il est plus commode de considérer cette équation sous sa forme matricielle :

$$\mathbf{x}(k,\ell) = \mathbf{H}(k)\mathbf{s}(k,\ell) + \mathbf{n}(k,\ell), \tag{2.28}$$

où $\mathbf{x}(k,\ell) = [x_1(k,\ell), ..., x_M(k,\ell)]^T$, $\mathbf{H}(k) = [\mathbf{h}_1(k), ..., \mathbf{h}_Q(k)]$ avec $\mathbf{h}_q(k) = [h_{1,q}(k), ..., h_{M,q}(k)]^T$, $\mathbf{s}(k,\ell) = [s_1(k,\ell), ..., s_Q(k,\ell)]^T$ et $\mathbf{n}(k,\ell) = [n_1(k,\ell), ..., n_M(k,\ell)]^T$.

Modélisation probabiliste Similairement au débruitage monocanal, il sera utile de modéliser les coefficients TFD des sources de parole et du bruit comme des variables aléatoires indépendantes suivant une distribution Gaussienne complexe multivariée centrée circulaire :

$$\mathbf{s}(k,\ell) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{\Phi}_{\mathbf{s}}(k,\ell)), \tag{2.29}$$

$$\text{avec} \quad \mathbf{\Phi}_{\mathbf{s}}(k,\ell) = \begin{bmatrix} \phi_{s_1}(k,\ell) & 0 & ... & 0 \\ 0 & \phi_{s_2}(k,\ell) & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \phi_{s_Q}(k,\ell) \end{bmatrix}, \tag{2.30}$$

$$\text{et} \quad \mathbf{n}(k,\ell) \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{\Phi}_{\mathbf{n}}(k,\ell)), \tag{2.31}$$

où $\mathbf{\Phi}_{\mathbf{s}}(k,\ell)$ et $\mathbf{\Phi}_{\mathbf{n}}(k,\ell)$ sont les matrices de covariance des sources de parole et du bruit, respectivement. Cette dernière peut être décomposée en une matrice de cohérence (matrice de covariance normalisée), notée $\mathbf{\Gamma}_{\mathbf{n}}(k,\ell)$ contenant

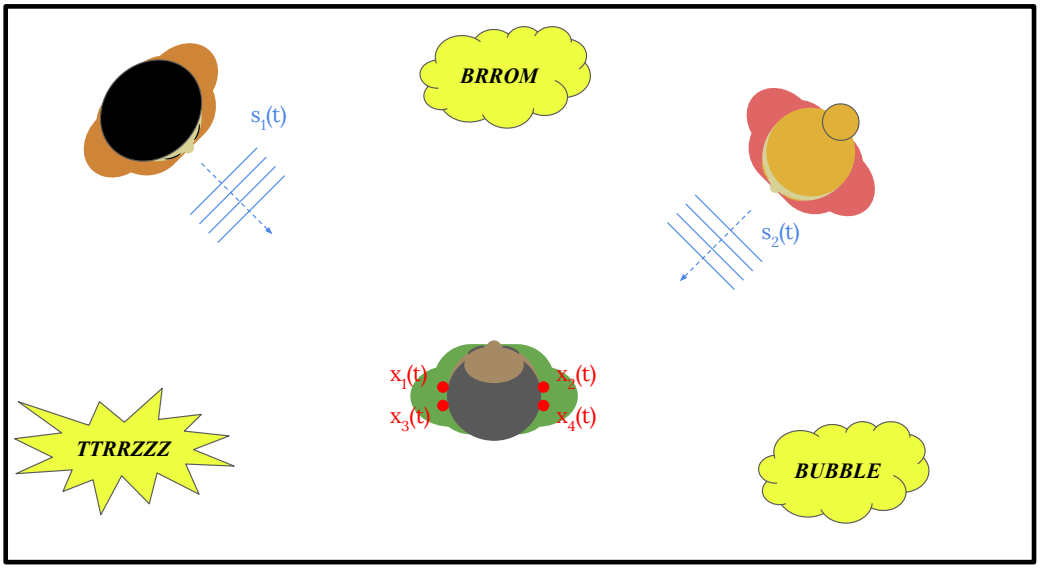


FIGURE 2.4 – Schéma du scénario de la scène sonore considérée dans cette partie vue du dessus. Celle-ci est composée d'un auditeur (au centre) portant deux prothèses auditives composées chacune de deux microphones (en rouge) et de deux interlocuteurs dont la voix est figurée sous forme d'onde sonore en bleu.

l'information spatiale du bruit, et un facteur d'échelle, $\phi_n(k,\ell)$, contenant l'information de la puissance du bruit :

$$\mathbf{\Phi_n}(k,\ell) = \phi_n(k,\ell)\mathbf{\Gamma_n}(k,\ell). \tag{2.32}$$

Structure du beamformer Dans l'idéal, la sortie du beamformer correspond à la somme des sources d'intérêt sans le bruit :

$$y(k,\ell) = \mathbf{d}(k)^H \mathbf{s}(k,\ell), \tag{2.33}$$

où $.^H$ est la transposée hermitienne et $\mathbf{d}(k)^H \in \mathbb{C}^{1\times Q}$ contient la réponse en fréquence désirée pour chaque source de parole. L'estimateur de $y(k,\ell)$, noté $\hat{y}(k,\ell)$, est défini comme une combinaison linéaire des coefficients de Fourier des différents microphones :

$$\hat{y}(k,\ell) = \mathbf{w}(k,\ell)^H \mathbf{x}(k,\ell), \tag{2.34}$$

avec $\mathbf{w}(k,\ell) \in \mathbb{C}^M$, le vecteur contenant les poids à déterminer pour obtenir une telle estimation.

Simplification à une source d'intérêt Il est assez courant de considérer $Q = 1$, soit parce que dans l'application considérée, il y a en effet qu'une seule source de parole présente, soit en faisant l'hypothèse de parcimonie des sources de parole dans le plan T-F, *i.e.* qu'une seule source est présente par point T-F [Rickard and Yilmaz, 2002]. Dans ce cas, plusieurs sources peuvent être présentes dans la scène sonore mais leur énergie est regroupée sur peu de points T-F et rarement les mêmes. On simplifie alors l'Eq. (2.28) en posant $\mathbf{h}(k) = \mathbf{h}_1(k)$ et $s(k,\ell) = s_1(k,\ell)$:

$$\mathbf{x}(k,\ell) = \mathbf{h}(k)s(k,\ell) + \mathbf{n}(k,\ell). \tag{2.35}$$

Dans la suite, nous allons exposer les différentes stratégies pour régler $\mathbf{w}(k,\ell)$.

Le beamformer *aligneur*

Dans un premier temps, limitons nous à un scénario composé d'une unique source d'intérêt, $Q = 1$. Considérons un beamformer dont l'objectif est uniquement de préserver la réponse en fréquence du signal cible dans la sortie. Comme illustré en Fig. 2.5, une onde sonore (plane) atteint chaque microphone avec un certain retard (et potentiellement une certaine atténuation) selon son angle d'arrivée par rapport au réseau. L'objectif du beamformer revient donc à réaligner les signaux des microphones en phase et en amplitude de sorte à ce qu'un signal provenant de la direction cible ressorte en phase avec un gain de 1. En injectant le modèle en Eq. (2.35) dans l'expression du beamformer en (2.34) et en respectant la contrainte de reconstruction parfaite du signal cible $\mathbf{w}(k)^H\mathbf{h}(k) = 1$, on peut écrire :

$$\begin{aligned}\hat{y}(k,\ell) &= \underbrace{\mathbf{w}(k)^H\mathbf{h}(k)}_{=1} s(k,\ell) + \mathbf{w}(k)^H\mathbf{n}(k,\ell), \\ &= s(k,\ell) + \mathbf{w}(k)^H\mathbf{n}(k,\ell).\end{aligned} \tag{2.36}$$

Pour remplir cette contrainte, on peut trouver une solution triviale de la forme :

$$\mathbf{w}_0(k) = \frac{\mathbf{h}(k)}{||\mathbf{h}(k)||^2}. \tag{2.37}$$

Il faut noter que n'importe quel vecteur $\mathbf{w}$ dont le produit scalaire avec $\mathbf{h}(k)$ est égal à 1 fonctionne et qu'il en existe une infinité. En l'occurrence, rien n'assure que $\mathbf{w}_0(k)$ soit le filtre idéal pour minimiser la composante de bruit $\mathbf{w}(k)^H\mathbf{n}(k,\ell)$ dans la sortie du beamformer, ou tout autre critère.

On peut alors regarder la réponse en fréquence en sortie du beamformer pour chaque direction (θ,ϕ), appelée le diagramme de directivité, noté $P(\theta,\phi,k)$:

$$P(\theta,\phi,k) = \mathbf{w}(k)^H\mathbf{h}(\theta,\phi,k), \tag{2.38}$$

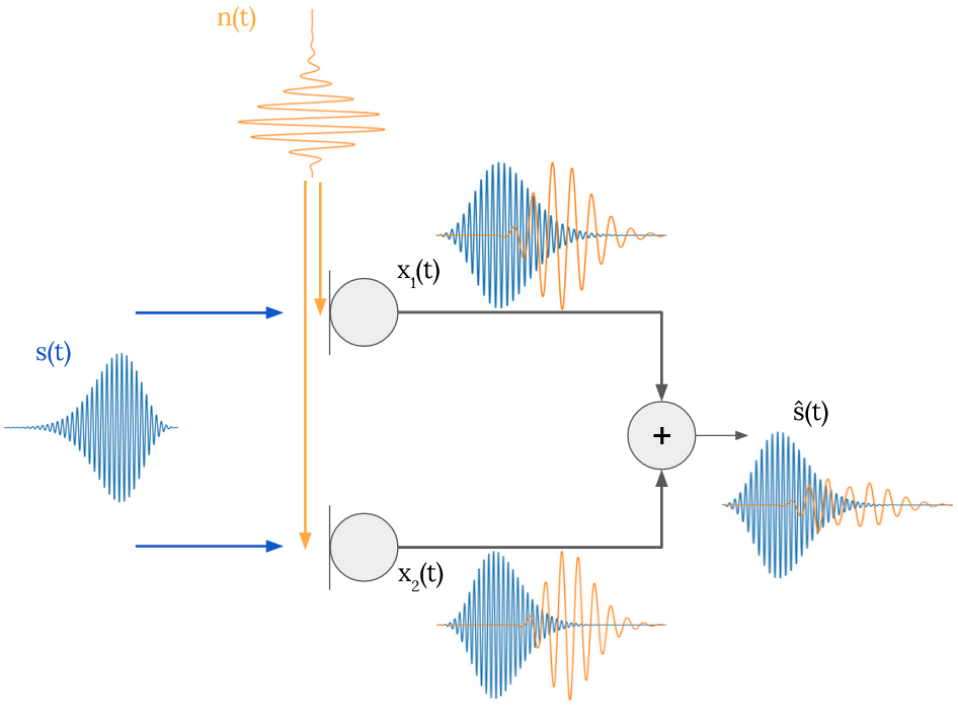
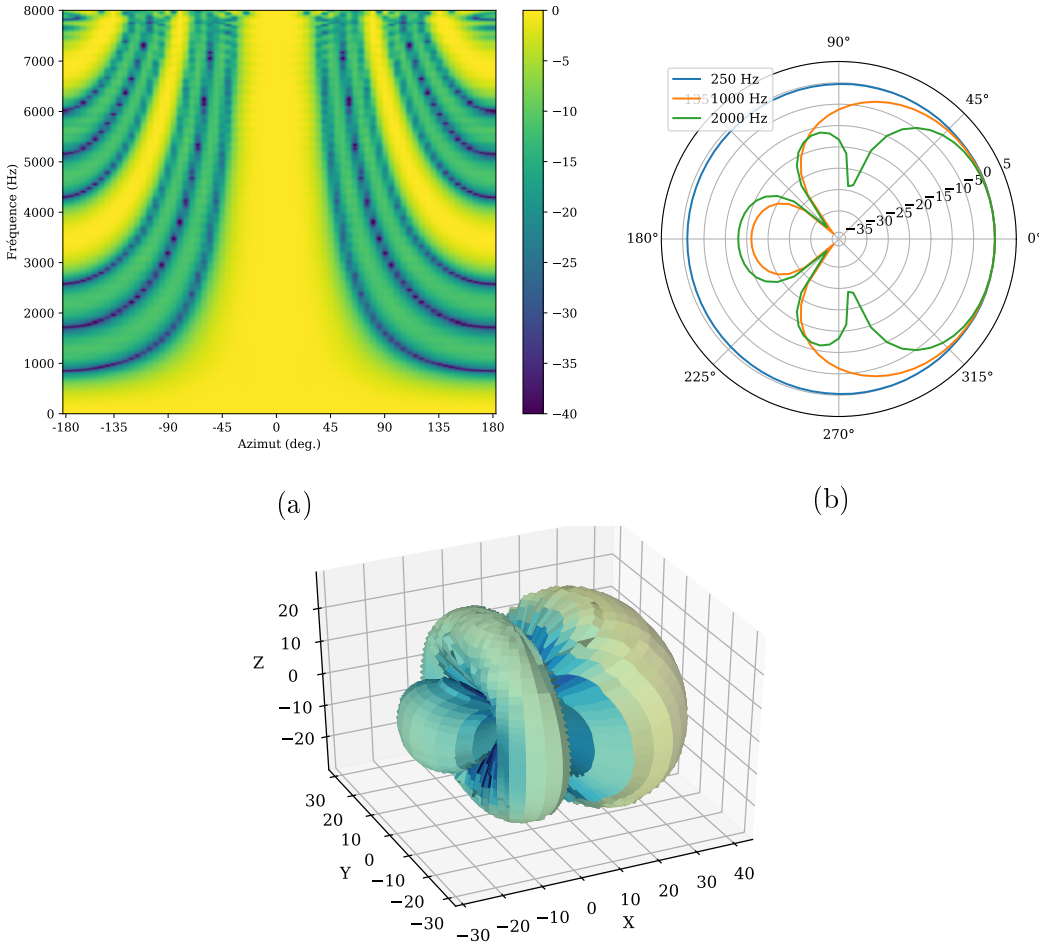


FIGURE 2.5 – Schéma d'un beamformer aligneur composé de deux microphones.

où $\mathbf{h}(\theta, \phi, k) \in \mathbb{C}^M$ est le vecteur contenant les fonctions de transfert entre une source située à la direction d'azimut-élévation (θ, ϕ) et les microphones ; on fait apparaitre explicitement la dépendance de $\mathbf{h}$ en (θ, ϕ) dans ce cas particulier où on on parcourt l'ensemble des directions possibles. Nous avons représenté en Fig. 2.6 le diagramme de directivité de puissance de plusieurs manières pour un beamformer aligneur afin d'offrir la représentation la plus complète de cet objet multidimensionnel. Le réseau utilisé pour cette stimulation est constitué de quatre microphones en ligne. On observe qu'une onde plane provenant d'une direction autre que la direction de visée subit un filtrage en peigne par le beamformer. On comprend alors aisément que plus le nombre de microphones est élevé plus les phénomènes de filtrage en peigne vont se cumuler et augmenter l'atténuation des sons arrivant hors-axe. On remarque aussi que la réponse en fréquence dépend de la direction.

Le beamformer à RSB maximal (MSNR)

On vient de voir qu'il est possible d'assurer une reconstruction parfaite de la source dans la sortie du beamformer en s'assurant de respecter la contrainte $\mathbf{w}^H \mathbf{h}(k) = 1$. Tout en la respectant, il est possible de trouver une solution qui



(a) (b)

(c)

FIGURE 2.6 – Diagramme de directivité de puissance exprimé en dB pour un réseau de microphone uniforme linéaire, sur le plan horizontal et en fonction de la fréquence (a), sur le plan horizontal en diagramme polaire pour les fréquences 0,25, 1 et 2 kHz (b), et enfin pour la fréquence 2 kHz pour toutes les directions (c).

maximise ou minimise un critère donné. Le plus direct que l'on peut considérer pour un algorithme de débruitage est le RSB de sortie, noté $\xi_{\text{o}}(k,\ell)$. Celui-ci peut s'écrire comme suit :

$$\xi_{\text{o}}(k,\ell) = \frac{\mathbb{E}\left[|\mathbf{w}(k,\ell)^H \mathbf{h}(k) s(k,\ell)|^2\right]}{\mathbb{E}\left[|\mathbf{w}(k,\ell)^H \mathbf{n}(k,\ell)|^2\right]} \tag{2.39}$$

$$= \frac{|\mathbf{w}(k,\ell)^H \mathbf{h}(k)|^2}{\mathbf{w}(k,\ell)^H \mathbf{\Gamma}_{\mathbf{n}}(k,\ell)\mathbf{w}(k,\ell)} \xi(k,\ell), \tag{2.40}$$

avec $\xi(k,\ell) = \frac{\phi_s(k,\ell)}{\phi_n(k,\ell)}$, le RSB d'entrée. Maximiser ce critère revient à minimiser son dénominateur, le numérateur correspondant à la contrainte. Ainsi, chercher le filtre maximisant le RSB, noté $\mathbf{w}_{\text{MSNR}}(k,\ell)$, revient à poser le problème d'optimisation sous contrainte suivant :

$$\mathbf{w}_{\text{MSNR}}(k,\ell) = \underset{\mathbf{w}}{\text{argmin}} \left\{\mathbf{w}^H \mathbf{\Gamma}_{\mathbf{n}}(k,\ell)\mathbf{w}\right\} \quad \text{s.t. } \mathbf{w}^H \mathbf{h}(k) = 1. \tag{2.41}$$

La fonction de coût est quadratique et $\mathbf{\Gamma}_{\mathbf{n}}(k,\ell)$ est définie positive ce qui nous permet d'utiliser la méthode des multiplieurs de Lagrange. On peut alors écrire la solution sous la forme suivante :

$$\mathbf{w}_{\text{MSNR}}(k,\ell) = \frac{\mathbf{\Gamma}_{\mathbf{n}}^{-1}(k,\ell)\mathbf{h}(k)}{\mathbf{h}(k)^H \mathbf{\Gamma}_{\mathbf{n}}^{-1}(k,\ell)\mathbf{h}(k)}. \tag{2.42}$$

On peut remarquer que la solution est indépendante des variances du signal cible et du bruit. Celle-ci est appelée le beamformer à RSB maximal (de l'anglais *Maximum Signal-to-Noise Ratio*, MSNR).

En pratique, $\mathbf{\Gamma}_{\mathbf{n}}(k,\ell)$ est difficile à estimer et il est courant de l'approximer par $\mathbf{I}$ ou par $\mathbf{\Gamma}_{\text{diff}}(k)$, respectivement les matrices de cohérence correspondant à un bruit spatialement décorrélé [3] ou diffus [4] [Thiemann et al., 2016, Zohourian et al., 2018, Marquardt and Doclo, 2018]. Dans le premier cas, $\mathbf{w}_{\text{MSNR}}(k,\ell)$ devient $\mathbf{w}_0(k)$ et dans le deuxième cas le beamformer résultant maximise alors l'indice de directivité (ID) [Stadler and Rabinowitz, 1993]. Le DI est défini comme le rapport entre la puissance du diagramme de directivité [5] dans la

[3] Correspondant par exemple à du bruit thermique électronique dans les préamplificateurs des microphones.

[4] Correspondant par exemple à du bruit de fond ou de la réverbération [Schwarz and Kellermann, 2015].

[5] Défini en Eq. (2.38).

direction cible (θ_0, ϕ_0) et l'intégrale sur toutes les directions (θ, ϕ) :

$$\begin{aligned}\mathrm{DI}(k) &= \frac{|P(\theta_0,\phi_0,k)|^2}{\frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi}|P(\theta,\phi,k)|^2\,\sin(\phi)\;d\phi d\theta} \\ &= \frac{|\mathbf{w}(k)^H\mathbf{h}(\theta_0,\phi_0,k)|^2}{\mathbf{w}(k)^H\left(\frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi}\mathbf{h}(\theta,\phi,k)\mathbf{h}(\theta,\phi,k)^H\,\sin(\phi)\;d\phi d\theta\right)\mathbf{w}(k)} \\ &= \frac{|\mathbf{w}(k)^H\mathbf{h}(k)|^2}{\mathbf{w}(k)^H\mathbf{\Gamma}_{\mathrm{diff}}(k)\mathbf{w}(k)},\end{aligned} \tag{2.43}$$

où $\mathbf{\Gamma}_{\mathrm{diff}}(k)$ est la matrice de cohérence d'un bruit composé d'une somme d'onde plane d'égale intensité provenant de toutes les directions, appelé bruit diffus. On note la similitude de cette expression avec l'expression du RSB en sortie du beamformer en Eq. (2.40). L'approximation de $\mathbf{\Gamma}_{\mathbf{n}}(k,\ell)$ par $\mathbf{\Gamma}_{\mathrm{diff}}(k)$ est courante [Thiemann et al., 2016, Zohourian et al., 2018, Marquardt and Doclo, 2018] car les performances d'intelligibilité de la parole sont similaires sans avoir besoin d'estimer la matrice de cohérence en temps-réel ce qui est coûteux en ressource de calcul et risque d'introduire des erreurs d'estimation supplémentaires [Maj et al., 2006, Völker et al., 2015].

Par ailleurs, on peut montrer en utilisant l'identité de Woodbury qu'il est strictement équivalent d'utiliser la matrice de covariance du mélange, notée $\mathbf{\Phi}_{\mathbf{x}}(k,\ell)$, à la place de $\mathbf{\Gamma}_{\mathbf{n}}(k,\ell)$ dans la solution du beamformer MSNR en Eq. (2.42). Cela revient alors à minimiser la variance en sortie du beamformer sous contrainte de conserver une réponse en fréquence plate pour la direction de visée. Cette solution est appelé le beamformer sans distorsion à variance minimale (de l'anglais *Minimum Variance Distortionless Response*, MVDR) [Haykin and Liu, 2009, Chap. 9] ou sans distorsion à puissance minimale (de l'anglais *Minimum Power Distortionless Response*, MPDR) [Gannot et al., 2017] selon les auteurs. Ceci peut sembler une aubaine car il est trivial de voir que $\mathbf{\Phi}_{\mathbf{x}}(k,\ell)$ est beaucoup plus facile à estimer que $\mathbf{\Gamma}_{\mathbf{n}}(k,\ell)$. Cependant, bien que strictement égal, le beamformer MVDR est plus sensible aux erreurs d'estimation de $\mathbf{h}(k)$ [Cox, 1973, Ehrenberg et al., 2010], le rendant inutilisable en pratique. Notons que le beamformer MSNR est souvent appelé MVDR par abus de langage et c'est ce que nous ferons dans la suite du document, on note donc $\mathbf{w}_{\mathrm{MVDR}} = \mathbf{w}_{\mathrm{MSNR}}$.

Enfin, ce beamformer se généralise bien au modèle multi-locuteur [Suzuki et al., 1999]. Il est alors appelé le beamformer linéairement contraint à variance minimale (de l'anglais *Linearly Constrained Minimum Variance*, LCMV) et le problème d'optimisation s'écrit alors sous la forme :

$$\mathbf{w}_{\mathrm{LCMV}}(k,\ell) = \underset{\mathbf{w}}{\mathrm{argmin}}\left\{\mathbf{w}^H\mathbf{\Gamma}_{\mathbf{n}}(k,\ell)\mathbf{w}\right\}\ \text{s.t. } \mathbf{w}^H\mathbf{H}(k) = \mathbf{d}(k)^H, \tag{2.44}$$

où $\mathbf{d}(k) \in \mathbb{C}^Q$ contient les fonctions de transfert à appliquer à chaque source visée. La solution est dérivée de manière similaire et s'exprime comme suit :

$$\mathbf{w}_{\text{LCMV}}(k,\ell) = \mathbf{\Gamma}_{\mathbf{n}}^{-1}(k,\ell)\mathbf{H}(k)\left(\mathbf{H}(k)^H\mathbf{\Gamma}_{\mathbf{n}}^{-1}(k,\ell)\mathbf{H}(k)\right)^{-1}\mathbf{d}(k). \tag{2.45}$$

Cas particulier : Q=2 Dans le cas où $Q = 2$, il a été montré [Hadad et al., 2016] que le beamformer LCMV est équivalent à deux beamformers MVDR visant respectivement vers chacune des deux sources :

$$\mathbf{w}_{\text{LCMV}} = \frac{1}{1-\Gamma}\left(\mathbf{w}_{\text{MVDR, 1}}\left(d_1 - \frac{d_2\gamma_{1,2}}{\gamma_2}\right) + \mathbf{w}_{\text{MVDR, 2}}\left(d_2 - \frac{d_1\gamma_{1,2}^*}{\gamma_1}\right)\right) \tag{2.46}$$

où $\mathbf{w}_{\text{MVDR, q}}$ est le filtre du beamformer MVDR visant la $q^{\text{ème}}$ source et avec :

$$\gamma_q = \mathbf{h}_q^H\mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{h}_q \quad \forall q \in \{1,2\}, \tag{2.47}$$

$$\gamma_{1,2} = \mathbf{h}_1^H\mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{h}_2, \tag{2.48}$$

$$\Gamma = \frac{|\gamma_{1,2}|^2}{\gamma_1\gamma_2}. \tag{2.49}$$

Par soucis de brièveté, nous avons omis les indices k et ℓ.

Limites Il est important de noter que le nombre de sources considérées, Q, doit être inférieur au nombre de microphones, M, de sorte qu'il reste au moins un degré de liberté dans l'optimisation de $\mathbf{w}$ alloué au débruitage. Dans le cas des prothèses auditives binaurales, la taille standard du réseau de microphones est $M = 4$ (deux microphones sur chaque appareil) [Oreinos and Buchholz, 2013, Moore et al., 2019a], permettant de considérer jusqu'à trois sources d'intérêt dans une scène. Pour des problèmes complexes, il arrive que certains auteurs se permettent de considérer un réseau de microphone plus grand, $M = 6$, en ajoutant un microphone dans la conque ou sur le corps de l'appareil [Kayser et al., 2009]. Dans la suite, on va voir que l'on peut relâcher les contraintes linéaires de sorte à s'affranchir de la limite $Q < M$.

Le filtre de Wiener multicanal (MWF)

On vient de voir des beamformers conçus pour maximiser le RSB et le DI. Il est possible aussi d'en concevoir un qui minimise la puissance de l'erreur entre le signal désiré, $s(k,\ell)$ ou $y(k,\ell)$ suivant le modèle, et son estimation. Similairement à l'approche monocanal, le problème d'optimisation peut s'écrire de la manière suivante :

$$\mathbf{w}_{\text{MWF}}(k,\ell) = \underset{\mathbf{w}}{\text{argmin}}\left\{J_{\text{MWF}}(\mathbf{w},k,\ell)\right\}, \tag{2.50}$$

avec

$$J_{\text{MWF}}(\mathbf{w}, k, \ell) = \mathbb{E}\left[|s(k,\ell) - \mathbf{w}^H \mathbf{x}(k,\ell)|^2\right]. \tag{2.51}$$

La fonction de coût $J_{\text{MWF}}(\mathbf{w}, k, \ell)$ étant quadratique positive, son minimum se situe au point annulant son gradient. De telle sorte, on peut déterminer analytiquement la solution, appelée filtre de Wiener multicanal (de l'anglais *Multi-channel Wiener Filter*, MWF) :

$$\mathbf{w}_{\text{MWF}}(k,\ell) = \mathbf{\Phi}_{\mathbf{x}}^{-1}(k,\ell)\mathbf{h}(k)\phi_s(k,\ell). \tag{2.52}$$

En utilisant l'identité de Woodbury, le MWF peut être exprimé comme un beamformer MVDR suivi d'un filtre de Wiener [Brooks and Reed, 1972] :

$$\mathbf{w}_{\text{MWF}}(k,\ell) = w_{\text{WF}}(k,\ell)\ \mathbf{w}_{\text{MSNR}}(k,\ell). \tag{2.53}$$

Dans la suite, on retire les indices (k,ℓ) par soucis de brièveté. Du fait de l'indépendance de la source et du bruit, on peut diviser $J_{\text{MWF}}(\mathbf{w})$ en deux termes correspondant à la préservation de la source et la réduction du bruit. On peut alors ajouter un facteur de compromis, noté μ, de sorte à donner plus ou moins d'importance à une des deux fonctions du beamformer [Doclo and Moonen, 2002, Haykin and Liu, 2009] :

$$J_{\text{SDW}}(\mathbf{w}) = \mathbb{E}\left[|s - \mathbf{w}^H\mathbf{h}s|^2\right] + \mu\mathbb{E}\left[|\mathbf{w}^H\mathbf{n}|^2\right]. \tag{2.54}$$

Cette approche est appelée le MWF à distorsion de parole pondéré (de l'anglais *Speech-Distortion-Weighted-MWF*, SDW-MWF) et sa solution est :

$$\mathbf{w}_{\text{SDW}} = \left(\mathbf{h}\mathbf{h}^H\phi_s + \mu\mathbf{\Phi}_{\mathbf{n}}\right)^{-1}\mathbf{h}\phi_s. \tag{2.55}$$

De nouveau, il est possible de décomposer le SDW-MWF en un beamformer MVDR et un filtre de Wiener paramétrique :

$$\mathbf{w}_{\text{SDW}} = w_{\text{PWF}}\ \mathbf{w}_{\text{MVDR}}, \tag{2.56}$$

où w_{PWF} est le filtre de Wiener paramétrique :

$$w_{\text{PWF}} = \frac{\xi_{\text{MVDR}}}{\mu + \xi_{\text{MVDR}}}. \tag{2.57}$$

avec ξ_{MVDR}, le RSB en sortie du beamformer MVDR :

$$\begin{aligned}\xi_{\text{MVDR}} &= \frac{\mathbb{E}\left[|\mathbf{w}_{\text{MVDR}}^H\mathbf{h}s|^2\right]}{\mathbb{E}\left[|\mathbf{w}_{\text{MVDR}}^H\mathbf{n}|^2\right]} \\ &= \xi\ \mathbf{h}^H\mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{h}.\end{aligned} \tag{2.58}$$

On peut alors identifier trois régimes de fonctionnement selon la valeur associée à μ :

— $\mu \to 0$: $\mathbf{w}_{\text{SDW}} \to \mathbf{w}_{\text{MVDR}}$ [6] ;
— $\mu < 1$: privilégie la préservation de la cible au détriment du débruitage ;
— $\mu = 1$: $\mathbf{w}_{\text{SDW}} = \mathbf{w}_{\text{WMF}}$;
— $\mu > 1$: augmente la distorsion de la parole au profit d'un débruitage plus important.

Il n'existe pas de consensus clair afin de régler de manière optimal ce paramètre. Plusieurs propositions consistent à utiliser une estimation de la présence de la parole basée sur une estimation du RSB [Thiergart and Habets, 2014, Ngo et al., 2009, Bagheri and Giacobello, 2019].

Le MWF peut être facilement étendu au modèle multi-locuteur (MMWF) en modifiant légèrement la fonction de coût de sorte à faire apparaitre un terme pour chaque source cible [Markovich-Golan et al., 2012b] :

$$J_{\text{MMWF}}(\mathbf{w}) = \sum_{q=1}^{Q} \lambda_q \mathbb{E}\left[|d_q^* s_q - \mathbf{w}^H \mathbf{h}_q s_q|^2\right] + \mathbb{E}\left[|\mathbf{w}^H \mathbf{n}|^2\right], \tag{2.59}$$

où d_q^* est le gain complexe à appliquer à la $q^{\text{ème}}$ source, ceux-ci peuvent véhiculer les indices de localisation par exemple et λ_q le poids à accorder à la préservation de la $q^{\text{ème}}$ source.[7] Pour la suite des calculs, il est plus commode de placer le facteur de compromis entre débruitage et préservation des cibles devant chacun des termes concernant les sources cibles plutôt qu'un seul paramètre devant le terme visant à réduire le bruit comme en Eq. (2.54). Ainsi, le compromis peut être réglé indépendamment pour chacune des cibles. La solution analytique de ce problème d'optimisation peut être obtenue de la même manière que pour le MWF :

$$\mathbf{w}_{\text{MMWF}} = \left(\mathbf{H}\mathbf{\Lambda}\mathbf{\Phi}_{\mathbf{s}}\mathbf{H}^H + \mathbf{\Phi}_{\mathbf{n}}\right)^{-1} \mathbf{H}\mathbf{\Lambda}\mathbf{\Phi}_{\mathbf{s}}\mathbf{d}, \tag{2.60}$$

où $\mathbf{\Lambda} = \text{diag}\{\lambda_1, ..., \lambda_Q\}$. En utilisant l'identité de Woodbury, on peut reformuler la solution de la manière suivante :

$$\mathbf{w}_{\text{MMWF}} = \mathbf{\Phi}_{\mathbf{n}}^{-1}\mathbf{H}\left(\mathbf{\Lambda}^{-1}\mathbf{\Phi}_{\mathbf{s}}^{-1} + \mathbf{H}^H\mathbf{\Phi}_{\mathbf{n}}^{-1}\mathbf{H}\right)^{-1}\mathbf{d}. \tag{2.61}$$

Sous cette forme, il est direct de voir que le beamformer LCMV, présenté en Eq. (2.45), est un cas particulier du MMWF pour lequel le terme $\mathbf{\Lambda}^{-1}\mathbf{\Phi}_{\mathbf{s}}^{-1}$ devient négligeable devant $\mathbf{H}^H\mathbf{\Phi}_{\mathbf{n}}^{-1}\mathbf{H}$ *i.e.* pour un RSB d'entrée qui tend vers l'infini ou lorsque l'on fait tendre les λ_q vers l'infini.

Comme nous l'avions noté dans l'analyse du beamformer LCMV, celui-ci limite le nombre de sources d'intérêt suivant l'inégalité $Q < M$. Le MMWF permet de relâcher cette contrainte et de considérer potentiellement $Q > M$

[6] Il est nécessaire de prendre garde à conserver l'inversibilité de la matrice $\left(\mathbf{h}\mathbf{h}^H\phi_s + \mu\mathbf{\Phi}_{\mathbf{n}}\right)$.
[7] Si $Q = 1$, $\lambda_1 = \frac{1}{\mu}$.

sources d'intérêt, en accordant plus ou moins d'importante à la préservation de telle ou telle source selon son RSB grâce au paramètre λ_q. Cela nécessite d'avoir une estimation des puissances spectrales du bruit et de chacune des sources de parole. On utilise classiquement les estimateurs suivants [Ye and DeGroat, 1995, Thiergart et al., 2013] :

$$\hat{\phi}_{s_q} = \mathbf{w}_{\text{LCMV},q}^H \left(\hat{\mathbf{\Phi}}_{\mathbf{x}} - \hat{\phi}_n \mathbf{\Gamma}_{\mathbf{n}} \right) \mathbf{w}_{\text{LCMV},q} \tag{2.62}$$

$$\hat{\phi}_n = \frac{1}{M-Q} \text{Tr} \left\{ \mathbf{P} \ \hat{\mathbf{\Phi}}_{\mathbf{x}} \mathbf{\Gamma}_{\mathbf{n}}^{-1} \right\}, \tag{2.63}$$

où $\mathbf{w}_{\text{LCMV},q}$ est le beamformer LCMV préservant la $q^{\text{ème}}$ source et annulant toutes les autres et $\mathbf{P} = \mathbf{I} - \mathbf{H}(\mathbf{H}^H \mathbf{\Gamma}_{\mathbf{n}}^{-1} \mathbf{H})^{-1} \mathbf{H}^H \mathbf{\Gamma}_{\mathbf{n}}^{-1}$ où l'on remarque que :

$$(\mathbf{H}^H \mathbf{\Gamma}_{\mathbf{n}}^{-1} \mathbf{H})^{-1} \mathbf{H}^H \mathbf{\Gamma}_{\mathbf{n}}^{-1} = \begin{bmatrix} \mathbf{w}_{\text{LCMV},1}^H \\ \vdots \\ \mathbf{w}_{\text{LCMV},Q}^H \end{bmatrix}. \tag{2.64}$$

Enfin, il faut noter que ces estimateurs ne sont valables que pour $Q < M$.

2.2.3 Beamformers préservant les indices de localisation de la cible

Les beamformers présentés précédemment fournissent en sorti un signal mono. Afin de préserver les indices de localisation de la cible, cette sortie doit être filtrée par les HRTFs h_{L} et h_{R} contenant les indices de localisation souhaités comme illustré en Fig. 2.7. Définissons d'abord la structure du beamformer binaural :

$$\text{Sortie gauche :} \quad y_{\text{L}} = \mathbf{w}_{\text{L}}^H \ \mathbf{x} \tag{2.65}$$

$$\text{Sortie droite :} \quad y_{\text{R}} = \mathbf{w}_{\text{R}}^H \ \mathbf{x}, \tag{2.66}$$

avec y_{L} et y_{R} les sorties du beamformer pour l'oreille gauche et droite, respectivement. Cela permet de réécrire le problème d'optimisation du MWF comme suit :

$$\mathbf{w}_{\text{MWF, L}} = \underset{\mathbf{w}}{\text{argmin}} \left\{ \mathbb{E} \left[\left| h_{\text{L}} s - \mathbf{w}^H \mathbf{x} \right|^2 \right] \right\} \tag{2.67}$$

$$\mathbf{w}_{\text{MWF, R}} = \underset{\mathbf{w}}{\text{argmin}} \left\{ \mathbb{E} \left[\left| h_{\text{R}} s - \mathbf{w}^H \mathbf{x} \right|^2 \right] \right\} \tag{2.68}$$

et le problème d'optimisation du beamformer MVDR ainsi :

$$\mathbf{w}_{\text{MVDR, L}} = \underset{\mathbf{w}}{\text{argmin}} \left\{ \mathbf{w}^H \mathbf{\Gamma}_{\mathbf{n}} \mathbf{w} \right\} \quad \text{s.t. } \mathbf{w}^H \mathbf{h} = h_{\text{L}} \tag{2.69}$$

$$\mathbf{w}_{\text{MVDR, R}} = \underset{\mathbf{w}}{\text{argmin}} \left\{ \mathbf{w}^H \mathbf{\Gamma}_{\mathbf{n}} \mathbf{w} \right\} \quad \text{s.t. } \mathbf{w}^H \mathbf{h} = h_{\text{R}}. \tag{2.70}$$

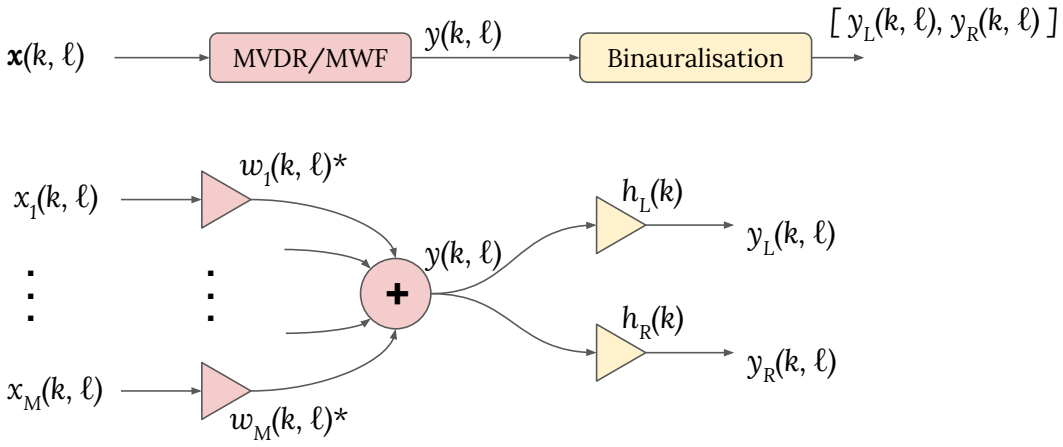


FIGURE 2.7 – Un beamformer préservant les indices de localisation de la cible est équivalent à binauraliser la sortie d'un beamformer standard.

Ce faisant, le son en sortie est perçu comme provenant d'une source virtuelle placée à la direction souhaitée (en général à la direction de la cible). En d'autres termes, les indices de localisation du bruit résiduel sont détruits et remplacés par les indices de localisation de la cible. Il est possible de le montrer [Cornelis et al., 2010] en définissant la fonction de transfert interaurale (de l'anglais *Interaural Transfer Function*, ITF) du signal de parole en entrée, notée ITF^s, contenant aussi bien l'ILD que l'ITD :

$$\mathrm{ITF}^s = \frac{h_\mathrm{L}}{h_\mathrm{R}}, \tag{2.71}$$

et en sortie pour la parole et le bruit, notées respectivement ITF^s_o et ITF^n_o :

$$\mathrm{ITF}^s_o = \frac{\mathbf{w}_\mathrm{L}^H \mathbf{h}}{\mathbf{w}_\mathrm{R}^H \mathbf{h}} = \frac{h_\mathrm{L}}{h_\mathrm{R}} = \mathrm{ITF}^s \tag{2.72}$$

$$\text{et} \quad \mathrm{ITF}^n_o = \frac{\mathbf{w}_\mathrm{L}^H \mathbf{n}}{\mathbf{w}_\mathrm{R}^H \mathbf{n}} = \frac{h_\mathrm{L}}{h_\mathrm{R}} = \mathrm{ITF}^s, \tag{2.73}$$

et ce, que ce soit pour le beamfomer MVDR ou le MWF.

2.2.4 Beamformers préservant les indices de localisation du bruit

Dans la sous-section précédente, on s'est intéressé à la préservation des indices de localisation de la cible, modélisée comme une source sonore provenant d'une unique direction. Quant au bruit, il peut être de nature très différente

et ne peut pas être modélisé de cette manière dans le cas général. Il peut être tantôt une source de parole interférant, auquel cas, modélisé comme une source venant d'une direction unique, mais en général il va être considéré comme étant spatialement plus diffus. Or, pour étudier la préservation des indices de localisation, nous utilisions jusqu'à lors l'ITF qui porte en elle l'ITD et l'ILD. Ces quantités sont pertinentes pour une source correspondant à une onde provenant d'une direction bien définie mais pas pour des sources diffuses.

On va alors distinguer plusieurs stratégies adoptées par les algorithmes proposés dans la littérature. D'une part, les algorithmes de type MWF avec estimation partielle du bruit (MWF-N) qui cherchent à préserver une version atténuée du bruit dans la sortie, sans chercher à préserver en particulier les indices de localisation. D'autre part, les algorithmes cherchant à préserver les indices interauraux, soit d'une source spatialement localisée, soit d'une source spatialement diffuse via un nouveau critère : la cohérence interaurale.

Le MWF avec estimation partielle du bruit (MWF-N)

L'idée est de changer légèrement la fonction de coût du SDW-MWF, présentée en Eq. (2.54) (p.48), en remplaçant le terme minimisant la puissance du bruit en sortie par un autre minimisant la puissance de l'erreur entre ce même bruit et une version atténuée d'un facteur η du bruit reçu à un microphone de référence [Van den Bogaert et al., 2008]. La fonction de coût du problème d'optimisation pour l'oreille gauche peut alors s'écrire comme suit :

$$J_{\text{MWF-N, L}}(\mathbf{w}) = \mathbb{E}\left[|h_{\text{L}}s - \mathbf{w}^H\mathbf{h}s|^2\right] + \mu\ \mathbb{E}\left[|\eta n_{\text{L}} - \mathbf{w}^H\mathbf{n}|^2\right], \tag{2.74}$$

où n_{L} est le bruit dans le microphone servant de référence pour l'oreille gauche. La fonction de coût pour l'oreille droite, notée $J_{\text{MWF-N, R}}(\mathbf{w})$, est définie similairement. La solution est dérivée de la même manière que pour le SDW-MWF (voir Eq. (2.55)) et peut s'exprimer comme suit :

$$\mathbf{w}_{\text{MWF-N, L}}(k,\ell) = (1-\eta)\ \mathbf{w}_{\text{SDW}}(k,\ell)\ h_{\text{L}}(k)^* + \eta\ \mathbf{q}_{\text{L}}, \tag{2.75}$$

où $\mathbf{q}_{\text{L}}$ est un vecteur nul à l'exception d'un élément égal à 1 correspondant au microphone de référence pour l'oreille gauche. Une démarche similaire donne la solution pour l'oreille droite.

Le paramètre η permet de régler le compromis entre débruitage et préservation des indices de localisation du bruit. La manière de le régler a été l'objet de quelques recherches et il n'y a pas encore de consensus clair. Il a été montré [Van den Bogaert et al., 2008] que $\eta = 0{,}2$ permet de rétablir la localisation de l'ensemble de la scène sonore. Par ailleurs, il a été montré [Marquardt and Doclo, 2018, Gössling et al., 2017] qu'il existe une expression de la MSC [8]

[8] La MSC est définie en Eq. (1.6) (p. 18).

en fonction de η. En général, une recherche exhaustive permet de trouver la solution. Les auteurs identifient un cas particulier pour lequel il est possible de trouver une solution analytique de η en fonction de la cohérence interaurale souhaitée en sortie. Il s'agit du cas où $\mu \to 0$, *i.e.* qu'aucune distorsion de la parole n'est tolérée, cette solution est appelée le beamformer MVDR-N (équivalent à désactiver le WF dans le MWF).

D'une manière moins théoriquement fondée, il a été proposé [Thiemann et al., 2016] de régler η de manière adaptative en fonction du RSB en entrée, noté $\xi(k,\ell)$. La proposition consiste à utiliser un classifieur binaire pour déterminer si le point temps-fréquence (T-F) est dominé par la parole ou par le bruit. Ainsi, lorsque la parole domine, on délivre la sortie du beamformer MVDR ($\eta(k,\ell) = 0$), dans le cas inverse c'est une version atténuée du signal d'entrée qui est délivré en sortie ($\eta(k,\ell) = \gamma(k)$) :

$$\mathbf{w}_{\text{SBB, L}}(k,\ell) = \begin{cases} h_{\text{L}}(k)^{*}\mathbf{w}_{\text{MVDR}}(k) & \text{si } \xi(k,\ell) > 1 \\ \gamma(k)\mathbf{q}_{\text{L}} & \text{sinon.} \end{cases} \tag{2.76}$$

avec $\gamma(k) = (\mathbf{w}_{\text{MVDR}}(k)^{H}\mathbf{\Gamma}_{\text{diff}}(k)\mathbf{w}_{\text{MVDR}}(k))^{-1}$, le gain [9] du beamformer MVDR.

Dans un but d'analyse, on propose la réécriture de l'Eq. (2.76) de la manière suivante :

$$\mathbf{w}_{\text{SBB, L}}(k,\ell) = \frac{\xi(k,\ell)^{\beta}}{1+\xi(k,\ell)^{\beta}} h_{\text{L}}(k)^{*}\mathbf{w}_{\text{MVDR}}(k) + \frac{\gamma(k)}{1+\xi(k,\ell)^{\beta}}\mathbf{q}_{\text{L}} \tag{2.77}$$

où $\beta \to \infty$. Cette reformulation de la proposition de [Thiemann et al., 2016] permet d'identifier une sorte de filtre de Wiener paramétrique en aval du beamformer MVDR. Cependant, celui-ci se distingue par l'utilisation du RSB d'entrée plutôt que de celui de sortie du beamformer MVDR alors qu'il est en aval de celui-ci. On peut se demander pourquoi utiliser le RSB d'entrée comme critère de classification et non celui en sortie du MVDR. En effet, on peut se retrouver dans la situation où un point T-F est classifié comme étant dominé par le bruit alors que celui-ci est dominé par la parole en sortie du beamformer MVDR, du fait du débruitage de celui-ci. Par ailleurs, utiliser le RSB de sortie du beamformer permettrait de rapprocher cette solution du MWF-N original.

De plus, notre reformulation fait apparaitre un nouveau paramètre, β, réglé à une valeur extrême dans la proposition originale. On peut alors se poser deux questions :

— est-ce que remplacer $\xi(k,\ell)$ par $\xi_{\text{MVDR}}(k,\ell) = \xi(k,\ell)\gamma(k)$ réduit la distorsion de la parole ?

— est-ce qu'une valeur de β moins extrême réduit la distorsion de la parole ?

[9] Le rapport entre le RSB de sortie et d'entrée du beamformer.

Le MWF-N et ses variants que nous avons présentés cherchent à préserver les indices de localisation du bruit de manière aveugle en réinjectant dans la sortie simplement une version atténuée de celui-ci. Dans la suite, nous allons présenter des beamformers qui font apparaitre explicitement les indices de localisation du bruit, notamment interauraux, dans la procédure d'optimisation.

Le beamformer LCMV avec préservation des ITFs des sources interférant (LCMV-ITF)

Plaçons nous dans un cas particulier pour lequel la composante de bruit $\mathbf{n}(k,\ell)$ est composée d'une somme de sources de parole non désirée, dites interférant. Le principal problème du beamformer LCMV est qu'il est limité par le nombre de contraintes linéaires (autant que de sources). Or, Q dans ce scénario peut être grand au vu de la limite indépassable de $2M-2$ contraintes dans le cas d'un beamformer LCMV binaural. Chaque contrainte mobilise un degré de liberté de la variable d'optimisation, $\mathbf{w}$, qui ne peut servir au débruitage. Pour réduire le nombre de contraintes, il a été proposé [Koutrouvelis et al., 2016] de préserver l'ITF plutôt que l'ATF des sources interférant. Ainsi, on préserve les indices de localisation interauraux en laissant la détermination du niveau sonore de celles-ci à la discrétion de la procédure de débruitage.

Afin de prendre en compte les indices de localisation interauraux dans l'optimisation, il est nécessaire d'optimiser $\mathbf{w}_{\mathrm{L}}$ et $\mathbf{w}_{\mathrm{R}}$ conjointement. Pour cela, on les considère comme une seule variable d'optimisation $\mathbf{w}$ de la manière suivante :

$$\mathbf{w} = \begin{bmatrix} \mathbf{w}_{\mathrm{L}} \\ \mathbf{w}_{\mathrm{R}} \end{bmatrix}. \tag{2.78}$$

Définissons alors l'ITF à l'entrée pour la $q^{\text{ème}}$ source interférant :

$$\mathrm{ITF}_q = \frac{h_{q,\mathrm{L}}}{h_{q,\mathrm{R}}} \quad \forall q \in \{2,...,Q\}, \tag{2.79}$$

et l'ITF à la sortie du beamformer :

$$\mathrm{ITF}_q^o = \frac{\mathbf{w}_{\mathrm{L}}^H \mathbf{h}_q}{\mathbf{w}_{\mathrm{R}}^H \mathbf{h}_q} \quad \forall q \in \{2,...,Q\}. \tag{2.80}$$

Par conséquent, préserver l'ITF de la $q^{\text{ème}}$ source interférant à la sortie du beamformer peut s'écrire comme suit :

$$\mathrm{ITF}_q^o = \mathrm{ITF}_q \qquad \forall q \in \{2,...,Q\} \tag{2.81}$$

$$\iff \mathbf{w}_{\mathrm{L}}^H \mathbf{h}_q h_{q,\mathrm{R}} - \mathbf{w}_{\mathrm{R}}^H \mathbf{h}_q h_{q,\mathrm{L}} = 0 \tag{2.82}$$

$$\iff \mathbf{w}^H \begin{bmatrix} \mathbf{h}_q h_{q,\mathrm{R}} \\ -\mathbf{h}_q h_{q,\mathrm{L}} \end{bmatrix} = 0, \tag{2.83}$$

ce qui correspond bien à une contrainte linéaire dans la formulation du problème d'optimisation du beamformer LCMV. La préservation de l'ITF d'une source consiste donc à mobiliser un degré de liberté de l'optimisation contrairement à la préservation de ses ATFs, en mobilisant deux.

Résumons alors les objectifs et contraintes de cet algorithme :

— préserver les ATFs de la source cible ;
— préserver les ITFs des sources interférant (sans contrôle de niveau sonore) ;
— réduire le bruit ambiant.

On peut alors écrire le problème d'optimisation sous la forme suivante :

$$\mathbf{w}_{\text{LCMV-ITF}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ \mathbf{w}^H \mathbf{P} \mathbf{w} \right\} \ \textbf{s.t.} \ \mathbf{w}^H \mathbf{C} = \mathbf{f}^H \tag{2.84}$$

avec

$$\mathbf{w}_{\text{LCMV-ITF}} = \begin{bmatrix} \mathbf{w}_{\text{LCMV-ITF, L}} \\ \mathbf{w}_{\text{LCMV-ITF, R}} \end{bmatrix}, \tag{2.85}$$

$$\mathbf{P} = \begin{bmatrix} \mathbf{\Gamma_n} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Gamma_n} \end{bmatrix}, \tag{2.86}$$

$$\mathbf{C} = \begin{bmatrix} \mathbf{h}_1 & \mathbf{0} & \mathbf{h}_2 h_{2,\text{R}} & \dots & \mathbf{h}_Q h_{Q,\text{R}} \\ \mathbf{0} & \mathbf{h}_1 & -\mathbf{h}_2 h_{2,\text{L}} & \dots & -\mathbf{h}_Q h_{Q,\text{L}} \end{bmatrix} \tag{2.87}$$

et

$$\mathbf{f}^H = \begin{bmatrix} h_{1,\text{L}} & h_{1,\text{R}} & 0 & \dots & 0 \end{bmatrix}. \tag{2.88}$$

Notons qu'avec cette approche, le nombre de contraintes est de $2+Q$ au lieu de $2Q$ pour le beamformer LCMV binaural et que l'on récupère donc $Q-2$ degrés de liberté pour débruiter. Remarquons aussi que cette fois-ci l'optimisation est conjointe, *i.e.* $\mathbf{w}_\text{L}$ et $\mathbf{w}_\text{R}$ ne peuvent être déterminés indépendamment l'un de l'autre.

MWF avec préservation de la cohérence interaurale (MWF-IC)

Lorsque le bruit est trop complexe pour être considéré comme la somme d'un nombre limité d'ondes planes, l'ITF n'est plus un indice pertinent sur lequel s'appuyer pour concevoir un beamformer. Il a été proposé de considérer l'IC du bruit, notée IC_n^{in}, dans ce genre de scénario [Marquardt et al., 2013]. Celle-ci est définie comme la corrélation croisée normalisée entre le bruit capté

à gauche et à droite, noté respectivement n_{L} et n_{R} :

$$\mathrm{IC}_n^{\mathrm{in}} = \frac{\mathbb{E}\left[n_{\mathrm{L}}\ n_{\mathrm{R}}^*\right]}{\sqrt{\mathbb{E}\left[|n_{\mathrm{L}}|^2\right]}\ \sqrt{\mathbb{E}\left[|n_{\mathrm{R}}|^2\right]}} \tag{2.89}$$

$$= \frac{\mathbf{q}_{\mathrm{L}}^T \mathbf{\Phi_n} \mathbf{q}_{\mathrm{R}}}{\sqrt{\mathbf{q}_{\mathrm{L}}^T \mathbf{\Phi_n} \mathbf{q}_{\mathrm{L}} \mathbf{q}_{\mathrm{R}}^T \mathbf{\Phi_n} \mathbf{q}_{\mathrm{R}}}} \tag{2.90}$$

$$= \frac{\mathbf{q}_{\mathrm{L}}^T \mathbf{\Gamma_n} \mathbf{q}_{\mathrm{R}}}{\sqrt{\mathbf{q}_{\mathrm{L}}^T \mathbf{\Gamma_n} \mathbf{q}_{\mathrm{L}} \mathbf{q}_{\mathrm{R}}^T \mathbf{\Gamma_n} \mathbf{q}_{\mathrm{R}}}}. \tag{2.91}$$

A partir de cette définition, l'IC en sortie du beamformer peut être écrite comme suit :

$$\mathrm{IC}_n^{\mathrm{out}}(\mathbf{w}_{\mathrm{L}}, \mathbf{w}_{\mathrm{R}}) = \frac{\mathbf{w}_{\mathrm{L}}^H \mathbf{\Phi_n} \mathbf{w}_{\mathrm{R}}}{\sqrt{\mathbf{w}_{\mathrm{L}}^H \mathbf{\Phi_n} \mathbf{w}_{\mathrm{L}} \mathbf{w}_{\mathrm{R}}^H \mathbf{\Phi_n} \mathbf{w}_{\mathrm{R}}}} \tag{2.92}$$

$$= \frac{\mathbf{w}_{\mathrm{L}}^H \mathbf{\Gamma_n} \mathbf{w}_{\mathrm{R}}}{\sqrt{\mathbf{w}_{\mathrm{L}}^H \mathbf{\Gamma_n} \mathbf{w}_{\mathrm{L}} \mathbf{w}_{\mathrm{R}}^H \mathbf{\Gamma_n} \mathbf{w}_{\mathrm{R}}}}. \tag{2.93}$$

Notons que $\mathrm{IC}_n^{\mathrm{in}}$ et $\mathrm{IC}_n^{\mathrm{out}}(\mathbf{w}_{\mathrm{L}}, \mathbf{w}_{\mathrm{R}})$ sont à valeurs complexes. Alors, on définit la fonction de coût visant à la préserver comme suit [Marquardt et al., 2013] :

$$J_{\mathrm{IC}}(\mathbf{w}_{\mathrm{L}}, \mathbf{w}_{\mathrm{R}}) = |\mathrm{IC}_n^{\mathrm{out}}(\mathbf{w}_{\mathrm{L}}, \mathbf{w}_{\mathrm{R}}) - \mathrm{IC}_n^{\mathrm{in}}|^2. \tag{2.94}$$

Afin de prendre en compte les indices de localisation interauraux dans l'optimisation, il est nécessaire d'optimiser $\mathbf{w}_{\mathrm{L}}$ et $\mathbf{w}_{\mathrm{R}}$ conjointement. Pour cela, on considère ces variables comme une seule variable d'optimisation $\mathbf{w}$ de la manière suivante :

$$\mathbf{w} = \begin{bmatrix} \mathbf{w}_{\mathrm{L}} \\ \mathbf{w}_{\mathrm{R}} \end{bmatrix}. \tag{2.95}$$

Ce faisant, on peut réécrire le problème d'optimisation menant au MWF de manière conjointe :

$$\mathbf{w}_{\text{MWF-IC}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ J_{\mathrm{BSDW}}(\mathbf{w}) + \lambda J_{\mathrm{IC}}(\mathbf{w}) \right\} \tag{2.96}$$

avec

$$J_{\mathrm{BSDW}}(\mathbf{w}) = \mathbf{w}^H \mathbf{R} \mathbf{w} - \mathbf{w}^H \mathbf{r} - \mathbf{r}^H \mathbf{w} \tag{2.97}$$

où

$$\mathbf{R} = \begin{bmatrix} \mathbf{h}\mathbf{h}^H \phi_s + \mu \mathbf{\Phi_n} & \mathbf{0}_{M \times M} \\ \mathbf{0}_{M \times M} & \mathbf{h}\mathbf{h}^H \phi_s + \mu \mathbf{\Phi_n} \end{bmatrix}, \tag{2.98}$$

$$\mathbf{r} = \begin{bmatrix} \mathbf{h}\ h_{\mathrm{L}}^* \\ \mathbf{h}\ h_{\mathrm{R}}^* \end{bmatrix} \phi_s \tag{2.99}$$

et λ est le paramètre arbitrant le compromis entre débruitage et préservation des indices de localisation du bruit.

Malheureusement, le terme $J_{\text{IC}}(\mathbf{w})$ est non-convexe et donc il n'existe pas de solution analytique comme on obtenait jusqu'à lors. Cette optimisation est réalisée grâce à un algorithme itératif comme la méthode quasi-Newton de Broyden-Fletcher-Goldarb-Shanno [Habets and Naylor, 2010, Itturriet and Costa, 2019]. Celle-ci est infaisable en temps-réel dans les prothèses auditives aujourd'hui.

Notons enfin que dans le cas particulier où le bruit est une source localisée dans l'espace, il a été montré [Itturriet and Costa, 2019] que le MWF-IC permet aussi de préserver l'ITD de cette source interférant.

Le beamformer MVDR-IC post-filtré

Comme on l'a vu précédemment, le MWF-IC dépend des données d'entrée et doit être calculé pour chaque point T-F. La résolution du problème d'optimisation n'étant pas faisable en temps-réel, l'idée est de le simplifier afin de le rendre indépendant des données d'entrée. Ceci permettrait de calculer le filtre en amont [Marquardt and Doclo, 2018].

Dérivation du beamformer MVDR-IC Pour ce faire, on doit faire deux simplifications dans la fonction de coût du MWF-IC. Rappelons là :

$$J_{\text{MWF-IC}}(\mathbf{w}) = \mathbf{w}^H\mathbf{R}\mathbf{w} - \mathbf{w}^H\mathbf{r} - \mathbf{r}^H\mathbf{w} + \lambda J_{\text{IC}}(\mathbf{w}) \tag{2.100}$$

où

$$\mathbf{R} = \begin{bmatrix} \mathbf{h}\mathbf{h}^H\phi_s + \mu\mathbf{\Phi}_{\mathbf{n}} & \mathbf{0}_{M\times M} \\ \mathbf{0}_{M\times M} & \mathbf{h}\mathbf{h}^H\phi_s + \mu\mathbf{\Phi}_{\mathbf{n}} \end{bmatrix}. \tag{2.101}$$

Premièrement, le facteur de distorsion de la parole, μ, est réglé à une valeur proche de zéro, *e.g.* 10^{-5}, de sorte à toujours pouvoir inverser $\mathbf{R}$. Ceci équivaut à n'accepter presque aucune distorsion de la parole et donc à faire tendre le MWF vers un beamformer MVDR. Deuxièmement, la variance de la cible est supposée égale à 1 et la matrice de covariance du bruit, $\mathbf{\Phi}_{\mathbf{n}}$, est remplacée par la matrice de cohérence d'un bruit spatialement diffus, $\mathbf{\Gamma}_{\text{diff}}$. On peut alors réécrire une version simplifiée de $\mathbf{R}$, notée $\tilde{\mathbf{R}}$, comme suit :

$$\tilde{\mathbf{R}} = \begin{bmatrix} \mathbf{h}\mathbf{h}^H + \mu\mathbf{\Gamma}_{\text{diff}} & \mathbf{0}_{M\times M} \\ \mathbf{0}_{M\times M} & \mathbf{h}\mathbf{h}^H + \mu\mathbf{\Gamma}_{\text{diff}} \end{bmatrix}. \tag{2.102}$$

En d'autres termes, on peut l'interpréter dans le cadre probabiliste que l'on s'était donné comme un scénario où la cible est supposée être un bruit blanc provenant de la direction de visée et où la composante de bruit est spectralement blanche et spatialement diffuse. Par conséquent, la première simplification

fait devenir le SDW-MWF en un beamformer MVDR et la seconde fait devenir le beamformer MVDR en un beamformer maximisant le DI. Dans le terme $J_{\text{IC}}(\mathbf{w})$, $\mathbf{\Gamma_n}$ est aussi remplacée par $\mathbf{\Gamma}_{\text{diff}}$ de sorte à le rendre indépendant du temps :

$$\tilde{J}_{\text{IC}}(\mathbf{w}) = \left| \frac{\mathbf{w}_{\text{L}}^H \mathbf{\Gamma}_{\text{diff}} \mathbf{w}_{\text{R}}}{\sqrt{\mathbf{w}_{\text{L}}^H \mathbf{\Gamma}_{\text{diff}} \mathbf{w}_{\text{L}} \mathbf{w}_{\text{R}}^H \mathbf{\Gamma}_{\text{diff}} \mathbf{w}_{\text{R}}}} - \text{IC}_v^{\text{in}} \right|^2 . \tag{2.103}$$

Ainsi, la fonction de coût obtenue ne dépend plus de la dimension temporelle et le filtre correspondant, noté $\mathbf{w}_{\text{MVDR-IC}}$, peut être calculé hors-ligne grâce à un algorithme itératif.

L'ajout d'un filtre en aval Les simplifications faites dans la fonction de coût de l'optimisation ont pour conséquence une diminution de l'amélioration du RSB en sortie comparé à un MWF standard (-7 dB en moyenne). Pour répondre à cet inconvénient, un filtre de Wiener est calculé pour les sorties gauche et droite des beamformers :

$$w_{\text{WF, L}} = \frac{\xi_o^{\text{L}}}{1 + \xi_o^{\text{L}}} \quad \text{et} \quad w_{\text{WF, R}} = \frac{\xi_o^{\text{R}}}{1 + \xi_o^{\text{R}}} \tag{2.104}$$

où ξ_o^{L} est le RSB en sortie du beamformer gauche. Il doit être mentionné que le filtre w_{WF} est fonction du temps et que rien n'assure que le même soit appliqué aux deux oreilles. Par ailleurs, si les filtres gauche et droit ne sont pas égaux, on va augmenter les distorsions d'ILD et de IC. Pour éviter cela, les auteurs [Marquardt and Doclo, 2018] ont proposé de moyenner les deux filtres dans le domaine des dB, *i.e.* appliquer la moyenne géométrique dans le domaine linéaire :

$$\bar{w}_{\text{WF}} = \sqrt{w_{\text{WF, L}}\ w_{\text{WF, R}}}. \tag{2.105}$$

Alors, le filtre obtenu est appliqué à la sortie du beamformer MVDR-IC. Le filtre résultant de la mise en série de ces deux filtrages est alors appelé le beamformer MVDR préservant l'IC post-filtré (MVDR-ICP) :

$$\mathbf{w}_{\text{MVDR-ICP}} = \bar{w}_{\text{WF}}\ \mathbf{w}_{\text{MVDR-IC}}. \tag{2.106}$$

Les résultats montrent une amélioration du RSB de 6 dB comparé au beamformer MVDR-IC sans trop affecter la différence de MSC (de l'ordre de 0,03) [Marquardt and Doclo, 2018].

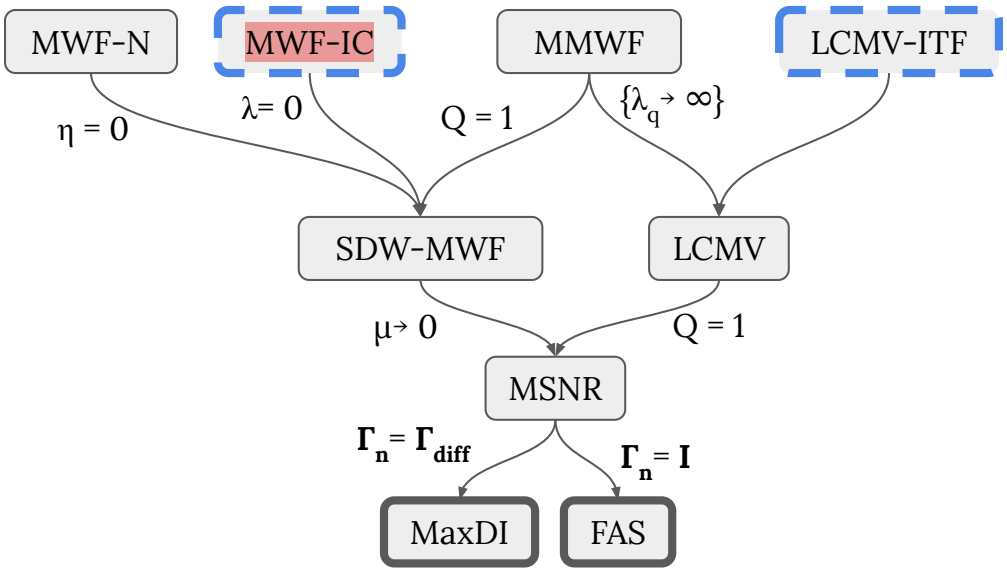


FIGURE 2.8 – Les algorithmes de beamforming présentés sous forme de réseau selon leurs liens. Ceux pour lesquels il n'existe pas de solution analytique sont surlignés en rouge, ceux qui nécessitent une optimisation conjointe (binaurale) sont encerclés en pointillé bleu.

2.3 Estimation des fonctions de transfert acoustiques

Dans la section précédente, nous avons passé en revue les algorithmes de beamforming conçus pour les prothèses auditives. Tous nécessitent de connaître la ou les ATFs (ou au moins la ou les fonctions de transfert relative (de l'anglais *Relative Transfer Function*s, RTFs)) de la/les source(s) cible(s).

Dans cette section, nous présentons les différentes techniques pour estimer ces quantités. Celles-ci peuvent être séparées en deux catégories : d'une part, les techniques basées sur les signaux d'entrée $\mathbf{x}(k, \ell)$ pour lesquels l'estimation se fait donc en temps-réel (mais ne considère qu'une source cible) ; et d'autre part, les techniques consistant à déterminer ces filtres en amont pour toutes les directions et de sélectionner le moment venu celui qui correspond à la/les source(s) que l'on souhaite viser.

2.3.1 Estimation en temps-réel

Modèle des signaux

Rappelons le modèle des signaux dans le domaine de la TFCT pour un scénario composé d'une source ponctuelle de parole et une source de bruit :

$$\mathbf{x}(k,\ell) = \mathbf{h}(k)s(k,\ell) + \mathbf{n}(k,\ell). \tag{2.107}$$

Les algorithmes présentés se basent sur deux hypothèses importantes :
- la source de parole n'est pas tout le temps active, la source de bruit si (voir Eq. (2.17) et (2.18) p.36) ;
- la matrice de covariance de la source est de rang 1, *i.e.* peu de réverbération.

L'indépendance statistique de la source et du bruit nous permet d'exprimer la matrice de covariance des données sous la forme :

$$\mathbf{\Phi}_{\mathbf{x}}(k,\ell) = \underbrace{\mathbf{h}(k)\mathbf{h}(k)^H\phi_s(k,\ell)}_{\mathbf{\Phi}_{\mathbf{s}}(k,\ell)} + \mathbf{\Phi}_{\mathbf{n}}(k,\ell). \tag{2.108}$$

En utilisant l'hypothèse de parcimonie du signal de parole dans le domaine de la TFCT, on peut estimer la matrice de covariance du bruit, notée $\hat{\mathbf{\Phi}}_{\mathbf{n}}$, lors des phases d'absence de la parole et la matrice de covariance du mélange, $\hat{\mathbf{\Phi}}_{\mathbf{x}}$, lors de sa présence :

$$\hat{\mathbf{\Phi}}_{\mathbf{x}}(k,\ell) = \frac{1}{L_{\mathcal{H}_1}} \sum_{\ell \backslash \mathcal{H}_1} \mathbf{x}(k,\ell)\mathbf{x}(k,\ell)^H \tag{2.109}$$

$$\hat{\mathbf{\Phi}}_{\mathbf{n}}(k,\ell) = \frac{1}{L_{\mathcal{H}_0}} \sum_{\ell \backslash \mathcal{H}_0} \mathbf{n}(k,\ell)\mathbf{n}(k,\ell)^H, \tag{2.110}$$

où $\mathcal{H}_0$ et $\mathcal{H}_1$ font respectivement référence aux hypothèses d'absence et de présence de la parole au point temps-fréquence (k,ℓ), $L_{\mathcal{H}_1}$ est le nombre de trames temporelles considérées dans le calcul de la moyenne pour laquelle la $\mathcal{H}_1$ est valide. $L_{\mathcal{H}_1}$ est définie similairement.

Soustraction des matrices de covariance

La fonction de transfert relative (de l'anglais *Relative Transfer Function*, RTF), notée $\tilde{\mathbf{h}}$, est définie comme l'ATF normalisée par rapport à un microphone de référence :

$$\tilde{\mathbf{h}}(k) = \frac{\mathbf{h}(k)}{h_r(k)}, \tag{2.111}$$

où $h_r(k)$ est le $r^{\text{ème}}$ élément de $\mathbf{h}(k)$ correspondant au microphone de référence. On peut alors définir son estimation, notée $\hat{\tilde{\mathbf{h}}}_{\text{CS}}$, à partir de l'estimation de la matrice de covariance de la source dans l'espace de microphones :

$$\hat{\mathbf{\Phi}}_{\mathbf{s}}(k,\ell) = \hat{\mathbf{\Phi}}_{\mathbf{x}}(k,\ell) - \hat{\mathbf{\Phi}}_{\mathbf{n}}(k,\ell), \tag{2.112}$$

où

$$\hat{\mathbf{\Phi}}_{\mathbf{s}}(k,\ell) \approx \mathbf{h}(k)\mathbf{h}(k)^H \phi_s(k,\ell). \tag{2.113}$$

Pour ce faire, on vient sélectionner la $r^{\text{ème}}$ colonne de $\hat{\mathbf{\Phi}}_{\mathbf{s}}(k,\ell)$ grâce au vecteur $\mathbf{q}_r = [0,...,1,...,0]^T$ composé de 0 à l'exception d'un 1 placé à la $r^{\text{ème}}$ ligne. Il faut ensuite normaliser le vecteur résultant par le $r^{\text{ème}}$ élément de la diagonal de $\hat{\mathbf{\Phi}}_{\mathbf{s}}(k,\ell)$ [Talmon et al., 2009, Serizel et al., 2014] :

$$\hat{\tilde{\mathbf{h}}}_{\text{CS}}(k,\ell) = \frac{\hat{\mathbf{\Phi}}_{\mathbf{s}}(k,\ell)\mathbf{q}_r}{\mathbf{q}_r^H \hat{\mathbf{\Phi}}_{\mathbf{s}}(k,\ell)\mathbf{q}_r}. \tag{2.114}$$

Blanchiment de la matrice de covariance du bruit

Cette méthode consiste à utiliser la propriété de symétrie hermitienne de la matrice de covariance du bruit. L'estimateur de RTF qui en découle, noté $\hat{\tilde{\mathbf{h}}}_{\text{CW}}(k,\ell)$, peut s'écrire comme suit [Serizel et al., 2014, Markovich et al., 2009] :

$$\hat{\tilde{\mathbf{h}}}_{\text{CW}}(k,\ell) = \frac{\hat{\mathbf{\Phi}}_{\mathbf{n}}^{\frac{1}{2}}(k,\ell)\hat{\mathbf{v}}_1(k,\ell)}{\mathbf{q}_r^H \hat{\mathbf{\Phi}}_{\mathbf{n}}^{\frac{1}{2}}(k,\ell)\hat{\mathbf{v}}_1(k,\ell)}, \tag{2.115}$$

où $\hat{\mathbf{\Phi}}_{\mathbf{n}}^{\frac{1}{2}}(k,\ell)$ est la décomposition de Cholesky de la matrice de covariance du bruit et $\hat{\mathbf{v}}_1(k,\ell)$ est le vecteur propre associé à la plus grande valeur propre de $\mathbf{\Phi}_{\mathbf{b}}(k,\ell)$, définie comme suit :

$$\mathbf{\Phi}_{\mathbf{b}}(k,\ell) = \hat{\mathbf{\Phi}}_{\mathbf{n}}^{-\frac{1}{2}}(k,\ell)\hat{\mathbf{\Phi}}_{\mathbf{x}}(k,\ell)\left(\hat{\mathbf{\Phi}}_{\mathbf{n}}^{-\frac{1}{2}}(k,\ell)\right)^H. \tag{2.116}$$

Plus de détails sur le fondement de cette méthode sont disponibles en annexe A.

[Markovich-Golan and Gannot, 2015] ont montré que cette méthode a de meilleures performances que la méthode dite de soustraction des matrices de covariance, quelque soit la direction d'arrivée de la source, du niveau d'interférence, de la plage d'estimation ($L_{\mathcal{H}_0}$ et $L_{\mathcal{H}_1}$), ainsi que du nombre de microphones. Néanmoins, elle est plus coûteuse en calcul car elle nécessite le calcul de la décomposition de Cholesky de la matrice de covariance du bruit ainsi que de réaliser la décomposition en valeurs propres de la matrice de covariance du vecteur signal *blanchi*.

2.3.2 Mesure en amont

Nous avons vu deux algorithmes permettant d'estimer la RTF de la source cible dans un scénario où un seul locuteur est présent. Cependant, dans un scénario multi-locuteurs ou pour les algorithmes d'estimation de DOA (que nous présenterons dans la suite), il est nécessaire de connaitre *a priori* les ATFs d'un grand nombre de direction, ce qui ne peut être obtenu par les méthodes présentées précédemment. Alors, celles-ci peuvent être mesurées à l'avance sur le sujet lui-même ou sur un mannequin en laboratoire [Kayser et al., 2009, Oreinos and Buchholz, 2013, Moore et al., 2019a]. La mesure d'un grand nombre d'ATFs est très contraignante en pratique, elle nécessite de se placer en chambre anéchoïque avec un système d'acquisition calibré et un temps important de sorte à mesurer les ATFs dans un maximum de directions (entre 72 et 2000). Pour cette raison, il n'est pas possible de généraliser la mesure d'ATF à tous les malentendant•e•s appareillé•e•s et il est en général préféré d'utiliser un mannequin pour une mesure la plus fiable possible (pas de mouvement de tête ou autre bruit parasite lors de la mesure comme des bruits de respiration ou de déglutitions, rappelons que les microphones sont placés dans/sur les oreilles).

Cependant, cette estimation, consistant en une mesure en amont de l'utilisation, peut contenir des erreurs dues à une variation du positionnement de l'appareil sur l'oreille à l'usage ou à une trop grande différence morphologique entre le mannequin servant à la mesure et l'auditeur. Il a été montré que l'individualisation des ATFs est importante pour les algorithmes de débruitage binauraux avec des prothèses derrière l'oreille (de l'anglais *Behind-The-Ear*, BTE) [Moore et al., 2019a] et monauraux avec des prothèses dans l'oreille (de l'anglais *In-The-Ear*, ITE) [Harder, 2015]. Dans le premier cas, lorsque les microphones sont placés de part et d'autre de la tête, on peut supposer que la différence est due à la variance de la largeur de la tête qui pourrait entrainer un déphasage entre les signaux des microphones. Cette hypothèse est confortée par la diminution des performances de débruitage lorsque la cible est située sur le coté de l'auditeur, où le déphasage est maximal. Dans le second cas, lorsque les microphones sont placés à l'intérieur de l'oreille, c'est l'influence du pavillon de l'oreille, très différent d'un individus à l'autre, qui modifie fortement les indices spectraux des ATFs.

Pour l'estimation de DOA, la nécessité d'individualiser les ATFs semble moins critique [Zohourian et al., 2017, Zohourian et al., 2018] bien que le niveau de preuve est faible (des résultats sous forme de moyenne, sans écart-type, pour un seul sujet).

2.4 Estimation de la localisation des sources

Dans cette section nous présentons les algorithmes principaux utilisés dans le cadre des prothèses auditives pour estimer la DOA dans un scénario composé d'une seule source cible par point T-F. Rappelons le modèle de mélange de signaux dans les microphones dans le domaine de la TFCT :

$$\mathbf{x}(k,\ell) = \mathbf{h}(k)s(k,\ell) + \mathbf{n}(k,\ell). \tag{2.117}$$

Dans ce cadre applicatif, nous limitons la recherche de la direction d'arrivée à son azimut et supposons que son élévation est de 0°, *i.e.* la source est située sur le plan horizontal.

L'estimation de DOA peut être le fruit de plusieurs stratégies d'optimisation. Le problème général consiste à trouver l'angle $\hat{\theta}(k,\ell)$ qui maximise une certaine fonction-objectif, notée $J(\theta,k,\ell)$:

$$\hat{\theta}(k,\ell) = \underset{\theta}{\operatorname{argmax}} \left\{ J(\theta,k,\ell) \right\}. \tag{2.118}$$

Ce problème est résolu par une recherche exhaustive. Dans la suite, nous omettons les indices k et ℓ par soucis de brièveté sauf si nécessaire.

Par soucis de cohérence, nous n'abordons ici que les algorithmes optimaux au sens du maximum de vraisemblance fondés sur des modèles semblables à ceux utilisés pour le beamforming dans la section 2.2. Néanmoins, il existe un certain de nombre de travaux proposant des algorithmes d'estimation de DOA moins théoriquement fondés et dont les performances ne sont pas meilleures, la•e lecteur•rice pourra se référer aux travaux suivant sur ce sujet [Goetze et al., 2007, Boyd et al., 2013, Braun et al., 2015, Archer-Boyd et al., 2015, Thiergart et al., 2016, Zohourian et al., 2018].

2.4.1 Cible déterministe et bruit aléatoire

Supposons que la source s est une variable complexe déterministe et que le bruit $\mathbf{n}$ est une variable aléatoire suivant une distribution Gaussienne multivariée centrée isotropique complexe de matrice de covariance $\mathbf{\Phi_n}$. Le modèle de vraisemblance des données peut s'écrire comme suit :

$$P(\mathbf{x}|\theta,s,\mathbf{\Phi_n}) = \frac{1}{\pi^M |\mathbf{\Phi_n}|} e^{-(\mathbf{x}-\mathbf{h}(\theta)s)^H \mathbf{\Phi_n}^{-1}(\mathbf{x}-\mathbf{h}(\theta)s)}, \tag{2.119}$$

avec $|.|$ le déterminant et $\mathbf{h}(\theta)$ les ATFs correspondant à une source située à l'azimut θ. La fonction de coût associée, au sens du maximum de vraisemblance, peut s'écrire alors comme suit :

$$J_{\text{DML}}(\theta) = \frac{\mathbf{h}(\theta)^H \mathbf{\Gamma_n}^{-1} \hat{\mathbf{\Phi}}_\mathbf{x} \mathbf{\Gamma_n}^{-1} \mathbf{h}(\theta)}{\mathbf{h}(\theta)^H \mathbf{\Gamma_n}^{-1} \mathbf{h}(\theta)}, \tag{2.120}$$

avec

$$\hat{\mathbf{\Phi}}_{\mathbf{x}} = \frac{1}{L}\sum_{\lambda=1}^{L}\mathbf{x}_\lambda\mathbf{x}_\lambda^H. \tag{2.121}$$

Les détails de calcul sont disponibles en annexe B.

On remarque que dans le cas où $\mathbf{\Gamma_n} = \mathbf{I}$, $J_{\text{DML}}(\theta)$ devient $J_{\text{SRP}}(\theta)$ la fonction de coût associée à l'algorithme puissance de la réponse visée (de l'anglais *Steered Response Power*, SRP) [Brandstein et al., 2001, Chapt. 8], un algorithme largement utilisé pour la localisation sonore en général :

$$\begin{aligned} J_{\text{SRP}}(\theta) &= \frac{\mathbf{h}(\theta)^H\hat{\mathbf{\Phi}}_{\mathbf{x}}\mathbf{h}(\theta)}{||\mathbf{h}(\theta)||^2} \\ &= \frac{1}{L}\sum_{\lambda=1}^{L}\left|\frac{\mathbf{h}(\theta)^H}{||\mathbf{h}(\theta)||}\mathbf{x}_\lambda\right|^2 \\ &= \frac{1}{L}\sum_{\lambda=1}^{L}\left|\mathbf{w}_{\text{SRP}}^H\mathbf{x}_\lambda\right|^2, \end{aligned} \tag{2.122}$$

où

$$\mathbf{w}_{\text{SRP}} = \frac{\mathbf{h}(\theta)}{||\mathbf{h}(\theta)||}, \tag{2.123}$$

est appelé le *match-filter*, très proche de $\mathbf{w}_0$, le filtre du beamformer aligneur défini en Eq. (2.37) (p.42). D'après [Zohourian et al., 2018], l'usage de l'estimation imparfaite [10] de $\mathbf{\Gamma_n}$ plutôt que de prendre naïvement $\mathbf{I}$ entraine des erreurs d'estimation qui ne compensent le gain que pourrait apporter une plus grande précision du modèle.

2.4.2 Cible stochastique

L'estimateur DML, présenté précédemment, suppose le coefficient de Fourier de la parole déterministe et ceux du bruit aléatoires. Supposons maintenant que les deux sont des variables aléatoires, leur somme suit donc aussi une distribution Gaussienne. Alors, le modèle de vraisemblance des données peut s'écrire comme suit :

$$P(\mathbf{x}|\theta, \mathbf{\Phi}_{\mathbf{x}}) = \frac{1}{\pi^M|\mathbf{\Phi}_{\mathbf{x}}|}e^{-\mathbf{x}^H\mathbf{\Phi}_{\mathbf{x}}^{-1}\mathbf{x}}, \tag{2.124}$$

où $|.|$ désigne le déterminant et $\mathbf{\Phi}_{\mathbf{x}}$ la matrice de covariance des signaux des microphones. L'estimateur associé à ce modèle est nommé l'estimateur maximisant la vraisemblance du modèle stochastique (ou *Stochastic Maximum Likelihood*) (SML). On obtient alors la fonction de coût associée, au sens du

[10] avec la méthode de [Lotter and Vary, 2006].

maximum de vraisemblance, notée $J_{\mathrm{SML}}(\theta)$:

$$J_{\mathrm{SML}}(\theta) = -\log \big|\mathbf{h}(\theta)\mathbf{w}_{\mathrm{MVDR}}(\theta)^H \hat{\mathbf{\Phi}}_{\mathbf{x}} \mathbf{w}_{\mathrm{MVDR}}(\theta)\mathbf{h}(\theta)^H + \phi_n(\mathbf{I} - \mathbf{h}(\theta)\mathbf{w}_{\mathrm{MVDR}}(\theta)^H)\mathbf{\Gamma}_{\mathbf{n}}\big|, \tag{2.125}$$

où $\mathbf{w}_{\mathrm{MVDR}}(\theta)$ est le filtre du beamformer MVDR visant la direction d'azimut θ sur le plan horizontal, défini en Eq. (2.42, p.45), et $\hat{\mathbf{\Phi}}_{\mathbf{x}}$ est l'estimation de la matrice de covariance des microphones tel que défini en Eq. (2.121). Plus de détails sont disponibles en annexe C.

Cette approche a été développée par [Ye and DeGroat, 1995] et évaluée dans le contexte des prothèses auditives par [Zohourian et al., 2018]. Elle se montre sensible aux erreurs d'estimation des ATFs et de la matrice de covariance du bruit. Cependant, elle est particulièrement efficace pour estimer la DOA sur l'ensemble de la trame en prenant le maximum des estimations sur toutes les fréquences. Enfin, il faut noter que le calcul du déterminant dans la fonction de coût rend cette méthode particulièrement lourde en calcul par rapport aux autres [Zohourian et al., 2018].

2.5 Interaction des étages de réduction du bruit et compensation de niveau sonore

Nous avons montré en section 2.1 (p.28) que le compresseur de dynamique présente des difficultés à remplir son rôle en présence de bruit et de réverbération. Pour répondre à ce problème, l'ajout d'un algorithme de débruitage, souvent multicanal, est placé en amont de sorte à « nettoyer » le signal utile du bruit ou de la réverbération tardive. Néanmoins, ces deux briques, conçues séparément, répondent à des objectifs antagonistes. En effet, tandis que l'étage de débruitage cherche à augmenter le RSB, il a été montré que le compresseur fait exactement l'inverse en atténuant les sons fort (dominé par le signal) par rapport aux sons faibles (dominé par le bruit) [Hagerman and Olofsson, 2004, Souza et al., 2006, Naylor and Johannesson, 2009, Corey, 2019]. Par ailleurs, les deux sont connus pour détériorer les performances de localisation [Keidser et al., 2006, Van den Bogaert et al., 2006, Wiggins and Seeber, 2011, Hassager et al., 2017b] et la capacité à s'orienter vers un nouveau locuteur [Schwartz and Shinn-Cunningham, 2013, Brimijoin et al., 2014]. Pour ces raisons, il est raisonnable de penser que la concaténation sérielle des deux traitements n'est pas optimale pour maximiser la qualité de l'écoute selon les critères précédemment édictés.

Dans cette section, nous discutons les phénomènes observés dus à l'interaction entre l'étage de débruitage et de compression lorsqu'ils sont combinés en

série. Puis, nous présentons et discutons les autres manières de combiner ces deux familles de traitement proposées dans la littérature.

Cette question reste relativement peu traitée, alors contrairement aux sections précédentes de ce chapitre, nous tentons ici une revue exhaustive des travaux qui lui ont été dédiées. En outre, il nous a semblé intéressant d'en considérer l'aspect sociologique. En effet, nous avons pu distinguer les travaux effectués par des équipes expertes en traitement du signal de ceux menés par des équipes expertes en psychoacoustique, dite « clinicienne », sur la base des affiliations institutionnelles et des cultures épistémiques [Knorr-Cetina, 2003]. Ces deux communautés peuvent se différencier dans leurs méthodes par deux aspects : d'une part, les traiteurs de signaux ont tendance à plus fonder théoriquement leur algorithmes que leurs homologues cliniciens tandis que ces derniers privilégient une évaluation plus poussée, notamment sur des patients malentendants, ce qui demande une expertise à part entière. Nous avons résumé ceci en Fig. 2.9 sous forme de graphe chronologique où chaque connexion correspond à une citation. Nous pouvons alors remarquer que la communauté clinicienne s'est plus emparée du sujet avec la production de quatre articles de revue tandis que la communauté du traitement du signal a produit deux articles de conférence et un article dans une revue. Il est notable que [Corey and Singer, 2017] n'a fait référence à aucun des travaux précédents que nous avons pu identifier.

2.5.1 Combinaison sérielle

Dans le cas d'une prothèse auditive disposant d'un seul microphone et donc d'un algorithme de débruitage monocanal, on peut décrire la sortie, notée $y(k,\ell)$, comme le résultat du produit d'un gain, noté $g_{\mathrm{s}}(k,\ell)$, et de l'entrée, notée $x(k,\ell)$. Dans le cas d'un combinaison sérielle des traitements de débruitage et de compression, on peut écrire :

$$g_{\mathrm{s}}(k,\ell) = w_{\mathrm{WF}}(k,\ell)\; g_{\mathrm{CD}}(k,\ell), \tag{2.126}$$

où $w_{\mathrm{WF}}(k,\ell)$ est le gain du filtre de Wiener calculé à partir du signal d'entrée et $g_{\mathrm{CD}}(k,\ell)$ le gain de compression calculé à partir de la sortie du filtre de Wiener, comme illustré en Fig. 2.10 (a).

Il a été montré que pour un test de préférence, le débruitage seul était préféré à la mise en série d'un débruiteur et d'un compresseur [Anderson et al., 2009]. Dans une étude plus récente, [Brons et al., 2015] évaluent quatre modèles de prothèses auditives en activant ou pas la compression et/ou la fonction de débruitage.[11] Ils évaluent alors, pour un RSB assez élevé (4 dB), l'intelligibilité

[11] Il faut noter que nous n'avons pas le détail des algorithmes utilisés dans chacune d'entre elle, sinon que le débruiteur repose sur un filtre de Wiener.

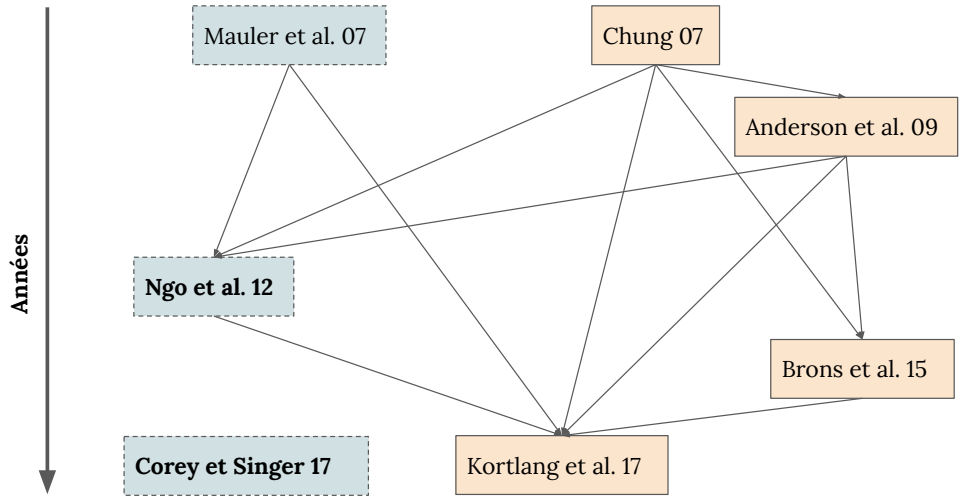


FIGURE 2.9 – Graphe des connections par citations entre les articles de recherche proposant et/ou évaluant différentes combinaisons d'étage de réduction de bruit et de compression de dynamique. Les boites bleues avec un contour pointillé désignent les articles issus de la communauté du traitement du signal et les boites orange avec un contour en trait plein désignent ceux issus de la communauté clinicienne des concepteurs de prothèses auditives. Les références en gras désignent les travaux incluant du beamforming tandis que les autres ne considèrent que des algorithmes monocanaux.

de la parole, l'inconfort dû au bruit, la distorsion de la parole et la préférence globale. Les résultats ne révèlent pas de préférence significativement plus élevée pour le son traité par la prothèse auditive (débruitage et compression) par rapport au son non traité, ce qu'ils attribuent à la réduction du RSB par la compression. Aussi, ils trouvent que le classement des algorithmes en fonction du niveau de bruit résiduel et d'inconfort dû au bruit sont les mêmes, suggérant ainsi que la mesure du premier permet de renseigner sur le deuxième, en première approximation. Concernant l'intelligibilité de la parole, ils n'observent pas de variations significatives ce qui peut être dû aux conditions expérimentales (un RSB élevé) et au fait que le filtre de Wiener est connu pour ne pas l'améliorer [McCreery et al., 2012a, Völker et al., 2015, Chong and Jenstad, 2018]. Enfin, malgré le coté « boîtes noires » des combinaisons d'algorithmes de débruitage et de compression dans des prothèses auditives commercialisées, [Brons et al., 2015] pointent des disparités de préférence pour tel ou tel appareils, ce qu'ils interprètent comme une nécessité de co-concevoir ces deux briques essentielles de traitement dans les prothèses auditives.

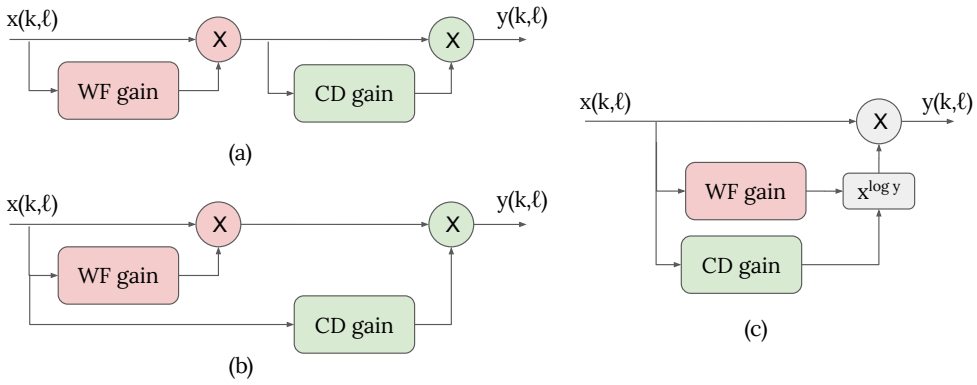


FIGURE 2.10 – Schémas blocs des combinaisons de débruitage et de compression testées dans [Kortlang et al., 2017] : (a) sérielle, (b) parallèle et (c) multiplicative.

2.5.2 Autres tentatives de combinaisons

Quelques propositions ont été faites pour combiner débruitage et compression autrement qu'en série. Nous traitons indépendamment les propositions considérant des algorithmes de débruitage monocanaux et multicanaux.

Débruitage monocanal

Le travail liminaire sur le sujet est produit par [Mauler et al., 2007]. C'est une étude avant tout théorique, limitée au cas monocanal. Les auteurs y développent la solution optimale de l'estimateur de l'amplitude logarithmique de la source compressée au sens des moindres carrés [Ephraim and Malah, 1985]. En reprenant les notations de la section 2.2, le problème d'optimisation s'écrit comme suit :

$$w_{\text{CLSA-MMSE}} = \underset{w}{\operatorname{argmin}} \left\{ \mathbb{E}\left[\left| \log |t^{1-\frac{1}{R}} s^{\frac{1}{R}}| - \log |wx| \right|^2 \middle| x \right] \right\}, \tag{2.127}$$

avec t et R le seuil et le ratio du compresseur, respectivement, voir Eq. (2.5) (p.30). Ils montrent que la solution optimale, notée $w_{\text{CLSA-MMSE}}$, correspond à la concaténation sérielle du débruiteur et du compresseur. Ils développent aussi l'estimateur de l'amplitude de la source compressée au sens des moindres carrés :

$$w_{\text{CSA-MMSE}} = \underset{w}{\operatorname{argmin}} \left\{ \mathbb{E}\left[\left| |t^{1-\frac{1}{R}} s^{\frac{1}{R}}| - |wx| \right|^2 \middle| x \right] \right\}, \tag{2.128}$$

dont la solution, notée $w_{\text{CSA-MMSE}}$, diffère de la concaténation sérielle mais dont la différence avec celle-ci est trop petite pour être perceptible. Enfin, ils développent l'estimateur de l'amplitude logarithmique de la source compressée au sens du maximum *a posteriori* :

$$|\hat{s}|_{\text{CSA-MAP}} = \underset{|\check{s}|}{\operatorname{argmax}} \left\{ p\left(|\check{s}| \middle| x\right) \right\}, \tag{2.129}$$

où $|\check{s}|$ désigne l'amplitude idéale du signal compressé et $p(|\check{s}||x)$ sa fonction de densité de probabilité. Cette solution diffère aussi de celle obtenue par concaténation sérielle des deux traitements. Néanmoins, les auteurs rapportent par des tests d'écoute informels que la différence est petite et en général non-perceptible lorsqu'elle est appliquée à un signal réel.

De manière moins théoriquement fondée, il a été proposé de combiner les algorithmes de débruitage et de compression en parallèle en sommant leur gain [12] résultant en dB [Chung, 2007, Anderson et al., 2009, Kortlang et al., 2017, Chen et al., 2021]. De sorte à être plus compatible avec notre cadre d'analyse, nous proposons de le présenter dans le domaine linéaire :

$$g_{\text{p}}(k,\ell) = w_{\text{WF}}(k,\ell)\ \tilde{g}_{\text{CD}}(k,\ell), \tag{2.130}$$

où $w_{\text{WF}}(k,\ell)$ et $\tilde{g}_{\text{CD}}(k,\ell)$ sont les gains du filtre de Wiener et du compresseur calculé à partir du signal bruité d'entrée. Cette configuration est illustrée en Fig. 2.10 (b). Nous trouvons que la présentation de la recombinaison des gains de débruitage et de compression en dB est trompeuse. En effet, la combinaison présentée comme parallèle revient uniquement à calculer le gain de compression à partir du signal bruité plutôt que débruité. L'application des gains au signal d'origine est, quant à elle, la même, comme on peut le voir en Eq. (2.126) et (2.130) ainsi qu'en Fig. 2.10 (a et b) mais ce qui est loin d'être explicite dans [Kortlang et al., 2017]. Présenté ainsi, on ne comprend pas spécialement la motivation derrière ce choix. En matière d'intelligibilité de la parole, certains trouvent des performances égales à celles de la combinaison sérielle [Chung, 2007, Kortlang et al., 2017] quand d'autres trouvent une diminution de celles-ci [Anderson et al., 2009]. De même, pour un test de préférence subjectif, certains trouvent une amélioration par rapport à la combinaison sérielle [Chung, 2007, Kortlang et al., 2017] quand d'autres [Anderson et al., 2009] trouvent des résultats équivalent à la combinaison sérielle et significativement moins bons par rapport à l'utilisation du débruiteur seul. Récemment, [Chen et al., 2021] ont choisi d'étendre le travail de [Kortlang et al., 2017] en ajoutant un compresseur double (rapide et lent en parallèle en dB), introduit par [Moore and Glasberg, 1988], à la combinaison parallèle et rapportent une amélioration de la préférence subjective par rapport à l'algorithme original.

[12] La manière de recombiner les gains n'est pas claire dans [Chung, 2007].

[Kortlang et al., 2017] proposent aussi une combinaison mutliplicative en dB, dont le gain total a appliqué au signal bruité, noté $g_{\mathrm{m}}(k,\ell)$, s'exprime dans le domaine linéaire comme suit :

$$g_{\mathrm{m}}(k,\ell) = \tilde{g}_{\mathrm{CD}}(k,\ell)^{\log_{10} w_{\mathrm{WF}}(k,\ell)}, \tag{2.131}$$

et est illustré en Fig. 2.10 (c). Remarquons qu'en Eq. (2.131), $\tilde{g}_{\mathrm{CD}}(k,\ell)$ et $w_{\mathrm{WF}}(k,\ell)$ peuvent être échangés sans conséquence sur le résultat et que la base du logarithme, quant à elle, est importante. Cette architecture présente des performances similaires aux autres méthodes en compréhension de la parole et semble un peu plus préférée chez les malentendants en matière de qualité subjective, sans pour autant atteindre un résultat statistiquement significatif.

Les algorithmes de débruitage basés sur un filtre de Wiener étant connus pour être inefficaces dans l'amélioration de l'intelligibilité de la parole, on comprend le caractère mitigé des résultats des études présentées ici. Pour un effet mesurable, il faut se tourner vers des algorithmes de débruitage multicanaux.

Débruitage multicanal

A notre connaissance, il n'existe que deux études proposant des combinaisons du beamformer et du compresseur autres que sérielle. La première [Ngo et al., 2012] considère un scénario composé d'une seule source d'intérêt tandis que la seconde [Corey and Singer, 2017] est plus souple et peut comprendre plusieurs locuteurs dans la scène sonore.

De sorte à améliorer le RSB en sortie sans trop dégrader la distorsion de la parole, [Ngo et al., 2012] ont proposé de compresser indépendamment l'estimation de la source de parole et du bruit. Ainsi, on peut débruiter le plus agressivement possible grâce à un MWF paramétrique basé sur la probabilité de présence de la parole (PPP) et on n'est plus obligé de procéder à un compromis entre débruitage et préservation du bruit dans l'estimation de la parole comme dans tous les algorithmes présentés en section 2.2.4 (p.51). Nous ne détaillons pas ici les calculs de cet algorithme mais nous l'illustrons sous forme de schéma bloc en Fig. 2.11 (c). Cependant, nous devons tout de même aborder l'originalité de celui-ci qui est le traitement de l'estimation de la parole par un CD différent suivant si la parole est estimée comme étant présente ou absente. En effet, un compresseur différent est appliqué à la source suivant si elle est estimée comme étant présente ou absente au point temps-fréquence donné, et si elle est présente ou absente dans la trame temporelle en général. Les gains de compression à appliquer à la source sont sommés dans le domaine des dB en pondérant par la probabilité de présence de la parole par point T-F, noté

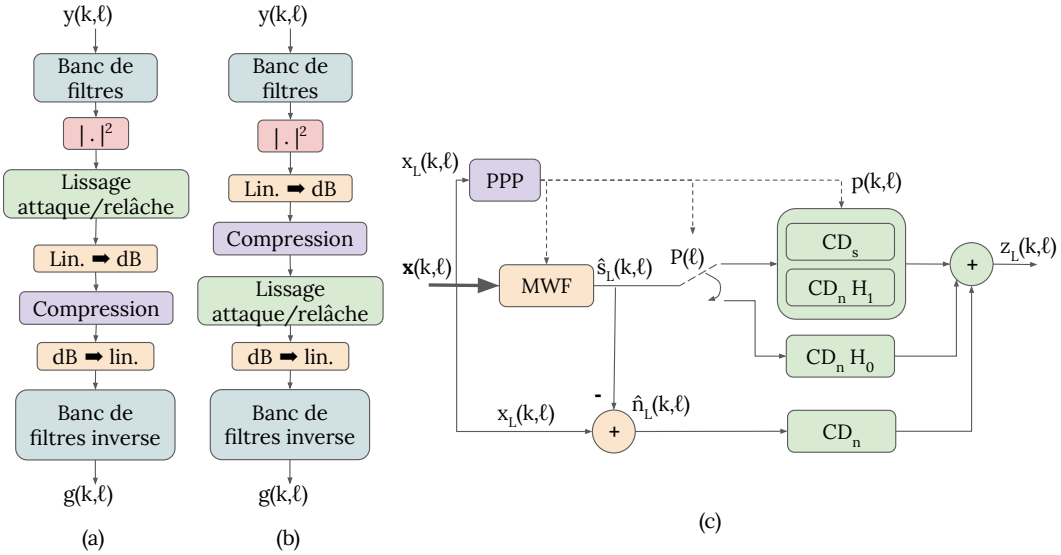


FIGURE 2.11 – Schéma bloc du calcul du gain de compression pour un compresseur état de l'art dans les prothèses auditives (a) [Hassager et al., 2017b] et chez [Ngo et al., 2012]. En (c), le schéma bloc de l'algorithme combinant algorithme de débruitage et compresseurs proposé par [Ngo et al., 2012].

$p(k,\ell)$, et sur toute la trame, noté $P(\ell)$:

$$\begin{aligned} G_{\text{Flex}}(k,\ell) = P(\ell)\,[p(k,\ell)G_s(k,\ell) + (1-p(k,\ell))G_{n,H_1}(k,\ell)] \\ + (1-P(\ell))G_{n,H_0}(k,\ell), \end{aligned} \tag{2.132}$$

où $G_{\text{Flex}}(k,\ell)$ est le gain en dB à appliquer à l'estimation de la parole, $G_s(k,\ell)$, $G_{n,H_1}(k,\ell)$ et $G_{n,H_0}(k,\ell)$ ceux destinés à la parole lorsqu'elle est, respectivement, présente à ce point T-F, absente à ce point T-F mais présente à une autre fréquence, et enfin, absente sur toute la trame. Ensuite, le gain est lissé dans le domaine des dB par le filtre récursif avec les constantes d'attaque et de relâche. Il faut noter que cette organisation des étapes dans le compresseur n'est pas standard dans la littérature des prothèses auditives [Hassager et al., 2017b, May et al., 2018], nous avons mis en évidence ces différences en Fig. 2.11 (a et b). En particulier, le lissage placé ici joue aussi bien le rôle de lissage dans la détection d'enveloppe que de lissage de l'estimation de PPP, deux paramètres dont les réglages sont pourtant antagonistes.

On peut identifier d'autres limitations ou perspectives d'amélioration de cette architecture : (i) tout d'abord, l'algorithme n'est pas mis en regard avec le MWF-N alors qu'il y ressemble beaucoup, avec son estimation du bruit. Ainsi, on pourrait envisager d'évaluer l'influence sur la localisation auditive et la pré-

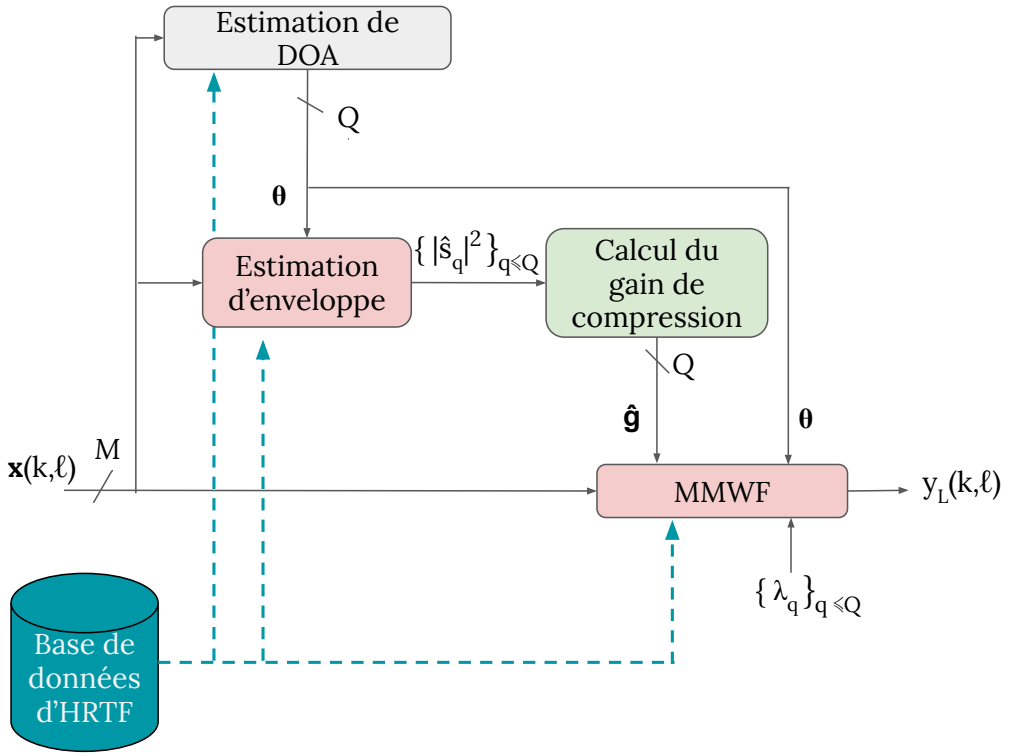


FIGURE 2.12 – Schéma bloc de l'algorithme combinant beamforming et compression pour plusieurs locuteurs proposé par [Corey and Singer, 2017].

servation des indices de localisation par l'ensemble de la chaine, ce qui n'a pas été fait dans l'étude d'origine. (ii), par ailleurs, le réseau de microphone utilisé ne comporte que deux microphones et n'est pas binaural, entrainant un gain de débruitage très limité. Celui-ci pourrait facilement être amélioré en considérant un réseau de microphone binaural de quatre [Oreinos and Buchholz, 2013], voire six microphones [Kayser et al., 2009]. Enfin, l'estimateur de bruit, ou nullformer, utilisé minimise la puissance de l'erreur entre le bruit et son estimation mais ne considère pas les indices de localisation du bruit explicitement dans sa fonction de coût. On peut imaginer la conception d'un nullformer qui les considérerait, à l'image des beamformers LCMV-ITC et MWF-IC par rapport au MWF-N présentés en section 2.2.4 (p.51). En effet, un débruitage parfait n'est pas souhaitable pour préserver les informations de localisation présentes hors-axe et pour pouvoir rester disponible à l'arrivée d'un nouvel interlocuteur ou un signal d'alerte [Brimijoin et al., 2014].

Plus récemment, [Corey and Singer, 2017] ont proposé une extension à un scénario multi-locuteur, avec une approche dite « démixage/remixage », toutefois sans faire référence à [Ngo et al., 2012]. Leur algorithme illustré en Fig. 2.12,

que l'on nommera ici le filtre de Wiener multicanal multicibles compressées (CMMWF), est principalement basé sur le MMWF [Markovich-Golan et al., 2012b]. Commençons par définir la sortie optimale voulue, notée $y_{\text{CMMWF}}(k,\ell)$ composée des Q sources de parole spatialisées et compressées :

$$y_{\text{CMMWF}}(k,\ell) = \sum_{q=1}^{Q} h_q(k) g_q(k,\ell) s_q(k,\ell), \tag{2.133}$$

avec, pour la $q^{\text{ème}}$ source, $h_q(k)$ la réponse en fréquence incluant les indices de localisation et $g_q(k,\ell)$ le gain de compression à lui appliquer. On définit alors son estimation de la même manière qu'un beamformer :

$$\hat{y}_{\text{CMMWF}}(k,\ell) = \mathbf{w}_{\text{CMMWF}}(k,\ell)^H \mathbf{x}(k,\ell), \tag{2.134}$$

et en utilisant la même procédure que pour le MMWF (voir Eq. (2.60, p.49)), on obtient la solution suivante en omettant les indices k et ℓ par soucis de brièveté :

$$\mathbf{w}_{\text{CMMWF}} = \left(\mathbf{H}\mathbf{\Lambda}\mathbf{\Phi}_{\mathbf{s}}\mathbf{H}^H + \mathbf{\Phi}_{\mathbf{n}}\right)^{-1} \mathbf{H}\mathbf{\Lambda}\mathbf{\Phi}_{\mathbf{s}}\mathbf{d}, \tag{2.135}$$

avec $\mathbf{d} = \left[h_1^* g_1^*, ..., h_Q^* g_Q^*\right]^T$ et les notations du modèle en Eq. (2.28, p.40). Pour estimer les enveloppes de chaque source, nécessaire au calcul du gain de compression, un beamformer MVDR est utilisé en approximant la matrice de cohérence du bruit par $\mathbf{\Gamma}_{\text{diff}}$. Enfin, notons aussi que, contrairement à [Ngo et al., 2012], il n'est pas question de réinjecter une version atténuée et compressée du bruit. Il en résulte que dans le cas simple locuteur, le CMMWF redevient une combinaison sérielle d'un MWF et d'un CD.

Cette étude limite l'évaluation à des métriques objectives et utilise un réseau binaural constitué de six microphones [Kayser et al., 2009]. Par ailleurs, les paramètres de préservations des sources, λ_q, sont, soit réglés à 1, soit à $\rightarrow \infty$. Dans le deuxième cas, le CMMWF devient le beamformer LCMV avec compression indépendante (CLCMV).

Leur proposition est comparée à la version idéale du signal de sortie et au signal correspondant à la sortie du compresseur appliqué au mélange bruité. Malheureusement, le beamformer n'est pas comparé à la concaténation sérielle du MMWF et du CD (à part pour le cas où $Q = 1$). Dans un scénario composé de deux locuteurs, le CLCMV et le CMMWF obtiennent peu ou prou les mêmes performances en débruitage et compression, et ces performances sont bonnes sur tous les critères. Mais lorsqu'il y a cinq locuteurs, le CLCMV ne débruite que très peu et parvient à bien compresser les sources et maintenir la distorsion de la parole faible tandis que le CMMWF parvient à bien débruiter mais ne compresse plus très bien et augmente la distorsion de la parole. Enfin, dans un scénario composé d'un locuteur et de bruit ambiant, le CMMWF est meilleur en

débruitage que le CLCMV mais ce dernier compresse mieux et dégrade moins la distorsion de la parole. En effet, dans ce scénario, le CMMWF compresse les sources de parole autant que le CD appliqué au mélange bruité.

2.6 Conclusion

Dans ce chapitre, nous avons étudié les deux principaux traitements présent dans les prothèses auditives : compression et débruitage. En premier lieu, le compresseur de dynamique a pour objectif de ramener les sons faibles dans la plage d'audibilité de l'auditeur. Ceci est réalisé par une amplification non-linéaire différente par bande de fréquences. Nous avons vu que celui-ci permet de rétablir la compréhension de la parole dans un environnement calme. En revanche, ses performances se dégradent en présence de bruit ou de réverbération. Les performances de localisation peuvent être aussi affectées ainsi que la capacité à se concentrer sur une source de parole dans une scénario multi-locuteur.

Pour pallier ces problèmes, il est d'usage de rajouter un algorithme de débruitage en amont du compresseur pour « nettoyer » la parole. Cependant, celui-ci peut aussi empirer les performances de localisation et il est donc nécessaire d'opérer un compromis entre performances de débruitage et de localisation. Aussi, les algorithmes récents de débruitage nécessitent la connaissance *a priori*, ou l'estimation, de quantités relatives à la scène sonore (*e.g.* DOA, matrice de covariance) parfois difficile à obtenir.

De cette analyse, nous pouvons retirer trois grandes lignes : (i) le compresseur et le débruiteur ont des objectifs antagonistes en matière de RSB ; (ii) tous deux distordent les indices de localisation et les propositions de la littérature jusqu'à lors consistent à faire un compromis entre les performances dans leur tâche respective et préservation des indices de localisation au niveau de chaque traitement, et (iii) la compression de dynamique opérée sur le signal total distord les caractéristiques de chacune des sources. Nous avons vu alors qu'il existe quelques travaux qui traitent tout ou parti de ces problématiques en proposant des manières différentes d'associer ces traitements. Cependant, la littérature reste peu abondante à ce sujet. En particulier, la plupart des travaux traitant de l'association entre algorithme de débruitage et compression considèrent un algorithme de débruitage monocanal, connu pour ne pas améliorer l'intelligibilité de la parole dans ce contexte, à l'inverse des algorithmes multicanaux de type *beamforming*. Enfin, notons que les algorithmes de compression et de débruitage sont traités principalement par deux communautés de recherche distinctes : la psychoacoustique pour la première et le traitement du signal pour le second. Les méthodes, habitudes et niveaux de preuves ne sont pas tout à

fait les mêmes dans ces deux communautés. Par ailleurs, là où la communauté de traitement de signal préfère en général dériver un algorithme à partir d'un modèle (de production et/ou de réception du son) explicite, la communauté en psychoacoustique n'a pas cette habitude. Ceci a pour conséquence que les algorithmes de compression et de débruitage sont difficiles à mettre en regard, car issus de méthodes de conceptions très différentes.

Dans le chapitre suivant, nous proposons une manière d'unifier les formalismes sous-jacents des algorithmes de débruitage et de compression de dynamique usuels. De celui-ci, nous développons un problème d'optimisation permettant d'obtenir une solution analytique unique intégrant à la fois les fonctions de débruitage et compression.

Chapitre 3

Débruitage et correction du recrutement de la sonie conjointes

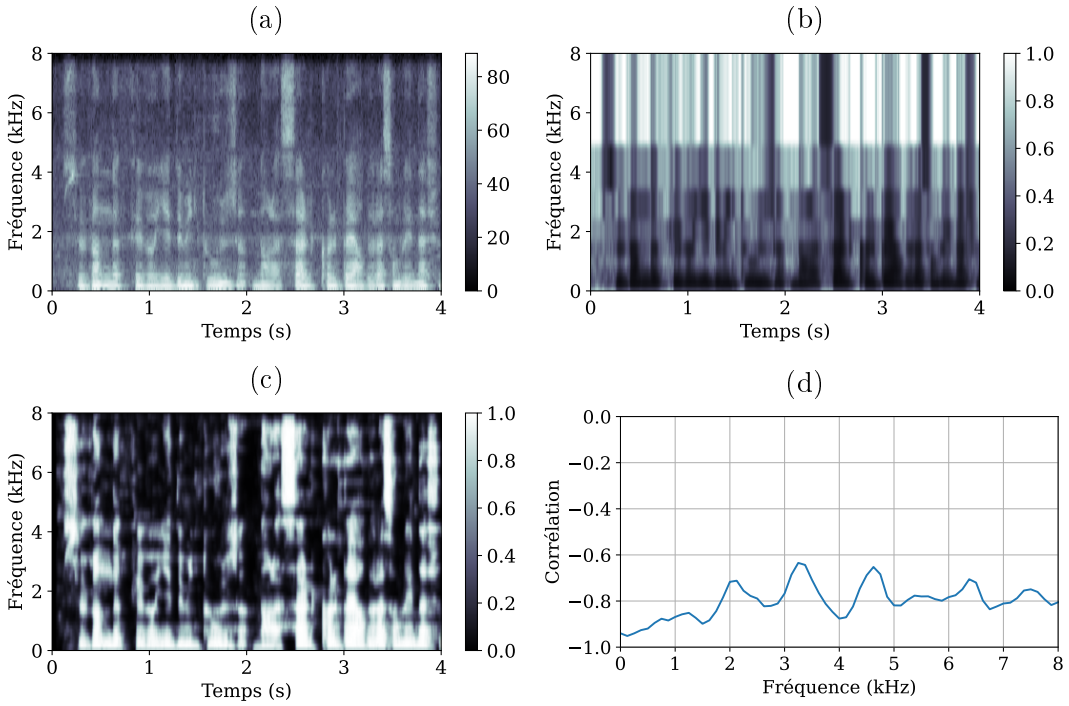


FIGURE 3.1 – Spectrogramme d'un signal de parole bruité à un RSB de 20 dB (a), gain du CD (b) et masque du filtre de Wiener (c) dans le domaine de la TFCT calculé à partir de (a), et coefficient de corrélation entre (b) et (c) en fonction de la fréquence.

3.1 Introduction

Dans le chapitre précédent, nous avons discuté les méthodes de compression de dynamique et de débruitage employées dans les prothèses auditives. Nous avons noté le caractère contraire des objectifs de l'algorithme de débruitage et du compresseur. Nous en fournissant l'exemple illustré en Fig. 3.1 avec un filtre de Wiener suivi d'un CD. En Fig. 3.1a est représenté le spectrogramme d'une source de parole bruitée ; tandis qu'en Fig. 3.1b et 3.1c sont représentés dans le domaine de la TFCT les poids à appliquer à la parole bruitée, entre 0 et 1, correspondant respectivement à la compression et au débruitage. On observe aisément que lorsque l'un est proche de 1, l'autre est proche de 0, et inversement. Nous mettons en évidence cette corrélation inverse en Fig. 3.1d où la corrélation sur la dimension temporelle est calculée pour chaque fréquence. Cette corrélation est particulièrement proche de -1 pour les fréquences inférieures à 1 kHz, soit là où la parole concentre son énergie. On s'attend donc à ce que les poids se contre-balancent et que les performances de débruitage et

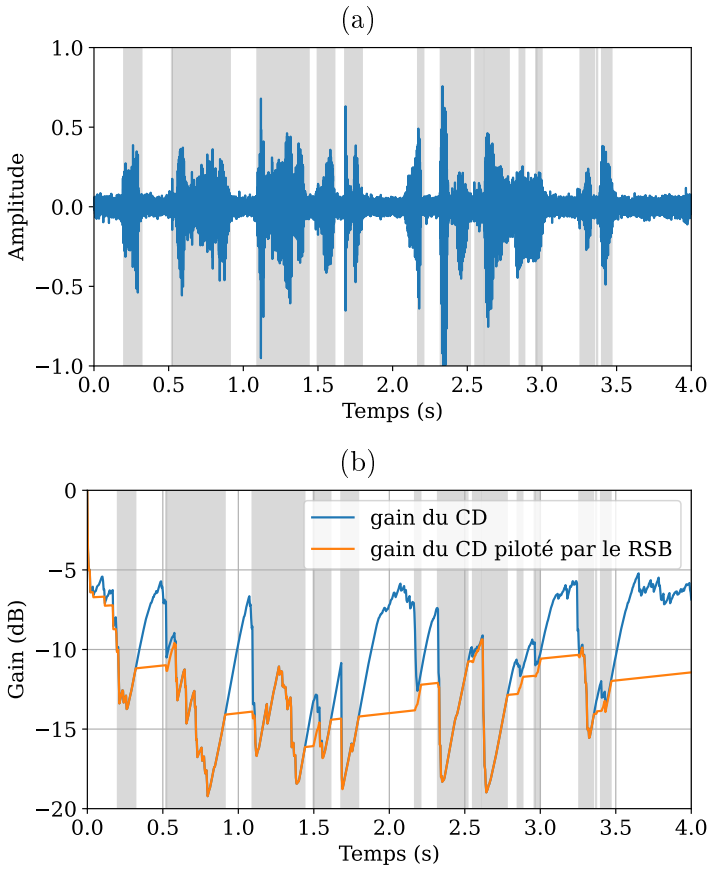


FIGURE 3.2 – Parole bruitée (a) et gain des CDs classique (bleu) et piloté par le RSB (orange) proposé par [May et al., 2018] (b) à 2,8 kHz. Les zones grises illustrent la détection d'activité de la parole avec [Cohen, 2002].

de réduction de dynamique en pâtissent.

Par ailleurs, nous avons présenté les méthodes de compression et de débruitage visant à préserver les caractéristiques acoustiques de la scène sonore, comme les indices de localisation ou l'enveloppe. Notamment, nous avons pointé en sous-section 2.1.2 (p.31) les limites du compresseur de dynamique piloté par le RSB (rsbCD) proposé par [May et al., 2018] du fait de la forte dépendance des performances au contenu de la parole et au seuil de détection de sa présence. Nous en apportons un exemple illustré en Fig. 3.2. En Fig. 3.2a est représenté la forme d'onde de la parole bruité et en Fig. 3.2b les gains de compression, à 2,8 kHz, calculés à partir de cette dernière avec un CD classique et rsbCD. Les zones grises représentent les périodes où la parole est détectée. On observe qu'en dehors de ces zones, le gain du rsbCD est presque figé, dû à la grande constante de temps de relâche (2 s). Cette stratégie permet de ne pas suram-

plifier les périodes d'absence de la parole. Dans l'exemple, la différence entre les deux courbes de gain peut atteindre -8 dB (autour de 2 s). Néanmoins, le gain du rsbCD dans ces périodes est dépendant de la puissance du signal dans la dernière trame de signal identifié comme contenant de la parole. De plus, le gain est systématiquement dans une période de croissance très forte lorsqu'intervient cette détection, rendant la valeur de gain à prolonger d'autant plus incertaine. Dans cette exemple, le gain peut varier de -14 à -10 dB, selon les périodes d'absence de la parole, ce qui est tout à fait audible. Il faut rajouter que le gain est calculé indépendamment par bande de fréquence. Il en résulte que la variation incontrolée du gain à prolonger sur le bruit ne se traduit pas uniquement comme un changement de niveau, mais aussi par une altération du timbre.

Dans ce chapitre, nous proposons d'unifier les tâches de débruitage et de compression de la dynamique au sein d'un même problème d'optimisation. Nous espérons ainsi dépasser les limites posées par des compromis locaux, *i.e.* à l'échelle de l'algorithme de débruitage et du CD.

Aussi, commençons par expliciter les objectifs que l'on attend du traitement effectué par les prothèses auditives :

1. réduire la plage de dynamique de la source de parole cible ;
2. préserver la dynamique originale du bruit ;
3. augmenter le RSB en sortie ;
4. préserver les caractéristiques spatiales de la scène sonore.

Ces quatre objectifs guideront par la suite la conception et l'évaluation de la méthode proposée.

Tout d'abord, en section 3.2, nous développons la fusion des formalismes permettant de développer un modèle des signaux duquel on peut poser un problème d'optimisation englobant les tâches de compression et de débruitage. Pour les rendre compatibles et conserver une solution analytique, il est nécessaire de faire quelques approximations. Nous privilégions cette approche de sorte à pouvoir comparer la solution obtenue avec les méthodes de la littérature. En section 3.3, nous comparons la méthode proposée aux méthodes état-de-l'art grâce à des critères objectifs correspondant aux buts sus-mentionnés. Enfin, en section 3.4, nous évaluons ces méthodes à l'aune de deux critères perceptifs : l'intelligibilité de la parole en présence d'un fort bruit et une notation subjective de la préservation de la scène sonore.

3.2 Méthode proposée

3.2.1 Formulation du problème et solution

Avant toute chose, nous visons ici à traiter de débruitage et compression de dynamique au sein d'un même problème d'optimisation. Cela pose plusieurs difficultés car ces deux algorithmes ont été développés séparément dans des paradigmes différents. Premièrement, la compression est définie dans le domaine des dB là où le débruitage est exprimé dans le domaine linéaire. Deuxièmement, les deux sont définis dans le domaine temps-fréquence avec des résolutions fréquentielles très différentes. En effet, là où l'échantillonnage en fréquence est linéaire pour le débruitage, le compresseur utilise un banc de filtres non linéaire, suivant en général une échelle logarithmique (plus grossier en haute fréquence). Troisièmement, la fonction de compression à proprement parlé est définie en deux parties : une amplification non-linéaire pour la plage audible et une linéaire pour les faibles niveaux, empêchant d'amplifier trop fortement lorsque le signal est absent. Le point de transition entre ces deux régimes, appelé le seuil T_b, est réglé très bas dans les prothèses auditives (entre 20 et 40 $\mathrm{dB_{SPL}}$ [Kuk and Ludvigsen, 2003, Souza et al., 2006, Kowalewski et al., 2020]). À des fins d'analyse et pour éviter de manipuler des expressions définies par parties dans la suite, on suppose que le niveau d'entrée est toujours supérieur à ce seuil [Mauler et al., 2007], ce qui nous permet de considérer uniquement la fonction non-linéaire d'amplification. Dans le but d'unifier les notations, on réécrit les Eq. (2.5, p.30) et (2.2) dans le domaine linéaire et en fonction de la fréquence plutôt qu'en fonction de la bande fréquentielle :

$$g(k,\ell) = g_0(b(k)) \left(\frac{\sum\limits_{\kappa \in \mathcal{B}_{b(k)}} |y(\kappa,\ell)|^2}{\#\mathcal{B}_{b(k)} t_{b(k)}} \right)^{\frac{1}{2}\left(\frac{1}{R_{b(k)}}-1\right)} \tag{3.1}$$

où $t_b = \rho_0 10^{\frac{T_b}{10}}$ et $g_0(b(k)) = 10^{\frac{G_{0,b(k)}}{20}}$. On rappelle que $y(k,\ell)$ est le signal d'entrée générique du CD.

On définit, dans le domaine de la TFCT, le signal que l'on souhaite obtenir en sortie de la prothèse auditive gauche, noté $z_{\mathrm{L}}(k,\ell)$. Le signal pour l'oreille droite est défini de manière similaire et par simplicité, dans la suite nous ne parlerons que de l'oreille gauche, sans perte de généralité. On souhaite que $z_{\mathrm{L}}(k,\ell)$ soit la somme de la source de parole s'étalant dans la plage de dynamique réduite de l'auditeur·rice et une version atténuée du bruit [Rhebergen et al., 2009, Hassager et al., 2017b]. En effet, dans une scène sonore complexe, la variance du bruit $\phi_n(k,\ell)$ varie peu au cours du temps. Il n'est donc pas nécessaire d'avoir un compresseur qui suit ces moindres variations à chaque

instant. En outre, il est bien établi que les indices de localisation sont stables avec un gain de compression peu variable [Schwartz and Shinn-Cunningham, 2013, Hassager et al., 2017a, May et al., 2018]. On peut alors écrire l'expression suivante :

$$z_{\mathrm{L}}(k,\ell) = h_{\mathrm{L}}(k)\check{s}(k,\ell) + \eta\, n_{\mathrm{L}}(k,\ell), \tag{3.2}$$

et $\check{s}(k,\ell)$ la source de parole compressée [Mauler et al., 2007] :

$$\check{s}(k,\ell) = \underbrace{g_0(b(k))t_{b(k)}^{1-\frac{1}{R_{b(k)}}}}_{\tau(k)} |s(k,\ell)|^{\frac{1}{R_{b(k)}}} e^{\jmath\theta^s(k,\ell)}, \tag{3.3}$$

où $\jmath$ est l'unité imaginaire et $\theta^s(k,\ell)$ est la phase de la source de parole.

L'estimateur de $z_{\mathrm{L}}(k,\ell)$, noté $\hat{z}_{\mathrm{L}}(k,\ell)$, est alors construit comme une combinaison linéaire des observations, pondérée par les poids $\mathbf{w}(k,\ell) \in \mathbb{C}^M$:

$$\hat{z}_{\mathrm{L}}(k,\ell) = \mathbf{w}(k,\ell)^H \mathbf{x}(k,\ell). \tag{3.4}$$

Dans la suite, nous omettons les indices k, et ℓ par soucis de brièveté. Le filtre $\mathbf{w}$ peut être réglé de sorte à répondre au problème d'optimisation suivant :

$$\mathbf{w}_{\mathrm{p}} = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{ J_{\mathrm{p}}(\mathbf{w}) \right\}, \tag{3.5}$$

avec

$$\begin{aligned} J_{\mathrm{p}}(\mathbf{w}) &= \mathbb{E}\left[|z_{\mathrm{L}} - \hat{z}_{\mathrm{L}}|^2 \right], \\ &= \mathbb{E}\left[|h_{\mathrm{L}}\tau|s|^{\frac{1}{R}} e^{\jmath\theta_s} + \eta n_{\mathrm{L}} - \mathbf{w}^H\mathbf{x}|^2 \right], \\ &= \mathbb{E}\left[|(h_{\mathrm{L}}\tau|s|^{\frac{1}{R}-1} - \mathbf{w}^H\mathbf{h})s|^2 \right] + \mathbb{E}\left[|(\eta\mathbf{q}_{\mathrm{L}} - \mathbf{w})^H\mathbf{n}|^2 \right], \end{aligned} \tag{3.6}$$

où $\mathbf{q}_{\mathrm{L}}$ est un vecteur nul à l'exception de l'élément correspondant au microphone de référence pour l'oreille gauche réglé à 1. Le premier terme de la dernière équation est très difficile à manipuler. On propose alors de faire l'approximation suivante :

$$|s|^{\frac{1}{R}-1} \approx \phi_s^{\frac{1}{2}(\frac{1}{R}-1)}. \tag{3.7}$$

Ainsi,

$$J_{\mathrm{p}}(\mathbf{w}) \approx \mathbb{E}\left[|(h_{\mathrm{L}}\tau\phi_s^{\frac{1}{2}(\frac{1}{R}-1)} - \mathbf{w}^H\mathbf{h})s|^2 \right] + \mathbb{E}\left[|(\eta\mathbf{q}_{\mathrm{L}} - \mathbf{w})^H\mathbf{n}|^2 \right]. \tag{3.8}$$

On note cette approximation $J_{\tilde{\mathrm{p}}}(\mathbf{w})$ que l'on peut réécrire sous la forme :

$$J_{\tilde{\mathrm{p}}}(\mathbf{w}) = \mathbb{E}\left[|(h_{\mathrm{L}}g - \mathbf{w}^H\mathbf{h})s|^2 \right] + \mathbb{E}\left[|(\eta\mathbf{q}_{\mathrm{L}} - \mathbf{w})^H\mathbf{n}|^2 \right], \tag{3.9}$$

avec

$$g = g_0 \left(\frac{\phi_s}{t} \right)^{\frac{1}{2}(\frac{1}{R}-1)}. \tag{3.10}$$

On reconnait ici le gain du compresseur de dynamique (tel que défini en Eq. (3.1)) appliqué à la source de parole. C'est donc au prix de l'approximation en Eq. (3.7) que l'on peut simplifier $J_{\mathrm{p}}(\mathbf{w})$ en $J_{\tilde{\mathrm{p}}}(\mathbf{w})$ de sorte à rendre une expression intraitable en une fonction de coût d'optimisation standard dans la littérature des prothèses auditives dont on connait la solution analytique [Van den Bogaert et al., 2008] :

$$\mathbf{w}_{\tilde{\mathrm{p}}} = h_{\mathrm{L}}^* g \mathbf{w}_{\mathrm{MWF}} + \eta(\mathbf{q}_{\mathrm{L}} - h_{\mathrm{L}}^* \mathbf{w}_{\mathrm{MWF}}), \tag{3.11}$$

où $\mathbf{w}_{\mathrm{MWF}}$ est le MWF. Cette solution est donc très similaire dans sa forme au beamformer MWF-N, se différenciant uniquement par le facteur g. De sorte à améliorer la robustesse du MWF, on utilise sa forme décomposée en beamformer sans distorsion et variance minimale (MVDR) suivi d'un filtre de Wiener [Brooks and Reed, 1972] :

$$\mathbf{w}_{\mathrm{MWF}} = \frac{\xi_o}{1+\xi_o} \underbrace{\frac{\mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{h}}{\mathbf{h}^H\mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{h}}}_{\mathrm{MVDR}}, \tag{3.12}$$

où $\xi_o = \phi_s/\phi_n \mathbf{h}^H\mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{h}$ est le RSB en sortie du beamformer MVDR.

3.2.2 Calcul du gain de compression

Il reste que calculer g est difficile car nous n'avons pas accès à la variance de la source cible. Néanmoins, un algorithme de débruitage comme le beamformer MVDR ou le MWF peut fournir une estimation de $s_{\mathrm{L}} = h_{\mathrm{L}}s$, la source de parole reçu au microphone de référence pour l'oreille gauche (s_{R} pour l'oreille droite).

Nous allons montrer ici que le gain de compression du compresseur appairé prenant en entrée s_{L} et s_{R}, noté $\bar{g}$, se révèle être un bon estimateur de g, le gain de compression prenant en entrée la source de parole sans les indices de localisation. Rappelons la définition de $\bar{g}$: de sorte que la sortie du compresseur à droite et à gauche ne dépasse pas le seuil de douleur et préserve l'ILD, on prend le minimum des gains de compression calculés à partir des signaux droite et gauche, g_{L} et g_{R} :

$$\bar{g} = \min\{g_{\mathrm{L}}, g_{\mathrm{R}}\}. \tag{3.13}$$

En utilisant Eq. (3.1), l'expression précédente devient :

$$\bar{g} \approx g_0 \min \left\{ \left(\frac{\sum_{\kappa \in \mathcal{B}} |h_{\mathrm{L}}(\kappa) s(\kappa)|^2}{\#\mathcal{B}t} \right)^{\frac{1}{2}\left(\frac{1}{R}-1\right)}, \left(\frac{\sum_{\kappa \in \mathcal{B}} |h_{\mathrm{R}}(\kappa) s(\kappa)|^2}{\#\mathcal{B}t} \right)^{\frac{1}{2}\left(\frac{1}{R}-1\right)} \right\}. \tag{3.14}$$

Sachant que $R > 1$, on peut écrire :

$$\bar{g} \approx g_0 \left(\frac{\max \left\{ \sum_{\kappa \in \mathcal{B}} |h_{\mathrm{L}}(\kappa) s(\kappa)|^2, \sum_{\kappa \in \mathcal{B}} |h_{\mathrm{R}}(\kappa) s(\kappa)|^2 \right\}}{\#\mathcal{B}t} \right)^{\frac{1}{2}\left(\frac{1}{R}-1\right)}. \tag{3.15}$$

Définissons l'HRTF ipsilatérale comme celle entre la source et l'oreille la plus proche, notée h_{I}. En considérant que $|h_{\mathrm{L}}| > |h_{\mathrm{R}}|$ pour toutes les sources situées à gauche de l'auditeur·rice et inversement, alors $|h_{\mathrm{I}}| = \max\{|h_{\mathrm{L}}|, |h_{\mathrm{R}}|\}$. On illustre cette propriété en Fig. 3.3(a, b et c)) avec les amplitudes des HRTFs en dB sur le plan horizontal. L'expression précédente peut alors être réécrite comme suit :

$$\bar{g} \approx g_0 \left(\frac{\sum_{\kappa \in \mathcal{B}} |h_{\mathrm{I}}(\kappa) s(\kappa)|^2}{\#\mathcal{B}t} \right)^{\frac{1}{2}\left(\frac{1}{R}-1\right)}. \tag{3.16}$$

On suppose que $|h_{\mathrm{I}}(\kappa)|$ est approximativement constant au sein de chaque bande fréquentielle. Cette approximation est illustrée en Fig. 3.3(c, d et e) sur le plan horizontal, en utilisant un banc de filtres par bande d'octave. En particulier, la Fig. 3.3d représente la moyenne par bande fréquentielle de $|h_{\mathrm{I}}(\kappa)|$. La Fig. 3.3e représente, quant à elle, la différence en dB entre les deux amplitudes. Celle-ci est d'au maximum de l'ordre de 10 dB à de rares fréquences (autour de 3.2 kHz). La constante issue de l'approximation, une fois sortie de la somme et élevée[1] à la puissance $\frac{1}{2}\left(\frac{1}{R}-1\right)$, est proche de 1. Cette dernière approximation est illustrée en Fig. 3.3f où est représentée la Fig. 3.3d élevée à la dite puissance. On observe que le résultat se situe autour de 0 dB (±3 dB) pour toutes les fréquences et toutes les directions. Il faut noter que le placement des microphones au-dessus du pavillon permet d'asseoir cette approximation car ils ne captent pas les indices spectraux relatifs à ce dernier [Kayser et al., 2009, Oreinos and Buchholz, 2013]. Alors, on peut écrire :

$$\bar{g} \approx g_0 \left(\frac{\sum_{\kappa \in \mathcal{B}} |s(\kappa)|^2}{\#\mathcal{B}t} \right)^{\frac{1}{2}\left(\frac{1}{R}-1\right)}. \tag{3.17}$$

[1] Élévation qui correspond à la fonction de compression.

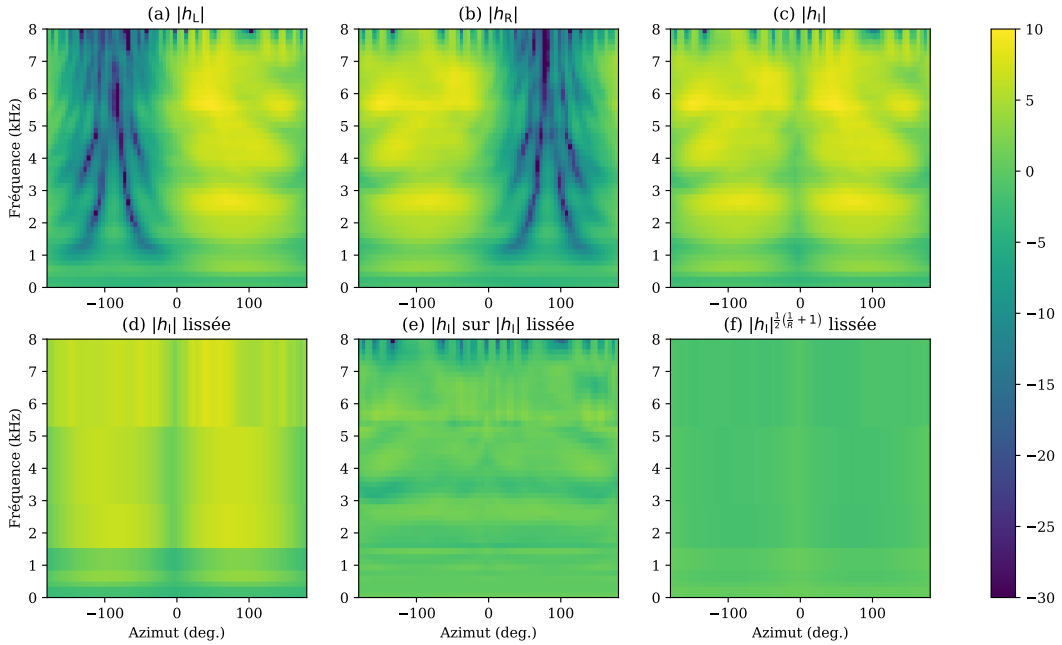


FIGURE 3.3 – Amplitude en dB d'HRTF du plan horizontal pour l'oreille gauche (a), droite (b), l'HRTF ipsilatérale (c), cette dernière passée au travers du banc de filtres par bande d'octave du CD (d). La différence en dB entre (c) et (d) est présentée en (e) et enfin en (f) l'HRTF ipsilatérale passée au travers du banc de filtres puis compressée avec $R = 2$.

Nous avons expliciter ici comment les choix du banc de filtres du compresseur et l'appairage gauche/droite permettent d'obtenir une estimation de la puissance de la source malgré l'ambiguïté de niveau initiale.

Nous avons bien conscience que certaines des approximations qui ont été faites sont substantielles. Notre propos est d'expliciter un modèle de sortie idéal souvent évoqué [Rhebergen et al., 2009, Hassager et al., 2017b] mais rarement formalisé tel que nous l'avons fait en Eq. (3.2), et de développer, au prix de ces approximations, une solution qui utilise le même formalisme que les solutions que l'on trouve dans la littérature. Ceci n'est pas négligeable car cela nous permet de comparer notre solution à celles préexistantes et aussi d'expliciter le rôle de certaines étapes, comme le choix dans le compresseur du banc de filtres, l'appairage gauche/droite et le filtrage de l'enveloppe. La contribution principale ici est de traiter de compression et de beamforming au sein d'un même formalisme.

[Ngo et al., 2012] ont déjà proposé une méthode à l'architecture similaire

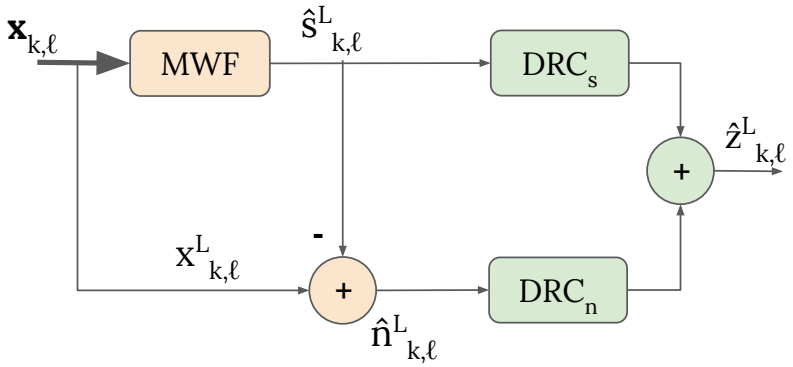


FIGURE 3.4 – Schéma bloc de la méthode proposée.

TABLE 3.1 – Récapitulatif des conditions algorithmiques testées dans l'évaluation objective.

Condition	Débruitage	Compression
Normo-entendant (NH)		
CD appairés		CD
MWF-N + CD	MWF-N	CD
MWF-N + rsbCD	MWF-N	rsbCD [May et al., 2018]
Ngo [Ngo et al., 2012]	Combinaison bilatérale	
Proposée	Combinaison binaurale	

mais nous apportons ici un éclairage nouveau sur les modèles des signaux d'entrée et de sortie, ainsi que sur les hypothèses sous-jacentes permettant d'aboutir à une telle solution. Un résumé de l'implémentation de l'algorithme est illustré en Fig. 3.4. En pratique, le paramètre η est le gain de compression appairé, calculé à partir de l'estimation du bruit lissée avec de grandes constantes de temps.

3.3 Évaluation objective

Dans cette section, on compare la méthode proposée aux méthodes état-de-l'art grâce à des critères objectifs. On veut ainsi évaluer l'influence de chaque méthode sur les caractéristiques des signaux en sortie que nous avons présenté en section 3.1.

3.3.1 Méthode

Dispositif

Chaque stimulus est composé d'une phrase prononcée par un locuteur masculin et un bruit d'ambiance de cafétéria. Le locuteur est virtuellement spatialisé sur le plan horizontal à l'azimut 0° (en face de l'auditeur) en utilisant les ATFs des microphones avant et arrière de prothèses auditives placées derrière les oreilles ($M = 4$) de [Kayser et al., 2009]. L'enregistrement d'ambiance de cafétéria provient de la même base de données d'ATFs. Le bruit ambiant seul est fixé au niveau de 50 dB(A) et le niveau de la source de parole est modifié de sorte à évaluer les RSBs suivants : 0, 5 et 10 dB.

En Tab. 3.1 sont rassemblées les conditions algorithmiques considérées dans cette évaluation. Tous les algorithmes sont intégrés à une chaîne de traitement de signal telle que décrite en sous-section 1.3.3 avec une trame de longueur 8 ms, 50 % de recouvrement et une fréquence d'échantillonnage de 16 kHz. Les compresseurs sont paramétrés avec les réglages décrits en Tab. 3.4. Le compresseur de dynamique piloté par le RSB (rsbCD) [May et al., 2018] est implémenté au plus proche de la description faite par les auteurs, à l'exception du détecteur de présence de la parole où nous utilisons l'estimateur de PPPM de [Souden et al., 2010] avec l'implémentation proposée en annexe D. Dans les conditions incluant un beamformer MWF-N, le paramètre de compromis entre la sortie du beamformer et le signal du microphone de référence est réglé à 0,2 comme préconisé par [Van den Bogaert et al., 2008] de sorte à rétablir la localisation des sources hors-axe. Dans toutes les conditions incluant un CD à l'exception de la méthode proposée par [Ngo et al., 2012], les constantes de temps du filtre de lissage du CD appliqué à la parole sont réglées à 10 ms et 60 ms. Dans la condition *Proposée*, le compresseur appliqué au bruit est réglé avec des constantes de temps d'attaque et de relâche de 10 ms et 2000 ms, respectivement, tandis que celles de la condition *Ngo* sont réglées à 10 ms et 20 ms [Ngo et al., 2012]. La moyenne des ATFs sur le plan horizontal nous permet d'obtenir l'estimation de la matrice de cohérence d'un bruit spatialement diffus, $\mathbf{\Gamma_n}$, nécessaire au beamforming [Lotter and Vary, 2006].

Chaque stimulus dure environ 5,5 secondes dont 1,5 secondes de bruit seul au début. Au moment de la présentation du stimulus, la première seconde n'est pas lu car c'est le temps nécessaire à certains algorithmes pour s'initialiser (en particulier la condition *Ngo*). Lors de cette phase d'initialisation, des artéfacts non représentatifs en situation d'usage à long terme peuvent apparaître.

Critère d'évaluation

À la fin de la section 3.1, nous avons défini quatre objectifs que doit remplir la chaîne de traitement des prothèses auditives. Nous présentons ici les métriques objectives que nous avons choisies pour les évaluer.

Amélioration du RSB On définit le rapport signal à bruit (RSB) en sortie destinée à l'oreille gauche de la manière suivante :

$$\mathrm{RSB_L} = 10 \log_{10} \frac{\sum_t s_{\mathrm{out,\,L}}(t)^2}{\sum_t n_{\mathrm{out,\,L}}(t)^2}, \tag{3.18}$$

et de même manière pour l'oreille droite. Les RSB sont moyennés entre oreille droite et gauche puis référencés par rapport au RSB d'entrée, permettant d'obtenir l'amélioration du RSB, notée ΔRSB.

Compression effective La réduction de la dynamique, ou son augmentation, est définie par le taux de compression effectif (de l'anglais *Effective Compression Ratio*, ECR). Cette métrique consiste à faire le rapport entre la dynamique d'entrée et de sortie. Nous utilisons ici la définition fournie en Eq. (2.11, p.33). On utilise le même banc de filtres que pour les CDs (voir Tab. 3.4 (p.93)).

Erreur quadratique moyenne de l'enveloppe Pour compléter l'ECR dans l'évaluation de la préservation des caractéristiques acoustiques du bruit, nous définissons un critère de préservation de son enveloppe. Celle-ci est définie comme :

$$\mathrm{env}_n(t) = \mathrm{LP}_{50\mathrm{Hz},6}(|n_{\mathrm{out}}(t)|), \tag{3.19}$$

où $n_{\mathrm{out}}(t)$ est la composante de bruit en sortie et $\mathrm{LP}_{50\mathrm{Hz},6}(.)$ est un filtre passe-bas d'ordre 6 et de fréquence de coupure 50 Hz. Cette enveloppe est alors normalisée par sa valeur efficace puis soustraite par l'enveloppe idéale pour donner l'erreur d'enveloppe. L'erreur quadratique moyenne (EQM) est alors définie comme la valeur efficace de cette erreur d'enveloppe.

Cohérence interaurale Pour évaluer la largeur de la scène sonore, on utilise la cohérence interaurale (IC). Pour plus de détail, nous renvoyons à la sous-sous-section 1.2.5 (p.17).

Ici, on s'intéresse à la cohérence interaurale de la composante de bruit ambiant. Celle-ci est définie ici comme l'absolue du maximum de la corrélation-croisée normalisée entre les composantes de bruit droite et gauche en sortie de

la prothèse auditive filtrée par un passe-bande, notée $\tilde{n}_{\text{out, L}}(t)$ et $\tilde{n}_{\text{out, R}}(t)$:

$$\text{IC}_{\tilde{n}} = \max_{\tau} \left| \frac{\sum_t \tilde{n}_{\text{out, L}}(t+\tau)\tilde{n}_{\text{out, R}}(t)}{\sqrt{\sum_t |\tilde{n}_{\text{out, L}}(t)|^2 \sum_t |\tilde{n}_{\text{out, R}}(t)|^2}} \right|, \tag{3.20}$$

avec $|\tau| < 1$ ms. Le passe-bande est accordé sur la fréquence 2 kHz comme dans [Hassager et al., 2017b] car cette fréquence correspond à la zone de transition entre les indices interauraux.[2]

3.3.2 Résultats

Tout d'abord, on observe en Fig. 3.5a que la compression seule diminue bien le RSB par rapport à l'entrée et ce, de plus en plus en fonction du niveau de la source de parole. En ajoutant un MWF-N en amont, les performances s'améliorent mais continuent de décroître en augmentant le niveau de la source. On ne remarque pas ici d'effet significatif du rsbCD. Enfin, la méthode proposée permet bien d'améliorer le RSB, et ce de manière constante pour tous les RSBs d'entrée testés. En revanche, la méthode de [Ngo et al., 2012] obtient des performances en deçà des méthodes de débruitage binaurales, dû au fait que le traitement est effectué de manière indépendante à droite et à gauche.

On observe bien l'influence de la compression calculée sur le mélange bruité en Fig. 3.5c et Fig. 3.5e. Plus le RSB est élevé plus la source de parole influence de manière importante la dynamique de la composante de bruit en augmentant sa dynamique (ECR<1). La méthode proposée améliore significativement ce critère par rapport à toutes les autres méthodes, sans pour autant totalement rejoindre les performances idéales. La méthode de [Ngo et al., 2012] ne permet pas d'améliorer significativement ce critère. Bien que significatif, la taille d'effet des performances d'ECR observées est à confirmer avec l'évaluation perceptive. Toutefois, on peut regarder plus précisément l'enveloppe de la composante de parole et la puissance de son erreur avec celle idéalement attendue. En Fig. 3.6, on met en évidence que la parole de forte intensité peut déformer l'enveloppe du bruit avec les algorithmes état-de-l'art (voir entre 1.1 et 1.2 s et dans une moindre mesure autour de 0.9 s). La méthode proposée permet de corriger cet effet et de réduire en moyenne l'erreur de l'enveloppe, voir Fig. 3.5e.

En Fig. 3.5d est représentée la cohérence interaurale du bruit à 2 kHz. On observe que les méthodes basées sur une combinaison sérielle des algorithmes de débruitage et de compression ne permettent pas de respecter tout-à-fait la cohérence interaurale du bruit comparée à la méthode proposée, du fait du

[2] ITD pour les basses fréquences, ILD pour les hautes fréquences.

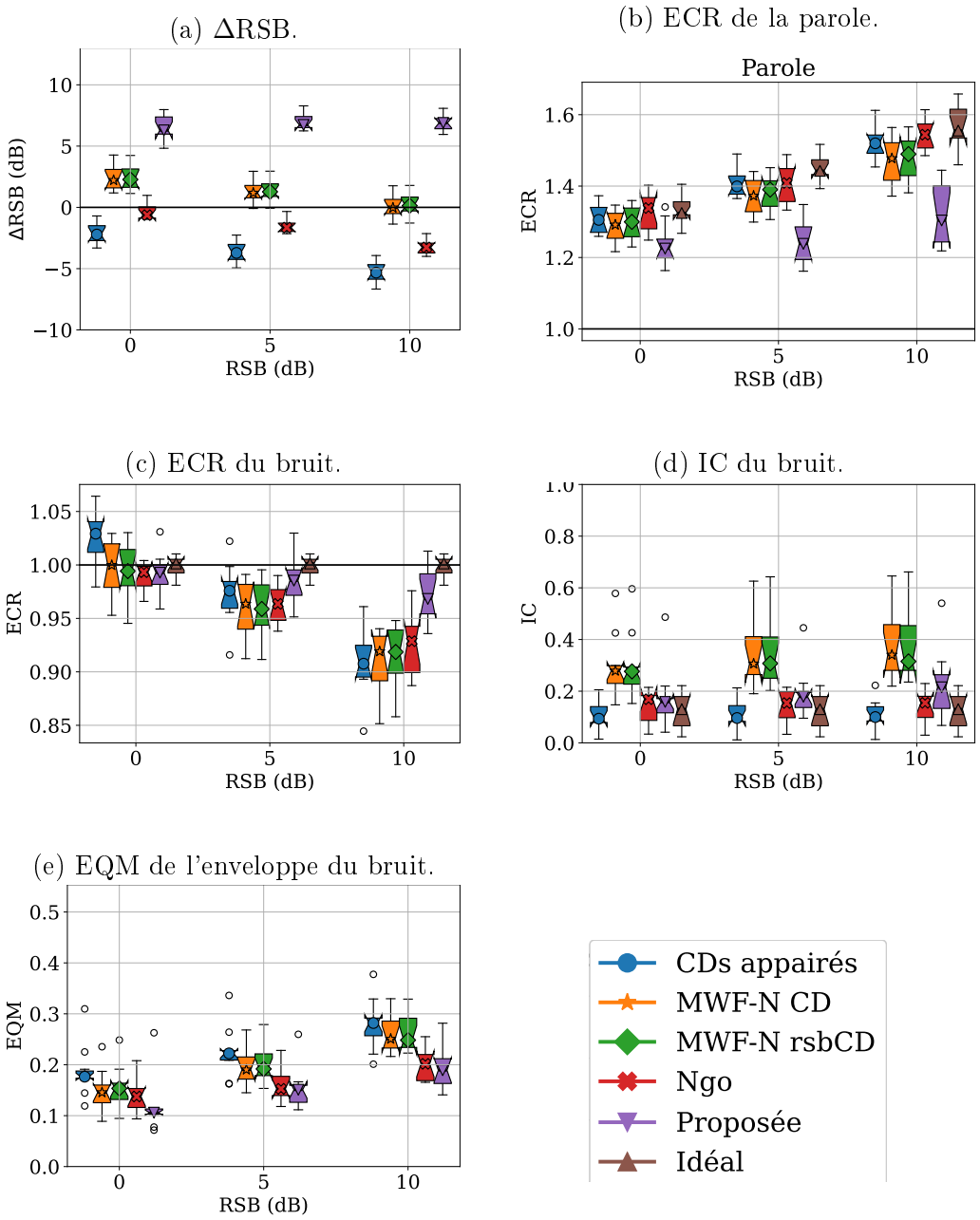


FIGURE 3.5 – Métriques objectives pour chaque condition algorithmique et les RSBs 0, 5 et 10 dB. Chaque boîte à moustaches agrège les résultats sur dix extraits. Pour simplifier la lecture des résultats, les boîtes à moustaches sont légèrement éclatées sur l'axe horizontale autour du RSB testé.

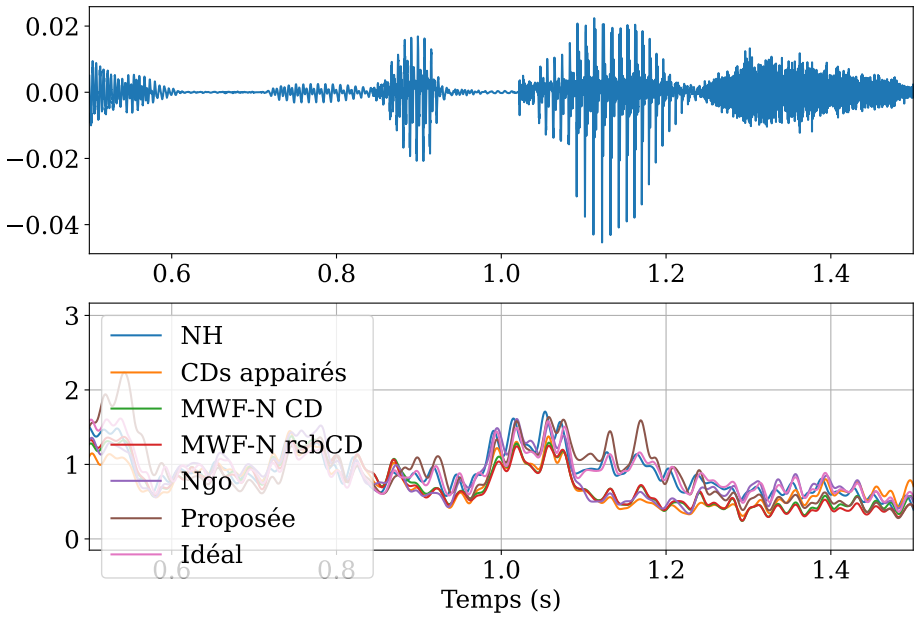


FIGURE 3.6 – Exemple d'enveloppe du bruit pour chaque condition testée (bas) avec en regard le signal de parole au même moment (haut).

compromis fait entre débruitage et préservation des indices de localisation effectué localement au niveau du MWF-N plutôt que sur l'ensemble de la chaîne. La méthode de [Ngo et al., 2012] semble obtenir une IC correcte mais c'est au détriment de l'amélioration du RSB. La méthode que nous proposons dans ce chapitre permet, quant à elle, de préserver la cohérence interaurale tout en améliorant le RSB.

Enfin, notons que la méthode proposée augmente la dynamique de la parole, voir Fig. 3.5b. Elle semble donc moins bien compresser la parole que les autres méthodes, ce qui pourrait se révéler néfaste pour son intelligibilité. Il semble que ce soit davantage dû au débruitage plus agressif que ce critère diminue. Le test de compréhension de la parole permettra de compléter cette analyse.

3.4 Évaluation perceptive

Dans cette section, on décrit la méthode d'évaluation perceptive de la compréhension de la parole et de préservation de la scène sonore. Dans la suite, on appellera une condition d'écoute de l'expérience une combinaison de traitements appliqués au signal capté par les microphones des prothèses auditives.

TABLE 3.2 – Récapitulatif des conditions de test pour le test d'intelligibilité.

Condition	Débruitage	Compression	HLS [a]
Normo-entendant (NH)			
Non-appareillé			✓
CDs appairés		CD	✓
MVDR-N + CD	MVDR-N	CD	✓
MVDR-N + rsbCD	MVDR-N	rsbCD	✓
Ngo [Ngo et al., 2012]	Combinaison bilatérale		✓
Proposée	Combinaison binaurale		✓

[a] simulation de perte auditive (de l'anglais *Hearing Loss Simulation*, HLS) [Grimault et al., 2018].

Celles-ci sont résumées en Tab. 3.2. Les expériences sont menées sur des sujets normo-entendants et une simulation de perte auditive (de l'anglais *Hearing Loss Simulation*, HLS) est appliquée lorsque c'est nécessaire [Hu et al., 2011, Grimault et al., 2018].

3.4.1 Méthode

Dispositif

On reprend l'essentiel du dispositif utilisé pour l'évaluation objective, décrite en sous-section 3.3.1 à l'exception du filtre de Wiener en aval du beamformer MVDR dans la méthode proposée et les conditions incluant un MWF-N qui est inactivé car il est difficile d'obtenir une estimation du RSB en sortie du beamformer MVDR pour les RSBs considérés dans le test de compréhension de la parole (voir ci-dessous).

Les pertes auditives sont simulées grâce à un modèle proposé par [Grimault et al., 2018] suivant l'audiogramme de référence $N3$ de [Bisgaard et al., 2010], correspondant à une perte modérée (voir Tab. 3.3).

Les stimuli sont diffusés grâce à une carte son Focusrite Scarlett 2i2 et un casque Sennheiser HD650. Pour le test d'intelligibilité, le système de diffusion a été calibré de sorte que le bruit ambiant seul soit délivré au niveau de 50 dB(A) à l'oreille du sujet. Celui-ci reste fixe et le niveau de la source de parole est modifié de sorte à évaluer les RSBs suivants : -17, -14,5, -11, -9,5 dB, permettant de couvrir l'ensemble de la plage de transition entre une absence et une totale compréhension de la parole pour les conditions testées. Pour le test de préférence, le RSB est fixé à 0 dB.

TABLE 3.3 – Réglage de la simulation de perte auditive correspondant à l'audiogramme standard $N3$ défini par [Bisgaard et al., 2010].

Fréq. (kHz)	0,25	0,375	0,5	0,75	1	1,5	2	3	4	6
Perte (dB)	35	35	35	35	40	45	50	55	60	65

TABLE 3.4 – Réglage des CDs pour un audiogramme type $N3$ [Bisgaard et al., 2010] issus de [Kowalewski et al., 2020].

Fréquences (kHz)	0,125	0,25	0,5	1	2	4	8
Seuil (dB_{SPL})	43	43	41	41	37	31	28
Gain (dB)	35	35	35	40	50	60	60
Ratio	1,7	1,7	1,8	2,1	2,6	2,8	2,4

Sujets

21 sujets (12 femmes et 9 hommes, âgés entre 22 et 37 ans avec une moyenne de 28,4 ans et un écart-type de 3,52) normo-entendants ne rapportant pas la connaissance de perte auditive ont pris part à l'expérience. Chaque sujet a pris connaissance de ses droits sur les données recueillies ainsi que leur finalité et a accordé son consentement par écrit.

Critère d'évaluation

Pourcentage de mots reconnus

Afin d'évaluer la compréhension de la parole, on mesure la proportion de mots reconnus. Pour cela, on utilise le corpus *French Matrix* [Jansen et al., 2012] (voir Tab. 3.5). Celui-ci permet de construire des phrases composées de *prénom, verbe, numéro, objet* et *couleur* en piochant aléatoirement dans une liste de dix items différents.

Chaque combinaison de RSB et d'algorithme est répétée six fois avec une phrase différente à chaque fois. Ainsi, on obtient trente réponses binaires de reconnaissance de mots par sujet, RSB et algorithme. Le pourcentage de réponses correctes est donc mesuré avec un pas de quantification de 3,33 %.

Il est demandé au sujet de répéter les mots entendus provenant du locuteur situé virtuellement en face de lui. Le test commence par une phase d'entrainement où le sujet se familiarise avec la tâche en écoutant un exemple pour chaque algorithme testé diffusé au RSB de -10 dB, les réponses associées ne sont pas utilisées pour l'analyse des résultats. Lors du test, l'ordre de présentation des algorithmes est randomisée. La durée du test de compréhension de la parole est d'environ 30 minutes.

TABLE 3.5 – Ensemble des 50 mots de la *French Matrix*.

Nom	Verbe	Numéro	Objet	Couleur
Agnès	achète	deux	anneaux	blanc
Charlotte	attrape	trois	ballons	bleus
Émile	demande	cinq	classeurs	bruns
Étienne	déplace	six	crayons	gris
Eugène	dessine	sept	jetons	jaunes
Félix	propose	huit	livres	mauves
Jean-Luc	ramasse	neuf	pions	noirs
Julien	ramène	onze	piquets	roses
Michel	reprend	douze	rubans	rouges
Sophie	voudrait	quinze	vélos	verts

Speech Reception Threshold (SRT)

Le seuil de compréhension de la parole (*Speech Reception Threshold*, SRT) est défini comme le RSB pour lequel une certaine proportion (en général 50 %) de mots sont reconnus [Levitt, 1971]. Il permet de fournir un critère synthétique pour classer les performances des algorithmes en s'abstrayant du RSB de présentation. Ici, on estime le SRT en ajustant une fonction sigmoïde sur le pourcentage de mots reconnus en fonction du RSB pour chaque sujet comme dans [Brand and Kollmeier, 2002, Moore et al., 2019b]. Celle-ci est définie comme suit :

$$\sigma(L, L_{50}, s_{50}) = \frac{1}{1 + e^{4 s_{50}(L_{50} - L)}}, \tag{3.21}$$

où L est le RSB, L_{50} est le SRT à 50 % et s_{50} la pente de la sigmoïde. Les deux paramètres, L_{50} et s_{50}, sont alors réglés de manière à maximiser la fonction de vraisemblance du modèle sigmoïde, notée $\mathcal{L}(L_{50}, s_{50})$, avec les données recueillies pour chaque sujet de la manière suivante :

$$\mathcal{L}(L_{50}, s_{50}) = \prod_{i=1}^{I} \sigma(L_i, L_{50}, s_{50})^{c(i)} (1 - \sigma(L_i, L_{50}, s_{50}))^{1-c(i)}, \tag{3.22}$$

où i est l'indice du mot, I est le nombre de mot évalués, L_i le RSB de présentation du stimulus et $c(i)$ est 1 si la réponse est juste et 0 sinon.

Évaluation subjective de la préservation de la scène sonore

Afin d'évaluer la préservation de la scène sonore, on utilise le test multi stimuli avec référence cachée et une ancre (de l'anglais *MUlti Stimulus test with Hidden Reference and Anchor*, MUSHRA) [International Telecommunications

TABLE 3.6 – Valeurs-p pour le SRT.

Condition	NH	CDs appairés	MVDR-N +CD	MVDR-N +rsbCD	Ngo
NH	-				
CDs appairés	0,63	-			
MVDR-N+CD	<0,01	<0,01	-		
MVDR-N+rsbCD	<0,01	<0,01	0,99	-	
Ngo	<0,01	0,01	<0,01	<0,01	-
Proposée	<0,01	<0,01	0,92	0,98	<0,01

Union-Recommendation and BS.1534-1., 2003] et en particulier l'implémentation webMUSHRA [Schoeffler et al., 2018]. Ce protocole consiste à noter l'effet de chaque algorithme appliqué à un même stimulus. Parmi les algorithmes, une condition de référence et une ancre basse sont incluses. Comme référence, on choisi de prendre la condition normo-entendante avec un RSB supérieur de 6 dB, correspondant au gain en RSB moyen du beamformer MVDR-N. L'ancre basse est une condition irréaliste visant à forcer le sujet à la noter toujours au plus bas. On choisi pour cela d'appliquer la simulation de perte auditive, le compresseur et un moyennage des signaux droite et gauche de manière à présenter à l'auditeur·rice un signal dénué d'indices de localisation interauraux.

Il est demandé aux sujets de noter la préservation globale (niveau, timbre et spatialisation) de la scène sonore, à l'exception du locuteur situé en face, sur une échelle de 0 à 100, annotée par tranche de 20 par les qualificatifs : *bad, poor, fair, good* et *excellent*, par rapport au signal de référence. Le sujet peut écouter le stimulus autant de fois qu'il le souhaite et comparer les différents algorithmes comme il l'entend. Cinq répétitions du test sont effectuées avec une phrase et un bruit différent à chaque fois.

La condition *CDs appairés* n'est pas testée ici car l'absence de débruitage la rend trop différente des autres conditions. En effet, nous avons observé lors de l'élaboration du test que les sujets avaient tendance à noter de manière semblable tous les algorithmes incluant du débruitage si cette condition était incluse.

3.4.2 Résultats

Compréhension de la parole

Les performances de compréhension de la parole sont illustrées en Fig. 3.7 et 3.8 sous forme de pourcentages de mots reconnus en fonction du RSB, pour chaque condition ; ainsi qu'en Fig. 3.9 pour le SRT en fonction des conditions d'écoute. Les valeurs-p par paire de conditions pour le SRT sont rapportées en Tab. 3.6.

L'estimation de SRT de la condition *Ngo* est à considérer avec précaution

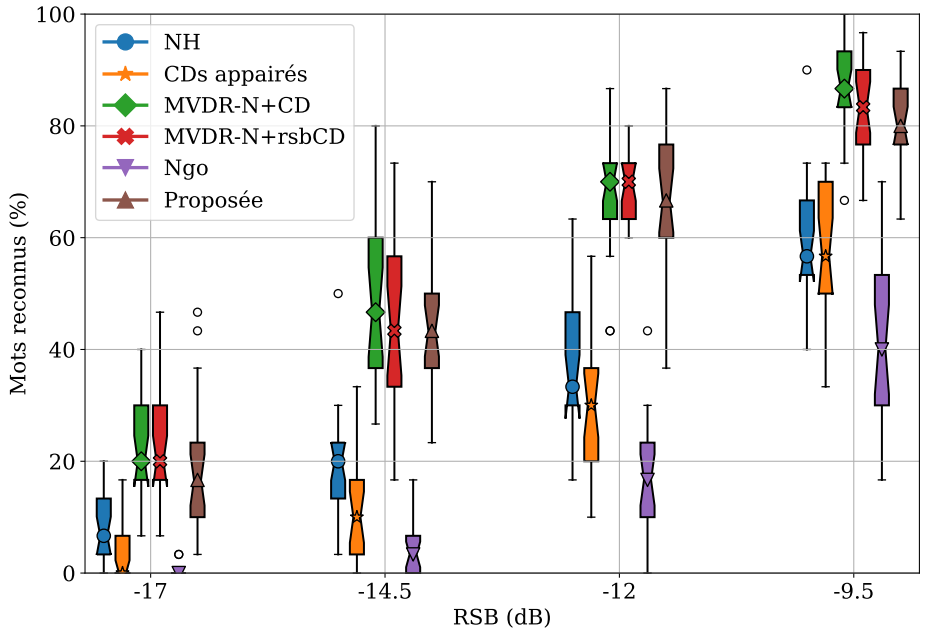


FIGURE 3.7 – Pourcentage de mots reconnus en fonction du RSB pour les six conditions testées. Chaque boîte à moustaches agrège les résultats de tous les sujets. Pour chaque sujet, trente mots ont été évalués (pas de quantification de 3,33 %). Les limites hautes et basses des boîtes représentent respectivement le troisième et le premier quartile, *i.e.* la plage inter-quartile (IQR). La médiane est représentée par le marqueur à l'intérieure. Les moustaches indiquent les valeurs ne dépassant pas le premier ou troisième quartile ± 1.5 le IQR et les cercles en dehors des boîtes représentent les horsains.[3]

car les RSBs choisis pour l'évaluation ne permettent pas de saisir toute la phase de transition de la sigmoïde. Néanmoins, comme nous le verrons dans la suite, les résultats sont suffisamment différents des autres conditions pour ne pas avoir d'ambigüité sur leur interprétation.

Pour le SRT, on analyse la significativité statistique sur la moyenne grâce à une Analyse de la Variance de mesures répétées (rANOVA) à un facteur. Ce test fait l'hypothèse de normalité multivariée (sphéricité) et d'homoscédasticité de chaque groupe de données. On teste alors au préalable que nos données remplissent bien ces critères grâce à un test de Mauchly pour la normalité et de Bartlett pour l'homoscédasticité. Pour compenser l'augmentation du risque de surdétection dû aux tests de plusieurs conditions, on emploie un test *post-hoc* de Tukey. Pour considérer que la condition est remplie, on se fixe comme critère un seuil de valeur-p de 0,05. Pour le pourcentage de la parole, on analyse la

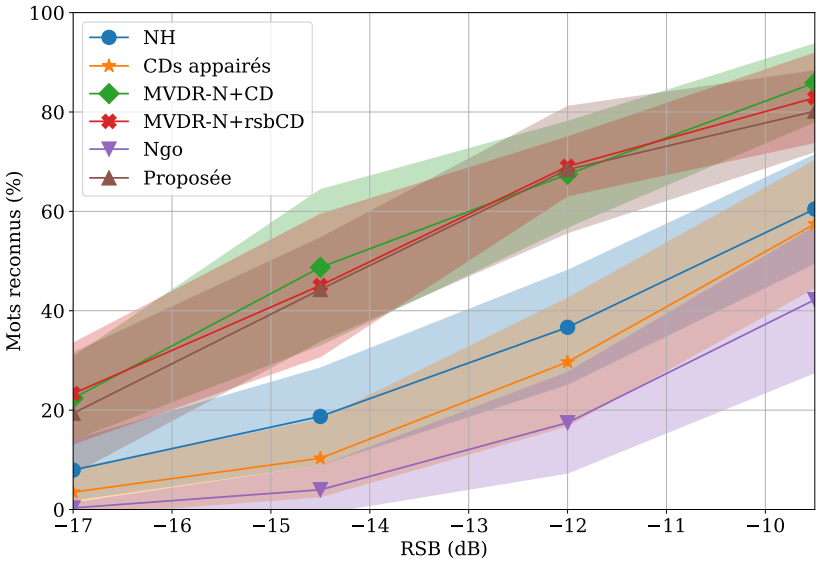


FIGURE 3.8 – Pourcentage de mots reconnus en fonction du RSB pour les six conditions testées (même données que dans la Fig. 3.7 mais représentées différemment). Le trait plein représente la moyenne et la surface l'écart-type.

significativité statistique indépendamment pour chaque RSB testé grâce à une ANOVA de Friedman par rangs. Contrairement à l'rANOVA, celle-ci s'intéresse à la médiane et s'affranchit ainsi des hypothèses sur la loi de distribution et sur l'égalité des variances. En effet, contrairement au SRT, une majeure partie des paires de conditions ne remplissent pas la condition homoscédasticité et certaines ne remplissent pas non plus la condition de normalité. Ceci est tout à fait attendu car le score est borné entre 0 et 100 %, les distributions viennent donc s'écraser sur ces limites. Pour compenser l'augmentation du risque de surdétection dû aux tests de plusieurs conditions, on emploie un test *post-hoc* de Conover.

Tout d'abord, comparons la condition *normo-entendante* et *CDs appairés*. La différence de SRT est est faible en moyenne (0,5 dB) et statistiquement non-significative (p=0,63). La valeur-p est particulièrement élevée dans ce cas dû au test *post-hoc* de Tukey. Cependant, la pente est plus raide pour la condition *CDs appairés* menant à des performances de compréhension inférieures dans la partie basse de la sigmoïde. On peut compléter cette analyse en regardant les résultats de pourcentage de mots reconnus par RSB, illustré en Fig. 3.7. Cette fois-ci, la différence des moyennes et la significativité statistique deviennent plus marquées entre ces deux conditions pour tous les RSBs testés, à l'exception du plus élevé, -9,5 dB.

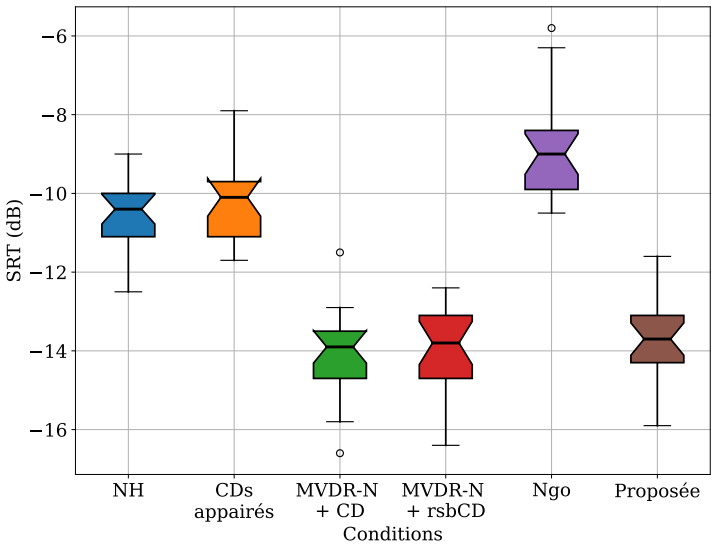


FIGURE 3.9 – SRT à 50 % pour les six conditions. L'estimation est effectuée en ajustant une fonction sigmoïde sur les données pour chaque sujet selon la méthode de [Brand and Kollmeier, 2002].

Les conditions *MVDR-N+CD* et *MVDR-N+rsbCD* montrent les mêmes performances en compréhension de la parole. La stratégie de pilotage des constantes de temps du CD introduite par [May et al., 2018] ne permet donc pas d'améliorer les performances de compréhension de la parole à ce niveau de RSB. On peut attribuer cela à deux phénomènes : (i) cette stratégie repose sur l'hypothèse que l'enveloppe du signal à l'entrée du compresseur est significativement plus grande en présence de la voix que sans ; (ii) à un RSB si bas, les erreurs d'estimation de détection de la parole sont importantes et dégradent donc le fonctionnement de l'algorithme.

Tant en matière de SRT que de pourcentage de mots reconnus pour tous les RSBs testés, la solution proposée montre des résultats statistiquement non-significativement différents de ceux obtenus avec les méthodes adoptant une concaténation sérielle des étages de beamforming et de compression (voir Tab. 3.6).

Les conditions *Proposée* et *Ngo*, bien que basées sur une combinaison des étages de beamforming et de compression très semblables, ont des performances de compréhension de la parole très différentes. Notre proposition obtient des scores équivalents aux solutions sérielles (*MVDR-N+CD* et *MVDR-N+rsbCD*) alors que la condition *Ngo* est très en dessous de toutes les autres conditions (1,5 dB de SRT au dessus de la condition *normo-entendante*). Cela s'explique très bien par l'implémentation du beamformer. En effet, *Ngo* s'appuie sur un

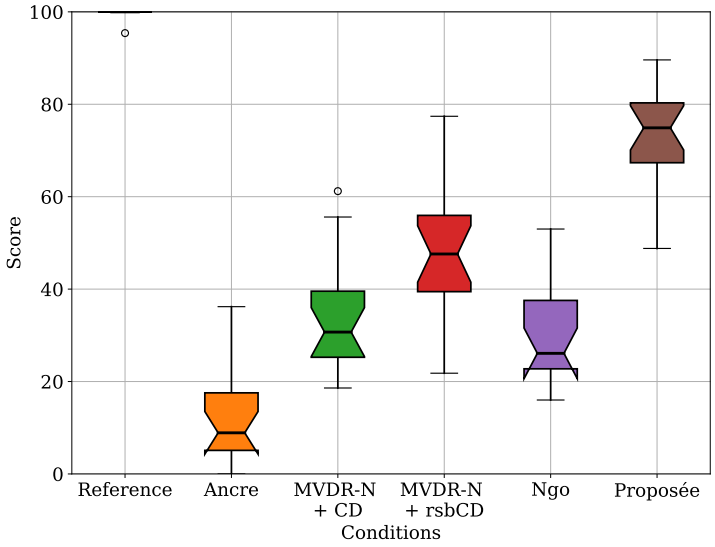


FIGURE 3.10 – Score du test de préférence MUSHRA pour les six conditions. Chaque boîte à moustaches agrège les moyennes des résultats de chaque sujet.

MWF local et non binaural, *i.e.* au niveau de chaque oreille et donc basé sur les signaux de deux microphones seulement. De plus, l'implémentation du MWF, décrite en [Ngo et al., 2012], ne permet pas d'être efficace à des RSBs d'entrée aussi bas. De fait, presque tous les points temps-fréquence sont classés comme contenant uniquement du bruit, ce qui ne permet ni d'avoir une bonne estimation de la matrice de covariance de la source, ni du gain de compression (voir Eq. (2.132, p.71) et les détails de l'implémentation dans [Ngo et al., 2012]).

Test de qualité subjective

Les résultats du test MUSHRA sont illustrés en Fig. 3.10. La significativité statistique sur la moyenne est obtenue grâce à une Analyse de la Variance de mesures répétées (rANOVA) à un facteur. Ce test fait l'hypothèse de sphéricité et d'homoscédasticité de chaque groupe de données. On vérifie alors au préalable que nos données remplissent bien ces critères grâce à un test de Mauchly pour la sphéricité et de Bartlett pour l'homoscédasticité. Pour compenser l'augmentation du risque de surdétection dû aux tests de plusieurs conditions, on emploie un test *post-hoc* de Tukey. Pour considérer que la condition est remplie, on se fixe comme critère un seuil de valeur-p de 0,05. Les conditions de référence et d'ancre basse sont exclues de cette analyse et les quatre conditions testées remplissent bien ces critères. Les sujets sont parvenus à distinguer la référence cachée en la notant systématiquement à 100, à l'exception d'un

sujet. Comme souhaité dans le protocole expérimental, l'ancre obtient la note médiane la plus basse avec 8,9.

Les différences des moyennes de chaque paire de conditions sont statistiquement significatives ($p < 0,005$) à l'exception de la paire *Ngo* et *MVDR-N+CD* ($p = 0,89$). Les sujets ont noté ces deux dernières assez bas, 26,2 et 30,7 respectivement, correspondant au qualificatif *poor*. La condition *Ngo* a la particularité de rendre la localisation des sources perçues très instable du fait que le traitement est effectué indépendamment à droite et à gauche.

On peut ensuite noter que les sujets ont en moyenne préféré l'ajout d'un compresseur piloté par le RSB (*MVDR-N+CD*), contrairement à ce qui a été observé pour le test de compréhension de la parole. Cet ajout permet d'augmenter la note de 17 points en moyenne. Bien que cet algorithme altère le timbre de la scène sonore dû à la dépendance de l'atténuation au dernier segment de parole détecté, il améliore significativement le RSB dans ces conditions (RSB d'entrée de 0 dB suivi d'un beamformer MVDR-N) contrairement au test de parole (RSB plus bas) [May et al., 2018]. Les sujets ont donc vraisemblablement préféré une amélioration du RSB à une dégradation du timbre.

Enfin, on remarque que la solution proposée se démarque positivement des autres, tant en relatif qu'en absolu. En effet, la solution proposée est majoritairement préférée (86 % des sujets) par rapport aux autres mais obtient aussi une note absolue de 75, avec les trois-quarts des notes se trouvant au dessus de 68, correspondant à la tranche haute du qualificatif *good*.

3.5 Conclusion

La compression de dynamique et le débruitage multicanal sont les traitements principaux dans les prothèses auditives. Leur interaction est peu étudiée et la manière de les associer n'est pas trivial. Dans ce chapitre, nous avons proposé de traiter conjointement de compensation des pertes auditives et de débruitage au sein d'un même problème d'optimisation. Cela explicite en particulier le rôle du banc de filtre et de l'appairage gauche/droite dans la compression de la parole. La solution proposée a été comparée, sur des critères aussi bien objectifs que perceptifs, aux algorithmes faisant référence dans la littérature, en particulier le beamformer MVDR-N suivi du rsbCD. L'évaluation objective a permis de montrer que la méthode proposée parvient à augmenter le RSB en sortie pour des RSBs d'entrée supérieur à 0 dB, tout en préservant la cohérence interaurale et l'enveloppe de la composante de bruit beaucoup mieux que les autres méthodes.

L'évaluation perceptive consiste en un test de compréhension de la parole en présence d'un bruit diffus de cafétéria, ainsi qu'un test de préférence subjective

où les sujets évaluent la préservation du reste de la scène sonore. Les résultats montrent que par rapport à la combinaison sérielle du beamformer et du CD, la solution proposée permet de mieux préserver la scène sonore globale tout en obtenant des performances de compréhension de la parole similaires. Cette solution, bien que très proche de celle de [Ngo et al., 2012], dans l'architecture globale, montre des performances bien meilleures sur les deux critères. Nous avons aussi montré que la stratégie du CD piloté par le RSB était préférée par les sujets pour un RSB d'entrée de 0 dB mais qu'elle était inefficace pour améliorer la compréhension de la parole pour la gamme de RSBs critiques (< -10 dB), ce qui va dans le sens de [May et al., 2018] et [Kowalewski et al., 2020].

Dans la méthode proposée, l'estimation de la composante de bruit se fait de manière indépendante à droite et à gauche. Il est donc impossible de prendre en compte les indices interauraux pour celle-ci. Le chapitre suivant sera dédié à l'investigation de cet aspect. Par ailleurs, la manière d'estimer le gain de compression appliqué à la source de parole n'est pas triviale et la solution proposée ici peut surement être améliorée.

Chapitre 4

Estimateur de bruit préservant les indices interauraux

4.1 Introduction

Dans le chapitre précédent, nous avons vu qu'une optimisation conjointe du débruitage et de la compression de dynamique faisait apparaitre dans la solution une architecture parallèle où la source de parole et la composante de bruit ambiant sont séparées et traitées indépendamment avant d'être recombinées en sortie. Lorsqu'on examine de plus près l'estimateur de la composante de bruit, on se rend compte que c'est un *nullformer*, *i.e.* un filtre spatial réduisant le niveau en fonction de la direction, voir les diagrammes de directivité en Fig. 4.2(a et b). On peut interpréter le diagramme de directivité en sortie du filtre comme les nouvelles HRTFs de l'auditeur (hors de l'axe de visée) et, de même la manière, les réponses impulsionnelles associées à ce diagramme de directivité comme les nouvelles HRIRs (voir Fig. 4.1(a et b)).

Nous allons voir dans ce chapitre qu'en l'état, le nullformer est capable de préserver globalement les HRTFs sur l'ensemble de la sphère mais qu'il distord en grande partie les indices interauraux (ITD et ILD). Or, la détermination des filtres des nullformers droite et gauche étant effectuée indépendamment de chaque coté, il est impossible de les prendre en compte dans la formulation actuelle du problème d'optimisation. On s'inspirera donc de la littérature sur les beamformers préservant les indices interauraux, tel que le MWF avec préservation de la fonction de transfert interaurale (MWF-ITF) [Cornelis et al., 2010], pour développer des nullformers en ce sens. Nous verrons qu'il existe plusieurs obstacles à dépasser pour rendre ce problème d'optimisation simple à résoudre. En particulier, nous proposerons en section 4.2 de réinterpréter le nullformer à l'aune d'un formalisme déterministe, plutôt que probabiliste. Ceci nous permettra de développer en section 4.3 deux algorithmes originaux basés sur la préservation de la fonction de transfert interaurale (de l'anglais *Interaural Transfer Function*, ITF), encodant l'ITD et l'ILD. Cette réinterprétation ouvre aussi la voie à d'autres possibles applications que nous discuterons. Enfin, les nullformers préservant l'ITF proposés seront comparés en section 4.4 au nullformer utilisé dans le chapitre précédent en matière de réduction du niveau dans la direction cible, d'erreur d'ITD et d'erreur d'ILD.

4.2 Reformulation déterministe du nullformer

Dans cette section, on montre l'équivalence du nullformer utilisé dans le chapitre précédent, développé dans un cadre probabiliste, et de celui visant à annuler la direction cible et à minimiser la différence entre son diagramme de directivité et les HRTFs originales des microphones de références pour les oreilles droite et gauche.

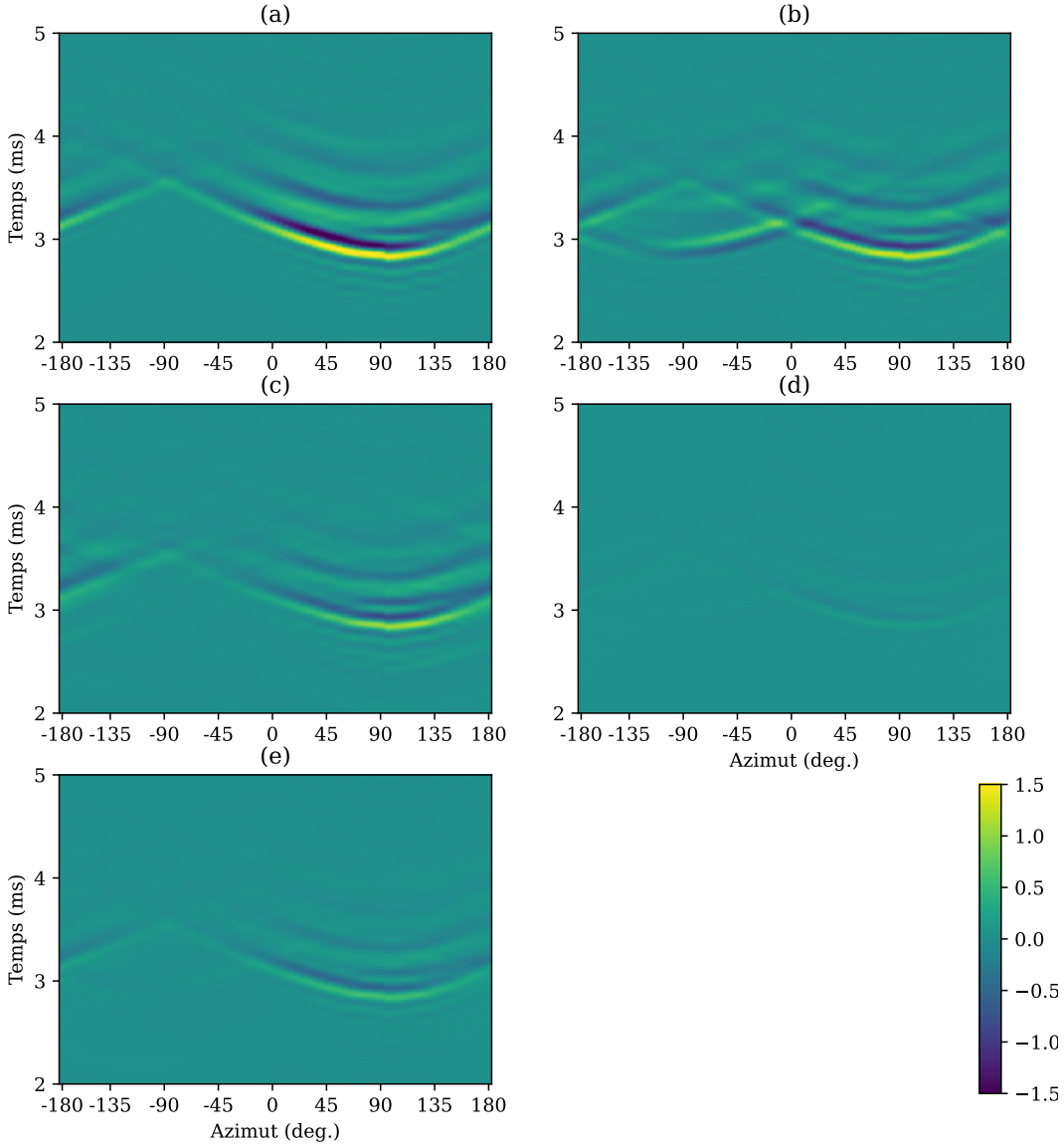


FIGURE 4.1 – HRIR gauche sur le plan horizontal au microphone de référence (a) et en sortie du H-NF (b), du JLC-NF (c), du ITF1-NF (d) et du ITF2-NF (e).

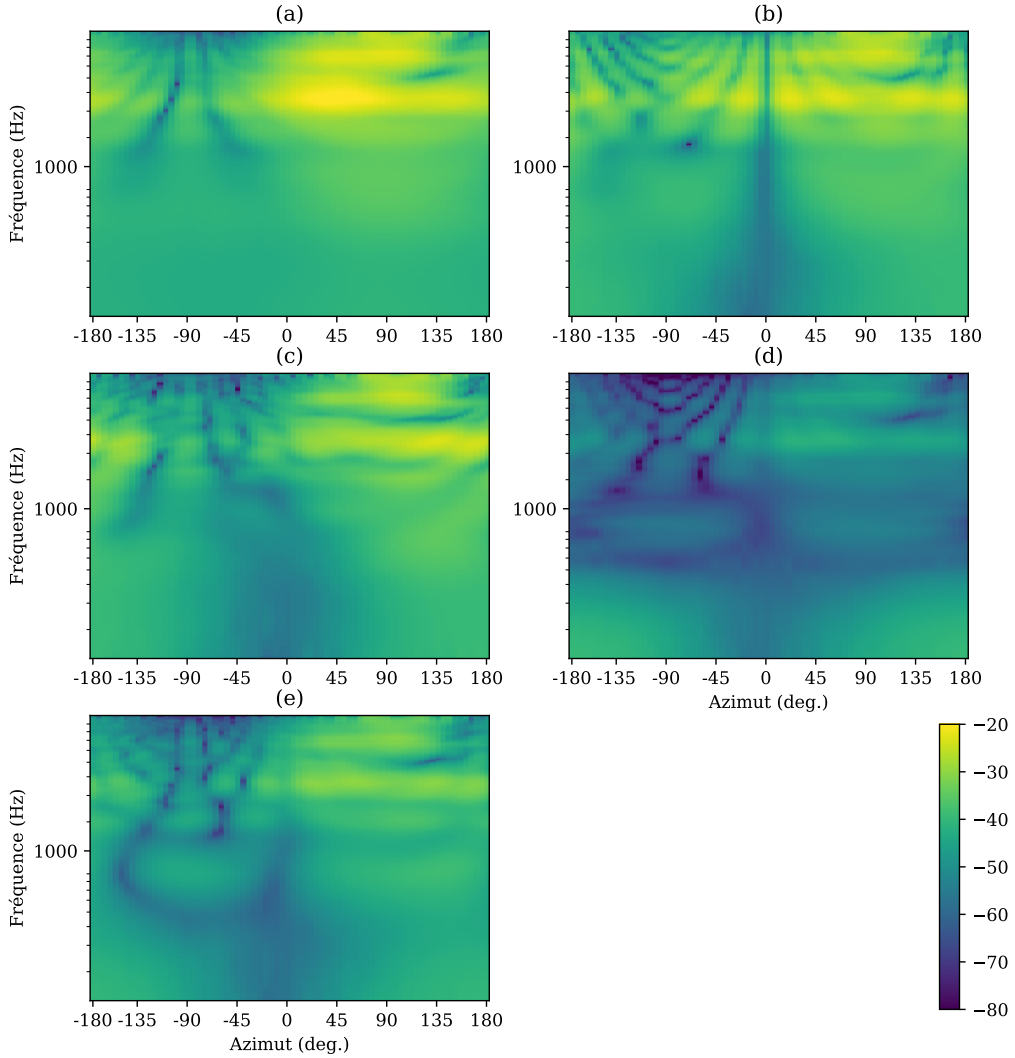


FIGURE 4.2 – Amplitude en dB de l'HRTF gauche sur le plan horizontal au microphone de référence (a) et en sortie du H-NF (b), du JLC-NF (c), du ITF1-NF (d) et du ITF2-NF (e).

4.2.1 Preuve de l'équivalence

On rappelle l'expression du filtre de nullforming, pour l'oreille gauche, employé dans le chapitre précédent (voir Eq. (3.11, p.83)) :

$$\mathbf{w}_{\text{NF, L}} = \mathbf{q}_{\text{L}} - \left(\xi \mathbf{h}\mathbf{h}^H + \mathbf{\Gamma_n}\right)^{-1} \xi \mathbf{h} h_{\text{L}}^*, \tag{4.1}$$

où $\xi \in \mathbb{R}_+$ est le RSB d'entrée, $\mathbf{\Gamma_n} \in \mathbb{C}^{M\times M}$ la matrice de covariance normalisée du bruit, $\mathbf{h} \in \mathbb{C}^M$ les ATFs de la direction cible, h_{L} l'élément de $\mathbf{h}$ correspondant au microphone de référence pour l'oreille gauche et $\mathbf{q}_{\text{L}}$ le vecteur nul à l'exception de l'élément correspondant au microphone de référence pour l'oreille gauche, réglé à 1. Sauf mention contraire, nous reprenons les notations mathématiques du chapitre précédent et nous omettons l'indice fréquentiel k. En utilisant $\mathbf{h}h_{\text{L}}^* = \mathbf{h}\mathbf{h}^H\mathbf{q}_{\text{L}}$, on peut réécrire $\mathbf{w}_{\text{NF, L}}$ sous la forme suivante :

$$\mathbf{w}_{\text{NF, L}} = \left(\xi \mathbf{h}\mathbf{h}^H + \mathbf{\Gamma_n}\right)^{-1} \mathbf{\Gamma_n}\mathbf{q}_{\text{L}}. \tag{4.2}$$

Le nullformer pour l'oreille droite peut être exprimé similairement.

On propose ici le problème d'optimisation suivant : minimiser la puissance du diagramme de directivité dans la direction cible, notée d_0, ainsi que la différence entre le diagramme de directivité et les HRTFs du microphone de référence. On rappelle que le diagramme de directivité pour l'oreille gauche est défini comme suit :

$$P_{\text{L}}(d) = \mathbf{w}_{\text{L}}^H \mathbf{h}_d, \tag{4.3}$$

avec d l'indice de direction et $\mathbf{h}_d$ les ATFs pour une source située à la direction d (et donc $\mathbf{h} = \mathbf{h}_{d_0}$). La fonction de coût peut alors s'écrire comme suit :

$$J_{\text{H-NF, L}}(\mathbf{w}_{\text{L}}) = \alpha_0 |P_{\text{L}}(d_0)|^2 + \sum_d w_d \left|h_{d,\text{L}} - P_{\text{L}}(d)\right|^2, \tag{4.4}$$

$$= \alpha_0 |\mathbf{w}_{\text{L}}^H \mathbf{h}|^2 + \sum_d w_d \left|h_{d,\text{L}} - \mathbf{w}_{\text{L}}^H \mathbf{h}_d\right|^2, \tag{4.5}$$

où $\alpha_0 \in \mathbb{R}_+$ est un paramètre pour arbitrer entre atténuation dans la direction cible et préservation des HRTFs, $h_{d,\text{L}}$ est l'HRTF gauche vers laquelle on veut tendre, ici c'est celle du microphone de référence pour l'oreille gauche, et w_d est le poids correspondant à la surface représentée par la direction d dans le cas où l'échantillonnage de l'espace n'est pas régulier, ce qui est la norme lorsqu'on considère la sphère et non plus uniquement le plan horizontal [Harder, 2015]. En annulant le gradient de $J_{\text{H-NF, L}}(\mathbf{w}_{\text{L}})$, on obtient la solution du nullformer

préservant les HRTFs (H-NF) que l'on peut écrire comme suit :

$$\mathbf{w}_{\text{H-NF, L}} = \left(\alpha_0 \mathbf{h}\mathbf{h}^H + \sum_d w_d \mathbf{h}_d \mathbf{h}_d^H\right)^{-1} \sum_d w_d \mathbf{h}_d h_{d,\text{L}}^*, \tag{4.6}$$

$$= \left(\alpha_0 \mathbf{h}\mathbf{h}^H + \hat{\mathbf{\Gamma}}_\mathbf{n}\right)^{-1} \hat{\mathbf{\Gamma}}_\mathbf{n} \mathbf{q}_\text{L}, \tag{4.7}$$

où $\hat{\mathbf{\Gamma}}_\mathbf{n}$ est l'estimation de la matrice de cohérence d'un bruit spatialement diffus selon [Stadler and Rabinowitz, 1993, Lotter and Vary, 2006] :

$$\hat{\mathbf{\Gamma}}_\mathbf{n} = \sum_d w_d \mathbf{h}_d \mathbf{h}_d^H. \tag{4.8}$$

On note alors l'égalité entre l'Eq. (4.2) et l'Eq. (4.7) lorsque l'estimateur de la matrice de cohérence du bruit correspond à celle d'un bruit diffus et $\alpha_0 = \xi$.

4.2.2 Analyse du résultat

Cette reformulation permet de comprendre que le nullformer considérant un bruit spatialement diffus peut être aussi vu comme celui préservant au mieux les HRTFs d'une source quelque soit sa direction d'arrivée. Il n'est donc plus seulement le plus adapté au traitement d'une scène sonore composée d'un bruit spatialement diffus, mais devient alors le nullformer à employer lorsque l'on n'a pas d'information sur la composition de la scène sonore en dehors de la source cible.

En Fig. 4.2(a et b) sont illustrées l'amplitude des HRTFs du microphone de référence de l'oreille gauche et du diagramme de directivité en sortie du nullformer sur le plan horizontal. On observe bien une atténuation dans la direction frontale, ainsi que la reproduction de l'ombrage acoustique de la tête (au dessus de 1 kHz pour les azimuts négatifs) -bien que distordu- et des indices spectraux (au dessus de 3 kHz pour les azimuts positifs).

Perspectives On remarque que, le fait de se placer dans ce cadre déterministe rend explicite l'apparition de l'HRTF cible, qui jusqu'ici était implicitement celle du microphone de référence. Cela ouvre la possibilité d'envisager de prendre des HRTFs cibles différentes des originales. En effet, [Dieudonné and Francart, 2018] ont montré l'intérêt d'ajouter de l'ILD en basse fréquence sur les performances de localisation et de compréhension de la parole pour des malentendants appareillés d'une prothèse auditive d'un coté et d'un implant cochléaire de l'autre.

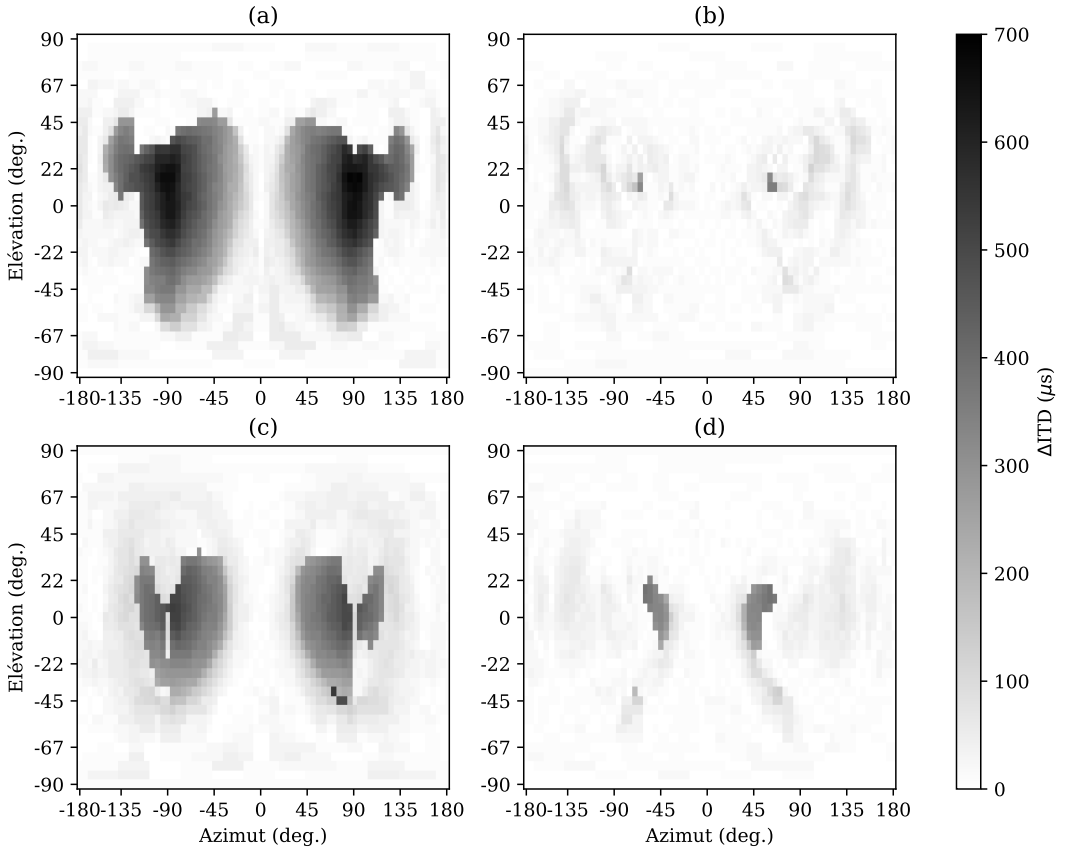


FIGURE 4.3 – Planisphère de l'erreur absolue d'ITD du H-NF (a), du JLC-NF (b), du ITF1-NF (c) et du ITF2-NF (d).

4.3 Nullformers préservant les indices interauraux

En Fig. 4.2(a et b), on remarque que le nullformer présenté jusqu'ici est efficace pour réduire le niveau dans la direction frontale tout en préservant relativement bien l'amplitude de l'HRTF. Cependant, lorsqu'on s'intéresse à l'ILD (voir Fig. 4.4a) et plus particulièrement à l'ITD (voir Fig. 4.3a), on note que l'erreur est très importante, en particulier sur l'hémisphère frontale. En regardant de plus près les réponses impulsionnelles, illustrées en Fig. 4.1(a et b), on observe que le nullformer mélange les microphones droite et gauche en opposition de phase de façon à annuler le signal qui provient de la direction frontale. De ce fait, un front d'onde parasite apparait, correspondant à la contri-

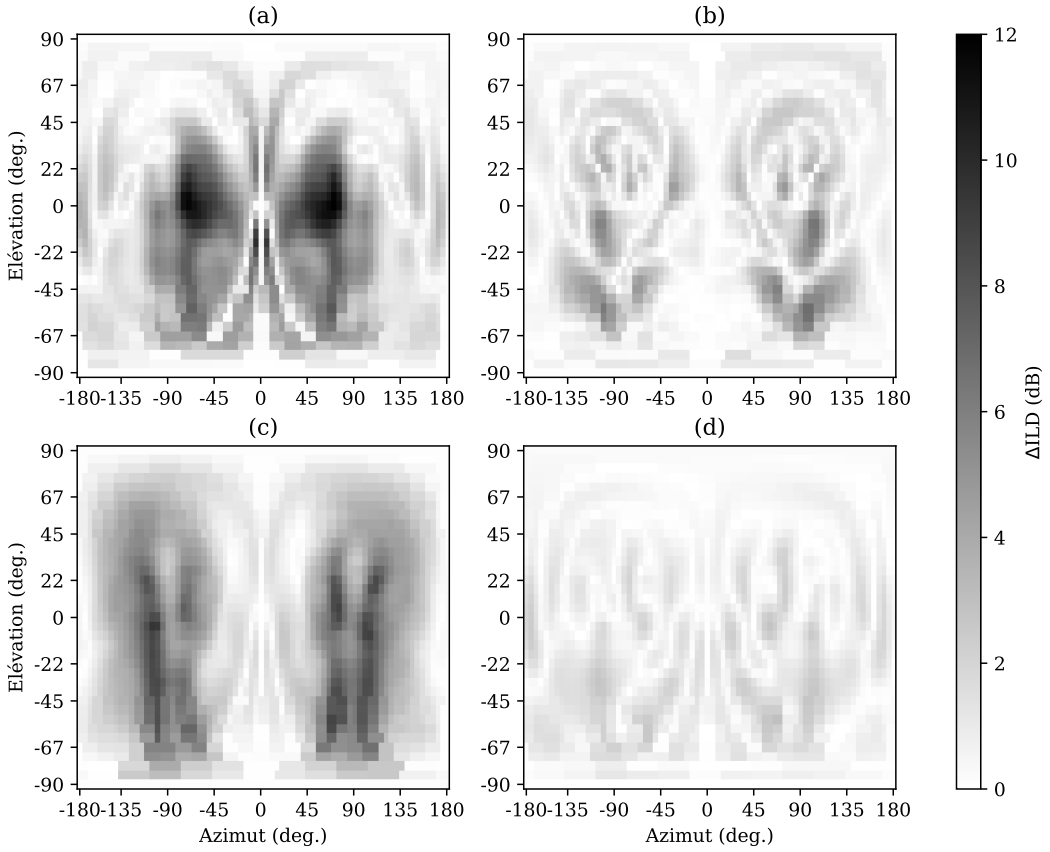


FIGURE 4.4 – Planisphère de l'erreur absolue d'ILD du H-NF (a), du JLC-NF (b), du ITF1-NF (c) et du ITF2-NF (d).

bution des microphones contralatéraux (placés à droite dans l'exemple donné). Celui-ci est particulièrement important pour les directions contralatérales car il est du même ordre de grandeur et en avance de phase par rapport au front d'onde attendu pour les HRIRs de cette oreille. En effet, lorsqu'un son nous parvient avec différents temps d'arrivée, nous avons tendance à considérer le premier comme porteur de l'information de direction, on appelle cela l'effet de précédence ou effet de Haas [Gardner, 1968]. Le problème d'optimisation permettant d'obtenir $\mathbf{w}_{\text{H-NF, L}}$ et $\mathbf{w}_{\text{H-NF, R}}$ étant formulé indépendamment à droite et à gauche, il est impossible de prendre en compte les indices interauraux.

Dans cette section, nous proposons différents algorithmes permettant de prendre en compte l'information interaurale dans le processus d'optimisation. Pour ce faire, nous allons nous inspirer de ce qui a été fait dans la littérature

pour les beamformers préservant les indices de localisation de la scène sonore, comme le MWF-ITF introduit par [Klasen et al., 2006], préservant l'ITF. Celle-ci est définie comme le rapport entre l'HRTF gauche et l'HRTF droite, portant ainsi aussi bien l'information d'ILD que d'ITD pour une direction donnée. Le MWF-ITF a montré son efficacité à préserver les indices de localisation interauraux pour les sources de bruits localisées dans l'espace [Doclo et al., 2006, Van den Bogaert et al., 2007]. Celui-ci est développé à partir d'un formalisme probabiliste dont nous allons voir qu'il ne permet pas de prendre correctement en compte un bruit spatialement diffus. Cet obstacle peut être levé en utilisant le formalisme déterministe développé dans la section précédente.

Afin de pouvoir prendre en compte les indices interauraux dans le problème d'optimisation, nous réalisons une optimisation conjointe des filtres droite et gauche, $\mathbf{w}_\mathrm{R}$ et $\mathbf{w}_\mathrm{L}$, en les considérant comme une seule et même variable d'optimisation en les concaténant de la manière suivante :

$$\mathbf{w} = \begin{bmatrix} \mathbf{w}_\mathrm{L} \\ \mathbf{w}_\mathrm{R} \end{bmatrix}. \tag{4.9}$$

4.3.1 Nullformer préservant les ITFs dans un cadre probabiliste (ITF1-NF)

Comme nous l'avons décrit en section 2.2 (p.34), les filtres de beamforming sont la plupart du temps basés sur un MWF, lui même développé dans un cadre probabiliste. C'est pourquoi, il est courant de voir le MWF avec préservation de la fonction de transfert interaurale (MWF-ITF) dérivé à partir d'une définition probabiliste de l'ITF [Cornelis et al., 2010, Marquardt, 2015], qui pour l'entrée est définie comme suit :

$$\mathrm{ITF1}_\mathrm{in} = \frac{\mathbb{E}\left[n_\mathrm{L} n_\mathrm{R}^*\right]}{\mathbb{E}\left[|n_\mathrm{R}|^2\right]} = \gamma, \tag{4.10}$$

et en sortie :

$$\mathrm{ITF1}_\mathrm{out}(\mathbf{w}) = \frac{\mathbf{w}_\mathrm{L}^H \mathbf{n}}{\mathbf{w}_\mathrm{R}^H \mathbf{n}}. \tag{4.11}$$

La terme de coût associé à la préservation de cette ITF est alors défini de la manière suivante [Cornelis et al., 2010] :

$$J_\mathrm{ITF1}(\mathbf{w}) = \mathbb{E}\left[\left|\mathrm{ITF1}_\mathrm{out}(\mathbf{w}) - \mathrm{ITF1}_\mathrm{in}\right|^2\right], \tag{4.12}$$

$$= \mathbb{E}\left[\left|\frac{\mathbf{w}_\mathrm{L}^H \mathbf{n}}{\mathbf{w}_\mathrm{R}^H \mathbf{n}} - \gamma\right|^2\right], \tag{4.13}$$

et est souvent simplifié de la manière suivante [Doclo et al., 2006, Van den Bogaert et al., 2007, Cornelis et al., 2010] :

$$J_{\mathrm{ITF1}}(\mathbf{w}) = \frac{\mathbb{E}\left[|\mathbf{w}_{\mathrm{L}}^H\mathbf{n} - \gamma\mathbf{w}_{\mathrm{R}}^H\mathbf{n}|^2\right]}{\mathbb{E}\left[|\mathbf{w}_{\mathrm{R}}^H\mathbf{n}|^2\right]}, \tag{4.14}$$

qui est vraie lorsque la composante de bruit est composée d'une source localisée dans l'espace. Ce terme ajouté à la fonction de coût initiale du nullformer ne permet pas d'aboutir à une solution analytique. Il est alors d'usage d'omettre le dénominateur permettant d'aboutir à l'approximation quadratique suivante :

$$\tilde{J}_{\mathrm{ITF1}}(\mathbf{w}) = \mathbb{E}\left[\left|\mathbf{w}_{\mathrm{L}}^H\mathbf{n} - \mathrm{ITF1}_{\mathrm{in}}\mathbf{w}_{\mathrm{R}}^H\mathbf{n}\right|^2\right] \tag{4.15}$$

$$= \mathbf{w}_{\mathrm{L}}^H\boldsymbol{\Gamma}_{\mathbf{n}}\mathbf{w}_{\mathrm{L}} - |\mathrm{ITF}_{\mathrm{in}}|^2\mathbf{w}_{\mathrm{R}}^H\boldsymbol{\Gamma}_{\mathbf{n}}\mathbf{w}_{\mathrm{R}} - \mathrm{ITF}_{\mathrm{in}}\mathbf{w}_{\mathrm{R}}^H\boldsymbol{\Gamma}_{\mathbf{n}}\mathbf{w}_{\mathrm{L}} - \mathrm{ITF1}_{\mathrm{in}}^*\mathbf{w}_{\mathrm{L}}^H\boldsymbol{\Gamma}_{\mathbf{n}}\mathbf{w}_{\mathrm{R}} \tag{4.16}$$

$$= \mathbf{w}^H\begin{bmatrix} \boldsymbol{\Gamma}_{\mathbf{n}} & -\gamma^*\boldsymbol{\Gamma}_{\mathbf{n}} \\ -\gamma\boldsymbol{\Gamma}_{\mathbf{n}} & |\gamma|^2\boldsymbol{\Gamma}_{\mathbf{n}} \end{bmatrix}\mathbf{w}. \tag{4.17}$$

Cette approximation est valable lorsque le bruit est composé d'une source localisée. Dans notre cas, cela ne fait pas du tout sens. Néanmoins, c'est cette version de l'algorithme qui est souvent retrouvé dans la littérature [Cornelis et al., 2010, Marquardt, 2015]. Afin d'éviter la confusion, nous intégrons cette version de l'algorithme dans notre évaluation de sorte à mettre en lumière ses limites.

4.3.2 Nullformer préservant les ITFs dans un cadre déterministe (ITF2-NF)

L'ITF est historiquement définie comme le rapport entre les HRTFs gauche et droite [Klasen et al., 2006, Doclo et al., 2006, Van den Bogaert et al., 2007]. Ici, on fait aussi apparaitre sa dépendance à la DOA, notée $d \in \mathcal{D}$ où $\mathcal{D}$ est l'ensemble des directions considérées, pour l'ITF aux microphones de référence pour les oreilles droite et gauche :

$$\mathrm{ITF2}_{\mathrm{in}}(d) = \frac{h_{d,\mathrm{L}}}{h_{d,\mathrm{R}}} = \gamma_d, \tag{4.18}$$

et en sortie des nullformers droite et gauche :

$$\mathrm{ITF2}_{\mathrm{out}}(d, \mathbf{w}) = \frac{\mathbf{w}_{\mathrm{L}}^H\mathbf{h}_d}{\mathbf{w}_{\mathrm{R}}^H\mathbf{h}_d}. \tag{4.19}$$

Grâce à la formulation du problème d'optimisation de nullforming déterministe plutôt que probabiliste, introduite en section 4.2, nous pouvons écrire le terme

d'erreur associé à cette définition de l'ITF de la manière suivante :

$$J_{\mathrm{ITF2}}(\mathbf{w}) = \sum_d w_d \left|\mathrm{ITF2}_{\mathrm{out}}(d,\mathbf{w}) - \mathrm{ITF2}_{\mathrm{in}}(d)\right|^2, \tag{4.20}$$

$$= \sum_d w_d \left|\frac{\mathbf{w}_{\mathrm{L}}^H \mathbf{h}_d}{\mathbf{w}_{\mathrm{R}}^H \mathbf{h}_d} - \frac{h_{d,L}}{h_{d,R}}\right|^2, \tag{4.21}$$

$$= \sum_d w_d \frac{\left|\mathbf{w}_{\mathrm{L}}^H \mathbf{h}_d - \gamma_d \mathbf{w}_{\mathrm{R}}^H \mathbf{h}_d\right|^2}{\left|\mathbf{w}_{\mathrm{R}}^H \mathbf{h}_d\right|^2}. \tag{4.22}$$

Afin d'obtenir une solution analytique, on simplifie l'expression en retirant son dénominateur similairement à [Klasen et al., 2006, Van den Bogaert et al., 2007, Cornelis et al., 2010, Marquardt et al., 2015] de sorte à obtenir un terme de coût quadratique :

$$\tilde{J}_{\mathrm{ITF2}}(\mathbf{w}) = \sum_d w_d \left|\mathbf{w}_{\mathrm{L}}^H \mathbf{h}_d - \gamma_d \mathbf{w}_{\mathrm{R}}^H \mathbf{h}_d\right|^2, \tag{4.23}$$

$$\begin{aligned} = \sum_d w_d (&\mathbf{w}_{\mathrm{L}}^H \mathbf{h}_d \mathbf{h}_d^H \mathbf{w}_{\mathrm{L}} + |\gamma_d|^2 \mathbf{w}_{\mathrm{R}}^H \mathbf{h}_d \mathbf{h}_d^H \mathbf{w}_{\mathrm{R}} \\ &- \gamma_d \mathbf{w}_{\mathrm{R}}^H \mathbf{h}_d \mathbf{h}_d^H \mathbf{w}_{\mathrm{L}} - \gamma_d^* \mathbf{w}_{\mathrm{R}}^H \mathbf{h}_d \mathbf{h}_d^H \mathbf{w}_{\mathrm{R}}), \end{aligned} \tag{4.24}$$

$$= \mathbf{w}^H \begin{bmatrix} \sum_d w_d \mathbf{h}_d \mathbf{h}_d^H & -\sum_d w_d \gamma_d^* \mathbf{h}_d \mathbf{h}_d^H \\ -\sum_d w_d \gamma_d \mathbf{h}_d \mathbf{h}_d^H & \sum_d w_d |\gamma_d|^2 \mathbf{h}_d \mathbf{h}_d^H \end{bmatrix} \mathbf{w}. \tag{4.25}$$

On remarque que les Eq. (4.17) et (4.25) ne sont pas égales sauf si $|\mathcal{D}| = 1$. Comme expliqué plus haut, cela est attendu dans la mesure où, casser la fraction en Eq. (4.14) revient à faire cette hypothèse.

4.3.3 Correction de l'approximation quadratique

Lorsqu'on regarde les réponses impulsionnelles correspondant aux diagrammes de directivité en sortie des ITF1-NF et ITF2-NF, illustrées en Fig. 4.5(a et b), on remarque que le front d'onde parasite est mieux atténué à droite qu'à gauche. Cette dissymétrie est due à l'approximation quadratique opérée en Eq. (4.25). Pour y remédier, on propose de considérer l'ITF inverse, *i.e.* le rapport entre l'HRTF droite et l'HRTF gauche. En effet, la position au numérateur du terme correspondant à l'oreille gauche est purement arbitraire. On peut alors définir de même manière un terme de coût correspondant à la préservation de cette

ITF inverse et l'ajouter au terme de coût initial :

$$\hat{J}_{\text{ITF2}}(\mathbf{w}) = \sum_d w_d \left|\mathbf{w}_{\text{L}}^H \mathbf{h}_d - \gamma_d \mathbf{w}_{\text{R}}^H \mathbf{h}_d\right|^2 + \sum_d w_d \left|\mathbf{w}_{\text{R}}^H \mathbf{h}_d - \gamma_d^{-1} \mathbf{w}_{\text{L}}^H \mathbf{h}_d\right|^2 \quad (4.26)$$

$$= \mathbf{w}^H \begin{bmatrix} \sum_d w_d(|\gamma_d|^{-2}+1)\mathbf{h}_d\mathbf{h}_d^H & -\sum_d w_d(\gamma_d+\gamma_d^{-1})^*\mathbf{h}_d\mathbf{h}_d^H \\ -\sum_d w_d(\gamma_d+\gamma_d^{-1})\mathbf{h}_d\mathbf{h}_d^H & \sum_d w_d(|\gamma_d|^2+1)\mathbf{h}_d\mathbf{h}_d^H \end{bmatrix} \mathbf{w}. \quad (4.27)$$

$\hat{J}_{\text{ITF1}}(\mathbf{w})$ est défini similairement. Cette « symétrisation » de la fonction de coût de l'ITF permet de compenser le déséquilibre introduit par l'approximation quadratique. On peut voir le résultat de cette méthode illustré en Fig. 4.5(c et d) où l'on observe le rétablissement de la symétrie entre les deux HRIRs et l'adoucissement du front d'onde parasite des deux cotés plutôt que du seul coté droit.

En utilisant la correction de l'approximation quadratique introduite précédemment, on obtient finalement les problèmes d'optimisation suivants pour le nullformer préservant les ITFs selon la définition probabiliste (ITF1-NF) :

$$\mathbf{w}_{\text{ITF1-NF}} = \underset{\mathbf{w}}{\text{argmin}} \left\{ J_{\text{NF}}(\mathbf{w}) + \alpha_{\text{ITF}} \hat{J}_{\text{ITF1}}(\mathbf{w}) \right\}, \quad (4.28)$$

et pour le nullformer préservant les ITFs selon la définition déterministe (ITF2-NF) :

$$\mathbf{w}_{\text{ITF2-NF}} = \underset{\mathbf{w}}{\text{argmin}} \left\{ J_{\text{NF}}(\mathbf{w}) + \alpha_{\text{ITF}} \hat{J}_{\text{ITF2}}(\mathbf{w}) \right\}, \quad (4.29)$$

avec $\alpha_{\text{ITF}} \in \mathbb{R}_+$ un paramètre de pondération de la préservation des ITFs dans le processus d'optimisation, ainsi que $J_{\text{NF}}(\mathbf{w})$ la version binaurale de $J_{\text{H-NF, L}}(\mathbf{w}_{\text{L}})$ et $J_{\text{H-NF, R}}(\mathbf{w}\text{R})$:

$$J_{\text{NF}}(\mathbf{w}) = \mathbf{w}^H \mathbf{Q} \mathbf{w} - \mathbf{w}^H \mathbf{v} - \mathbf{v}^H \mathbf{w}, \quad (4.30)$$

où

$$\mathbf{Q} = \begin{bmatrix} \alpha_0 \mathbf{h}\mathbf{h}^H + \hat{\mathbf{\Gamma}}_{\mathbf{n}} & \mathbf{0}_{M\times M} \\ \mathbf{0}_{M\times M} & \alpha_0 \mathbf{h}\mathbf{h}^H + \hat{\mathbf{\Gamma}}_{\mathbf{n}} \end{bmatrix}, \quad (4.31)$$

et

$$\mathbf{v} = \begin{bmatrix} \hat{\mathbf{\Gamma}}_{\mathbf{n}} \mathbf{q}_{\text{L}} \\ \hat{\mathbf{\Gamma}}_{\mathbf{n}} \mathbf{q}_{\text{R}} \end{bmatrix}. \quad (4.32)$$

Dans ces trois dernières sous-sections, nous avons développé deux méthodes de nullforming préservant les indices interauraux basées sur la préservation de la fonction de transfert interaurale, encodant aussi bien l'ILD que l'ITD. Le terme de coût associé à cette fonction est non-convexe et ne permet donc pas

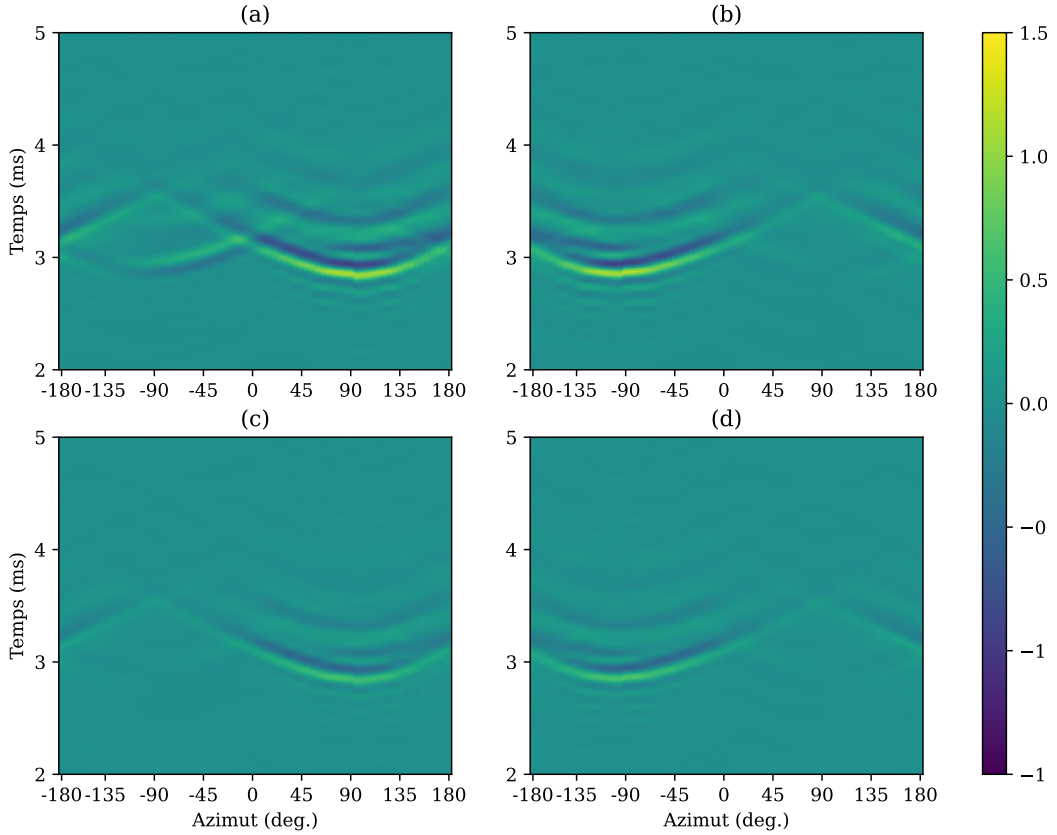


FIGURE 4.5 – Réponses impulsionnelles sur le plan horizontal en sortie du ITF2-NF pour les oreilles droite (b et d) et gauche (a et c) avec (c et d) et sans (a et b) la correction de l'approximation quadratique introduite en Eq. (4.26).

de trouver une solution analytique du filtre en l'état. Pour y parvenir, nous avons choisi de faire une approximation quadratique de ce terme comme il est d'usage. Cette approximation introduit une dissymétrie des diagrammes de directivités résultant. Nous avons alors proposé une méthode originale pour retrouver cette symétrie.

L'approximation quadratique n'est pas la seule possible. Dans la sous-section suivante, nous développons une autre méthode permise par le formalisme déterministe introduit en section 4.2.

4.3.4 Nullformer linéairement contraint (JLC-NF)

On introduit ici une nouvelle stratégie pour rendre convexe le problème d'optimisation incluant la préservation des ITFs. Pour cela, nous considérons l'hypothèse suivante : un $\hat{\mathbf{w}}$ qui minimise $J_{\text{NF}}(\mathbf{w}) + J_{\text{ITF2}}(\mathbf{w})$ a pour propriété d'annuler l'erreur d'ITF pour un sous-ensemble de P directions, noté $\tilde{\mathcal{D}}$. On peut l'écrire ainsi :

$$J_{\text{NF}}(\hat{\mathbf{w}}) + J_{\text{ITF2}}(\hat{\mathbf{w}}) \approx \min\{J_{\text{NF}}(\mathbf{w}) + J_{\text{ITF2}}(\mathbf{w})\}, \tag{4.33}$$

$$\text{avec} \quad \text{ITF2}_{\text{out}}(d, \hat{\mathbf{w}}) - \text{ITF2}_{\text{in}}(d) = 0 \quad \forall d \in \tilde{\mathcal{D}}, \tag{4.34}$$

et reformuler le système d'équations linéaires de la manière suivante :

$$\frac{\hat{\mathbf{w}}_{\text{L}}^H \mathbf{h}_d}{\hat{\mathbf{w}}_{\text{R}}^H \mathbf{h}_d} - \frac{h_{d,L}}{h_{d,R}} = 0 \qquad \forall d \in \tilde{\mathcal{D}}, \tag{4.35}$$

$$\mathbf{w}_{\text{L}}^H \mathbf{h}_d - \frac{h_{d,L}}{h_{d,R}} \mathbf{w}_{\text{R}}^H \mathbf{h}_d = 0 \qquad \forall d \in \tilde{\mathcal{D}}, \tag{4.36}$$

$$\mathbf{w}^H \begin{bmatrix} \mathbf{h}_d \\ -\frac{h_{d,L}}{h_{d,R}} \mathbf{h}_d \end{bmatrix} = 0 \qquad \forall d \in \tilde{\mathcal{D}}, \tag{4.37}$$

$$\mathbf{w}^H \underbrace{\begin{bmatrix} \mathbf{h}_{d_1}, & \ldots, & \mathbf{h}_{d_P} \\ -\frac{h_{d_1,L}}{h_{d_1,R}} \mathbf{h}_{d_1}, & \ldots, & -\frac{h_{d_P,L}}{h_{d_P,R}} \mathbf{h}_{d_P} \end{bmatrix}}_{\mathbf{C}} = \mathbf{0}_{1\times P}. \tag{4.38}$$

Cela permet d'utiliser la fonction de coût quadratique originale, $J_{\text{NF}}(\mathbf{w})$, mais en restreignant la recherche de la solution à un sous-espace où l'erreur d'ITF est nulle pour le sous-ensemble de directions $\tilde{\mathcal{D}}$. Or, on remarque que l'Eq. (4.38) fait apparaître que ce sous-espace est un hyper-plan de dimension $2M - P$ et donc permet de ramener le problème d'optimisation à une forme quadratique linéairement contrainte. On nomme la solution à ce problème le nullformer conjointement linéairement contraint (de l'anglais *jointly linearly constrained nullformer*, JLC-NF) que l'on peut écrire comme suit :

$$\mathbf{w}_{\text{JLC-NF}} = \underset{\mathbf{w}}{\text{argmin}} \{J_{\text{NF}}(\mathbf{w})\} \quad \text{s.t.} \quad \mathbf{w}^H \mathbf{C} = \mathbf{0}_{1\times P}, \tag{4.39}$$

avec $\mathbf{C} \in \mathbb{C}^{2M\times P}$ définie en Eq. (4.38) et $J_{\text{NF}}(\mathbf{w})$ défini en Eq. (4.30). En employant les multiplieurs de Lagrange, on obtient alors la solution analytique suivante :

$$\mathbf{w}_{\text{JLC-NF}} = \mathbf{Q}^{-1}(\mathbf{I} - \mathbf{C}(\mathbf{C}^H \mathbf{Q}^{-1} \mathbf{C})^{-1} \mathbf{C}^H \mathbf{Q}^{-1})\mathbf{v}. \tag{4.40}$$

On note qu'en l'absence de contraintes, *i.e.* $|\mathcal{D}| = 0$, le filtre $\mathbf{w}_{\text{JLC-NF}}$ se réduit bien à $\left[\mathbf{w}_{\text{H-NF, L}}^T \mathbf{w}_{\text{H-NF, R}}^T\right]^T$.

4.4 Évaluation

Dans cette section, on compare les performances des H-NF, ITF1-NF, ITF2-NF et JLC-NF -respectivement définis en Eq. (4.7), (4.28), (4.29) et (4.40)- en matière d'atténuation dans la direction frontale et de préservation des indices de localisation sur toute la sphère. Pour ce faire, nous avons besoin de mesures d'ATFs de prothèses auditives d'une grande précision spatiale afin de calculer le diagramme de directivité en sortie des nullformers. Or, les bases de données d'ATFs de prothèses auditives sont rarement d'une grande précision spatiale [Kayser et al., 2009, Denk et al., 2018, Moore et al., 2019a], en particulier en élévation. A notre connaissance, le jeu de données de [Oreinos and Buchholz, 2013] est le seul publiquement disponible avec une précision de 5° aussi bien en azimut qu'en élévation. Celui-ci ne contient les mesures que d'un seul mannequin, pour les deux microphones placés au dessus de chaque pavillon, ainsi qu'un à l'entrée de chaque canal auditif, soit six microphones au total.

4.4.1 Paramètres

Pour le JLC-NF, le nombre de directions pour former le système de contraintes linéaires a été fixé à six ($|\mathcal{D}| = 6$), soit la moitié des degrés de liberté de l'optimisation ($2M = 12$). Afin de sélectionner un ensemble $\mathcal{D}$, on tire sans remise aléatoirement trois directions sur l'hémisphère gauche qui sont ensuite symétrisées sur l'hémisphère droite, de sorte à obtenir six directions. L'azimut, noté θ, et l'élévation, notée ϕ, de ces directions sont tirées selon les lois de probabilité suivantes :

$$\theta \sim \mathcal{N}(90°, 30°), \tag{4.41}$$
$$\phi \sim \mathcal{N}(0°, 30°), \tag{4.42}$$

où $\mathcal{N}(\mu, \sigma)$ est une loi normale de moyenne μ et variance σ. Les azimuts et élévations tirées sont limitées respectivement aux angles $[0°, 180°]$ et $[-60°, 60°]$.

Pour tous les algorithmes, le paramètre de pondération du terme d'annulation, α_0, a été fixé à 1 (*cf.* Eq. (4.4, p.107)) et Eq. (4.30, p.114)). Pour les ITF1-NF et ITF2-NF, le paramètre de pondération de la préservation des ITFs, noté α_{ITF}, est réglé à 1, sauf mention contraire.

4.4.2 Critères d'évaluation

Annulation dans la direction cible Pour mesurer la capacité du nullformer à atténuer dans la direction cible, on utilise le rapport d'amplification entre

la direction cible et toutes les autres directions,[1] en dB :

$$\Delta\mathrm{RSB} = \mathrm{RSB}_{\mathrm{out}} - \mathrm{RSB}_{\mathrm{in}} \tag{4.43}$$

avec

$$\mathrm{RSB}_{\mathrm{in}} = 10\log_{10}\frac{\sum_k |h_{\mathrm{L}}(k)|^2}{\sum_k \mathbf{q}_{\mathrm{L}}^H \hat{\mathbf{\Gamma}}_{\mathbf{n}}(k)\mathbf{q}_{\mathrm{L}}}, \tag{4.44}$$

et

$$\mathrm{RSB}_{\mathrm{out}} = 10\log_{10}\frac{\sum_k |\mathbf{w}_{\mathrm{L}}(k)^H \mathbf{h}(k)|^2}{\sum_k \mathbf{w}_{\mathrm{L}}(k)^H \hat{\mathbf{\Gamma}}_{\mathbf{n}}(k)\mathbf{w}_{\mathrm{L}}(k)}, \tag{4.45}$$

où k est l'indice de fréquence, qu'il est nécessaire de faire apparaître ici.

Erreur d'ITD Afin de mesurer la préservation de l'ITD sur toute la sphère, on moyenne l'erreur absolue d'ITD par direction entre l'entrée et la sortie du nullformer :

$$\Delta\mathrm{ITD} = \sum_d w(d)|\mathrm{ITD}_{\mathrm{out}}(d) - \mathrm{ITD}_{\mathrm{in}}(d)|, \tag{4.46}$$

où $w(d)$ sont les poids suivant la méthode de [Harder, 2015] permettant de prendre en compte le sur-échantillonnage des pôles de la grille de mesure de [Oreinos and Buchholz, 2015]. L'ITD est estimée grâce à la méthode de seuillage adaptatif des HRIRs décrite par [Katz and Noisternig, 2014] et identifiée comme étant la plus perceptivement pertinente [Andreopoulou and Katz, 2017].

Erreur d'ILD La préservation de l'ILD est quantifiée comme la moyenne de l'erreur absolue d'ILD par direction entre l'entrée et la sortie du nullformer :

$$\Delta\mathrm{ILD} = \sum_d w(d)|\mathrm{ILD}_{\mathrm{out}}(d) - \mathrm{ILD}_{\mathrm{in}}(d)|, \tag{4.47}$$

avec

$$\mathrm{ILD}_{\mathrm{in}}(d) = 10\log_{10}\frac{\sum_k |h_{d,\mathrm{L}}(k)|^2}{\sum_k |h_{d,\mathrm{R}}(k)|^2}, \tag{4.48}$$

et

$$\mathrm{ILD}_{\mathrm{out}}(d) = 10\log_{10}\frac{\sum_k |P_{\mathrm{L}}(d,k)|^2}{\sum_k |P_{\mathrm{R}}(d,k)|^2}, \tag{4.49}$$

où k est l'indice de fréquence, qu'il est nécessaire de faire apparaître ici.

[1] Ceci équivaut à une différence de RSB en considérant un bruit spatialement diffus.

4.4.3 Sélection du meilleur sous-espace d'optimisation pour le JLC-NF

Cinq cents ensembles de directions $\mathcal{D}$ ont été générés avec la méthode de tirage aléatoire décrite en sous-section 4.4.1 de sorte à produire autant de filtres de JLC-NF. L'histogramme des directions tirées est illustré en Fig. 4.7(a) et les performances en matière de préservation de l'ITD sont mises en regard avec celles d'atténuation dans la direction cible en Fig. 4.6. On remarque que les résultats sont concentrés autour d'une ΔITD de 25 à 40 μs et d'une ΔRSB de -9 à -7 dB. A titre de comparaison, le H-NF obtient un score de ΔITD de 200 μs et une ΔRSB de -14.2 dB. Néanmoins, la ΔITD des JLC-NF est très variable et peut atteindre des valeurs extrêmes jusqu'à 300 μs, sans pour autant améliorer la ΔRSB.

On propose de sélectionner les directions qui ont permis d'obtenir une ΔITD inférieure à 50 μs. L'histogramme correspondant est visible en Fig. 4.7(b). On peut observer que la moyenne des azimuts est décalée vers l'hémisphère avant (entre $\pm 90°$) par rapport au tirage total. On n'observe pas de pareil tendance pour les élévations.

Se pose alors la question de trouver un ensemble de direction $\mathcal{D}$ de manière objective. Pour cela, on propose d'utiliser l'algorithme de partitionnement des K-moyennes sur les directions qui ont permis d'obtenir une ΔITD inférieure à 50 μs, ce qui est de l'ordre de la plus petite différence perceptible (de l'anglais *Just Noticeable Difference*, jnd) (voir la sous-section 1.2.2 (p.11)). Comme le tirage des sous-ensembles de six directions a été fait symétriquement selon le plan médian, il suffit de chercher trois centroïdes sur l'hémisphère gauche. Les directions obtenues (et symétrisées) sont rassemblées en Tab. 4.1. Le JLC-NF utilisant cet ensemble de direction est calculé et ses performances sont illustrées en Fig. 4.6 et en Fig. 4.8. On remarque que cette méthode permet de trouver un filtre ayant les mêmes propriétés que les filtres dont les sous-ensembles de directions ont permis de le générer.

TABLE 4.1 – Liste des directions obtenues avec la méthode des K-moyennes sur l'ensemble des directions qui ont permis d'obtenir une ΔITD inférieure à 50 μs pour les JLC-NF.

Azimut (°)	-120	-85	-50	50	85	120
Élévation (°)	-10	15	-5	-5	15	-10

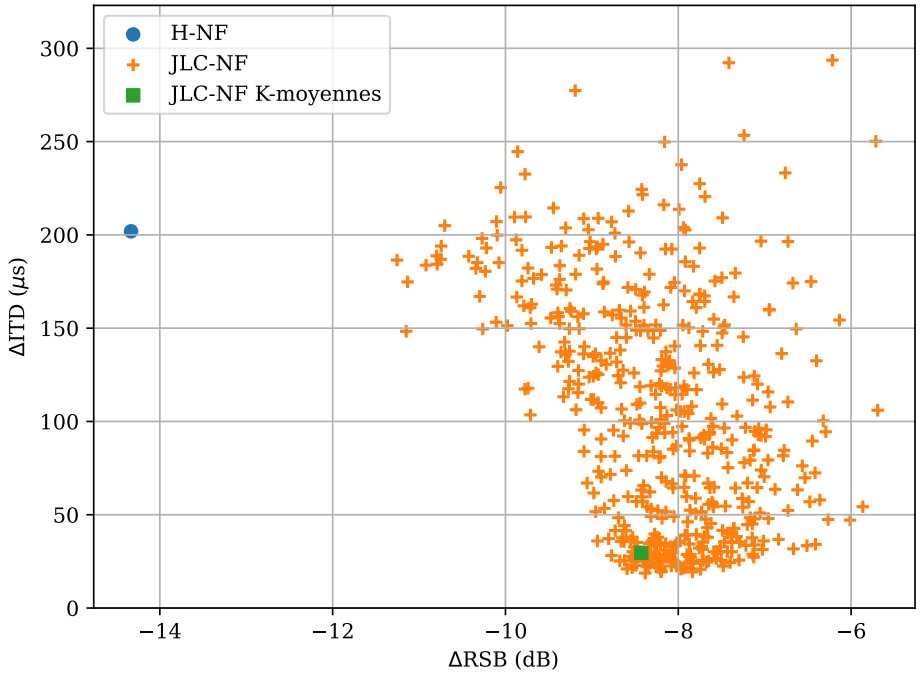


FIGURE 4.6 – ΔITD en fonction du ΔRSB pour le H-NF et 500 JLC-NF obtenus dont le sous-ensemble de directions servant à construire les contraintes linéaires ont été tirées aléatoirement suivant la distribution illustrée en Fig. 4.7a et respectant une symétrie suivant le plan médian.

4.4.4 Résultats

En premier lieu, on observe sans surprise que les nullformers visant à préserver les ITFs introduits dans ce chapitre réduisent les performances d'annulation du signal dans la direction frontale par rapport au H-NF (voir Fig. 4.8 et Fig. 4.6) d'environ 6 dB pour le JLC-NF et même jusqu'à 10 dB pour le ITF2-NF avec $\alpha_{\text{ITF}} = 1$ permettant d'obtenir des performances similaires en matière d'ITD.

En Fig. 4.8 et 4.9 sont illustrées les performances, pour les trois critères choisis, des méthodes H-NF, ITF1-NF et ITF2-NF en faisant varier l'importance donnée à la préservation des ITFs, ainsi que JLC-NF avec le jeu de contraintes obtenues par la méthodes des K-moyennes. On note que l'approche déterministe (ITF2-NF) gagne sur tous les tableaux par rapport à l'approche probabiliste (ITF1-NF). En effet, tout en obtenant des résultats similaires en atténuation de la direction frontale, elle parvient systématiquement à mieux préserver l'ITD et l'ILD sur l'ensemble de la sphère (voir pour plus de détails concernant l'erreur d'ITD et d'ILD en fonction de la direction, les Fig. 4.3(c

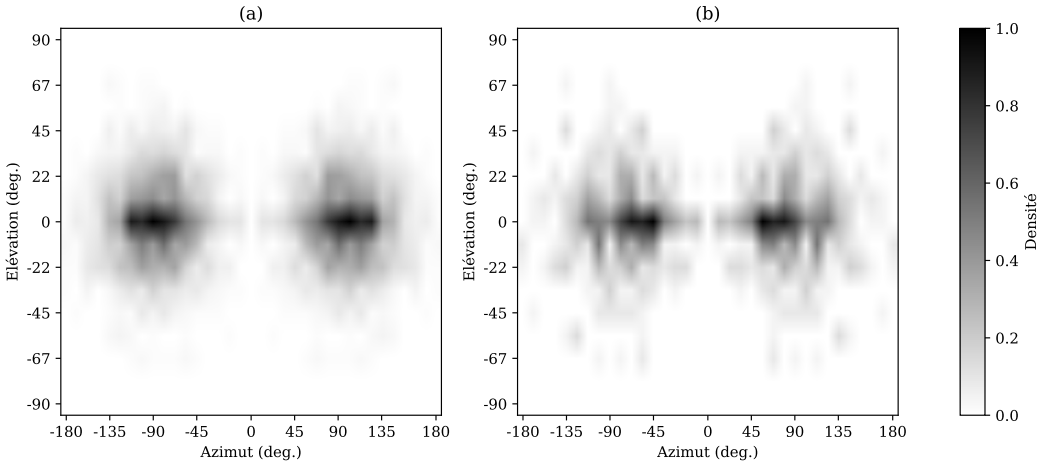


FIGURE 4.7 – L'histogramme de la distribution des directions choisies pour la génération JLC-NF (dont les résultats sont illustrés en Fig. 4.6) (a) et l'histogramme de la distribution du sous-ensemble de directions permettant d'avoir $\Delta\text{ITD} < 50\ \mu s$ (b).

et d) et Fig. 4.4(c et d), respectivement). Cela est attendu dans la mesure où l'approche probabiliste correspond à un cas où le bruit est localisé dans l'espace, filtré par une HRTF correspondant aux HRTFs moyennées sur toute la sphère. En particulier, il faut noter en Fig. 4.8 que la ΔITD et la ΔILD décroissent de manière monotone en fonction de α_{ITF} pour l'ITF2-NF alors que la ΔILD ne suit pas le même comportement pour l'ITF1-NF. On peut voir en détail ce phénomène en Fig. 4.2(d et e) et Fig. 4.4(c et d). Nous remarquons en particulier que l'exigence de faire diminuer la ΔITD, *i.e.* réduire le front d'onde parasite provenant des microphones contralatéraux, implique de réduire le niveau général ainsi que l'ILD, menant à une ΔILD plus grande.

On observe aussi en Fig. 4.9 que la stratégie employée pour obtenir le JLC-NF permet de trouver des solutions optimales au sens de Pareto [2] [Ehrgott, 2005] complémentaires avec le H-NF et le ITF2-NF. En particulier, cette stratégie permet de mieux préserver les indices interauraux à ΔRSB constant par rapport à l'ITF2-NF et *a fortiori* à l'ITF1-NF. Ce résultat est en partie attendu dans la mesure où l'on pouvait s'attendre à mieux préserver les indices interauraux en contraignant la solution à satisfaire une annulation totale de l'erreur d'ITF sur quelques directions, plutôt qu'en adoptant une approximation quadratique de $J_{\text{ITF2}}(\mathbf{w})$. En revanche, on ne pouvait pas prévoir à l'avance à quel point cette stratégie pénaliserait la tâche d'annulation du signal provenant de

[2] Dans une optimisation multi-critère, une solution optimale au sens de Pareto désigne une solution permettant de gagner sur au moins un critère sans perdre sur les autres.

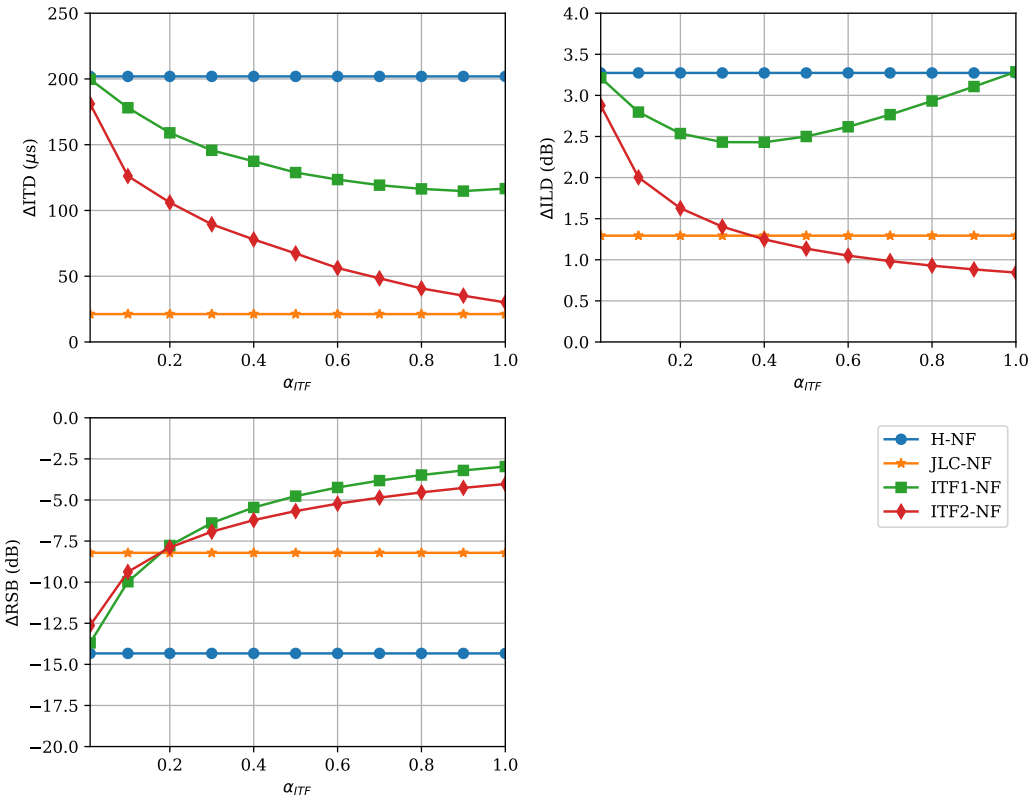
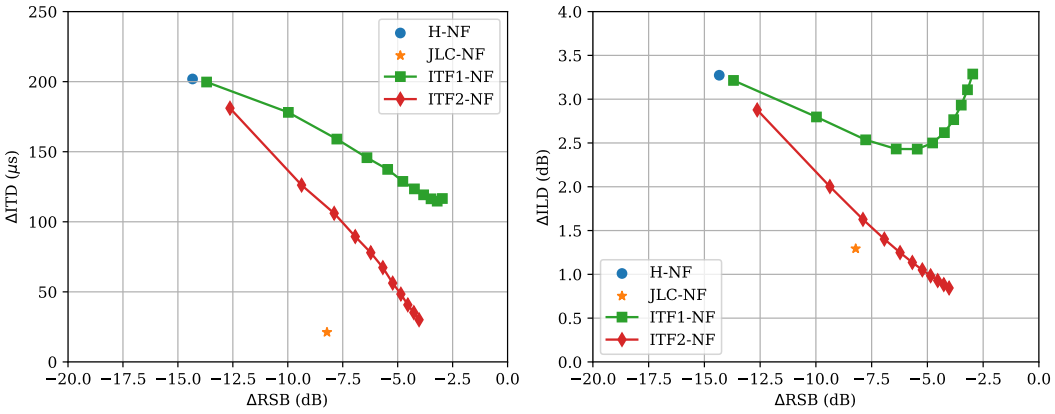


FIGURE 4.8 – ΔITD, ΔILD et ΔRSB en fonction de α_{ITF}, le paramètre de pondération du terme préservant les ITFs dans la détermination du filtre. Le JLC-NF est obtenu avec la méthode des K-moyennes décrite en sous-section 4.4.3. Les trois critères sont à minimiser.

la direction frontale par rapport à l'approximation quadratique de l'ITF2-NF. En Fig. 4.3(b et d), on compare plus en détail la ΔITD pour le JLC-NF et l'ITF2-NF. On observe alors que le JLC-NF permet d'obtenir une ΔITD très proche de la jnd (inférieure à environ 50 μs, voir la sous-section 1.2.2 (p.11) pour plus de détail) sur l'ensemble de la sphère alors que l'ITF2-NF ne parvient pas tout à fait à supprimer le front-d'onde parasite pour les azimuts autour de $\pm 45°$ bien que l'erreur soit très faible sur le reste de la sphère.

Toutefois, il faut noter que l'approximation quadratique dans $\hat{J}_{\text{ITF2}}(\mathbf{w})$ permet de réaliser explicitement et de manière déterministe le compromis entre annulation et préservation des indices interauraux contrairement au JLC-NF. En effet, on ne peut pas prévoir de manière certaine si un sous-ensemble de directions va nous donner une meilleure solution du JLC-NF qu'un autre, voir le nuage de point obtenu pour différent tirages aléatoires de JLC-NF en Fig. 4.6



(a) (b)

FIGURE 4.9 – ΔITD (a) et ΔILD (b) en fonction du ΔRSB. Le paramètre de pondération du terme préservant les ITFs dans la détermination des filtres des ITF1-NF et ITF2-NF, noté α_{ITF}, varie de 0.01 à 1 et le JLC-NF est obtenu avec la méthode des K-moyennes décrite en sous-section 4.4.3. Les trois critères sont à minimiser.

comparé aux ITF2-NF obtenus en faisant varier α_{ITF} en Fig. 4.9a.

4.5 Conclusion

Dans ce chapitre, nous avons mis en évidence les limites du H-NF employé dans le chapitre précédent en matière de préservation des indices de localisation interauraux, en particulier dans l'hémisphère frontale.

Nous avons alors proposé d'intégrer la préservation de l'ITF dans la détermination des filtres de nullforming. Cette fonction est non-convexe et déjà utilisée dans la littérature pour développer des algorithmes de beamforming préservant les indices interauraux. Nous nous en sommes inspiré pour développer l'ITF1-NF. Mais celui-ci montre de sévères limites dans la préservation de l'ILD. Nous nous sommes alors servi d'une reformulation du problème d'optimisation de nullforming dans un cadre déterministe de sorte à pouvoir considérer l'ITF de sources provenant de n'importe quelles directions. Nous avons alors proposé l'ITF2-NF, basé sur la même approximation quadratique que l'ITF1-NF, ainsi que le JLC-NF, basée sur une nouvelle stratégie consistant à restreindre le sous-espace de recherche de la solution de filtre de nullforming.

Les trois algorithmes introduits dans ce chapitre ont alors été comparés au H-NF en matière d'annulation dans la direction frontale et de préservation

des indices interauraux sur l'ensemble de la sphère. Pour cela, nous avons utilisé des métriques objectives et des mesures d'ATFs de prothèses auditives à haute résolution spatiale [Oreinos and Buchholz, 2015]. Pour tous les nullformers préservant l'ITF, la préservation des indices de localisation interauraux s'accompagne d'une réduction de l'atténuation dans la direction frontale. Le compromis entre ces deux objectifs peut être modulé suivant le RSB d'entrée par exemple. En effet, si le RSB est très grand il est préférable de privilégier la réduction de la cible dans l'estimateur du bruit par rapport aux indices de localisation et inversement. Nous avons aussi montré que le cadre déterministe permettait de dériver des algorithmes plus performants sur tous les aspects par rapport au cadre probabiliste usuellement employé dans la littérature du beamforming.

Limites Nous avons développé les nullformers dans la suite directe du chapitre 3. Dans ce dernier, nous avons limité le modèle de scène sonore à une source de parole cible et du bruit ambiant. Le travail de ce chapitre a donc hérité de cette hypothèse. Les algorithmes proposés ici permettent de préserver les indices de localisation interauraux de sources de parole possiblement d'intérêt supplémentaires hors-axe de visée. Néanmoins, celles-ci ne sont pas considérées par l'algorithme comme d'intérêt et donc ni compressées, ni amplifiées comme telles. Un tel scénario multi-locuteurs sera l'objet du chapitre suivant.

Aussi, la base de données d'ATFs que nous avons utilisée pour l'évaluation des algorithmes ne comprend qu'un seul mannequin. Il est donc difficile de généraliser les résultats obtenus, notamment la méthode de partitionnement employée dans le choix du sous-ensemble de directions nécessaire au JLC-NF. Aussi, les indices spectraux sont en parti distordus. Or, il est bien documenté que nous sommes capables de réapprendre des indices de localisation, notamment spectraux [Mendonça, 2014] (voir aussi la sous-section 1.2.4 (p.16)). Cependant, l'évaluation en condition réelle du réapprentissage des indices spectraux avec les nullformers est très compliquée à mener et est au-delà du champ d'étude considéré dans ce travail de thèse.

Perspectives La formulation déterministe des filtres de nullforming en section 4.2 (p.104) a fait apparaître explicitement l'HRTF cible vers laquelle on veut faire tendre l'HRTF en sortie du nullformer. Ceci ouvre la possibilité d'adapter l'HRTF cible en fonction de la perte auditive du patient appareillé. En effet, [Dieudonné and Francart, 2018] ont montré l'intérêt d'augmenter l'ILD en basse fréquence pour la localisation auditive et la compréhension de la parole.

L'élaboration des algorithmes et leur évaluation a nécessité des mesures d'ATFs de prothèses auditives d'une grande précision spatiale. Or, la plupart des bases de données existantes ne fournissent pas une telle précision et la seule à notre disposition est composée d'un seul sujet. Cela démontre l'intérêt d'avoir les ATFs avec une meilleure précision spatiale pour plus de sujets afin de pouvoir mieux généraliser nos résultats.

Chapitre 5

Amélioration du débruitage et réduction de la complexité algorithmique grâce à la parcimonie de la parole

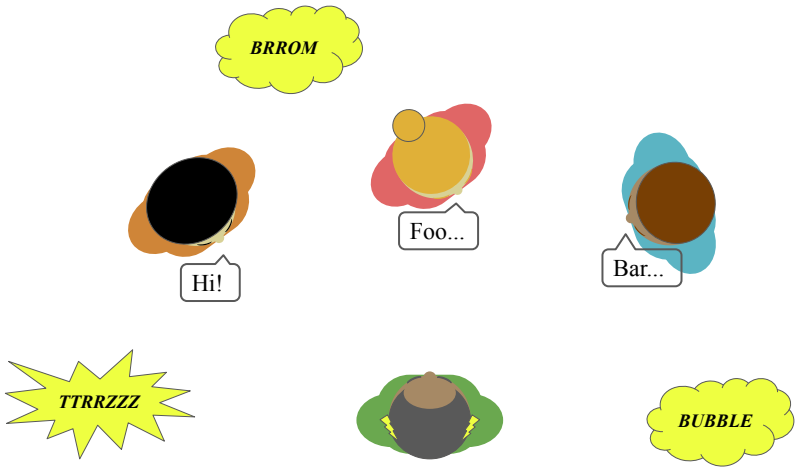


FIGURE 5.1 – Illustration de la scène sonore considérée dans ce chapitre, vue du dessus. Celle-ci est composée d'un auditeur (au centre) portant deux prothèses auditives (en jaune), de trois interlocuteurs et d'un bruit ambiant symbolisé par les bulles jaunes.

5.1 Introduction

Dans les deux chapitres précédents, nous avons restreint la composition de la scène sonore à une source de parole cible et un bruit ambiant. Il est pourtant courant que plusieurs sources d'intérêt localisées dans l'espace se trouvent actives simultanément. En effet, les situations d'interactions sociales impliquent souvent plusieurs personnes susceptibles de prendre la parole comme illustré en Fig. 5.1 (commensalité familiale, amicale ou professionnelle, réunion de travail, etc.). La solution générale proposée dans le chapitre 3 et améliorée dans le chapitre 4, permet de préserver les indices de localisation de l'ensemble de la scène sonore mais cherche à atténuer tout ce qui est hors de la direction cible. Or, dans un scénario considérant plusieurs sources sonores localisées dans l'espace (parole, signal d'alerte, etc.), il n'est pas simple de savoir *a priori* sur laquelle l'auditeur•rice va porter son attention, d'autant plus qu'elle peut basculer alternativement vers l'une ou l'autre au cours du temps. Pour cela, il est nécessaire de généraliser les algorithmes de réduction de bruit de sorte à préserver plusieurs sources à la fois. Les beamformers LCMV [Suzuki et al., 1999] et MMWF [Markovich-Golan et al., 2012b] sont respectivement les généralisations à plusieurs cibles des beamformers MVDR et MWF. En pratique, le beamformer LCMV semble être privilégié car il n'introduit pas de distorsion de la parole, contrairement au MMWF [Hadad et al., 2012, Hadad et al., 2016].

Notons que depuis quelques années, des travaux ont cherché à déterminer sur quelle source l'auditeur porte son attention, pour ainsi la privilégier dans la sortie du beamformer. Par exemple, on peut supposer que la direction du regard nous informe sur la zone de l'espace d'intérêt pour l'auditeur. Il a donc été proposé d'estimer la direction du regard grâce aux signaux nerveux captés au niveau du canal auditif et de diriger un beamformer vers celle-ci [Favre-Félix et al., 2018]. Il a aussi été proposé de mesurer directement l'activité cérébrale et de calculer la corrélation des signaux recueillis avec les sources en présence dans la scène sonore, préalablement séparées [Mesgarani and Chang, 2012]. On suppose alors que la source pour laquelle la corrélation est la plus forte est celle sur laquelle l'auditeur porte son attention. Celle-ci est alors privilégiée dans le processus d'optimisation permettant d'obtenir le filtre de beamforming [Van Eyndhoven et al., 2017, Aroudi and Doclo, 2019].

Un problème sur-contraint Néanmoins, les contraintes en matière d'efficacité calculatoire, de faible latence et de nombre de microphones, propres aux prothèses auditives, soulèvent des problèmes particuliers. Rappelons que l'objectif du beamformer LCMV est de minimiser la puissance du bruit à sa sortie, sous contrainte de préserver les sources d'intérêts localisées dans l'espace. Par conséquent, premièrement, les performances de réduction de bruit s'amenuisent avec le nombre de sources à préserver. En effet, l'ajout d'une contrainte correspond à retirer un degré de liberté au problème d'optimisation, réduisant ainsi la taille du sous-espace dans lequel la minimisation de la puissance du bruit est réalisée. Ce mécanisme est illustré en Fig. 5.2. Deuxièmement, l'ajout d'une contrainte dans le problème d'optimisation augmente la taille d'une matrice qu'il est nécessaire d'inverser pour obtenir le filtre. Ainsi, considérer plus de sources d'intérêt augmente le coût calculatoire du filtre.

Coût calculatoire Quelques travaux se sont attaqués au problème du coût calculatoire du beamformer LCMV. Par exemple, [Guo et al., 2014] s'appuient sur l'hypothèse d'une invariance temporelle de la localisation des sources de parole de telle sorte que le filtre de beamforming peut être mis à jour à chaque instant à l'aide d'une méthode itérative peu consommatrice en temps de calcul. Une autre étude [Markovich-Golan et al., 2012a] a proposé de supposer qu'uniquement un sous-ensemble de sources de parole change entre deux trames temporelles. Ainsi, il est possible de mettre à jour à moindre coût le filtre du beamformer LCMV sans le recalculer entièrement. Cependant, ces hypothèses ne sont pas appropriées dans le contexte des prothèses auditives. En effet, même si les sources de paroles ne bougent pas, l'auditeur peut être amené à bouger la tête rapidement et fréquemment. De plus, ces méthodes deviennent efficaces pour des réseaux de microphones bien plus grands que ceux utilisés dans les

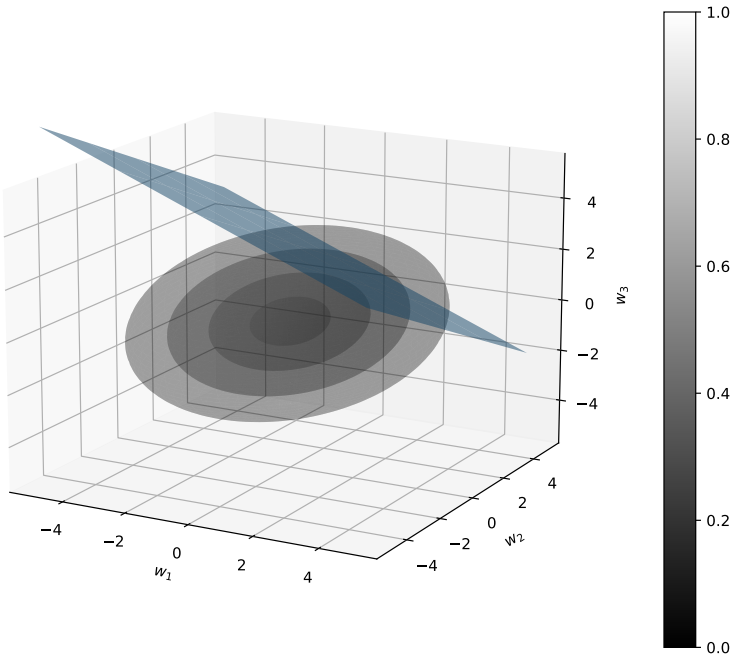


FIGURE 5.2 – Illustration d'un problème de minimisation d'une fonction de coût quadratique 3D et d'une contrainte linéaire. Les ellipsoïdes représentent les surfaces d'iso-valeurs de la fonction de coût dont les nuances de gris représentent sa valeur. La contrainte linéaire forme un sous-espace plan dans lequel se trouve la solution. L'ajout d'une contrainte supplémentaire engendrerait une droite plutôt qu'un plan.

prothèses auditives et ne traitent pas du problème de la réduction des performances de débruitage du beamformer LCMV lorsque le nombre de sources de parole devient proche du nombre de microphones.

Pour résoudre ce double problème, il est pratique d'exploiter la propriété de parcimonie des sources de parole dans le domaine de la TFCT [Rickard and Yilmaz, 2002]. Cette hypothèse de parcimonie est illustrée en Fig. 5.3, on observe sur le diagramme de dispersion des puissances à la fréquence 250 Hz que les deux sources de parole ne sont jamais simultanément de forte intensité (zone en haut à droite du graphe). Supposer qu'une seule source de parole est active à un point T-F donné permet de réduire le beamformer LCMV à un beamformer MVDR. Cette approche montre de bonnes performances de

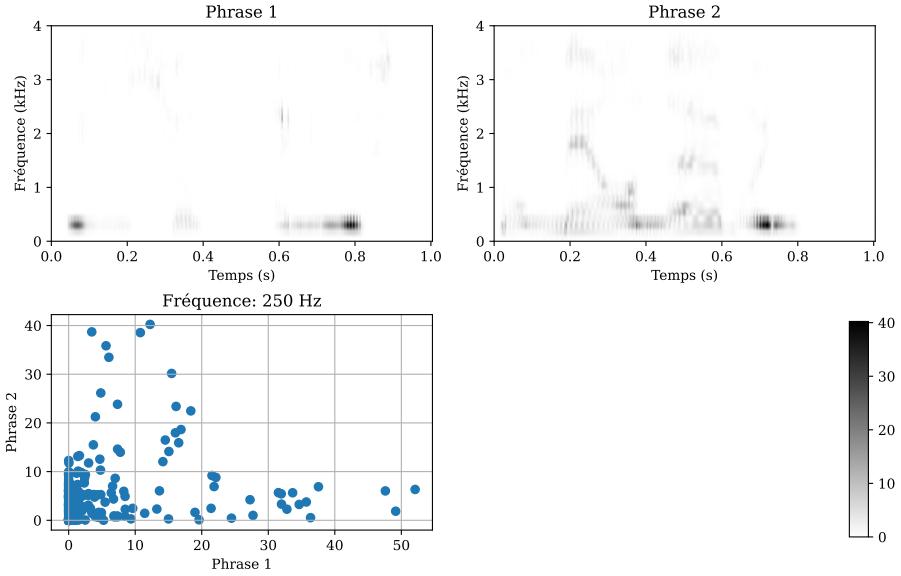


FIGURE 5.3 – Spectrogrammes de puissance de deux phrases (haut) prononcées par un locuteur masculin et le diagramme de dispersion entre les puissances des deux phrases au cours du temps pour la fréquence 250 Hz (bas).

réduction de bruit dans un scénario composé de deux locuteurs [Braun et al., 2015, Zohourian et al., 2018] tout en ayant une complexité algorithmique faible. Cependant, pour un nombre de sources supérieur à deux, cette hypothèse devient fausse pour une proportion non-négligeable de points T-F. Par exemple, dans un scénario composé de trois locuteurs, celle-ci est fausse pour environ 20 % des points [Jia et al., 2018]. De plus, les signaux de parole se recouvrent principalement en basses fréquences (voir Fig. 5.4), ce qui correspond à la plage où se concentre aussi leur énergie. Par conséquent, bien que l'hypothèse est valide sur une majorité du plan temps-fréquence, là où elle ne l'est pas correspond à la zone la plus critique du spectre de la parole. Dans l'exemple en Fig. 5.4, les points T-F concernés par le recouvrement sont en moyenne 20 dB plus énergique que le niveau moyen de la parole. Cette hypothèse peut être assouplie de telle sorte qu'on suppose que les sources de parole ne se recouvrent que rarement dans le domaine de la TFCT. A notre connaissance, seulement une étude a considéré une telle hypothèse et ce dans le cadre de la séparation de sources (sans bruit ambiant) avec un microphone Soundfield [1] plutôt que des prothèses auditives [Jia et al., 2018]. Ces différences de contexte applicatif

[1] Réseau de microphones composé de quatre capsules à directivité cardioïde disposées sur les faces d'un tétraèdre régulier. Ce réseau permet d'obtenir une représentation du champ

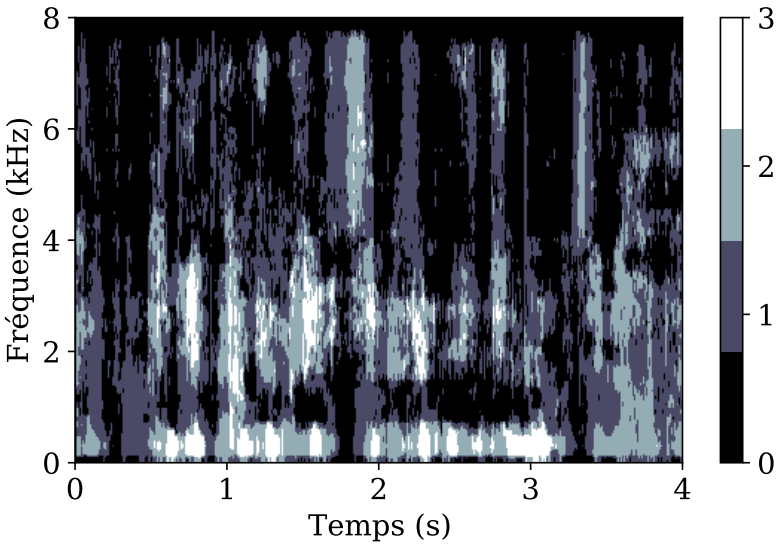


FIGURE 5.4 – Nombre de sources de paroles actives dans le domaine de la TFCT pour un mélange de trois phrases.

entraînent différentes implications qui ne se généralisent pas à un réseau quelconque, comme dans les prothèses auditives. Par ailleurs, le coût calculatoire n'était pas la motivation des auteurs et n'a donc pas été évalué.

Dans ce chapitre, nous proposons une méthode de beamforming basée sur l'hypothèse de non-recouvrement assouplie. De manière intéressante, celle-ci peut être considérée comme un cas particulier du MMWF [Markovich-Golan et al., 2012b]. Nous évaluons ses performances de réduction de bruit et de préservation des sources de parole selon des critères objectifs ainsi que sa complexité algorithmique. Ses performances sont comparées à celle du beamformer MVDR dirigé vers la source de parole la plus puissante à chaque point T-F, résultant de l'hypothèse de non-recouvrement des sources, et du beamformer LCMV pour lequel toutes les sources sont supposées être actives en tout point T-F.

5.2 Modèles des signaux

Nous considérons une scène sonore composée de Q sources de paroles, notées $s_q(t)$. La transformation entre la $g^{\text{ème}}$ source et le $m^{\text{ème}}$ microphone est modélisée par un filtre linéaire dont la réponse impulsionnelle est notée $h_{m,q}(t)$.

de pression acoustique dans le domaine des harmoniques sphériques (ordre 1).

On suppose aussi que le bruit est différent à chaque microphone, noté $n_m(t)$. Alors, le signal reçu au $m^{\text{ème}}$ microphone peut être écrit comme suit :

$$x_m(t) = \sum_{q=1}^{Q} (h_{m,q} \star s_q)(t) + n_m(t), \tag{5.1}$$

où $\star$ est le produit de convolution. Le modèle de mélange est habituellement exprimé dans le domaine de la TFCT. Lorsque la durée de $h_{m,q}(t)$ est inférieure à celle de la fenêtre d'analyse de la TFCT, le produit de convolution dans le domaine temporel peut être approximé par une simple multiplication dans le domaine de la TFCT [Avargel and Cohen, 2007] :

$$x_m(k,\ell) = \sum_{q=1}^{Q} h_{m,q}(k,\ell)\ s_q(k,\ell) + n_m(k,\ell), \tag{5.2}$$

où k et ℓ sont respectivement les indices de fréquence et de temps. Cette expression peut être réécrite sous forme matricielle en empilant les variables le long des axes des microphones et des sources de parole :

$$\mathcal{M}_1 : \quad \mathbf{x}(k,\ell) = \mathbf{H}(k,\ell)\ \mathbf{s}(k,\ell) + \mathbf{n}(k,\ell), \tag{5.3}$$

avec $\mathbf{H}(k,\ell) \in \mathbb{C}^{M\times Q}$ la matrice contenant les ATFs, $\mathbf{n}(k,\ell) \in \mathbb{C}^{M}$ et $\mathbf{s}(k,\ell) \in \mathbb{C}^{Q}$.

En supposant qu'une unique source est active à chaque point T-F [Rickard and Yilmaz, 2002], l'expression précédente peut être écrite comme suit :

$$\mathcal{M}_2 : \quad \mathbf{x}(k,\ell) = \mathbf{h}_{q(k,\ell)}(k)\ s_{q(k,\ell)}(k,\ell) + \mathbf{n}(k,\ell), \tag{5.4}$$

où $q(k,\ell)$ est l'indice de la source de parole active au point T-F k, ℓ.

L'hypothèse alternative proposée dans cette étude consiste à considérer toutes les configurations intermédiaires de $\kappa(k,\ell) = 0$ (aucune source présente) jusqu'à $\kappa(k,\ell) = Q$ (toutes les sources actives) au point T-F k, ℓ :

$$\mathcal{M}_3 : \quad \mathbf{x}(k,\ell) = \tilde{\mathbf{H}}(k,\ell)\tilde{\mathbf{s}}(k,\ell) + \mathbf{n}(k,\ell), \tag{5.5}$$

où $\tilde{\mathbf{H}}(k,\ell) \in \mathbb{C}^{M\times\kappa(k,\ell)}$ et $\tilde{\mathbf{s}}(k,\ell) \in \mathbb{C}^{\kappa(k,\ell)}$. Dans la suite, nous ferons référence aux modèles décrits en Eq. (5.3), (5.4) et (5.5) par $\mathcal{M}_1$, $\mathcal{M}_2$ et $\mathcal{M}_3$, respectivement.

De plus, $s_q(k,\ell)$ et $\mathbf{n}(k,\ell)$ sont modélisés comme des variables aléatoires suivant une distribution gaussienne complexe centrée circulaire de variance $\phi_{s_q}(k,\ell)$ et de matrice de covariance $\mathbf{\Phi}_{\mathbf{n}}(k,\ell)$, respectivement. Le bruit est supposé être spatialement cylindriquement diffus de sorte que $\mathbf{\Phi}_{\mathbf{n}}(k,\ell)$ peut

être factorisée en une matrice normalisée invariante dans le temps, notée $\mathbf{\Gamma_n}(k)$, et un facteur d'échelle, noté $\phi_n(k,\ell)$ [Gay and Benesty, 2000].

En pratique, la matrice $\mathbf{\Gamma_n}(k)$ est estimée en moyennant toutes les ATFs du plan horizontal [Lotter and Vary, 2006]. Enfin, les ATFs sont supposées connues pour toutes les directions et similairement à [Corey and Singer, 2017], nous faisons l'hypothèse que nous avons la connaissance parfaite des DOAs des sources présentes dans la scène sonore.

5.3 Méthodes de réduction de bruit

5.3.1 Détermination des algorithmes

Le signal attendu en sortie d'un beamformer idéal ne contient pas la composante de bruit et est uniquement composé de la somme des Q sources de parole filtrées par leur fonction de transfert correspondante, notée $g_q(k,\ell)$, contenant, par exemple, les indices de localisation désirés [Markovich-Golan et al., 2012b] :

$$y(k,\ell) = \mathbf{g}^H(k,\ell)\mathbf{s}(k,\ell), \tag{5.6}$$

où $\mathbf{g}(k,\ell) = \left[g_1^*(k,\ell), ..., g_Q^*(k,\ell)\right]^T \in \mathbb{C}^Q$. Ils peuvent varier au cours du temps si les sources se déplacent par rapport à l'auditeur·rice par exemple.

La sortie du beamformer, notée $\hat{y}(k,\ell)$, est construite comme une combinaison linéaire des signaux des microphones dans le domaine de la TFCT avec les poids $\mathbf{w}(k,\ell) \in \mathbb{C}^M$:

$$\hat{y}(k,\ell) = \mathbf{w}^H(k,\ell)\mathbf{x}(k,\ell). \tag{5.7}$$

Déterminer $\mathbf{w}_{\mathcal{M}_1}(k,\ell)$, les poids du beamformer pour le modèle $\mathcal{M}_1$, consiste à minimiser la variance de la composante de bruit dans la sortie du beamformer sous contrainte de préserver les réponses en fréquence des sources de parole ciblées :

$$\mathbf{w}_{\mathcal{M}_1}(k,\ell) = \underset{\mathbf{w}}{\operatorname{argmin}} \left\{\phi_n(k,\ell)\mathbf{w}^H\mathbf{\Gamma_n}(k)\mathbf{w}\right\} \tag{5.8}$$

$$\text{s.t.} \quad \mathbf{w}^H\mathbf{H}(k,\ell) = \mathbf{g}^H(k,\ell). \tag{5.9}$$

En utilisant les multiplieurs de Lagrange, on obtient :

$$\mathbf{w}_{\mathcal{M}_1}(k,\ell) = \mathbf{\Gamma_n}^{-1}(k)\mathbf{H}(k,\ell)\left(\mathbf{H}^H(k,\ell)\mathbf{\Gamma_n}^{-1}(k)\mathbf{H}(k,\ell)\right)^{-1}\mathbf{g}(k,\ell). \tag{5.10}$$

Cette solution est appelée le beamformer LCMV dans la littérature [Suzuki et al., 1999].

Le modèle $\mathcal{M}_2$ est un cas particulier de $\mathcal{M}_1$ pour lequel uniquement une source, notée $s_{q(k,\ell)}(k,\ell)$, est présente au point T-F k, ℓ. Nous pouvons alors écrire la solution de la manière suivante :

$$\mathbf{w}_{\mathcal{M}_2}(k,\ell) = \frac{\mathbf{\Gamma}_{\mathbf{n}}^{-1}(k)\mathbf{h}_{q(k,\ell)}(k)}{\mathbf{h}_{q(k,\ell)}^{H}(k)\mathbf{\Gamma}_{\mathbf{n}}^{-1}(k)\mathbf{h}_{q(k,\ell)}(k)}. \tag{5.11}$$

Cette solution correspond au beamformer MVDR et plus précisément au beamformer maximisant l'indice de directivité [Stadler and Rabinowitz, 1993] car nous considérons ici la matrice de cohérence d'un bruit spatialement diffus.

Enfin, le modèle $\mathcal{M}_3$ proposé mène, comme le modèle $\mathcal{M}_1$, au beamformer LCMV en remplaçant $\mathbf{H}$ par $\tilde{\mathbf{H}}$ et $\mathbf{g}$ par $\tilde{\mathbf{g}}$:

$$\mathbf{w}_{\mathcal{M}_3}(k,\ell) = \mathbf{\Gamma}_{\mathbf{n}}^{-1}(k)\tilde{\mathbf{H}}(k,\ell)\left(\tilde{\mathbf{H}}^{H}(k,\ell)\mathbf{\Gamma}_{\mathbf{n}}^{-1}(k)\tilde{\mathbf{H}}(k,\ell)\right)^{-1}\tilde{\mathbf{g}}(k,\ell). \tag{5.12}$$

Rappelons que les dimensions de $\tilde{\mathbf{H}}(k,\ell) \in \mathbb{C}^{M\times\kappa(k,\ell)}$ et $\tilde{\mathbf{g}}(k,\ell) \in \mathbb{C}^{\kappa(k,\ell)}$ varient en fonction de la fréquence et du temps même si les sources ne se déplacent pas. En faisant l'hypothèse que les signaux de parole peuvent se recouvrir dans le domaine de la TFCT mais que la plupart du temps ils ne le font pas, le nombre moyen de contraintes dans le problème d'optimisation devrait être plus faible que pour $\mathbf{w}_{\mathcal{M}_1}$, laissant plus de degré de liberté alloué à la réduction du bruit. En outre, contrairement à $\mathcal{M}_2$ pour lequel il est supposé qu'une seule source de parole est toujours présente, $\mathcal{M}_3$ considère le cas où aucune source n'est active, menant à $\mathbf{w}_3(k,\ell) = 0$.

5.3.2 Analyse de la solution

Dans ce paragraphe, nous analysons $\mathbf{w}_{\mathcal{M}_1}, \mathbf{w}_{\mathcal{M}_2}$ et $\mathbf{w}_{\mathcal{M}_3}$ comme des cas particuliers du filtre de Wiener multilocuteurs multicanal (MMWF) visant à minimiser la puissance du bruit à la sortie du beamformer ainsi que la distorsion entre les sources vocales idéales et leurs estimations [Markovich-Golan et al., 2012b]. En omettant les indices k et ℓ par souci de concision, nous pouvons écrire la détermination du filtre, noté $\mathbf{w}_{\text{MMWF}}$ comme le problème d'optimisation suivant :

$$\mathbf{w}_{\text{MMWF}} = \underset{\mathbf{w}}{\text{argmin}} \left\{ \mathbf{w}^{H}\mathbf{\Phi}_{\mathbf{n}}\mathbf{w} + \sum_{q=1}^{Q} \lambda_q \mathbb{E}\left[|g_q s_q - \mathbf{w}^{H}\mathbf{h}_q s_q|^2\right] \right\}, \tag{5.13}$$

où λ_q, $q \in \{1,...,Q\}$ contrôlent la quantité de distorsion de la parole. Ce problème admet la solution analytique suivante :

$$\mathbf{w}_{\text{MMWF}} = \mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{H}\left(\phi_n\mathbf{\Lambda}^{-1}\mathbf{\Phi}_{\mathbf{s}}^{-1} + \mathbf{H}^{H}\mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{H}\right)^{-1}\mathbf{g}, \tag{5.14}$$

Opération	nb. de produits	Opération	nb. de produits
$\mathbf{N} = \mathbf{\Gamma}_\mathbf{n}^{-1}\mathbf{H}$	M^2Q	$\mathbf{z} = \mathbf{D}^{-1}\mathbf{g}$	$\frac{Q^3}{6} + Q^2$
$\mathbf{D} = \mathbf{H}^H\mathbf{N}$	MQ^2	$\mathbf{w}_{\mathcal{M}_1} = \mathbf{N}\mathbf{z}$	MQ

TABLE 5.1 – Détail du nombre de produits requis pour calculer le filtre du beamformer LCMV. Résoudre $\mathbf{D}\mathbf{z} = \mathbf{g}$ nécessite $Q^3/6 + Q^2$ produits, en exploitant le fait que $\mathbf{D}$ est définie-positive [Press et al., 2007].

où $\mathbf{\Lambda} = \text{diag}\{\lambda_1, ..., \lambda_Q\}$ et $\mathbf{\Phi_s} = \text{diag}\{\phi_{s_1}, ..., \phi_{s_Q}\}$ est la matrice de covariance des sources de la parole. La manière optimale de définir $\mathbf{\Lambda}$ n'est pas simple. Plusieurs stratégies ont été proposées, par exemple fixer $\lambda_q = \text{sig}\left(\phi_{s_q}/\phi_n\right)$ avec $\text{sig}\left(.\right)$ une fonction sigmoïde [Thiergart and Habets, 2014], ou en définissant λ_q comme les probabilités de présence de la parole *a posteriori* [Bagheri and Giacobello, 2019], ou avec $\lambda_q \rightarrow \infty \; \forall q$, réduisant le MMWF au beamformer LCMV. Le beamformer proposé $\mathbf{w}_{\mathcal{M}_3}$ peut être interprété comme fixant $\lambda_q \rightarrow \infty$ si la source q est active, et $\lambda_q = 0$ sinon.

5.3.3 Analyse de la complexité algorithmique

Dans cette sous-section, nous analysons la complexité de calcul des beamformers présentés précédemment, définie comme le nombre de produits nécessaires pour calculer le filtre correspondant. Pour ce faire, nous supposons qu'il n'est pas possible de précalculer et de stocker les filtres. Par exemple, le nombre de $\mathbf{H}(k, \ell)$ possibles est égal au coefficient binomial $\binom{D}{Q}$ où D est le nombre de directions connues. Pour un plan horizontal échantillonné avec un pas de 5°, $D = 72$, conduisant à 59640 $\mathbf{H}$ possibles par fréquence avec $Q = 3$. Cela rend prohibitif le calcul en amont et le stockage de tous les filtres possibles $\mathbf{w}_{\mathcal{M}_1}$. En Tab. 5.1, nous fournissons les détails pour déterminer le nombre de produits requis pour calculer le beamformer LCMV $\mathbf{w}_{\mathcal{M}_1}$. Les résultats pour les trois beamformers sont présentés en Tab. 5.2.

Il convient de noter que le nombre de produits par trame temporelle requis par le calcul du beamformer proposé $\mathbf{w}_{\mathcal{M}_3}$ n'est plus constant, car il dépend du nombre de sources actives à chaque fréquence. Sa moyenne dépend des proportions de points T-F $\alpha_\kappa \in [0; 1]$ pour lesquels $\kappa \in \{0, ..., Q\}$ sources sont actives ($\sum_\kappa \alpha_\kappa = 1$).

Enfin, nous devons mentionner que le calcul du filtre LCMV dans le cas de deux locuteurs ($Q = 2$) peut être accéléré avec l'implémentation efficace proposée dans [Hadad et al., 2016]. La prise en compte de cette amélioration dans notre scénario plus général à trois locuteurs est laissée pour un travail futur.

Filtre	Nombre de produits moyen
$\mathbf{w}_{\mathcal{M}_1}$	$(MQ+Q^2)(M+1)+\frac{Q^3}{6}$
$\mathbf{w}_{\mathcal{M}_2}$	M^2+M
$\mathbf{w}_{\mathcal{M}_3}$	$\alpha_1(M^2+M)+\sum_{\kappa=2}^{Q}\alpha_\kappa\left((M\kappa+\kappa^2)(M+1)+\frac{\kappa^3}{6}\right)$

TABLE 5.2 – Nombre de produits moyen nécessaire au calcul des filtres de beamforming. α_κ désigne la proportion de point T-F pour lesquels κ sources de parole sont actives.

5.3.4 Détection de la parole

Afin de détecter quelle source vocale est présente ou non à chaque point T-F, nous proposons d'utiliser un détecteur d'activité vocale (de l'anglais *Voice Activity Detector*, VAD) basé sur le seuillage du RSB à la sortie d'un beamformer MVDR dirigé vers la source $q^{\text{ème}}$, noté $\xi_{\text{MVDR},q}(k,\ell)$ [Thiemann et al., 2016] :

$$\text{VAD}_q(k,\ell)=\begin{cases}1 & \text{si } \xi_{\text{MVDR},q}(k,\ell)>10^{\frac{\tau}{10}}\\ 0 & \text{sinon,}\end{cases} \tag{5.15}$$

où $\tau\in\mathbb{R}$ est le seuil de détection exprimé en dB. L'estimation du RSB à la sortie du beamformer MVDR visant la $q^{\text{ème}}$ source, notée $\hat{\xi}_{\text{MVDR},q}(k,\ell)$, est exprimée comme suit :

$$\hat{\xi}_{\text{MVDR},q}(k,\ell)=\frac{\hat{\phi}_{s_q}(k,\ell)}{\hat{\phi}_{n,q}(k,\ell)}\mathbf{h}_q^H(k)\mathbf{\Gamma}_{\mathbf{n}}^{-1}(k)\mathbf{h}_q(k), \tag{5.16}$$

où $\hat{\phi}_{s_q}(k,\ell)$ et $\hat{\phi}_{n,q}(k,\ell)$ sont respectivement les estimations des variances de la $q^{\text{ème}}$ source et du bruit en supposant que seule la $q^{\text{ème}}$ source est active [Thiemann et al., 2016] :

$$\hat{\phi}_{s_q}(k,\ell)=\mathbf{w}_{\text{MVDR},q}^H(k)(\mathbf{\Phi}_{\mathbf{x}}(k,\ell)-\hat{\phi}_{n,q}(k,\ell)\mathbf{\Gamma}_{\mathbf{n}}(k))\mathbf{w}_{\text{MVDR},q}(k) \tag{5.17}$$

$$\hat{\phi}_{n,q}(k,\ell)=\frac{1}{M-1}\text{Tr}\left\{(\mathbf{I}-\mathbf{h}_q(k)\mathbf{w}_{\text{MVDR},q}^H(k))\mathbf{\Phi}_{\mathbf{x}}(k,\ell)\mathbf{\Gamma}_{\mathbf{n}}^{-1}(k)\right\} \tag{5.18}$$

où $\mathbf{\Phi}_{\mathbf{x}}(k,\ell)$ est la matrice de covariance des signaux des microphones, estimée grâce à un filtre récursif, et :

$$\mathbf{w}_{\text{MVDR},q}(k)=\frac{\mathbf{\Gamma}_{\mathbf{n}}^{-1}(k)\mathbf{h}_q(k)}{\mathbf{h}_q^H(k)\mathbf{\Gamma}_{\mathbf{n}}^{-1}(k)\mathbf{h}_q(k)}. \tag{5.19}$$

5.4 Évaluation objective

Dans cette section, nous évaluons les trois algorithmes de débruitage en matière de réduction du bruit et de complexité algorithmique. Dans ce qui suit, les beamformers LCMV, MVDR et la méthode proposée font référence aux filtres $\mathbf{w}_{\mathcal{M}_1}$, $\mathbf{w}_{\mathcal{M}_2}$ et $\mathbf{w}_{\mathcal{M}_3}$, respectivement.

5.4.1 Méthodes

Génération des signaux

Les algorithmes sont testés en traitant des scènes auditives virtuelles composées de trois sources de parole d'une durée de 4 s et d'un bruit de cafétéria joué sur deux anneaux de haut-parleurs virtuels situés à des altitudes $\pm 45°$ mélangés à différents RSBs allant de 0 à 10 dB avec un pas de 2,5 dB. Les signaux de parole sont issus d'enregistrements de la radio France Culture, échantillonnés à 16 kHz et spatialisés sur le plan horizontal aux azimuts $\{-45°, 0°, 45°\}$. Les ATFs de prothèses auditives ($M = 4$) utilisées pour la génération de la scène auditive virtuelle et les algorithmes de beamforming proviennent de [Oreinos and Buchholz, 2013]. Au total, 40 exemples audio sont générés pour chaque RSB testé.

Les algorithmes sont intégrés dans une chaîne de traitement par trame de type ajout-recouvrement pondéré avec une taille de fenêtre de 128 échantillons (durée de 8 ms) et un recouvrement de 50 % (voir section 1.3 (p.19) pour plus de détails). Chaque trame est exprimée dans le domaine fréquentiel sans complétion de zéros. Le beamformer MVDR et la méthode proposée sont testées en utilisant le VAD basé sur le RSB idéal et estimé.

Critères d'évaluation

Pour évaluer la réduction du bruit, nous considérons le rapport signal-à-artefact (SAR) et les améliorations en matière de rapport signal-à-distorsion (ΔSDR) et de rapport signal-à-interférences (ΔSIR) [Vincent et al., 2006] qui sont définis respectivement comme le rapport entre la puissance du signal cible, tel que défini en Eq. (5.6), et (i) les artefacts générés par le beamforming, (ii) les autres composantes du signal de sortie, et (iii) la composante de bruit.

5.4.2 Résultats

Les résultats sont rapportés en Fig. 5.5 et Fig. 5.6. Tout d'abord, comparons le ΔSDR et la complexité de calcul en Fig. 5.5. Nous observons que le beamformer MVDR est plus de 5 fois moins complexe que le beamformer LCMV et qu'il

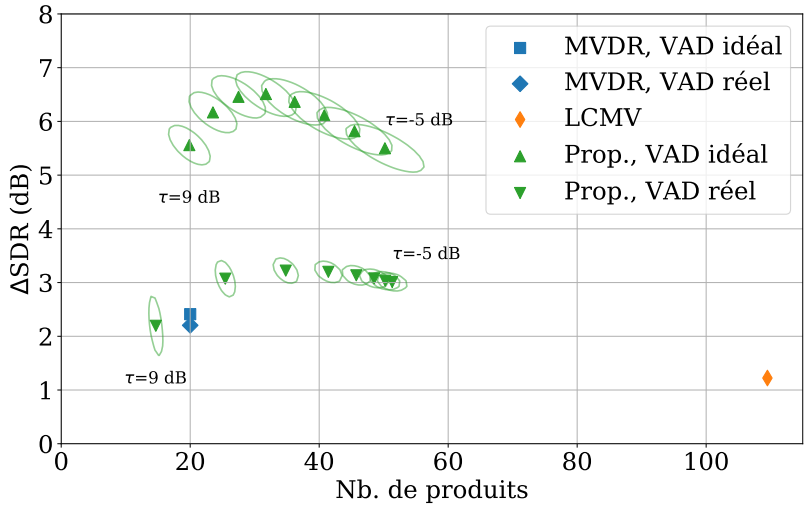


FIGURE 5.5 – Amélioration du SDR versus la complexité algorithmique exprimé en nombre de produit, moyenné sur le nombre de stimuli pour un RSB de 0 dB. Les ellipses illustrent l'écart-type d'une distribution gaussienne 2D ajustée sur les résultats de la méthode proposée avec un seuil de VAD allant de -5 à 9 dB avec un pas de 2 dB.

améliore le ΔSDR d'environ 1 dB par rapport à ce dernier. Dans le scénario testé, les deux algorithmes ont des performances de réduction de la distorsion très faibles. En utilisant un VAD basé sur un RSB oracle, l'algorithme proposé améliore les performances du ΔSDR de 6,5 dB avec un réglage optimal du seuil de détection tout en augmentant que légèrement la complexité algorithmique par rapport au beamformer MVDR (+50 %). On peut noter grâce aux ellipses représentant l'écart-type des données que le ΔSDR est négativement corrélé avec la complexité de calcul. En effet, plus les sources de parole se recouvrent, plus le nombre de contraintes dans l'optimisation est important, ce qui augmente le coût de calcul et réduit la taille du sous-espace sur lequel la réduction du bruit peut être effectuée, entraînant une performance plus faible en moyenne. Lorsque l'on utilise le VAD basé sur le RSB estimé (et non l'oracle), la complexité algorithmique reste inchangée mais les performances de la méthode proposée en matière de ΔSDR diminuent de manière significative, même si elle reste plus performante que les deux autres beamformers. Cela montre que le VAD a un impact important sur les performances de réduction du bruit de la méthode de beamforming proposée.

Maintenant, examinons de plus près les performances de débruitage en étudiant le ΔSIR et le SAR en fonction du RSB d'entrée, comme indiqué

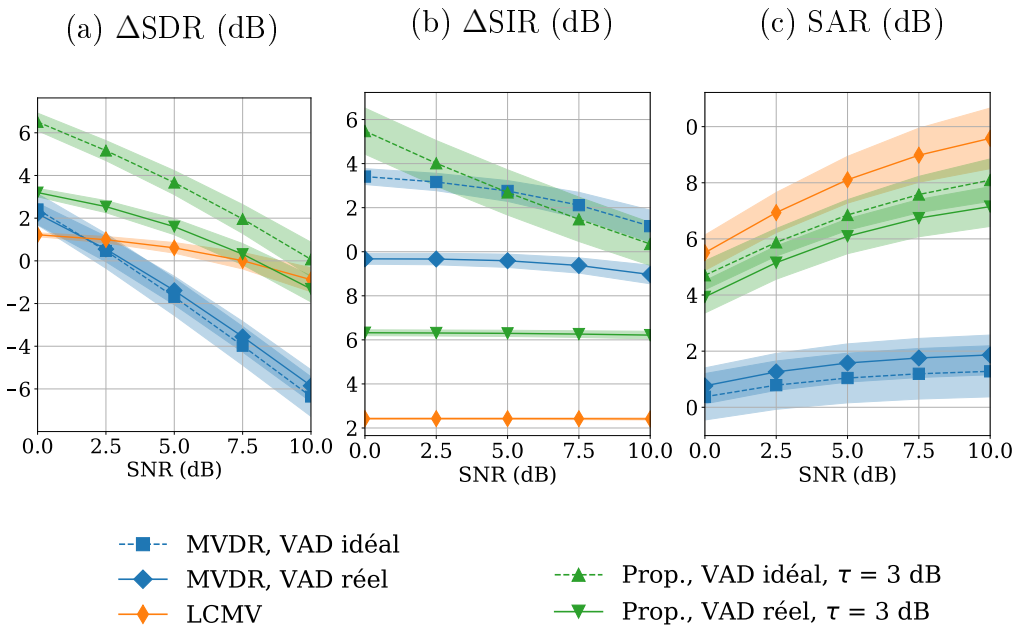


FIGURE 5.6 – ΔSDR (a), ΔSIR (b) and SAR (c) en fonction du RSB. Le trait plein illustre la moyenne et la surface l'écart-type. Nous montrons ici uniquement les performances de la méthode proposée avec le seuil du VAD τ=3 dB qui maximise le ΔSDR comme montré en Fig. 5.5.

dans Fig. 5.6. Le beamformer MVDR est très efficace pour réduire le bruit (ΔSIR=13 dB à RSB=0 dB avec le VAD idéal) comparé au beamformer LCMV (ΔSIR=2.5 dB). On s'y attend car le premier n'utilise qu'un seul degré de liberté de l'optimisation pour traiter la préservation des sources de parole. Cependant, en ne préservant qu'une source par point T-F, il introduit plus d'artefacts (SAR=10 dB à RSB=0 dB) comparé au beamformer LCMV (SAR=15 dB). Pour un RSB d'entrée élevé, la quantité d'artefacts introduite par le beamformer MVDR peut devenir plus importante que le niveau de la composante de bruit dans le mélange original, ce qui entraîne un ΔSDR négatif comme on peut le voir dans Fig. 5.6. La méthode proposée atteint une performance de SAR similaire à celle obtenue avec le beamformer LCMV, bien que légèrement inférieure. En ce qui concerne le ΔSIR, elle obtient une amélioration de 15 dB (à RSB=0 dB) pour un réglage de seuil maximisant le ΔSDR (τ=3 dB). Cependant, ce score diminue fortement lorsqu'on utilise le VAD basé sur le RSB estimé. En effet, celui-ci a tendance à faire beaucoup de faux positifs, réduisant ainsi le nombre de degrés de liberté pour le débruitage. Néanmoins, comme le montre la mesure de performance globale ΔSDR, la méthode proposée obtient des résultats similaires ou meilleurs par rapport aux deux autres

beamformers, tout en étant efficace en matière de complexité de calcul comme indiqué précédemment.

Enfin, notons que l'algorithme proposé est plus sensible aux erreurs d'estimation du VAD que le beamformer MVDR. En effet, ce dernier n'a besoin de connaître que la source la plus énergique alors que le premier a besoin de savoir précisément quelle source est active ou non.

5.5 Conclusion

Dans ce chapitre, nous avons proposé un nouvel algorithme de beamforming qui exploite la parcimonie des signaux de parole dans le domaine de la TFCT, d'une manière moins restrictive par rapport à l'hypothèse populaire de non-recouvrement totale des sources de parole [Rickard and Yilmaz, 2002]. Dans une scène sonore composée de trois locuteur·rice·s, les résultats expérimentaux montrent que les beamformers LCMV et MVDR présentent deux comportements extrêmes : le premier préserve bien les sources de parole mais ne permet pas une bonne réduction du bruit, tandis que le second réduit considérablement le bruit et la complexité de calcul mais introduit beaucoup d'artefacts. La méthode proposée parvient à être bénéfique à la fois en matière de réduction du bruit et de distorsion de la parole sans trop augmenter le coût de calcul par rapport au beamformer MVDR.

Limites et perspectives

Vers plus de sources et plus de microphones ? Nous avons limité l'étude à $Q = 3$ car nous avons utilisé un réseau de quatre microphones. En effet, il est rare que les prothèses auditives embarquent plus de quatre microphones étant donné le faible gain apporté par des microphones supplémentaires comparé au coût en ressource calculatoire qu'ils nécessitent (numérisation, passage dans le domaine fréquentiel, communication sans fil, etc.). Les travaux futurs devront étudier les performances de la méthode proposée pour un réseau de microphones plus grand [Kayser et al., 2009] ou tout simplement pour des configurations sous-déterminées ($Q > M$). Le VAD peut lui aussi être amélioré pour se rapprocher de la limite supérieure des performances de réduction du bruit.

Parcimonie harmonique Aussi, on peut noter qu'une nouvelle méthode de beamforming considérant des sources harmoniques dans son modèle a été proposée récemment [Jensen et al., 2020]. Celle-ci permet notamment de réaliser

simultanément débruitage et déréverbération même en présence d'un bruit assez fort (RSB = 0 dB). Cette méthode a pour l'instant été proposée dans un scénario considérant une seule source de parole voisée, mais elle pourrait être étendue à plusieurs sources. En principe, ce modèle plus rigide de structuration de la parole peut augmenter la robustesse du débruitage. Néanmoins, dans leur étude, [Jensen et al., 2020] considèrent une fenêtre d'analyse de 20 ms, correspondant à une résolution fréquentielle de 50 Hz, là où la contrainte de faible latence dans le contexte des prothèses auditives nous impose une fenêtre d'analyse inférieure à 8 ms, soit une résolution fréquentielle maximale de 125 Hz. Or, comme la fréquence fondamentale de la parole est de l'ordre de 100 Hz pour un homme, cette résolution peut se révéler insuffisante pour appliquer leur méthode. Toutefois, les auteurs ne rapportent pas la sensibilité de leur méthode à ce paramètre dans leur étude.

Mutualisation des résultats intermédiaires Reprenons l'architecture de l'algorithme de beamforming LCMV avec compression indépendante (CLCMV) introduit par [Corey and Singer, 2017], illustré en Fig. 2.12 (p.72). On remarque que le beamformer MVDR est utilisé à plusieurs étapes, aussi bien pour estimer les gains de compression à appliquer à chaque source que pour l'estimation des DOAs (voir annexe B (p.149)). On peut aussi noter que le beamformer LCMV peut être décomposé comme une somme de beamformers MVDR, comme il a été montré par [Hadad et al., 2016] dans le cas où $Q = 2$ et comme nous en apportons la preuve en annexe E (p.159) pour $Q = 3$. Cette décomposition est valable aussi bien pour le beamformer CLCMV. Aussi, nous montrons en annexe D (p.155) que l'estimateur de probabilité de présence de la parole multicanale (PPPM) se décompose également en un algorithme plus simple faisant apparaître des beamformers MVDR pointés en direction des sources cibles. Tous ces éléments suggèrent que l'on peut réduire le coût calculatoire du beamformer CLCMV, ainsi que des estimateurs de paramètres associés, en mutualisant les résultats intermédiaires fournis par les beamformers MVDR orientés vers chaque cibles.

Chapitre 6

Conclusion générale

6.1 Synthèse

Interaction entre réduction de bruit et compression de dynamique Nous avons analysé l'influence des algorithmes de débruitage et de compression de dynamique sur les performances de compréhension de la parole et de tâches relevant de l'écoute spatiale comme la localisation auditive mais aussi la perception plus globale de l'espace sonore, ou la réorientation vers une nouvelle source de parole.

Nous avons identifié que les deux familles d'algorithmes étudiées pouvaient entrer en interaction mais que leur étude et leur développement étaient souvent réalisés de manière indépendante. Nous avons alors analysé les quelques travaux qui avaient traité cette question et discuté leurs limites, que l'on peut résumer en deux points : (i) la plupart traitent la question en utilisant un filtre de débruitage monocanal, connu pour avoir un impact limité sur l'intelligibilité de la parole comparé aux algorithmes multicanaux ; et (ii) le manque de formalisme théorique commun entre les algorithmes de débruitage et la compression de dynamique empêche une co-conception motivée par une modélisation.

Unification de la réduction de bruit et de la compression de dynamique Nous avons alors proposé une façon d'unifier au sein d'un même problème d'optimisation, les tâches de débruitage et de compensation des pertes auditives, via la compression de dynamique. Pour ce faire, nous avons fait des choix de modélisation de sorte à fondre les deux problèmes au sein d'un formalisme commun. Le choix des approximations du modèle a été guidé par la volonté de conserver une interprétabilité du résultat et une solution analytique, de sorte à pouvoir la comparer explicitement avec les méthodes de beamforming et de compression classiques de la littérature. La méthode proposée a alors été comparée à ces dernières tant sur des critères objectifs, *i.e.* améliora-

tion du RSB, compression effective, préservation de la cohérence interaurale, que perceptifs, *i.e.* intelligibilité de la parole, test de préférence subjectif.

Les résultats ont montré que la méthode proposée préserve mieux les caractéristiques spatiales, d'enveloppe et de dynamique du bruit ambiant, que les méthodes basées sur une mise en cascade des algorithmes de débruitage et de compression. Aussi, le RSB est amélioré en sortie lorsque celui-ci supérieur à 0 dB en entrée. Perceptivement, cela se traduit par une meilleure perception globale de la scène sonore et des performances de compréhension de la parole similaires à celles obtenues avec l'approche état-de-l'art. Aussi, la comparaison avec la méthode de [Ngo et al., 2012] a montré que la binauralisation [1] des prothèses auditives était cruciale pour les performances de la solution proposée.

Préservation des indices interauraux de l'ensemble de la scène sonore L'algorithme que nous avons proposé, issu de l'unification des problèmes d'optimisation de débruitage et de compression de la parole, fait apparaître un estimateur de la composante de bruit. Bien que celui-ci permette de préserver globalement les HRTFs sur l'ensemble de la sphère, il distord particulièrement les indices de localisation interauraux, *i.e.* ITD et ILD, sur l'hémisphère frontal. Nous avons alors proposé trois algorithmes d'estimation de bruit intégrant la préservation de tels indices. En particulier, nous avons proposé une formulation déterministe du problème d'optimisation, plutôt que probabiliste, qui a permis de développer deux des trois algorithmes proposés. Ce sont aussi ces derniers qui ont obtenus les meilleurs performances d'annulation de la cible et de préservation des indices de localisation sur l'ensemble de la sphère.

Environnement multilocuteur : un cas très contraint Jusqu'alors, nous nous étions intéressés essentiellement à une scène sonore composée d'une source de parole cible dans la direction frontale et d'un bruit spatialement diffus. Or, le cas d'une situation multilocuteur est loin d'être anecdotique. Bien que des méthodes existent pour traiter ce cas, il reste très compliqué à appréhender étant donné le faible nombre de microphones utilisés dans les prothèses auditives binaurales (quatre, voire six) comparé au nombre de sources de parole potentiellement présentes dans la scène sonore. Par ailleurs, ces méthodes augmentent fortement en complexité avec le nombre de sources de parole. D'où un coût calculatoire supplémentaire potentiellement problématique, étant données les faibles capacités de calcul des prothèses auditives ainsi que l'exigence de faible latence du contexte applicatif. Nous avons alors proposé d'exploiter plus finement les propriétés de parcimonie de la parole dans le domaine temps-fréquence que ce qui avait été fait jusqu'alors. Nous avons aussi apporté la

[1] *i.e.* leur mise en communication.

preuve de la décomposition du beamformer LCMV en une somme de beamformers MVDR, pour le cas particulier où l'on considère trois sources, ainsi que la décomposition de l'estimateur de probabilité de présence de la parole multicanale (PPPM) en un estimateur monocanal prenant en entrée le résultat d'une somme de beamformers MVDR quelque soit le nombre de sources de parole. Ces deux dernières preuves sont un pas de plus pour montrer l'intérêt de mutualiser des résultats intermédiaires dans l'implémentation des prothèses auditives.

6.2 Perspectives

Prendre en compte la compression fréquentielle Une perte auditive peut devenir trop importante pour être compensée par un compresseur de dynamique. Dans ce cas, il peut être envisagé de déplacer le contenu fréquentiel devenu inaccessible dans une bande de fréquence où la perte peut-être compensée, c'est ce qu'on appelle la compression fréquentielle [Glista et al., 2009]. Ce traitement, fondamentalement non-linéaire, est moins souvent employé que la compression de dynamique et le bénéfice qu'il apporte n'est pas très bien déterminé [McCreery et al., 2012b]. Aussi, sa prise en compte dans le processus d'unification des traitements est rendue difficile du fait qu'il implique de mélanger les signaux de chaque bande fréquentielle. En effet, le formalisme que nous avons utilisé dans cette thèse considère essentiellement chaque bande indépendamment. C'est pourquoi, nous avons choisi de ne pas le traiter dans cette thèse. Néanmoins, son intégration dans le formalisme commun peut être une piste de recherche future.

Estimer les gains de compression Dans l'algorithme de [Corey and Singer, 2017], les gains de compression à appliquer à chaque source de parole sont estimés avec un beamformer MVDR. Alternativement, l'estimateur de l'enveloppe de la source de [Thiemann et al., 2016] pourrait être utilisé par exemple. Pour l'heure, il n'est pas clair de savoir si tel ou tel estimateur a un impact sur les performances finales de débruitage, d'intelligibilité de la parole ou encore de perception de la scène sonore.

Faire le deuil de l'interprétabilité Dans cette thèse, nous avons unifié le débruitage multicanal et la compression de dynamique au sein d'un même problème d'optimisation. L'écart entre le formalisme mathématique utilisé en psychoacoustique et celui utilisé en traitement de scène sonore acoustique est conséquent. Pour les relier, il est nécessaire de faire des choix en ayant recours à quelques approximations. Pour guider ces choix, nous avons privilégié de rester

dans le giron du formalisme du traitement de scène sonore acoustique, nous permettant une plus grande interprétabilité de la solution. Néanmoins, une piste pour le futur peut être de prendre l'autre voie : privilégier la justesse des modèles psychoacoustiques pour s'assurer que le problème d'optimisation que l'on traite est au plus proche des attentes de notre audition. Cette approche rend difficile le développement d'algorithmes basés sur une résolution analytique ou numérique du problème d'optimisation. Toutefois, il est envisageable d'utiliser des méthodes d'apprentissage machine de sorte à approcher la fonction que l'on cherche. En particulier, les modèles fonctionnels de localisation auditive [Baumgartner et al., 2014, Baumgartner et al., 2016] font intervenir des bancs de filtres particuliers et des fonctions non-linéaires tout à fait compatibles avec la rétropropagation du gradient de l'erreur nécessaire au développement de méthodes basées sur les réseaux de neurones profonds. Dans la même veine, des algorithmes de débruitage basés sur ces derniers ont récemment été proposés en intégrant dans la fonction de coût un modèle psychoacoustique plutôt qu'une simple erreur quadratique moyenne [Saddler et al., 2021, Zhang et al., 2021].

Annexe A

Estimation d'RTF en temps réel par blanchiment

Cette méthode consiste à utiliser la propriété hermitienne de la matrice de covariance du bruit. Il est ainsi possible d'utiliser la décomposition de Cholesky de $\mathbf{\Phi_n}(k,\ell)$, notée $\mathbf{\Phi}_{\mathbf{n}}^{\frac{1}{2}}(k,\ell)$:

$$\mathbf{\Phi_n}(k,\ell) = \left(\mathbf{\Phi}_{\mathbf{n}}^{\frac{1}{2}}(k,\ell)\right)^H \mathbf{\Phi}_{\mathbf{n}}^{\frac{1}{2}}(k,\ell). \tag{A.1}$$

On peut alors définir une version du vecteur de signaux dans lequel le bruit est spatialement blanc *i.e.* sa matrice de covariance est l'identité :

$$\mathbf{b}(k,\ell) = \mathbf{\Phi}_{\mathbf{n}}^{-\frac{1}{2}}(k,\ell)\mathbf{x}(k,\ell). \tag{A.2}$$

En effet, il est direct de montrer que la matrice de covariance de $\mathbf{b}(k,\ell)$, notée $\mathbf{\Phi_b}(k,\ell)$, peut s'exprimer de la manière suivante :

$$\mathbf{\Phi_b}(k,\ell) = \phi_s(k,\ell)\mathbf{g}(k)\mathbf{g}(k)^H + \mathbf{I}, \tag{A.3}$$

où $\mathbf{g}(k)$ est l'ATF dans la version du signal *blanchi* :

$$\mathbf{g}(k) = \mathbf{\Phi}_{\mathbf{n}}^{-\frac{1}{2}}(k,\ell)\mathbf{h}(k). \tag{A.4}$$

Maintenant, appliquons la décomposition en composantes principales à $\mathbf{\Phi_b}(k,\ell)$:

$$\mathbf{\Phi_b}(k,\ell) = \mathbf{V}(k,\ell)\mathbf{\Lambda}(k,\ell)\mathbf{V}(k,\ell)^H, \tag{A.5}$$

où $\mathbf{V}(k,\ell)$ forme une base orthonormale et $\mathbf{\Lambda}(k,\ell)$ est la matrice diagonale composée de valeurs propres classées par ordre décroissant. En mettant cette expression en regard avec l'Eq. (A.3), il est clair que la première colonne de $\mathbf{V}(k,\ell)$, notée $\mathbf{v}_1(k,\ell)$, contient $\mathbf{g}(k)$ à un facteur près.

Afin d'obtenir l'estimation de la RTF, notée $\hat{\tilde{\mathbf{h}}}_{\mathrm{CW}}(k,\ell)$, il suffit alors de *déblanchir* $\mathbf{v}_1(k,\ell)$ et de le normaliser par son $r^{\text{ème}}$ élément :

$$\hat{\tilde{\mathbf{h}}}_{\mathrm{CW}}(k,\ell) = \frac{\hat{\mathbf{\Phi}}_{\mathbf{n}}^{\frac{1}{2}}(k,\ell)\hat{\mathbf{v}}_1(k,\ell)}{\mathbf{q}_r^H \hat{\mathbf{\Phi}}_{\mathbf{n}}^{\frac{1}{2}}(k,\ell)\hat{\mathbf{v}}_1(k,\ell)}. \tag{A.6}$$

Annexe B

Estimateur de DOA considérant une cible déterministe et un bruit aléatoire

B.1 Modèle des signaux

Supposons que la source s est une valeur complexe déterministe et que le bruit $\mathbf{n}$ est une variable aléatoire suivant une distribution Gaussienne multivariée centrée isotropique complexe de matrice de covariance $\mathbf{\Phi_n}$. Le modèle de vraisemblance des données peut s'écrire comme suit :

$$P(\mathbf{x}|\theta, s, \mathbf{\Phi_n}) = \frac{1}{\pi^M |\mathbf{\Phi_n}|} e^{-(\mathbf{x}-\mathbf{h}(\theta)s)^H \mathbf{\Phi_n}^{-1}(\mathbf{x}-\mathbf{h}(\theta)s)}, \tag{B.1}$$

avec $\mathbf{h}(\theta)$ les ATFs correspondant à une source située à l'azimut θ.

B.2 Formulation du problème et solution

Définissons alors la fonction de log-vraisemblance $\mathcal{L}(\mathbf{X}|\theta, s, \mathbf{\Phi_n})$ avec $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_L]$ l'ensemble des observations des microphones sur les L trames tem-

porelles passées :

$$\begin{aligned}\mathcal{L}(\mathbf{X}|\theta, s, \mathbf{\Phi_n}) &= \log \prod_{\lambda=1}^{L} P(\mathbf{x}_\lambda|\theta, s_\lambda, \mathbf{\Phi_n}) \\ &\stackrel{c}{=} -L \log |\mathbf{\Phi_n}| - \sum_{\lambda=1}^{L} (\mathbf{x}_\lambda - \mathbf{h}(\theta)s_\lambda)^H \mathbf{\Phi_n}^{-1} (\mathbf{x}_\lambda - \mathbf{h}(\theta)s_\lambda) \\ &\stackrel{c}{=} -LM \log \phi_n - \frac{1}{\phi_n} \sum_{\lambda=1}^{L} (\mathbf{x}_\lambda - \mathbf{h}(\theta)s_\lambda)^H \mathbf{\Gamma_n}^{-1} (\mathbf{x}_\lambda - \mathbf{h}(\theta)s_\lambda),\end{aligned} \tag{B.2}$$

où $\stackrel{c}{=}$ signifie l'égalité à une constante près. Constante que l'on peut omettre car la solution du problème d'optimisation minimisant cette fonction de coût est invariante à celle-ci. On peut obtenir l'estimation de ϕ_n, notée $\hat{\phi}_n$, en annulant la dérivée de la log-vraisemblance par rapport à ϕ_n :

$$\begin{aligned}&\frac{\delta \mathcal{L}}{\delta \phi_n} = 0 \\ \Leftrightarrow \quad &\hat{\phi_n} = \frac{1}{LM} \sum_{\lambda=1}^{L} (\mathbf{x}_\lambda - \mathbf{h}(\theta)s_\lambda)^H \mathbf{\Gamma_n}^{-1} (\mathbf{x}_\lambda - \mathbf{h}(\theta)s_\lambda).\end{aligned} \tag{B.3}$$

De la même manière, on peut obtenir l'estimateur de s_λ, noté $\hat{s}_\lambda$:

$$\begin{aligned}&\frac{\delta \mathcal{L}}{\delta s_\lambda^*} = 0 \\ \Leftrightarrow \quad &\hat{s}_\lambda = \frac{\mathbf{h}(\theta)^H \mathbf{\Gamma_n}^{-1} \mathbf{x}_\lambda}{\mathbf{h}(\theta)^H \mathbf{\Gamma_n}^{-1} \mathbf{h}(\theta)}.\end{aligned} \tag{B.4}$$

Il est remarquable que la solution est la même que celle du beamformer MVDR, présenté en Eq. (2.42) (p.45). Cela montre que ce beamformer est la solution optimale au sens du maximum de vraisemblance pour ce modèle.

En introduisant l'Eq. (B.4) et l'Eq. (B.3) en Eq. (B.2), on obtient $\hat{\mathcal{L}}(\mathbf{X}|\theta, \mathbf{\Gamma_n})$ à partir de laquelle on va dériver la fonction de coût de l'estimateur maximisant la vraisemblance du modèle à source déterministe (ou *Deterministic Maximum*

Likelihood) (DML), notée $J_{\text{DML}}(\theta)$:

$$\begin{aligned}
\hat{\theta}_{\text{DML}}(k,\ell) &= \underset{\theta}{\text{argmax}} \left\{ \hat{\mathcal{L}}(\mathbf{X}|\theta, \mathbf{\Gamma_n}) \right\} \\
&= \underset{\theta}{\text{argmax}} \left\{ -LM \log \sum_{\lambda=1}^{L} (\mathbf{x}_\lambda - \mathbf{h}(\theta)\hat{s}_\lambda)^H \mathbf{\Gamma}_\mathbf{n}^{-1} (\mathbf{x}_\lambda - \mathbf{h}(\theta)\hat{s}_\lambda) \right\} \\
&= \underset{\theta}{\text{argmax}} \left\{ -\sum_{\lambda=1}^{L} (\mathbf{x}_\lambda - \mathbf{h}(\theta)\hat{s}_\lambda)^H \mathbf{\Gamma}_\mathbf{n}^{-1} (\mathbf{x}_\lambda - \mathbf{h}(\theta)\hat{s}_\lambda) \right\} \\
&= \underset{\theta}{\text{argmax}} \left\{ \underbrace{\frac{\mathbf{h}(\theta)^H \mathbf{\Gamma}_\mathbf{n}^{-1} \hat{\mathbf{\Phi}}_\mathbf{x} \mathbf{\Gamma}_\mathbf{n}^{-1} \mathbf{h}(\theta)}{\mathbf{h}(\theta)^H \mathbf{\Gamma}_\mathbf{n}^{-1} \mathbf{h}(\theta)}}_{J_{\text{DML}}(\theta)} \right\},
\end{aligned} \tag{B.5}$$

avec

$$\hat{\mathbf{\Phi}}_\mathbf{x} = \frac{1}{L} \sum_{\lambda=1}^{L} \mathbf{x}_\lambda \mathbf{x}_\lambda^H. \tag{B.6}$$

On remarque que dans le cas où $\mathbf{\Gamma_n} = \mathbf{I}$, $J_{\text{DML}}(\theta)$ devient $J_{\text{SRP}}(\theta)$ la fonction de coût associée à l'algorithme SRP [Brandstein et al., 2001, Chapt. 8], un algorithme largement utilisé pour la localisation sonore en général :

$$\begin{aligned}
J_{\text{SRP}}(\theta) &= \frac{\mathbf{h}(\theta)^H \mathbf{\Gamma}_\mathbf{n}^{-1} \hat{\mathbf{\Phi}}_\mathbf{x} \mathbf{\Gamma}_\mathbf{n}^{-1} \mathbf{h}(\theta)}{\mathbf{h}(\theta)^H \mathbf{\Gamma}_\mathbf{n}^{-1} \mathbf{h}(\theta)} \\
&= \frac{\mathbf{h}(\theta)^H \hat{\mathbf{\Phi}}_\mathbf{x} \mathbf{h}(\theta)}{||\mathbf{h}(\theta)||^2} \\
&= \frac{1}{L} \sum_{\lambda=1}^{L} \left| \frac{\mathbf{h}(\theta)^H}{||\mathbf{h}(\theta)||} \mathbf{x}_\lambda \right|^2 \\
&= \frac{1}{L} \sum_{\lambda=1}^{L} \left| \mathbf{w}_{\text{SRP}}^H \mathbf{x}_\lambda \right|^2,
\end{aligned} \tag{B.7}$$

où

$$\mathbf{w}_{\text{SRP}} = \frac{\mathbf{h}(\theta)}{||\mathbf{h}(\theta)||}, \tag{B.8}$$

est appelé le *match-filter*, très proche de $\mathbf{w}_0$, le filtre du beamformer aligneur défini en Eq. (2.37) (p.42). D'après [Zohourian et al., 2018], l'usage de l'estimation de $\mathbf{\Gamma_n}$ plutôt que de prendre naïvement $\mathbf{I}$, entraine des erreurs d'estimation qui ne compensent pas le gain que pourrait apporter une plus grande complexité du modèle.

Annexe C

Estimateur de DOA considérant une cible stochastique

C.1 Modèle des signaux

L'estimateur DML, présenté en annexe B supposait le coefficient de Fourier de la parole déterministe et ceux du bruit aléatoires. Supposons maintenant que les deux sont des variables aléatoires indépendantes, leur somme suit donc aussi une distribution Gaussienne. Alors, le modèle de vraisemblance des données peut s'écrire comme suit :

$$P(\mathbf{x}|\theta, \mathbf{\Phi_x}) = \frac{1}{\pi^M |\mathbf{\Phi_x}|} e^{-\mathbf{x}^H \mathbf{\Phi_x}^{-1} \mathbf{x}}. \tag{C.1}$$

L'estimateur associé à ce modèle est nommé l'estimateur maximisant la vraisemblance du modèle stochastique (de l'anglais *Stochastic Maximum Likelihood*, SML).

C.2 Formulation du problème et solution

On va alors chercher l'azimut θ maximisant la log-vraisemblance, notée $\mathcal{L}(\mathbf{X}|\theta, \mathbf{\Phi_x})$, pour ce modèle. Celle-ci s'exprime comme suit :

$$\mathcal{L}(\mathbf{X}|\theta, \mathbf{\Phi_x}) = \log \prod_{\lambda=1}^{L} P(\mathbf{x}_\lambda|\theta, \mathbf{\Phi_x}) \tag{C.2}$$

$$= -LM \log \pi - L \log \left|\mathbf{\Phi_x}\right| - L\mathrm{Tr}\left\{\mathbf{\Phi_x}^{-1} \hat{\mathbf{\Phi}}_\mathbf{x}\right\}. \tag{C.3}$$

Après quelques manipulations [Ye and DeGroat, 1995], on obtient la fonction de coût, notée $J_{\mathrm{SML}}(\theta)$, associée à $\mathcal{L}(\mathbf{X}|\theta, \mathbf{\Phi_x})$:

$$J_{\mathrm{SML}}(\theta) = -\log \big|\mathbf{h}(\theta)\mathbf{w}_{\mathrm{MVDR}}(\theta)^H \hat{\mathbf{\Phi}}_{\mathbf{x}} \mathbf{w}_{\mathrm{MVDR}}(\theta)\mathbf{h}(\theta)^H + \hat{\phi_n}(\mathbf{I} - \mathbf{h}(\theta)\mathbf{w}_{\mathrm{MVDR}}(\theta)^H)\mathbf{\Gamma_n}\big|,$$

où $\mathbf{w}_{\mathrm{MVDR}}(\theta)$ est le filtre du beamformer MVDR visant la direction d'azimut θ sur le plan horizontal.

Annexe D

Décomposition de l'estimateur de PPPM en une somme de beamformers MVDR et d'un estimateur de PPP

La connaissance de la probabilité de présence de la parole (PPP) en présence de bruit a été largement utilisée dans le cadre des algorithmes de débruitage monocanaux [Cohen, 2002] et multicanaux [Ngo et al., 2009, Bagheri and Giacobello, 2019, Thiergart and Habets, 2014] pour améliorer leurs performances. L'estimateur de la PPP dans le bruit basé sur un seul microphone (monocanal) a été largement étudié et utilisé [Middleton and Esposito, 1968, Cohen, 2003] avant d'être étendu au cas multicanal [Souden et al., 2010].

Nous montrons ici l'équivalence entre l'estimateur de la probabilité de présence de la parole multicanale (PPPM) et un estimateur de PPP monocanal basé sur la sortie de Q beamformers MVDR visant chacune des Q sources de parole, lorsque les ATFs de celles-ci sont courtes. Nous montrons aussi que l'estimateur de PPPM peut être décomposé comme la mise en série d'un beamformer MVDR avec un estimateur de PPP monocanal, lorsque la matrice de covariance de la parole est de rang 1, *i.e.* une seule source cible localisée, sans réverbération. Ceci est utile lorsqu'on souhaite estimer la PPP d'une source dont l'ATF est connue. En particulier, si plusieurs sources de parole sont présentes dans une même scène sonore et que celles-ci doivent être considérées indépendamment [Thiergart and Habets, 2014]. Par ailleurs, dans le cas où l'estimateur de PPPM s'intègre dans un système comprenant déjà un beamformer MVDR, ce résultat permet d'utiliser un estimateur de PPP monocanal plutôt que multicanal, moins complexe, réduisant ainsi le coût calculatoire du système global. À notre connaissance, ces résultats sont originaux.

D.1 Modèle des signaux

Nous considérons ici le modèle présenté en 2.2.2 (p.39) dont nous rappelons la version du modèle de mélange dans le domaine de la TFCT pour Q sources de parole :

$$\mathbf{x}(k,\ell) = \mathbf{H}(k)\mathbf{s}(k,\ell) + \mathbf{n}(k,\ell). \tag{D.1}$$

En utilisant la propriété d'indépendance de la source et du bruit, on peut exprimer la matrice de covariance des signaux des microphones comme suit :

$$\mathbf{\Phi}_{\mathbf{x}}(k,\ell) = \tilde{\mathbf{\Phi}}_{\mathbf{s}}(k,\ell) + \mathbf{\Phi}_{\mathbf{n}}(k,\ell), \tag{D.2}$$

où la matrice $\tilde{\mathbf{\Phi}}_{\mathbf{s}}(k,\ell)$ peut être décomposée de la manière suivante :

$$\tilde{\mathbf{\Phi}}_{\mathbf{s}}(k,\ell) = \mathbf{H}(k)\mathbf{\Phi}_{\mathbf{s}}(k,\ell)\mathbf{H}(k)^H, \tag{D.3}$$

avec $\mathbf{\Phi}_{\mathbf{s}}(k,\ell) = \text{diag}\{\phi_{s_1}(k,\ell), ..., \phi_{s_Q}(k,\ell)\}$, la matrice de covariance de la parole, et $\mathbf{\Phi}_{\mathbf{n}}(k,\ell)$, celle du bruit :

$$\mathbf{\Phi}_{\mathbf{n}}(k,\ell) = \phi_n(k,\ell)\mathbf{\Gamma}_{\mathbf{n}}(k). \tag{D.4}$$

D.2 Preuve de l'équivalence

En premier lieu, rappelons l'expression de l'estimateur de PPP monocanal [Middleton and Esposito, 1968] :

$$p_{\text{PPP}}(k,\ell) = \left(1 + \frac{q(k,\ell)}{1-q(k,\ell)}(1+\xi(k,\ell))e^{-\gamma(k,\ell)\frac{\xi(k,\ell)}{1+\xi(k,\ell)}}\right)^{-1}, \tag{D.5}$$

où $\xi(k,\ell) = \frac{\phi_s(k,\ell)}{\phi_n(k,\ell)}$ est le RSB *a priori*, $\gamma(k,\ell) = \frac{|x(k,\ell)|^2}{\phi_n(k,\ell)}$ est appelé le RSB *a posteriori* et $q(k,\ell)$ est la probabilité *a priori* d'absence de la parole. L'estimateur de PPPM [Souden et al., 2010] admet, quant à lui, la solution suivante :

$$p_{\text{PPPM}}(k,\ell) = \left(1 + \frac{q(k,\ell)}{1-q(k,\ell)}(1+\zeta(k,\ell))e^{-\frac{\beta(k,\ell)}{1+\zeta(k,\ell)}}\right)^{-1}, \tag{D.6}$$

avec

$$\beta(k,\ell) = \mathbf{x}(k,\ell)^H\mathbf{\Phi}_{\mathbf{n}}^{-1}(k,\ell)\tilde{\mathbf{\Phi}}_{\mathbf{s}}(k,\ell)\mathbf{\Phi}_{\mathbf{n}}^{-1}(k,\ell)\mathbf{x}(k,\ell), \tag{D.7}$$

et $\zeta(k,\ell)$, appelé le RSB multicanal *a priori*, défini comme suit :

$$\zeta(k,\ell) = \text{Tr}\left\{\mathbf{\Phi}_{\mathbf{n}}^{-1}(k,\ell)\tilde{\mathbf{\Phi}}_{\mathbf{s}}(k,\ell)\right\}. \tag{D.8}$$

On rappelle aussi l'expression du beamformer MVDR, orienté vers la $i^{\text{ème}}$ source de parole :

$$\mathbf{w}_{\text{MVDR},i}(k) = \frac{\mathbf{\Gamma}_{\mathbf{n}}^{-1}(k)\mathbf{h}_i(k)}{\mathbf{h}_i(k)^H\mathbf{\Gamma}_{\mathbf{n}}^{-1}(k)\mathbf{h}_i(k)}, \tag{D.9}$$

où $\mathbf{h}_i(k)$ est la $i^{\text{ème}}$ colonne de $\mathbf{H}(k)$, soit les ATFs entre la $i^{\text{ème}}$ source de parole et les microphones. Pour plus de détail, nous renvoyons la•e lecteur•rice vers la section 2.2.2 (p.43).

Dans la suite, nous omettons les indices de fréquentiel et temporel (k, ℓ), par souci de lisibilité. En utilisant les Eq. (D.3), (D.4) et (D.9), nous montrons que ζ est en réalité la somme des RSBs en sortie des beamformers MVDR orientés vers les sources de parole, notés $\xi_{\text{MVDR},i}$:

$$\begin{aligned}
\zeta &= \text{Tr}\left\{\mathbf{\Phi}_{\mathbf{n}}^{-1}\tilde{\mathbf{\Phi}}_{\mathbf{s}}\right\} \\
&= \text{Tr}\left\{\mathbf{\Phi}_{\mathbf{n}}^{-1}\mathbf{H}\mathbf{\Phi}_{\mathbf{s}}\mathbf{H}^H\right\} \\
&= \text{Tr}\left\{\mathbf{H}^H\mathbf{\Phi}_{\mathbf{n}}^{-1}\mathbf{H}\mathbf{\Phi}_{\mathbf{s}}\right\} \\
&= \sum_i \mathbf{h}_i^H\mathbf{\Phi}_{\mathbf{n}}^{-1}\mathbf{h}_i\phi_{s_i} \\
&= \sum_i \xi_{\text{MVDR},i}.
\end{aligned} \tag{D.10}$$

Maintenant, introduisons la reformulation de ζ en Eq. (D.10) ainsi que le modèle des matrices de covariance des sources de parole et du bruit exprimés en Eq. (D.3) et (D.4) dans l'expression de β en Eq. (D.7) :

$$\begin{aligned}
\beta &= \mathbf{x}^H\mathbf{\Phi}_{\mathbf{n}}^{-1}\tilde{\mathbf{\Phi}}_{\mathbf{s}}\mathbf{\Phi}_{\mathbf{n}}^{-1}\mathbf{x} \\
&= \mathbf{x}^H\mathbf{\Phi}_{\mathbf{n}}^{-1}\mathbf{H}\mathbf{\Phi}_{\mathbf{s}}\mathbf{H}^H\mathbf{\Phi}_{\mathbf{n}}^{-1}\mathbf{x} \\
&= \sum_i \phi_n^{-2}\phi_{s_i}|\mathbf{h}_i^H\mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{x}|^2 \\
&= \sum_i \phi_n^{-1}\xi_i|\mathbf{h}_i^H\mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{x}|^2,
\end{aligned} \tag{D.11}$$

où $\xi_i = \dfrac{\phi_{s_i}}{\phi_n}$ est le RSB *a priori* de la $i^{\text{ème}}$ source. En introduisant le facteur

$\left(\frac{\mathbf{h}_i^H \boldsymbol{\Gamma}_\mathbf{n}^{-1} \mathbf{h}_i}{\mathbf{h}_i^H \boldsymbol{\Gamma}_\mathbf{n}^{-1} \mathbf{h}_i}\right)^2$ à chaque terme de la somme, on peut écrire :

$$\begin{aligned}\beta &= \sum_i (\mathbf{h}_i^H \boldsymbol{\Gamma}_\mathbf{n}^{-1} \mathbf{h}_i)^2 \phi_n^{-1} \xi_i \frac{|\mathbf{h}_i^H \boldsymbol{\Gamma}_\mathbf{n}^{-1} \mathbf{x}|^2}{(\mathbf{h}_i^H \boldsymbol{\Gamma}_\mathbf{n}^{-1} \mathbf{h}_i)^2} \\ &= \sum_i \xi_{\text{MVDR},i} \frac{|\mathbf{w}_{\text{MVDR},i}^H \mathbf{x}|^2}{\phi_{n,\text{MVDR},i}} \\ &= \sum_i \xi_{\text{MVDR},i} \gamma_{\text{MVDR},i}, \end{aligned} \tag{D.12}$$

où $\gamma_{\text{MVDR},i}$ est le RSB *a posteriori* en sortie du beamformer MVDR orienté vers la $i^{\text{ème}}$ source, définit comme suit :

$$\gamma_{\text{MVDR},i} = \frac{|\mathbf{w}_{\text{MVDR},i}^H \mathbf{x}|^2}{\phi_{n,\text{MVDR},i}}, \tag{D.13}$$

avec $\phi_{n,\text{MVDR},i}$, la variance du bruit en sortie du beamformer MVDR visant la $i^{\text{ème}}$ source :

$$\phi_{n,\text{MVDR},i} = \frac{\phi_n}{\mathbf{h}_i^H \boldsymbol{\Gamma}_\mathbf{n}^{-1} \mathbf{h}_i}. \tag{D.14}$$

En introduisant l'Eq. (D.12) et l'Eq. (D.10) en Eq. (D.6), on obtient une nouvelle formulation de l'estimateur de PPPM :

$$p_{\text{PPPM}} = \left(1 + \frac{q}{1-q}\left(1 + \sum_j \xi_{\text{MVDR},j}\right) e^{-\sum_i \frac{\xi_{\text{MVDR},i}}{1+\sum_j \xi_{\text{MVDR},j}} \gamma_{\text{MVDR},i}}\right)^{-1}. \quad \blacksquare \tag{D.15}$$

Cette expression est plus simple que celle originellement fournie par [Souden et al., 2010] (voir Eq. (D.6)). Dans le cas où l'algorithme est intégré à un système employant déjà par ailleurs les Q beamformers MVDR, alors cette nouvelle expression de p_{PPPM} permet de réduire drastiquement le coût calculatoire.

Cas particulier, $Q = 1$

Dans le cas particulier où une seule source cible est présente, soit $Q = 1$, alors l'Eq. (D.15) se réduit et devient l'expression suivante :

$$p_{\text{PPPM}} = \left(1 + \frac{q}{1-q}\left(1 + \xi_{\text{MVDR}}\right) e^{-\frac{\xi_{\text{MVDR}}}{1+\xi_{\text{MVDR}}} \gamma_{\text{MVDR}}}\right)^{-1}. \quad \blacksquare \tag{D.16}$$

Il devient alors clair que sous cette forme, l'estimateur de PPPM est l'estimateur de PPP (voir l'Eq. (D.5)) qui prend en entrée le signal issu du beamformer MVDR.

Annexe E

Décomposition du beamformer LCMV en une somme de beamformers MVDR pour trois sources cibles

Dans cette annexe, on montre comment décomposer le filtre du beamformer LCMV en une somme de filtres de beamforming MVDR pour trois sources cibles. La démonstration pour deux sources cibles a été apportée par [Hadad et al., 2016], voir Eq. (2.46, p.47). A notre connaissance, la démonstration pour trois sources cibles est originale. Tout d'abord, on rappelle l'expression du filtre du beamformer LCMV pour $Q = 3$ sources cibles en employant les notations utilisées dans le chapitre 2 :

$$\mathbf{w}_{\text{LCMV}} = \mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{H}(\mathbf{H}^H\mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{H})^{-1}\mathbf{g}, \tag{E.1}$$

$$= \mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{H}(\mathbf{H}^H\mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{H})^{-1}\begin{bmatrix} g_1 \\ g_2 \\ g_3 \end{bmatrix}, \tag{E.2}$$

$$\begin{aligned} = \mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{H}(\mathbf{H}^H\mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{H})^{-1}\begin{bmatrix} g_1 \\ 0 \\ 0 \end{bmatrix} + \mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{H}(\mathbf{H}^H\mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{H})^{-1}\begin{bmatrix} 0 \\ g_2 \\ 0 \end{bmatrix} \\ + \mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{H}(\mathbf{H}^H\mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{H})^{-1}\begin{bmatrix} 0 \\ 0 \\ g_3 \end{bmatrix}, \end{aligned} \tag{E.3}$$

$$= \mathbf{w}_{\text{LCMV, 1}} + \mathbf{w}_{\text{LCMV, 2}} + \mathbf{w}_{\text{LCMV, 3}}, \tag{E.4}$$

où $\mathbf{w}_{\text{LCMV, q}}$ est le beamformer LCMV préservant la $q^{\text{ème}}$ source à un facteur g_q près et annulant les deux autres. Explicitons maintenant l'inversion de la

matrice $\mathbf{H}^H\mathbf{\Gamma}_\mathbf{n}^{-1}\mathbf{H}$:

$$\mathbf{D}^{-1} = (\mathbf{H}^H\mathbf{\Gamma}_\mathbf{n}^{-1}\mathbf{H})^{-1}, \tag{E.5}$$

$$= \begin{bmatrix} \delta_1 & \delta_{1,2} & \delta_{1,3} \\ \delta_{2,1} & \delta_2 & \delta_{2,3} \\ \delta_{3,1} & \delta_{3,2} & \delta_3 \end{bmatrix}^{-1}, \tag{E.6}$$

$$= \frac{1}{|\mathbf{D}|} \begin{bmatrix} \delta_2\delta_3 - |\delta_{2,3}|^2 & \delta_{1,3}\delta_{2,3}^* - \delta_{1,2}\delta_3 & \delta_{1,2}\delta_{2,3} - \delta_{1,3}\delta_2^* \\ \delta_{2,3}\delta_{1,3}^* - \delta_{1,2}^*\delta_3 & \delta_1\delta_3 - |\delta_{1,3}|^2 & \delta_{1,3}\delta_{1,2}^* - \delta_1\delta_{2,3} \\ \delta_{1,2}^*\delta_{2,3}^* - \delta_2\delta_{1,3}^* & \delta_{1,2}\delta_{1,3}^* - \delta_1\delta_{2,3}^* & \delta_1\delta_2 - |\delta_{1,2}|^2 \end{bmatrix}. \tag{E.7}$$

où $|\mathbf{D}|$ est le déterminant de $\mathbf{D}$:

$$|\mathbf{D}| = \delta_1\delta_2\delta_3 + \delta_{1,2}\delta_{2,3}\delta_{1,3}^* + \delta_{1,3}\delta_{1,2}^*\delta_{2,3}^* - \delta_2|\delta_{1,3}|^2 - \delta_1|\delta_{2,3}|^2 - \delta_3|\delta_{1,2}|^2, \tag{E.8}$$

avec $\delta_q = \mathbf{h}_q^H\mathbf{\Gamma}_\mathbf{n}^{-1}\mathbf{h}_q$ et $\delta_{i,j} = \mathbf{h}_i^H\mathbf{\Gamma}_\mathbf{n}^{-1}\mathbf{h}_j$. On peut alors développer les $\mathbf{w}_{\text{LCMV, q}}$ de la manière suivante :

$$\mathbf{w}_{\text{LCMV, 1}} = \mathbf{\Gamma}_\mathbf{n}^{-1}\mathbf{H}(\mathbf{H}^H\mathbf{\Gamma}_\mathbf{n}^{-1}\mathbf{H})^{-1} \begin{bmatrix} g_1 \\ 0 \\ 0 \end{bmatrix}, \tag{E.9}$$

$$= \frac{g_1}{|\mathbf{D}|} \begin{bmatrix} \mathbf{\Gamma}_\mathbf{n}^{-1}\mathbf{h}_1 & \mathbf{\Gamma}_\mathbf{n}^{-1}\mathbf{h}_2 & \mathbf{\Gamma}_\mathbf{n}^{-1}\mathbf{h}_3 \end{bmatrix} \begin{bmatrix} \delta_2\delta_3 - |\delta_{2,3}|^2 \\ \delta_{2,3}\delta_{1,3}^* - \delta_{1,2}^*\delta_3 \\ \delta_{1,2}^*\delta_{2,3}^* - \delta_2\delta_{1,3}^* \end{bmatrix}, \tag{E.10}$$

$$= \frac{g_1}{|\mathbf{D}|} \left((\delta_2\delta_3 - |\delta_{2,3}|^2)\mathbf{\Gamma}_\mathbf{n}^{-1}\mathbf{h}_1 + (\delta_{2,3}\delta_{1,3}^* - \delta_{1,2}^*\delta_3)\mathbf{\Gamma}_\mathbf{n}^{-1}\mathbf{h}_2 + (\delta_{1,2}^*\delta_{2,3}^* - \delta_2\delta_{1,3}^*)\mathbf{\Gamma}_\mathbf{n}^{-1}\mathbf{h}_3 \right). \tag{E.11}$$

et similairement :

$$\mathbf{w}_{\text{LCMV, 2}} = \frac{g_2}{|\mathbf{D}|} \left((\delta_{1,3}\delta_{2,3}^* - \delta_{1,2}\delta_3)\mathbf{\Gamma}_\mathbf{n}^{-1}\mathbf{h}_1 + (\delta_1\delta_3 - |\delta_{1,3}|^2)\mathbf{\Gamma}_\mathbf{n}^{-1}\mathbf{h}_2 + (\delta_{1,2}\delta_{1,3}^* - \delta_1\delta_{2,3}^*)\mathbf{\Gamma}_\mathbf{n}^{-1}\mathbf{h}_3 \right), \tag{E.12}$$

$$\mathbf{w}_{\text{LCMV, 3}} = \frac{g_3}{|\mathbf{D}|} \left((\delta_{1,2}\delta_{2,3} - \delta_{1,3}\delta_2^*)\mathbf{\Gamma}_\mathbf{n}^{-1}\mathbf{h}_1 + (\delta_{1,3}\delta_{1,2}^* - \delta_1\delta_{2,3})\mathbf{\Gamma}_\mathbf{n}^{-1}\mathbf{h}_2 + (\delta_1\delta_2 - |\delta_{1,2}|^2)\mathbf{\Gamma}_\mathbf{n}^{-1}\mathbf{h}_3 \right). \tag{E.13}$$

Enfin, en introduisant l'Eq. (E.8) en Eq. (E.11), (E.12) et (E.13) et réorganisant les termes, on peut écrire l'Eq. (E.4) comme une somme de $\mathbf{w}_{\text{MVDR, q}}$ à un

facteur près :

$$\begin{aligned}\mathbf{w}_{\mathrm{LCMV}} = \mathbf{w}_{\mathrm{MVDR,\,1}} &\left(\frac{g_1}{1+\eta_1} + \delta_1 \frac{g_2(\delta_{1,3}\delta_{2,3}^* - \delta_{1,2}\delta_3) + g_3(\delta_{1,2}\delta_{2,3} - \delta_{1,3}\delta_2^*)}{|\mathbf{D}|}\right) \\ + \mathbf{w}_{\mathrm{MVDR,\,2}} &\left(\frac{g_2}{1+\eta_2} + \delta_2 \frac{g_1(\delta_{2,3}\delta_{1,3}^* - \delta_{1,2}^*\delta_3) + g_3(\delta_{1,3}\delta_{1,2}^* - \delta_1\delta_{2,3})}{|\mathbf{D}|}\right) \\ + \mathbf{w}_{\mathrm{MVDR,\,3}} &\left(\frac{g_3}{1+\eta_3} + \delta_3 \frac{g_1(\delta_{1,2}^*\delta_{2,3}^* - \delta_2\delta_{1,3}^*) + g_2(\delta_{1,2}\delta_{1,3}^* - \delta_1\delta_{2,3}^*)}{|\mathbf{D}|}\right), \end{aligned} \quad \blacksquare \tag{E.14}$$

où $\mathbf{w}_{\mathrm{MVDR,\,q}} = \mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{h}_q/(\mathbf{h}_q^H\mathbf{\Gamma}_{\mathbf{n}}^{-1}\mathbf{h}_q)$ est le filtre du beamformer MVDR dirigé vers la $q^{\text{ème}}$ source et :

$$\eta_1 = \frac{\delta_{1,2}\delta_{2,3}\delta_{1,3}^* + \delta_{1,3}\delta_{1,2}^*\delta_{2,3}^* - \delta_2|\delta_{1,3}|^2 - \delta_3|\delta_{1,2}|^2}{\delta_1\delta_2\delta_3 - \delta_1|\delta_{2,3}|^2}, \tag{E.15}$$

$$\eta_2 = \frac{\delta_{1,2}\delta_{2,3}\delta_{1,3}^* + \delta_{1,3}\delta_{1,2}^*\delta_{2,3}^* - \delta_1|\delta_{2,3}|^2 - \delta_3|\delta_{1,2}|^2}{\delta_1\delta_2\delta_3 - \delta_2|\delta_{1,3}|^2}, \tag{E.16}$$

$$\eta_3 = \frac{\delta_{1,2}\delta_{2,3}\delta_{1,3}^* + \delta_{1,3}\delta_{1,2}^*\delta_{2,3}^* - \delta_2|\delta_{1,3}|^2 - \delta_1|\delta_{2,3}|^2}{\delta_1\delta_2\delta_3 - \delta_3|\delta_{1,2}|^2}. \tag{E.17}$$

Bibliographie

[Aaronson and Hartmann, 2014] Aaronson, N. L. and Hartmann, W. M. (2014). Testing, correcting, and extending the Woodworth model for interaural time difference. *The Journal of the Acoustical Society of America*, 135(2) :817–823.

[Akeroyd, 2014] Akeroyd, M. A. (2014). An Overview of the Major Phenomena of the Localization of Sound Sources by Normal-Hearing, Hearing-Impaired, and Aided Listeners. *Trends in Hearing*, 18 :1–7.

[Alexander and Rallapalli, 2017] Alexander, J. M. and Rallapalli, V. (2017). Acoustic and perceptual effects of amplitude and frequency compression on high-frequency speech. *The Journal of the Acoustical Society of America*, 142(2) :908–923.

[Algazi et al., 2001a] Algazi, V., Duda, R., Thompson, D., and Avendano, C. (2001a). The CIPIC HRTF database. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, pages 99–102, New Platz, NY, USA. IEEE.

[Algazi et al., 2001b] Algazi, V. R., Avendano, C., and Duda, R. O. (2001b). Elevation localization and head-related transfer function analysis at low frequencies. *The Journal of the Acoustical Society of America*, 109(3) :1110–1122.

[Algazi et al., 2002] Algazi, V. R., Duda, R. O., Duraiswami, R., Gumerov, N. A., and Tang, Z. (2002). Approximating the head-related transfer function using simple geometric models of the head and torso. *The Journal of the Acoustical Society of America*, 112(5) :2053–2064.

[Allen et al., 2008] Allen, K., Carlile, S., and Alais, D. (2008). Contributions of talker characteristics and spatial location to auditory streaming. *The Journal of the Acoustical Society of America*, 123(3) :1562–1570.

[Andersen et al., 2018] Andersen, A. H., de Haan, J. M., Tan, Z.-H., and Jensen, J. (2018). Refinement and validation of the binaural short time objective intelligibility measure for spatially diverse conditions. *Speech Communication*, 102 :1–13.

[Anderson et al., 2009] Anderson, M. C., Arehart, K. H., and Kates, J. M. (2009). The Acoustic and Peceptual Effects of Series and Parallel Processing. *EURASIP Journal on Advances in Signal Processing*, 2009(1) :1–20.

[Andreopoulou and Katz, 2017] Andreopoulou, A. and Katz, B. F. G. (2017). Identification of perceptually relevant methods of inter-aural time difference estimation. *The Journal of the Acoustical Society of America*, 142(2) :588–598.

[ANSI, 1997] ANSI (1997). ANSI S3.5 : SII—Speech intelligibility index standard.

[ANSI, 2003] ANSI (2003). ANSI S3.22 : Specification of Hearing Aid Characteristics.

[Arbogast et al., 2005] Arbogast, T. L., Mason, C. R., and Kidd, G. (2005). The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 117(4) :2169–2180.

[Archer-Boyd et al., 2015] Archer-Boyd, A. W., Whitmer, W. M., Brimijoin, W. O., and Soraghan, J. J. (2015). Biomimetic direction of arrival estimation for resolving front-back confusions in hearing aids. *The Journal of the Acoustical Society of America*, 137(5) :EL360–EL366.

[Aroudi and Doclo, 2019] Aroudi, A. and Doclo, S. (2019). Cognitive-driven Binaural LCMV Beamformer Using EEG-based Auditory Attention Decoding. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 406–410, Brighton, United Kingdom. IEEE.

[Arweiler and Buchholz, 2011] Arweiler, I. and Buchholz, J. M. (2011). The influence of spectral characteristics of early reflections on speech intelligibility. *The Journal of the Acoustical Society of America*, 130(2) :996–1005.

[Avargel and Cohen, 2007] Avargel, Y. and Cohen, I. (2007). On Multiplicative Transfer Function Approximation in the Short-Time Fourier Transform Domain. *IEEE Signal Processing Letters*, 14(5) :337–340.

[Bagheri and Giacobello, 2019] Bagheri, S. and Giacobello, D. (2019). Exploiting Multi-Channel Speech Presence Probability in Parametric Multi-Channel Wiener Filter. In *Interspeech 2019*, pages 101–105. ISCA.

[Bahu, 2016] Bahu, H. (2016). *Localisation auditive en contexte de synthèse binaurale non-individuelle*. PhD thesis, Université Pierre et Marie Curie - Paris VI, Paris, France.

[Batteau, 1967] Batteau, D. W. (1967). The Role of the Pinna in Human Localization. *Proc. R. Soc. London. Series B, Biological Sciences*, 168 :158–180.

[Baumgartner et al., 2014] Baumgartner, R., Majdak, P., and Laback, B. (2014). Modeling sound-source localization in sagittal planes for human listeners. *The Journal of the Acoustical Society of America*, 136(2) :791–802.

[Baumgartner et al., 2016] Baumgartner, R., Majdak, P., and Laback, B. (2016). Modeling the Effects of Sensorineural Hearing Loss on Sound Localization in the Median Plane. *Trends in Hearing*, 20 :233121651666200.

[Benesty et al., 2016] Benesty, J., Chen, J., and Pan, C. (2016). *Fundamentals of Differential Beamforming*. SpringerBriefs in Electrical and Computer Engineering. Springer Singapore, Singapore.

[Best et al., 2010] Best, V., Kalluri, S., McLachlan, S., Valentine, S., Edwards, B., and Carlile, S. (2010). A comparison of CIC and BTE hearing aids for three-dimensional localization of speech. *International Journal of Audiology*, 49(10) :723–732.

[Bisgaard et al., 2010] Bisgaard, N., Vlaming, M. S. M. G., and Dahlquist, M. (2010). Standard Audiograms for the IEC 60118-15 Measurement Procedure. *Trends in Amplification*, 14(2) :113–120.

[Blauert, 1969] Blauert, J. (1969). Sound Localization in the Median Plane. *Acta Acustica United with Acustica*, 22 :205–213.

[Blauert, 2013] Blauert, J., editor (2013). *The Technology of Binaural Listening*. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Boyd et al., 2013] Boyd, A. W., Whitmer, W. M., Brimijoin, W. O., and Akeroyd, M. A. (2013). Improved estimation of direction of arrival of sound sources for hearing aids using gyroscopic information. In *Proceedings of Meetings on Acoustics*, volume 19, pages 1–9, Montréal. Acoustical Society of America.

[Boyd et al., 2012] Boyd, A. W., Whitmer, W. M., Soraghan, J. J., and Akeroyd, M. A. (2012). Auditory externalization in hearing-impaired listeners : The effect of pinna cues and number of talkers. *The Journal of the Acoustical Society of America*, 131(3) :EL268–EL274.

[Brand and Kollmeier, 2002] Brand, T. and Kollmeier, B. (2002). Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *The Journal of the Acoustical Society of America*, 111(6) :2801–2810.

[Brandstein et al., 2001] Brandstein, M., Ward, D., Lacroix, A., and Venetsanopoulos, A., editors (2001). *Microphone Arrays*. Digital Signal Processing. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Braun et al., 2015] Braun, S., Zhou, W., and Habets, E. A. P. (2015). Narrowband direction-of-arrival estimation for binaural hearing aids using relative

transfer functions. In *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5, New Paltz, NY, USA. IEEE.

[Brimijoin et al., 2014] Brimijoin, W. O., Whitmer, W. M., McShefferty, D., and Akeroyd, M. A. (2014). The Effect of Hearing Aid Microphone Mode on Performance in an Auditory Orienting Task. *Ear and Hearing*, 35 :204–212.

[Bronkhorst and Plomp, 1989] Bronkhorst, A. H. and Plomp, R. (1989). Binaural speech intelligibility in noise for hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 86(4) :1374–1383.

[Bronkhorst, 2015] Bronkhorst, A. W. (2015). The cocktail-party problem revisited : early processing and selection of multi-talker speech. *Atten Percept Psychophys*, 77 :1465–1487.

[Bronkhorst and Plomp, 1988] Bronkhorst, A. W. and Plomp, R. (1988). The effect of head-induced interaural time and level differences on speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 83(4) :1508–1516.

[Brons et al., 2015] Brons, I., Houben, R., and Dreschler, W. A. (2015). Acoustical and Perceptual Comparison of Noise Reduction and Compression in Hearing Aids. *Journal of Speech, Language, and Hearing Research*, 58(4) :1363–1376.

[Brooks and Reed, 1972] Brooks, L. and Reed, I. (1972). Equivalence of the Likelihood Ratio Processor, the Maximum Signal-to-Noise Ratio Filter, and the Wiener Filter. *IEEE Transactions on Aerospace and Electronic Systems*, AES-8(5) :690–692.

[Brungart and Iyer, 2012] Brungart, D. S. and Iyer, N. (2012). Better-ear glimpsing efficiency with symmetrically-placed interfering talkers. *The Journal of the Acoustical Society of America*, 132(4) :2545–2556.

[Chen et al., 2021] Chen, Y., Wong, L. L. N., Kuehnel, V., Qian, J., Voss, S. C., and Shangqiguo, W. (2021). Can Dual Compression Offer Better Mandarin Speech Intelligibility and Sound Quality Than Fast-Acting Compression ? *Trends in Hearing*, 25 :13.

[Cherry, 1953] Cherry, E. C. (1953). Some Experiments on the Recogntition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, 25(5) :975–979.

[Chong and Jenstad, 2018] Chong, F. Y. and Jenstad, L. M. (2018). A critical review of hearing-aid single-microphone noise-reduction studies in adults and children. *Disability and Rehabilitation : Assistive Technology*, 13(6) :600–608.

[Chung, 2007] Chung, K. (2007). Effective compression and noise reduction configurations for hearing protectors. *The Journal of the Acoustical Society of America*, 121(2) :1090–1101.

[Cohen, 2002] Cohen, I. (2002). Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. *IEEE Signal Processing Letters*, 9(4) :113–116.

[Cohen, 2003] Cohen, I. (2003). Noise spectrum estimation in adverse environments : improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing*, 11(5) :466–475.

[Corey, 2019] Corey, R. M. (2019). *Microphone Array Processing for Augmented Listening*. PhD thesis, University of Illinois, Urbana-Champaign.

[Corey and Singer, 2017] Corey, R. M. and Singer, A. C. (2017). Dynamic range compression for noisy mixtures using source separation and beamforming. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 289–293, New Paltz, NY. IEEE.

[Cornelis et al., 2010] Cornelis, B., Doclo, S., Van dan Bogaert, T., Moonen, M., and Wouters, J. (2010). Theoretical Analysis of Binaural Multimicrophone Noise Reduction Techniques. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2) :342–355.

[Cox, 1973] Cox, H. (1973). Resolving power and sensitivity to mismatch of optimum array processors. *The Journal of the Acoustical Society of America*, 54(3) :771–785.

[Cox et al., 1988] Cox, R. M., Matesich, J. S., and Moore, J. N. (1988). Distribution of short-term rms levels in conversational speech. *The Journal of the Acoustical Society of America*, 84(3) :1100–1104.

[Crochiere, 1980] Crochiere, R. (1980). A weighted overlap-add method of short-time Fourier analysis/Synthesis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(1) :99–102.

[Denk et al., 2018] Denk, F., Ernst, S. M. A., Ewert, S. D., and Kollmeier, B. (2018). Adapting Hearing Devices to the Individual Ear Acoustics : Database and Target Response Correction Functions for Various Device Styles. *Trends in Hearing*, 22 :1–19.

[Denk et al., 2019] Denk, F., Ewert, S. D., and Kollmeier, B. (2019). On the limitations of sound localization with hearing devices. *The Journal of the Acoustical Society of America*, 146(3) :1732–1744.

[Dieudonné and Francart, 2018] Dieudonné, B. and Francart, T. (2018). Head shadow enhancement with low-frequency beamforming improves sound localization and speech perception for simulated bimodal listeners. *Hearing Research*.

[Doclo et al., 2006] Doclo, S., Klasen, T. J., Van den Bogaert, T., Wouters, J., and Moonen, M. (2006). Theorical analysis of binaural cue preservation using multi-channel Wiener filtering and interaural transfer functions. In *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, page 4.

[Doclo and Moonen, 2002] Doclo, S. and Moonen, M. (2002). GSVD-based optimal filtering for single and multimicrophone speech enhancement. *IEEE Transactions on Signal Processing*, 50(9) :2230–2244.

[Drennan et al., 2005] Drennan, W. R., Gatehouse, S., Howell, P., Tasell, D. V., and Lund, S. (2005). Localization and Speech-Identification Ability of Hearing-Impaired Listeners Using Phase-Preserving Amplification. *Ear and Hearing*, 26(5) :461–472.

[Duda and Martens, 1998] Duda, R. O. and Martens, W. L. (1998). Range dependence of the response of a spherical head model. *The Journal of the Acoustical Society of America*, 104(5) :3048–3058.

[Durlach, 1963] Durlach, N. I. (1963). Equalization and Cancellation Theory of Binaural Masking-Level Differences. *The Journal of the Acoustical Society of America*, 35(8) :1206–1218.

[Durlach et al., 1992] Durlach, N. I., Rigopulos, A., Pang, X. D., Woods, W. S., Kulkarni, A., Colburn, H. S., and Wenzel, E. M. (1992). On the Externalization of Auditory Images. *Presence : Teleoperators and Virtual Environments*, 1(2) :251–257.

[Edmonds and Culling, 2005] Edmonds, B. A. and Culling, J. F. (2005). The spatial unmasking of speech : evidence for within-channel processing of interaural time delay. *The Journal of the Acoustical Society of America*, 117(5) :3069–3078.

[Edmonds and Culling, 2006] Edmonds, B. A. and Culling, J. F. (2006). The spatial unmasking of speech : Evidence for better-ear listening. *The Journal of the Acoustical Society of America*, 120(3) :1539–1545.

[Eggermont, 1977] Eggermont, J. J. (1977). Electrocochleography and Recruitment. *Annals of Otology, Rhinology & Laryngology*, 86(2) :12.

[Ehrenberg et al., 2010] Ehrenberg, L., Gannot, S., Leshem, A., and Zehavi, E. (2010). Sensitivity analysis of MVDR and MPDR beamformers. In *2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel*, pages 416–420, Eilat, Israel. IEEE.

[Ehrgott, 2005] Ehrgott, M. (2005). *Multicriteria optimization*. Springer, Berlin Heidelberg, 2. ed edition.

[Elko and Anh-Tho Nguyen Pong, 1995] Elko, G. and Anh-Tho Nguyen Pong (1995). A simple adaptive first-order differential microphone. In *Procee-*

dings of 1995 Workshop on Applications of Signal Processing to Audio and Accoustics, pages 169–172, New Paltz, NY, USA. IEEE.

[Ellinger et al., 2017] Ellinger, R. L., Jakien, K. M., and Gallun, F. J. (2017). The role of interaural differences on speech intelligibility in complex multi-talker environmentsa). *The Journal of the Acoustical Society of America*, 141(2) :EL170–EL176.

[Ephraim and Malah, 1984] Ephraim, Y. and Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6) :1109–1121.

[Ephraim and Malah, 1985] Ephraim, Y. and Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2) :443–445.

[Ernst et al., 2018] Ernst, S. M. A., Kortlang, S., Grimm, G., Bisitz, T., Kollmeier, B., and Ewert, S. D. (2018). Binaural model-based dynamic-range compression. *International Journal of Audiology*, pages 1–12.

[Favre-Félix et al., 2018] Favre-Félix, A., Graversen, C., Hietkamp, R. K., Dau, T., and Lunner, T. (2018). Improving Speech Intelligibility by Hearing Aid Eye-Gaze Steering : Conditions With Head Fixated in a Multitalker Environment. *Trends in Hearing*, 22.

[Füllgrabe, 2015] Füllgrabe, C. (2015). Age-group differences in speech identification despite matched audiometrically normal hearing : contributions from auditory temporal processing and cognition. *Frontiers in Aging Neuroscience*, 6 :25.

[Fontaine et al., 2017] Fontaine, M., Liutkus, A., Girin, L., and Badeau, R. (2017). Explaining the parameterized wiener filter with alpha-stable processes. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 51–55, New Paltz, NY. IEEE.

[French and Steinberg, 1947] French, N. R. and Steinberg, J. C. (1947). Factors Governing the Intelligibility of Speech Sounds. *The Journal of the Acoustical Society of America*, 90(19) :31.

[Gannot et al., 2017] Gannot, S., Vincent, E., Markovich-Golan, S., and Ozerov, A. (2017). A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(4) :692–730.

[Gardner, 1968] Gardner, M. B. (1968). Historical Background of the Haas and/or Precedence Effect. *The Journal of the Acoustical Society of America*, 43(6) :1243–1248.

[Gates and Mills, 2005] Gates, G. A. and Mills, J. H. (2005). Presbycusis. *The Lancet*, 366(9491) :1111–1120.

[Gay and Benesty, 2000] Gay, S. L. and Benesty, J., editors (2000). *Acoustic Signal Processing for Telecommunication*. Springer US, Boston, MA.

[Gerlach et al., 2021] Gerlach, L., Payá-Vayá, G., and Blume, H. (2021). A Survey on Application Specific Processor Architectures for Digital Hearing Aids. *Journal of Signal Processing Systems*, pages 1–16.

[Giannoulis et al., 2012] Giannoulis, D., Massberg, M., and Reiss, J. D. (2012). Digital dynamic range compressor design—A tutorial and analysis. *Journal of the Audio Engineering Society*, 60(6) :399–408.

[Glasberg and Moore, 1990] Glasberg, B. R. and Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, 47(1-2) :103–138.

[Glista et al., 2009] Glista, D., Scollie, S., Bagatto, M., Seewald, R., Parsa, V., and Johnson, A. (2009). Evaluation of nonlinear frequency compression : Clinical outcomes. *International Journal of Audiology*, 48(9) :632–644.

[Glyde et al., 2013a] Glyde, H., Buchholz, J. M., Dillon, H., Cameron, S., and Hickson, L. (2013a). The importance of interaural time differences and level differences in spatial release from masking. *The Journal of the Acoustical Society of America*, 134(2) :EL147–EL152.

[Glyde et al., 2013b] Glyde, H., Cameron, S., Dillon, H., Hickson, L., and Seeto, M. (2013b). The effects of hearing impairment and aging on spatial processing. *Ear and hearing*, 34(1) :15–28.

[Goetze et al., 2007] Goetze, S., Rohdenburg, T., Hohmann, V., Kollmeier, B., and Kammeyer, K.-D. (2007). Direction of arrival estimation based on the dual delay line approach for binaural hearing aid microphone arrays. In *2007 International Symposium on Intelligent Signal Processing and Communication Systems*, pages 84–87, Xiamen, China. IEEE.

[Goldsworthy and Greenberg, 2004] Goldsworthy, R. L. and Greenberg, J. E. (2004). Analysis of speech-based speech transmission index methods with implications for nonlinear operations. *The Journal of the Acoustical Society of America*, 116(6) :3679–3689.

[Grimault et al., 2018] Grimault, N., Irino, T., Dimachki, S., Corneyllie, A., Patterson, R. D., and Garcia, S. (2018). A Real Time Hearing Loss Simulator. *Acta Acustica united with Acustica*, 104(5) :904–908.

[Gössling et al., 2017] Gössling, N., Marquardt, D., and Doclo, S. (2017). Performance analysis of the extended binaural MVDR beamformer with partial noise estimation in a homogeneous noise field. In *Hands-free Speech Communications and Microphone Arrays (HSCMA), 2017*, pages 1–5. IEEE.

[Guo et al., 2014] Guo, X., Xu, B., Rao, Z., Wan, Q., Feng, Z., and Shen, Y. (2014). Low-Complexity Iterative Adaptive Linearly Constrained Minimum Variance Beamformer. *Circuits, Systems, and Signal Processing*, 33(3) :987–997.

[Habets and Naylor, 2010] Habets, E. A. and Naylor, P. A. (2010). An online quasi-Newton algorithm for blind SIMO identification. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2662–2665, Dallas, TX. IEEE.

[Hadad et al., 2016] Hadad, E., Doclo, S., and Gannot, S. (2016). The Binaural LCMV Beamformer and its Performance Analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3) :543–558.

[Hadad et al., 2012] Hadad, E., Gannot, S., and Doclo, S. (2012). Binaural Linearly Constrained Minimum Variance Beamformer for Hearing Aid Applications. In *International Workshop on Acoustic Signal Enhancemoment (IWAENC)*, page 4, Aachen.

[Hagerman, 1982] Hagerman, B. (1982). Sentences for Testing Speech-Intelligibility in Noise. *Scandinavian Audiology*, 11.

[Hagerman and Olofsson, 2004] Hagerman, B. and Olofsson, k. (2004). A Method to Measure the Effect of Noise Reduction Algorithms Using Simultaneous Speech and Noise. *Acta Acustica united with Acustica*, 90(2) :356–361.

[Hamacher et al., 2005] Hamacher, V., Chalupper, J., Eggers, J., Fischer, E., Kornagel, U., Puder, H., and Rass, U. (2005). Signal processing in high-end hearing aids : state of the art, challenges, and future trends. *EURASIP Journal on Applied Signal Processing*, 2005 :2915–2929.

[Harder, 2015] Harder, S. (2015). *Individualized directional microphone optimization in hearing aids based on reconstructing the 3D geometry of the head and ear from 2D images*. PhD thesis, Technical University of Denmark (DTU), Kongens Lyngby. OCLC : 931880689.

[Hartmann et al., 2005] Hartmann, W. M., Rakerd, B., and Koller, A. (2005). Binaural Coherence in Rooms. *Acta Acustica united with Acustica*, 91 :12.

[Hartmann and Wittenberg, 1996] Hartmann, W. M. and Wittenberg, A. (1996). On the Externalization of Sound Images. *The Journal of the Acoustical Society of America*, 99(6) :3678–3688.

[Hassager et al., 2016] Hassager, H. G., Gran, F., and Dau, T. (2016). The role of spectral detail in the binaural transfer function on perceived externalization in a reverberant environment. *The Journal of the Acoustical Society of America*, 139(5) :2992–3000.

[Hassager et al., 2017a] Hassager, H. G., May, T., Wiinberg, A., and Dau, T. (2017a). Preserving spatial perception in rooms using direct-sound driven

dynamic range compression. *The Journal of the Acoustical Society of America*, 141(6) :4556–4566.

[Hassager et al., 2017b] Hassager, H. G., Wiinberg, A., and Dau, T. (2017b). Effects of hearing-aid dynamic range compression on spatial perception in a reverberant environment. *The Journal of the Acoustical Society of America*, 141(4) :2556–2568.

[Haykin and Liu, 2009] Haykin, S. S. and Liu, K. J. R. (2009). *Handbook on array processing and sensor networks*. Wiley, Hoboken, NJ. OCLC : 845468569.

[Hendrickx et al., 2017] Hendrickx, E., Stitt, P., Messonnier, J.-C., Lyzwa, J.-M., Katz, B. F., and de Boishéraud, C. (2017). Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis. *The Journal of the Acoustical Society of America*, 141(3) :2011–2023.

[Hofman et al., 1998] Hofman, P. M., Van Riswick, J. G., and Van Opstal, A. J. (1998). Relearning sound localization with new ears. *Nature neuroscience*, 1(5) :417–421.

[Hu et al., 2011] Hu, H., Sang, J., Lutman, M. E., and Bleeck, S. (2011). Simulation of hearing loss using compressive gammachirp auditory filters. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5428–5431, Prague. IEEE.

[Häusler et al., 1983] Häusler, R., Colburn, S., and Marr, E. (1983). Sound Localization in Subjects with Impaired Hearing. *Acta Oto-laryngologica*, 400.

[Ibrahim et al., 2012] Ibrahim, I., Parsa, V., Macpherson, E., and Cheesman, M. (2012). Evaluation of speech intelligibility and sound localization abilities with hearing aids using binaural wireless technology. *Audiology Research*, 3(1) :10.

[International Telecommunications Union-Recommendation and BS.1534-1., 2003] International Telecommunications Union-Recommendation and BS.1534-1. (2003). Method for the subjective assessment of intermediate quality level of coding systems.

[Itturriet and Costa, 2019] Itturriet, F. P. and Costa, M. H. (2019). Perceptually Relevant Preservation of Interaural Time Differences in Binaural Hearing Aids. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(4) :753–764.

[Jansen et al., 2012] Jansen, S., Luts, H., Wagener, K. C., Kollmeier, B., Del Rio, M., Dauman, R., James, C., Fraysse, B., Vormès, E., Frachet, B., Wouters, J., and van Wieringen, A. (2012). Comparison of three types of French speech-in-noise tests : A multi-center study. *International Journal of Audiology*, 51(3) :164–173.

[Jensen et al., 2020] Jensen, J., Karimian-Azari, S., Christensen, M., and Benesty, J. (2020). Harmonic beamformers for speech enhancement and dereverberation in the time domain. *Speech Communication*, 116 :1–11.

[Jerlvall and Lindblad, 1978] Jerlvall, L. and Lindblad, A. (1978). The influence of attack time and release time on speech intelligibility. A study of the effects of AGC on normal hearing and hearing impaired subjects. *Scandinavian audiology. Supplementum*, 6 :341–353.

[Jia et al., 2018] Jia, M., Sun, J., and Zheng, X. (2018). Multiple Speech Source Separation Using Inter-Channel Correlation and Relaxed Sparsity. *Applied Sciences*, 8(123) :23.

[Kates, 2005] Kates, J. M. (2005). Principles of Digital Dynamic-Range Compression. *Trends in Amplification*, 9(2) :45–76.

[Katz and Noisternig, 2014] Katz, B. F. and Noisternig, M. (2014). A comparative study of interaural time delay estimation methods. *The Journal of the Acoustical Society of America*, 135(6) :3530–3540.

[Kayser et al., 2009] Kayser, H., Ewert, S. D., Anemüller, J., Rohdenburg, T., Hohmann, V., and Kollmeier, B. (2009). Database of Multichannel In-Ear and Behind-the-Ear Head-Related and Binaural Room Impulse Responses. *EURASIP Journal on Advances in Signal Processing*, 2009(1) :10.

[Keating and King, 2013] Keating, P. and King, A. J. (2013). Developmental plasticity of spatial hearing following asymmetric hearing loss : context-dependent cue integration and its clinical implications. *Frontiers in Systems Neuroscience*, 7.

[Keidser et al., 2009] Keidser, G., O'Brien, A., Hain, J.-U., McLelland, M., and Yeend, I. (2009). The effect of frequency-dependent microphone directionality on horizontal localization performance in hearing-aid users. *International Journal of Audiology*, 48(11) :789–803.

[Keidser et al., 2006] Keidser, G., Rohrseitz, K., Dillon, H., Hamacher, V., Carter, L., Rass, U., and Convery, E. (2006). The effect of multi-channel wide dynamic range compression, noise reduction, and the directional microphone on horizontal localization performance in hearing aid wearers. *International Journal of Audiology*, 45(10) :563–579.

[Klasen et al., 2006] Klasen, T., Doclo, S., Van den Bogaert, T., Moonen, M., and Wouters, J. (2006). Binaural Multi-Channel Wiener Filtering for Hearing Aids : Preserving Interaural Time and Level Differences. In *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, volume 5, pages V–145–V–148, Toulouse, France. IEEE.

[Klockgether and van de Par, 2016] Klockgether, S. and van de Par, S. (2016). Just noticeable differences of spatial cues in echoic and anechoic acous-

tical environments. *The Journal of the Acoustical Society of America*, 140(4) :EL352–EL357.

[Knorr-Cetina, 2003] Knorr-Cetina, K. (2003). *Epistemic cultures : how the sciences make knowledge*. Harvard Univ. Press, Cambridge, Mass., 3. print edition. OCLC : 254506278.

[Korhonen et al., 2015] Korhonen, P., Lau, C., Kuk, F., Keenan, D., and Schumacher, J. (2015). Effects of Coordinated Compression and Pinna Compensation Features on Horizontal Localization Performance in Hearing Aid Users. *Journal of the American Academy of Audiology*, 26(1) :80–92.

[Kortlang et al., 2017] Kortlang, S., Chen, Z., Gerkmann, T., Kollmeier, B., Hohmann, V., and Ewert, S. D. (2017). Evaluation of combined dynamic compression and single channel noise reduction for hearing aid applications. *International Journal of Audiology*, 57(sup3) :S43–S54.

[Kortlang et al., 2016] Kortlang, S., Grimm, G., Hohmann, V., Kollmeier, B., and Ewert, S. D. (2016). Auditory Model-Based Dynamic Compression Controlled by Subband Instantaneous Frequency and Speech Presence Probability Estimates. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10) :1759–1772.

[Koutrouvelis et al., 2016] Koutrouvelis, A. I., Hendriks, R. C., Jensen, J., and Heusdens, R. (2016). Improved multi-microphone noise reduction preserving binaural cues. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 460–464, Shanghai. IEEE.

[Kowalewski et al., 2020] Kowalewski, B., Dau, T., and May, T. (2020). Perceptual Evaluation of Signal-to-Noise-Ratio-Aware Dynamic Range Compression in Hearing Aids. *Trends in Hearing*, 24 :14.

[Kuhn, 1977] Kuhn, G. F. (1977). Model for the interaural time differences in the azimuthal plane. *The Journal of the Acoustical Society of America*, 62(1) :157–167.

[Kuk and Ludvigsen, 2003] Kuk, F. and Ludvigsen, C. (2003). Reconsidering the Concept of the Aided Threshold for Nonlinear Hearing Aids. *Trends in Amplification*, 7(3) :77–97.

[Kuk, 1996] Kuk, F. K. (1996). Theoretical and Practical Considerations in Compression Hearing Aids. *Trends in Amplification*, 1(1) :5–39.

[Langendijk and Bronkhorst, 2002] Langendijk, E. H. A. and Bronkhorst, A. W. (2002). Contribution of spectral cues to human sound localization. *The Journal of the Acoustical Society of America*, 112(4) :1583–1596.

[Lavandier et al., 2012] Lavandier, M., Jelfs, S., Culling, J. F., Watkins, A. J., Raimond, A. P., and Makin, S. J. (2012). Binaural prediction of speech

intelligibility in reverberant rooms with multiple noise sources. *The Journal of the Acoustical Society of America*, 131(1) :218–231.

[Leglaive et al., 2019] Leglaive, S., Simsekli, U., Liutkus, A., Girin, L., and Horaud, R. (2019). Speech enhancement with variational autoencoders and alpha-stable distributions. In *ICASSP 2019 - IEEE International Conference on Acoustics Speech and Signal Processing*, page 6, Brighton, United Kingdom. IEEE.

[Levitt, 1971] Levitt, H. (1971). Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, 49(2B) :467–477.

[Llave and Leglaive, 2021a] Llave, A. and Leglaive, S. (2021a). Joint denoising and dynamic range compression in binaural hearing aids. *Soumis à The Journal of the Acoustical Society of America*, page 12.

[Llave and Leglaive, 2021b] Llave, A. and Leglaive, S. (2021b). On the Speech Sparsity for Computational Efficiency and Noise Reduction in Hearing Aids. In *13th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, page 5, Tokyo, Japan. IEEE.

[Llave et al., 2020] Llave, A., Leglaive, S., and Seguier, R. (2020). Localization Cues Preservation in Hearing Aids by Combining Noise Reduction and Dynamic Range Compression. In *12th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, page 9, Auckland, New Zealand. IEEE.

[Llave and Séguier, 2019] Llave, A. and Séguier, R. (2019). Influence sur les indices de localisation du beamforming et de la compression de dynamique dans les audioprothèses. In *XXVIIème colloque GRETSI (GRETSI 2019)*, Lille, France.

[Loizou, 2013] Loizou, P. C. (2013). *Speech Enhancement*. CRC Press, 2nd edition.

[Lotter and Vary, 2006] Lotter, T. and Vary, P. (2006). Dual-Channel Speech Enhancement by Superdirective Beamforming. *EURASIP Journal on Advances in Signal Processing*, 2006(1) :14.

[Ludvigsen et al., 1993] Ludvigsen, C., Elberling, C., and Keidser, G. (1993). Evaluation of a noise reduction method : Comparison between observed scores and scores predicted from STI. *Scandinavian Audiology*, 38.

[Luo et al., 2002] Luo, F.-L., Yang, J., Pavlovic, C., and Nehorai, A. (2002). Adaptive null-forming scheme in digital hearing aids. *IEEE Transactions on signal processing*, 50(7) :1583–1590.

[Macpherson and Middlebrooks, 2002] Macpherson, E. A. and Middlebrooks, J. C. (2002). Listener weighting of cues for lateral angle : The duplex theory

of sound localization revisited. *The Journal of the Acoustical Society of America*, 111(5) :2219.

[Maj et al., 2006] Maj, J.-B., Royackers, L., Wouters, J., and Moonen, M. (2006). Comparison of adaptive noise reduction algorithms in dual microphone hearing aids. *Speech Communication*, 48(8) :957–970.

[Majdak et al., 2013] Majdak, P., Walder, T., and Laback, B. (2013). Effect of long-term training on sound localization performance with spectrally warped and band-limited head-related transfer functions. *The Journal of the Acoustical Society of America*, 134(3) :2148–2159.

[Markovich et al., 2009] Markovich, S., Gannot, S., and Cohen, I. (2009). Multichannel Eigenspace Beamforming in a Reverberant Noisy Environment With Multiple Interfering Speech Signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6) :1071–1086.

[Markovich-Golan and Gannot, 2015] Markovich-Golan, S. and Gannot, S. (2015). Performance Analysis of the Covariance Substraction Method for Relative Transfer Function Estimation and Comparison to the Covariance Whitening Method. In *ICASSP 2015 - 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 5, Brisbane, QLD, Australia.

[Markovich-Golan et al., 2012a] Markovich-Golan, S., Gannot, S., and Cohen, I. (2012a). Low-Complexity Addition or Removal of Sensors/Constraints in LCMV Beamformers. *IEEE Transactions on Signal Processing*, 60(3) :1205–1214.

[Markovich-Golan et al., 2012b] Markovich-Golan, S., Gannot, S., and Cohen, I. (2012b). A weighted multichannel Wiener filter for multiple sources scenarios. In *2012 IEEE 27th Convention of Electrical and Electronics Engineers in Israel*, pages 1–5, Eilat, Israel. IEEE.

[Marquardt, 2015] Marquardt, D. (2015). *Development and evaluation of psychoacoustically motivated binaural noise reduction and cue preservation techniques*. PhD thesis, Von der Fakultät für Medizin und Gesundheitswissenschaften der Carl von Ossietzky Universität Oldenburg, Oldenburg.

[Marquardt and Doclo, 2018] Marquardt, D. and Doclo, S. (2018). Interaural Coherence Preservation for Binaural Noise Reduction Using Partial Noise Estimation and Spectral Postfiltering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(7) :1261–1274.

[Marquardt et al., 2013] Marquardt, D., Hohmann, V., and Doclo, S. (2013). Coherence preservation in multi-channel Wiener filtering based noise reduction for binaural hearing aids. In *2013 IEEE International Conference on*

Acoustics, Speech and Signal Processing, pages 8648–8652, Vancouver, BC, Canada. IEEE.

[Marquardt et al., 2015] Marquardt, D., Hohmann, V., and Doclo, S. (2015). Interaural Coherence Preservation in Multi-Channel Wiener Filtering-Based Noise Reduction for Binaural Hearing Aids. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12) :2162–2176.

[Marrone et al., 2008a] Marrone, N., Mason, C. R., and Kidd, G. (2008a). The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms. *The Journal of the Acoustical Society of America*, 124(5) :3064–3075.

[Marrone et al., 2008b] Marrone, N., Mason, C. R., and Kidd, G. (2008b). Evaluating the Benefit of Hearing Aids in Solving the Cocktail Party Problem. *Trends in Amplification*, 12(4) :300–315.

[Mauler et al., 2007] Mauler, D., Nagathil, A. M., and Martin, R. (2007). On Optimal Estimation of Compressed Speech for Hearing Aids. In *Interspeech*, page 4.

[May et al., 2018] May, T., Kowalewski, B., and Dau, T. (2018). Signal-to-Noise-Ratio-Aware Dynamic Range Compression in Hearing Aids. *Trends in Hearing*, 22 :1–12.

[McAulay and Malpass, 1980] McAulay, R. and Malpass, M. (1980). Speech enhancement using a soft-decision noise suppression filter. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(2) :137–145.

[McCreery et al., 2012a] McCreery, R. W., Venediktov, R. A., Coleman, J. J., and Leech, H. M. (2012a). An Evidence-Based Systematic Review of Directional Microphones and Digital Noise Reduction Hearing Aids in School-Age Children With Hearing Loss. *American Journal of Audiology*, 21(2) :295.

[McCreery et al., 2012b] McCreery, R. W., Venediktov, R. A., Coleman, J. J., and Leech, H. M. (2012b). An Evidence-Based Systematic Review of Frequency Lowering in Hearing Aids for School-Age Children With Hearing Loss. *American Journal of Audiology*, 21(2) :313.

[Mendonça, 2014] Mendonça, C. (2014). A review on auditory space adaptations to altered head-related cues. *Frontiers in Neuroscience*, 8 :14.

[Mesgarani and Chang, 2012] Mesgarani, N. and Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397) :233–236.

[Middlebrooks, 1999a] Middlebrooks, J. C. (1999a). Individual differences in external-ear transfer functions reduced by scaling in frequency. *The Journal of the Acoustical Society of America*, 106(3) :1480–1492.

[Middlebrooks, 1999b] Middlebrooks, J. C. (1999b). Virtual localization improved by scaling nonindividualized external-ear transfer functions in frequency. *The Journal of the Acoustical Society of America*, 106(3) :1493–1510.

[Middlebrooks et al., 2017] Middlebrooks, J. C., Simon, J. Z., Popper, A. N., and Fay, R. R., editors (2017). *The Auditory system at the cocktail party*. Number Volume 60 in Springer Handbook of Auditory Research. Springer International Publishing, Cham. OCLC : 990257975.

[Middleton and Esposito, 1968] Middleton, D. and Esposito, R. (1968). Simultaneous optimum detection and estimation of signals in noise. *IEEE Transactions on Information Theory*, 14(3) :434–444.

[Mills, 1958] Mills, A. W. (1958). On the Minimum Audible Angle. *The Journal of the Acoustical Society of America*, 30(237).

[Møller, 1992] Møller, H. (1992). Fundamentals of binaural technology. *Applied acoustics*, 36(3-4) :171–218.

[Moore et al., 2019a] Moore, A. H., de Haan, J. M., Pedersen, M. S., Naylor, P. A., Brookes, M., and Jensen, J. (2019a). Personalized signal-independent beamforming for binaural hearing aids. *The Journal of the Acoustical Society of America*, 145(5) :2971–2981.

[Moore et al., 2019b] Moore, A. H., Naylor, P. A., and Brookes, M. (2019b). Improving robustness of adaptive beamforming for hearing devices. In *International Symposium on Auditory and Audiological Research*, volume 7, page 12.

[Moore and Glasberg, 1988] Moore, B. C. J. and Glasberg, B. R. (1988). A comparison of four methods of implementing automatic gain control (AGC) in hearing aids. *British Journal of Audiology*, 22(2) :93–104.

[Naylor and Johannesson, 2009] Naylor, G. and Johannesson, R. B. (2009). Long-Term Signal-to-Noise Ratio at the Input and Output of Amplitude-Compression Systems. *Journal of the American Academy of Audiology*, 20(3) :161–171.

[Ngo et al., 2012] Ngo, K., Spriet, A., Moonen, M., Wouters, J., and Holdt Jensen, S. (2012). A combined multi-channel Wiener filter-based noise reduction and dynamic range compression in hearing aids. *Signal Processing*, 92(2) :417–426.

[Ngo et al., 2009] Ngo, K., Spriet, A., Moonen, M., Wouters, J., and Jensen, S. H. (2009). Incorporating the Conditional Speech Presence Probability in Multi-Channel Wiener Filter Based Noise Reduction in Hearing Aids. *EURASIP Journal on Advances in Signal Processing*, 2009(1) :11.

[Nilsson et al., 1994] Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994). Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, 95(2) :1085–1099.

[Oreinos and Buchholz, 2013] Oreinos, C. and Buchholz, J. M. (2013). Measurement of Full 3D Set of HRTFs for In-Ear and Hearing Aid Microphones on a Head and Torso Simulator. *Acta Acustica united with Acustica*, 99 :836–844.

[Oreinos and Buchholz, 2015] Oreinos, C. and Buchholz, J. M. (2015). Objective analysis of ambisonics for hearing aid applications : Effect of listener's head, room reverberation, and directional microphones. *The Journal of the Acoustical Society of America*, 137(6) :3447–3465.

[Popelka et al., 2016] Popelka, G. R., Moore, B. C. J., Fay, R. R., and Popper, A. N., editors (2016). *Hearing Aids*, volume 56 of *Springer Handbook of Auditory Research*. Springer International Publishing, Cham.

[Press et al., 2007] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, r. P. (2007). *Numerical Recipes*. Cambridge University Press, third edition.

[Rhebergen et al., 2009] Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (2009). The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interrupted noise. *The Journal of Acoustical Society of America*, 126(6) :10.

[Rickard and Yilmaz, 2002] Rickard, S. and Yilmaz, z. (2002). On the Approximate W-Disjoint Orthogonality of Speech. In *ICASSP 2002 - 2002 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 4, Orlando, USA.

[Saddler et al., 2021] Saddler, M. R., Francl, A., Feather, J., Qian, K., Zhang, Y., and McDermott, J. H. (2021). Speech Denoising with Auditory Models. In *Interspeech 2021*, pages 2681–2685. ISCA.

[Schoeffler et al., 2018] Schoeffler, M., Bartoschek, S., Stöter, F.-R., Roess, M., Westphal, S., Edler, B., and Herre, J. (2018). webMUSHRA — A Comprehensive Framework for Web-based Listening Tests. *Journal of Open Research Software*, 6 :8.

[Schwartz and Shinn-Cunningham, 2013] Schwartz, A. H. and Shinn-Cunningham, B. G. (2013). Effects of dynamic range compression on spatial selective auditory attention in normal-hearing listeners. *The Journal of the Acoustical Society of America*, 133(4) :2329–2339.

[Schwarz and Kellermann, 2015] Schwarz, A. and Kellermann, W. (2015). Coherent-to-Diffuse Power Ratio Estimation for Dereverberation.

IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(6) :1006–1018.

[Serizel et al., 2014] Serizel, R., Moonen, M., Van Dijk, B., and Wouters, J. (2014). Low-rank Approximation Based Multichannel Wiener Filter Algorithms for Noise Reduction with Application in Cochlear Implants. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4) :785–799.

[Sherbecoe and Studebaker, 2004] Sherbecoe, R. L. and Studebaker, G. A. (2004). Supplementary formulas and tables for calculating and interconverting speech recognition scores in transformed arcsine units. *International Journal of Audiology*, 43(8) :7.

[Shinn-Cunningham et al., 2000] Shinn-Cunningham, B. G., Santarelli, S., and Kopco, N. (2000). Tori of confusion : Binaural localization cues for sources within reach of a listener. *The Journal of the Acoustical Society of America*, 107(3) :1627–1636.

[Skottun et al., 2001] Skottun, B. C., Shackleton, T. M., Arnott, R. H., and Palmer, A. R. (2001). The ability of inferior colliculus neurons to signal differences in interaural delay. *Proceedings of the National Academy of Sciences*, 98(24) :14050–14054.

[Souden et al., 2010] Souden, M., Jingdong Chen, Benesty, J., and Affes, S. (2010). Gaussian Model-Based Multichannel Speech Presence Probability. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5) :1072–1077.

[Souza et al., 2006] Souza, P. E., Jenstad, L. M., and Boike, K. T. (2006). Measuring the acoustic effects of compression amplification on speech in noise. *The Journal of the Acoustical Society of America*, 119(1) :41–44.

[Souza and Turner, 1998] Souza, P. E. and Turner, C. W. (1998). Multichannel Compression, Temporal Cues, and Audibility. *Journal of Speech, Language, and Hearing Research*, 41(2) :315–326.

[Spagnol et al., 2010] Spagnol, S., Geronazzo, M., and Avanzini, F. (2010). Structural Modeling Of Pinna-Related Transfer Functions. *Sound and Music Computing Conference*, pages 422–428.

[Stadler and Rabinowitz, 1993] Stadler, R. W. and Rabinowitz, W. M. (1993). On the potential of fixed arrays for hearing aids. *The Journal of Acoustical Society of America*, 94(3) :1332–1342.

[Steinberg and Gardner, 1937] Steinberg, J. C. and Gardner, M. B. (1937). The Dependence of Hearing Impairment on Sound Intensity. *The Journal of the Acoustical Society of America*, 9(11) :11–23.

[Stelmachowicz et al., 2001] Stelmachowicz, P. G., Pittman, A. L., Hoover, B. M., and Lewis, D. E. (2001). Effect of stimulus bandwidth on the perception of /s/ in normal- and hearing-impaired children and adults. *J. Acoust. Soc. Am.*, 110(4) :8.

[Stone and Moore, 2003] Stone, M. A. and Moore, B. C. J. (2003). Tolerable Hearing Aid Delays. III. Effects on Speech Production and Perception of Across-Frequency Variation in Delay. *Ear and Hearing*, 24(2) :175–183.

[Stone and Moore, 2007] Stone, M. A. and Moore, B. C. J. (2007). Quantifying the effects of fast-acting compression on the envelope of speech. *The Journal of the Acoustical Society of America*, 121(3) :10.

[Stone and Moore, 2008] Stone, M. A. and Moore, B. C. J. (2008). Effects of spectro-temporal modulation changes produced by multi-channel compression on intelligibility in a competing-speech task. *The Journal of the Acoustical Society of America*, 123(2) :1063–1076.

[Studebaker, 1985] Studebaker, G. A. (1985). A "Rationalized" Arcsine Transform. *Journal of Speech, Language, and Hearing Research*, 28(3) :455–462.

[Suzuki et al., 1999] Suzuki, Y., Tsukui, S., Asano, F., Nishimura, R., and Sone, T. (1999). New Design Method of a Binaural Microphone Array Using Multiple Constraints. *IEICE Trans. Fundamentals*, 82(4) :588–596.

[Taal et al., 2010] Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217, Dallas, TX, USA. IEEE.

[Talmon et al., 2009] Talmon, R., Cohen, I., and Gannot, S. (2009). Relative Transfer Function Identification Using Convolutive Transfer Function Approximation. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(4) :546–555.

[Thiemann et al., 2016] Thiemann, J., Müller, M., Marquardt, D., Doclo, S., and van de Par, S. (2016). Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene. *EURASIP Journal on Advances in Signal Processing*, 2016(1) :11.

[Thiergart and Habets, 2014] Thiergart, O. and Habets, E. A. P. (2014). An Informed Parametric Spatial Filter Based on Instantaneous Direction-of-Arrival Estimates. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12) :15.

[Thiergart et al., 2016] Thiergart, O., Huang, W., and Habets, E. A. P. (2016). A low complexity weighted least squares narrowband DOA estimator for arbitrary array geometries. In *2016 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP), pages 340–344, Shanghai. IEEE.

[Thiergart et al., 2013] Thiergart, O., Taseska, M., and Habets, E. A. P. (2013). An Informed MMSE Filter based on Multiple Instantaneous Direction-of-Arrival Estimates. In *Proc. 21st Eur. Signal Process. Conf. (EUSIPCO 13)*, page 5.

[Vaillancourt et al., 2011] Vaillancourt, V., Laroche, C., Giguère, C., Beaulieu, M.-A., and Legault, J.-P. (2011). Evaluation of Auditory Functions for Royal Canadian Mounted Police Officers. *Journal of the American Academy of Audiology*, 22(06) :313–331.

[Van den Bogaert et al., 2011] Van den Bogaert, T., Carette, E., and Wouters, J. (2011). Sound source localization using hearing aids with microphones placed behind-the-ear, in-the-canal, and in-the-pinna. *International Journal of Audiology*, 50(3) :164–176.

[Van den Bogaert et al., 2008] Van den Bogaert, T., Doclo, S., Wouters, J., and Moonen, M. (2008). The effect of multimicrophone noise reduction systems on sound source localization by users of binaural hearing aids. *The Journal of the Acoustical Society of America*, 124(1) :484–497.

[Van den Bogaert et al., 2006] Van den Bogaert, T., Klasen, T. J., Moonen, M., Van Deun, L., and Wouters, J. (2006). Horizontal localization with bilateral hearing aids : Without is better than with. *The Journal of the Acoustical Society of America*, 119(1) :515–526.

[Van den Bogaert et al., 2007] Van den Bogaert, T., Wouters, J., Doclo, S., and Moonen, M. (2007). Binaural cue preservation for hearing aids using an interaural transfer function multichannel Wiener filter. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–565. IEEE.

[Van Eyndhoven et al., 2017] Van Eyndhoven, S., Francart, T., and Bertrand, A. (2017). EEG-Informed Attended Speaker Extraction From Recorded Speech Mixtures With Application in Neuro-Steered Hearing Prostheses. *IEEE Transactions on Biomedical Engineering*, 64(5) :1045–1056.

[Van Opstal, 2016] Van Opstal, J. (2016). *The auditory system and human sound-localization behavior*. Academic Press.

[Vincent et al., 2006] Vincent, E., Gribonval, R., and Févotte, C. (2006). Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing, Institute of Electrical and Electronics Engineers*, page 10.

[Völker et al., 2015] Völker, C., Warzybok, A., and Ernst, S. M. A. (2015). Comparing Binaural Pre-processing Strategies III : Speech Intelligibility of Normal-Hearing and Hearing-Impaired Listeners. *Trends in Hearing*, 19 :18.

[Wiggins and Seeber, 2011] Wiggins, I. M. and Seeber, B. U. (2011). Dynamic-range compression affects the lateral position of sounds. *The Journal of the Acoustical Society of America*, 130(6) :3939–3953.

[Wiggins and Seeber, 2012] Wiggins, I. M. and Seeber, B. U. (2012). Effects of dynamic-range compression on the spatial attributes of sounds in normal-hearing listeners. *Ear and hearing*, 33(3) :399–410.

[Wiggins and Seeber, 2013] Wiggins, I. M. and Seeber, B. U. (2013). Linking dynamic-range compression across the ears can improve speech intelligibility in spatially separated noise. *The Journal of the Acoustical Society of America*, 133(2) :1004–1016.

[Wong et al., 2018] Wong, L. L. N., Chen, Y., Wang, Q., and Kuehnel, V. (2018). Efficacy of a Hearing Aid Noise Reduction Function. *Trends in Hearing*, 22 :14.

[Ye and DeGroat, 1995] Ye, H. and DeGroat, D. (1995). Maximum likelihood DOA estimation and asymptotic Cramer-Rao bounds for additive unknown colored noise. *IEEE Transactions on Signal Processing*, 43(4) :938–949.

[Zahorik, 2002] Zahorik, P. (2002). Direct-to-reverberant energy ratio sensitivity. *The Journal of the Acoustical Society of America*, 112(5) :2110–2117.

[Zahorik, 2005] Zahorik, P. (2005). Auditory Distance Perception in Humans : A Summary of Past and Present Research. *Acta Acustica united with Acustica*, 91 :12.

[Zakarauskas and Cynader, 1993] Zakarauskas, P. and Cynader, M. S. (1993). A computational theory of spectral cue localization. *The Journal of the Acoustical Society of America*, 94(3) :1323–1331.

[Zhang et al., 2021] Zhang, X., Ren, X., Zheng, X., Chen, L., Zhang, C., Guo, L., and Yu, B. (2021). Low-Delay Speech Enhancement Using Perceptually Motivated Target and Loss. In *Interspeech 2021*, pages 2826–2830. ISCA.

[Zohourian et al., 2018] Zohourian, M., Enzner, G., and Martin, R. (2018). Binaural Speaker Localization Integrated Into an Adaptive Beamformer for Hearing Aids. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(3) :515–528.

[Zohourian et al., 2017] Zohourian, M., Martin, R., and Madhu, N. (2017). New insights into the role of the head radius in model-based binaural speaker localization. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 221–225, Kos, Greece. IEEE.

DOCTORAT BRETAGNE LOIRE / MATHSTIC

CentraleSupélec

Titre : Amélioration de la compréhension de la parole et de l'écoute spatiale pour les malentendants appareillés

Mot clés : Prothèse auditive, formation de voie, localisation auditive, compression de dynamique, intelligibilité de la parole

Résumé : Les prothèses auditives ont pour but de faire recouvrir les principales capacités auditives, au premier rang de laquelle : l'intelligibilité de la parole. Cela est assuré principalement par deux tâches : compenser la perte auditive et réduire le niveau de bruit. La réduction de bruit et la compensation de perte auditive sont effectuées l'une à la suite de l'autre. Or, toutes deux ont des objectifs antagonistes et introduisent des artefacts néfastes à l'appréhension d'une scène sonore complexe dans sa globalité.

Dans un premier temps, nous unifions le formalisme sous-jacent aux algorithmes de débruitage et de compensation de perte de sorte à développer une solution explicite au problème dans son ensemble, pour une scène sonore composée d'une source de parole et d'un bruit ambiant.

Dans une second temps, nous nous employons à mieux préserver les indices de localisation interauraux pour toutes les directions de l'espace. Pour cela, nous développons trois méthodes basées sur des approximations du terme de coût associé à la préservation de la fonction de transfert interaurale.

Enfin, nous élargissons notre modèle de scène sonore à plusieurs sources de parole et du bruit ambiant. Le contexte des prothèses auditives rend ce cas difficile à traiter du fait du nombre réduit de microphones. Nous proposons d'exploiter la propriété de parcimonie de la parole dans le domaine temps-fréquence pour dépasser cet obstacle.

Title: Improvement of the speech intelligibility and the spatial hearing for aided listeners

Keywords: Hearing aids, beamforming, auditory localization, dynamic range compression, speech intelligibility

Abstract: Hearing aids are designed to restore the essential abilities of hearing, the most important of which is speech intelligibility. This is achieved mainly through two functions: compensating for hearing loss and reducing the noise level. Noise reduction and hearing loss compensation are performed one after the other. However, both have antagonistic objectives and introduce artifacts that are detrimental to the apprehension of a complex auditory scene in its entirety.

In a first step, we unify the formalism underlying the denoising and loss compensation algorithms in order to develop an explicit solution to the problem as a whole, for an auditory scene composed of one speech source and an ambient noise.

In a second step, we focus on a better preservation of the interaural localization cues for all spatial directions. For this purpose, we develop three methods based on approximations of the cost function related to the interaural transfer function preservation.

Finally, we extend the auditory scene model to several speech sources and ambient noise. The context of hearing aids makes this case difficult to handle due to the small number of microphones. We propose to exploit the sparsity property of speech in the time-frequency domain to overcome this obstacle.