Apprentissage profond pour le rehaussement de la parole dans les antennes acoustiques ad-hoc

Nicolas Furnon

UNIVERSITÉ DE LORRAINE

Thèse de Doctorat

Nicolas Furnon

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Lorraine
Mention Informatique

École doctorale : IAEM

Unité de recherche : Laboratoire Lorrain de Recherche en Informatique et ses Applications UMR 7503

Soutenue le 14 décembre 2021
Thèse N° :

Apprentissage profond pour le rehaussement de la parole dans les antennes acoustiques ad-hoc

JURY

Rapporteur :	**Marc Delcroix**, Directeur de recherche, NTT Corporation, Kyoto, Japon
Rapporteur :	**Mathieu Lagrange**, Chargé de recherche CNRS, LS2N, Ecole Centrale de Nantes, France
Examinatrice :	**Ann Spriet**, Directrice de recherche, Goodix Technology Inc., Louvain, Belgique
Examinateur (Président) :	**Joël Ducourneau**, Professeur, LEMTA, Université de Lorraine, France
Invité :	**Romain Serizel**, Maître de conférences, Université de Lorraine, Inria, Loria, France
Directrice de thèse :	**Irina Illina**, Maître de conférences, HDR, Université de Lorraine, Inria, Loria, France
Directeur de thèse :	**Slim Essid**, Professeur, LTCI, Télécom Paris, France

Résumé

Un grand nombre d'appareils que nous utilisons au quotidien embarque un ou plusieurs microphones afin de rendre possible leur utilisation par commande vocale. Le réseau de microphones que l'on peut former avec ces appareils est ce qu'on appelle une antenne acoustique ad-hoc (AAAH). Une étape de rehaussement de la parole est souvent appliquée afin d'optimiser l'exécution des commandes vocales. Pour cela, les AAAH, de par leur flexibilité d'utilisation, leur large étendue spatiale et la diversité de leurs enregistrements, offrent un grand potentiel. Ce potentiel est néanmoins difficilement exploitable à cause de la mobilité des appareils, leur faible puissance et les contraintes en bande passante. Ces limites empêchent d'utiliser les algorithmes de rehaussement de la parole « classiques » qui reposent sur un nœud de fusion et requièrent de fortes puissances de calcul.

Cette thèse propose de rallier le domaine de l'apprentissage profond à celui des AAAH, en conciliant la puissance de modélisation des réseaux de neurones (RN) à la flexibilité d'utilisation des AAAH. Pour cela, nous présentons un système distribué de rehaussement de la parole. Il est distribué en cela que la contrainte d'un centre de fusion est levée. Des signaux dits compressés, échangés entre les nœuds, permettent de véhiculer l'information spatiale tout en réduisant la consommation en bande passante. Des RN sont utilisés afin d'estimer les coefficients d'un filtre de Wiener multicanal. Une analyse empirique détaillée de ce système est conduite à la fois sur données synthétiques et sur données réelles afin de valider son efficacité et de mettre en évidence l'intérêt d'utiliser conjointement des RN et des algorithmes distribués classiques de rehaussement de la parole. Nous montrons ainsi que notre système obtient des performances équivalentes à celles de l'état de l'art, tout en étant plus flexible et en réduisant significativement la complexité algorithmique.

Par ailleurs, nous développons notre solution pour l'adapter à des conditions d'utilisation propres aux AAAH. Nous étudions son comportement lorsque le nombre d'appareils de l'AAAH varie, et nous comparons l'influence de deux mécanismes d'attention, l'un d'attention spatiale et l'autre d'auto-attention. Les deux mécanismes d'attention rendent notre système résilient à un nombre variable d'appareils et les poids du mécanisme d'auto-attention révèlent l'utilité de l'information convoyée par chaque signal.

Nous analysons également le comportement de notre système lorsque les signaux des différents appareils sont désynchronisés. Nous proposons une solution pour améliorer les performances de notre système en conditions asynchrones, en présentant un autre mécanisme d'attention. Nous montrons que ce mécanisme d'attention permet de retrouver un ordre de grandeur du décalage d'horloge entre les appareils d'une AAAH.

Enfin, nous montrons que notre système est une solution viable pour la séparation de sources de parole. Même avec des RN d'architecture simple, il est capable d'exploiter efficacement l'information spatiale enregistrée par tous les appareils d'une AAAH dans une configuration typique de réunion.

Abstract

More and more devices we use in our daily life are embedded with one or more microphones so that they can be voice controlled. Put together, these devices can form a so-called ad-hoc microphone array (AHMA). A speech enhancement step is often applied on the recorded signals to optimise the execution of the voice commands. To this effect, AHMAs are of high interest because of their flexible usage, their wide spatial coverage and the diversity of their recordings. However, it is challenging to exploit the potential of AHMAs because devices that compose them may move and have a limited power and bandwidth capacity. Because of these limits, the speech enhancement solutions deployed in "classic" microphone arrays, relying on a fusion center and high processing loads, cannot be afforded.

This thesis combines the modelling power of deep neural networks (DNNs) with the flexibility of use of AHMAs. To this end, we introduce a distributed speech enhancement system, which does not rely on a fusion center. So-called compressed signals are sent among the nodes and convey the spatial information recorded by the whole AHMA, while reducing the bandwidth requirements. DNNs are used to estimate the coefficients of a multichannel Wiener filter. We conduct an empirical analysis of this system, both on synthesized and real data, in order to validate its efficiency and to highlight the benefits of jointly using DNNs and distributed speech enhancement algorithms. We show that our system performs comparatively well compared with a state-of-the-art solution, while being more flexible and significantly reducing the computation cost.

Besides, we develop our solution to adapt it to the typical usage conditions of AHMAs. We study its behaviour when the number of devices in the AHMA varies. We introduce and compare a spatial attention mechanism and a self-attention mechanism. Both mechanisms make our system robust to a varying number of devices. We show that the weights of the self-attention mechanism reveal the utility of the information carried by each signal.

We also analyse our system when the signals recorded by different devices are not synchronised. We propose a solution to improve its performance in such conditions by introducing a temporal attention mechanism. We show that this mechanism can help estimating the sampling time offset between the several devices of the AHMA.

Lastly, we show that our system is also efficient for source separation. It can efficiently process the spatial information recorded by the whole AHMA in a typical meeting scenario and alleviate the needs of a complex DNN architecture.

Chercheur, chercheuse – car quelle autre population
A le temps et l'envie, sinon l'obligation
Plus qu'un autre récit, de lire cet ouvrage ? –
Avant de t'enhardir à tourner cette page,
Ne te hâte pas trop, écoute mon message.
Thalès, Socrate et tous ces philosophes sages
Estimaient la science comme on apprécie l'art.
Nous prétendrons ainsi que l'informaticien
N'a à jalouser ni Raphaël, ni Mozart
Et use du clavier non moins qu'un musicien.

Admirez les lignes de la belle Vénus
De Milo ; j'admire celles de C++.

Honneurs vont au danseur ; "Erreur" voit le codeur.
Oui, informatique et esthétique sont sœurs,
Croyez-en ce calembour sinon mon labeur.

Remerciements

Trop de remerciements tuent les remerciements. Afin de ne noyer ni les lecteurs ni mon message de reconnaissance dans d'interminables remerciements, je ne les adresserai qu'à mes trois encadrants, Romain, Irina et Slim. Qu'on me pardonne de ne pas m'épancher en déclarations d'amitié, d'amour et de foi, mais s'il est tout-à-fait probable que sans l'amitié, l'amour et la foi, cette thèse aurait été moins épanouissante, il est absolument certain que sans Slim, Romain et Irina, elle n'aurait même pas commencé.

Trois encadrants plutôt qu'un, certains diront que c'est trois fois plus d'avis divergents, trois fois plus de critiques et trois fois plus de corrections à prendre en compte. C'est vrai... Mais c'est oublier que c'est aussi trois fois plus de conseils, trois fois plus de points de vue, et trois fois plus de relations, qui enrichissent le suivi de thèse d'une manière difficilement quantifiable. Merci à vous trois de vous être si bien complétés, par vos approches scientifiques et vos sensibilités relationnelles différentes.

Merci pour l'encadrement que vous m'avez proposé. Le mot est éloquent : vous avez délimité le cadre du possible et marqué les frontières à ne pas franchir. Vous m'avez ainsi gardé éloigné de certains extrêmes, entre l'excès d'optimisme et l'excès d'indolence.

Merci pour la liberté que vous m'avez laissée sur les problématiques traitées et sur les solutions considérées. Et comme il n'y a pas de liberté sans confiance, merci pour votre confiance, un peu vertigineuse au début de ma thèse. Je me souviens qu'il me paraissait bien téméraire de votre part d'interpréter mes résultats sans avoir relu mon code ni même envisagé l'éventualité qu'il pouvait être erroné.

Enfin, merci de m'avoir montré l'exemple de ce qu'est un bon chercheur. Au-delà de la passion, de la curiosité et de la persévérance, de toutes vos qualités, s'il en est une que j'aimerais avoir acquise à votre contact, c'est celle de la rigueur. Cette rigueur qui ne tolère pas d'explication bancale ni de conclusion hâtive qu'une expérience ciblée ne puisse valider. Mon penchant d'ingénieur à qui satisfait ce qui fonctionne, a été redressé par votre esprit de chercheur, qui préfère l'explication au résultat.

De manière individuelle, merci Irina d'avoir accepté d'être ma directrice de thèse alors que le rehaussement de la parole n'était pas au centre de ta recherche. Merci pour le temps que tu as pris pour moi dans un emploi du temps déjà bien chargé, pour les solutions suggérées au cours des réunions et pour tes conseils, que ce soit sur l'entraînement de réseaux de neurones, sur mon organisation de travail ou sur la rédaction de ma thèse.

Merci à Slim de m'avoir accompagné, à distance, mais avec quelle intensité ! Qu'il est formateur de t'avoir comme encadrant, lorsque tu remets tout en question, pointes les limites d'une approche, lances un débat, ébranles une théorie vaseuse ou proposes des solutions dont je n'ai jamais entendu parler. Merci pour ta franchise qui m'a plus d'une fois remis sur le bon chemin. Dans tes critiques comme dans tes compliments, elle m'a donné confiance en toi et en moi.

Et enfin, merci Romain. Il me faudrait des pages pour t'exprimer ma reconnaissance pour ta disponibilité et ton aide. Que ce petit mot, merci, ne cache pas l'immense reconnaissance que j'ai pour ton accompagnement ni ma très grande admiration pour tes compétences de chercheur. Et si la tradition d'ajouter une page de remerciements au début du manuscrit retire un peu de la spontanéité de ces mots, ils n'en sont pas moins sincères.

Cette thèse a été financée par l'Agence Nationale de la Recherche dans le cadre du projet DiSCogs (ANR-17-CE23-0026-01). Une partie des expériences présentées dans ce manuscrit ont été effectuées sur Grid5000, un serveur soutenu par un groupe à intérêts scientifiques de l'Inria, du CNRS, de RENATER et d'autres universités et organisations (cf https://www.grid5000). D'autres expériences ont été effectuées sur le serveur EXPLOR de l'Université de Lorraine.

Table des matières

Liste des figures

Liste des tableaux

Liste des acronymes

AAAH	antenne acoustique ad-hoc
RN	réseau de neurones
RNN	RN récurrent
CRNN	RN convolutionnel récurrent
RI	réponse impulsionnelle
TF	temps-fréquence
DAV	détecteur d'activité vocale
TFCT	transformée de Fourier à court terme
MAS	matrice d'autocorrélation spatiale
MRI	masque de ratio idéal
RSB	rapport signal-à-bruit
dB	décibels
RNMuN	réseau de neurones multinœud
RNMoN	réseau de neurones mono-nœud
BMP	bruit modulé par la parole
TR	temps de réverbération
FE	fréquence d'échantillonnage
FIR	filtre à réponse finie
DSB	*delay-and-sum beamformer*
MPDR	*minimum power distortionless response*
MVDR	*minimum variance distortionless response*
LCMP	*linearly constrained minimum power*
LCMV	*linearly constrained minimum variance*
MWF	*multichannel Wiener filter*
SDW-MWF	*speech distortion weighted multichannel Wiener filter*
GSC	*generalized sidelobe canceller*
DANSE	*distributed adaptive node-specific signal estimation*
DBSA	*dual-based subgradient algorithm*
ADMM	*alternative direction method of multipliers*
PDMM	*primal-dual method of multipliers*
SNR	*signal to noise ratio*
SAR	*sources to artifacts ratio*
SIR	*source to interferences ratio*
SDR	*source to distortion ratio*
SI-SDR	*scale-invariant signal to distortion ratio*
SI-SIR	*scale-invariant signal to interference ratio*
SI-SAR	*scale-invariant signal to artifacts ratio*

STOI	*short-time objective intelligibility*
LSTM	*long short-term memory*
GRU	*gated recurrent unit*
SE	*squeeze and excitation*
CBAM	*convolutional block attention module*
SA	*self attention*

Première partie

Mise en contexte

1. Introduction

1.1. Contexte

Si la parole a longtemps été le propre de l'homme, il faut bien reconnaître qu'elle ne l'est plus exclusivement depuis que les machines se la sont appropriées. Aujourd'hui, celles-ci sont capables de transcrire et de parler plusieurs langues, bien que ces capacités soient encore de vastes sujets de recherche (Vadwala et al., 2017; Malik et al., 2021; Ning et al., 2019; Tan et al., 2021c). Grâce aux dernières avancées technologiques, en particulier grâce à l'essor de l'apprentissage automatique, de plus en plus de nos interactions avec les machines reposent sur des commandes vocales. On peut ainsi demander la météo, fermer ses volets, programmer sa route sur système de navigation ou encore envoyer un message en parlant directement avec les appareils ou les applications responsables de ces différentes tâches. Les commandes sont d'autant mieux comprises que la parole émise est claire, c'est-à-dire dénuée de bruit ou d'interférences, qui la déforment et rendent plus difficile son interprétation par les machines. Lorsque la commande est émise en présence de bruit, il convient donc de supprimer ce bruit, ce qui est effectué au cours d'une étape dite de *rehaussement de la parole* (Loizou, 2007). Le rehaussement de la parole regroupe de nombreux autres traitements, comme la déréverbération (Naylor and Gaubitch, 2010; Delcroix et al., 2007), l'égalisation (Westerlund et al., 2005; Gentet et al., 2020) ou le rehaussement d'enveloppe (Lorenzi et al., 1999). Dans cette thèse, nous utiliserons ce terme essentiellement pour désigner la réduction de bruit, qui est une des composantes principales du rehaussement de la parole.

Que ce soit dans les téléphones, dans les systèmes de navigation, dans les ordinateurs portables ou dans les appareils auditifs, de nombreux microphones nous entourent et peuvent enregistrer la scène acoustique dans laquelle ils sont actifs. Lorsque plusieurs de ces appareils qui embarquent des microphones sont présents dans une même scène et sont considérés comme une seule entité, ils forment ce qu'on appelle une *antenne acoustique ad-hoc (AAAH)*. La figure 1.1 représente une AAAH dans un environnement domestique. Les AAAH présentent de nombreux avantages pour le rehaussement de la parole et constituent donc un moyen prometteur pour cette tâche.

Un autre axe de progrès majeur pour le rehaussement de la parole est l'utilisation de l'apprentissage automatique. L'apprentissage automatique est un des domaines de recherche à la base de l'intelligence artificielle qui a permis de grandes avancées dans des applications aussi variées que la conduite autonome (Badue et al., 2021), la reconnaissance de la parole (Malik et al., 2021), les diagnostics médicaux (Dai et al., 2019) ou la cybersécurité (Fraley and Cannady, 2017). Il a également été appliqué avec succès dans le domaine du rehaussement de la parole (Vincent et al., 2018), où une grande majorité des algorithmes le font intervenir aujourd'hui.

Cette thèse est dédiée au rehaussement de la parole dans les AAAH, et cherche à profiter des forces de l'apprentissage automatique dans ce contexte. Les avantages qu'il y a à considérer des AAAH, et les problématiques qui lui sont inhérentes, sont présentés en section 1.2. La section 1.3 décrit l'utilisation de l'apprentissage profond pour le rehaussement de la parole et les limites des solutions existantes sont indiquées en section 1.4. Les contributions de la thèse sont détaillées en section 1.5. En fin de chapitre, nous listons les différents articles publiés dans le cadre de cette thèse ainsi que les données que nous avons rendues publiques (section 1.7).

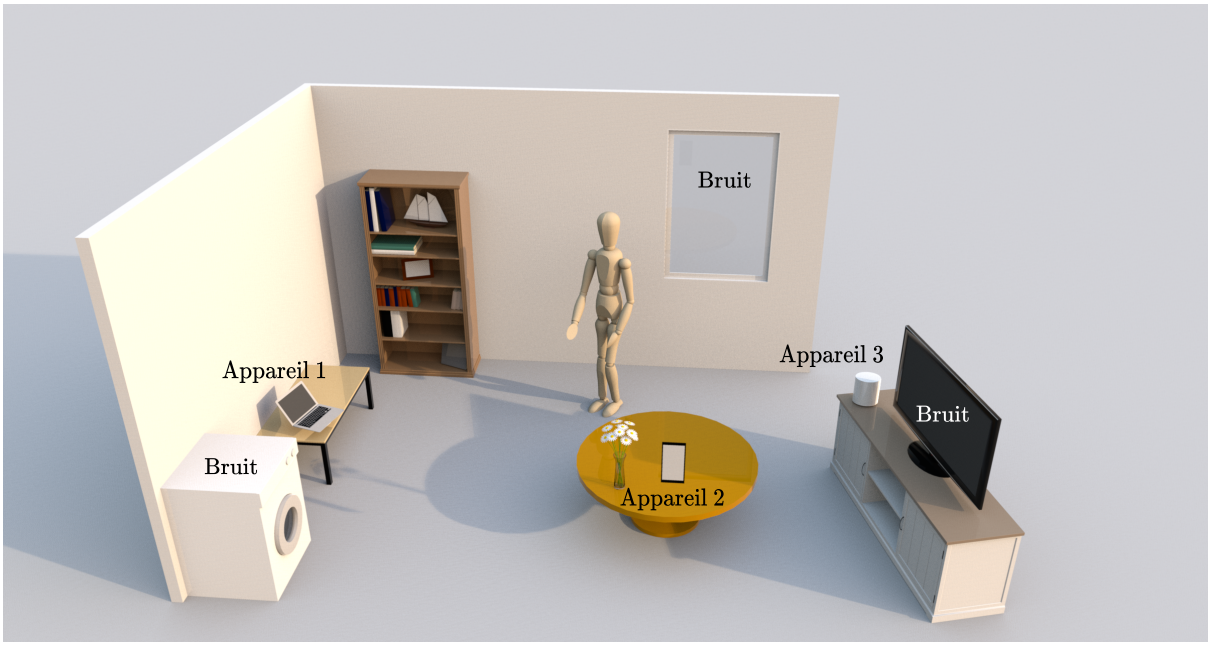


FIGURE 1.1. – Exemple d'AAAH formée par des appareils du quotidien (enceinte connectée, ordinateur, télévision connectée).

1.2. Les antennes de microphones

Nous commençons cette section par évoquer ce que sont les antennes acoustiques avant de décrire les caractéristiques des AAAH, leurs avantages et les difficultés qu'elles présentent.

1.2.1. Les antennes acoustiques

Une antenne acoustique regroupe plusieurs microphones, embarqués sur le même appareil. La présence de plusieurs microphones permet d'effectuer ce qu'on appelle un filtrage spatial, c'est-à-dire de combiner les signaux enregistrés par ces microphones de sorte à restituer le plus fidèlement possible le signal d'intérêt. Dans le contexte du rehaussement de la parole, ce signal d'intérêt est le signal de la parole émis par la personne que l'on veut entendre.

Le terme technique pour se référer à la combinaison des signaux dans le but de rehausser la parole est celui de *formation de voies*. Il désigne le fait que les sons venant d'une direction donnée (appelée direction d'arrivée) sont privilégiés, et que ceux venant d'autres directions sont atténués. Pour cela, un faisceau d'écoute est créé ; de la même manière qu'un faisceau de lumière éclaire uniquement dans la direction du faisceau, un faisceau d'écoute n'enregistre principalement que les sons venant de la direction du faisceau. En fonction de la fréquence de ces sons, de la géométrie de l'antenne, de la stratégie de combinaison des signaux, différents faisceaux peuvent être créés. Les motifs de deux de ces faisceaux pour une antenne acoustique circulaire uniforme (c'est-à-dire dont les microphones sont disposés à intervalles réguliers sur un cercle) sont représentés en figure 1.2, où les lobes principaux sont pointés dans la direction de la source cible. On peut y voir que le faisceau est plus resserré pour les hautes fréquences que pour les basses fréquences. Une présentation plus détaillée de différents filtres spatiaux qu'il est possible d'implémenter dans une antenne acoustique est donnée en section 2.3.1.

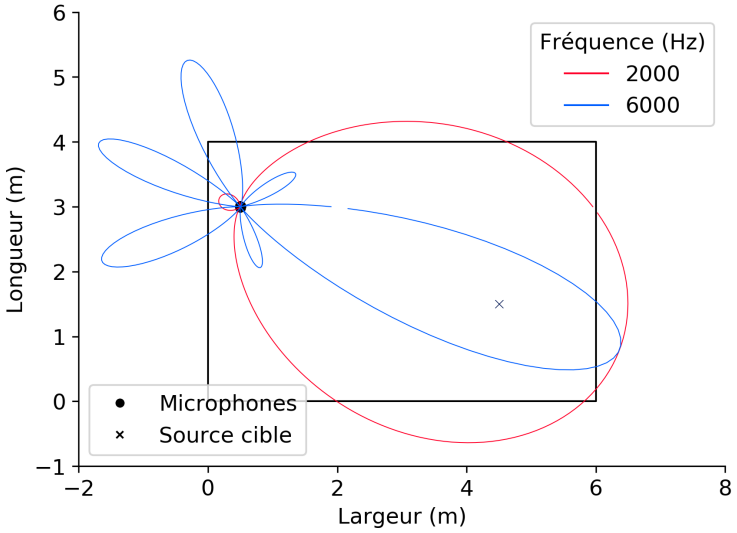
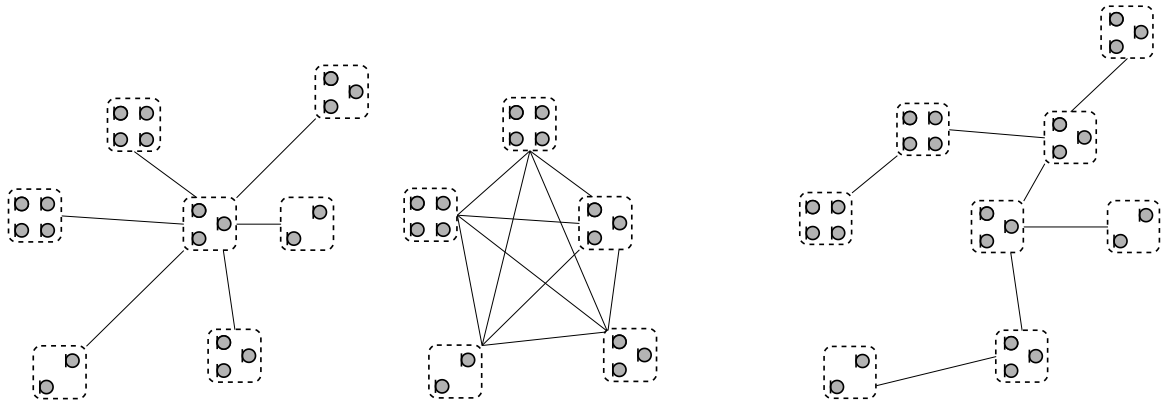


FIGURE 1.2. – Motifs de formateurs de voies d'une antenne circulaire uniforme à différentes fréquences. La croix désigne la source cible et le cercle désigne l'antenne de microphones circulaire.

1.2.2. Les antennes acoustiques ad-hoc

On parle de réseau d'antennes, ou de réseaux de capteurs acoustiques lorsque plusieurs antennes acoustiques sont mises en commun et utilisées dans un but précis, par exemple de rehaussement de la parole. Les AAAH font partie de la catégorie assez large des réseaux d'antennes, mais avec cette caractéristique que les appareils qui la constituent ne sont pas nécessairement conçus pour être des antennes acoustiques. Par exemple, la vocation d'un ordinateur n'est pas d'enregistrer des scènes acoustiques. S'il contient des microphones, c'est parce que certaines de ses utilisations, en particulier la visio-conférence, en requièrent. Or beaucoup de ces appareils embarquant des microphones nous entourent : les ordinateurs, mais aussi les téléphones, les enceintes connectées, les aides auditives, les montres intelligentes, etc. La figure 1.1 illustre une AAAH formée par certains de ces appareils dans un contexte typique d'utilisation, où une personne communique avec un ordinateur, mais dont la commande est perturbée par les bruit environnants (bruit de la rue, bruit du lave-linge). Que ce soit pour interagir avec une machine par commande vocale ou pour améliorer la qualité du signal transmis par une aide auditive, tous ces microphones constituent une richesse qui fait la force des AAAH, mais ils relèvent également de nombreuses difficultés. Ces avantages et difficultés sont discutés dans la suite.

Avantages des antennes acoustiques ad-hoc Le premier avantage des AAAH est lié au fait que plus de microphones sont à la disposition des algorithmes de rehaussement de la parole. Avoir plus de microphones laisse plus de liberté dans la conception du filtrage spatial que l'on souhaite appliquer, par exemple en choisissant les microphones les plus pertinents. Cela permet aussi de créer des formateurs de voies aux faisceaux plus resserrés, et donc de mieux atténuer le bruit venant des directions indésirables. Par ailleurs, les AAAH offrent une couverture spatiale bien supérieure à celle des antennes acoustiques classiques. Dans le contexte d'une conférence par exemple, où de multiples sources de parole peuvent être actives, il serait coûteux et peu flexible de couvrir la salle d'antennes acoustiques classiques dans le but d'avoir une estimation de chacune des sources. Cependant, il paraît tout-à-fait envisageable de recourir à tous les microphones présents dans les ordinateurs ou téléphones des personnes présentes pour effectuer cette tâche. De plus, cette large couverture spatiale augmente la probabilité que l'un des microphones de



(a) Topologie centralisée (b) Topologie toute connectée (c) Topologie en arbre

FIGURE 1.3. – Exemples de topologies d'AAAH.

l'AAAH soit proche de la source d'intérêt, et donc que l'algorithme de rehaussement de la parole dispose d'un enregistrement de bonne qualité de cette source.

De manière générale, le concept de *comportement émergent* a été introduit pour désigner le fait que l'utilité de l'ensemble des appareils est supérieure à la somme des utilités des appareils considérés séparément (Elson and Römer, 2003). Ce concept s'applique parfaitement au cas des AAAH, mais encore faut-il savoir exploiter la richesse des AAAH. C'est un des défis qu'aborde cette thèse.

Sous réserve que des algorithmes puissent tirer profit de la multiplication des appareils et des microphones présents dans les AAAH, rehausser la parole avec ces AAAH offrirait deux autres avantages. Le premier avantage est la disponibilité de ces antennes, que l'on n'a pas à acheter, ni même à apporter dans la scène où opère l'algorithme de rehaussement de la parole, puisqu'elles sont très probablement déjà présentes avec les appareils qui nous entourent. Cela mène au second avantage, celui de la flexibilité. En effet, il est facile de déplacer les appareils, mobiles par nature, pour les positionner à des endroits stratégiques, par exemple proches d'une source de bruit ou à proximité de la source d'intérêt. C'est d'ailleurs ce qui est fait par certaines entreprises d'aides auditives qui vendent des microphones distants à attacher sur les vêtements des locuteurs que l'on veut mieux entendre (Wolfe et al., 2021). En principe, il est possible de se passer de ces microphones distants en les remplaçant par le téléphone que nous portons très probablement avec nous.

Difficultés dans les antennes acoustiques ad-hoc Si la flexibilité est un atout des AAAH, pouvoir l'exploiter oppose de nombreuses difficultés. La première est que pour pouvoir fonctionner en toutes circonstances, un algorithme opérant dans une AAAH ne doit pas dépendre de la topologie de l'antenne. La topologie d'une antenne est déterminée par les positions des appareils et les connexions qui les relient entre eux. Différentes topologies sont représentées en figure 1.3. Un algorithme opérant dans une AAAH ne doit donc dépendre ni du nombre d'appareils constituant l'antenne, ni des liens reliant les différents appareils entre eux, ni de la position des appareils dans la scène acoustique. De plus, il est préférable de concevoir un algorithme dont la consommation en bande passante par appareil et en énergie n'augmente pas (trop) lorsqu'un appareil est ajouté à l'antenne. Enfin, la plupart des filtres spatiaux reposent sur un centre de fusion, qui reçoit et traite tous les signaux de l'antenne (typiquement le nœud central sur la figure 1.3(a)). Cela constitue une contrainte qui n'est pas viable dans une AAAH. En effet, le centre de fusion pourrait à tout moment disparaître, par exemple si l'utilisateur de cet appareil quitte la pièce,

ce qui mettrait à défaut le traitement de toute l'AAAH. Un des requis pour un algorithme de rehaussement de la parole dans une AAAH est donc qu'il ne repose pas sur la présence d'un centre de fusion.

Deux autres difficultés doivent être considérées lorsque l'on travaille dans des AAAH. La première est la question de la *synchronisation* des signaux. Les microphones, embarqués sur des appareils différents, ne partagent pas la même horloge d'échantillonnage. Les signaux qu'ils enregistrent n'ont donc pas la même référence temporelle et ne sont pas alignés dans le temps, ce qui peut dégrader les performances du rehaussement de la parole (Schmalenstroeer and Haeb-Umbach, 2018b). La deuxième difficulté a la même cause que l'asynchronisation des microphones : puisque les microphones d'appareils différents ne partagent pas la même implémentation matérielle, leur réponse à une onde sonore diffère, ce qui handicape le filtrage spatial (Oak and Kellermann, 2005). Les opérations pour égaliser les réponses des microphones sont appelées des opérations de *calibration*. Ce terme de calibration peut également décrire une autre nécessité dans les AAAH, celle de connaître les positions des microphones et des sources, car certains formateurs de voies reposent sur cette connaissance. Étant donné que les appareils de l'AAAH sont susceptibles de bouger, la calibration ne peut être déterminée à l'avance et doit pouvoir être effectuée tout au long de l'utilisation de l'AAAH. Notons que les problématiques de la calibration des microphones et de leur synchronisation sont étroitement liées, les solutions de calibration reposant souvent sur une bonne synchronisation des signaux (Plinge et al., 2016).

En conclusion de cette section, malgré de nombreux défis à relever pour pleinement exploiter leurs atouts, les AAAH offrent de nombreux avantages pour le rehaussement de la parole, en particulier de par la richesse d'information qu'elles enregistrent et par la flexibilité de leur utilisation. Différentes solutions proposées pour pallier les difficultés liées à l'utilisation des AAAH sont présentées en section 2.4.

1.3. Apprentissage profond pour le rehaussement de la parole

Des solutions existent déjà pour le rehaussement de la parole dans les AAAH (Bertrand and Moonen, 2010a; Tavakoli et al., 2017). Cependant, elles ne sont pas applicables directement car elles reposent sur la connaissance de paramètres dont on ne dispose pas en pratique, comme le moment ou l'endroit où la source cible est active. Si certaines solutions ont été proposées pour estimer ces paramètres (Martin, 2002; Gerkmann and Hendriks, 2011), l'apprentissage profond est aujourd'hui une des méthodes les plus efficaces pour ce faire et, plus généralement, pour effectuer du rehaussement de la parole.

Utilisation de l'apprentissage profond pour le rehaussement de la parole Les techniques d'apprentissage profond ont permis des avancées remarquables dans de nombreux domaines, en particulier dans le domaine du rehaussement de la parole. Dans le cas où seul un microphone enregistre la scène acoustique, les réseaux de neurones (RN) peuvent déterminer un masque qui, appliqué sur un mélange de parole et de bruit, atténue les zones où le bruit est présent et augmente la qualité ou l'intelligibilité du signal ainsi masqué (Narayanan and Wang, 2013; Weninger et al., 2014; Wang and Wang, 2018). Ils peuvent également servir à directement estimer le signal de parole à partir du mélange bruité (Xu et al., 2014; Park and Lee, 2016; Tan and Wang, 2019; Pandey and Wang, 2021). Les RN peuvent aussi être intégrés à des algorithmes de rehaussement de la parole, et prédire les paramètres requis par ces algorithmes (Zhang and Wu, 2012; Tu et al., 2019; Li et al., 2016; Wang et al., 2021a; Tammen and Doclo, 2021).
Bien qu'ayant montré des résultats convaincants, ces solutions n'utilisent qu'un seul signal, ce qui limite leurs performances et n'exploite pas l'information spatiale fournie par un appareil qui

possèderait plusieurs microphones (comme la plupart de nos téléphones, enceintes connectées, tablettes, etc. aujourd'hui). D'autres recherches ont montré que les RN pouvaient profiter de l'information spatiale enregistrée par plusieurs microphones d'un même appareil pour améliorer les performances de rehaussement de la parole (Jiang et al., 2014; Araki et al., 2015; Sainath et al., 2017; Perotin et al., 2018; Yoshioka et al., 2018; Chakrabarty and Habets, 2019; Liu et al., 2020).
Une présentation plus détaillée des solutions de rehaussement de la parole basées sur l'apprentissage profond est proposée en sections 2.2.2 et 2.3.2. Il est incontestable que le rehaussement de la parole a largement bénéficié de l'introduction de ce type d'algorithmes. Cependant, plusieurs facteurs limitent leur utilisation dans le contexte des AAAH, voire rendent leur utilisation impossible.

1.4. Limites des solutions existantes

Nous relevons plusieurs limites aux solutions de rehaussement de la parole aujourd'hui considérées comme faisant partie de l'état de l'art. Elles sont listées ici.

- *Connaissance oracle des paramètres*
  Une grande partie des solutions proposées pour le rehaussement de la parole supposent connues certaines grandeurs, comme la position de la source de parole ou les instants auxquels la source de parole est active. En pratique cependant, on ne dispose pas (toujours) de ces connaissances.
- *Information spatiale sous-exploitée*
  Même si, comme présenté au paragraphe précédent, de nombreuses propositions ont été faites pour exploiter l'information spatiale enregistrée par plusieurs microphones, celle-ci est très redondante car issue d'un même appareil. De rares solutions sont présentées spécifiquement pour les AAAH (Qian et al., 2018; Wang et al., 2020; Luo et al., 2020a), mais leurs approches sont centralisées.
- *Centralisation du calcul*
  On ne trouve encore que peu de recherches avec des RN qui exploitent toute l'information spatiale enregistrée par une AAAH. A notre connaissance, celles qui le font sont centralisées (Qian et al., 2018; Wang et al., 2020; Luo et al., 2020a), c'est-à-dire qu'elles fournissent tous les canaux disponibles au même RN. Au-delà des problèmes de passage à l'échelle (*scalability* en anglais), cela soulève des problèmes de dépendance à un centre de fusion que nous avons déjà évoqués et qui rendent ces solutions moins pertinentes. Les solutions distribuées de rehaussement de la parole existantes reposent quant à elles sur des connaissances oracles de certains paramètres (Bertrand and Moonen, 2010a; Zeng and Hendriks, 2014; Koutrouvelis et al., 2018) qui ne sont pas disponibles en pratique.
- *Complexité des RN*
  Si les performances de rehaussement de la parole augmentent grâce à l'introduction de techniques d'apprentissage profond, elles le sont à l'aide de RN de plus en plus complexes et comptant de plus en plus de paramètres (Subakan et al., 2021). Cela demande de grandes capacités de mémoire et de puissance de calcul (donc d'énergie pour un long temps d'utilisation) aux appareils sur lesquels ces RN sont actifs, ce que ne peuvent pas assurer les appareils de petites tailles des AAAH.
- *Spécialisation des RN*
  Pour qu'un système basé sur l'apprentissage profond soit performant, il est nécessaire de disposer de grands jeux de données d'entraînement. Celles-ci doivent également être variées

et représentatives de toutes les configurations dans lesquelles le système sera utilisé. Or des données utiles pour l'entraînement (composées d'entrées de bonne qualité et dont la vérité terrain est connue) sont difficiles et coûteuses à acquérir. Dans le cas où elles ne sont disponibles qu'en petite quantité, il est probable que le système ait de faibles performances ou qu'il se spécialise uniquement sur les configurations vues à l'entraînement, et donc qu'il généralise mal sur les autres.

- *Manque de transparence*
  Une des principales critiques adressées à l'apprentissage profond est le manque de transparence qu'il impose. La complexité des RN les transforme en « boîtes noires » dont il est souvent très difficile de comprendre la logique. Les résultats ne sont pas interprétables, il est donc difficile de s'y fier.

1.5. Contributions

Notre travail de recherche vise à concilier l'utilisation de RN avec celle d'AAAH pour le rehaussement de la parole. Conscients des limites décrites dans la section précédente, nous cherchons à les dépasser et à proposer un système de rehaussement de la parole qui opère dans les AAAH afin de profiter de leur flexibilité et étendue spatiale. Nos contributions sont donc les suivantes :

- *Estimation des paramètres de rehaussement de la parole*
  Nous estimons certains paramètres d'un algorithme de rehaussement de la parole à l'aide de RN. Différentes architectures de RN sont comparées et l'entraînement des RN est optimisé pour leur utilisation dans des AAAH.
- *Exploitation de toute l'information spatiale*
  Nous proposons un algorithme de rehaussement de la parole qui opère dans des AAAH, capable d'exploiter l'information spatiale enregistrée par tous ses appareils. Nous concevons une solution qui puisse utiliser efficacement cette information avec des RN dont les performances sont améliorées grâce à la richesse de l'information qui leur est fournie. Nous montrons par ailleurs que notre système permet non seulement de rehausser la parole (c'est-à-dire de restituer la parole d'un mélange bruité), mais aussi de séparer des sources de parole.
- *Distribution du calcul*
  Notre solution est une solution distribuée, qui ne repose sur aucun centre de fusion et limite les échanges entre les différents appareils d'une AAAH. Cela la rend flexible et diminue la charge de calcul sur chacun des appareils de l'AAAH.
- *Utilisation de RN simples*
  Afin d'estimer les paramètres de nos algorithmes, nous utilisons des RN qui permettent d'obtenir des performances comparables aux performances obtenues lorsqu'ils sont remplacés par des connaissances oracles. En fournissant toute l'information spatiale aux RN, nous leur permettons d'obtenir de bonnes performances sans qu'ils reposent sur une architecture complexe ; les RN que nous utilisons ont des architectures beaucoup plus petites et légères que celles de l'état de l'art, ce qui les rend plus aptes à opérer sur les appareils d'une AAAH.
- *Résilience*
  En évaluant notre système dans une grande variété de conditions acoustiques, nous montrons qu'il est résilient aux configurations acoustiques dans lesquelles il pourrait opérer. Nous l'évaluons dans des conditions propres aux AAAH, en particulier lorsque les signaux

ne sont pas synchrones. Enfin, bien que les données d'entraînement soient simulées, nous évaluons aussi les RN sur des données réelles.

- *Flexibilité de notre système*
  Plutôt que de reposer sur une approche de bout-en-bout, notre solution se base sur une approche classique du traitement du signal. Cela la rend plus flexible car il est possible d'ajuster les caractéristiques des signaux traités. Par ailleurs, cela facilite la tâche des RN qui ne sont plus qu'un maillon dans une chaîne de traitements, et qui peuvent donc reposer sur une architecture plus simple. Enfin, nous proposons une solution pour que les RN intégrés à notre système s'adaptent à différentes topologies d'AAAH, en particulier pour qu'ils puissent opérer dans des contextes avec un nombre variable d'appareils.
- *Interprétabilité des résultats*
  En plus de reposer sur une approche classique, qui permet de mieux contrôler le comportement de notre système, nous proposons des architectures de RN dont certains mécanismes sont interprétables, ce qui constitue un premier pas vers l'interprétabilité de l'ensemble du RN.
- *Mise à disposition de nos ressources*
  Des jeux de données et le code à la base de nos expériences ont été rendus publics. Cela participe à la reproductibilité de nos résultats et facilite l'approfondissement de nos travaux.

1.6. Plan de thèse

Cette thèse se divise en trois parties, elles-mêmes constituées de plusieurs chapitres.

La **partie I**, dont fait partie cette introduction, présente le contexte de la thèse.
Dans le **chapitre 2**, nous formulons la problématique en termes mathématiques et introduisons les notations utilisées au cours du document. Les solutions de rehaussement de la parole déjà existantes, opérant sur un ou plusieurs signaux d'enregistrement, sont décrites. Les solutions apportées aux défis propres aux AAAH sont également présentées, avant que ne soit abordée la question des métriques utilisées pour quantifier les résultats obtenus au cours de la thèse.

La **partie II** présente notre solution pour un rehaussement de la parole distribué opérationnel dans les AAAH.
Le **chapitre 3** présente notre système distribué de rehaussement de la parole, appelé Tango. Après une validation de son principe de fonctionnement sur un premier corpus qui montre l'intérêt d'opérer dans des AAAH, il est évalué sur un second corpus qui permet d'ajuster certains paramétrages d'entraînement et d'inférence. Le chapitre se termine par une comparaison avec une solution de l'état-de-l'art.
Le **chapitre 4** détaille les performances de Tango sur une grande variété de conditions acoustiques et met en valeur sa résilience. Tango y est également évalué sur des données réelles car toutes les évaluations précédentes avaient été effectuées sur données simulées. La dernière section du chapitre met en évidence comment Tango favorise la coopération des différents appareils d'une AAAH et en quoi il est capable d'exploiter la richesse de l'information spatiale enregistrée par ces différents appareils.

La **partie III** étend l'application de Tango à des conditions d'utilisation plus spécifiques aux AAAH.
Le **chapitre 5** analyse le comportement de Tango lorsque certains appareils d'une AAAH

disparaissent. Suite à cette analyse empirique, nous proposons une solution pour rendre notre système plus résilient aux situations où certains liens entre les appareils d'une AAAH sont rompus.

Le **chapitre 6** propose une évaluation de Tango lorsque les appareils d'une AAAH ne sont pas synchrones. En particulier, nous analysons l'influence de la dérive d'horloge et du décalage d'horloge à la fois sur le filtrage spatial et sur les performances des RN intégrés à notre système. Nous montrons que les RN que fait intervenir Tango sont sensibles aux décalages d'horloges et nous proposons une solution simple pour pallier la baisse de performances des RN.

Le **chapitre 7** montre que notre système peut être adapté à la séparation de sources simultanées. En exploitant la connaissance *a priori* de la configuration spatiale dans un contexte de réunion, nous concevons une solution qui fait transiter efficacement l'information spatiale entre les appareils afin de faciliter la tâche des RN dans le processus de séparation de sources. Nous étudions également la résilience de cette solution à un nombre variable de sources dans le même contexte de réunion.

Le chapitre 8 conclut cette thèse en résumant ses différentes contributions et en ouvrant sur différentes pistes de recherches inspirées par nos travaux.

1.7. Publications associées à la thèse

La plupart de nos contributions ont été publiées sous la forme d'un article de journal et de trois articles de conférence. Nous avons également rendue disponible une partie du code et des données ayant servi à nos expériences.

1.7.1. Article de revue

- Furnon Nicolas, Serizel Romain, Essid Slim, and Illina Irina. DNN-based mask estimation for distributed speech enhancement in spatially unconstrained microphone arrays. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29 (2021) : 2310-2323.

1.7.2. Articles de conférence

- Delebecque Louis, Serizel Romain, Furnon Nicolas. Towards an efficient computation of masks for multichannel speech enhancement. *Soumise à ICASSP 2022.*
- Furnon Nicolas, Serizel Romain, Essid Slim, and Illina Irina. Attention-based distributed speech enhancement for unconstrained microphone arrays with varying number of nodes. In *2021 29th European Signal Processing Conference (EUSIPCO).*
- Furnon Nicolas, Serizel Romain, Illina Irina, and Essid Slim. Distributed speech separation in spatially unconstrained microphone arrays. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4490-4494). IEEE.
- Furnon Nicolas, Serizel Romain, Illina Irina, and Essid Slim. DNN-based distributed multichannel mask estimation for speech enhancement in microphone arrays. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4672-4676). IEEE.

1.7.3. Logiciels et jeux de données

- Corpus DISCO : Le code pour le jeu de données DISCO (DIstributed Semi-COnstrained microphone array) est disponible à l'adresse https://github.com/nfurnon/disco/tree/master/dataset_generation/gen_disco.
- Corpus MEETIT : Le code pour le jeu de données MEETIT (MEETing InTerferences) est disponible à l'adresse https://github.com/nfurnon/disco/tree/master/dataset_generation/gen_meetit.
- Bruits du corpus DISCO : Les bruits téléchargés de Freesound puis traités pour créer les mélanges du corpus DISCO sont disponibles à l'adresse https://zenodo.org/record/4019030
- Poids de RN : Les poids des RN utilisés pour notre article de revue sont disponibles à l'adresse https://zenodo.org/record/4019041

2. Contexte et état de l'art

Ce chapitre formalise le contexte scientifique du rehaussement de la parole en présentant les notations utilisées par la suite dans la section 2.1. Le reste du chapitre rend compte de différentes techniques de rehaussement de la parole, en les catégorisant selon le nombre de microphones qu'ils utilisent. Ainsi, la section 2.2 décrit les techniques de rehaussement de la parole à partir d'un seul enregistrement. La section 2.3 décrit les techniques de rehaussement de la parole lorsque plusieurs signaux sont enregistrés par un même appareil (ou un même nœud). Enfin, la section 2.4 présente les techniques de rehaussement de la parole dans les antennes acoustiques ad-hoc, c'est-à-dire lorsque plusieurs nœuds contenant plusieurs microphones enregistrent la scène.

2.1. Formulation du problème et notations

On considère une scène acoustique dans un environnement fermé, où les sons émis par les différentes sources sont réverbérés sur les murs de la pièce.

2.1.1. Cas mono-canal

Soit $y(\mathfrak{t})$ la variation de pression mesurée par un microphone à l'instant $\mathfrak{t}$. Dans le cas où seule une source est active, son signal se propage dans la pièce en se réverbérant sur les différent murs avant d'arriver sur le microphone, ce qui se traduit mathématiquement par une convolution :

$$y(\mathfrak{t}) = (c * h)(\mathfrak{t}) \tag{2.1}$$

où c est le son émis par la source, h est la réponse impulsionnelle (RI) de la pièce et où le signe $*$ décrit la convolution. La convolution dans le domaine temporel des deux fonctions c et h est exprimée dans l'équation (2.2) :

$$(c * h)(\mathfrak{t}) = \int_{-\infty}^{\infty} c(\tau) \cdot h(\mathfrak{t} - \tau) d\tau \,. \tag{2.2}$$

En pratique, un microphone ne mesure les variations de pression qu'à des instants quantifiés t, déterminés par la fréquence d'échantillonnage d'une horloge. Ainsi dans la suite de cette thèse, les signaux ne sont pas des variables continues, mais discrètes et on peut réécrire l'équation (2.2) :

$$(c * h)(t) = \sum_{\tau=0}^{\infty} c(\tau) \cdot h(t - \tau) \,. \tag{2.3}$$

La RI d'une pièce caractérise son comportement acoustique à toutes les fréquences pour des positions de source et de microphone données. Si la réverbération d'une pièce n'est pas trop forte, elle peut être bénéfique à la compréhension de la parole de par la redondance qu'apportent les échos (Moore, 2012, 6ème édition). Puisque dans cette thèse nous nous intéressons essentiellement à des situations où la réverbération des pièces (bureaux, salons, etc) n'est pas trop élevée, nous ne chercherons pas à annuler les effets de la réverbération, bien que cela soit parfois désirable

(Nakatani et al., 2010; Carbajal et al., 2020) voire nécessaire dans les cas où la réverbération est forte (Kinoshita et al., 2009).

Si N sources sont présentes, leurs contributions s'ajoutent, si bien que le signal mesuré par le microphone est donné par :

$$y(t) = \sum_{j=1}^{N} x_j(t)\,, \tag{2.4}$$

où $x_j(t)$ est le signal réverbéré de la source $j \in [\![1; N]\!]$. Dans le cas particulier où seules deux sources sont présentes, l'une de parole et l'autre de bruit, on peut alors représenter le signal mesuré par

$$y(t) = s(t) + n(t)\,, \tag{2.5}$$

où $s(t)$ est l'image réverbérée du signal de parole et $n(t)$ celle du signal de bruit.

Il peut être intéressant d'appliquer une transformation à y pour le projeter dans un espace où certaines grandeurs sont plus facilement analysables. Bien que différentes projections puissent être utilisées dans le contexte du rehaussement de la parole, comme la transformation en ondelettes (Seok and Bae, 1997; Baugé et al., 2013; Andén et al., 2015; Dash et al., 2021) ou la décomposition en modes empiriques (Huang et al., 1998; Flandrin et al., 2004; Zão et al., 2014), l'essentiel des travaux sur le traitement du signal audio repose sur une analyse dans le domaine fréquentiel. La transformation vers le domaine fréquentiel la plus couramment utilisée est la transformation de Fourier discrète, qui représente le signal temporel sur des bandes de fréquences dont les centres sont espacés linéairement. Pour des signaux discrets, elle se calcule selon l'équation suivante :

$$\mathcal{F}(y)(f) = \sum_{t=0}^{\mathcal{T}-1} e^{-2\iota\pi f \frac{t}{\mathcal{T}}} y(t)\,, \tag{2.6}$$

avec ι le nombre imaginaire tel que $\iota^2 = -1$, f la variable de fréquence et $\mathcal{T}$ le nombre d'échantillons dans y. Puisque seule l'évolution fréquentielle du signal y est caractérisée dans l'équation (2.6), on utilise la transformée de Fourier à court terme (TFCT) du signal pour représenter à la fois le contenu fréquentiel et le contenu temporel des signaux. Le signal est divisé en courtes trames successives de longueur T, et séparées de H échantillons. H est appelé *pas d'avancement* et est souvent pris inférieur à T, si bien que deux trames successives sont partiellement superposées. Chaque trame est multipliée par une fenêtre $v(t)$ avant d'être transformée dans le domaine de Fourier. La TFCT d'un signal dépend ainsi de deux variables, temporelle et fréquentielle :

$$\mathrm{TFCT}(y)(f, l) = \sum_{\kappa=0}^{T-1} y(lH + \kappa) \cdot v(\kappa) e^{-2\iota\pi f \frac{\kappa}{T}}\,. \tag{2.7}$$

Plutôt que de garder une répartition linéaire des bandes de fréquences, il est possible de les regrouper en des plages de fréquences dont la largeur dépend de la fréquence centrale de la plage. Il existe par exemple l'échelle de Bark (Zwicker, 1961), l'échelle à largeur de bande rectangulaire équivalente (Moore and Glasberg, 1996) ou l'échelle de Mel (Stevens et al., 1937; Logan, 2000) qui toutes s'inspirent du système auditif humain.

Afin d'alléger les formulations, nous utiliserons par la suite la même notation pour désigner un signal à la fois dans le domaine temporel et dans le domaine de la TFCT. Là où il n'y aura pas de confusion, la dépendance aux variables t et f sera abandonnée, toujours dans un souci de concision. Ainsi, le scalaire (complexe) y désigne le contenu du mélange à la fréquence f et à la trame t.

Différentes solutions pour restituer la meilleure estimation possible de la parole à partir de la seule observation y sont présentées en section 2.2.

2.1.2. Cas multicanal

Lorsque M microphones différents, intégrés à la même antenne acoustique, enregistrent la scène acoustique, ils fournissent un signal dit *multicanal*, que l'on regroupe dans le vecteur :

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \\ \vdots \\ y_M \end{bmatrix} . \tag{2.8}$$

où y_m est l'enregistrement du m-ème microphone. Par la suite, les lettres minuscules en gras représenteront des vecteurs, et les majuscules en gras représenteront des matrices.

Le rehaussement de la parole multicanal se base sur l'utilisation de filtres, appelés *formateurs de voies* (*beamformer* en anglais). Le signal filtré résulte d'une convolution entre le filtre $\mathbf{w}$ et le signal temporel $\mathbf{y}$, ce qu'on peut plus simplement calculer, sous l'approximation de bande étroite (Kowalski et al., 2010), par un produit scalaire dans le domaine complexe :

$$y_{filt} = \mathbf{w}^H \mathbf{y} \tag{2.9}$$

$$= \sum_{m=1}^{M} w_m^* \cdot y_m \, . \tag{2.10}$$

L'exposant $\cdot^H$ désigne la transposée hermitienne (la transposée du complexe conjugué) et l'exposant $\cdot^*$ désigne le conjugué d'un scalaire complexe. Les variables $\mathbf{w}$ et $\mathbf{y}$ étant complexes, le filtre $\mathbf{w}$ modifie l'amplitude et la phase de chaque canal dans $\mathbf{y}$ avant de les sommer. Différents types de formateurs de voies sont décrits dans la section 2.3.1.

La formation de voies peut plus généralement être appelée *filtrage spatial*, car elle exploite l'information spatiale, disponible par le fait que plusieurs microphones sont présents en des endroits distincts de l'espace. On peut quantifier l'information spatiale à l'aide des matrices d'autocorrélation spatiale (MAS). Celle de $\mathbf{y}$ dans l'équation (2.8) se calcule par :

$$\mathbf{R}_{\mathrm{y}} = \mathbb{E}\{\mathbf{y}\mathbf{y}^H\} \, , \tag{2.11}$$

avec $\mathbb{E}\{\cdot\}$ l'opérateur d'espérance. $\mathbf{R}_{\mathrm{y}}$ peut également être appelée *matrice de covariance* si $\mathbf{y}$ est d'espérance nulle. La MAS décrit la corrélation entre les différents signaux de $\mathbf{y}$. On notera de même $\mathbf{R}_{\mathrm{s}}$ et $\mathbf{R}_{\mathrm{n}}$ les MAS de la parole et du bruit respectivement. Dans le cas de corrélations croisées entre deux signaux, on parle de matrice d'intercorrélation spatiale. Par exemple, la matrice d'intercorrélation spatiale entre $\mathbf{y}$ et $\mathbf{s}$ est notée :

$$\mathbf{R}_{\mathrm{ys}} = \mathbb{E}\{\mathbf{y}\mathbf{s}^H\} \, . \tag{2.12}$$

2.1.3. Cas multinœud

Plusieurs antennes acoustiques peuvent être présentes dans une salle. Soit K le nombre de ces antennes, que l'on peut également appeler *nœuds*. Chacune compte M_k microphones et mesure un signal $\mathbf{y}_k$ similaire à l'équation (2.8) :

$$\mathbf{y}_k = \begin{bmatrix} y_{k,1} \\ y_{k,2} \\ \vdots \\ y_{k,M_k} \end{bmatrix} . \tag{2.13}$$

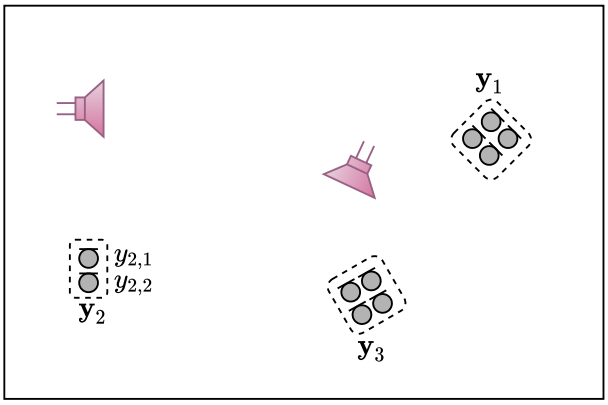


FIGURE 2.1. – Illustration d'une antenne de microphones avec $K = 3$ nœuds et deux sources.

On rassemble alors tous ces signaux en un signal $\mathbf{y}$ dit multinœud :

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_K \end{bmatrix}. \tag{2.14}$$

La figure 2.1 représente une antenne de microphones avec $K = 3$ nœuds et deux sources. Dans cet exemple, $M_1 = 2$, $M_2 = 4$ et $M_3 = 4$.

2.2. Rehaussement de la parole mono-canal

Cette section présente les solutions de rehaussement de la parole mono-canal, c'est-à-dire les solutions qui n'utilisent qu'un seul signal pour estimer le signal cible qu'il contient. Nous ferons la distinction entre les approches dites « classiques » (section 2.2.1) et celles basées sur l'apprentissage profond (section 2.2.2). Les approches classiques furent les premières à être utilisées et reposent sur des modèles mathématiques dont les paramètres ne nécessitent pas d'étape préalable dite d'apprentissage. Les méthodes basées sur l'apprentissage profond, au contraire, reposent sur une modélisation des signaux qui nécessite un apprentissage sur des jeux de données. Leur forte capacité de modélisation a permis de grandes améliorations du rehaussement de la parole dans les dernières années et ce sont ces méthodes qui sont essentiellement utilisées actuellement, quoique souvent en combinaison avec les approches classiques.

2.2.1. Approches classiques

Soustraction spectrale La méthode la plus intuitive pour restituer la parole est de soustraire le bruit n au mélange y dans l'équation (2.5) (Boll, 1979; Berouti et al., 1979; Lim, 1978; Martin, 1994; Yoshioka et al., 2009). Puisque la transformée de Fourier est linéaire, soustraire le bruit dans le domaine temporel revient à le soustraire dans le domaine spectral. Par extension, de par les avantages de la TFCT, la soustraction se fait souvent dans le plan temps-fréquence (TF). Bien que simple, cette méthode souffre de plusieurs inconvénients. Tout d'abord, si le spectre du signal mesuré est bien la somme des spectres de la parole et du bruit, cela n'est pas le cas de leurs amplitudes (c'est-à-dire le module du spectre). Or la plupart des applications de cette méthode soustraient l'amplitude du bruit, voire la moyenne de l'amplitude du bruit sur une

plage temporelle donnée (Boll, 1979), ce qui ne permet pas une restitution parfaite du signal de parole et n'augmente pas forcément l'intelligibilité de la parole (Lim, 1978) voire amène des dégradations indésirables (Vaseghi, 2008, Chapitre 11). La deuxième limite de cette méthode est que le bruit ne peut être estimé que lorsque la source de parole est inactive. On suppose donc que la source de bruit est stationnaire pour pouvoir soustraire l'estimation du bruit sur les périodes où les deux sources sont actives (Yoshioka et al., 2009). Cette hypothèse est très forte et fausse pour de nombreux bruits quotidiens, comme le cliquetis d'un clavier, l'interférence de plusieurs interlocuteurs ou les aboiements d'un chien.

Estimation des paramètres nécessaires à la soustraction spectrale La plupart des méthodes citées précédemment reposent sur une estimation fiable soit du bruit, soit du rapport signal-à-bruit (RSB, voir section 2.6.1). Plus que la méthode elle-même, c'est cette estimation du bruit qui est cruciale à la qualité du rehaussement de la parole. Ephraim et Malah proposent une méthode qui estime le RSB de manière récursive (Ephraim and Malah, 1984), ce qui permet de limiter le bruit musical (Scalart et al., 1996). Un détecteur d'activité vocale (DAV), indiquant les périodes où seul le bruit est présent, et celles où le bruit et la parole sont actifs, permet d'avoir une estimation précise du bruit, mais à condition qu'il soit stationnaire (Haigh and Mason, 1993; Tanyer and Ozer, 2000; Sharma et al., 2021). Dans le cas où le bruit est moins stationnaire, Martin propose une méthode qui estime la puissance du bruit en suivant les valeurs minimales de la puissance du mélange lissé (Martin, 1994). En se basant sur cette technique mais en la combinant avec une moyenne récursive, Cohen développe un algorithme appelé IMCRA (*improved minima controlled recursive averaging*, moyenne récursive contrôlée par les minimums améliorée), qu'il montre plus robuste dans les conditions de faible RSB (Cohen, 2003). Des estimations basées sur la moindre erreur quadratique moyenne permettent également d'alléger les calculs tout en augmentant les performances (Hendriks et al., 2010; Gerkmann and Hendriks, 2011).

Estimations statistiques de la parole et du bruit Afin d'éviter les distorsions apportées par la soustraction spectrale du bruit, il est possible de directement estimer le signal de la parole. Ainsi, McAulay et Malpass estiment la parole en maximisant la probabilité d'observer le signal mesuré étant donné la parole estimée (McAulay and Malpass, 1980). Ephraim et Malah proposent d'estimer l'amplitude de la parole dans le domaine TF en minimisant l'erreur quadratique moyenne entre la densité de probabilité des amplitudes estimée et réelle. Pour cela, ils supposent que les densités de probabilité des coefficients du bruit et de la parole suivent des distributions gaussiennes (Ephraim and Malah, 1984, 1985). Les coefficients TF de la parole peuvent également être estimés à l'aide de modèles de Markov cachés, ce qui permet une meilleure réduction des bruits non stationnaires (Varga and Moore, 1990; Ephraim and Van Trees, 1995; Mohammadiha et al., 2013). D'autres travaux font l'hypothèse que les coefficients TF des signaux suivent d'autres lois de distribution pour estimer la parole au sens des moindres carrés (Martin, 2002; Erkelens et al., 2007; Mohammadiha et al., 2013), afin de traduire plus fidèlement la distribution de la parole qui n'est pas aussi stationnaire que ce que modélise une distribution gaussienne. Enfin, une modélisation non pas des densités moyennes mais des densités de probabilités *a posteriori* peut être recherchée, ce qui peut être plus simple dans certaines conditions (Loizou, 2007, 2$^{\text{ème}}$ édition) ou mener à de meilleurs résultats (Lotter and Vary, 2005; Wolfe and Godsill, 2003).

Masquage spectral Le principe du masquage spectral repose sur un effet psychoacoustique selon lequel dans des mélanges, les régions TF sont souvent dominées par une des sources qui est la seule entendue (Moore, 2012, 6$^{\text{ème}}$ édition). Lorsque cette source est celle du bruit, il convient

donc de masquer la zone TF correspondante afin de minimiser la contribution du bruit dans le signal global. Pour cela, on applique un gain sur chaque point TF du spectre bruité (Virag, 1999; Hu and Wang, 2004; Wang, 2005; Jensen and Hendriks, 2011) :

$$\hat{s}(t, f) = m(t, f) \cdot y(t, f)\,. \tag{2.15}$$

$m(t, f)$ est la valeur du masque à la trame t et à la fréquence f. Le masquage spectral a longtemps été considéré comme la meilleure méthode pour le rehaussement de la parole dans le contexte de l'analyse de scènes acoustiques (Wang, 2005; Hummersone et al., 2014). En général, les valeurs du masque sont comprises entre 0 et 1 afin d'atténuer la contribution des points TF dans lesquels le bruit est dominant, en ne restituant que les zones dans lesquelles la parole domine. Lorsque le masque est contraint à ne prendre comme valeur que 0 ou 1, on parle de masque binaire (Wang, 2005). Bien que ces masques demandent moins de mémoire pour être sauvegardés et permettent des calculs plus efficaces, de nombreuses études ont montré qu'ils n'étaient pas idéaux en terme d'intelligibilité de la parole car ils apportent ce qu'on appelle du bruit musical (Jensen and Hendriks, 2011; Liang et al., 2014; Stöter et al., 2018; Chakrabarty and Habets, 2019). D'autres types de masques ont donc été proposés comme cibles, comme le masque de Wiener idéal (Liutkus and Badeau, 2015), le masque de ratio idéal (MRI) (Srinivasan et al., 2006) et le masque de ratio optimal (Liang et al., 2014).

Projection des différentes composantes du signal dans des espaces distincts Plutôt que de supposer que le mélange mesuré est la somme du bruit et de la parole, on peut supposer qu'il existe un espace dans lequel ces deux composantes sont disjointes. En projetant le signal bruité dans les espaces (orthogonaux) de la parole et du bruit, on peut alors ne sélectionner que la parole (Ephraim and Van Trees, 1995). Différents outils d'algèbre linéaires permettent une telle approche, comme la décomposition en valeurs singulières lorsqu'on considère les signaux temporels (Dendrinos et al., 1991; Jensen et al., 1995). La factorisation en matrices non-négatives est une autre technique qui décompose le signal bruité en un produit de matrices non-négatives (Lee and Seung, 1999). Puisque l'amplitude du signal dans le domaine de la TFCT est une matrice positive, il est possible de la factoriser en un produit entre une matrice dite de bases et une matrice dite d'activations. N'activer que les vecteurs de la matrice de bases correspondant aux composantes du signal cible permet de débruiter le signal. La factorisation en matrices positives est très populaire dans le domaine plus général de la séparation de sources, en particulier pour des applications musicales (Cho et al., 2003; Virtanen, 2007; Févotte et al., 2009; Nakano et al., 2010)

Approches multitrames et multibandes A cheval entre les traitements mono-canaux et multicanaux, les approches multitrames et multibandes appliquent des filtres multicanaux au signal d'un seul microphone. Pour chaque point TF, ils considèrent le signal mesuré sur plusieurs trames ou pour plusieurs bandes de fréquences, ce qui permet d'exploiter plus d'information spectrale ou temporelle. Ainsi, Benesty et Huang exploitent la corrélation inter-trames d'un signal pour construire un formateur de voies à variance minimale et réponse sans distorsion (MVDR : *minimum variance distortionless response*) (Benesty and Huang, 2011). Andersen et al. estiment la parole en équilibrant la distorsion de la parole et la réduction du bruit (Andersen and Moonen, 2017). Ranjbaryan et Abutalebi utilisent l'information passée pour mieux estimer la densité de probabilité *a posteriori* (Ranjbaryan and Abutalebi, 2021). De la même manière, l'information inter-bandes peut être exploitée pour améliorer les qualités du filtrage (Avargel and Cohen, 2007; Huang et al., 2014). Ces types de filtres permettent par ailleurs de limiter les distorsions consé-

quentes au fait que l'hypothèse de bande étroite n'est pas vérifiée (Vincent et al., 2018).

Si les méthodes mono-canales décrites dans cette section ont permis de considérables avancées dans le domaine du rehaussement de la parole, elles sont souvent mises à mal lorsque le bruit n'est pas stationnaire. Par ailleurs, la plupart d'entre elles négligent l'information apportée par la phase, bien que certaines approches la prennent explicitement en compte, en particulier celles basées sur la factorisation en matrices positives (Parry and Essa, 2007; Badeau, 2011). Ces deux limites sont repoussées par les travaux de recherche plus récents basés sur l'apprentissage profond.

2.2.2. Approches basées sur l'apprentissage profond

L'utilisation de données préalables pour paramétrer un modèle a déjà été présentée il y a longtemps, par exemple par Tamura et Waibel en 1988 (Tamura and Waibel, 1988), ou par Xie et Van Compernolle en 1996 (Xie and Van Compernolle, 1996). Il a néanmoins fallu attendre que les processeurs de calcul soient plus puissants, et que de plus grandes bases de données soient disponibles, pour que l'utilisation de l'apprentissage profond devienne prépondérante dans le domaine du rehaussement de la parole.

Prédiction de masques TF et de spectrogrammes Une des premières utilisations des réseaux de neurones pour le rehaussement de la parole fut de prédire des masques TF à appliquer sur le spectrogramme du signal bruité (Narayanan and Wang, 2013; Weninger et al., 2014; Kolbæk et al., 2017; Liang et al., 2020; Cui and Bao, 2021). La phase peut également être explicitement considérée afin d'affiner le rehaussement par l'utilisation du masque de ratio idéal complexe (Williamson et al., 2016; Xia et al., 2017; Hu et al., 2020) ou le masque idéal sensible à la phase (Erdogan et al., 2015). Assez logiquement, les réseaux de neurones ont par la suite été utilisés pour prédire directement soit l'amplitude de la parole (Xu et al., 2014; Park and Lee, 2016; Fu et al., 2017; Tan and Wang, 2018), soit son spectre (c'est-à-dire partie réelle et imaginaire conjointement, ou alors l'amplitude et la phase conjointement) (Tan and Wang, 2019; Pandey and Wang, 2019a; Koizumi et al., 2020; Pandey and Wang, 2021). Ce sont des tâches plus compliquées mais rendues possibles par des réseaux plus complexes capables également d'exploiter l'information de la phase.

Prédiction des paramètres requis par les algorithmes classiques Plutôt que d'utiliser un masque TF comme filtre, ce qui peut apporter des distorsions, il est également possible d'exploiter la puissance de modélisation des réseaux de neurones pour prédire les paramètres internes des algorithmes classiques présentés dans la section précédente (2.2.1). Ces approches ont l'avantage de nécessiter des réseaux souvent plus simples et de profiter de l'efficacité des algorithmes classiques. Ainsi, des réseaux de neurones sont utilisés pour estimer le DAV (Zhang and Wu, 2012; Dinkel et al., 2021; Giri et al., 2021), pour l'algorithme IMCRA (Tu et al., 2019; Nian et al., 2021), pour le rehaussement de la parole par approche des moindres erreurs quadratiques moyennes (Li et al., 2016; Nicolson and Paliwal, 2019), pour l'estimation du gain de Wiener (Wang et al., 2021a) ou pour le filtre de Kalman (Roy et al., 2021). De même, Tammen et Doclo ont combiné l'efficacité du filtrage MVDR multitrame à la puissance de modélisation des réseaux de neurones (Tammen and Doclo, 2021). Enfin, des auto-encodeurs peuvent être utilisés pour modéliser la variance de la parole et la reconstruire ensuite par un filtre de Wiener, le bruit étant estimé par une factorisation en matrices positives (Leglaive et al., 2018; Pariente et al., 2019).

Rehaussement de bout-en-bout Bien que certains des modèles précédents soient explicitement capables d'exploiter la phase du signal bruité, ils dépendent de la représentation intermédiaire fournie au réseau, le plus souvent la TFCT du signal temporel. Les derniers travaux de recherche tendent à ne plus dépendre de cette représentation intermédiaire, et à directement fournir la forme d'onde du signal bruité, et de restituer le signal de parole. On dit de ces approches qu'elles sont de bout-en-bout et elles permettent de s'assurer que la phase du signal est vue (Rethage et al., 2018), et de laisser au réseau le soin de modéliser le signal par une autre représentation que la transformée de Fourier (Luo and Mesgarani, 2019). Certaines de ces approches de bout-en-bout reposent sur un masquage implicite où les signaux passent par un encodeur, avant d'être masqués puis décodés (Luo and Mesgarani, 2019; Pandey and Wang, 2019b; Takeuchi et al., 2020; Hu et al., 2020). Néanmoins, d'autres méthodes modélisent directement la parole débruitée, souvent à l'aide de réseaux antagonistes génératifs (Goodfellow et al., 2014; Pascual et al., 2017; Qin and Jiang, 2018) bien que cela ne soit pas indispensable et que d'autres architectures aient également été proposées (Rethage et al., 2018; Macartney and Weyde, 2018).

Dépendance aux données Le principal inconvénient des approches supervisées basées sur l'apprentissage profond est qu'elles nécessitent des données étiquetées, c'est-à-dire dont la parole et le bruit sont disponibles séparément. Obtenir de telles données réelles est coûteux et il est difficile d'en obtenir de grands corpus d'apprentissage. Les réseaux de neurones (RN) entraînés sur de tels corpus risquent donc de mal généraliser sur d'autres données, car il est important d'avoir une grande variété de données d'apprentissage pour une bonne généralisation (Kolbæk et al., 2016; Maciejewski et al., 2018; Kadıoğlu et al., 2020). Pour pallier cette difficulté, il est possible de simuler les données d'apprentissage, en mélangeant *a posteriori* des signaux enregistrés séparément. L'avantage est qu'il est beaucoup plus facile de créer de grands jeux de données où l'on dispose de la vérité terrain. L'inconvénient est que les RN entraînés sur les données simulées risquent d'avoir de faibles performances sur les données réelles. Toutefois, il a été montré que ce risque est limité et que cette méthode d'apprentissage peut être suffisante pour de bonnes performances sur données réelles (Perotin et al., 2019; Taherian et al., 2020; Llombart et al., 2021). C'est pourquoi c'est la méthode que nous retiendrons au cours de cette thèse.

Une autre approche est celle de l'apprentissage non supervisé, qui ne nécessite pas de données étiquetées. Cela peut passer soit par l'utilisation d'une architecture de réseau adaptée à ce genre de données, par exemple un réseau antagoniste génératif (Higuchi et al., 2017) ou un auto-encodeur (Bie et al., 2021; Neri et al., 2021). Il est sinon possible de chercher à séparer des mélanges de mélanges et de forcer le débruitage de ces mélanges par la même opération (Wisdom et al., 2020; Saito et al., 2021; Fujimura et al., 2021). Nous ne développerons cependant pas plus ces méthodes car elles ne seront pas utilisées au cours de la thèse.

2.3. Rehaussement de la parole multicanal

Il a été largement (dé)montré que disposer de plusieurs microphones permet d'augmenter les performances du rehaussement de la parole (Van Compernolle et al., 1990; Meyer and Simmer, 1997; Souden et al., 2011; Corey et al., 2019; Ceolini et al., 2020) grâce à l'information spatiale que contiennent les enregistrements et au filtrage spatial qui amène moins de distorsion. Dans cette section, on suppose que plusieurs microphones intégrés dans la même antenne enregistrent le signal $\mathbf{y} = \mathbf{s}+\mathbf{n}$ avec $\mathbf{s}$ les composantes de parole dans $\mathbf{y}$ et $\mathbf{n}$ celles de tous les bruits indésirables. Puisque tous les microphones appartiennent à la même antenne, on peut raisonnablement les supposer synchronisés. Enfin, on suppose que les différentes sources acoustiques sont en champ lointain, c'est-à-dire que la distance entre les sources et les microphones est largement supérieure

à la taille caractéristique de l'antenne.

2.3.1. Approches classiques

Les filtres multicanaux les plus couramment utilisés sont présentés dans les paragraphes suivants. Sans perte de généralité, on considérera toujours le premier microphone de chaque antenne comme le microphone dit de référence, bien que cela ait fait l'objet de nombreuses discussions (Lawin-Ore and Doclo, 2012; Araki et al., 2018; Zhang et al., 2020).

2.3.1.1. Les filtres à contraintes linéaires

Formateur retardateur-sommateur Le formateur de voies le plus simple est sans doute le formateur retardateur-sommateur (DSB : *delay-and-sum beamformer*) (Van Veen and Buckley, 1988; Vincent et al., 2018). Ce filtre déphase les signaux afin d'aligner les composantes venant d'une direction donnée (celle de la source cible) et fait la moyenne de tous les signaux ainsi décalés, ce qui rehausse les composantes de la cible (par interférences constructives) et diminue relativement les autres composantes (supposées de moyenne nulle).

En l'absence de réverbération, le signal de la source cible mesuré par les différents microphones est proportionnel au vecteur $\mathbf{d}$ de l'équation (2.16), appelé *vecteur d'orientation*, ou encore fonction de transfert acoustique. Le filtre DSB revient simplement à multiplier les signaux mesurés par le conjugué du vecteur d'orientation, comme dans l'équation (2.17).

$$\mathbf{d}(f) = \left[1, \quad e^{-2\imath\pi f\tau_1}, \quad \cdots, \quad e^{-2(M-1)\imath\pi f\tau_{M-1}}\right]^T \tag{2.16}$$

$$\hat{s} = \mathbf{d}^H\mathbf{y}\,, \tag{2.17}$$

avec τ_i le délai qu'il a fallu à l'onde sonore pour atteindre le microphone i après avoir atteint le microphone de référence.

Une illustration de cette méthode pour une antenne linéaire uniforme est donnée en Figure 2.2, dont les $M = 3$ microphones sont espacés d'une distance l, lorsque la source cible est située à un angle d'arrivée θ.

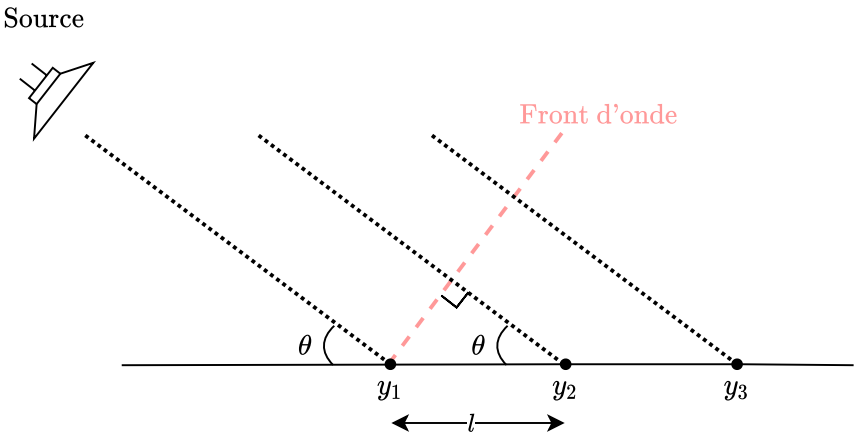
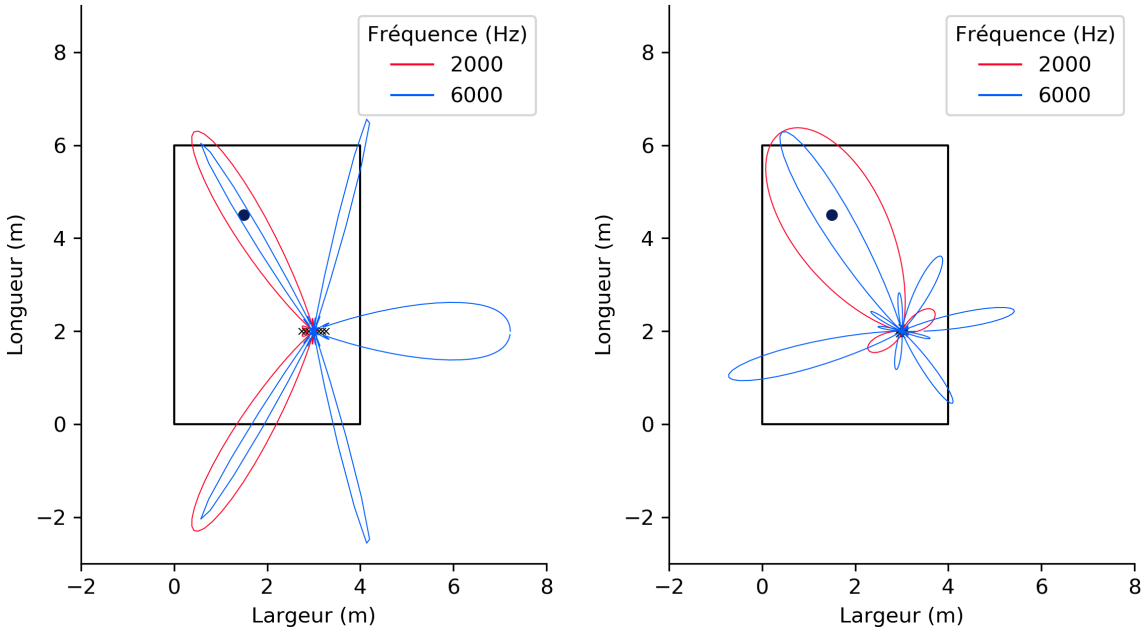


FIGURE 2.2. – Représentation d'une antenne de microphones linéaire et uniforme. Pour une source émettant à un angle θ et des microphones espacés de l, le temps d'arrivée relatif entre le microphone 1 et le microphone i est $\tau_i = i\frac{l\cos(\theta)}{c}$ avec c la célérité du son dans l'air.

Il est toutefois possible d'utiliser des antennes non-linéaires. Des antennes circulaires uniformes par exemple permettent de supprimer l'invariance par rotation autour de l'axe de l'antenne

linéaire (Taylor, 1952; Meyer, 2001). Pour prendre en compte à la fois l'angle d'azimuth et d'élévation, une antenne à 2 dimensions (une maille de microphones) (Flanagan et al., 1985), ou 3 dimensions (une sphère de microphones) doit être utilisée (Rafaely et al., 2010). La figure 2.3 illustre la différence de motifs de formateurs de voies entre une antenne linéaire uniforme de 8 microphones et une antenne circulaire uniforme de 8 microphones, aux fréquences de 2000 Hz et 6000 Hz. Dans les cas de l'antenne linéaire, les microphones sont espacés de $l = 7,5$ cm. Dans le cas de l'antenne circulaire, les microphones sont à 7,5 cm du centre de l'antenne. On peut y voir que le formateur de voies sur l'antenne linéaire a un motif symétrique par rapport à l'axe de l'antenne, qui fait rehausser les sons venant d'une direction indésirable.



(a) Antenne linéaire (b) Antenne circulaire

FIGURE 2.3. – Motifs des faisceaux d'un formateur retardateur-sommateur sur deux antennes de microphones uniformes, l'une linéaire et l'autre circulaire, aux fréquences 2000 Hz et 6000 Hz.

Filtres à puissance ou variance minimale et réponse sans distorsion Le filtre à puissance minimale et réponse sans distorsion (MPDR : *minimum power distortionless response*) (Capon, 1969; Van Trees, 2004) minimise la puissance du signal filtré $\mathbf{w}_{\text{MPDR}}^H \mathbf{y}$ sous la contrainte que la parole filtrée ne doit pas être distordue. Puisque le signal de la parole suit le chemin caractérisé par le vecteur d'orientation $\mathbf{d}$, cela se traduit par

$$\mathbf{w}_{\text{MPDR}} = \underset{\mathbf{w}}{\text{argmin}}\, \mathbb{E}\{||(\mathbf{w}^H\mathbf{y})(\mathbf{w}^H\mathbf{y})^H||^2\} \quad \text{tel que} \quad \mathbf{w}^H\mathbf{d} = 1 \tag{2.18}$$

$$= \underset{\mathbf{w}}{\text{argmin}}\, \mathbf{w}^H\mathbf{R}_{\text{y}}\mathbf{w} \quad \text{tel que} \quad \mathbf{w}^H\mathbf{d} = 1\,. \tag{2.19}$$

La solution de l'équation (2.19) est donnée par :

$$\mathbf{w}_{\text{MPDR}} = \frac{\mathbf{R}_{\text{y}}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{R}_{\text{y}}^{-1}\mathbf{d}}\,. \tag{2.20}$$

Pour améliorer les performances de réduction de bruit de ce filtre, il est possible de minimiser non pas la puissance du signal filtré, mais celle du bruit filtré. On peut alors réécrire l'équation

(2.19) en remplaçant la MAS du mélange par celle du bruit :

$$\mathbf{w}_{\text{MVDR}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathbf{w}^H \mathbf{R}_{\text{n}} \mathbf{w} \qquad \text{tel que} \qquad \mathbf{w}^H \mathbf{d} = 1\,, \tag{2.21}$$

dont la solution est :

$$\mathbf{w}_{\text{MVDR}} = \frac{\mathbf{R}_{\text{n}}^{-1}\mathbf{d}}{\mathbf{d}^H \mathbf{R}_{\text{n}}^{-1}\mathbf{d}}\,. \tag{2.22}$$

Ce filtre est appelé le formateur à variance minimale et réponse sans distorsion (MVDR : *minimum variance distortionless response*) (Cox et al., 1987; Affes and Grenier, 1997; Souden et al., 2010). Il a l'avantage d'être plus robuste que le MPDR (Cox, 1973) mais nécessite l'estimation de la MAS du bruit, donc l'utilisation d'un DAV.

Filtres aux contraintes linéaires et à puissance ou variance minimale On peut ajouter plus de contraintes aux équations (2.19) et (2.21) si l'on sait que L_s sources cibles sont présentes aux angles $\{\theta_i\}_{i=1..L_s}$ et L_n sources interférentes aux angles $\{\phi_i\}_{i=1..L_n}$. En construisant la matrice $\mathbf{D}$ dont les $L = L_s + L_n$ colonnes rassemblent les vecteurs d'orientation correspondant à chacune des sources, on peut réécrire l'équation (2.18) avec les nouvelles contraintes :

$$\mathbf{w}_{\text{LCMP}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathbf{w}^H \mathbf{R}_{\text{y}} \mathbf{w} \qquad \text{tel que} \qquad \mathbf{D}\mathbf{w}^H = \mathbf{q}\,, \tag{2.23}$$

où $\mathbf{q}$ rassemble les gains que l'on veut appliquer à chaque source (typiquement, $q_i = 1$ pour les sources de parole et $q_j = 0$ pour les sources de bruit). La solution de (2.23) est donnée par :

$$\mathbf{w}_{\text{LCMP}} = \frac{\mathbf{R}_{\text{y}}^{-1}\mathbf{D}}{\mathbf{D}^H \mathbf{R}_{\text{y}}^{-1}\mathbf{D}}\mathbf{q}\,. \tag{2.24}$$

Le filtre $\mathbf{w}_{\text{LCMP}}$ est appelé formateur aux contraintes linéaires et à puissance minimale (LCMP : *linearly constrained minimum power*) (Frost, 1972).

De même qu'avec le MVDR, il est possible de minimiser la puissance du bruit uniquement, en réécrivant l'équation (2.23) :

$$\mathbf{w}_{\text{LCMV}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathbf{w}^H \mathbf{R}_{\text{n}} \mathbf{w} \qquad \text{tel que} \qquad \mathbf{D}\mathbf{w}^H = \mathbf{q}\,, \tag{2.25}$$

qui a pour solution le formateur aux contraintes linéaires et à variance minimale (LCMV : *linearly constrained minimum variance*) suivant (Cox, 1973; Er and Cantoni, 1983) :

$$\mathbf{w}_{\text{LCMV}} = \frac{\mathbf{R}_{\text{n}}^{-1}\mathbf{D}}{\mathbf{D}^H \mathbf{R}_{\text{n}}^{-1}\mathbf{D}}\mathbf{q}\,. \tag{2.26}$$

Ce filtre est plus robuste que le LCMP mais nécessite de connaître les statistiques du bruit, donc d'avoir un DAV fiable. Il est largement utilisé pour le rehaussement de la parole (Buckley and Griffiths, 1986; Zhao et al., 2012; Markovich-Golan et al., 2012b; Koutrouvelis et al., 2018; Zhang et al., 2018).

L'annulateur de lobe latéral généralisé Il est possible de rassembler tous les filtres présentés précédemment sous une seule représentation, celle de l'annulateur de lobe latéral généralisé (GSC : *generalized sidelobe canceller*) (Griffiths and Jim, 1982). Breed et Strauss ont en effet montré que la structure du GSC était une forme générique du LCMV, qui est lui-même une généralisation du MVDR (Breed and Strauss, 2002). Le GSC a en plus l'avantage d'être constitué d'une annulation de bruit adaptative qui permet de suivre des changements de position des sources de bruit (Affes and Grenier, 1997).

Pour conclure et résumer cette partie, nous rassemblons les différents filtres présentés dans le tableau 2.1 en précisant les puissances qu'ils minimisent et sous quelles contraintes.

TABLEAU 2.1. – Résumé des principaux filtres à contraintes linéaires. $\mathbb{I}$ est la matrice identité.

Filtre	Grandeur minimisée	Contraintes	Solution
DSB	$\mathbb{I}$	$\mathbf{w}^H\mathbf{d} = 1$	$\mathbf{d}$ (2.16)
MPDR	$\mathbf{R}_\mathrm{y}$	$\mathbf{w}^H\mathbf{d} = 1$ (2.19)	$\frac{\mathbf{R}_\mathrm{y}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{R}_\mathrm{y}^{-1}\mathbf{d}}$ (2.20)
MVDR	$\mathbf{R}_\mathrm{n}$	$\mathbf{w}^H\mathbf{d} = 1$ (2.21)	$\frac{\mathbf{R}_\mathrm{n}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{R}_\mathrm{n}^{-1}\mathbf{d}}$ (2.22)
LCMP	$\mathbf{R}_\mathrm{y}$	$\mathbf{w}^H\mathbf{D} = \mathbf{q}$ (2.23)	$\frac{\mathbf{R}_\mathrm{y}^{-1}\mathbf{D}}{\mathbf{D}^H\mathbf{R}_\mathrm{y}^{-1}\mathbf{D}}\mathbf{q}$ (2.24)
LCMV	$\mathbf{R}_\mathrm{n}$	$\mathbf{w}^H\mathbf{D} = \mathbf{q}$ (2.25)	$\frac{\mathbf{R}_\mathrm{n}^{-1}\mathbf{D}}{\mathbf{D}^H\mathbf{R}_\mathrm{n}^{-1}\mathbf{D}}\mathbf{q}$ (2.26)

2.3.1.2. Les filtres à maximisation de RSB

Une autre catégorie de filtres cherche non pas à minimiser la variance d'un signal filtré, mais à maximiser le RSB du signal filtré (Cox et al., 1987). On cherche donc :

$$\mathbf{w}_\mathrm{maxRSB} = \underset{\mathbf{w}}{\mathrm{argmax}}\, \frac{\mathbf{w}^H\mathbf{R}_\mathrm{s}\mathbf{w}}{\mathbf{w}^H\mathbf{R}_\mathrm{n}\mathbf{w}}\,. \tag{2.27}$$

La solution à cette équation est appelée *filtre à RSB maximum*. Résoudre l'équation (2.27) revient à chercher $\mathbf{w}$ tel que :

$$\mathbf{R}_\mathrm{n}^{-1}\mathbf{R}_\mathrm{s}\mathbf{w} = \lambda_\mathrm{max}\mathbf{w}\,, \tag{2.28}$$

avec λ_max le maximum de la fonction $\mathbf{w} \mapsto \frac{\mathbf{w}^H\mathbf{R}_\mathrm{s}\mathbf{w}}{\mathbf{w}^H\mathbf{R}_\mathrm{n}\mathbf{w}}$. La solution à (2.27) est donc le vecteur propre de $\mathbf{R}_\mathrm{n}^{-1}\mathbf{R}_\mathrm{s}$ correspondant à la plus haute valeur propre (λ_max en l'occurence). C'est pour cela que ce filtre est également appelé formateur de voies à valeurs propres généralisées (Warsitz and Haeb-Umbach, 2007). L'avantage de ce filtre est qu'il maximise directement le RSB de sortie du filtre. L'inconvénient est qu'il demande de résoudre le problème de décomposition en valeurs propres de la matrice $\mathbf{R}_\mathrm{n}^{-1}\mathbf{R}_\mathrm{s}$ en plus de la nécessité de connaître les deux MAS de la parole et du bruit.

2.3.1.3. Le filtre de Wiener multicanal et ses variantes

Le filtre de Wiener classique Tous les filtres présentés précédemment en section 2.3.1.1 requièrent la connaissance des positions des sources relatives à l'antenne de microphones. Cette connaissance n'est pas toujours disponible, et limite l'application de ces filtres. Par ailleurs, plutôt que de chercher à minimiser la puissance du bruit ou du mélange filtré, on peut chercher à minimiser l'erreur quadratique moyenne entre le signal désiré et le signal filtré. En prenant comme signal désiré s_1, la parole enregistrée par le microphone de référence, cela se traduit par :

$$\mathbf{w}_\mathrm{MWF} = \underset{\mathbf{w}}{\mathrm{argmin}}\, \mathbb{E}\{||s_1 - \mathbf{w}^H\mathbf{y}||^2\}\,. \tag{2.29}$$

Le filtre qui minimise cette fonction de coût est le filtre de Wiener multicanal (MWF : *multichannel Wiener filter*) (Widrow et al., 1967; Doclo and Moonen, 2002). C'est la solution optimale au sens des moindres carrés. En annulant la dérivée de la fonction (convexe) de (2.29), on trouve la solution suivante :

$$\mathbf{w}_\mathrm{MWF} = \mathbf{R}_\mathrm{y}^{-1}\mathbf{R}_\mathrm{ys}\mathbf{e}_1\,, \tag{2.30}$$

avec $\mathbf{e}_1 = [1, 0, \cdots, 0]$ le vecteur qui sélectionne le (premier) canal de référence. Si le bruit et la parole sont décorrélés et de moyenne nulle, on a $\mathbf{R}_{\text{ys}} = \mathbf{R}_{\text{s}}$ et $\mathbf{R}_{\text{y}} = \mathbf{R}_{\text{s}} + \mathbf{R}_{\text{n}}$. Ainsi l'équation (2.30) est équivalente à :

$$\mathbf{w}_{\text{MWF}} = (\mathbf{R}_{\text{s}} + \mathbf{R}_{\text{n}})^{-1}\mathbf{R}_{\text{s}}\mathbf{e}_1 \,. \tag{2.31}$$

En factorisant l'équation (2.30), on peut montrer que le MWF est équivalent à un MVDR suivi d'un filtre mono-canal si le bruit suit une loi de distribution gaussienne (Simmer et al., 2001; Balan and Rosca, 2002).

Le MWF a l'avantage de ne pas reposer sur la connaissance des angles d'arrivée. Aucune connaissance *a priori* sur la position des microphones et des sources n'est requise. En revanche, il nécessite un DAV et le bruit doit être relativement stationnaire pour permettre une estimation précise de la MAS de la parole sur les périodes où la parole est active. La formulation (2.29) peut se décliner en différentes variantes qui permettent d'ajuster certains critères ou de rendre l'implémentation du filtre plus robuste. Elles sont présentées ci-dessous.

Filtre de Wiener de rang 1 Lorsque seule une source de parole est présente, en l'absence de réverbération, on peut supposer que le signal de parole tel que mesuré par les microphones dépend uniquement du vecteur d'orientation :

$$\mathbf{s} = s_{\text{src}}\mathbf{d} \,, \tag{2.32}$$

où s_{src} est le signal de parole non réverbéré. Ainsi la MAS de la parole $\mathbf{R}_{\text{s}}$ se calcule par (cf. équation (2.11)) :

$$\begin{aligned}
\mathbf{R}_{\text{s}} =& \mathbb{E}\{\mathbf{s}\mathbf{s}^H\} && (2.33)\\
=& \mathbb{E}\{s_{\text{src}}\mathbf{d}\mathbf{d}^H s_{\text{src}}^*\} && (2.34)\\
=& \sigma_s^2\mathbf{d}\mathbf{d}^H \,. && (2.35)
\end{aligned}$$

La dernière ligne est vraie car on suppose la fonction de transfert acoustique constante au cours du temps (donc les sources et microphones statiques). $\sigma_s^2 = \mathbb{E}\{s_{\text{src}}s_{\text{src}}^*\}$ est la puissance du signal de parole.

En réinjectant (2.35) dans (2.31) et à l'aide de l'identité de Woodbury (Woodbury, 1950), on trouve l'expression du filtre de Wiener de rang 1 suivante (Doclo et al., 2006; Souden et al., 2010) :

$$\mathbf{w}_{\text{r1-MWF}} = \frac{\mathbf{R}_{\text{n}}^{-1}\mathbf{R}_{\text{s}}\mathbf{e}_1}{1 + tr\{\mathbf{R}_{\text{n}}^{-1}\mathbf{R}_{\text{s}}\}} \,, \tag{2.36}$$

où $tr\{\cdot\}$ désigne l'opérateur de la trace, donc la somme des termes diagonaux. Notons que puisque $\mathbf{R}_{\text{n}}^{-1}\mathbf{R}_{\text{s}}$ est de rang 1 [1], la trace de cette matrice est égale à sa seule valeur propre.

L'avantage de cette formulation, strictement équivalente à la formulation en (2.31) sous la condition (forte) que la matrice $\mathbf{R}_{\text{s}}$ est bien de rang 1, est que son implémentation est plus robuste. En pratique en effet, on vient forcer le rang de $\mathbf{R}_{\text{s}}$ en prenant $\mathbf{R}_{\text{s}} = \mathbf{a}\mathbf{a}^H$ avec $\mathbf{a}$ le vecteur propre de $\mathbf{R}_{\text{s}}$ correspondant à sa plus grande valeur propre. On est ainsi assuré que $\mathbf{R}_{\text{s}}$ est bien déterminée (Cornelis et al., 2010).

Filtre de Wiener multicanal pondéré par la distorsion de la parole Si le bruit et la parole sont décorrélés, la fonction de coût dans (2.29) peut être décomposée en deux termes :

$$\mathcal{J}(\mathbf{w}) = \mathbb{E}\{||s_1 - \mathbf{w}^H\mathbf{s}||^2\} + \mu\mathbb{E}\{||\mathbf{w}^H\mathbf{n}||^2\} \,. \tag{2.37}$$

1. Par invariance du rang d'une matrice par multiplication avec une matrice inversible.

Le paramètre $\mu \in \mathbb{R}^+$ permet de pondérer la réduction du bruit (exprimée dans le terme $\mathbb{E}\{||\mathbf{w}^H\mathbf{n}||^2\}$) par la distorsion de la parole (exprimée dans le terme $\mathbb{E}\{||s_1 - \mathbf{w}^H\mathbf{s}||^2\}$). Choisir un μ élevé revient à accorder plus d'importance à la réduction de bruit, quitte à ce que cela déforme le signal de parole. Inversement, prendre un μ faible assure une distorsion minimale de la parole, quitte à filtrer moins de bruit. Ainsi, le cas extrême $\mu = 0$, pour lequel aucune distorsion n'est appliquée au signal de la parole, donne la même solution que le MVDR.

Le vecteur $\mathbf{w}$ qui minimise (2.37) est le filtre de Wiener multicanal pondéré par la distorsion de la parole (SDW-MWF : *speech distortion weighted multichannel Wiener filter*) (Doclo and Moonen, 2002; Doclo et al., 2007), donné par l'expression :

$$\mathbf{w}_{\text{SDW-MWF}} = (\mathbf{R}_\text{s} + \mu\mathbf{R}_\text{n})^{-1}\mathbf{R}_\text{s}\mathbf{e}_1 \,. \tag{2.38}$$

Le MWF et le SDW-MWF sont équivalents pour $\mu = 1$. Par ailleurs, Doclo et al. démontrent que comme dans le cas du MWF, le SDW-MWF peut être décomposé en un MVDR auquel est appliqué un gain (Doclo et al., 2010a). De la même manière que pour le MWF, on peut supposer la MAS de la parole de rang 1, ce qui donne l'expression du SDW-MWF de rang 1 suivante :

$$\mathbf{w}_{\text{r1-SDW-MWF}} = \frac{\mathbf{R}_\text{n}^{-1}\mathbf{R}_\text{s}\mathbf{e}_1}{\mu + tr\{\mathbf{R}_\text{n}^{-1}\mathbf{R}_\text{s}\}} \,. \tag{2.39}$$

Filtre de Wiener multicanal basé sur la décomposition en valeurs propres généralisée des matrices d'autocorrélations spatiales En supposant $\mathbf{R}_\text{y}$ et $\mathbf{R}_\text{n}$ de rang plein, on peut trouver [2] une matrice $\mathbf{Q}$ inversible qui diagonalise conjointement $\mathbf{R}_\text{s}$ et $\mathbf{R}_\text{n}$ (Doclo and Moonen, 2002; Serizel et al., 2014) :

$$\mathbf{R}_\text{y} = \mathbf{Q}\mathbf{\Sigma}_\text{y}\mathbf{Q}^H \,, \tag{2.40}$$
$$\mathbf{R}_\text{n} = \mathbf{Q}\mathbf{\Sigma}_\text{n}\mathbf{Q}^H \,, \tag{2.41}$$
$$\mathbf{R}_\text{n}^{-1}\mathbf{R}_\text{y} = \mathbf{Q}^{-H}\mathbf{\Sigma}_\text{n}^{-1}\mathbf{\Sigma}_\text{y}\mathbf{Q}^H \,, \tag{2.42}$$

$$\mathbf{R}_\text{s} = \mathbf{Q}(\mathbf{\Sigma}_\text{y} - \mathbf{\Sigma}_\text{n})\mathbf{Q}^H \,, \tag{2.43}$$
$$\mathbf{R}_\text{s} = \mathbf{Q}\mathbf{\Sigma}_\text{s}\mathbf{Q}^H \,. \tag{2.44}$$

Avec ces nouvelles formulations, il est possible de réexprimer le SDW-MWF par :

$$\mathbf{w}_{\text{GEVD-SDW-MWF}} = \frac{\mathbf{Q}^{-H}\mathbf{\Sigma}\mathbf{Q}^H}{\mathbf{Q}^{-H}\mathbf{\Sigma}\mathbf{Q}^H + \mu}\mathbf{e}_1 \tag{2.45}$$
$$= \mathbf{Q}^{-H}\text{diag}\Big\{\frac{\frac{\sigma_{s_i}}{\sigma_{n_i}}}{\frac{\sigma_{s_i}}{\sigma_{n_i}} + \mu}\Big\}\mathbf{Q}^H\mathbf{e}_1 \,, \tag{2.46}$$

avec $\mathbf{\Sigma} = \mathbf{\Sigma}_\text{n}^{-1}\mathbf{\Sigma}_\text{s} = \text{diag}\big\{\frac{\sigma_{s_i}}{\sigma_{n_i}}\big\}$ et $\text{diag}\{a_i\}$ la matrice diagonale avec les termes $\{a_i\}_{i=1..M}$ dans la diagonale.

Cette formulation permet une implémentation du filtre plus stable car on calcule $\mathbf{\Sigma}_\text{s} = \mathbf{\Sigma}_\text{n}\big(\mathbf{\Sigma}_\text{n}^{-1}\mathbf{\Sigma}_\text{y} - \mathbb{I}\big)$ directement dans l'espace des vecteurs propres, où cette matrice est mieux

2. Une approximation de cette matrice peut se trouver numériquement (Cardoso and Souloumiac, 1994), la diagonalisation conjointe de deux matrices n'étant assurée que si elles commutent (Cardoso and Souloumiac, 1996).

définie. Par ailleurs, Serizel et al. montrent que prendre la formule (2.45) revient à considérer non plus la parole au premier microphone comme signal de référence, mais la projection de ce signal dans l'espace des vecteurs propres : $\mathbf{t}_1 = q_{11}^* \mathbf{Q}^{-H} \mathbf{e}_1$ avec q_{11} le premier élément du premier vecteur de $\mathbf{Q}$ (Serizel et al., 2014). Dans cet espace, les signaux à RSB plus élevé sont plus mis à contribution. Pour augmenter la réduction de bruit, on peut ne considérer que les r premiers éléments diagonaux de $\mathbf{\Sigma}$, ce qui revient à ne considérer que les r signaux aux RSB les plus élevés. Cela se fait cependant au prix d'une distorsion de la parole plus élevée. Il est ainsi possible de jouer à la fois sur le terme μ et sur le rang r de la matrice $\mathbf{\Sigma}$ pour ajuster le compromis entre la réduction de bruit et la distorsion de la parole. Dans le cas où $\mathbf{\Sigma}$ est de rang 1, on fera référence au filtre correspondant par l'acronyme r1-GEVD-SDW-MWF.

2.3.2. Approches basées sur l'apprentissage profond

Il a été montré que l'usage de l'information spatiale augmente les performances finales de techniques reposant sur le masquage spatial (Jiang et al., 2014; Li et al., 2020a; Tan et al., 2021a) ou sur l'utilisation d'auto-encodeurs (Araki et al., 2015). Ces résultats montrent que les RN peuvent extraire de ces représentations multicanales de l'information utile au rehaussement de la parole et ainsi améliorer leurs performances de rehaussement de la parole. Nous présentons dans la suite quatre grandes approches du rehaussement de la parole multicanal basées sur l'apprentissage profond.

Prédiction de masques TF Plutôt que d'utiliser le masque TF directement comme filtre, il est préférable de l'utiliser comme remplacement du DAV nécessaire aux formateurs de voies. De nombreux travaux estiment ces masques TF avec des RN à partir de signaux multicanaux (Heymann et al., 2016; Erdogan et al., 2016; Perotin et al., 2018; Togami, 2019; Martín-Doñas et al., 2020), ce qui a l'avantage de profiter de la capacité de modélisation des RN et de la robustesse des formateurs de voies. De manière très similaire, on peut également utiliser des RN pour prédire des signaux rehaussés qui eux-mêmes servent à calculer les MAS nécessaires aux formateurs de voies (Nugraha et al., 2016; Ochiai et al., 2020).

Prédiction des coefficients des formateurs de voies Une autre approche consiste à directement prédire les coefficients du filtre multicanal (Koyama et al., 2020; Pfeifenberger et al., 2019). Toutes les opérations de calcul de MAS et du filtre sont remplacées par des couches de RN, et l'apprentissage se fait par une fonction de coût basée sur la reconstruction du signal. Cela a l'avantage de considérer le signal filtré plutôt que le masque en sortie de réseau, et de se dédouaner des inversions de matrices, souvent coûteuses, requises dans les formulations des filtres (voir par exemple les équations (2.26) et (2.39)). Dans cet état d'esprit, Pfeifenberger et al. ont utilisé des réseaux de neurones complexes (Pfeifenberger et al., 2019), et Li et al. ont remplacé les différentes parties d'un GSC par des couches apprenables tout en gardant la structure globale et la logique du GSC (Li et al., 2021).

Approches de bout-en-bout De même que dans le cas mono-canal, il est possible d'effectuer le rehaussement de la parole multicanal de bout-en-bout, c'est-à-dire de fournir la forme d'onde bruitée au RN et d'en obtenir la forme d'onde rehaussée (Luo et al., 2019, 2020a; Liu et al., 2020). Afin d'alléger un peu la complexité du RN, la TFCT peut toutefois être considérée comme représentation intermédiaire en entrée et sortie du RN (Martín-Doñas et al., 2017).

2.4. Rehaussement de la parole dans les antennes acoustiques ad-hoc

Comme présenté en introduction, les antennes de microphones ad-hoc ont un fort potentiel, au prix de plusieurs défis. L'une des principales difficultés est d'avoir un algorithme distribué, c'est-à-dire qui ne repose pas sur la présence d'un centre de fusion ou sur des calculs trop gourmands en énergie ou en bande passante. Nous présentons ces algorithmes dans la première partie de cette section. En réponse aux autres grandes difficultés rencontrées dans les antennes acoustiques ad-hoc (AAAH), les solutions de calibration, de synchronisation et de réduction de bande passante sont présentées ensuite.

2.4.1. Algorithmes distribués de rehaussement de la parole

De nombreux algorithmes ont été développés afin de se défaire de la contrainte d'un centre de fusion dans le contexte du rehaussement de la parole dans les AAAH. Un algorithme de bavardage (Boyd et al., 2006) par exemple permet de calculer le MVDR de manière distribuée (Zeng and Hendriks, 2014, 2015). Valable sous la contrainte forte que le bruit soit décorrélé entre les nœuds de l'antenne, cette approche a été généralisée à l'aide de la méthode du passage de message (Heusdens et al., 2012), qui autorise que le bruit soit partiellement corrélé entre les microphones. Koutrouvelis et al. proposent de considérer que seuls les bruits enregistrés par les microphones d'un même nœud soient des bruits corrélés. La corrélation inter-nœud étant nulle, il est possible de diagonaliser par blocs la MAS de bruit et de développer un algorithme qui ne repose ni sur un centre de fusion, ni sur une topologie particulière afin de calculer un formateur de voies centralisé (Koutrouvelis et al., 2018) ou spécifique à chaque nœud (Guo et al., 2021a).

D'autres algorithmes, basés sur une logique de diffusion permettent de supprimer la contrainte de la présence d'un centre de fusion, comme proposé par Lopes et Sayed (Lopes and Sayed, 2008) ou O'Connor et al. (O'Connor and Kleijn, 2014). Une autre solution consiste à transformer le problème aux contraintes tel que présenté dans les équations (2.21) et (2.25) en un problème complexe solvable à l'aide d'algorithmes d'optimization distribués, comme le DBSA (*dual-based subgradient algorithm*) (Bertrand and Moonen, 2011a), l'ADMM (*alternative direction method of multipliers*) (Li et al., 2020b) ou le PDMM (*primal-dual method of multipliers*) (Tavakoli et al., 2016; Sherson et al., 2016).

Enfin, des approches basées sur des signaux dits compressés permettent de se dédouaner de la contrainte d'un centre de fusion, et de limiter les besoins en bande passante dans le cas où chaque nœud doit estimer un signal cible propre : seul un nombre restreint de signaux est échangé entre les nœuds plutôt que l'ensemble des signaux enregistrés par tous les microphones de chaque nœud (Bertrand and Moonen, 2011b, 2013; Markovich-Golan et al., 2015; Ranjbaryan et al., 2018; Ranjbaryan and Abutalebi, 2020; Guo et al., 2021b; Musluoglu and Bertrand, 2021). Il a été prouvé que ces approches distribuées convergent vers les solutions centralisées par un processus itératif. Ainsi le LCMV a pu être distribué (Bertrand and Moonen, 2011b, 2013), puis le MWF (Doclo et al., 2009; Bertrand and Moonen, 2010a,b) et le GSC (Markovich-Golan et al., 2012b).

Parce qu'une grande partie de nos travaux repose sur lui, l'algorithme distribué, adaptatif et spécifique à chaque nœud (DANSE : *distributed adaptive node-specific signal estimation*) (Bertrand and Moonen, 2010a,b) est décrit dans une partie à part, en section 3.1.

2.4.2. Solutions aux autres défis posés par les antennes acoustiques ad-hoc

Bien que les points abordés par la suite ne soient pas spécifiques au rehaussement de la parole, ils l'impactent tout de même et il n'est pas rigoureux de parler des AAAH sans les traiter. Nous présentons ainsi dans les grandes lignes certaines solutions proposées aux défis de la calibration des microphones, de leur synchronisation et des limites en bande passante.

Calibration des microphones Comme vu dans l'introduction, la calibration des microphones recouvre en réalité deux concepts.

En réponse à la *localisation des microphones et des sources*, trois grandes approches se dégagent. La première approche, plus intuitive, se base sur la corrélation entre un signal joué connu à l'avance et celui mesuré par les microphones (Sachar et al., 2004; Raykar et al., 2004). C'est une méthode assez contrainte dans l'application car il faut un haut-parleur pour jouer le son et calibrer les signaux avant l'utilisation des microphones. Une deuxième catégorie, moins contrainte, repose uniquement sur la connaissance que certains événements sonores, comme de la parole parlée, sont présents dans la scène acoustique (Liu et al., 2007; Chen et al., 2007; Gaubitch et al., 2013; Gburrek et al., 2021b). Enfin la dernière catégorie fonctionne en présence de bruit diffus, donc en l'absence de toute source sonore localisée, en se basant sur le fait que la cohérence du bruit entre microphones doit avoir la forme d'un sinus cardinal (McCowan et al., 2008; Hennecke et al., 2009; Taghizadeh et al., 2015). Plus récemment, les RN ont également été utilisés, par exemple pour estimer les distances entre microphones et sources (Gburrek et al., 2021a). Une étude plus détaillée des différentes approches est proposée par Plinge et al. (Plinge et al., 2016).

En réponse à l'*égalisation des réponses fréquentielles des microphones*, les premières solutions se sont concentrées sur les gains des microphones (Madhu and Martin, 2011; Gaubitch et al., 2014), mais d'autres travaux ont également pris en compte la phase des signaux et calibré toute la réponse fréquentielle des microphones (Wang et al., 2021b). Le calibrage de la réponse fréquentielle repose souvent sur la création d'un signal de référence qui peut être un signal issu d'un formateur de voies (Oak and Kellermann, 2005; Hu et al., 2019).

Synchronisation des microphones Les effets de la désynchronisation des microphones dans une antenne de microphones ont été assez largement étudiés, en particulier son impact sur le rehaussement de la parole. On peut ainsi trouver des études théoriques et empiriques sur l'impact de la désynchronisation sur les formateurs de voies (Zeng and Hendriks, 2015; Cherkassky et al., 2015; Schmalenstroeer and Haeb-Umbach, 2018b). On peut extraire deux grandes approches basées sur les signaux pour resynchroniser les signaux enregistrés par des microphones dont les fréquences d'échantillonnage diffèrent (nous n'aborderons pas les approches reposant sur le matériel). La première méthode consiste à envoyer par un nœud maître un signal de référence sur lequel les autres microphones s'alignent (Schenato and Fiorentin, 2011; Rajan and van der Veen, 2011; Ceolini et al., 2020). La deuxième approche est dite aveugle car elle n'a pas de référence de temps à disposition mais uniquement les signaux enregistrés par les microphones (Markovich-Golan et al., 2012a; Wang and Doclo, 2016; Cherkassky and Gannot, 2017; Chinaev et al., 2021). Les solutions reposent souvent sur la corrélation entre les signaux mesurés, si bien que les nœuds doivent s'échanger les signaux qu'ils enregistrent. Cette approche est moins intrusive (aucun signal de référence ne doit être envoyé) et plus robuste au bruit mais nécessite une plus grande bande passante pour l'envoi des signaux (Zeng and Hendriks, 2015). Précisons enfin que certains algorithmes de rehaussement de la parole sont robustes aux désynchronisations de microphones et ne nécessitent donc aucune compensation (Chiba et al., 2014; Corey and Singer, 2018).

Limitation en bande passante Afin de réduire les besoins en bande passante, trois cas de figure se distinguent. Dans le premier cas de figure, chaque nœud de l'AAAH dispose d'une certaine puissance de calcul et peut traiter localement les signaux mesurés en envoyant le résultat de ce traitement sous forme de signal compressé, à l'image des travaux de Bertrand et Moonen décrits en détail en section 3.1 (Zhang et al., 2003; Bertrand and Moonen, 2010a). Cette technique manque toutefois de flexibilité et est limitée par la puissance de calcul des nœuds car tout est opéré localement. Dans le cas où les signaux doivent être envoyés sans traitement préalable, la consommation de bande passante peut être limitée en encodant les signaux sur un nombre restreint de bits, en accordant un nombre supérieur de bits aux signaux ou aux parties de signal les plus utiles (Srinivasan and Den Brinker, 2009; Doclo et al., 2010b; Amini et al., 2019; Drude et al., 2021). On atteint le cas extrême de cette approche lorsqu'un signal n'est pas envoyé (aucun bit ne lui est alloué) ; on parle alors de sélection de canaux où ne sont échangés que les signaux les plus utiles à la tâche considérée pour une contrainte en bande passante donnée (Bertrand and Moonen, 2010; Szurley et al., 2011; Casebeer et al., 2021). Là encore, des RN peuvent être utilisés pour sélectionner un nombre restreint de canaux (Casebeer et al., 2021).

2.5. Vers des solutions en temps réel

On désigne sous l'appellation *temps réel* des algorithmes dont le temps de calcul est inférieur à la latence (Reddy et al., 2020), c'est-à-dire inférieur à la durée de la fenêtre temporelle considérée par l'algorithme. En fonction des applications des solutions décrites dans les sections précédentes, il est souvent nécessaire d'avoir une latence très faible. Par exemple, les utilisateurs d'appareils auditifs trouvent nuisible un décalage de 10 ms entre le moment où un son est émis (donc vu par l'utilisateur) et celui où il est reçu (donc entendu) (Agnew and Thornton, 2000).

La première condition que doit remplir un algorithme pour pouvoir fonctionner en temps réel est qu'il doit être causal, c'est-à-dire que sa sortie à un instant t_0 ne doit pas dépendre d'instants futurs $t > t_0$. A moins de rajouter des échantillons en début de signal (de rajouter des 0s par exemple), la fenêtre d'analyse de la TFCT amène nécessairement de la latence, puisqu'il faut connaître tous les points de la fenêtre d'analyse avant de calculer la TFCT. Afin de diminuer la latence, il est donc possible de réduire la taille de la fenêtre d'analyse de la TFCT (Naithani et al., 2017). Les approches de bout-en-bout, qui ont supprimé la TFCT notamment dans le but de réduire la latence (Luo et al., 2019; Défossez et al., 2020; Pandey and Wang, 2021), ont d'ailleurs montré des résultats supérieurs avec des fenêtres d'analyse plus courtes (Luo et al., 2020a; Pariente et al., 2020b).

Pour une latence donnée, un algorithme sera à temps réel si son temps de calcul est plus court que la latence. Cela dépend de l'implémentation logicielle de l'algorithme et de l'architecture matérielle sur laquelle il opère. Augmenter la puissance des plateformes matérielles sur lesquelles opèrent les algorithmes étant hors de notre expertise, nous nous concentrerons ici sur les implémentations logicielles. Le temps de calcul des solutions les plus récentes dépend en grande partie du temps de calcul des RN utilisés. Pour réduire le temps de calcul, il faut donc réduire la complexité des RN. Dans cette optique, on peut les utiliser pour des tâches plus simples, au sein d'un algorithme classique (Valin et al., 2020). Une autre solution consiste à utiliser des données d'entrée plus simples que la TFCT (Valin, 2018; Li and Horaud, 2019; Bhat et al., 2019; Haruta and Ono, 2021) ou rendre le RN moins complexe soit dès la construction (Luo et al., 2021; Braun et al., 2021; Chen et al., 2021), soit par compression (Fedorov et al., 2020; Tan et al., 2021a; Tan and Wang, 2021).

2.6. Du difficile choix des métriques

2.6.1. Les différentes métriques objectives

Deux options s'offrent pour quantifier le résultat d'une opération de rehaussement de la parole. La première est de faire des tests d'écoute, qui permettent de prendre en compte la subjectivité des individus et les effets psychoacoustiques. On peut alors par exemple mesurer la note d'opinion moyenne (ITU, 2002), la préférence des individus pour un signal plutôt qu'un autre ou encore le seuil de compréhension de la parole (Plomp and Mimpen, 1979). Ce sont les tests les plus fiables si un nombre suffisamment élevé de personnes y prend part afin de tenir compte de la variabilité inter-individus. Cependant, ces tests sont très coûteux en terme de temps et de moyens et ne peuvent être réalisés qu'à de rares occasions. Ils ne constituent donc pas une mesure envisageable pour l'ensemble des expériences faites au cours d'une thèse, que nous évaluons à l'aide d'autres métriques.

Il faut pour cela des mesures plus automatiques, que l'on nommera mesures objectives. Motivées pour la plupart par des résultats de psychoacoustique, elles visent à quantifier à quel point un signal est de bonne qualité ou à quel point la parole qu'il contient est intelligible. De nombreuses mesures objectives ont ainsi été proposées, et nous n'en présenterons ici qu'une partie.

Rapport signal-à-bruit Le rapport signal-à-bruit (RSB) quantifie le rapport de la puissance du signal cible sur la puissance du signal de bruit. Dans le cas de figure de l'équation (2.5), cela se traduirait par :

$$\mathrm{RSB} = \frac{P_\mathrm{s}}{P_\mathrm{n}}\,, \tag{2.47}$$

où P_s et P_n sont respectivement les puissances de parole et de bruit. Le RSB est souvent considéré en décibels (dB) :

$$\mathrm{RSB}_\mathrm{db} = 10\log_{10}\big(\mathrm{RSB}\big)\,. \tag{2.48}$$

Sous l'une ou l'autre des formulations, un RSB plus élevé correspond à de meilleures performances de rehaussement de la parole. Cette métrique est simple mais ne rend pas compte de plusieurs effets psychoacoustiques, notamment que l'oreille humaine ne perçoit pas les sons de la même façon en fonction de leur contenu spectral ou du masquage du bruit. Pour pallier cela, le RSB pondéré spectralement et le RSB segmental (Tribolet et al., 1978) ont été introduits afin de quantifier les fluctuations spectrales et temporelles des signaux.

Distorsion de la parole Une réduction du bruit s'accompagne souvent d'une réduction de la parole. Afin de quantifier à quel point la parole a été dégradée par la réduction de bruit, on utilise une autre métrique, la distorsion de la parole. Elle qualifie le rapport entre la puissance de la parole non filtrée et celle de la parole filtrée :

$$\mathrm{DP} = \frac{P_\mathrm{s}}{P_{\hat{\mathrm{s}}}}\,, \tag{2.49}$$

où $P_{\hat{\mathrm{s}}}$ est la composante de parole dans le signal filtré. De même que pour le RSB, il est possible de considérer la distorsion de la parole en dB et de prendre la métrique segmentale ou pondérée spectralement. En revanche, contrairement au RSB, on cherche à minimiser la distorsion de la parole, dont les valeurs en dB sont toujours positives.

Métriques de séparation de sources On relève deux inconvénients des métriques précédentes. Le premier est qu'elles ne se basent que sur la puissance des signaux, et non leur forme d'onde ou, à défaut, leur TFCT. Le deuxième est qu'elles sanctionnent toute différence entre les puissances de référence et celles du signal filtré, alors que certaines différences, par exemple celles liées à la réverbération, peuvent être, sinon souhaitées du moins acceptables et sans répercussion sur les métriques. Vincent et al. ont ainsi présenté quatre métriques basées sur la corrélation entre les signaux désirés et les signaux filtrés (Vincent et al., 2006). Ils décomposent le signal filtré en la somme du signal désiré, des interférences, des artéfacts et du bruit :

$$\hat{s} = s_{\text{cible}} + e_{\text{interf}} + e_{\text{arté}} + e_{\text{bruit}} \tag{2.50}$$

avec s_{cible}, e_{interf}, $e_{\text{arté}}$ et e_{bruit} les quatre composantes : signal cible, interférences, artéfacts et bruit. A partir de ces quatre composantes quatre métriques sont définies :

- le rapport source-à-distorsion (SDR : *source to distortion ratio*)
  $\text{SDR} = 10\log_{10}\frac{||s_{\text{cible}}||^2}{||e_{\text{interf}}+e_{\text{arté}}+e_{\text{bruit}}||^2}$
- le rapport source-à-interférences (SIR : *source to interferences ratio*)
  $\text{SIR} = 10\log_{10}\frac{||s_{\text{cible}}||^2}{||e_{\text{interf}}||^2}$
- le rapport sources-à-artéfacts (SAR : *sources to artifacts ratio*)
  $\text{SAR} = 10\log_{10}\frac{||s_{\text{cible}}+e_{\text{interf}}+e_{\text{bruit}}||^2}{||e_{\text{arté}}||^2}$
- le rapport sources-à-bruit (SNR : *signal to noise ratio*)
  $\text{SNR} = 10\log_{10}\frac{||s_{\text{cible}}+e_{\text{interf}}||^2}{||e_{\text{bruit}}||^2}$

En se basant sur le travail de Vincent et al. (2006), Le Roux et al. (2019) ont proposé le SIR (resp. SDR) à invariance d'échelle (SI-SIR : *scale-invariant signal to interference ratio*) (resp. SI-SDR), depuis lors largement utilisé pour la séparation de sources (Zeghidour and Grangier, 2020; Tzinis et al., 2020; Subakan et al., 2021). L'avantage de ces métriques est qu'elles permettent une comparaison plus rigoureuse de différents algorithmes de séparation de sources. Une explication plus détaillée du calcul du SI-SDR est donnée en annexe A.

STOI La mesure d'intelligibilité objective à court-terme (STOI : *short-time objective intelligibility*) a été proposée par Taal et al. dans le but de maximiser la corrélation entre cette métrique et l'intelligibilité de la parole (Taal et al., 2010). Elle est basée sur la corrélation dans le domaine TF des enveloppes normalisées de la parole et du mélange et est comprise entre 0 (mauvaise intelligibilité) et 1 (bonne intelligibilité).

L'inconvénient de toutes ces métriques est qu'elles sont intrusives, c'est-à-dire qu'elles nécessitent au moins la parole (et le bruit) non mélangés, ce dont on ne dispose pas lorsque les mélanges ne sont pas synthétiques. Pour pallier cela, deux solutions sont possibles. Des métriques non intrusives existent comme le rapport d'énergie des modulations de parole et de réverbération (Falk et al., 2010) et l'aire moyenne du spectre de modulation (Chen et al., 2013), mais elles ne sont pas aussi précises que les métriques intrusives (Santos et al., 2013). Autrement, il a récemment été proposé d'utiliser des RN pour quantifier la qualité d'un signal (Gamper et al., 2019; Reddy et al., 2021; Serrà et al., 2021). Ces approches ne sont pas encore unanimement adoptées, d'abord parce qu'elles sont très récentes, et ensuite parce qu'elles sont soumises aux données et conditions d'entraînement des RN, difficilement contrôlables pour ceux qui n'ont pas eux-mêmes entraîné les RN.

Notons toutefois que les métriques de SIR, SAR et SDR ne nécessitent pas de connaître les composantes séparées de parole et de bruit dans le signal filtré (contrairement au RSB par exemple, qui nécessite d'avoir accès à $\tilde{s}$ et $\tilde{n}$, les composantes en parole et en bruit dans le mélange débruité). Cela les rend moins intrusives.

2.6.2. Les références des métriques

En plus des métriques elles-mêmes, les signaux choisis comme références pour calculer les métriques ont une grande influence sur la valeur de la métrique (Drude et al., 2019). Par exemple, la source non convoluée est commune à tous les microphones dans la pièce et permet donc une évaluation équitable des résultats aux différents microphones voire nœuds. En revanche, avec une telle référence, la réverbération est considérée comme une distorsion du signal et il est plus difficile de faire la part des choses entre les distorsions apportées par le traitement du signal et celles apportées par la réverbération. Prendre les sources convoluées a les avantages et inconvénients inverses. Comme aucune alternative n'est parfaite, la solution proposée par Drude et al. (2019) et retenue dans cette thèse est de considérer plusieurs métriques (et références) à la fois, afin d'affiner les analyses et de compenser les manques d'une métrique par les informations apportées par une autre.

2.7. Conclusion

Après avoir formalisé le problème du rehaussement de la parole, ce chapitre a présenté différentes solutions existantes du rehaussement de la parole. Ces solutions ont été catégorisées selon le nombre de microphones disponibles pour effectuer le rehaussement de la parole. Nous avons ainsi distingué les approches mono-canales des approches multicanales et des approches appliquées aux antennes acoustiques ad-hoc. La dernière section du chapitre a abordé les métriques couramment utilisées pour quantifier les solutions de rehaussement de la parole.

L'usage de RN a permis de grandement améliorer les performances de rehaussement de la parole mono-canal et multicanal, mais peu de solutions existent dans le contexte des AAAH. L'objectif de cette thèse est d'intégrer l'utilisation de RN dans les AAAH afin de profiter à la fois de la diversité d'information enregistrée par les AAAH, et de la puissance des RN. Une solution dans ce sens est présentée dans le chapitre 3.

Deuxième partie

Tango : Un nouveau système de rehaussement de la parole distribué dans les antennes acoustiques ad-hoc.

Cette partie présente notre solution pour un rehaussement de la parole distribué opérationnel dans les antennes acoustiques ad-hoc (AAAH). Le système, baptisé Tango, est présenté en chapitre 3, où il est également évalué sur deux corpus différents, simulant les conditions dans lesquelles des AAAH sont typiquement utilisées. Il y est également comparé à un système de l'état de l'art de rehaussement de la parole de bout-en-bout. Dans le chapitre 4, nous développons une étude plus détaillée des performances de Tango dans une grande variété de conditions acoustiques qui permettent d'analyser sa résilience. Il y est également évalué sur des données réelles. Enfin nous montrons que notre solution permet d'exploiter l'information spatiale enregistrée par tous les nœuds d'une AAAH et mettons en évidence la coopération entre les nœuds de l'antenne.

3. Présentation et validation de Tango

On se place dans le contexte d'une AAAH. Dans un premier temps, on supposera l'antenne toute-connectée et ses nœuds synchronisés. Comme présenté en introduction, la difficulté avec les AAAH est d'exploiter la multiplication des microphones et leur large étendue spatiale sans pour autant surcharger les appareils par des calculs trop lourds ni la bande passante par l'envoi de trop de données.

Ce chapitre présente une approche originale de rehaussement de la parole distribuée qui permet de réduire la puissance de calcul nécessaire sur chaque nœud de l'antenne. Le système, nommé Tango, est évalué sur des corpus d'antennes acoustiques ad-hoc. Nous commençons par décrire l'algorithme de rehaussement de la parole distribué, adaptatif et spécifique à chaque nœud DANSE (Bertrand and Moonen, 2010a,b) dont est inspiré notre propre algorithme (section 3.1). Dans la section 3.2, Tango est présenté, en mettant en évidence les principales contributions de cette nouvelle approche. Ensuite, la section 3.3 présente une évaluation expérimentale de l'approche proposée sur un premier corpus et démontre son efficacité à exploiter l'information spatiale. Afin d'optimiser les performances du système, nous l'évaluons dans une troisième partie sur un autre corpus plus représentatif de la réalité que le premier (section 3.5). Les deux dernières sections comparent cette approche à une solution de l'état de l'art (section 3.6) puis analysent ses performances sur des données réelles (section 4.4).

3.1. DANSE : rehaussement de la parole distribué, adaptatif et spécifique à chaque nœud

L'acronyme DANSE définit un algorithme d'estimation du signal distribué, adaptatif et spécifique à chaque nœud, qui a été introduit par Bertrand et Moonen en 2010 (Bertrand and Moonen, 2010a,b), bien que le même algorithme ait été proposé par Doclo et al. dans des conditions plus restreintes en 2009 (Doclo et al., 2009). Nous présentons dans cette section le cas particulier où un seul signal d'intérêt doit être restitué à chaque nœud.

On considère K nœuds de M_k microphones chacun, tel que $\sum_{k=1}^{K} M_k = M$. Les signaux au nœud k sont représentés par $\mathbf{y}_k = [y_{11}, \quad \ldots, \quad y_{1M_k}]^T$.

L'algorithme est itératif : à chaque itération i, chaque nœud k estime le signal désiré $\hat{s}_k^i$ qui converge vers la solution centralisée. Ainsi, si $\hat{s}_k^{\text{ctr}}$ est la solution du filtre centralisé (c'est-à-dire le filtre dans lequel chaque nœud a accès à tous les signaux de tous les autres nœuds), on a $\lim_{i\to\infty} \hat{s}_k^i = \hat{s}_k^{\text{ctr}}$.

A chaque itération i de l'algorithme, le nœud k reçoit de tous les autres nœuds $j \neq k$ un signal dit *compressé* z_j^i. Les signaux $\{z_j^i\}_{j\neq k}$ sont rassemblés dans le vecteur $\mathbf{z}_{-k}^i$, si bien que le nœud k dispose à l'itération i des signaux :

$$\tilde{\mathbf{y}}_k^i = \begin{bmatrix} \mathbf{y}_k \\ \mathbf{z}_{-k}^i \end{bmatrix}. \tag{3.1}$$

On note de même respectivement $\tilde{\mathbf{s}}^i$ et $\tilde{\mathbf{n}}^i$ les composantes de parole et de bruit de $\tilde{\mathbf{y}}_k^i$.

La matrice d'autocorrélation spatiale (MAS) de $\tilde{\mathbf{y}}_k^i$ est $\mathbf{R}_{\tilde{\mathbf{y}},k}^i$, celles de $\tilde{\mathbf{s}}_k^i$ et $\tilde{\mathbf{n}}_k^i$ sont $\mathbf{R}_{\tilde{\mathbf{s}},k}^i$ et $\mathbf{R}_{\tilde{\mathbf{n}},k}^i$ respectivement. Ces deux dernières sont calculées à l'aide d'un détecteur d'activité vocale (DAV)

binaire de la manière suivante :

$$\begin{cases} \text{si DAV} = 0 & \mathbf{R}^i_{\tilde{\mathrm{n}},k} = \mathbb{E}\{\tilde{\mathbf{y}}^i_k \tilde{\mathbf{y}}^{iH}_k\} = \mathbb{E}\{\tilde{\mathbf{n}}^i_k \tilde{\mathbf{n}}^{iH}_k\} \quad \text{(car la parole est absente)} \\ \text{si DAV} = 1 & \mathbf{R}^i_{\tilde{\mathrm{s}},k} = \mathbb{E}\{\tilde{\mathbf{y}}^i_k \tilde{\mathbf{y}}^{iH}_k\} - \mathbf{R}^i_{\tilde{\mathrm{n}},k} \quad \text{(en supposant le bruit stationnaire)} \end{cases} \tag{3.2}$$

On cherche le filtre de Wiener multicanal (MWF : *multichannel Wiener filter*) appliqué à $\tilde{\mathbf{y}}^i_k$ qui minimise l'erreur quadratique moyenne entre le signal filtré et le signal de référence de chaque nœud s_{k1} (pris ici encore comme le premier signal de chaque nœud [1]) :

$$\mathbf{w}^{i+1}_k = \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathbb{E}\{||s_{k1} - \mathbf{w}^H \tilde{\mathbf{y}}^i_k||^2\}\,. \tag{3.3}$$

Similairement à (2.30) et (2.31), il est donné par :

$$\mathbf{w}^{i+1}_k = {\mathbf{R}^i_{\tilde{\mathrm{y}},k}}^{-1} \mathbf{R}^i_{\tilde{\mathrm{y}}\mathrm{s},\,\mathrm{k}} \mathbf{e}_1 \tag{3.4}$$

$$= (\mathbf{R}^i_{\tilde{\mathrm{s}},k} + \mathbf{R}^i_{\tilde{\mathrm{n}},k})^{-1} \mathbf{R}^i_{\tilde{\mathrm{s}},k} \mathbf{e}_1\,. \tag{3.5}$$

Les premiers coefficients de $\mathbf{w}^{i+1}_k$ sont appliqués à $\mathbf{y}_k$ tandis que les derniers sont appliqués à $\mathbf{z}^i_{-k}$. On peut diviser le filtre $\mathbf{w}^{i+1}_k$ en deux termes, notés $\mathbf{w}^{i+1}_{kk}$ et $\mathbf{g}^{i+1}_{k-k}$:

$$\mathbf{w}^{i+1}_k = \begin{bmatrix} \mathbf{w}^{i+1}_{kk} \\ \mathbf{g}^{i+1}_{k-k} \end{bmatrix}, \tag{3.6}$$

et le signal estimé au nœud k est :

$$\hat{s}^i_k = {\mathbf{w}^{i+1}_k}^H \tilde{\mathbf{y}}^i_k \tag{3.7}$$

$$= {\mathbf{w}^{i+1}_{kk}}^H \mathbf{y}_k + {\mathbf{g}^{i+1}_{k-k}}^H \mathbf{z}^i_{-k}\,. \tag{3.8}$$

On calcule à partir de l'équation (3.8) le signal compressé du nœud k et de l'itération $i+1$:

$$z^{i+1}_k = {\mathbf{w}^{i+1}_{kk}}^H \mathbf{y}_k\,, \tag{3.9}$$

qui peut être envoyé à tous les autres nœuds où peut commencer une nouvelle itération. La figure 3.1 décrit l'algorithme DANSE sous forme schématique.

En pratique, les itérations se confondent avec les trames, c'est-à-dire que les filtres sont mis à jour à chaque nouvelle trame, et un seul signal compressé est envoyé par nœud et par trame.

Les nœuds peuvent mettre à jour leur filtre soit de manière séquentielle (Bertrand and Moonen, 2010a), soit de manière simultanée (Bertrand and Moonen, 2010b). Le pseudo-code de la version simultanée est porté dans l'algorithme 1. Bertrand et Moonen montrent que dans un cas comme dans l'autre, les signaux $\{\hat{s}^i_k\}_{k=1..K}$ convergent vers les résultats de MWF centralisés, c'est-à-dire vers les mêmes signaux que l'on aurait calculés si chaque nœud disposait de tous les signaux de tous les nœuds. Cet algorithme DANSE permet donc de réduire d'un facteur M_k les coûts en bande passante par rapport au cas où chaque nœud envoie tous ses signaux à tous les autres nœuds. Par ailleurs, chaque nœud k n'a plus que $M_k + (K-1) < M$ signaux, ce qui permet d'alléger les calculs, en particulier ceux d'inversion des MAS, qui sont en $\mathcal{O}(N^3)$.

La contrainte initiale d'une antenne toute-connectée a pu être d'abord remplacée par celle d'une topologie en arbre (Szurley et al., 2013) puis complètement supprimée (Szurley et al., 2016), mais au prix d'une convergence plus lente. Enfin, Hassani et al. ont montré que la même logique pouvait être utilisée pour le MWF avec décomposition en valeurs propres généralisée (Hassani et al., 2015).

1. Puisque chaque nœud a une référence différente, les filtres sont différents à tous les nœuds.

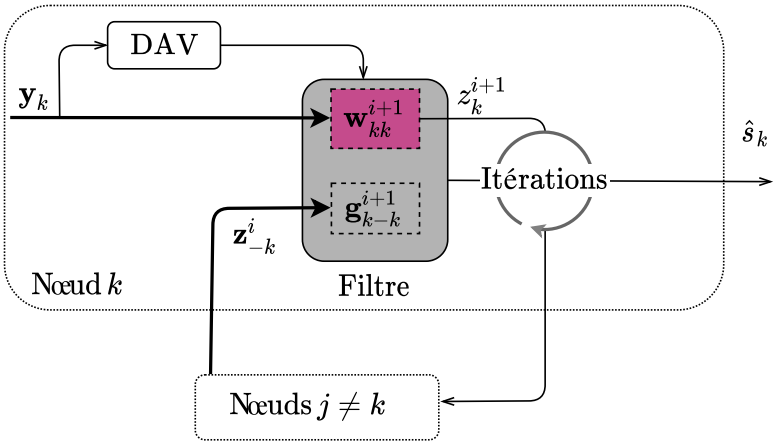


FIGURE 3.1. – Schéma de DANSE.

Algorithme 1 Pseudo-code de DANSE avec mise à jour séquentielle des filtres. Les notations sont celles de la section 3.1.

Initialiser $u \leftarrow 0$
Initialiser aléatoirement $\mathbf{w}_k^i \leftarrow \left[{\mathbf{w}_{kk}^i}^T, {\mathbf{g}_{-k}^i}^T \right]^T$
pour $i = 0 \cdots \mathcal{T}$ ▷ Rappel : itérations et trames sont confondues
  pour $k = 1 \cdots K$
    Envoyer $z_k^i = {\mathbf{w}_{kk}^i}^H \mathbf{y}_k^i$ aux nœuds distants
    Recevoir $\mathbf{z}_{-k}^i$ des nœuds distants
    A l'aide d'un DAV, calculer $\mathbf{R}_{\hat{\mathrm{s}},k}^i, \mathbf{R}_{\hat{\mathrm{n}},k}^i$
    si $k = u$ **alors**
      Calculer $\mathbf{w}_k^{i+1}$ (Eq. (3.4))
    sinon
      $\mathbf{w}_k^{i+1} \leftarrow \mathbf{w}_k^i$
    fin si
    Calculer $\hat{s}_k^i$ (Eq. (3.7))
    $i \leftarrow i + 1$
    $u \leftarrow (u \bmod K) + 1$
  fin pour
fin pour

3.2. Présentation de Tango

Notre algorithme Tango part de la même logique que l'algorithme DANSE décrit en section 3.1, notamment de l'idée de propager l'information spatiale sous forme de signaux compressés. Cette section présente quatre caractéristiques de Tango qui sont autant de contributions de cette thèse. Elles sont mises en évidence en les comparant à DANSE. Premièrement, le système de convergence itératif de DANSE est remplacé par un traitement par blocs, lui-même scindé en deux étapes de filtrage successives. Deuxièmement, un masque temps-fréquence (TF) est utilisé plutôt qu'un DAV pleine bande. Troisièmement, des réseaux de neurones (RN) sont utilisés afin d'estimer la présence ou non de parole dans chaque point TF du signal pour calculer les MAS. Enfin, les signaux compressés envoyés pour calculer le filtre interne sont également utilisés par les RN afin d'améliorer la prédiction des masques TF. Ces quatre caractéristiques sont décrites plus en détail dans les sections suivantes.

La figure 3.2 représente l'algorithme Tango dans le cas où deux nœuds sont présents. Elle

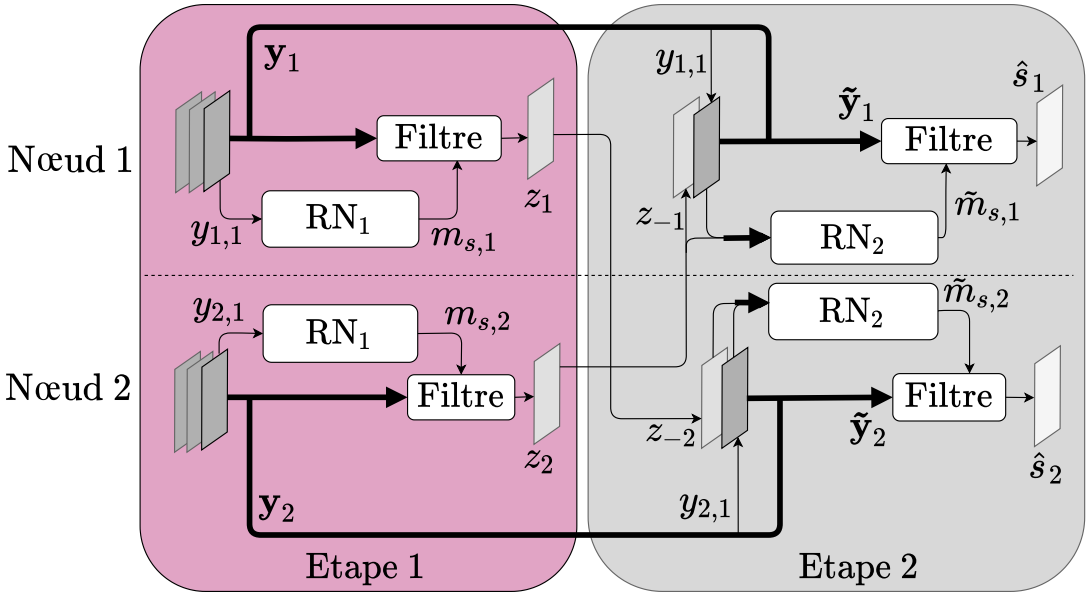


FIGURE 3.2. – Schéma de Tango dans un contexte à deux nœuds. Les flèches épaisses représentent des signaux multicanaux. Les flèches simples représentent des signaux mono-canaux.

illustre le fait que le filtrage de Tango se fait en deux étapes. A l'issue de la première étape, un signal compressé z_k est estimé au nœud k. Ce signal est envoyé à tous les autres nœuds $j \neq k$, qui l'utilisent à la deuxième étape d'abord pour prédire un masque TF, puis pour calculer un second filtre multicanal. La sortie de ce second filtre multicanal est l'estimation définitive du signal cible spécifique au nœud j, s_j. Les masques nécessaires au nœud k pour le calcul du filtre des étapes 1 et 2 sont respectivement notés $m_{s,k}$ et $\tilde{m}_{s,k}$. Ils sont prédits par des RN. Sur la figure 3.2, l'indice après l'acronyme « RN » fait référence au nombre de signaux fournis en entrée du RN.

3.2.1. Traitement par blocs non itératif

L'algorithme DANSE actualise les filtres $\mathbf{w}_k^i$ et $\mathbf{w}_{kk}^i$ à chaque nouvelle trame, ce qui lui confère un caractère adaptatif. Bien qu'il soit garanti que ces filtres convergent vers les filtres centralisés sous réserve que le DAV soit oracle, cette convergence n'est pas assurée en pratique, d'une part parce que le DAV (ou masque TF) n'est pas oracle, et d'autre part parce que les MAS sont estimées par récurrence, et l'espérance mathématique de l'équation (2.11) est remplacée par une moyenne temporelle. Dans ce travail, nous choisissons de mettre l'accent sur l'intégration de RN dans un algorithme distribué, et sur l'exploitation de l'information spatiale convoyée par les signaux compressés. Afin d'éviter la problématique de la convergence de l'algorithme en pratique, nous avons simplifié l'algorithme DANSE par deux aspects, en remettant l'étude de la convergence à un travail ultérieur.

Premièrement, les MAS sont estimées en bloc, c'est-à-dire par la moyenne temporelle des signaux sur toute leur durée [2]. Cela n'est pas réalisable en pratique, car il faudrait un traitement non causal (ou avec très grande latence). Soit $\mathcal{T}$ le nombre total de trames dans $\mathbf{y}_k$. Avec

$$\hat{\mathbf{s}}_k = \mathbf{m}_{\mathrm{s},k} \odot \mathbf{y}_k \tag{3.10}$$

2. Comme on le verra en sections 3.3.2 et 3.4.1, les signaux sont de durée relativement courte, entre 5 s et 10 s.

où $\odot$ est l'opérateur de multiplication terme-à-terme et $\mathbf{m}_{\mathrm{s},k}$ les masques associés à la parole de chaque signal de $\mathbf{y}_k$ (voir section 3.2.2), la MAS de la parole dans $\mathbf{y}_k$ est calculée de la manière suivante :

$$\mathbf{R}_{\mathrm{s},k}(f) = \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \hat{\mathbf{s}}_k(t,f)\hat{\mathbf{s}}_k^H(t,f)\,. \tag{3.11}$$

Celle du bruit est calculée de manière similaire :

$$\mathbf{R}_{\mathrm{n},k}(f) = \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \hat{\mathbf{n}}_k(t,f)\hat{\mathbf{n}}_k^H(t,f)\,, \tag{3.12}$$

où $\hat{\mathbf{n}}_k = \mathbf{m}_{\mathrm{n},k} \odot \mathbf{y}_k = (1 - \mathbf{m}_{\mathrm{s},k}) \odot \mathbf{y}_k$. Le calcul sur toute la fenêtre temporelle permet d'estimer précisément le contenu spectral des MAS, mais résulte en un filtre statique qui ne varie pas au cours du temps.

Le second aspect est que l'estimation itérative des filtres $\mathbf{w}_k$ et $\mathbf{w}_{kk}$ est remplacée par une estimation en deux étapes distinctes et successives, comme représenté en figure 3.2. Le filtre $\mathbf{w}_{kk}$ n'est donc pas estimé en même temps que $\mathbf{w}_k$ (cf. équation (3.6)), mais au cours de la première étape de filtrage local.

3.2.2. Utilisation de masques temps-fréquence

L'algorithme original de DANSE repose sur un DAV pleine bande, c'est-à-dire sur une fonction qui prend la valeur 1 lorsque la source cible est active et la valeur 0 lorsqu'elle ne l'est pas. Comme le DAV est pleine bande, quand la sortie vaut 1, la parole est considérée comme active sur toutes les bandes fréquentielles quel que soit son contenu fréquentiel et celui des interférences. Par ailleurs, le DAV est commun à tous les nœuds, et ne prend donc pas en compte la variabilité spatiale de l'activité de la parole. Dans Tango, un masque TF propre à chaque nœud k remplace donc le DAV. Pour chaque point TF, sa valeur cible est la suivante (Weninger et al., 2014) :

$$m_{\mathrm{s},k1}(t,f) = \frac{|s_{k,1}(t,f)|}{|s_{k,1}(t,f)| + |n_{k,1}(t,f)|}\,. \tag{3.13}$$

$s_{k,1}$ et $n_{k,1}$ sont respectivement les composantes de parole et de bruit dans le mélange du premier microphone du nœud k. Puisque ces grandeurs ne sont pas connues, le masque doit être estimé (cf. section 3.2.3).

Le premier avantage d'utiliser un masque TF est que l'information spectrale est utilisée en plus de l'information temporelle, ce qui permet une estimation plus précise des MAS. Un autre avantage est que la stationnarité du bruit n'est plus nécessaire pour l'estimation des MAS puisque, comme le montre l'équation (3.11), la MAS de la parole peut être estimée sans recourir à celle du bruit. L'intérêt, ou non, d'utiliser un masque TF est évalué en section 3.3.

Une illustration graphique de la différence entre un DAV et un masque TF est représentée en figure 3.3.

3.2.3. Utilisation de réseaux de neurones pour l'estimation des masques temps-fréquence

Puisque les signaux de parole et de bruit ne sont pas connus, il n'est pas possible de calculer directement le masque $m_{\mathrm{s},k1}$ de l'équation (3.13). Plutôt que d'estimer les composantes $s_{k,1}$ et $n_{k,1}$, nous proposons d'utiliser des RN pour prédire directement les masques TF nécessaires aux calculs des MAS. Par ailleurs, comme expliqué en section 3.2.4, nous proposons également d'utiliser des RN qui prédisent les masques TF à partir de plusieurs canaux afin d'améliorer leurs

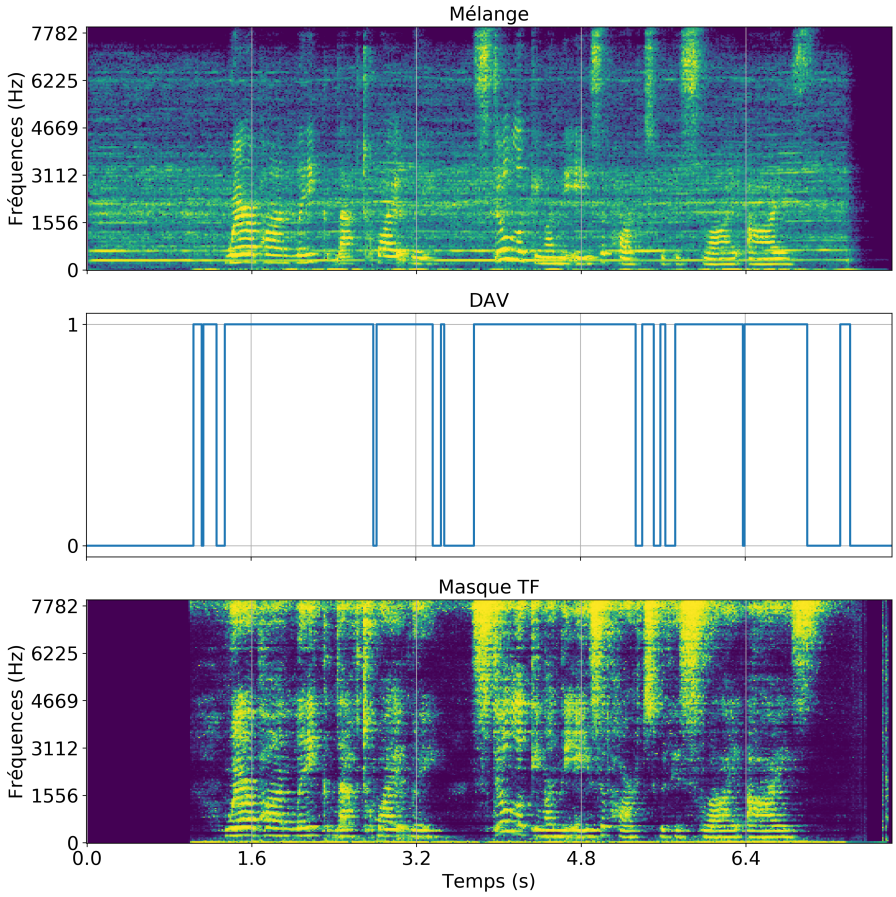


FIGURE 3.3. – Comparaison entre un DAV et un masque TF.

performances. Pour ne pas augmenter la complexité des RN, nous présentons une architecture de RN dont les premières couches sont des couches de convolution, de complexité presque indépendante du nombre de canaux en entrée. C'est donc une architecture adaptée à des contextes où plusieurs signaux sont utilisés pour prédire les masques. Différentes architectures sont comparées en section 3.3.2 pour estimer les masques TF nécessaires aux filtres des deux étapes.

3.2.4. Exploitation des signaux compressés pour l'estimation des masques

Aussi bien Doclo et al. que Bertrand et al. montrent par leurs expériences que les signaux compressés $\mathbf{z}_{-k}$ augmentent les performances d'un MWF opérant sur un seul nœud (Doclo et al., 2009; Bertrand and Moonen, 2009). Cela tient au fait que $\mathbf{z}_{-k}$ apporte une information très bénéfique, car elle provient de nœuds différents du nœud local (Bertrand and Moonen, 2009), et qui est de surcroît une estimation du signal cible (Doclo et al., 2009) [3].

Nous proposons d'utiliser les signaux compressés non seulement pour augmenter les performances du filtre, mais aussi pour mieux prédire les masques TF par les RN. Ainsi à la deuxième étape de Tango, le RN dispose de K fois plus de signaux pour prédire le masque, puisqu'il dispose d'un mélange local et de $K-1$ signaux compressés, reçus des $K-1$ nœuds distants. De plus, ces signaux sont pré-filtrés puisqu'ils sont issus de la première étape de filtrage. Ils apportent donc potentiellement une information plus fiable. Le masque de la seconde étape sera noté $\tilde{m}_{\mathrm{s},k1}$ dans la suite. Les RN prédisant les masques à partir du mélange local et des signaux compressés seront appelés réseaux de neurones *multi-nœuds* par la suite (RNMuN). La figure 3.4 représente un de ces RNMuN dans Tango appliqué à une antenne de deux nœuds. L'influence de l'information supplémentaire sur les performances du RN est analysée en sections 3.3.4 et 4.5.

3.2.5. Représentations schématique et algorithmique

Afin de visualiser plus facilement le fonctionnement de Tango, il est représenté sous forme graphique dans la figure 3.5 et sous forme algorithmique dans l'algorithme 2.

Algorithme 2 Pseudo-code de Tango.

procedure ETAPE 1
  pour $k = 1 \cdots K$
    Estimer $m_{\mathrm{s},k1}$ avec un RN à partir de $y_{k,1}$ (sur toute sa longueur)
    Calculer $\mathbf{R}_{\mathrm{s},k}$, $\mathbf{R}_{\mathrm{n},k}$ à partir de $\mathbf{y}_k$ (sur toute sa longueur)
    Calculer $\mathbf{w}_{kk} = \left(\mathbf{R}_{\mathrm{s},k} + \mu\mathbf{R}_{\mathrm{n},k}\right)^{-1}\mathbf{R}_{\mathrm{s},k}\mathbf{e}_1$
    Calculer $z_k = {\mathbf{w}_{kk}}^H\mathbf{y}_k$
  fin pour
fin procedure

procedure ETAPE 2
  pour $k = 1 \cdots K$
    Recevoir $\mathbf{z}_{-k}$ des nœuds distants
    Estimer $\tilde{m}_{\mathrm{s},k1}$ avec un RNMuN à partir de $\tilde{\mathbf{y}}_k = [y_{k,1}, \mathbf{z}_{-k}^T]^T$.
    Calculer $\mathbf{R}_{\tilde{\mathrm{s}},k}$, $\mathbf{R}_{\tilde{\mathrm{n}},k}$ à partir de $\tilde{\mathbf{y}}_k$ (sur toute sa longueur)
    Calculer $\mathbf{w}_k = \left(\mathbf{R}_{\tilde{\mathrm{s}},k} + \mu\mathbf{R}_{\tilde{\mathrm{n}},k}\right)^{-1}\mathbf{R}_{\tilde{\mathrm{s}},k}\mathbf{e}_1$
    Calculer $\hat{s}_k = \mathbf{w}_k^H\tilde{\mathbf{y}}_k$
  fin pour
fin procedure

3. En termes de théorie de l'information, on pourrait dire que l'entropie d'un jeu de signaux $\{y_{k,1}, z_j\}$ plus élevée que celle d'un jeu $\{y_{k,1}, y_{k,2}\}$ (Shannon, 1948).

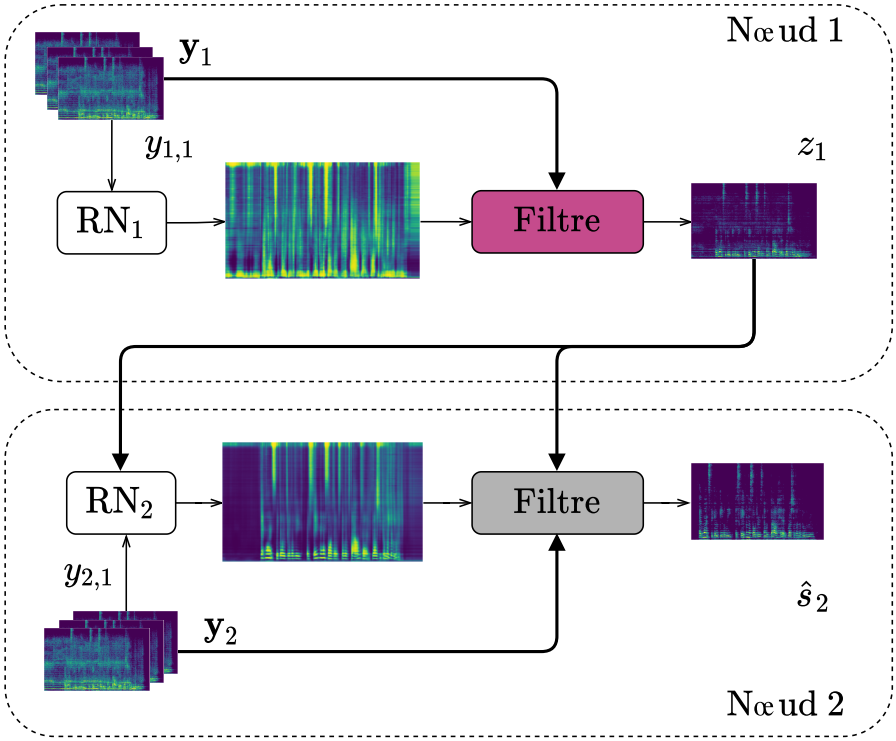


FIGURE 3.4. – Mise en évidence de l'utilisation des signaux compressés dans le système Tango pour la prédiction de masques TF, ici dans un contexte d'antenne à deux nœuds.

3.3. Validation de principe

Trois aspects du système Tango sont analysés dans cette section. Tout d'abord, nous montrons l'intérêt d'utiliser des masques TF plutôt qu'un DAV. Ensuite, nous montrons l'efficacité des couches convolutives pour les RNMuN. Enfin, nous montrons qu'utiliser les signaux compressés pour prédire les masques par des RN permet d'améliorer les performances de Tango, au point d'obtenir des performances supérieures à celles obtenues avec un DAV oracle.

3.3.1. Systèmes comparés

Pour analyser ces trois aspects, trois jeux d'expérience sont effectués et présentés dans la suite. Pour chacun d'entre eux, les filtres des différentes étapes de Tango sont toujours des MWF dont l'implémentation est basée sur la décomposition en valeurs propres généralisée (cf. équation (2.45)) en réduisant la matrice $\mathbf{\Sigma}$ à une matrice de rang 1. Cela augmente la réduction de bruit, quitte à augmenter la distorsion de la parole. Cette implémentation est également plus robuste dans les zones de faible rapport signal-à-bruit (RSB) (Serizel et al., 2014).

Utilisation de masques TF Le système Tango est comparé en conditions oracles, avec soit un DAV oracle, soit un masque TF oracle pour calculer les MAS de la parole et du bruit. Le masque TF oracle choisi est le masque de ratio idéal (MRI).

Utilisation de couches convolutives dans l'architecture des RN Deux architectures de RN sont comparées. La première est celle proposée par Heymann et al. (Heymann et al., 2016). Il s'agit d'un RN récurrent (RNN) composé d'une couche d'unités à longue mémoire à court terme

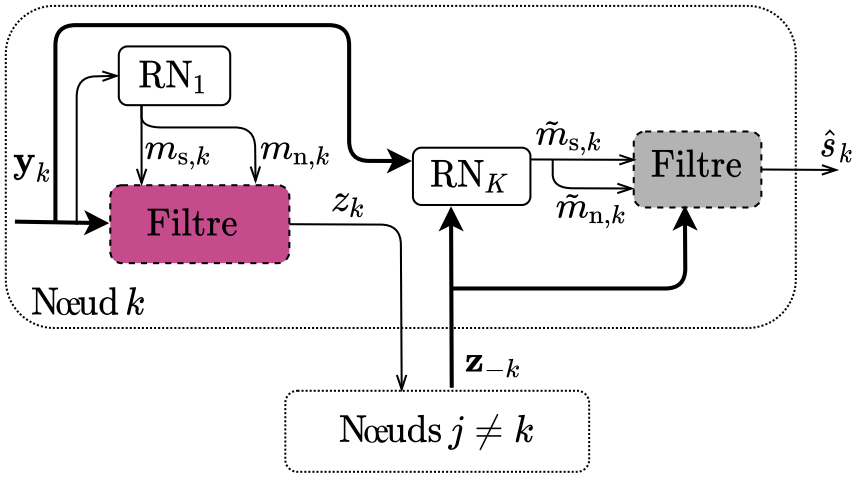
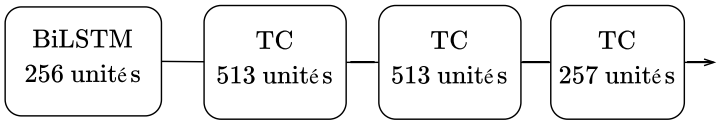
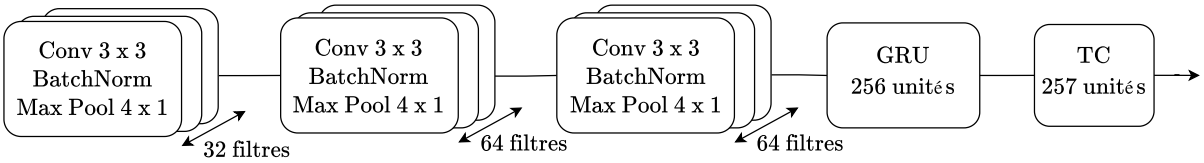


FIGURE 3.5. – Schéma de Tango.

(LSTM : *long short-term memory*) bidirectionnelle, suivie de trois couches toutes-connectées. La seconde architecture est un RN convolutionnel récurrent (CRNN) composé de trois blocs de convolution, suivis d'une couche d'unités récurrentes seuillées (GRU : *gated recurrent unit*) et d'une couche toute-connectée. Les blocs de convolution sont formés d'une couche de convolution avec un noyau de taille 3×3, suivie d'un normalisation par lots (*BatchNorm*) et d'un regroupement par sélection du maximum (*MaxPool*). Afin de conserver la résolution temporelle des tenseurs à l'entrée de la couche récurrente, les couches de MaxPool ont un noyau 4×1 pour que le regroupement ne soit appliqué que sur l'axe des fréquences. Les trois convolutions ont respectivement 32, 64 et 64 filtres. La couche GRU comporte 256 neurones et la couche toute-connectée en compte 257. La fonction d'activation de cette dernière est une sigmoïde. 21 trames de spectrogramme d'amplitudes sont fourniers à l'une et l'autre des architectures. La trame du milieu de la sortie du RN est retenue pour construire le masque de l'ensemble du signal. Ces deux architectures sont représentées en figure 3.6.



(a) Architecture du RNN de Heymann et al. (Heymann et al., 2016) adaptée à notre configuration.



(b) Architecture de notre CRNN.

FIGURE 3.6. – Représentation graphique des deux architectures de RN comparées.

Utilisation des signaux compressés pour la prédiction de masques Afin de mettre en évidence l'intérêt d'utiliser les signaux compressés pour prédire les masques TF, deux types de RN sont comparés. Le premier type rassemble les RN dits mono-nœuds, qui ne disposent que d'un mélange local pour prédire le masque. Ils seront dénommés « RNN$_1$ » ou « CRNN$_1$ » en fonction de s'il s'agit de RNN ou de CRNN. Le second type rassemble les RNMuN qui voient les signaux

compressés en plus du mélange local. Ils seront dénommés « RNN$_K$ » ou « CRNN$_K$ » avec K le nombre total de signaux dont ils disposent.

3.3.2. Corpus d'évaluation

Un premier corpus est créé afin d'évaluer les performances de Tango et d'étudier les trois facteurs présentés précédemment. Il comporte 11000 configurations différentes, dont 1000 sont réservées uniquement pour la validation, et 1000 autres pour l'évaluation. La durée des signaux étant prise aléatoirement entre 5 s et 10 s, cela correspond à un peu plus de 20 heures d'entraînement et 2 heures d'évaluation. Des réponses impulsionnelles (RI) sont simulées à l'aide de `Roomsimove` [4] et les signaux sources sont convolués avec ces RI. Dans le jeu d'entraînement, la largeur, longueur et hauteur des salles sont tirées aléatoirement de manière uniforme entre 3 m et 5 m, 3 m et 8 m et 2 m et 3 m respectivement. Deux nœuds de quatre microphones et séparés de 1 m sont placés dans la salle. Les microphones sont placés en carré à 10 cm du centre de chaque nœud. Le point $\mathcal{O}$, centre de la ligne reliant les deux nœuds, est placé aléatoirement dans la salle, sous la contrainte que les nœuds sont éloignés d'au moins 1 m du mur le plus proche. Deux sources, l'une de parole et l'autre de bruit, sont placées à 2.5 m de $\mathcal{O}$, en formant un angle $\alpha \in [25, 90]°$. Tous les microphones et sources sont à une hauteur de 1.5 m. La parole source est issue du corpus LibriSpeech (Panayotov et al., 2015) et le bruit est du bruit modulé par la parole (BMP). Les signaux sources de parole et de bruit sont décrits plus en détail en sections 3.4.1.1 et 3.4.1.2 respectivement. Le RSB des signaux sources est pris aléatoirement entre -5 dB et +15 dB. Le temps de réverbération (TR) est pris aléatoirement entre 300 ms et 600 ms.

L'ensemble d'évaluation est similaire à celui d'entraînement, à ceci près que la largeur et la longueur des salles sont contraintes à prendre des valeurs entières dans les mêmes intervalles que ceux d'entraînement. La hauteur de la salle est fixée à 2.5 m. L'angle α ne peut prendre (aléatoirement) que les valeurs fixes de $\{25, 45, 90\}°$. Enfin, le bruit est issu du corpus CHiME (Barker et al., 2015) dans l'environnement d'une rue ou d'une cafétéria.

Tous les signaux sont échantillonnés à une fréquence d'échantillonnage de 16 kHz. La transformée de Fourier à court terme (TFCT) est prise avec des fenêtres de Hann de 512 échantillons et un pas d'avancement de 256 échantillons. Les caractéristiques principales des configurations de ce corpus sont illustrées sur la figure 3.7. Par la suite, nous désignerons ce corpus sous le nom de *Majorette*.

3.3.3. Métriques

Comme décrit en section 2.6, aucune métrique ne reflète parfaitement à elle seule la réponse des algorithmes de rehaussement de la parole. Tout au long de cette thèse, nous quantifierons donc les différents résultats à l'aide de quatre métriques :

— **ΔSIR$_{\text{img}}$** Il s'agit de la différence entre le SIR de sortie et le SIR d'entrée, exprimé en dB. Les références nécessaires au calcul du SIR sont les signaux images (les signaux convolués, différents à chaque nœud), d'où l'indice « $_{\text{img}}$ ».

— **SAR$_{\text{img}}$** Il s'agit du SAR, exprimé en dB, dont les références sont les signaux images.

— **SAR$_{\text{src}}$** Il s'agit du SAR, exprimé en dB, dont les références sont les signaux sources (les signaux non convolués, communs à tous les nœuds), d'où l'indice « $_{\text{src}}$ ».

— **STOI$_{\text{img}}$** Il s'agit du STOI dont les références sont les signaux images.

4. homepages.loria.fr/evincent/software/Roomsimove_1.4.zip

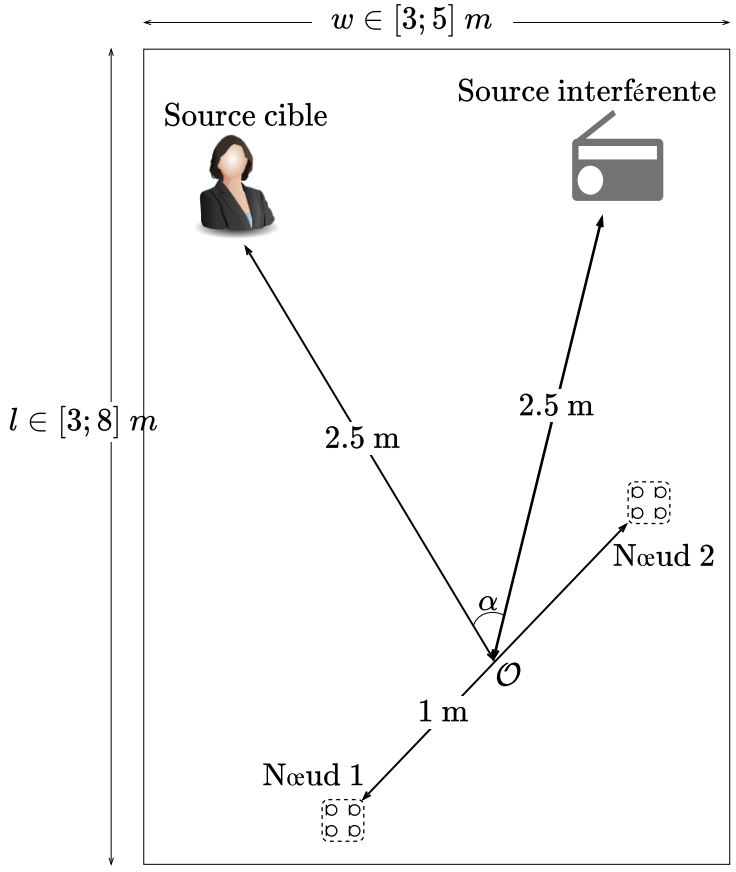


FIGURE 3.7. – Représentation graphique du corpus d'entraînement

Ce choix est motivé par plusieurs considérations. Tout d'abord, il est nécessaire de regarder à la fois le SIR et le SAR pour estimer à quel point la réduction de bruit a impacté la distorsion de la parole. Le STOI est également montré car c'est une métrique couramment utilisée dans le domaine du rehaussement de la parole et facilite la comparaison avec d'autres systèmes. De plus, il est pour certains plus révélateur de l'intelligibilité de la parole, là où le SAR et le SIR caractérisent la similarité entre le signal de sortie et ceux de référence, sans considération de la qualité ni de l'intelligibilité de la parole restituée. Par ailleurs, nous avons observé au cours d'expériences préliminaires que le SAR était très sensible aux références choisies pour le calculer, et qu'il ne correspondait pas toujours aux impressions subjectives lors d'écoutes informelles. Il s'avère en fait que le SAR est sensible aux distorsions liées à la réverbération. C'est pour cela que nous reportons les SAR_{img} et SAR_{src}, dont la différence reflète les distorsions liées à la réverbération.

3.3.4. Résultats et analyse

Les résultats sont indiqués dans le tableau 3.1. Nous y reportons la moyenne sur les 1000 configurations de test des métriques mesurées au niveau du nœud avec le meilleur SIR en sortie, accompagnée de l'intervalle de confiance à 95%. Le premier bloc du tableau présente les résultats lorsqu'un DAV ou des masques oracles sont utilisés. Le deuxième bloc présente les résultats lorsque des réseaux de neurones mono-nœuds (RNMoN) sont utilisés aux deux étapes de filtrage. Le troisième bloc présente les résultats lorsque des RNMoN sont utilisés à la première étape, et des RNMuN à la deuxième étape.

	$\Delta\text{SIR}_{\text{img}}$ (dB)	SAR_{img} (dB)	SAR_{src} (dB)	STOI_{img}
DAV oracle	16,4 ± 0,5	4,7 ± 0,3	4,0 ± 0,4	0,69 ± 0,01
Masques TF oracles	19,5 ± 0,4	8,0 ± 0,3	4,9 ± 0,4	0,76 ± 0,01
RNN_1	15,6 ± 0,5	6,5 ± 0,3	4,2 ± 0,4	0,72 ± 0,01
CRNN_1	16,1 ± 0,5	5,8 ± 0,3	4,2 ± 0,4	0,71 ± 0,01
RNN_2	15,7 ± 0,6	6,5 ± 0,3	4,1 ± 0,4	0,72 ± 0,01
CRNN_2	18,6 ± 0,5	6,4 ± 0,3	4,8 ± 0,4	0,73 ± 0,01

TABLEAU 3.1. – Résultats du rehaussement de la parole de Tango sur la base de données *Majorette*.

Les trois blocs permettent de répondre aux trois questions de la section 3.3.1. L'intérêt des masques TF est mis en évidence dans le premier bloc. La nette amélioration des performances avec les masques TF est permise par l'information supplémentaire qu'apportent les masques sur le plan spectral, qui conduit à des estimations des MAS plus précises.

En ce qui concerne l'intérêt d'utiliser une architecture de RN purement récurrente ou avec des couches convolutives, les deux derniers blocs montrent que le RNN est à peu près équivalent au CRNN en termes de rehaussement de la parole lorsque seuls les mélanges sont utilisés pour prédire les masques (RNN_1 Vs CRNN_1). Néanmoins, l'avantage d'utiliser des CRNN plutôt que des RNN est plus important et significatif dans le dernier bloc (RNN_2 Vs CRNN_2), où les RNMuN sont utilisés à la deuxième étape de filtrage. Le tableau 3.2 montre un autre avantage des structures convolutives : leur nombre de paramètres entraînables n'augmente qu'à peine lorsque le nombre de canaux en entrée double. Au contraire, pour utiliser l'information spatiale avec un RNN, il faut concaténer les canaux sur l'axe des fréquences, ce qui revient à presque doubler le nombre de paramètres du RN.

Architecture	Nombre de paramètres
RNN_1	1 717 773
$CRNN_1$	911 109
RNN_2	2 244 109
$CRNN_2$	911 397

TABLEAU 3.2. – Nombre de paramètres entraînables des différents RN utilisés.

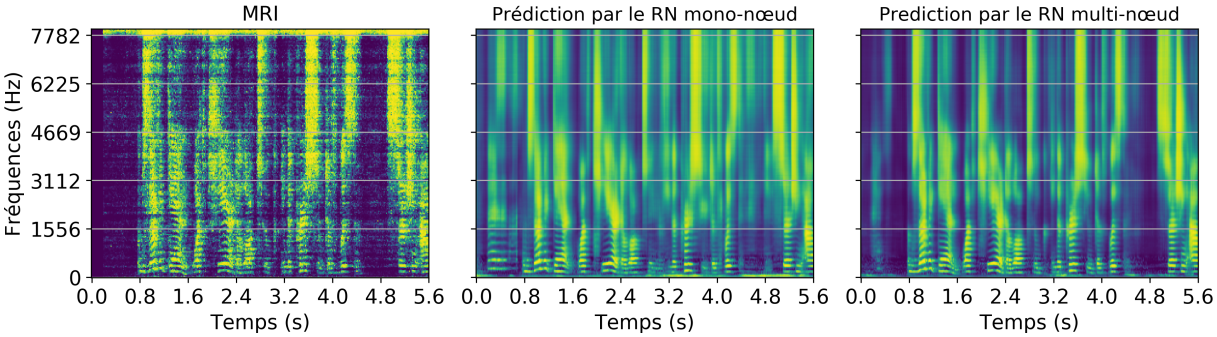


FIGURE 3.8. – Comparaison des masques prédits par un RNMoN et par un RN multi-nœud avec le MRI.

Par ailleurs, le tableau 3.1 montre qu'utiliser des couches convolutives permet de mieux exploiter l'information spatiale convoyée par les signaux compressés. En particulier, le ΔSIR_{img} augmente de près de 3 dB lorsque les signaux compressés sont fournis au RN. Notons qu'utiliser le CRNN multi-nœud permet de surpasser les performances obtenues avec un DAV oracle. Puisqu'avec des RNMuN le SIR augmente et que le SAR ne diminue pas, les signaux compressés sont utiles pour mieux réduire le bruit sans distordre la parole. D'après la figure 3.8, il semble que les masques prédits par le CRNN multi-nœud sont plus précis dans les zones de faible RSB, qui ont des valeurs proches de 0, alors que beaucoup de points TF correspondants ont des valeurs assez élevées dans le masque prédit par le réseau mono-nœud. Cela s'explique par le fait que les signaux compressés sont des signaux pré-filtrés qui apportent donc non seulement plus d'information (spatiale), mais en plus une information plus précise sur l'activité spectro-temporelle de la parole.

3.4. Description du corpus DISCO

Afin d'évaluer le système Tango sur des données plus représentatives de la réalité que celles simulées dans le corpus *Majorette*, un autre corpus a été simulé. Il rassemble de nombreux types de configurations que l'on pourrait rencontrer au quotidien lorsque quatre nœuds sont présents dans une salle. Il est nommé DISCO (DIStributed Semi-COnstrained microphone arrays). Les signaux sources mélangés et les salles simulées pour l'obtention des RI sont présentés dans la suite.

3.4.1. Présentation des signaux sources

Les jeux d'entraînement, de validation et d'évaluation du corpus présenté dans cette section ne diffèrent que par les signaux sources considérés. Ces signaux sont décrits dans cette section. Les configurations acoustiques (dimensions des pièces, positions des sources et microphones, temps de réverbération) sont identiques pour les différents jeux et présentées en section 3.4.2. Nous

supposons que les mélanges à débruiter proviennent de deux sources ponctuelles, l'une de parole et l'autre de bruit. Tous les signaux sources sont de durée prise aléatoirement entre 5 s et 10 s.

3.4.1.1. Signaux de parole

Les signaux de parole proviennent du jeu de données LibriSpeech (Panayotov et al., 2015). Il s'agit initialement de textes anglais lus à voix haute aussi bien par des hommes que par des femmes, pour des personnes mal-voyantes. Les livres et chapitres lus sont divisés en signaux de quelques secondes et leur transcription est disponible. Les enregistrements, faits à l'aide d'un microphone de proximité, sont de qualités variées. C'est pourquoi les signaux sont répartis par Panayotov et al. en deux catégories. La première est dite *propre* ; elle rassemble les enregistrements clairs et permettant une bonne reconnaissance automatique de la parole (moins de 13% de taux d'erreur sur les mots). La seconde catégorie est dite *autre* ; elle rassemble les signaux de moins bonne qualité (de 13% à 40% de taux d'erreur sur les mots). Dans le cadre du rehaussement de la parole, nous n'avons retenu que les signaux de la première catégorie afin d'avoir une vérité terrain de bonne qualité.

Les signaux de cette catégorie sont regroupés en trois jeux de données dans le corpus LibriSpeech, l'un d'entraînement (`train-clean`), l'autre de développement (`dev-clean`) et le dernier d'évaluation (`test-clean`). Le jeu d'entraînement `train-clean` est lui-même divisé en deux dossiers, l'un contenant 360 heures de données (`train-clean-360`), et l'autre 100 (`train-clean-100`). Etant donné que les RN utilisés pour le rehaussement de la parole ne nécessitent pas plus de 100 heures de données d'entraînement, nous n'avons retenu que le jeu de données de l'ensemble `train-clean-100` pour créer les jeux d'entraînement et de validation. Les signaux d'évaluation sont tous issus de l'ensemble `test-clean`. Le tableau 3.3 résume la répartition des signaux de parole des différents jeux d'entraînement, validation et évaluation du corpus Disco.

Jeu dans Disco	Dossier dans LibriSpeech	Nombre d'heures disponibles dans LibriSpeech	Nombre de signaux créés dans Disco	Nombre d'heures créées dans Disco
Entraînement	`train-clean-100`	100	10 000	30
Validation			1 000	3
Evaluation	`test-clean`	5,4	1 000	3

Tableau 3.3. – Répartition des signaux de parole dans les jeux d'entraînement, validation et évaluation du corpus Disco.

3.4.1.2. Signaux de bruit

Deux types de bruits sont considérés pour corrompre la parole. Les bruits du premier type sont des BMP et ceux du second sont des bruits réels.

Bruit modulé par la parole Le BMP est créé à partir des signaux prononcés par des locuteurs du dossier `train-clean-360` de LibriSpeech. Aucun de ces locuteurs n'apparaît donc dans les ensembles de parole de la source cible des jeux d'entraînement, validation ou d'évaluation car les locuteurs du dossier `train-clean-360` sont différents de ceux des dossiers `train-clean-100` et `test-clean`. De même, les locuteurs utilisés pour créer les BMP des signaux d'entraînement ne sont jamais sélectionnés pour créer les BMP des signaux d'évaluation, et inversement.

Pour créer un échantillon de BMP par lequel sera corrompu un signal de parole cible donné, au moins cinq signaux d'au moins cinq locuteurs différents, choisis aléatoirement dans le dossier `train-clean-360` de LibriSpeech, sont mis bout à bout de telle sorte que le signal ainsi concaténé soit aussi long que le signal cible. Dans le domaine de la transformée de Fourier à long terme, l'amplitude du signal ainsi concaténé est conservée, mais la phase est remplacée par une phase aléatoire suivant une loi uniforme dans l'intervalle $[-\pi; \pi]$. Une fois que le signal est transformé dans le domaine temporel, on obtient un signal dont l'amplitude spectrale est similaire à l'amplitude spectrale (moyenne) de la parole, mais dont l'allure temporelle est plus proche d'un bruit coloré (Bryson and Johansen, 1965). La figure 3.9 représente la forme d'onde d'un tel signal ainsi que son amplitude et sa phase dans le domaine fréquentiel.

Bruits réels Les bruits réels sont tous issus de la plateforme en ligne Freesound [5] qui regroupe une multitude d'enregistrements sous licence Creative Commons (Font et al., 2013). Afin d'utiliser des sons représentatifs des bruits du quotidien, nous avons téléchargé tous les fichiers correspondants aux catégories suivantes : machine à laver, aspirateur, (robot-) mixeur, ventilation, lave-vaisselle, bébé, imprimante, eau. Après avoir vérifié que les signaux téléchargés automatiquement correspondaient bien à leur catégorie supposée, nous avons ré-échantillonné les signaux à 16 kHz. Les différents canaux des signaux multicanaux ont été moyennés afin de n'avoir que des signaux monocanaux. Puis nous avons supprimé les zones de silence et divisé les longs signaux en parties de 10 s. Enfin, en nous assurant que les signaux téléversés sur le site par un même utilisateur étaient rangés dans le même jeu, nous avons réparti les signaux ainsi obtenus dans un jeu d'entraînement et un jeu d'évaluation. Du fait d'un grand déséquilibre entre les différentes catégories de bruit, les catégories contenant peu d'échantillons ont été réservées pour le jeu d'évaluation uniquement. Seules celles contenant plus de deux heures d'enregistrement ont été utilisées pour le jeu d'entraînement et celui de validation. Afin d'éviter qu'une catégorie soit sur-représentée, nous avons gardé au maximum quatre heures d'enregistrement pour chaque catégorie. Le tableau 3.4 résume les catégories téléchargées et la durée des enregistrements disponibles par catégorie.

Le code pour télécharger les fichiers à partir de l'interface de programme d'application Freesound est disponible en ligne [6]. Les fichiers après le traitement décrit, tels qu'utilisés pour être mélangés avec la parole source sont également disponibles [7].

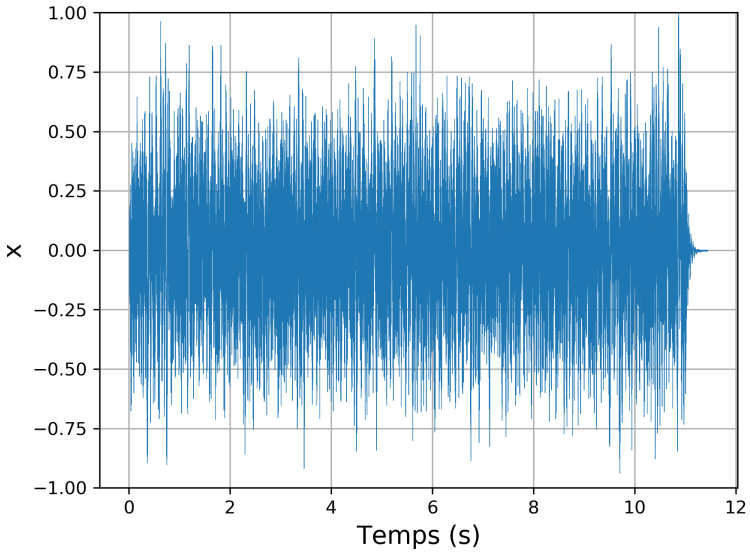
3.4.2. Présentation des configurations acoustiques

Les signaux utilisés pour entraîner les RN sont créés à partir de RI simulées avec la boîte à outils Pyroomacoustics (Scheibler et al., 2018). Les salles simulées sont toujours en forme de parallélépipède rectangle, dont la longueur l est prise aléatoirement entre 3 m et 8 m ; la largeur w est prise aléatoirement entre 3 m et 5 m ; la hauteur est prise aléatoirement entre 2.5 m et 3 m. Deux sources sont présentes, l'une de parole (c'est-à-dire que le signal source est un signal de parole non bruité) et l'autre de bruit (le signal est soit du BMP, soit un bruit de Freesound). Chaque configuration comporte quatre nœuds de quatre microphones chacun. Les microphones sont disposés en carré à 5 cm du centre du nœud.

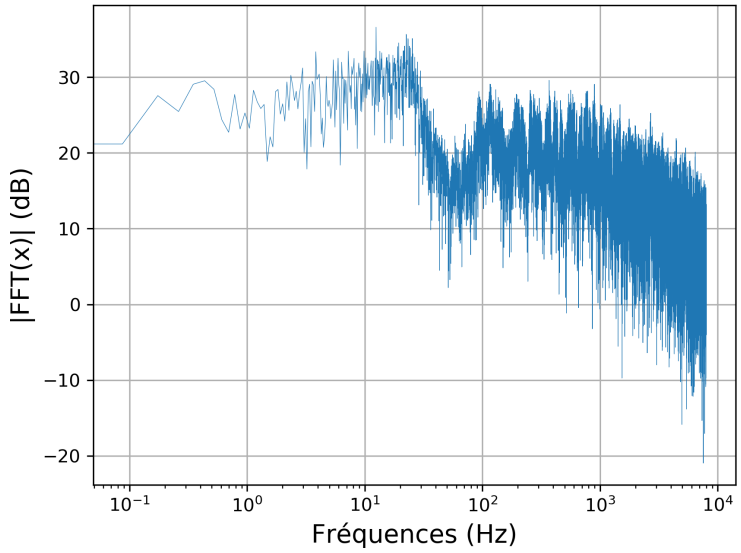
Le TR de chaque salle est pris aléatoirement entre 150 ms et 400 ms, comme dans une salle de taille moyenne à faible ou moyenne réverbération. Le signal de bruit source est multiplié par un gain aléatoire entre -6 dB et 0 dB. Cela mène à des RSB compris majoritairement entre -10 dB et

5. https://freesound.org/
6. https://github.com/nfurnon/disco/tree/master/dataset_generation/pre_generation
7. https://zenodo.org/record/4019030

(a) Forme d'onde d'un BMP



(b) Amplitude d'un BMP



(c) Phase d'un BMP

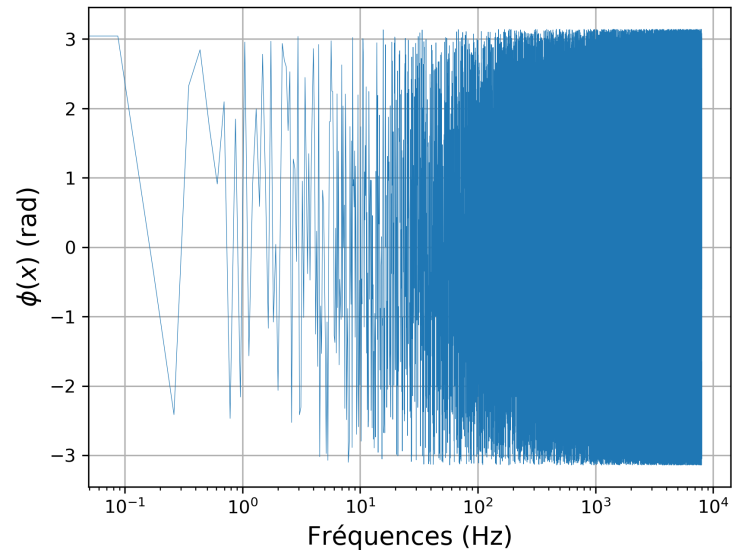


FIGURE 3.9. – Forme d'onde, amplitude et phase d'un bruit modulé par la parole.

(h)	Entraînement	Evaluation
Bébé	0	0,68
(Robot-)mixeur	0	0,39
Lave-vaisselle	2,19	0,11
Rasoir/ brosse à dents électrique	0	0,35
Ventilation	4,07	0,78
Friture	0	0,42
Imprimante	2,12	0,3
Aspirateur	2,03	0,24
Lave-linge	4,2	1,29
Eau	3,15	0,34
Total	**17,76**	**4,91**

TABLEAU 3.4. – Durées (en heures) des différentes catégories de bruits réels dans les jeux de données simulées du corpus DISCO.

+10 dB après la convolution avec les RI en fonction du nœud de la pièce. La position des sources et nœuds dans la salle n'est quasiment pas contrainte. Les sources et les nœuds sont disposés aléatoirement dans la salle avec la seule contrainte qu'ils soient à une distance δ de plus de 50 cm les uns des autres et des murs. La hauteur des nœuds est comprise entre 0,7 m et 2 m et celle des sources entre 1,2 m et 2 m. La figure 3.10 représente les caractéristiques de ces salles.

3.5. Optimisation des performances de Tango sur le corpus DISCO

Plusieurs facteurs, dont l'impact a été mis de côté dans un premier temps afin de se concentrer sur l'efficacité du système proposé, influencent les performances globales de Tango. Dans cette section, nous étudions l'influence de trois facteurs. Tout d'abord, nous étudions s'il convient de masquer les signaux compressés par le masque du nœud qui les émet ou par le masque du nœud qui les reçoit et les utilise pour calculer les MAS. Ensuite, nous analysons l'importance d'utiliser, ou non, une grande variabilité de bruits pendant l'entraînement des RN. Enfin, nous étudions s'il vaut mieux entraîner les réseaux de neurones multinœuds avec les signaux compressés issus d'un filtre oracle (c'est-à-dire calculé à partir de masques TF oracles) ou d'un filtre réel (c'est-à-dire calculé à partir de masques TF prédits par le RNMoN de la première étape de filtrage). Afin de s'assurer que les conclusions de cette section sont suffisamment généralisables, nous n'évaluerons plus le système sur le corpus *Majorette* mais sur le corpus DISCO présenté en section 3.4.

3.5.1. Choix du masque à appliquer sur les signaux compressés

3.5.1.1. Description de la problématique

En transposant l'équation (3.11) à la deuxième étape de filtrage au nœud k, on peut exprimer la MAS de la parole contenue dans $\tilde{\mathbf{y}}$ de la manière suivante :

$$\mathbf{R}_{\tilde{\mathbf{s}},k}(f) = \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \tilde{\mathbf{s}}_k(t,f) \tilde{\mathbf{s}}_k^H(t,f)\,, \tag{3.14}$$

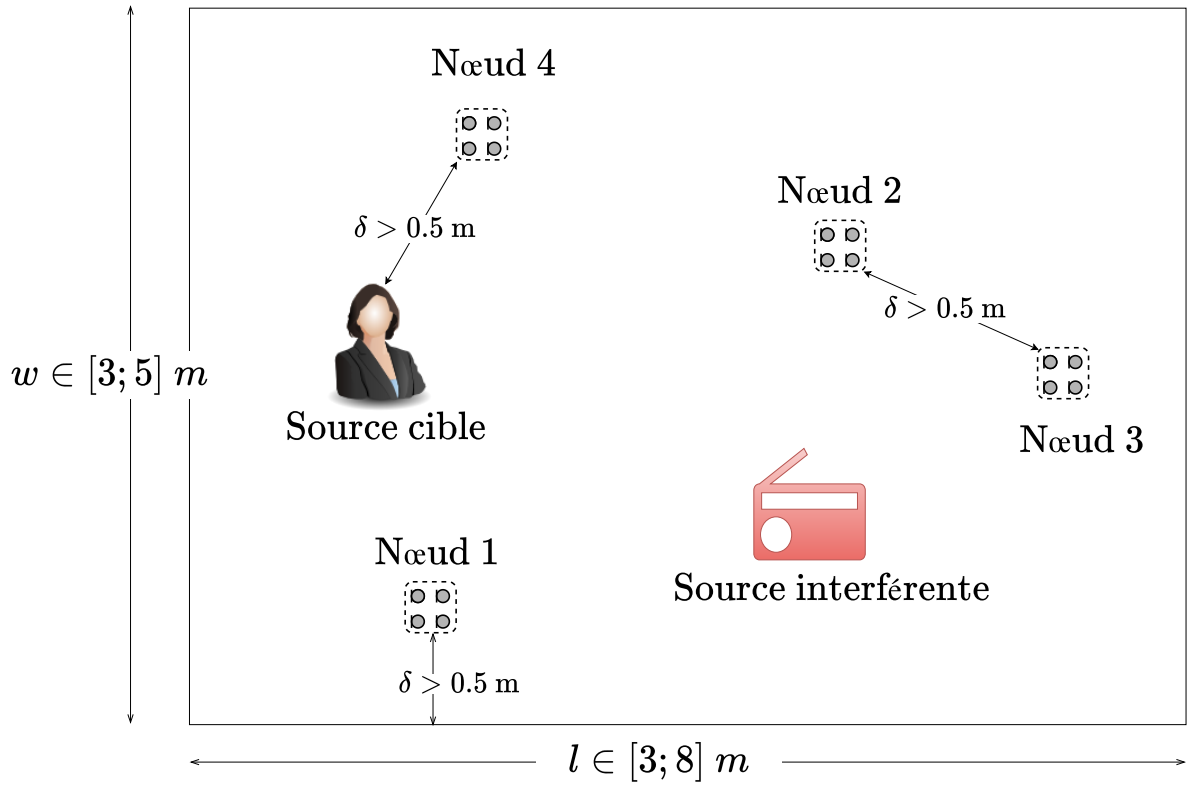


FIGURE 3.10. – Représentation des caractéristiques des salles du corpus DISCO.

avec $\tilde{\mathbf{s}}_k = \tilde{\mathbf{m}}_{\mathrm{s},k} \odot \tilde{\mathbf{y}}_k$ et où $\tilde{\mathbf{m}}_{\mathrm{s},k}$ regroupe les masques associés à la parole de chaque signal de $\tilde{\mathbf{y}}_k$. Rappelons ici l'équation (3.1), en supprimant les exposants d'itération :

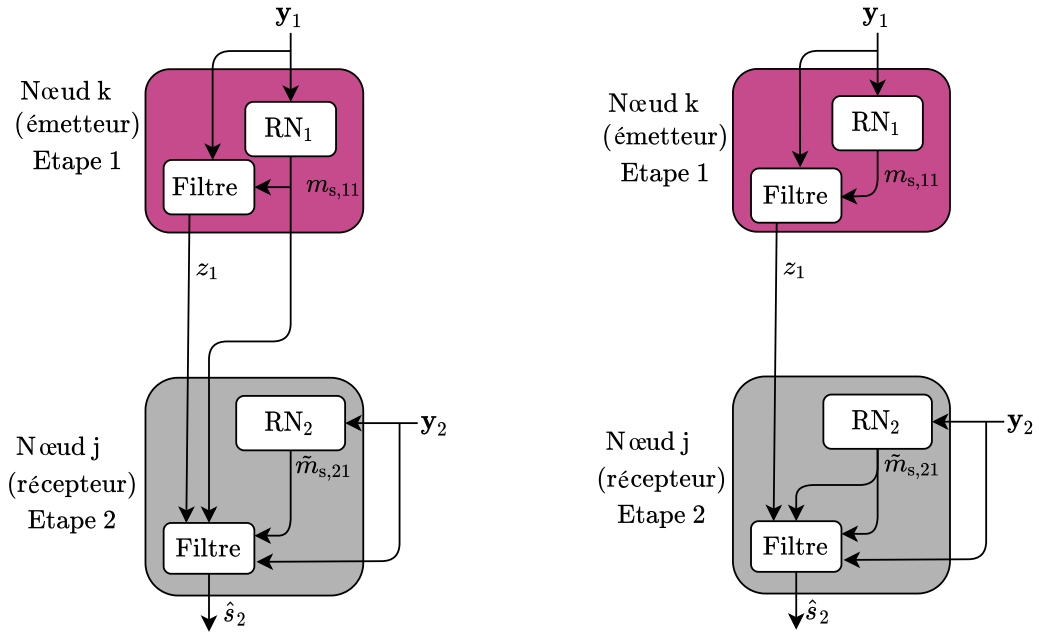
$$\tilde{\mathbf{y}}_k = \begin{bmatrix} \mathbf{y}_k \\ \mathbf{z}_{-k} \end{bmatrix}. \tag{3.15}$$

Etant donné que les signaux $\mathbf{y}_k$ de $\tilde{\mathbf{y}}_k$ proviennent du même nœud k, il est courant de prendre un seul masque à appliquer sur tous les canaux de $\mathbf{y}_k$ (Heymann et al., 2016; Weninger et al., 2014). On prendra par la suite le masque $m_{\mathrm{s},k1}$ associé au signal de référence, donc on peut simplifier l'équation (3.10) en :

$$\hat{\mathbf{s}}_k = m_{\mathrm{s},k1} \odot \mathbf{y}_k \,.$$

En revanche, les signaux $\mathbf{z}_{-k}$ dans $\tilde{\mathbf{y}}_k$ viennent de nœuds différents, voire éloignés du nœud k. Appliquer le même masque $m_{\mathrm{s},k1}$ n'est peut-être pas adapté car les nœuds peuvent observer des mélanges très différents. La question se pose alors de savoir s'il vaut mieux masquer les $\mathbf{z}_{-k}$ avec les masques $\{m_{\mathrm{s},j1}\}_{j\neq k}$ de tous les nœuds émetteurs, ou avec le masque $m_{\mathrm{s},k1}$ du nœud récepteur. L'avantage de la première variante est que les signaux compressés sont masqués par un masque adapté. L'inconvénient est que le masque doit être envoyé en même temps que le signal compressé, ce qui augmente de 50% les demandes en bande passante[8]. Dans la suite, on désignera sous le terme « émetteur » un nœud qui créé et envoie un signal compressé. On désignera sous le terme « récepteur » un nœud où les signaux compressés sont utilisés pour la seconde étape de filtrage. La figure 3.11 illustre la problématique de cette section.

8. Le masque ayant des valeurs réelles, il est encodé sur deux fois moins de bits que les signaux complexes z_j.



(a) Cas « distant » : masquage avec les masques du nœud émetteur

(b) Cas « local » : masquage avec le masque du nœud récepteur

FIGURE 3.11. – Représentation de comment masquer les signaux compressés avec le masque du nœud local peut économiser de la bande passante.

3.5.1.2. Expériences et résultats

Les deux variantes proposées ont été comparées en conditions oracles, c'est-à-dire avec des masques oracles à chaque étape de Tango. Les résultats sur le corpus DISCO sont reportés dans le tableau 3.5 où « local » fait référence au cas où les signaux compressés sont masqués avec le masque du nœud récepteur, et « distant » fait référence au cas où les signaux compressés sont masqués avec les masques de leur nœud émetteur.

TABLEAU 3.5. – Résultats du rehaussement de la parole lorsque les masques locaux ou distants sont utilisés pour masquer les signaux compressés. Les meilleurs résultats statistiquement significatifs sont renseignés en gras.

	$\Delta\mathrm{SIR}_{\mathrm{img}}$ (dB)	$\mathrm{SAR}_{\mathrm{img}}$ (dB)	$\mathrm{SAR}_{\mathrm{src}}$ (dB)	$\mathrm{STOI}_{\mathrm{img}}$
local	26,8 ± 0,4	**10,9 ± 0,2**	**9,6 ± 0,2**	**0,89 ± 0,004**
distant	26,1 ± 0,4	8,3 ± 0,2	9,0 ± 0,2	0,85 ± 0,004

Alors qu'on pourrait penser qu'utiliser les masques des nœuds distants conduit aux meilleurs résultats, cela n'est pas le cas en pratique. D'après les trois métriques $\mathrm{SAR}_{\mathrm{img}}$, $\mathrm{SAR}_{\mathrm{src}}$ et $\mathrm{STOI}_{\mathrm{img}}$, il vaut mieux utiliser le masque du nœud local, le $\Delta\mathrm{SIR}_{\mathrm{img}}$ ne permettant pas de départager les deux méthodes. Ce comportement s'explique probablement par le fait que le masque utilisé (oracle) « force » le signal compressé masqué à ressembler au signal de référence propre au nœud récepteur. Les résultats supérieurs avec le masque local ne reflèteraient donc pas réellement un meilleur rehaussement de la parole, mais une plus grande similitude entre le signal filtré et le signal de référence. Deux indices poussent à penser cela. Le premier est que le $\Delta\mathrm{SIR}_{\mathrm{img}}$, moins sensible à la référence considérée, n'est pas impacté par le choix du masque. Le deuxième est que la différence, très légèrement significative, entre les deux méthodes en termes de $\mathrm{SAR}_{\mathrm{src}}$, dont

les références sont communes aux deux méthodes, est plus faible qu'en termes de SAR_{img}, qui dépend de références différentes aux différents nœuds de l'antenne de microphones.

En conclusion de cette section, on peut masquer tous les signaux de $\tilde{\mathbf{y}}_k$ avec un seul masque, celui du nœud k, ce qui permet d'épargner de la bande passante, sans diminuer les performances de rehaussement de la parole.

3.5.2. Choix des signaux utilisés pour entraîner les réseaux de neurones

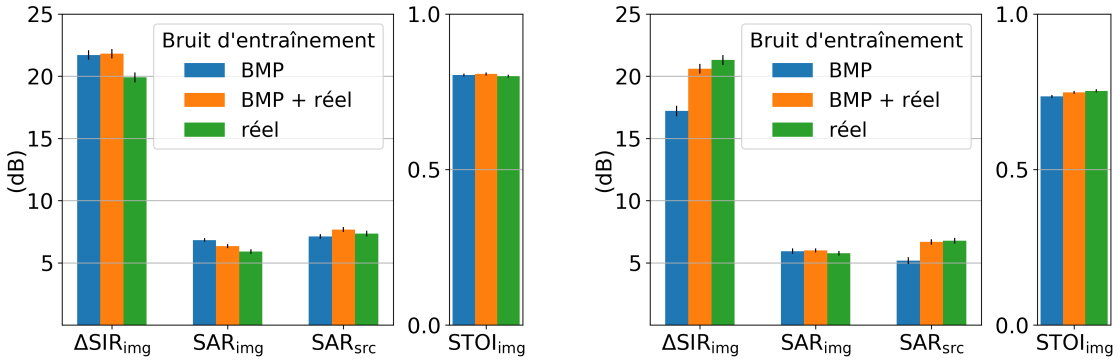
Les choix des signaux pour entraîner les RN de la section 3.3 n'ont pas été détaillés. Dans cette section, nous considérons différents bruits dans les mélanges pour voir quels types de bruits doivent être retenus pour entraîner les RN. Ensuite, nous étudions s'il est préférable d'entraîner les RNMuN avec des signaux compressés oracles (c'est-à-dire calculés lorsque le filtre de la première étape disposait de masques oracles) ou estimés (c'est-à-dire avec des masques prédits par un RN au cours de la première étape de filtrage).

3.5.2.1. Choix des bruits

Cette étude a pour but de savoir s'il vaut mieux spécialiser les RN sur un type de bruits en particulier ou si un entraînement plus général leur confère plus de résilience. Nous considérons deux catégories de bruits mélangés à la parole propre. La première catégorie rassemble des bruits dits réels, car ce sont des bruits réellement enregistrés dans notre quotidien. Ce sont les bruits présentés en section 3.4.1.2. Les bruits de la seconde catégorie sont des BMP ; il s'agit de bruits blancs dont on a modulé l'amplitude spectrale par celle de la parole, afin qu'il couvre majoritairement les régions fréquentielles où la parole est prépondérante. Parce qu'ils sont synthétiques, ces bruits sont faciles à obtenir, mais peu représentatifs des bruits réels.

Trois RN ont été entraînés. Le premier est entraîné uniquement avec des mélanges de BMP et de parole propre. Le deuxième ne voit que des mélanges avec des bruits réels. La moitié des mélanges utilisés pour l'entraînement du troisième est constituée de parole corrompue par du BMP. L'autre moitié des mélanges est constituée de parole corrompue par des bruits réels. Chacun de ces trois RN est évalué sur un corpus d'évaluation dont tous les signaux sont des mélanges de parole corrompue par du BMP et sur un second corpus dont tous les signaux sont des mélanges de parole corrompue par des bruits réels. Les signaux sont mélangés dans les salles du corpus Disco. Afin de simplifier l'étude, un même RNMoN est conservé aux deux étapes de filtrage de Tango. Ainsi, les signaux compressés à la seconde étape sont utilisés uniquement pour le calcul du filtre, mais pas pour l'estimation des masques. L'étude sur les signaux à considérer pour entraîner les RNMuN est faite dans la section suivante. Les résultats sur les deux corpus d'évaluation sont présentés dans les figures 3.12(a) et 3.12(b) où les tirets en haut de chaque barre représentent l'intervalle de confiance à 95%.

Si entraîner le RN avec de la parole corrompue par du BMP permet de bonnes performances lorsque l'interférence est également un mélange de parole et de BMP, ces performances généralisent mal en présence de bruits réels, comme le montre la baisse des résultats obtenus avec le réseau entraîné avec du BMP entre la figure 3.12(a) et la figure 3.12(b). De même, le RN entraîné uniquement avec de la parole corrompue par des bruits réels obtient certes les meilleurs résultats lorsque l'inférence est également de la parole corrompue par du bruit réel, mais d'une part, il ne se distingue pas de manière significative du RN entraîné avec les deux types de bruits, et d'autre part ses performances baissent de manière significative lorsque le bruit au test est du BMP. Seul le RN entraîné avec les deux types de bruits a toujours des résultats au moins aussi bons que le meilleur des deux autres RN entraînés avec des bruits spécifiques, et garde une performance constante sur les deux corpus. Ainsi, une plus grande variété de signaux pendant l'entraînement



(a) Evaluation avec de la parole corrompue par du BMP

(b) Evaluation avec de la parole corrompue par des bruits réels

FIGURE 3.12. – Résultats du rehaussement de la parole pour différents bruits d'entraînement et d'évaluation. Les tirets en haut de chaque barre représentent l'intervalle de confiance à 95%

des RN conduit à une plus grande résilience sans affecter la performance. Cette conclusion est similaire à ce qu'avaient observé Kolbæk et al. (Kolbæk et al., 2016). Nous sommes en plus en mesure de dire que retirer le BMP des bruits d'entraînement diminue la résilience des RN. C'est pourquoi dans toute la suite de cette thèse, les RN seront systématiquement entraînés sur des mélanges dont les bruits sont pour la moitié d'entre eux des BMP et pour l'autre moitié des bruits réels. Une étude plus détaillée des performances de Tango avec de tels RN en fonction des bruits d'évaluation est proposée en section 4.3.4.

3.5.2.2. Choix des signaux compressés pour entraîner le réseau multi-nœud

Deux possibilités s'offrent pour l'entraînement des RNMuN, qui ont besoin de signaux compressés en plus des mélanges locaux. Soit on utilise les signaux compressés issus de la première étape de filtrage lorsque les masques sont prédits par un RN. L'avantage de cette méthode est que les RNMuN sont entraînés dans des conditions similaires aux conditions d'évaluation. Soit on utilise les signaux compressés oracles, c'est-à-dire issus d'un filtre calculé à partir de masques oracles. Cette méthode a l'avantage d'être indépendante du RN utilisé à la première étape de filtrage, donc il n'est pas nécessaire de ré-entraîner un RNMuN à chaque fois qu'un nouveau RN est utilisé à la première étape. Deux RN ont été entraînés pour comparer ces deux possibilités. Le premier est entraîné avec des signaux compressés oracles. Le second est entraîné avec des signaux compressés estimés, c'est-à-dire calculés à l'aide du RNMoN qui a vu de la parole corrompue par du BMP et par des bruits réels. Les résultats de Tango avec des filtres calculés avec chacun de ces RN sur le corpus d'évaluation où la parole est corrompue par des bruits réels sont reportés dans le tableau 3.6. Une très faible différence ressort des résultats. Seuls le SAR$_{src}$ et le STOI$_{img}$ se distinguent légèrement entre les deux méthodes, et semblent indiquer qu'il vaut mieux entraîner un RNMuN avec des signaux compressés oracles plutôt qu'avec des signaux compressés prédits. Cela pourrait s'expliquer par le fait que les signaux compressés oracles soient moins bruités, et qu'ils assurent une meilleure convergence lors de l'entraînement des RNMuN. Toutefois, la différence de résultats entre les deux méthodes est faible et il est difficile de conclure avec plus de certitude. Dans la suite de cette thèse, nous entraînerons les RNMuN avec les signaux compressés oracles, d'une part parce que cette expérience semble indiquer de meilleures performances, et d'autre part parce que cette méthode est plus flexible, puisqu'elle ne demande pas de prédire

TABLEAU 3.6. – Résultats du rehaussement de la parole lorsque des signaux compressés oracles ou prédits sont utilisés pour entraîner les RNMuN. Les meilleurs résultats statistiquement significatifs sont indiqués en gras.

	$\Delta\text{SIR}_{\text{img}}$ (dB)	SAR_{img} (dB)	SAR_{src} (dB)	STOI_{img}
oracles	22,9 ± 0,5	**6,9 ± 0,1**	**8,5 ± 0,2**	**0,78 ± 0,004**
prédits	23,4 ± 0,5	6,6 ± 0,1	7,5 ± 0,2	0,76 ± 0,006

tous les signaux d'entraînement à chaque fois qu'un nouveau RN est utilisé à la première étape de filtrage.

3.6. Comparaison à l'état de l'art

Les sections précédentes ont montré que le système Tango profite d'une logique distribuée pour exploiter efficacement l'information spatiale. Si l'architecture en CRNN surpasse celle en RNN et amène des résultats *a priori* convenables, il est nécessaire de comparer notre méthode à une solution de l'état de l'art.

3.6.1. Systèmes comparés

Nous choisissons de comparer Tango au FaSNet de Luo et al. (Luo et al., 2019) qui est une solution de l'état de l'art, de bout-en-bout dont l'architecture a en plus été récemment revue afin d'en améliorer les résultats (Luo et al., 2020b). FaSNet repose sur une logique de formateur de faisceau, mais dont les différents filtres sont estimés en deux étapes par un RN dont les entrées sont les intercorrélations normalisées des différents signaux. Nous avons implémenté cette architecture à l'aide de la boîte à outils `Asteroid` (Pariente et al., 2020a).

L'architecture du CRNN dans Tango est celle présentée en section 3.3. Le RNMuN nécessaire à la seconde étape est entraîné avec les signaux compressés oracles. Cependant, même avec un seul RN, différents systèmes peuvent être utilisés pour rehausser la parole. En effet, comme vu dans la section 2.3.1.3 du chapitre précédent, il est possible de jouer à la fois sur le paramètre μ du SDW-MWF et sur le rang de la matrice $\mathbf{\Sigma}$ (cf. équation (2.45)) pour ajuster la réduction de bruit et la distorsion de la parole, sans pour autant changer de RN. Ainsi, nous comparons FaSNet avec quatre systèmes. Le premier utilise des masques oracles pour calculer le SDW-MWF avec $\mu = 1$ et la décomposition de rang 1 en valeurs propres généralisée de $\mathbf{R}_{\text{n}}^{-1}\mathbf{R}_{\text{s}}$ (le filtre est appelé r1-GEVD-SDW-MWF). Le deuxième système utilise le même filtre, mais avec des masques prédits par le CRNN. Le troisième système prédit les masques à partir d'un CRNN mais en choisissant $\mu = 5$ dans le SDW-MWF, en gardant la décomposition de rang 1 en valeurs propres généralisée de $\mathbf{R}_{\text{n}}^{-1}\mathbf{R}_{\text{s}}$. Enfin, le quatrième système prédit aussi les masques à partir d'un CRNN avec $\mu = 5$ dans le SDW-MWF, mais la matrice de décomposition en valeurs propres généralisée est prise de rang plein.

Tous les RN sont entraînés sur le même corpus d'apprentissage DISCO, avec le même nombre d'échantillons d'apprentissage, dont la moitié est un mélange de parole et BMP et l'autre moitié un mélange de parole et de bruits réels. Pendant l'inférence, FaSNet est appliqué sur chacun des nœuds et, pour chacune des 1000 configurations d'évaluation, les métriques au niveau du nœud (parmi les quatre présents) avec le meilleur SIR de sortie sont retenues pour comparaison avec Tango. De même, nous retenons les métriques obtenues avec Tango sur le nœud avec le meilleur SIR de sortie.

TABLEAU 3.7. – Comparaison de différentes variantes de Tango avec FaSNet (Luo et al., 2019) sur le corpus DISCO.

	$\Delta\text{SIR}_{\text{img}}$ (dB)	SAR_{img} (dB)	SAR_{src} (dB)	STOI_{img}
Tango (oracle) r1-GEVD-SDW-MWF $\mu = 1$	26,8 ± 0,4	10,9 ± 0,2	9,6 ± 0,2	0,89 ± 0,003
Tango (CRNN) r1-GEVD-SDW-MWF $\mu = 1$	22,9 ± 0,5	6,9 ± 0,1	8,5 ± 0,2	0,78 ± 0,004
Tango (CRNN) r1-GEVD-SDW-MWF $\mu = 5$	26,0 ± 0,4	6,7 ± 0,1	9,2 ± 0,2	0,74 ± 0,005
Tango (CRNN) GEVD-SDW-MWF $\mu = 5$	16,8 ± 0,5	13,2 ± 0,3	8,8 ± 0,3	0,83 ± 0,006
FaSNet (Luo et al., 2019)	17,5 ± 0,2	13,8 ± 0,2	6,7 ± 0,2	0,84 ± 0,005

3.6.2. Résultats et analyse

Les résultats des cinq systèmes sont reportés dans le tableau 3.7. En comparaison avec Tango oracle, FaSNet réduit moins le bruit, mais a une très bonne restitution de la parole, puisque les SAR_{img} et STOI_{img} sont élevés. Avec des masques prédits par le CRNN, Tango mène également à une réduction de bruit plus forte que FaSNet, mais au prix d'une assez forte distorsion de la parole puisque le SAR_{img} est plus faible de presque 7 dB et le STOI_{img} de 0,06, quoique le SAR_{src} soit supérieur, ce qui semble indiquer une plus forte déréverbération. Cependant, Tango présente l'avantage de permettre beaucoup de flexibilité et de pouvoir ajuster le compromis entre la réduction de bruit et la distorsion de la parole. En effet, en augmentant le paramètre μ, on peut décider d'insister plus sur la réduction de bruit, ce qui est confirmé par les résultats expérimentaux : avec le r1-GEVD-SDW-MWF et $\mu = 5$, le $\Delta\text{SIR}_{\text{img}}$ augmente de près de 3 dB par rapport au cas où $\mu = 1$. De manière intéressante, le SAR_{src} augmente légèrement, sans doute parce que certaines composantes de la réverbération sont associées au bruit dans le calcul des MAS. Cette hausse du SAR_{src} est néanmoins compensée par une baisse du STOI_{img} qui témoigne d'une plus faible intelligibilité de la parole. En choisissant la formulation du SDW-MWF avec $\mu = 5$ et en prenant la matrice de décomposition en valeurs propres généralisée de rang plein, Tango obtient des résultats très similaires à ceux de FaSNet. Il est donc possible de compenser une réduction trop agressive du bruit (lorsque μ augmente) en augmentant le rang de la matrice de décomposition en valeurs propres généralisée. En fonction du but désiré, il est donc possible d'ajuster la réduction de bruit et la distorsion de la parole dans Tango, une flexibilité que n'offre pas FaSNet.

Pour terminer cette section, nous comparons également la complexité des deux architectures de RN. Nous la quantifions en termes de nombre de paramètres entraînables et de consommation énergétique. Cette dernière est une grandeur importante dans les appareils à puissance et batterie limitées des AAAH, et attire de plus en plus l'attention dans le contexte du dérèglement climatique (Henderson et al., 2020; Anthony et al., 2020; Lacoste et al., 2019; Lottick et al., 2019). Puisque la consommation peut différer entre l'entraînement et l'inférence, nous séparons ces deux étapes en reportant la consommation des RN sur une époque (calculée sur la moyenne

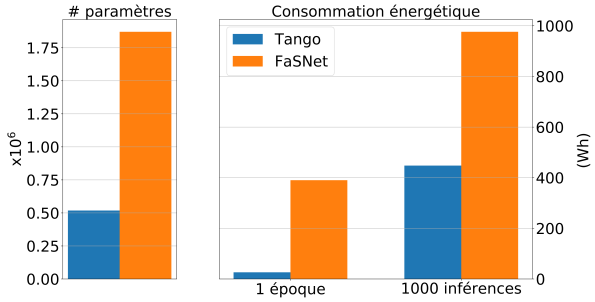


FIGURE 3.13. – Comparaison de la complexité de Tango et de FaSNet.

de 10 époques), ainsi que celle que représente le fait de filtrer les signaux de tout le jeu d'évaluation. Le nombre d'époques nécessaires aux deux architectures pour converger est sensiblement le même, compris entre 50 et 100. La consommation a été mesurée à l'aide de la boîte à outils `carbontracker` (Anthony et al., 2020) et les résultats sont représentés en figure 3.13.

La différence entre les deux systèmes est marquée. Tant à l'entraînement qu'à l'inférence, FaSNet, du fait de son nombre de paramètres bien supérieur, consomme entre 2 et 8 fois plus que Tango. Cela se traduit également par une durée d'entraînement ou d'inférence supérieure, que nous n'avons pas représentée ici. Si FaSNet est une solution envisageable pour le rehaussement de la parole, elle l'est moins dans le cadre des AAAH où une inférence rapide et une faible demande énergétique sont requises. Pour alléger ou accélérer les calculs, on pourrait « compresser » le RN, soit par distillation dans un RN (Hao et al., 2020; Chen et al., 2021), soit par quantification des poids du RN ou même par suppression des poids, des couches ou des canaux du RN (Tan et al., 2021b; Mishra et al., 2020). Notons toutefois que ces techniques doivent également considérer le matériel sur lequel les RN compressés opèrent, car toutes les architectures matérielles ne sont pas optimisées pour accélérer les calculs d'un RN compressé (Deng et al., 2020).

3.7. Conclusion

Dans ce chapitre, Tango, un système distribué de rehaussement de la parole, a été introduit. Nous avons montré qu'il permet d'utiliser des RN dans des antennes de microphones distribuées tout en exploitant l'information spatiale convoyée sous forme de signaux compressés d'un nœud à l'autre. Nous avons optimisé ses performances sur un corpus réaliste en choisissant le masque à appliquer sur les signaux compressés, les bruits vus à l'apprentissages et le type de signaux compressés pour entraîner les RNMuN. Comparé à l'état de l'art, Tango permet une grande flexibilité et légèreté que n'apportent pas les techniques de l'état de l'art de bout-en-bout, sans pour autant perdre en performance. Ces conclusions sont listées dans le tableau 3.8.

Section	Points-clé
3.1	Présentation de DANSE (Bertrand and Moonen, 2010a,b).
3.2	Présentation de Tango, une approche distribuée de rehaussement de la parole.
3.3	• Utiliser des masques TF amène de meilleures performances qu'un DAV. • Les couches de convolution permettent d'exploiter efficacement l'information de canaux supplémentaires. • Les signaux compressés permettent au RN de prédire plus précisément les masques TF.
3.4	Présentation du corpus d'antennes de microphones Disco.
3.5.1	On peut appliquer sur les signaux compressés le masque TF du nœud local.
3.5.2.1	La présence de BMP et de bruits réels dans les mélanges d'entraînement conduit à une meilleure généralisation des RN.
3.5.2.2	Utiliser les signaux compressés oracles permet un entraînement des RN plus flexible que les signaux prédits, sans pour autant perdre en performance.
3.6	Tango a des performances comparables à l'état de l'art mais apporte une flexibilité et une légèreté que n'ont pas les solutions de bout-en-bout.

Tableau 3.8. – Points-clé à retenir du chapitre 3.

4. Analyse expérimentale détaillée de Tango

Le chapitre précédent a présenté Tango, un système distribué de rehaussement de la parole. Nous avons montré que Tango obtenait de bonnes performances sur deux corpus d'antennes acoustiques ad-hoc et était compétitif par rapport à l'état de l'art. Dans ce chapitre, nous proposons d'étudier plus en détail le comportement de notre solution sur une grande variété de configurations spatiales et acoustiques simulées, et de le confronter à des données réelles. Après une rapide présentation des configurations simulées d'entraînement et de validation en section 4.1, la résilience aux configurations spatiales est étudiée en section 4.2. La résilience de Tango dans différentes conditions acoustiques est présentée en section 4.3, et le comportement de Tango sur données réelles est présenté en section 4.4. Enfin, nous montrons dans la section 4.5 en quoi les signaux compressés sont un support efficace pour l'échange d'information spatiale et qu'ils peuvent aussi être des estimations du bruit, en fonction des configurations spatiales.

4.1. Présentation des configurations d'évaluation

Pour évaluer Tango dans les conditions les plus réalistes et variées possibles, différents jeux de données ont été créés afin de représenter différents types de configurations dans lesquelles le système peut opérer. Deux autres configurations, en plus de celle présentée en section 3.4 sont simulées. Ces trois configurations seront également utilisées dans le but d'évaluer la résilience de Tango aux variations de configurations, c'est-à-dire de voir s'il est important d'entraîner les réseaux de neurones (RN) de Tango dans la même configuration spatiale que celle dans laquelle ils sont utilisés.

Les signaux sources des deux nouvelles configurations sont les mêmes que ceux de la première configuration présentée en section 3.4.1, ainsi que les dimensions des salles, les temps de réverbération et les rapports signal-à-bruit (RSB), mais la position des sources et des microphones est plus contrainte.

Configuration *aléatoire* La configuration présentée en section 3.4, de par sa conception peu contrainte, sera appelée configuration *aléatoire*. La figure 4.1 représente l'une des salles de la configuration *aléatoire*.

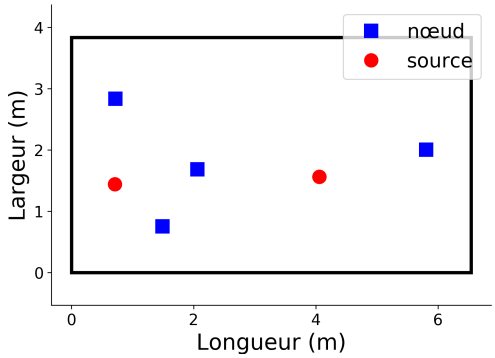


FIGURE 4.1. – Représentation 2D (vue du dessus) d'une instance de la configuration *aléatoire* dans le corpus DISCO.

Configuration *séjour* La deuxième configuration simule des scénarios plus contraints, proches d'une salle de séjour. Trois des quatre appareils enregistrant la scène acoustique sont disposés à moins de 50 cm d'un mur, comme s'ils étaient posés sur des meubles (buffet, commode, etc.). Le quatrième appareil est placé aléatoirement dans le reste de la pièce, à plus de 50 cm des murs et des autres appareils. Tous les appareils sont à une hauteur aléatoire entre 0,7 m et 0,95 m. Les deux sources sont placées aléatoirement dans la pièce, à plus de 50 cm des appareils et murs, et à une hauteur comprise entre 1.2 m et 2 m. La figure 4.2 représente l'une des salles de la configuration *séjour*.

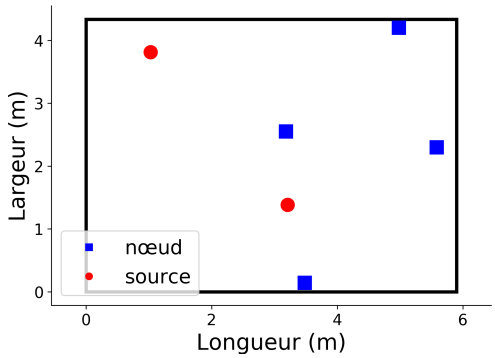


FIGURE 4.2. – Représentation 2D (vue du dessus) d'une instance de la configuration *séjour* dans le corpus DISCO.

Configuration *réunion* Le dernier scénario spatial simule une situation typique de réunion. Une personne est assise à une table et parle devant plusieurs appareils (ordinateur ou téléphone portable par exemple) posés en cercle sur la table et dont les microphones enregistrent la scène acoustique. Une autre source interférente est également située autour de la table. Nous avons donc disposé les quatre nœuds à $\frac{\pi}{2}$ rad les uns des autres, à une distance comprise entre 5 cm et 20 cm du bord d'une table. Les effets acoustiques de la table n'ont pas été pris en compte. La table (virtuelle donc) a un rayon compris entre 0,5 m et 1 m et une hauteur comprise entre 0,7 m et 0,8 m. Les deux sources sont disposées autour de la table, à un angle au moins supérieur à $\frac{\pi}{8}$ rad l'une de l'autre et à une hauteur comprise entre 1.15 m et 1.3 m. Elles sont placées à une distance inférieure à 50 cm du bord de la table, et à au moins 15 cm du mur le plus proche. La figure 4.3 représente l'une des salles de la configuration *réunion*.

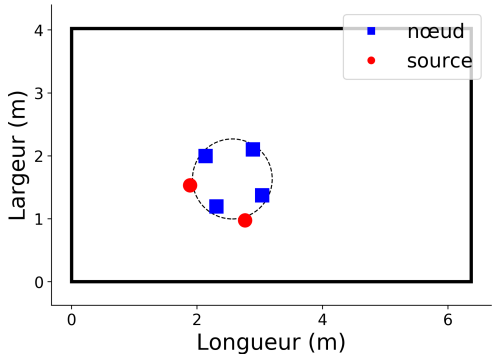


FIGURE 4.3. – Représentation 2D (vue du dessus) d'une instance de la configuration *réunion* dans le corpus DISCO.

4.2. Résilience aux variations de configurations spatiales

Dans ce chapitre ainsi que dans tout le reste de cette thèse, tous les RN utilisés sont des RN convolutionnels récurrents (CRNN) tels que décrits en section 3.3. De même, dans toute la suite de cette thèse, conformément aux conclusions de la section 3.5.2.1, les bruits utilisés pour corrompre la parole des mélanges d'apprentissage sont pour moitié des bruits modulés par la parole (BMP) et pour l'autre moitié les bruits réels décrits en section 3.4.1.2. Les filtres sont des filtres de Wiener multicanaux (MWF) basés sur la décomposition en valeurs propres généralisée des matrices d'autocorrélation spatiales avec $\mu = 1$ (cf. section 2.3.1.3) afin d'avoir une implémentation robuste même en conditions de faible RSB sans trop insister sur la réduction de bruit. Enfin, sauf mention contraire, les résultats présentés sont ceux obtenus au niveau du nœud avec le meilleur SIR_{img} de sortie. Ce nœud sera appelé par la suite « meilleur nœud ».

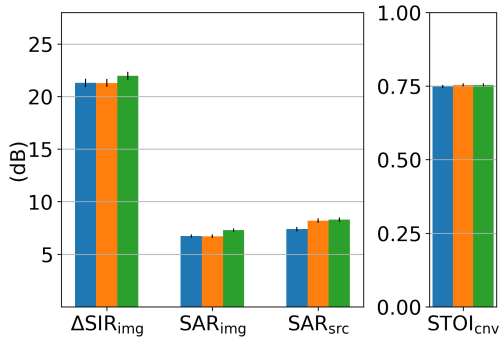
Les réseaux de neurones multinœuds (RNMuN) reçoivent des signaux compressés qui contiennent une information spatiale, donc dépendante de la configuration spatiale. Il se peut donc que le comportement des RNMuN dans des configurations spatiales variées diffère du comportement des réseaux de neurones mono-nœuds (RNMoN). C'est pourquoi la résilience de Tango à des conditions spatiales différentes est analysée en deux temps. Dans un premier temps, nous évaluons les performances de Tango lorsque des RNMoN sont utilisés aux deux étapes de filtrage. A la seconde étape de filtrage, les signaux compressés ne sont donc utilisés qu'au niveau du filtre et pas au niveau de la prédiction des masques. Dans un second temps, les RN à la seconde étape de filtrage sont remplacés par des RNMuN, afin de retrouver la configuration originale de Tango où les signaux compressés servent aussi à prédire les masques.

Pour effectuer ces analyses, dans chaque cas (RNMoN ou multinœuds), trois RN sont entraînés. Le premier est entraîné sur la configuration *aléatoire*. Le deuxième est entraîné sur la configuration *séjour*. Le dernier est entraîné sur la configuration *réunion*. Chacun de ces RN est évalué sur les trois jeux d'évaluation.

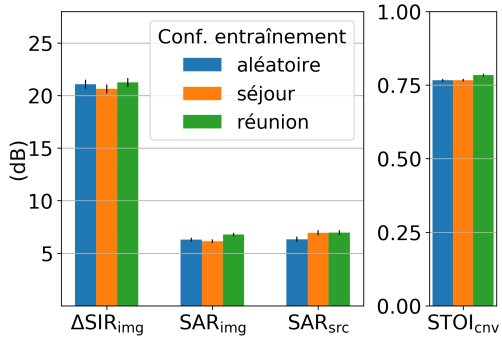
4.2.1. Performances avec des réseaux de neurones mono-nœuds

Les performance de Tango lorsque des RNMoN sont utilisés aux deux étapes de filtrage sont représentées dans la figure 4.4.

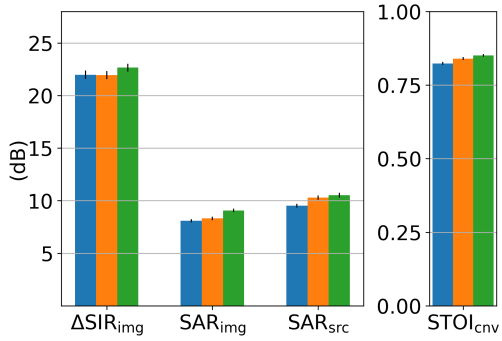
La première observation est que les performances de Tango, même avec des RNMoN, sont bonnes puisque le SIR_{img} augmente de plus de 20 dB quelle que soit la configuration d'entraînement et d'évaluation. Les SAR_{img} et SAR_{src} sont plus faibles, car la matrice de décomposition en valeurs propres généralisée du MWF est réduite à une matrice de rang 1 (cf. section 3.6). Ces



(a) Configuration *aléatoire*

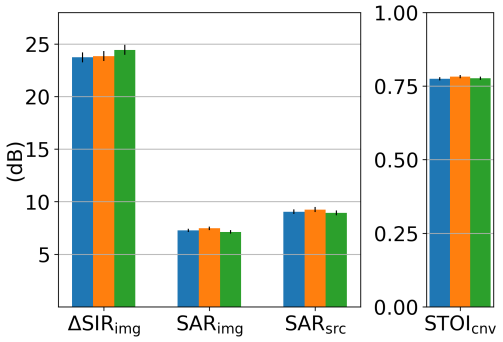


(b) Configuration *séjour*

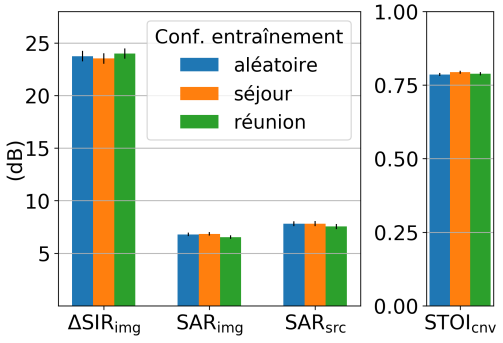


(c) Configuration *réunion*

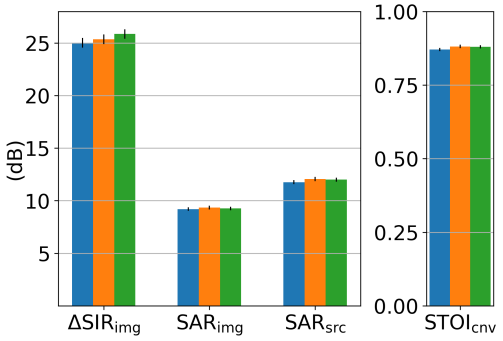
FIGURE 4.4. – Résultats du rehaussement de la parole de Tango dans les trois configurations spatiales avec des RNMoN entraînés sur une des trois configurations spatiales.



(a) Configuration *aléatoire*



(b) Configuration *séjour*



(c) Configuration *réunion*

FIGURE 4.5. – Résultats du rehaussement de la parole de Tango dans les trois configurations spatiales avec des RNMuN entraînés sur une des trois configurations spatiales.

métriques restent assez élevées et montrent que la parole n'est pas trop dégradée. De même, le STOI_{img}, autour de 75%, témoigne d'une bonne intelligibilité de la parole.

Par ailleurs, on peut observer que les performances sont très similaires sur les trois configurations spatiales. On peut noter des SAR légèrement supérieurs dans la configuration *réunion* que dans les deux autres configurations. Cela est probablement dû au fait que les microphones sont placés plus proches des sources, et donc que la parole est moins dégradée dans ce scénario.

Enfin, aucun des trois RN ne semble apporter d'amélioration significative par rapport aux deux autres. On observe certes que le RN entraîné sur la configuration *réunion* montre des performances très légèrement supérieures en termes de SIR_{img} et SAR_{img} dans les jeux d'évaluation *aléatoire* et *réunion*. La configuration *réunion* contient des signaux d'entraînement plus difficiles, avec des RSB plus faibles, car les microphones y sont plus proches des sources que dans les deux autres configurations spatiales. Cela pourrait expliquer que le RN soit mieux entraîné. Cependant, les différences sont si faibles qu'elles ne permettent pas vraiment de conclure que le RN entraîné sur la configuration *réunion* mène à de meilleures performances que les deux autres, d'autant moins que cela n'est pas confirmé avec les autres métriques.

4.2.2. Performances avec des réseaux de neurones multinœuds

Les performance de Tango lorsque des RNMuN sont utilisés à la seconde étape de filtrage sont représentées dans la figure 4.5. Tout d'abord, en comparant les figures 4.4 et 4.5, il est intéressant de noter qu'utiliser les signaux compressés pour prédire les masques par les RN permet d'augmenter les performances, ce qui confirme les résultats de la section 3.3.4. Cela sera détaillé en section 4.5.

Les conclusions de cette étude sont similaires à celles du paragraphe précédent. Les performances de Tango sont bonnes et similaires quelles que soient les configurations spatiales d'entraînement et d'évaluation. Avec les RNMuN, le léger avantage d'utiliser le RN entraîné sur la configuration spatiale *réunion* disparaît. Avec cette expérience, on peut confirmer que Tango est résilient aux variations de configurations spatiales et qu'entraîner ses RN sur l'une ou l'autre des configurations spatiales ne fait pas de différence. En vertu de quoi, dans la suite, les évaluations porteront toutes sur le même jeu d'évaluation, celui de la configuration *aléatoire*. De même, les RN seront tous entraînés sur le même jeu d'entraînement, celui de la configuration *aléatoire*.

4.3. Influence des conditions acoustiques

Dans la suite, les RN utilisés dans Tango sont des CRNN monocanaux à la première étape de filtrage et des CRNN multicanaux à la seconde étape de filtrage. Cette section analyse en détail les performances de Tango en fonction de trois facteurs. Le premier est la réverbération ; le second est le rapport source-à-interférences SIR des mélanges ; enfin, le troisième est la présence de bruit diffus.

4.3.1. Influence de la réverbération

Les performances de rehaussement de la parole de Tango en fonction du temps de réverbération (TR) sont présentées en figure 4.6. Deux jeux d'évaluation sont utilisés. Le premier est le jeu d'évaluation de la configuration *aléatoire*. Le second a les mêmes caractéristiques que le premier jeu d'évaluation, à la différence des TR qui sont tirés de manière aléatoire (uniforme) entre 400 ms et 700 ms. Ainsi Tango est aussi évalué sur des TR plus élevés que ceux que les RN ont vus pendant l'entraînement. Les résultats sont présentés en termes de $\Delta\text{SIR}_{\text{img}}$, SAR_{img}, SAR_{src} et STOI_{img}, ainsi qu'avec les valeurs absolues de SIR_{img} et STOI_{img} des signaux de sortie, afin

de mieux montrer l'influence de la réverbération sur les signaux de sortie. Ces deux métriques sont respectivement indiquées par $\text{SIR}_{\text{img, out}}$ et $\text{STOI}_{\text{img, out}}$ pour éviter les confusions avec les métriques calculées sur les signaux avant filtrage. Le SIR_{img} avant filtrage sera donc de même indiqué par $\text{SIR}_{\text{img, in}}$

Les quatre métriques indiquent le même phénomène : les performances de Tango sont quasiment constantes en fonction du TR. La légère baisse des SAR_{img}, SAR_{src} et $\text{STOI}_{\text{img, out}}$ est compensée par les performances quasiment constantes en termes de $\Delta\text{SIR}_{\text{img}}$ et $\Delta\text{STOI}_{\text{img}}$. Si les résultats absolus diminuent en fonction du TR, les résultats relatifs restent, eux, constants. Cela indique que la baisse de performance n'est pas directement liée à l'algorithme, mais au fait que les conditions initiales sont plus difficiles. Notons que les performances restent stables même lorsque la réverbération est supérieure à celle vue à l'entraînement (barres bleu foncé sur la figure).

En conclusion, Tango est résilient à une variété de conditions de réverbérations, même lorsque les conditions d'évaluation diffèrent légèrement des conditions d'entraînement.

4.3.2. Influence du SIR d'entrée

Dans cette section, nous analysons l'impact du SIR des signaux d'entrée sur les performances de Tango. Notons qu'en l'absence de bruit de microphones, le SIR tel que décrit par Vincent et al. est assimilable au RSB, bien qu'il ne soit pas calculé de la même manière (les deux métriques ne sont donc pas égales, mais s'interprètent de la même manière). Les résultats de rehaussement de la parole de Tango en fonction du SIR d'entrée $\text{SIR}_{\text{img, in}}$ sont représentés en figure 4.7.

Contrairement au TR, lorsque le $\text{SIR}_{\text{img, in}}$ augmente, la tâche de rehaussement de la parole est plus simple. C'est pourquoi on observe cette fois-ci une augmentation des performances absolues, en termes de $\text{SIR}_{\text{img, out}}$, SAR_{img}, SAR_{src} et $\text{STOI}_{\text{img, out}}$. En revanche, lorsqu'on considère les $\Delta\text{SIR}_{\text{img}}$ et $\Delta\text{STOI}_{\text{img}}$, les performances relatives diminuent (figures 4.7(b) et 4.7(f)). Cela a déjà été observé dans d'autres travaux de recherche (Souden et al., 2013; Li et al., 2016). Comme l'indiquent Kolbæk et al., on pourrait augmenter les performances d'un RN en l'entraînant sur la plage de SIR sur lequel il sera évalué (Kolbæk et al., 2016). Cela a néanmoins l'inconvénient de diminuer la capacité de généralisation du RN alors que ceux que nous utilisons semblent performants sur une large variété de SIR. Par ailleurs, cela requiert la connaissance *a priori* du SIR d'évaluation, ce qui n'est pas toujours possible.

Etant donné la relativement grande couverture spatiale des nœuds dans les salles simulées, les SIR d'entrée aux différents nœuds sont en réalité assez éloignés les uns des autres. Ainsi, les SIR des meilleurs nœuds de sortie, en abscisse des figures 4.7, ne sont pas les mêmes aux autres nœuds de l'antenne de microphones. La figure 4.8 représente l'histogramme des SIR au niveau du nœud avec le pire SIR en entrée et au niveau du nœud avec le meilleur SIR en entrée dans chaque configuration du jeu d'évaluation. On observe une différence de près de 5 dB entre les moyennes de SIR mesurés à ces deux nœuds. Il se peut donc que le comportement de Tango varie en fonction du nœud sur lequel on l'analyse. Une analyse plus précise est proposée dans ce sens en section 4.5.3.

4.3.3. Résilience à un bruit diffus

L'hypothèse faite jusqu'à maintenant que seule une source de bruit, ponctuelle, est présente dans la scène acoustique n'est presque jamais vérifiée en pratique. Même lorsqu'une seule source ponctuelle est active, il y a toujours un *bruit diffus* dans une pièce, que l'on pourrait décrire comme un bruit de fond. Le bruit diffus provient d'une ou plusieurs sources dont les signaux sont plusieurs fois réverbérés dans la pièce. Il n'est donc plus vraiment localisé et semble provenir de toutes les directions à la fois. Même si ce bruit diffus est en général de faible intensité, il

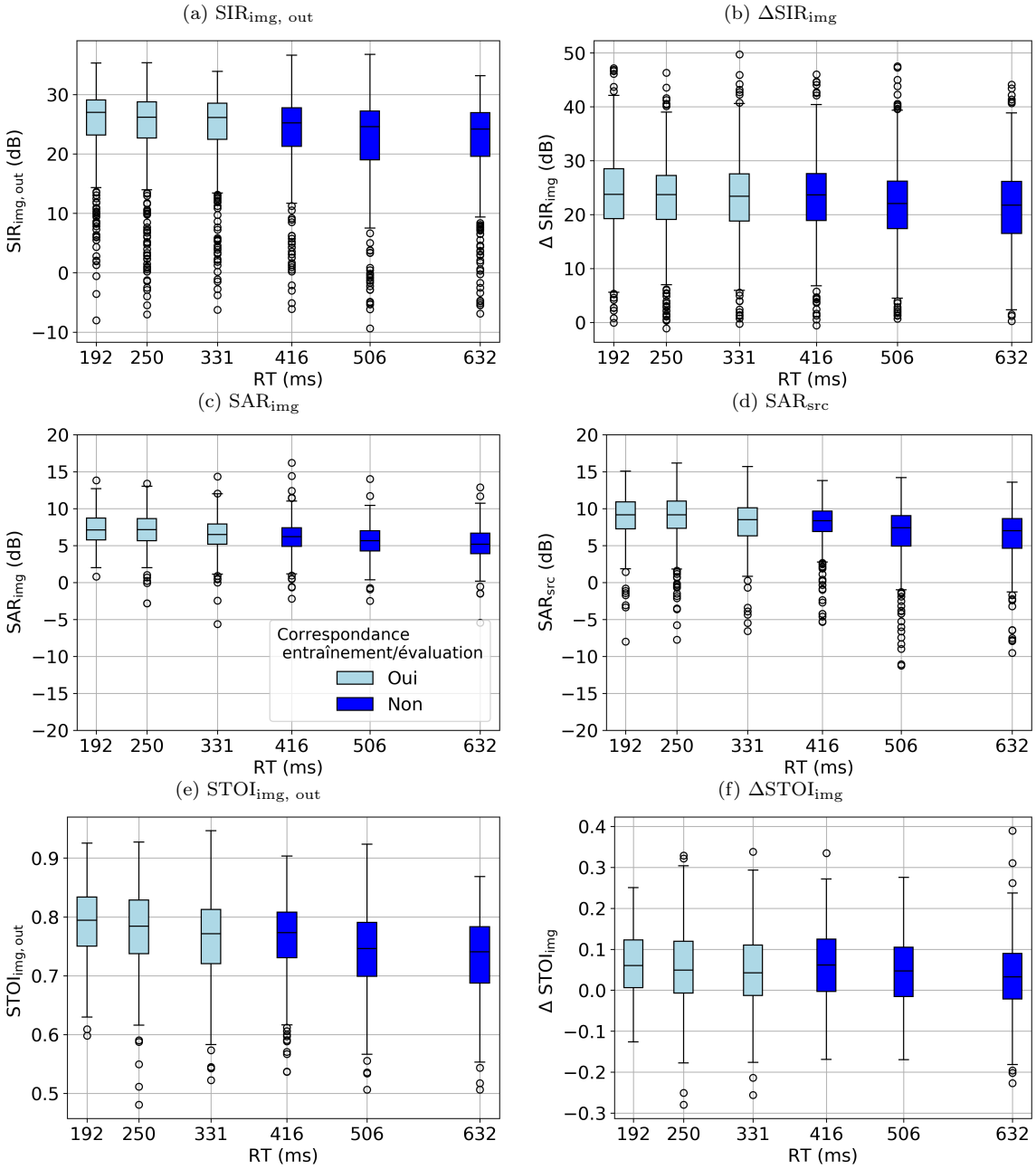


FIGURE 4.6. – Résultats du rehaussement de la parole au meilleur nœud de Tango en fonction du TR. Les résultats en bleu clair sont obtenus à des TR communs entre l'entraînement et l'évaluation. Les résultats en bleu foncé sont obtenus à des TR supérieurs à l'évaluation que ceux vus à l'entraînement.

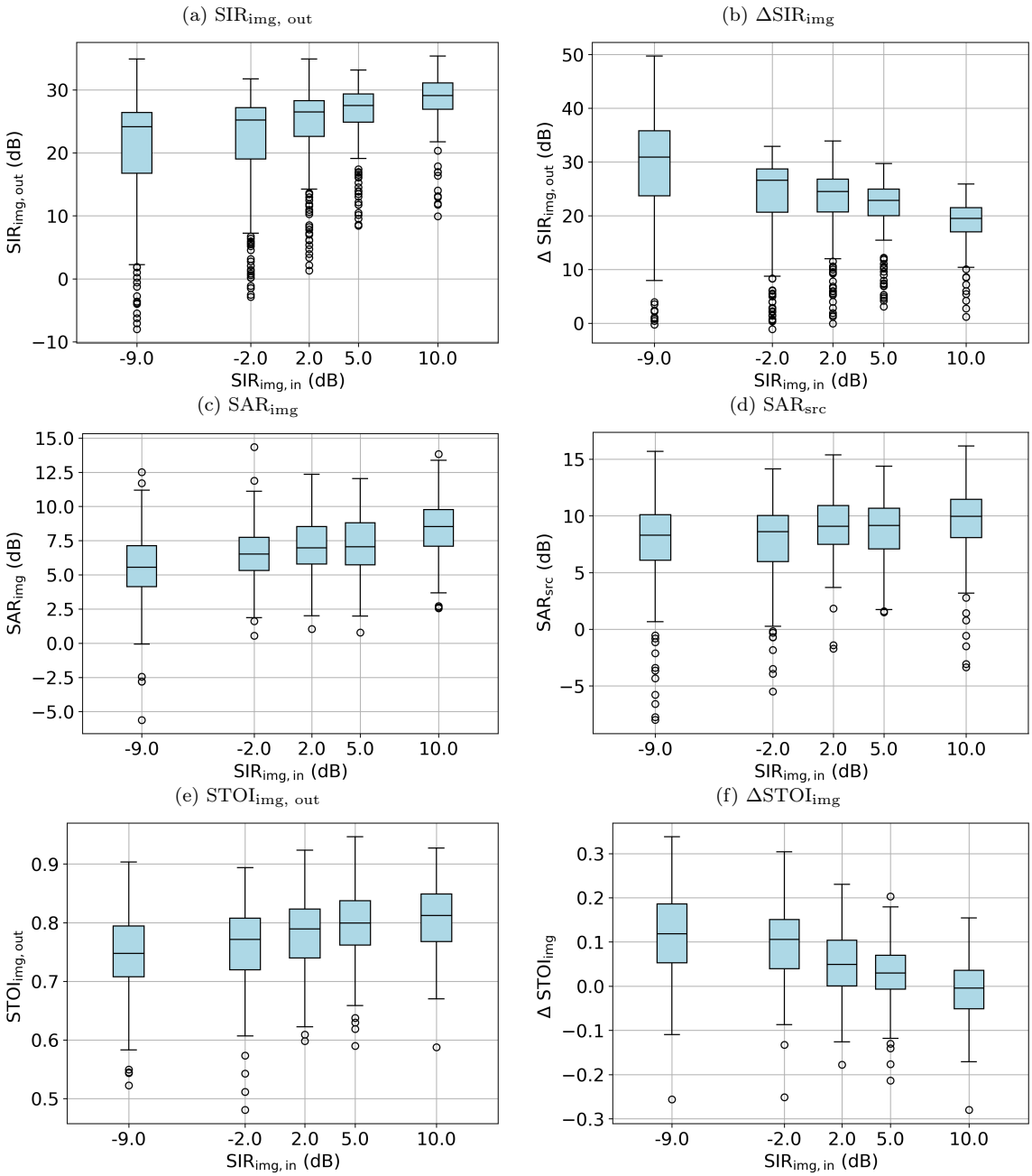


FIGURE 4.7. – Résultats du rehaussement de la parole au meilleur nœud de Tango en fonction du SIR d'entrée.

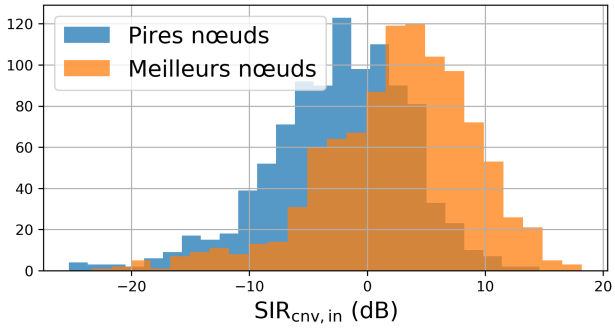


FIGURE 4.8. – Histogramme du SIR au niveau des meilleurs et des pires nœuds de chaque configuration du jeu d'évaluation.

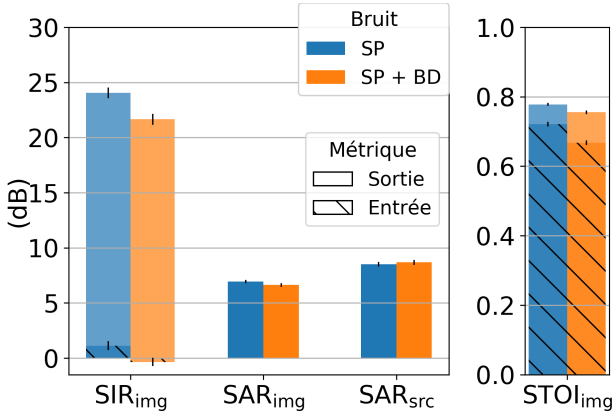


FIGURE 4.9. – Résultats de rehaussement de la parole de Tango avec bruit diffus (SP + BD) et sans bruit diffus (SP).

représente une difficulté pour les formateurs de voies qui ne peuvent pas atténuer un son venant de toutes les directions à la fois (à moins d'atténuer le signal cible également). Il convient donc d'étudier Tango dans le cas où un bruit diffus est présent, afin de voir si cela constitue une limite à son utilisation.

Un bruit diffus peut être simulé en convoluant une source de bruit ponctuelle avec la moyenne de la queue de réverbération de réponses impulsionnelles (RI) associées à plusieurs microphones répartis dans la pièce. Nous avons choisi 5 microphones, répartis aléatoirement dans les scènes simulées de la configuration spatiale *aléatoire*. Les signaux sources convolués sont des bruits ambiants issus d'environnements d'intérieur (maison, bureau, bibliothèque) du corpus TUT (Mesaros et al., 2016). Une fois convolués, les signaux sont ajoutés au mélange déjà créé de la configuration spatiale *aléatoire* à un RSB tiré aléatoirement entre 0 dB et 20 dB. La puissance de la parole nécessaire au calcul du RSB est prise comme la moyenne des puissances de la parole des seize microphones présents dans chaque salle. Les résultats de Tango sur le nouveau jeu d'évaluation ainsi créé sont présentés en figure 4.9. Nous y rappelons également les résultats obtenus dans le cas où seule une source ponctuelle est présente. La légende « SP » indique que seule une source ponctuelle est présente. La légende « SP + BD » indique que le bruit diffus est simulé en plus de la source ponctuelle. Afin de comparer les métriques avant et après filtrage, le $\text{SIR}_{\text{img, in}}$ et le $\text{STOI}_{\text{img, in}}$ sont représentés par les barres hachées. Les barres pleines représentent les métriques calculées sur les signaux en sortie de filtrage. Les RN nécessaires à Tango sont les mêmes que ceux des sections précédentes : ils n'ont pas été ré-entraînés dans les cas où un bruit diffus est présent.

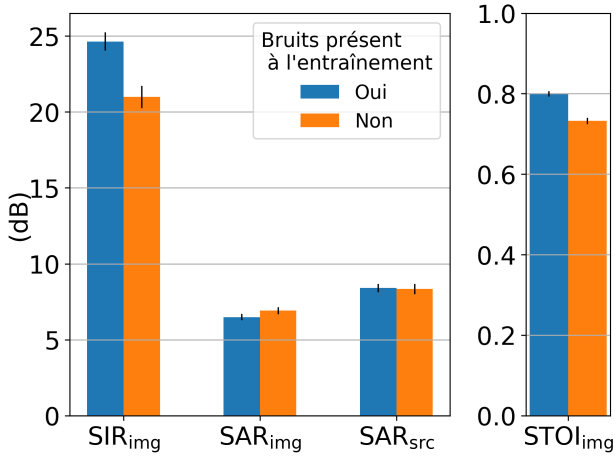


FIGURE 4.10. – Résultats de rehaussement de la parole de Tango lorsque les bruits d'interférence étaient présents, ou non, dans le jeu d'entraînement.

Même ainsi, les performances de Tango sont bonnes, et diminuent à peine en comparaison des cas où le bruit diffus n'est pas présent. La diminution des performances est plus marquée en termes de SIR qu'en termes de SAR. Intuitivement, cela pourrait s'expliquer par le fait que le faisceau du formateur de voies pointe toujours dans la direction de la source cible ; le signal de parole n'est donc pas distordu, le bruit diffus venant de cette direction « entre » dans le faisceau également. Par ailleurs, il est intéressant de noter que si le STOI diminue en présence de bruit diffus, la différence entre le STOI de sortie et le STOI d'entrée est plus importante en présence de bruit diffus qu'en son absence. Ainsi, la performance relative de Tango est meilleure en présence de bruit diffus.

En conclusion, nous pouvons affirmer que Tango est résilient à la présence de bruit diffus, même s'il ne peut pas annuler toutes les composantes de ce bruit non localisé.

4.3.4. Résilience à de nouveaux bruits

Il a été vu en section 3.5.2.1 que les RN les plus performants l'étaient lorsqu'ils étaient entraînés avec une grande variété de bruits. Nous avons également montré que certains des bruits d'évaluation n'ont pas été retenus dans le jeu d'entraînement, si bien que Tango est évalué sur des bruits avec lesquels les RN n'ont jamais été entraînés. Dans cette section, nous analysons plus en détail les performances de Tango sur l'ensemble d'évaluation, en séparant les résultats obtenus avec des bruits déjà vus pendant l'entraînement de ceux obtenus avec des bruits nouveaux. Les résultats sur ces deux catégories de bruits sont représentés en figure 4.10.

Malgré des résultats similaires en termes de SAR, les métriques $\Delta\text{SIR}_{\text{img}}$ et STOI_{img} montrent que les performances sont significativement plus faibles lorsque les interférences sont des bruits qui ne sont pas représentés dans le jeu d'entraînement. Cela confirme les résultats précédents qu'il est important d'avoir la plus grande variété possible de bruits dans le jeu d'entraînement. On peut toutefois nuancer les résultats moins bons sur les bruits non vus à l'entraînement en notant que les performances de Tango sur ces bruits restent bonnes, puisque le SIR_{img} augmente de plus de 20 dB sur cette catégorie de bruits, et que le STOI_{img} reste supérieur à 70%.

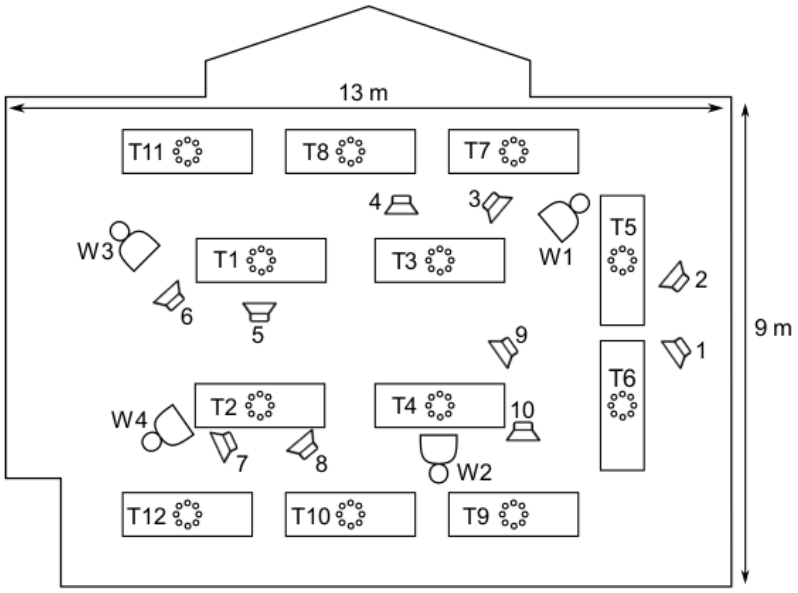


FIGURE 4.11. – Configuration spatiale de la pièce d'enregistrements réels. Les éléments T1−T12 et W1−W4 sont des antennes de microphones ; les sources sont indiquées par un symbole de haut-parleur. Figure issue de la documentation du corpus avec l'autorisation de l'auteur (Corey et al., 2019).

4.4. Evaluation en conditions réelles

Bien que réalistes et simulées à l'aide de modèles physiques reconnus, les RI utilisées jusqu'ici n'en sont pas moins synthétiques et ne reproduisent que de manière limitée la réverbération d'une pièce. Il a déjà été vu que les performances en conditions simulées sont surestimées par rapport à ce qu'elles sont en conditions réelles (Ceolini et al., 2020), c'est pourquoi nous évaluons dans cette section Tango avec des RI réelles.

4.4.1. Présentation des données réelles

Les RI proviennent du jeu de données de Corey et al. (Corey et al., 2019), et ont été mesurées dans une grande pièce dans laquelle étaient placés 10 sources et 160 microphones. Les 160 microphones étaient répartis sur 4 antennes de 16 microphones disposées sur quatre mannequin (W1−W4) et 12 antennes circulaires de 8 microphones (T1−T12) posées sur des tables. Le TR est approximativement donné à 800 ms. La figure 4.11, issue de la documentation du jeu de données (Corey et al., 2019), représente la pièce d'enregistrement. Le jeu de données comporte des mélanges, des signaux images non sommés correspondant aux mélanges, et des signaux sinusoïdaux à fréquences exponentielles qui permettent de calculer les RI associées à chaque microphone. Les mélanges comportent un bruit de conversation particulièrement élevé et pour lequel les RN n'ont pas été entraînés. Ils ne constituent donc pas un ensemble d'évaluation adapté à nos expérimentations. C'est pourquoi nous avons décidé de calculer les RI de chaque microphone à partir des signaux sinusoïdaux et de créer les mélanges avec les mêmes signaux sources que ceux utilisés dans le corpus DISCO (cf. section 3.4.1).

Pour cela, 1000 configurations d'évaluation ont été créées. Pour chaque configuration, quatre antennes de microphones et deux sources ont été choisies aléatoirement. Les antennes sont sé-

lectionnées parmi les 12 antennes circulaires (notées T1−T12 sur la figure 4.11). Sur leurs 8 microphones, 4 ont été choisis, en en prenant 1 sur 2 afin de créer une antenne de microphones en forme de carré. Les emplacements des deux sources ont été choisis aléatoirement parmi les 10 disponibles, sans considération de la distance entre les sources et les nœuds. Etant donné l'arrangement de la pièce, cela conduit à une plus grande variabilité de RSB car certaines sources sont très proches de certaines antennes de microphones, et d'autres plus éloignées.

4.4.2. Résultats et analyse

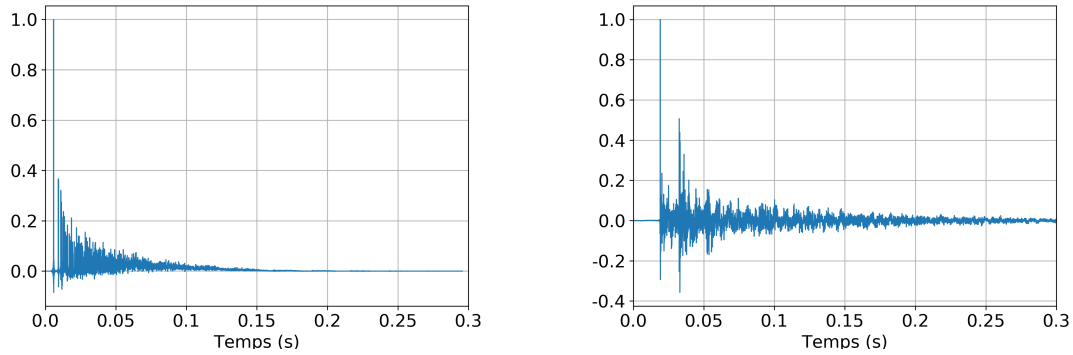
Les résultats obtenus avec Tango sur les données réelles sont indiqués dans le tableau 4.1. Nous y présentons également les performances obtenues en conditions simulées, ainsi que les résultats obtenus en conditions oracles. Lorsque des RN ont servi à prédire les masques, il s'agit des mêmes RN qu'utilisés jusqu'alors : ils n'ont pas été ré-entraînés dans des conditions réelles.

TABLEAU 4.1. – Résultats du rehaussement de la parole de Tango avec des RN entraînés sur données simulées, mais évalués sur données obtenues avec des RI soit simulées (« sim » dans le tableau), soit réelles (« réel »). Les masques nécessaires aux différentes étapes de Tango sont soit oracles (« MRI »), soit prédits par les RN (« CRNN »).

	$\Delta\text{SIR}_{\text{img}}$ (dB)	SAR_{img} (dB)	SAR_{src} (dB)	STOI_{img}
MRI sim	26,8 ± 0,4	10,9 ± 0,2	9,6 ± 0,2	0,89 ± 0,004
MRI réel	22,7 ± 0,3	7,9 ± 0,2	3,7 ± 0,3	0,79 ± 0,005
CRNN sim	22,9 ± 0,5	6,9 ± 0,1	8,5 ± 0,2	0,78 ± 0,004
CRNN réel	19,0 ± 0,4	3,2 ± 0,1	2,9 ± 0,2	0,61 ± 0,008

Evalué sur données réelles, Tango a des performances bien plus faibles que sur données simulées. Cet écart s'explique en partie par le fait que les données réelles sont plus difficiles que les données simulées, comme le montre la différence de performances avec les masques oracles. Cette difficulté peut être liée à la plus grande variété de RSB ou au TR plus élevé. Néanmoins, la différence entre données réelles et données simulées est plus importante lorsque des RN sont utilisés qu'avec des masques oracles. En particulier, les SAR sont très faibles, tout comme le STOI. Les RN ont donc des difficultés à généraliser sur les données réelles, qui diffèrent par de nombreux points des données simulées vues à l'entraînement. En effet, les configurations spatiales, les temps de réverbération, les caractéristiques des microphones et les conditions acoustiques diffèrent entre le jeu d'entraînement et ce jeu de données réelles. Il est donc compréhensible que les performances décroissent sur ces données.

Une première solution pour améliorer les performances sur données réelles serait de simuler un scénario plus proche de celui de Corey et al. (2019) de par la position des nœuds et des sources et de par le temps de réverbération. Néanmoins, cela n'aurait probablement qu'un effet limité car la différence entre les conditions réelles et simulées ne tient pas uniquement à la configuration spatiale. En effet, la figure 4.12 représente deux RI. L'une est simulée à l'aide de la boîte à outils `Pyroomacoustics` (Scheibler et al., 2018), comme dans DISCO (figure 4.12(a)) ; l'autre est calculée à l'aide des enregistrements réels de Corey et al. (figure 4.12(b)). On peut y voir que les RI simulées sont idéalisées dans le sens où les premières réverbérations sont clairement démarquées les unes des autres, alors qu'une sorte de bruit est présent tout au long de la RI réelle, qui dure d'ailleurs plus longtemps. Par ailleurs, la décroissance exponentielle de la RI, imposée par la simulation, n'est pas toujours observée dans les RI réelles. D'autres RI, non représentées ici, ont une enveloppe beaucoup moins régulière que celle des RI simulées. Ces différences, qui pointent les limites du modèle de la simulation de RI, peuvent expliquer les plus faibles résultats des RN sur données réelles.



(a) Exemple de RI simulée

(b) Exemple de RI réelle

FIGURE 4.12. – Exemples de RI simulée et réelle.

Ainsi, il faudrait également envisager d'incorporer des données réelles dans le jeu d'entraînement, voire de réapprendre sur des données réelles ou d'utiliser une architecture de RN plus puissante.

4.5. Exploitation de l'information spatiale

Nous avons brièvement vu en section 4.3.2 qu'utiliser les RNMuN à la seconde étape de filtrage augmentait les performances de Tango. Dans cette section, nous montrons plus en détail que les signaux compressés constituent effectivement une information utile pour mieux prédire les masques temps-fréquence (TF). Nous montrons également qu'ils profitent généralement mieux aux nœuds à faibles RSB et qu'il peut être préférable d'envoyer non pas l'estimation de la parole, mais celle du bruit, sous la forme des signaux compressés.

4.5.1. Performances au niveau des meilleurs nœuds en sortie

La figure 4.13 représente les résultats obtenus avec Tango lorsque des RNMoN sont utilisés aux deux étapes de filtrage (légendé « RN_1 »), et lorsque des RNMuN les remplacent à la seconde étape (légendé « RN_4 »). Ces deux configurations sont également comparées au cas où un détecteur d'activité vocale (DAV) oracle est utilisé pour calculer les matrices d'autocorrélation spatiale (MAS) du MWF. Les résultats de la figure sont ceux mesurés au niveau du meilleur nœud de sortie, c'est-à-dire au niveau du nœud avec le meilleur $\mathrm{SIR}_{\mathrm{img}}$ en sortie. Comme on avait déjà pu le constater, utiliser les signaux compressés pour prédire les masques TF permet d'augmenter significativement les performances. Le $\Delta\mathrm{SIR}_{\mathrm{img}}$ par exemple augmente de près de 3 dB, et le $\mathrm{SAR}_{\mathrm{src}}$ de 1.5 dB. Par ailleurs, les signaux compressés conduisent à des performances aussi bonnes que ce qu'un DAV oracle (donc sans erreur) permet d'atteindre en termes de $\Delta\mathrm{SIR}_{\mathrm{img}}$ et SAR. Seul le STOI semble indiquer de meilleures performances avec le DAV oracle, mais dans une mesure limitée. Dans l'ensemble, ces résultats montrent bien qu'il est utile d'exploiter l'information des signaux compressés pour prédire les masques TF et pour améliorer les performances de rehaussement de la parole.

4.5.2. Performances au niveau des meilleurs nœuds en entrée et en sortie

La coopération entre les différents nœuds d'une antenne acoustique est mise en évidence dans cette section. Pour cela, les résultats de Tango sont relevés au niveau de deux nœuds dans chaque

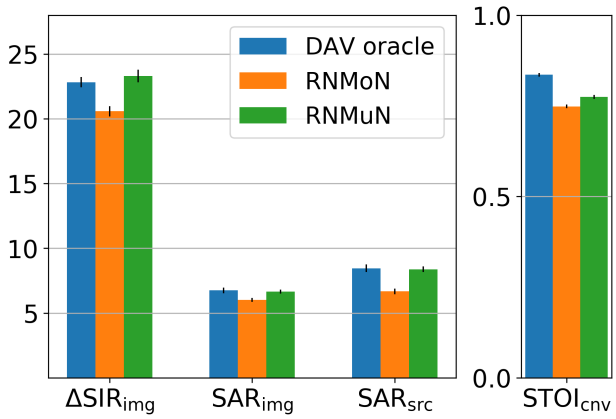


FIGURE 4.13. – Résultats de rehaussement de la parole de Tango avec des réseaux de neurones mononœuds (RNMoN) et des réseaux de neurones multinœuds (RNMuN), comparés aux résultats avec un DAV oracle.

configuration d'évaluation. Le premier nœud est le nœud avec le meilleur SIR$_{img}$ après filtrage ; il s'agit du même nœud que celui considéré jusqu'à présent et il est indiqué par les initiales « MS » dans le tableau 4.2 (meilleur en sortie). Le second nœud est le nœud avec le meilleur SIR$_{img}$ avant filtrage. Il est indiqué par les initiales « ME » (meilleur en entrée). Il n'est pas impossible qu'un nœud soit à la fois MS et ME. Par ailleurs, les métriques mesurées sur les signaux après la première étape de filtrage sont également relevées. Ces signaux sont issus d'un MWF appliqué sur tous les nœuds indépendamment les uns des autres, ce qui permet de mettre en évidence l'intérêt d'utiliser les signaux compressés. Les résultats sont présentés dans le tableau 4.2.

	SIR$_{img,\ out}$	ΔSIR$_{img}$	SAR$_{img}$	SAR$_{src}$	STOI$_{img}$
Etape 1 – ME	17,6 ± 0,4	15,0 ± 0,4	7,5 ± 0,2	7,5 ± 0,2	0,80 ± 0,004
Etape 1 – MS	19,0 ± 0,4	17,2 ± 0,3	7,5 ± 0,2	7,9 ± 0,2	0,80 ± 0,005
Etape 2 – ME	22,5 ± 0,5	19,9 ± 0,5	6,7 ± 0,2	8,3 ± 0,2	0,77 ± 0,004
Etape 2 – MS	24,0 ± 0,5	22,9 ± 0,5	6,9 ± 0,1	8,5 ± 0,2	0,78 ± 0,004

TABLEAU 4.2. – Résultats de rehaussement de la parole de Tango aux deux étapes de filtrage, aux meilleurs nœuds en entrée et en sortie.

La première remarque que l'on peut faire est que les meilleurs nœuds avant filtrage ne restent pas systématiquement les meilleurs nœuds après filtrage, puisque les résultats ne sont pas identiques pour les deux types de nœuds. Ce ne sont donc pas toujours les nœuds avec les meilleurs SIR$_{img,\ in}$ qui permettent le meilleur rehaussement de la parole (en tous cas en termes de SIR).

De plus, à part en termes de STOI, la différence de performance entre les meilleurs nœuds en entrée et en sortie est plus importante à la fin de la seconde étape qu'à la fin de la première étape. Cela signifie que certains nœuds profitent mieux des signaux compressés que les meilleurs nœuds en entrée. En fait les meilleurs nœuds en entrée sont ceux qui envoient des signaux compressés de bonne qualité. Ces signaux, envoyés aux autres nœuds, permettent aux nœuds qui les reçoivent de largement augmenter leur performance par rapport à la première étape.

On peut calculer le SIR$_{img,\ in}$ en soustrayant le ΔSIR$_{img}$ du SIR$_{img,\ out}$. Il est intéressant de noter que le SIR$_{img,\ in}$ des meilleurs nœuds de sortie de la seconde étape (égal à 1,1 dB) est inférieur au SIR$_{img,\ in}$ des meilleurs nœuds de sortie de la première étape (égal à 1,8 dB). Cela signifie que des nœuds avec de plus faibles SIR$_{img,\ in}$ sont devenus les meilleurs nœuds de sortie

grâce à l'échange des signaux compressés.

Ces observations montrent que les nœuds de l'antenne acoustique coopèrent grâce à Tango, où les « bons » nœuds permettent aux nœuds avec de plus faibles RSB de profiter de leur meilleure estimation du signal cible.

4.5.3. Performances au niveau des meilleurs et pires nœuds en entrée

Dans cette section, toujours dans l'optique de mettre en évidence l'intérêt d'utiliser les signaux compressés, nous reportons dans le tableau 4.3 les métriques mesurées aux deux étapes de filtrage de Tango, mais cette fois-ci au niveau des pires et meilleurs nœuds en entrée, indiqués respectivement par les initiales PE (pire en entrée) et ME (meilleur en entrée). Les mêmes nœuds sont donc considérés aux deux étapes, puisque le $\mathrm{SIR}_{\mathrm{img,\ in}}$ est indépendant du filtrage. Le $\mathrm{SIR}_{\mathrm{img,\ out}}$ est également reporté afin de mieux quantifier la performance absolue de chaque nœud, indépendamment du $\mathrm{SIR}_{\mathrm{img,\ in}}$.

	$\mathrm{SIR}_{\mathrm{img,\ out}}$	$\Delta\mathrm{SIR}_{\mathrm{img}}$	$\mathrm{SAR}_{\mathrm{img}}$	$\mathrm{SAR}_{\mathrm{src}}$	$\mathrm{STOI}_{\mathrm{img}}$
Etape 1 – ME	17,6 ± 0,4	15,0 ± 0,4	7,5 ± 0,2	7,5 ± 0,2	0,80 ± 0,004
Etape 1 – PE	12,7 ± 0,5	15,1 ± 0,4	5,4 ± 0,2	4,7 ± 0,2	0,72 ± 0,005
Etape 2 – ME	22,5 ± 0,5	19,9 ± 0,5	6,7 ± 0,2	8,3 ± 0,2	0,77 ± 0,004
Etape 2 – PE	19,9 ± 0,7	22,3 ± 0,6	4,6 ± 0,2	6,8 ± 0,2	0,75 ± 0,004

TABLEAU 4.3. – Résultats de rehaussement de la parole de Tango aux deux étapes de filtrage, aux pires et meilleurs nœuds en entrée.

Tout d'abord, en soustrayant le $\Delta\mathrm{SIR}_{\mathrm{img}}$ du $\mathrm{SIR}_{\mathrm{img,\ out}}$ aux deux types de nœuds, on retrouve bien que le $\mathrm{SIR}_{\mathrm{img,\ in}}$ au niveau des meilleurs nœuds est supérieur de 5 dB par rapport au $\mathrm{SIR}_{\mathrm{img,\ in}}$ des pires nœuds, ce qui pouvait déjà être observé sur la figure 4.8.

Par ailleurs, le tableau 4.3 met en évidence l'avantage de la deuxième étape : à tous les nœuds, toutes les métriques, exception faite du $\mathrm{SAR}_{\mathrm{img}}$ qui diminue probablement à cause de la déréverbération, indiquent une amélioration significative des performances entre l'étape 1 et l'étape 2. Cela confirme donc que les signaux compressés sont utiles, bien que cette expérience ne permette pas de dire à quel niveau du filtrage les signaux compressés sont le plus utiles (si c'est lié à une meilleure prédiction des masques TF ou à un meilleur formateur de voies).

Enfin, ces résultats montrent que la deuxième étape de filtrage profite plus aux pires nœuds qu'aux meilleurs nœuds. Ceci est le plus évident en considérant le $\mathrm{SIR}_{\mathrm{img,\ out}}$ qui augmente de plus de 7 dB au niveau des pires nœuds, alors qu'il augmente de 4,9 dB au niveau des meilleurs nœuds. Par conséquent, alors que les $\mathrm{SIR}_{\mathrm{img,\ out}}$ de l'étape 1 diffèrent de 4,9 dB entre les deux types de nœuds, ils ne diffèrent plus que de 2,6 dB à la suite de l'étape 2. L'échange de signaux compressés permet donc de réduire l'écart de performances entre les nœuds d'une antenne acoustique ad-hoc. Cela montre que les nœuds de l'antenne coopèrent, et que les nœuds avec le pire SIR en entrée profitent des signaux compressés des autres nœuds.

Néanmoins, cela semble indiquer que les meilleurs nœuds ne tirent pas un profit maximal des signaux compressés. Peut-être profiteraient-ils plutôt d'une estimation du bruit. Des travaux ont d'ailleurs fait état de l'intérêt d'utiliser le signal de bruit pour améliorer les performances de RN (Seltzer et al., 2013; Araki et al., 2015; Perotin et al., 2018). Dans la section suivante, nous proposons donc d'étudier si envoyer l'estimation du bruit sous forme de signal compressé, plutôt que l'estimation de la parole, permettrait de mieux exploiter l'échange de signaux entre les nœuds de l'antenne.

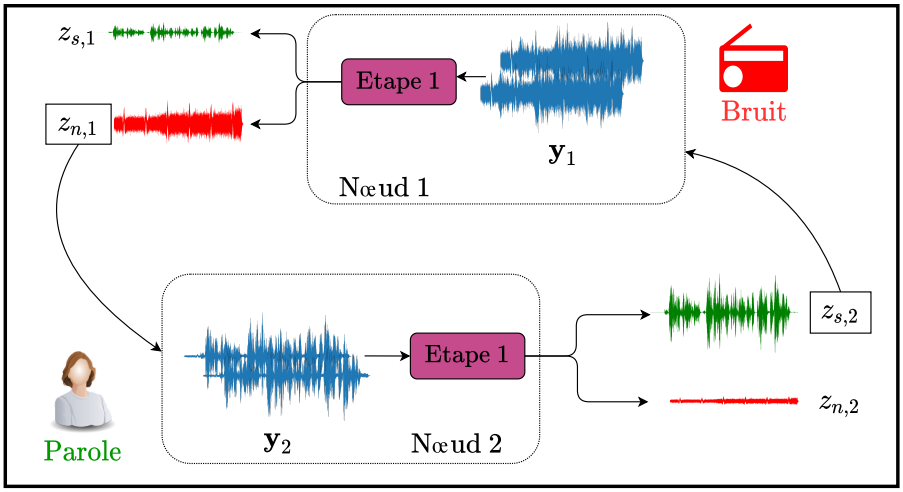


FIGURE 4.14. – Illustration de l'intérêt pour les nœuds éloignés de la source de bruit de recevoir l'estimation du bruit sous forme de signal compressé.

4.5.4. Intérêt d'envoyer l'estimation du bruit d'un nœud à l'autre

La figure 4.14 représente de manière schématique l'idée de cette section. Certains nœuds, comme le nœud 2 sur la figure, peuvent être proches de la source de parole et ont déjà un bon aperçu du signal de parole. Puisque le masque TF que doivent estimer les RN dépend aussi du signal de bruit (cf. équation (3.13)), il se peut que les RN placés sur des nœuds éloignés de la source de bruit gagneraient à recevoir une estimation du bruit sous forme de signaux compressés. Dans le cas de la figure 4.14, il est possible que le nœud 2 gagnerait à recevoir l'estimation du bruit de la part du nœud 1. Ceci semble d'autant plus vrai si le nœud 1 est placé proche de la source de bruit, qu'il peut donc bien estimer. Dans la suite, nous appellerons *estimation compressée de la parole* au nœud k le résultat de la première étape de filtrage de Tango au nœud k. Jusqu'alors elle était notée z_k, mais nous la renommerons $z_{s,k}$ pour éviter les confusions. De même, l'*estimation compressée du bruit* au nœud k sera notée $z_{n,k}$ et elle est l'estimation du bruit contenu dans $y_{k,1}$ à la fin de la première étape :

$$z_{n,k} = y_{k,1} - z_{s,k}\,.$$

Nous proposons ainsi d'envoyer l'estimation compressée du bruit plutôt que celle de la parole pour la prédiction du masque. Deux nouveaux RNMuN sont entraînés sur la configuration *aléatoire* : le premier a en entrée le mélange d'un nœud k et trois estimations compressées du bruit venant des trois autres nœuds. Précisons que si les signaux compressés de bruit sont utilisés par le RN pour prédire le masque, les estimations compressées de parole doivent également être échangées pour le calcul du filtre à la seconde étape de filtrage. Il conviendrait d'étudier le cas où ces signaux ne sont pas envoyés afin de ne pas surcharger la bande passante, mais cela dépasse le cadre de cette étude.

Le second RN voit en entrée le mélange du nœud k, les estimations compressées de la parole et les estimations compressées du bruit. Il dispose donc de 7 signaux pour prédire le masque.

Ces deux RNMuN, en plus du RNMuN jusqu'alors utilisé (prédisant les masques à partir du mélange local et des estimations compressées de la parole), sont comparés à la seconde étape de filtrage de Tango, avec le même RNMoN à la première étape de filtrage. Les résultats dans chacune de ces configurations sont représentés en figure 4.15. Comme dans la section précédente,

les résultats sont présentés en fonction du nœud sur lequel ils sont mesurés. La figure 4.15(a) présente les métriques mesurées sur les meilleurs nœuds de sortie (c'est-à-dire sur le nœud avec le meilleur SIR_{img} après filtrage). La figure 4.15(b) présente les métriques mesurées sur les meilleurs nœuds d'entrée. La figure 4.15(c) présente les métriques mesurées sur les pires nœuds d'entrée. La légende de la figure 4.15(b) indique les signaux utilisés à l'entraînement du RNMuN de la seconde étape. « z_s » indique que seules les estimations compressées de la parole (en plus du mélange local) ont été utilisées pour prédire les masques TF. « z_n » indique que seules les estimations compressées du bruit on été utilisées. « $z_s + z_n$ » indique que les deux types d'estimations compressées ont été utilisés.

Au niveau des meilleurs nœuds de sortie, envoyer l'une ou l'autre des estimations compressées ne fait pas de différence significative. Les trois RNMuN conduisent à des performances similaires. Au niveau des meilleurs nœuds d'entrée, une tendance semble se dégager, même si les différences ne sont pas significatives : Tango semble plus profiter des estimations compressées du bruit que des estimations compressées de la parole. Cette tendance s'inverse au niveau des pires nœuds, où il est significativement moins bon d'utiliser les signaux compressés du bruit pour prédire les masques que ceux de parole. Ceci s'explique par le fait que les RNMuN sur les pires nœuds ont déjà une bonne estimation du bruit de par le mélange local. Ils profitent donc plus des estimations compressées de la parole, reçues de nœuds où la parole est justement plus forte. Notons que les trois RNMuN utilisés ici ont été entraînés à partir de signaux mesurés sur tous les types de nœuds. Il est possible que spécialiser un RNMuN sur les pires ou meilleurs nœuds exacerberait les tendances, puisqu'ils apprendraient que les signaux compressés apportent toujours une estimation du signal le moins bien vu par le nœud sur lequel ils opèrent.

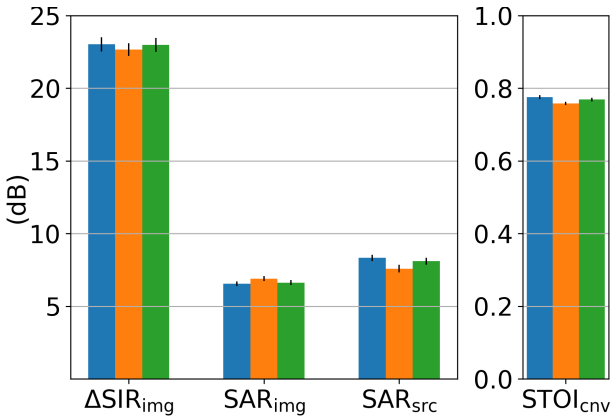
Dans les trois cas de figure, utiliser les deux types de signaux compressés ne fait jamais mieux qu'utiliser l'un ou l'autre des types de signaux compressés. Il semble même qu'envoyer les deux types de signaux compressés ne soit jamais mieux qu'envoyer l'estimation compressée de la parole uniquement, quel que soit le nœud sur lequel les métriques sont calculées. C'est une observation surprenante car plus d'information est disponible lorsque les deux types de signaux compressés sont envoyés. Une explication pour ce phénomène pourrait être que les RN, relativement simples, ne sont pas capables d'exploiter toute l'information disponible dans les canaux en entrée du RN.

En conclusion de ces expériences, le choix du type de signal compressé à envoyer d'un nœud à l'autre dépend de l'application de l'algorithme de rehaussement de la parole. Si le but visé est d'avoir le meilleur signal rehaussé, n'envoyer que l'estimation compressée de la parole suffit, comme le montre la figure 4.15(a). Si en revanche, chaque nœud doit avoir le meilleur résultats possible, il semble préférable au niveau des meilleurs nœuds d'entrée d'envoyer l'estimation compressée du bruit. Alors qu'envoyer les deux types de signaux paraissait une bonne option (aux dépends de la bande passante), les expériences n'ont pas pu le démontrer.

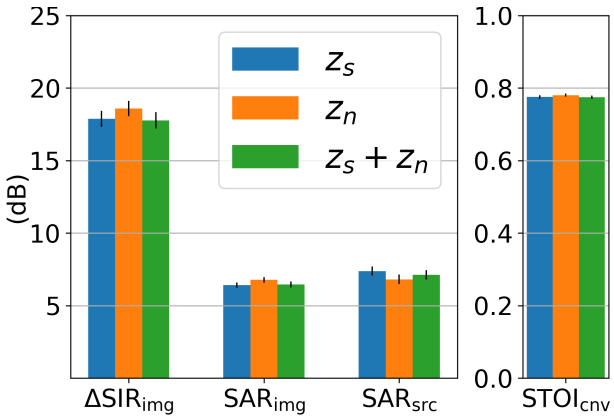
4.6. Conclusion

Ce chapitre a présenté une étude détaillée des performances de Tango, sur différentes configurations spatiales réalistes. Nous avons montré que Tango offre de bonnes performances sur les trois configurations spatiales simulées, et que les RN utilisés généralisent bien sur des configurations spatiales qu'ils n'ont pas vues à l'entraînement. De même, Tango conduit à des performances stables sur de grandes plages de TR et de SIR, ainsi que lorsqu'un bruit diffus est présent. Néanmoins, nous avons constaté que les RN généralisent mal sur des bruits et RI non vus à l'entraînement. En particulier, les RN entraînés sur des données simulées généralisent mal sur des données réelles. L'entraînement des RN devra donc être adapté aux données d'utilisation lors la mise en pratique de Tango. Par ailleurs, en évaluant les performances aux différents nœuds d'une

(a) Aux meilleurs nœuds de sortie



(b) Aux meilleurs nœuds d'entrée



(c) Aux pires nœuds d'entrée

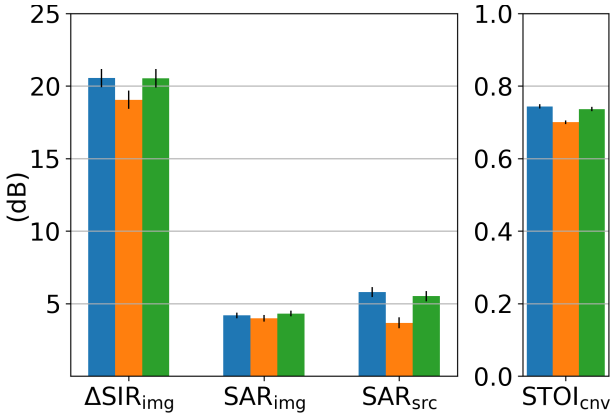


FIGURE 4.15. – Résultats de rehaussement de la parole de Tango lorsque les RNMuN sont entraînés avec différentes estimations compressées. « z_s » indique que seules les estimations compressées de la parole (en plus du mélange local) ont été utilisées pour prédire les masques TF. « z_n » indique que seules les estimations compressées du bruit on été utilisées. « $z_s + z_n$ » indique que les deux types d'estimations compressées ont été utilisés.

antenne acoustique, nous avons mis en évidence la coopération entre les nœuds permise par les échanges de signaux dans Tango. De plus, nous avons montré qu'il semble plus utile d'envoyer l'estimation compressée du bruit plutôt que celle de parole aux nœuds les plus proches de la source de parole.

Le tableau 4.4 rassemble les conclusions de ce chapitre.

Section	Points-clé
4.2	• Les performances de Tango sont bonnes sur une grande variété de configurations spatiales réalistes. • Tango généralise bien sur les différentes configurations spatiales, quelle que soit la configuration spatiale vue par ses RN pendant l'entraînement.
4.3.1	Tango est résilient à la réverbération.
4.3.2	Tango est résilient au SIR en entrée.
4.3.3	Tango est résilient à la présence d'un bruit diffus.
4.3.4	Tango est peu résilient à des types de bruits non vus à l'entraînement.
4.4	Les RN entraînés sur des données simulées généralisent mal sur des données réelles.
4.5.1	Les signaux compressés permettent de mieux prédire les masques TF et d'améliorer les performances du rehaussement de la parole.
4.5.2, 4.5.3	Les nœuds d'une antenne acoustique coopèrent.
4.5.4	Au niveau des nœuds les plus proches de la source de parole, il peut être utile d'envoyer l'estimation compressée du bruit plutôt que celle de parole.

TABLEAU 4.4. – Points-clé à retenir du chapitre 4.

Troisième partie

Extension de Tango à des cas d'applications pratiques

Nous avons présenté une solution distribuée de rehaussement de la parole et montré qu'elle était efficace dans un grand nombre de conditions acoustiques. Nous avons mis en évidence le fait que cette solution entraîne une coopération entre les nœuds des antennes acoustiques ad-hoc (AAAH) et qu'elle permet une exploitation de l'information spatiale enregistrée par tous les nœuds des AAAH. Toutefois, certaines spécificités des AAAH n'ont pas été considérées si bien qu'il n'est pas encore convaincant d'affirmer que la solution que nous proposons est réellement applicable dans les AAAH. Dans cette partie, nous prenons en compte le fait que le nombre de nœuds dans une AAAH peut varier, ce que nous présentons en chapitre 5. Nous considérons également les cas où les nœuds d'une AAAH ne sont pas synchronisés, ce qui est présenté en chapitre 6. Enfin, nous montrons dans le chapitre 7 que Tango est un algorithme qui peut être également utilisé dans des contextes de séparation de sources.

5. Extension à des antennes acoustiques ad-hoc avec un nombre variable de nœuds

Comme présenté dans l'introduction, la force des AAAH vient du fait qu'elles sont constituées d'appareils dont on dispose maintenant presque en permanence, présents en de nombreux points de l'espace, fournissant ainsi une représentation riche de la scène acoustique dans laquelle ils sont utilisés. Néanmoins, cette force des AAAH peut également limiter leur utilisation. Par exemple, la portée de réception ou d'émission des antennes (Wifi, bluetooth, téléphoniques) peut être réduite lorsque le niveau de batterie de l'appareil baisse. Dans d'autres situations, il est imaginable qu'un des appareils ait son champ d'émission ou de réception fortement occulté par un obstacle présent dans la pièce, coupant ainsi l'appareil du reste de l'AAAH. Dans ce chapitre, nous étudions la réponse de Tango à des situations où le nombre de nœuds dans une AAAH varie, et nous proposons une solution pour qu'il soit entièrement opérationnel dans des AAAH avec un nombre variable de nœuds.

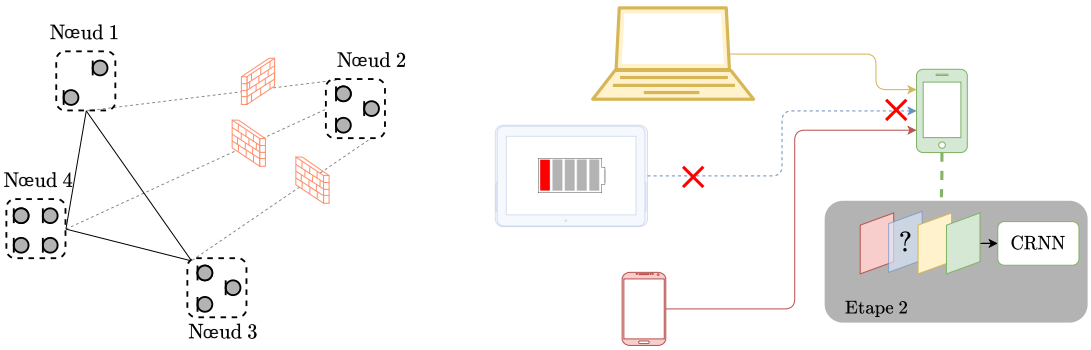
5.1. Présentation du contexte

5.1.1. Rupture de liens dans une antenne acoustique ad-hoc

Dans toute la suite de ce chapitre, nous utiliserons le terme *lien* pour définir la connexion d'un nœud avec tout le reste de l'AAAH. Lorsque le lien d'un nœud est rompu, ce nœud ne communique avec plus aucun autre nœud de l'antenne, mais il peut continuer sa tâche de rehaussement de la parole localement. Une telle situation peut arriver par exemple si l'état de batterie du nœud est trop faible pour recevoir et envoyer des signaux compressés, si son antenne est défectueuse ou si un obstacle empêche la propagation des signaux vers et depuis ce nœud. De telles situations sont représentées en figure 5.1. Les cas où les connexions de nœud à nœud ou de microphone à microphone (au niveau d'un nœud) peuvent être rompues, bien qu'intéressants, ne sont pas traités dans ce rapport.

Si un certain nombre d'algorithmes basés sur une approche classique s'adaptent à tous types de topologies d'AAAH (Szurley et al., 2016; Tavakoli et al., 2017; Koutrouvelis et al., 2018; Guo et al., 2021b), l'utilisation de réseaux de neurones (RN) est limitée dans les AAAH car la plupart des architectures de RN reposent sur un nombre constant de canaux. Dans les cas où des nœuds disparaissent de la salle, ou lorsque des liens entre les nœuds sont rompus, par exemple à cause d'une mauvaise connexion entre les appareils, il devient plus difficile d'utiliser ces RN.

Dans cette section, nous étudions la réponse de Tango lorsque des liens entre nœuds sont rompus, ce qui revient *de facto* à étudier les cas où le nombre de nœuds est variable. Puisque les réseaux de neurones multinœuds (RNMuN) dans Tango attendent autant de canaux qu'il y a de nœuds dans l'antenne, il n'est pas possible de les utiliser tels qu'ils l'ont été jusqu'à présent si le nombre de nœuds varie. Cette situation est représentée en figure 5.1(b).



(a) Déconnexion d'un nœud à cause d'un obstacle (b) Conséquences d'une rupture de lien à cause d'une batterie faible

FIGURE 5.1. – Illustrations de la problématique du chapitre 5.

5.1.2. Architectures de réseaux de neurones résilientes à un nombre variable de canaux

Une solution à apporter dans le cas où des liens entre nœuds viendraient à rompre serait de considérer une autre architecture de RN, en en prenant une qui soit résiliente à un nombre variable de canaux en entrée. L'une d'entre elles est proposée par Casebeer et al. qui utilisent des cellules récurrentes, dont l'entrée le long d'un axe (usuellement l'axe temporel) peut être de dimension arbitraire (Casebeer et al., 2018). En appliquant la récurrence le long de l'axe des canaux, le nombre de canaux peut donc être variable. Cette solution cependant suppose que l'ordre des canaux a une importance, ce qui est incompatible avec les applications non-contraintes des AAAH. Un autre moyen pour supprimer la dépendance au nombre de canaux en entrée est d'utiliser des couches dont les paramètres sont partagés par tous les canaux d'entrée, et de fusionner les sorties de ces couches partagées (Luo et al., 2020a; Wang et al., 2020; Zhang and Li, 2021). Cependant cette idée considère tous les canaux de la même manière, alors que leur répartition spatiale leur fait observer des signaux très différents. Une dernière solution a été récemment proposée par Zhang and Li (2021) et consiste à donner à la première couche du RN des paires de microphones où le microphone de référence constitue le premier signal de la paire. Si cette méthode est généralisable à un nombre arbitraire de canaux, les sorties de la première couche sont fusionnées par une opération de moyenne, qui lisse les spécificités intercanales.

5.1.3. Mécanismes d'attention

Afin que différentes parties des données d'entrée soient considérées de manière différente, le concept d'attention a été introduit dans le domaine des RN (Bahdanau et al., 2014; Cho et al., 2015). Dans le domaine plus précis du rehaussement de la parole, différents mécanismes d'attention ont été proposés. Deux grandes catégories en ressortent.

Les mécanismes de la première catégorie consistent à d'abord transformer les données d'entrée en un tenseur de dimension réduite qui est ensuite donné à un RN qui relève les parties les plus importantes du tenseur. Les mécanismes de cette catégorie suivent deux modèles. Le premier est celui du « compresser et stimuler » (SE : *squeeze and excitation*), d'abord introduit dans le contexte du traitement d'images (Hu et al., 2018) puis adapté au traitement des sons (Roy et al., 2018; Xia and Koishida, 2019; Lan et al., 2020a,b). Woo et al. ont quant à eux proposé le modèle d'attention par bloc convolutionnel (CBAM : *convolutional block attention module*) (Woo et al., 2018), composé d'un module d'attention sur les canaux et d'un module d'attention spatial. Le CBAM a lui aussi été réutilisé dans le domaine du rehaussement de la parole (Shi et al., 2020;

Zhao et al., 2021; Xue et al., 2021).

La deuxième grande catégorie de mécanismes d'attention regroupe les mécanismes dits d'« auto-attention » (SA : *self attention*) (Vaswani et al., 2017; Tolooshams et al., 2020; Nicolson and Paliwal, 2020) qui permettent d'accorder plus d'attention aux données d'entrée en fonction de leur contexte. Ce type de mécanismes est décrit plus en détail en section 5.2.

Précisons que ces deux catégories ne rassemblent pas tous les mécanismes d'attention utilisés dans le domaine du rehaussement de la parole, et que d'autres ont été présentés sans qu'ils puissent être rattachés à l'une ou l'autre des catégories (Giri et al., 2019; Hao et al., 2019; Ho et al., 2020).

5.2. Solution proposée dans des cas de ruptures de liens dans une antenne acoustique ad-hoc

Plutôt que de reposer sur une architecture invariable au nombre de canaux par construction, nous proposons de fixer un nombre maximal de canaux en entrée du RN en remplaçant tous les canaux manquants (ceux qui n'ont pas pu être envoyés par les nœuds déconnectés) par une valeur constante. De plus, afin de compenser la dégradation de performances du RN induite par le manque d'information, un mécanisme d'attention est introduit dans l'architecture des RNMuN de Tango. L'ambition d'utiliser un mécanisme d'attention est de sélectionner implicitement les canaux utiles en négligeant ceux qui n'ont pas été reçus. Pour cela, les canaux pertinents seraient multipliés par des poids élevés, tandis que les canaux non pertinents, peu informatifs, seraient multipliés par des poids faibles. Le tenseur d'entrée pondéré par les poids du mécanisme d'attention est ensuite fourni au RN convolutionnel récurrent (CRNN) afin d'estimer le masque temps-fréquence (TF) nécessaire au calcul des formateurs de voies de chaque étape de Tango.

Nous proposons d'utiliser et comparer deux mécanismes d'attention, de chacune des catégories décrites au paragraphe précédent : un mécanisme SE et un mécanisme SA. Dans cette section, les deux estimations compressées de la parole et du bruit sont envoyées par chaque nœud, comme en section 4.5.4. Bien que cela augmente la consommation en bande passante, nous avons montré en section 4.5.4 et 7.2 qu'envoyer à la fois les estimations compressées de la parole et celles du bruit permettait dans certaines conditions d'améliorer les performances des RNMuN. La réduction de la consommation en bande passante par l'envoi de moins de signaux est remise à des recherches ultérieures.

5.2.1. Utilisation d'un mécanisme « compresser et stimuler »

Dans un premier temps, le mécanisme d'attention considéré est un « compresser et stimuler » (SE) parce que c'est un mécanisme simple qui a déjà permis d'obtenir de bons résultats dans le traitement du signal monocanal (Lan et al., 2020a,b) et multicanal (Xia and Koishida, 2019). Le module de SE est constitué de deux blocs. Le premier bloc « compresse » l'information spectro-temporelle en un unique scalaire. Le tenseur d'entrée, à trois dimensions (canaux, temps, fréquences) est donc réduit à un vecteur uni-dimensionnel (canaux). Ce vecteur uni-dimensionnel est ensuite « stimulé » par le second bloc, constitué de deux couches toutes-connectées, dont la première divise la dimension des canaux par un facteur r appelé *facteur de réduction*. La deuxième couche redonne au vecteur sa dimension initiale. La sortie du mécanisme est ce que nous appellerons les *poids*, qui prennent des valeurs entre 0 et 1, et qui sont multipliés par le tenseur d'entrée.

Les précédents travaux faisant intervenir des mécanismes SE insèrent ces mécanismes après chaque bloc convolutionnel de leurs RN (Xia and Koishida, 2019; Lan et al., 2020a,b). Nous

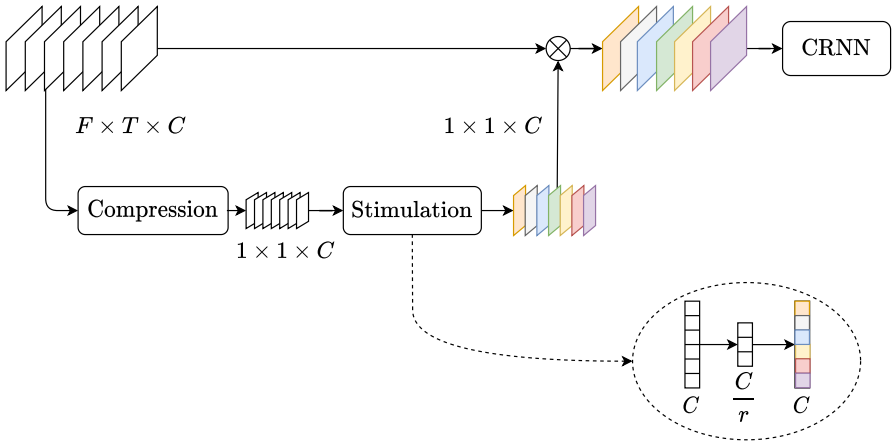


FIGURE 5.2. – Représentation graphique du mécanisme « compresser et stimuler » (SE) (Hu et al., 2018).

proposons d'insérer un unique mécanisme SE dès l'entrée du RN afin de pondérer les signaux reçus par une valeur proportionnelle à leur utilité. La figure 5.2 représente un mécanisme d'attention SE dans ce contexte, où C, T et F sont les dimensions respectives des canaux, du temps et de la fréquence. Les performances obtenues avec ce mécanisme sont présentées en section 5.3.2.

5.2.2. Utilisation d'un mécanisme d'auto-attention

Le mécanisme d'auto-attention (SA) (Vaswani et al., 2017) repose sur la transformation du tenseur d'entrée par trois fonction apprenables. Soit $\mathbf{y}$ le vecteur d'entrée, de dimension 1 et de taille d_y. Les trois transformations sont linéaires et notées $\mathbf{K} \in \mathbb{R}^{d_k \times d_y}$, $\mathbf{Q} \in \mathbb{R}^{d_k \times d_y}$ et $\mathbf{V} \in \mathbb{R}^{d_v \times d_y}$; elles transforment le tenseur d'entrées en :

$$\mathbf{k} = \mathbf{K}^T\mathbf{y} \tag{5.1}$$
$$\mathbf{q} = \mathbf{Q}^T\mathbf{y} \tag{5.2}$$
$$\mathbf{v} = \mathbf{V}^T\mathbf{y} \tag{5.3}$$

Les poids du mécanismes sont alors calculés de la manière suivante :

$$\mathbf{P} = softmax\left(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d_k}}\right) \tag{5.4}$$
$$= softmax\left(\tilde{\mathbf{P}}\right) \tag{5.5}$$

avec $\tilde{\mathbf{P}} = \frac{\mathbf{q}\mathbf{k}^T}{\sqrt{d_k}}$ et où $softmax$ est une fonction de $\mathbb{R}^N$ telle que, pour x_i la $i^{\text{ème}}$ valeur du vecteur $\mathbf{x} \in \mathbb{R}^N$:

$$softmax(\mathbf{x})_i = \frac{e^{x_i}}{\sum_{k=1}^{N} e^{x_k}}\,. \tag{5.6}$$

Dans le cas où, comme dans l'équation (5.4), $\mathbf{x}$ est une matrice, l'opérateur $softmax$ est appliqué sur chacune des lignes de $\mathbf{x}$. La sortie du mécanisme SA, qui est l'entrée du CRNN multinœud de Tango, est donnée par :

$$\mathbf{o} = \mathbf{P}\mathbf{v} \tag{5.7}$$

De manière plus qualitative, ce mécanisme d'attention vient activer les composantes de $\mathbf{v}$ d'après les valeurs de la matrice $\mathbf{P}$, qui dépend de la corrélation entre les deux projections $\mathbf{q}$ et $\mathbf{k}$. Cette matrice contient en quelque sorte l'information contextuelle de $\mathbf{y}$.

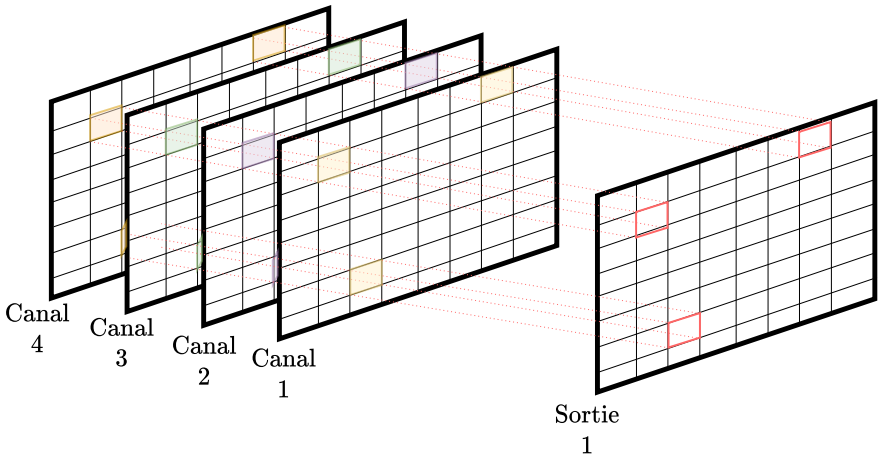


FIGURE 5.3. – Illustration graphique d'une convolution uni-dimensionnelle appliquée à un tenseur tri-dimensionnel. Les mêmes coefficients sont appliqués aux différents points TF d'un même canal d'entrée.

Le mécanisme d'auto-attention peut être étendu à un mécanisme d'attention à têtes multiples, composé de plusieurs « têtes » en parallèle, chacune des têtes étant un mécanisme d'auto-attention (Vaswani et al., 2017).

Dans cette thèse, nous proposons de modifier légèrement le mécanisme d'auto-attention pour qu'il puisse traiter des tenseurs tri-dimensionnels, comme ceux fournis à l'entrée de nos CRNN. Deux modifications sont apportées. La première n'est pas nouvelle, car elle a déjà été introduite par Zhang et al. (2019) pour le traitement d'images et par Tolooshams et al. (2020) pour le rehaussement de la parole. Elle consiste à remplacer les couches toutes connectées $\mathbf{K}$, $\mathbf{Q}$ et $\mathbf{V}$ par des couches de convolution uni-dimensionnelles, que l'on peut appliquer à des tenseurs tri-dimensionnels. Si par exemple la convolution est appliquée le long de l'axe temporel, la même opération est appliquée sur les $C \times F$ trames du tenseur. On notera d le nombre de canaux en sortie des convolutions $\mathbf{K}$ et $\mathbf{Q}$. La figure 5.3 illustre une convolution uni-dimensionnelle appliquée à un tenseur tri-dimensionnel.

La deuxième modification consiste à appliquer la transformation $\mathbf{V}$ (uni-dimensionnelle) sur un axe différent de celui de $\mathbf{K}$ et $\mathbf{Q}$. Cela permet de fournir une autre information pour le calcul des poids que ce qu'apportent les tenseurs $\mathbf{k}$ et $\mathbf{q}$. Afin de s'assurer que le tenseur de poids ait bien les mêmes dimensions que le tenseur d'entrée, le nombre de canaux en sortie de $\mathbf{V}$ est gardé égal au nombre de canaux en entrée. Le tenseur de sortie est calculé pour chaque bande de fréquences f de la manière suivante :

$$\mathbf{o}_f = \mathbf{v}_f \mathbf{P}_f^H \,, \tag{5.8}$$

où $\mathbf{o}_f$ est de dimensions $T \times C$, $\mathbf{P}_f$ est de dimensions $C \times C$ et $\mathbf{v}_f$ de dimensions $T \times C$. Ce sont les sous-matrices de $\mathbf{o}$, $\mathbf{P}$ et $\mathbf{v}$ respectivement, dans la bande de fréquences f.

L'intuition de ce nouveau mécanisme est d'extraire l'information temporelle du tenseur par les transformations $\mathbf{K}$ et $\mathbf{Q}$ et d'obtenir une représentation contextuelle dans le tenseur $\tilde{\mathbf{P}}$. La transformation $\mathbf{V}$ apporte quant à elle l'information spectrale, activée par la matrice des poids $\mathbf{P}$. Le mécanisme est représenté en figure 5.4 où est indiquée la dimension des différents tenseurs au cours du traitement.

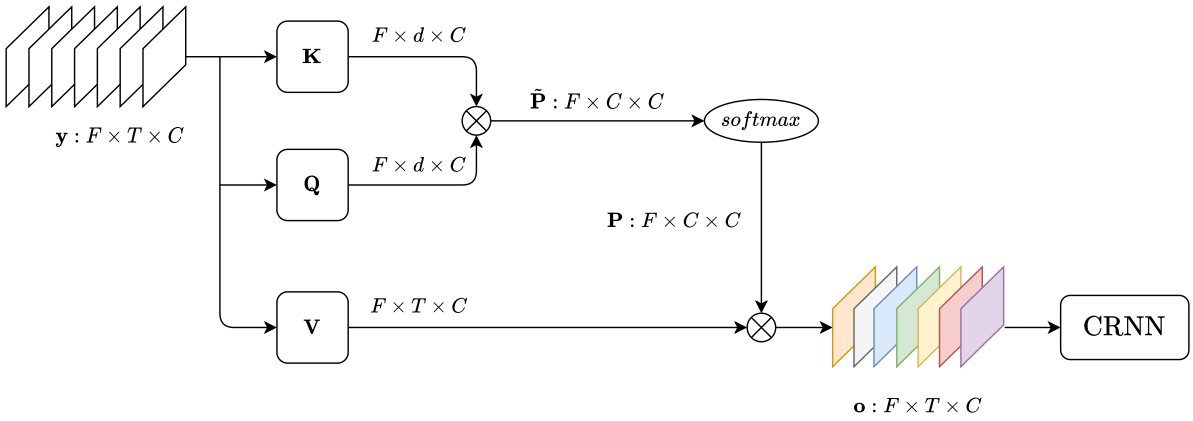


FIGURE 5.4. – Représentation graphique du mécanisme d'auto-attention tel qu'utilisé dans nos expériences.

5.3. Evaluation de la solution proposée

Les deux mécanismes d'attention présentés dans les paragraphes précédents sont utilisés en entrée du CRNN multinœud de la seconde étape de filtrage de Tango.

Notre méthode est évaluée sur le jeu d'évaluation de la configuration *aléatoire* présentée en section 3.4.2. Puisque quatre nœuds sont présents dans chaque situation, de 0 à 3 liens peuvent être rompus. Lorsque 3 liens sont rompus, plus aucun nœud ne reçoit de signal compressé et les RNMuN n'ont plus que le mélange local de réellement informatif. Les canaux manquants sont remplacés par une matrice constante égale à $-1 \cdot 10^{-7}$. Choisir une valeur négative, mais faible, permet d'indiquer une irrégularité, puisque les canaux sont normalement représentés par l'amplitude (positive) de la transformée de Fourier à court terme (TFCT) des signaux, sans pour autant perturber la dynamique des données.

5.3.1. Etude préliminaire

Dans cette partie, nous quantifions l'impact de canaux manquants sur la formation de voies uniquement. L'impact sur la prédiction des masques est étudiée dans les sections suivantes. Dans cette section, les mêmes masques prédits par les RN de la section 4.5.4 sont utilisés. Cependant, les formateurs de voies sont calculés non pas avec toutes les estimations compressées de la parole [1], mais uniquement avec celles qui ont réellement été reçues. Par exemple, si un nœud est déconnecté de l'antenne, les formateurs de voies à la seconde étape de filtrage des autres nœuds seront calculés à partir de leurs quatre mélanges locaux ($\mathbf{y}_k$ dans l'équation (3.1)) et des deux (et non trois) signaux compressés reçus.

Dans ces conditions, les performances de Tango sont représentées dans le tableau 5.1 lorsque 0, 1, 2 ou 3 liens sont rompus dans l'antenne du jeu d'évaluation. Afin de rendre compte des performances globales de Tango, les résultats correspondent aux moyennes sur tous les nœuds de chaque configuration d'évaluation.

La première observation que l'on peut faire est qu'un plus faible nombre de signaux compressés diminue le $\Delta\text{SIR}_{\text{img}}$. La diminution est assez limitée, mais presque constante et significative à chaque fois qu'un signal de moins est reçu. En revanche, cela n'est pas vérifié avec les autres métriques, puisque le SAR_{src} et le STOI_{img} diminuent peu quel que soit le nombre de signaux compressés, et que le SAR_{img} augmente même.

1. Les estimations compressées du bruit ne servent qu'à la prédiction du masque TF par le RNMuN.

# de liens rompus	$\Delta\text{SIR}_{\text{img}}$ (dB)	SAR_{img} (dB)	SAR_{src} (dB)	STOI_{img}
0	22,6 ± 0,2	5,6 ± 0,1	7,1 ± 0,1	0,65 ± 0,002
1	20,1 ± 0,2	6,0 ± 0,1	7,0 ± 0,1	0,64 ± 0,002
2	19,3 ± 0,2	6,4 ± 0,2	6,8 ± 0,1	0,64 ± 0,003
3	18,1 ± 0,2	6,9 ± 0,1	6,6 ± 0,1	0,63 ± 0,003

TABLEAU 5.1. – Résultats de rehaussement de la parole lorsque les signaux compressés manquent pour le calcul des formateurs de voies à la seconde étape de Tango

Cela est à rapprocher des discussions et résultats des sections 2.3.1.3 et 3.5.1. Comme argumenté dans la section 2.3.1.3, la décomposition en valeurs propres généralisée du GEVD-SDW-MWF définit implicitement un autre signal de référence dans la fonction de coût de l'équation (2.29). Le signal de référence défini est une combinaison linéaire des signaux d'entrée, dont les signaux à rapport signal-à-bruit (RSB) plus élevé sont plus mis à contribution. Lorsque les signaux compressés sont présents, comme ils ont un RSB supérieur, la référence implicite du GEVD-SDW-MWF est plus proche des signaux compressés que des signaux locaux. Or les résultats de la section 3.5.1 ont montré que le SAR_{img} pouvait croître sans que ce soit le cas du SAR_{src}, non pas parce que la déréverbération est plus faible ou parce que la qualité du signal est meilleure, mais parce que le signal ressemble plus à la référence image (convoluée) qu'à la référence source (non convoluée). Comme les signaux compressés proviennent de nœuds distants, ils ne sont pas trop proches de la référence image utilisée pour calculer les métriques sur le nœud local. Puisqu'ils prédominent dans la référence implicite du GEVD-MWF-SDW, cela explique que le SAR_{img} augmente alors que le nombre de signaux compressés reçus diminue et que les autres métriques décroissent.

Cette étude permettra de mieux analyser les résultats des sections suivantes où, pour mieux quantifier l'impact des ruptures de liens sur la prédiction des masques TF uniquement, tous les signaux compressés de parole seront utilisés pour calculer les formateurs de voies. Cela n'est certes pas réalisable en pratique (si les signaux sont disponibles pour le calcul des formateurs de voies, ils le sont également pour la prédiction des masques), mais permet de mieux dissocier l'impact de canaux manquants sur la prédiction de masques de l'impact de canaux manquants sur la formation de voies. On retiendra de cette étude que l'absence des signaux compressés fait diminuer la réduction de bruit, mais dans une mesure limitée.

5.3.2. Résilience à un nombre variable de canaux avec un mécanisme « compresser et stimuler »

Dans cette section, un mécanisme d'attention SE est introduit avant le CRNN multinœud à la seconde étape de filtrage de Tango. Ses poids sont multipliés avec le tenseur d'entrée (dont les canaux manquants sont remplacés par une valeur constante), et le résultat est fourni au CRNN. Cette architecture sera notée SECRNN.

5.3.2.1. Résilience à un nombre variable de canaux

Quatre systèmes sont comparés afin de déterminer l'impact du mécanisme SE sur les performances du RNMuN. Le premier système est un CRNN mono-nœud. C'est la solution la plus simple pour avoir un RN indépendant du nombre de nœuds, puisque le RN ne prédit le masque qu'à partir du mélange enregistré par le microphone de référence du nœud local. Les signaux compressés, reçus ou pas, ne servent donc pas à prédire les masques TF. Il sera noté « CRNN_1 ». Le deuxième système est un CRNN multinœud mais qui n'a pas été entraîné dans des conditions

où des liens sont rompus entre les nœuds. Le CRNN n'est donc pas entraîné à voir des canaux constants, de valeur négative, lorsque un signal n'est pas reçu. Ce système est noté « $\text{CRNN}_{7,0}$ » car le CRNN a sept canaux en entrée (un mélange local et 3×2 signaux compressés), mais entraîné avec aucun canal manquant. Le troisième système est la même architecture de CRNN multinœud, mais entraînée à voir des canaux manquants. A chaque configuration d'entraînement, entre 0 et 3 liens sont choisis aléatoirement et rompus dans l'antenne de microphones. Ce système est noté « $\text{CRNN}_{7,0-3}$ ». Enfin le quatrième système est le SECRNN entraîné dans des configurations où un nombre aléatoire de liens, entre 0 et 3, est rompu. Il est noté « $\text{SECRNN}_{7,0-3}$ ». Le facteur de réduction r du mécanisme SE est pris égal à 2. Le tableau 5.2 résume les configurations d'entraînement des quatre systèmes comparés.

Notation	RN mononœud/ multinœud	Mécanisme d'attention	# de liens manquants à l'entraînement
CRNN_1	mononœud	Non	–
$\text{CRNN}_{7,0}$	multinœud	Non	0
$\text{CRNN}_{7,0-3}$	multinœud	Non	De 0 à 3
$\text{SECRNN}_{7,0-3}$	multinœud	Oui	De 0 à 3

TABLEAU 5.2. – Résumé des conditions d'entraînement des systèmes comparés.

Les résultats obtenus avec ces quatre systèmes sont représentés en figure 5.5. La lettre L dénote le nombre de liens rompus dans l'antenne de microphones, c'est-à-dire le nombre de nœuds déconnectés du reste de l'antenne.

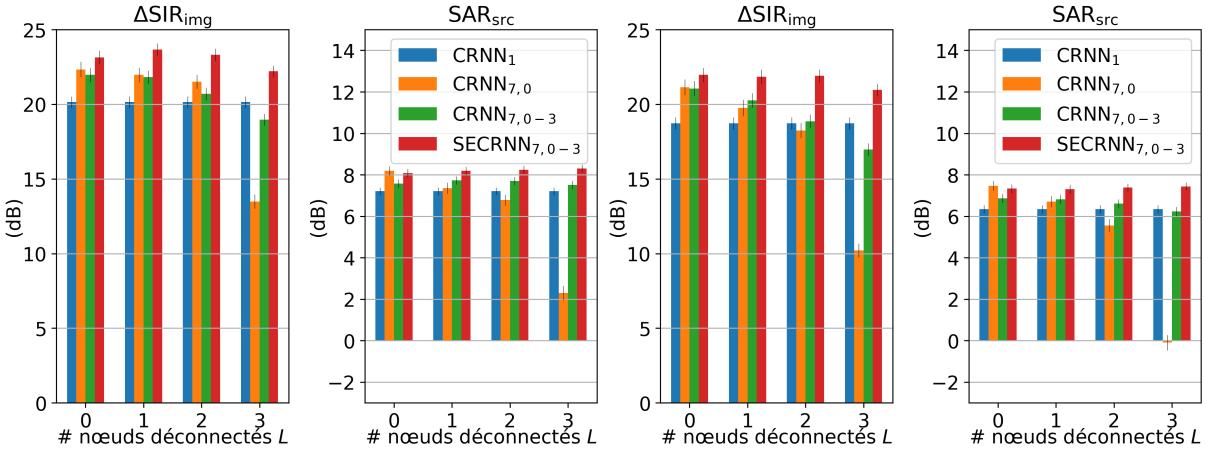
Afin d'analyser plus en détail le comportement de Tango à chaque nœud, les résultats sont représentés sur les quatre nœuds de chaque configuration, en les distinguant en fonction du SIR_{img} de sortie ; on représente sur la figure 5.5(a) les moyennes calculées sur le meilleur nœud de sortie de chaque configuration (c'est-à-dire sur le nœud avec le meilleur SIR_{img}). Les résultats de la figure 5.5(b) sont calculés sur le nœud avec le deuxième meilleur SIR_{img}. Les résultats de la figure 5.5(c) sont calculés sur le nœud avec le troisième meilleur SIR_{img}. Enfin les résultats de la figure 5.5(d) sont calculés sur le nœud avec le pire SIR_{img}.

Nous avons observé que le classement des nœuds est très fortement corrélé au nombre de signaux reçus. Les meilleurs nœuds sont les nœuds qui sont connectés avec le plus de nœuds, donc qui reçoivent le plus de signaux compressés. Les pires nœuds sont toujours ceux qui sont déconnectés du reste de l'antenne, donc ceux qui ne reçoivent aucun signal compressé, comme le nœud 2 sur la figure 5.1(a).

Par ailleurs, dans un souci de concision, étant donné que beaucoup de données sont déjà représentées sur les figures, seuls les $\Delta\text{SIR}_{\text{img}}$ et SAR_{src} sont rapportés. Les autres métriques jusqu'alors utilisées n'apportent pas d'information supplémentaire pertinente, et les figures peuvent ainsi être allégées.

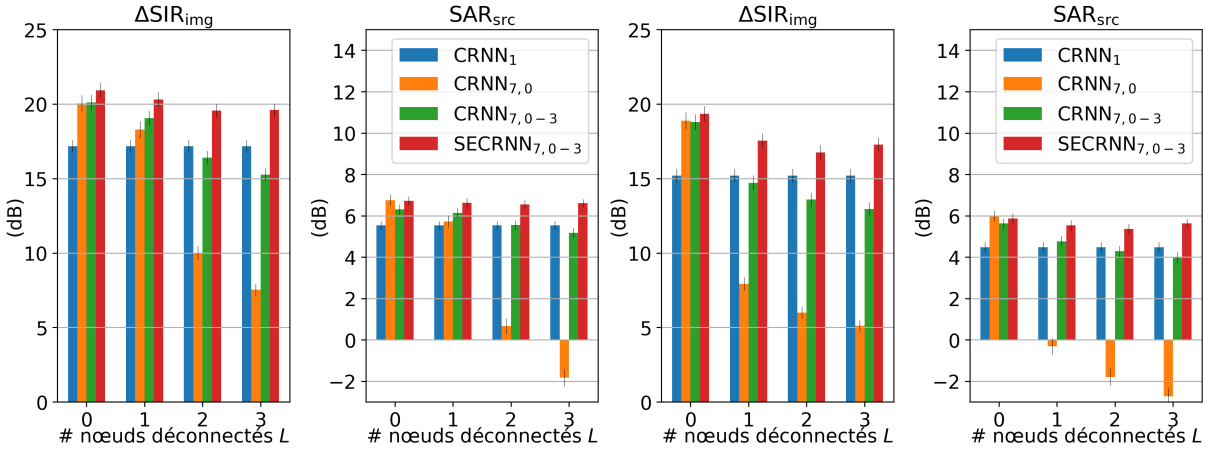
Les résultats obtenus avec le CRNN_1 peuvent servir de référence *a minima*. Lorsque les performances sont inférieures à celles obtenues avec ce système, il est préférable de ne pas du tout utiliser les signaux compressés pour prédire les masques TF. Les résultats de CRNN_1 ne varient pas en fonction du nombre de liens rompus, ce qui est logique car les RNMuN dans Tango ne reposent pas sur les signaux compressés pour prédire les masques TF.

Pour ce qui est du $\text{CRNN}_{7,0}$, ses performances dépendent beaucoup du fait que les signaux compressés soient reçus ou non. On remarque que les performances sont assez stables lorsqu'au moins un signal compressé est reçu, comme le montrent les métriques des figures 5.5(a) et 5.5(b) pour un nombre de liens rompus égal à 0, 1 ou 2. Cependant, les nœuds déconnectés du reste de



(a) Résultats au meilleur nœud de sortie

(b) Résultats au deuxième meilleur nœud de sortie



(c) Résultats au troisème meilleur nœud de sortie

(d) Résultats au pire nœud de sortie

FIGURE 5.5. – Résultats du rehaussement de la parole de Tango ($\Delta\text{SIR}_{\text{img}}$ et SAR_{src}) avec différents RN à la seconde étape de filtrage lorsque des nœuds sont déconnectés de l'antenne de microphones. Les résultats sont donnés aux différents nœuds de chaque configuration.

l'antenne ont de très faibles performances. Par exemple le troisième meilleur nœud (figure 5.5(c)) voit le $\Delta\text{SIR}_{\text{img}}$ chuter de 10 dB lorsque $L = 2$ et de 12.5 dB lorsque $L = 3$. De même, le SAR_{src} est inférieur à 1 dB pour $L = 2$ et même négatif pour $L = 3$.

Une plus grande résilience est obtenue en entraînant le RNMuN dans des conditions similaires à celles d'évaluation. Bien que les performances de $\text{CRNN}_{7,0-3}$ décroissent quand L croît, la diminution reste contenue. Le $\Delta\text{SIR}_{\text{img}}$ est toujours largement supérieur à 10 dB, même au niveau du pire nœud (donc ne recevant aucun signal compressé dès que $L \geq 1$) et le SAR_{src} n'est inférieur à 5 dB qu'au niveau du pire nœud pour $L \geq 1$.

Les meilleures performances sont obtenues avec le $\text{SECRNN}_{7,0-3}$. Quel que soit le nœud observé, les métriques restent élevées et stables même lorsque tous les nœuds sont déconnectés les uns des autres. Le $\Delta\text{SIR}_{\text{img}}$ diminue au maximum de 2 dB lorsque L passe de 0 à 3. Les variations de SAR_{src} dépassent quant à elles rarement l'intervalle de confiance. On observe même des performances supérieures avec le $\text{SECRNN}_{7,0-3}$ qu'avec les CRNN des autres systèmes pour $L = 0$, ce qui montre que le mécanisme d'attention ne permet pas seulement d'être résilient à un nombre variable de canaux, mais également d'améliorer la prédiction des masques TF. De plus, les performances avec le $\text{SECRNN}_{7,0-3}$ sont supérieures aux performances avec le CRNN_1 lorsque $L = 3$, donc lorsque les deux systèmes disposent d'autant d'information pour prédire les masques TF à la deuxième étape. Cela montre que le mécanisme d'attention ne sert pas uniquement à sélectionner les canaux utiles. Une étude pour expliquer les performances obtenues avec le mécanisme d'attention est proposée dans le paragraphe suivant. Dans l'ensemble, les expériences de ce paragraphe montrent que la méthode choisie pour répondre à la variation du nombre de nœuds dans une AAAH est une bonne solution.

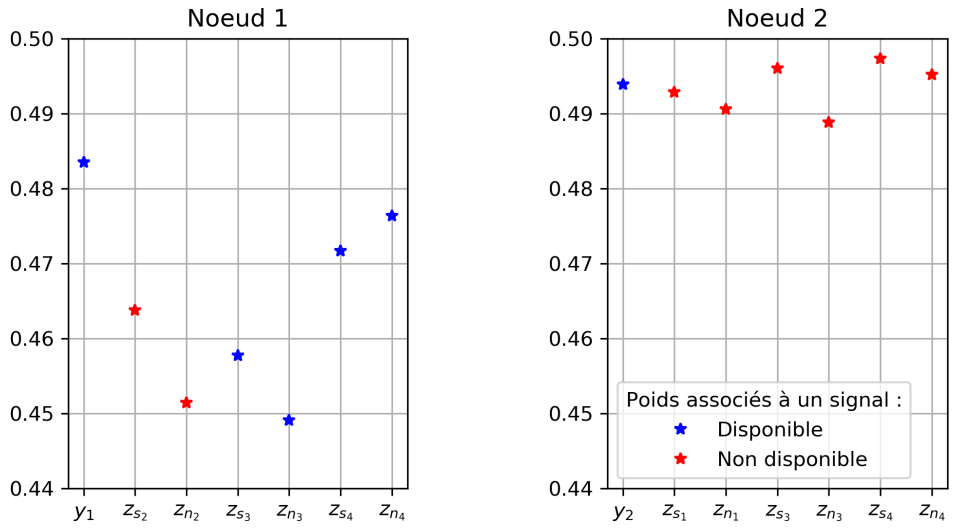
5.3.2.2. Dissociation des effets du mécanisme d'attention

L'intérêt d'utiliser le mécanisme d'attention à l'entrée du CRNN réside aussi dans le fait que les poids sont potentiellement interprétables. Comme dit dans l'introduction de cette section, l'idée d'utiliser des poids était d'accorder plus d'importance aux canaux « utiles » (c'est-à-dire aux canaux réellement reçus des autres nœuds) et d'en accorder moins à ceux qui n'ont pas été reçus. Cependant, nos observations n'ont pas permis de relever une quelconque correspondance entre la valeur des poids et celle des canaux.

La figure 5.6 illustre ceci de manière qualitative. Elle représente la valeur des poids d'une configuration où un seul lien est rompu, celui du $2^{\text{ème}}$ nœud. Les poids calculés sur le premier nœud (connecté au reste de l'antenne) sont représentés en figure 5.6(a) ; les poids calculés sur le nœud 2 (déconnecté du reste de l'antenne) sont représentés en figure 5.6(b). Les canaux manquants au niveau du nœud 1 sont les canaux 2 et 3. Si le mécanisme d'attention fonctionnait tel qu'initialement supposé, les poids appliqués sur ces canaux devraient être plus faibles, ce qu'on n'observe pas sur la figure 5.6(a). Le poids appliqué au premier canal du nœud 2, seul canal utile du nœud déconnecté, n'est pas non plus supérieur aux autres poids. Les autres configurations d'évaluation ne donnent pas plus lieu à une interprétation évidente des poids.

Dans ce paragraphe, nous proposons donc de dissocier les effets du mécanisme d'attention afin de comprendre comment il a permis d'améliorer les performances de Tango.

Nous considérons que le mécanisme SE intervient de deux manières. La première est par le fait de pondérer les canaux d'entrée par un poids. Cela revient à changer la dynamique des signaux en entrée. La deuxième intervention est le fait qu'un module apprenable (le SE) est ajouté au RN. Ce module peut avoir une influence sur l'apprentissage du reste du modèle, par exemple sur sa convergence d'apprentissage et sur la valeur du gradient, même si les valeurs en sa sortie ne sont pas pertinentes. Trois autres CRNN ont donc été entraînés afin de distinguer les effets de ces deux facteurs.



(a) Poids calculés sur un nœud connecté au reste de l'antenne

(b) Poids calculés sur un nœud déconnecté du reste de l'antenne

FIGURE 5.6. – Valeurs des poids d'un mécanisme SE calculés sur deux nœuds, l'un connecté et l'autre non. « z_{s_k} » est le signal compressé de parole reçu (ou non) du nœud k. « z_{n_k} » est le signal compressé de bruit reçu du nœud k.

Le premier modèle est un CRNN multinœud, sans mécanisme SE, mais dont les tenseurs d'entrée sont pondérés par des valeurs aléatoires comprises entre 0 et 1 à l'entraînement et à l'évaluation. Ce modèle est noté « aléa-CRNN$_{7,0-3}$ ». Le deuxième modèle est un SECRNN mais dont les poids du SE sont remplacés à chaque nouvelle propagation avant du RN (à l'entraînement comme à l'évaluation) par des valeurs aléatoires comprises entre 0 et 1. Il est noté « SE-aléa-CRNN$_{7,0-3}$ ». Enfin le dernier modèle est un SECRNN dont les poids ne sont pas multipliés avec le tenseur d'entrée. Cela revient donc à remplacer la valeur des poids par 1 avant multiplication avec le tenseur d'entrée. Ce modèle est noté « SE-1-CRNN$_{7,0-3}$ ». Les résultats de ces trois modèles, en plus de ceux obtenus avec le CRNN$_{7,0-3}$ et le SECRNN$_{7,0-3}$, sont représentés en figure 5.7. Cette figure représente les moyennes mesurées à tous les nœuds de toutes les configurations d'évaluation, afin d'avoir une vision plus succincte des résultats.

L'impact du module SE, indépendamment de la valeur de ses poids, peut être analysé en comparant les résultats obtenus avec aléa-CRNN$_{7,0-3}$ et SE-aléa-CRNN$_{7,0-3}$ d'une part, et avec CRNN$_{7,0-3}$ et le SE-1-CRNN$_{7,0-3}$ d'autre part. Dans les deux comparaisons, l'architecture qui comporte le mécanisme SE conduit à des performances supérieures à celles obtenues sans le mécanisme. Indépendamment de la valeur des poids multipliés avec le tenseur d'entrée, il semble donc profitable d'avoir ce module en plus dans le RN.

L'impact des poids est plus complexe à analyser. On peut étudier l'intérêt de pondérer le tenseur en entrée par des poids de valeur quelconque (aléatoire) en comparant les résultats obtenus avec aléa-CRNN$_{7,0-3}$ et CRNN$_{7,0-3}$, ainsi qu'avec SE-1-CRNN$_{7,0-3}$ et SE-aléa-CRNN$_{7,0-3}$. Il ressort de ces comparaisons que les poids aléatoires conduisent à des performances plus résilientes au nombre variable de canaux, car les métriques décroissent toujours moins vite lorsque les poids sont appliqués au tenseur d'entrée. Enfin, on peut étudier l'intérêt de pondérer le tenseur en entrée avec la bonne valeur des poids en comparant les résultats obtenus avec SE-1-CRNN$_{7,0-3}$, SE-aléa-CRNN$_{7,0-3}$ et SECRNN$_{7,0-3}$. Les SAR$_{\text{src}}$ obtenus avec SE-1-CRNN$_{7,0-3}$ et SECRNN$_{7,0-3}$ ne se distinguent de manière significative que lorsque $L = 3$, où il est préférable d'appliquer les bonnes valeurs de poids. Cette conclusion est vraie en termes de ΔSIR$_{\text{src}}$ quelle que soit la valeur

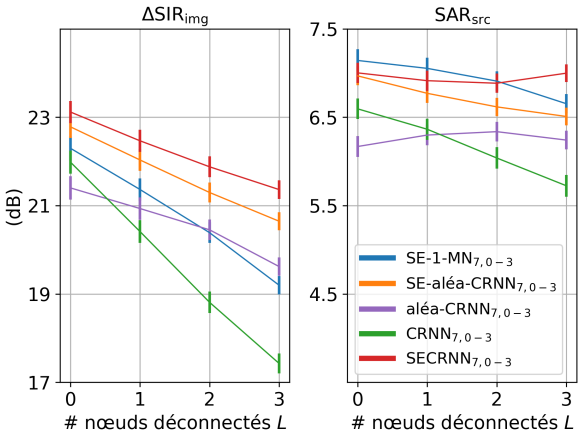


FIGURE 5.7. – Résultats de rehaussement de la parole de Tango ($\Delta\text{SIR}_{\text{img}}$ et SAR_{src}) avec un SECRNN lorsque des nœuds sont déconnectés de l'antenne de microphones. Différents modèles sont utilisés pour dissocier les effets des poids et du mécanisme d'attention SE seul dans le SECRNN.

de L. Néanmoins, les résultats obtenus avec SE-aléa-$\text{CRNN}_{7,0-3}$ sont à peine inférieurs à ceux obtenus avec le $\text{SECRNN}_{7,0-3}$. Cela laisse supposer que la pondération du tenseur en entrée a une influence non négligeable sur les performances du CRNN, mais que cette influence n'est pas liée à la valeur des poids eux-mêmes.

En conclusion, le mécanisme SE améliore les performances globales de Tango, tout en rendant les RNMuN de l'algorithme plus résilients à la variabilité de canaux. L'amélioration des performances semble surtout due au fait qu'un module supplémentaire soit présent dans l'architecture du RN, indépendamment de la valeur de ses sorties. La résilience du RN semble elle venir du fait que des poids sont appliqués sur les tenseurs en entrée, bien que la valeur de ces poids ne semble pas primordiale.

5.3.3. Résilience à un nombre variable de canaux avec un mécanisme d'auto-attention

Dans cette section, le module SE utilisé dans la section précédente est remplacé par un module SA. L'ambition de cette étude est de voir s'il est possible de rendre les poids du mécanisme plus interprétables, car l'interprétation des poids du SE n'a pas pu aboutir. Certaines expériences faisant intervenir des mécanismes SA ont déjà montré que les poids en sortie du mécanisme sont interprétables (Nicolson and Paliwal, 2020; Tolooshams et al., 2020), c'est pourquoi c'est cette architecture qui a été retenue.

Un nouveau modèle est entraîné dans les mêmes conditions que décrites dans la section 5.3 avec le module SA décrit en section 5.2 et illustré en figure 5.4. Un nombre aléatoire de liens est rompu dans chaque configuration d'entraînement. Les convolutions $\mathbf{K}$ et $\mathbf{Q}$ du mécanisme de SA sont appliquées sur l'axe temporel, et le nombre de canaux en sortie de convolution d est choisi égal à 42, ce qui double le nombre de trames en entrée. La convolution $\mathbf{V}$ est elle appliquée sur l'axe fréquentiel, avec un nombre de canaux en sortie de convolution maintenu à 257. Les résultats obtenus avec ce nouveau RNMuN, noté « $\text{SACRNN}_{7,0-3}$ sont représentés sur la figure 5.8. Ils sont accompagnés des résultats obtenus avec le $\text{CRNN}_{7,0-3}$ et le $\text{SECRNN}_{7,0-3}$. Ils correspondent là aussi à la moyenne prise sur tous les nœuds de chaque configuration.

Si les performances obtenues avec le SACRNN ne sont pas aussi bonnes qu'avec le SECRNN,

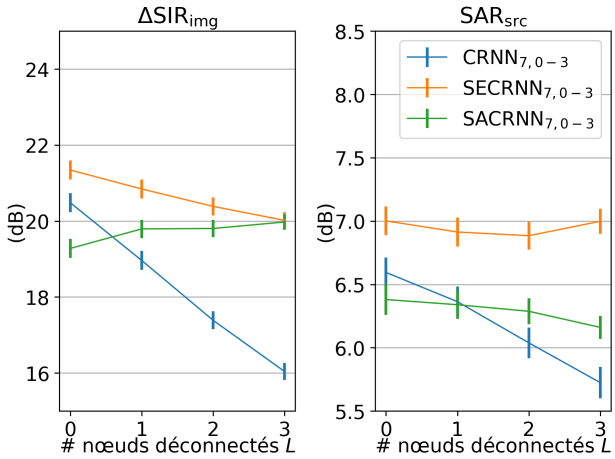


FIGURE 5.8. – Résultats de rehaussement de la parole de Tango ($\Delta\text{SIR}_{\text{img}}$ et SAR_{src}) avec un SACRNN lorsque des nœuds sont déconnectés de l'antenne de microphones.

elles sont stables en termes de SAR_{src} et le $\Delta\text{SIR}_{\text{img}}$ croît même lorsque L augmente. De plus, elles surpassent les performances du $\text{CRNN}_{7,0-3}$ dès qu'un nœud est déconnecté du reste de l'antenne en termes de $\Delta\text{SIR}_{\text{img}}$ et lorsque deux nœuds sont déconnectés en termes de SAR_{src}. Utiliser ce mécanisme d'attention peut être donc une solution envisageable pour augmenter la résilience de Tango face à un nombre variable de nœuds.

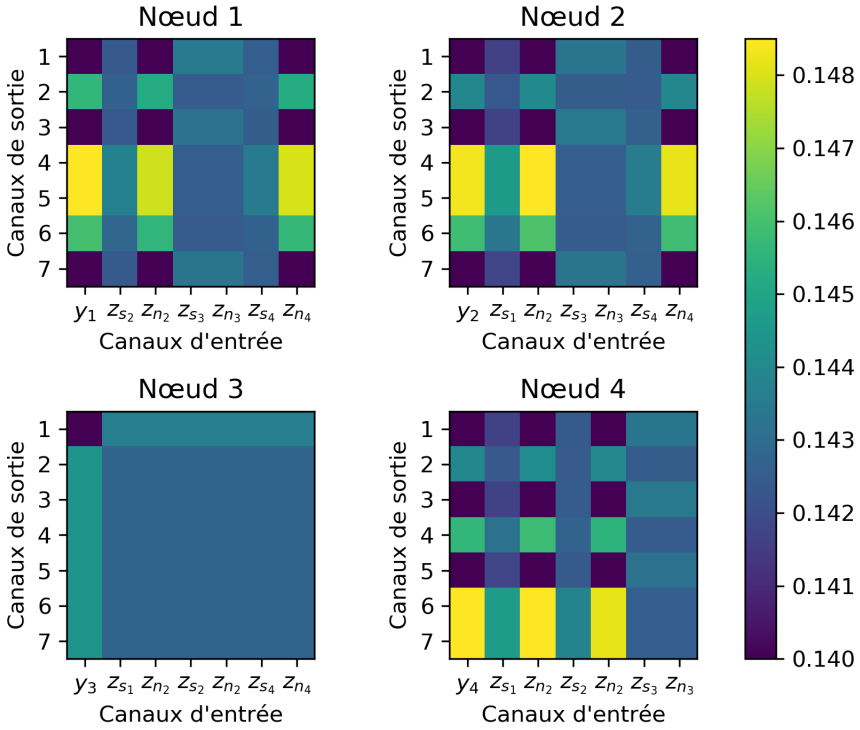
Pour ce qui est de l'interprétabilité des poids, la figure 5.9(a) représente les poids appliqués sur chacun des nœuds d'une des configurations d'évaluation, représentée en figure 5.9(b). Etant donné que les tenseurs de poids associés à chaque trame de sortie du masque TF sont de dimension $F \times C \times C$, cette figure représente la moyenne des poids sur toutes les bandes de fréquences et sur toutes les trames temporelles du masque prédit. Il en résulte une matrice $C \times C$. Rappelons ici l'équation (5.8) pour une meilleure lecture de la figure 5.9 :

$$\mathbf{o}_f = \mathbf{v}_f \mathbf{P}_f^H \,. \tag{5.9}$$

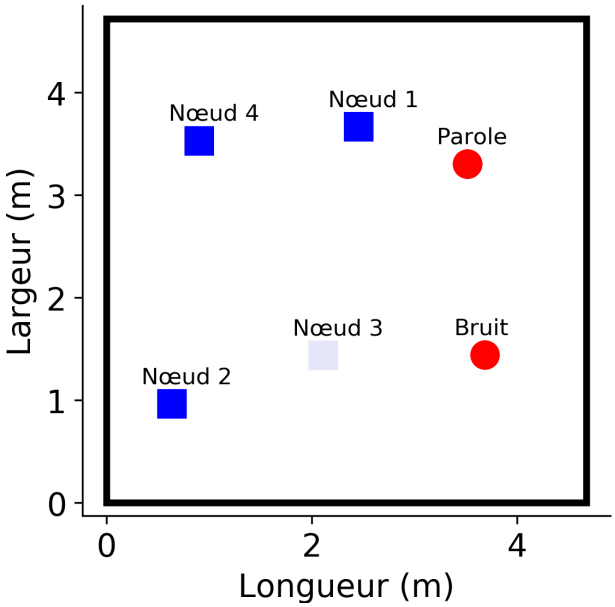
La $k^{\text{ème}}$ ligne de $\mathbf{P}_f$ est donc la moyenne des poids appliqués sur les canaux d'entrée pour obtenir la valeur du $k^{\text{ème}}$ canal de sortie, fourni au CRNN. Symétriquement, la $k^{\text{ème}}$ colonne est la moyenne des poids appliqués sur le $k^{\text{ème}}$ canal de $\mathbf{v}$. Sur la figure 5.9(a), les canaux d'entrée sont placés de telle sorte que la mixture locale est le premier canal, suivi des signaux compressés ordonnés dans l'ordre croissant des nœuds les émettant (ou non si le nœud émetteur n'est pas connecté au reste de l'antenne). L'estimation compressée de parole $z_{s,j}$ du nœud j est toujours placée (ou envoyée) avant l'estimation compressée de bruit $z_{n,j}$ du même nœud. Par exemple les canaux fournis au RNNMuN du nœud 3 sont :

$$\bar{\mathbf{y}}_3 = [y_{3,1}, \quad z_{s,1}, \quad z_{n,1}, \quad z_{s,2}, \quad z_{n,2}, \quad z_{s,4}, \quad z_{n,4}]^T \,. \tag{5.10}$$

Dans la configuration de la figure 5.9, un nœud avait été déconnecté du reste de l'antenne, le troisième. On remarque que quasiment tous les poids appliqués sur ce canal sont égaux, et que seul le mélange local (première colonne de la matrice de poids) est (à peine) plus sollicité que les autres. En revanche, sur les nœuds 1, 2, et 4, les poids appliqués sur les canaux reçus du nœud 3 sont de valeurs plus faibles que ceux appliqués sur les autres canaux. Les autres canaux semblent sollicités de manière assez équilibrée, même si cet exemple suggère que les estimations compressées de parole sont moins importantes que celles de bruit, sans qu'il soit évident de corréler cela avec la configuration spatiale. En particulier, au niveau du nœud 2 dont le mélange



(a) Moyenne des poids



(b) Configuration spatiale d'évaluation. Le RSB des sources non convoluées vaut 1.4 dB.

FIGURE 5.9. – (a) Moyennes sur le plan TF des poids d'un mécanisme de SA aux quatre nœuds de l'antenne de microphones représentée sur la figure (b) ; le troisième nœud est déconnecté de l'antenne.

est à un RSB de -1 dB, on s'attendrait à ce que les poids appliqués sur les estimations compressées de parole soient plus élevés que ceux appliqués sur les estimations compressées de bruit, ce qui n'est pas le cas.

Toutefois, bien que ces poids puissent être à première vue interprétés, une grande réserve est émise quant au caractère significatif des conclusions. Les valeurs des poids ne diffèrent en effet que de l'ordre du millième, comme le montre la barre de couleurs de la figure 5.9. Si la valeur relative des poids semble correspondre à la nature des signaux en entrée, il est peu probable que leur valeur absolue serve réellement au RN, car la dynamique des signaux en entrée diffère largement plus que de l'ordre du millième.

5.4. Conclusion

Ce chapitre était dédié à l'étude de la résilience de Tango face à un nombre variable de nœuds. Bien que l'algorithme Tango initial ne soit pas résilient à des ruptures de liens dans une AAAH, nous avons proposé une solution basée sur l'introduction d'un mécanisme d'attention à l'entrée du CRNN multinœud de Tango pour le rendre plus résilient. Cette solution a même permis d'augmenter ses performances lorsque tous les nœuds de l'antenne acoustique sont connectés. Deux mécanismes d'attention ont été comparés, un SE et un SA, dont les effets ont été analysés en détail et qui ont tous les deux permis de rendre Tango plus résilient.

On relève deux limites à l'utilisation des mécanismes d'attention. La première est que les valeurs des poids ne sont que partiellement interprétables. Avec le SE, aucune corrélation n'a pu être trouvée entre la valeur des poids et celle des canaux en entrée. Avec le SA, les valeurs semblent effectivement correspondre au fait qu'un canal ait été, ou non, reçu, mais les valeurs ne se distinguent qu'au millième près.

Par ailleurs, si les résultats obtenus avec les RN restent bons grâce aux mécanismes d'attention même lorsque plusieurs nœuds (voire tous) sont déconnectés de l'antenne, l'information apportée par le mécanisme d'attention est de faible intérêt : on sait avant même le mécanisme d'attention quels canaux ont été reçus et ceux qui ne l'ont pas été. Néanmoins, ces expériences ouvrent une porte vers l'utilisation de RN qui seraient capables de sélectionner les canaux les plus importants pour une tâche donnée. Avec ces RN, il deviendrait alors possible de réduire la consommation en bande passante en n'envoyant plus les signaux qui sont négligés par le RN. Elles montrent également qu'une légère modification de l'architecture du RN permet d'augmenter significativement ses performances et de le rendre plus résilient aux conditions d'utilisation spécifiques aux AAAH. Dans la suite de cette partie, nous considérons un autre mécanisme d'attention pour adapter Tango aux cas où les nœuds d'une AAAH ne sont pas synchronisés.

Le tableau 5.3 résume les conclusions tirées des réflexions et expériences de ce chapitre.

Section	Points-clé
5.3.1	Un plus faible nombre de signaux compressés à la seconde étape de Tango diminue les performances de filtrage, mais de manière limitée.
5.3.2	• Les RNMuN non-entraînés avec un nombre variable de nœuds ne sont pas résilients aux ruptures de lien. • Un mécanisme d'attention SE augmente les performances de Tango. • Un mécanisme d'attention SE rend les RNMuN de Tango résilients à des canaux manquants. • Les poids du SE ne sont pas interprétables dans nos évaluations.
5.3.3	• Un mécanisme d'attention SA rend les RNMuN de Tango résilients à des canaux manquants. • Les poids du SA sont interprétables. • Les poids du SA n'ont probablement pas d'influence sur les performances globales de Tango.

TABLEAU 5.3. – Points-clé à retenir du chapitre 5.

6. Extension de Tango pour la prise en charge des cas d'asynchronisation entre les nœuds d'une antenne acoustique ad-hoc

Ce chapitre est dédiée à l'analyse de l'impact de l'asynchronisation entre les nœuds d'une antenne acoustique ad-hoc (AAAH) en séparant les effets du décalage d'horloges de ceux de la dérive d'horloge. Après une présentation du contexte de ce chapitre en section 6.1, nous montrons l'impact de l'asynchronisation des nœuds sur le calcul des métriques en section 6.2 et sur les performances de Tango en section 6.3. Une solution pour compenser les baisses de performances liées à l'asynchronisation est proposée en section 6.4.

6.1. Introduction

6.1.1. Asynchronisation de deux signaux

Dans cette section, nous supposons que les signaux partagent la même fréquence d'échantillonnage (FE) nominale sur tous les nœuds. Dans le cas contraire, un ré-échantillonnage sur certains nœuds ramène le cas d'étude à notre hypothèse de départ.

Le terme « asynchronisation » de deux signaux décrit le fait que les $n^{\text{èmes}}$ échantillons de chacun des signaux ne correspondent pas à un même instant temporel. Soient y_1 et y_2 ces deux signaux. Bien que ces deux microphones partagent la même FE nominale, en pratique, leur FE réelle diffère de la FE nominale, ce qui introduit une *dérive d'horloge*, qui fluctue au cours du temps. De plus, ces deux microphones ne commencent pas à enregistrer au même instant, ce qui introduit un *décalage d'horloge* qui est lui fixe dans le temps. En prenant le début de l'acquisition de y_1, échantillonné à la fréquence f_{s_1} donc à la période d'échantillonnage T_{s_1}, comme référence pour le début de l'acquisition, le $n^{\text{ème}}$ échantillon de y_1 est enregistré à l'instant :

$$t_1 = n \cdot T_{s_1} \,.$$

Avec f_{s_2} la FE de y_2 et T_{s_2} sa période d'échantillonnage, le $n^{\text{ème}}$ échantillon de y_2 est enregistré à l'instant :

$$t_2 = n \cdot T_{s_2} + \tau_2 \,,$$

où τ_2 est l'écart entre le début d'acquisition du premier signal et celui du second signal. Il est appelé *décalage* d'horloge d'échantillonnage et peut être négatif dans le cas où l'acquisition de y_2 commence avant celle de y_1.

On peut écrire $f_{s_2} = (1 + \epsilon_2) f_{s_1}$. La grandeur $\epsilon_2 \ll f_{s_1}$ caractérise la dérive d'horloge d'échantillonnage du deuxième microphone. Le fait que f_{s_1} et f_{s_2} ne soient pas égales implique un « éloignement » au cours du temps des instants auxquels les $n^{\text{èmes}}$ échantillons des deux signaux sont acquis. Schmalenstroeer et al. (2015) ont montré que ϵ_2 varie peu au cours du temps, c'est pourquoi les variables f_{s_1} et f_{s_2} sont considérées indépendantes du temps. La différence entre les

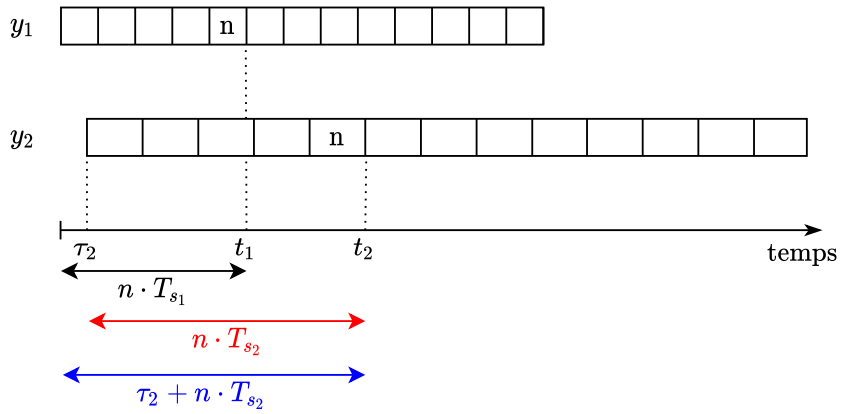


FIGURE 6.1. – Illustration de l'asynchronisation entre deux signaux

instants t_1 et t_2 traduit l'asynchronisation des deux signaux y_1 et y_2. On a :

$$t_2 = n \cdot T_{s_2} + \tau_2 \tag{6.1}$$
$$= \frac{n}{1+\epsilon_2} T_{s_1} + \tau_2 \tag{6.2}$$
$$\sim n \cdot (1-\epsilon_2) T_{s_1} + \tau_2 \qquad \text{pour } \epsilon_2 \text{ proche de } 0 \tag{6.3}$$
$$\sim t_1 - \underbrace{n \cdot \epsilon_2 T_{s_1}}_{\text{dérive}} + \underbrace{\tau_2}_{\text{décalage}} . \tag{6.4}$$

L'équation (6.4) met en évidence le fait que la dérive d'horloge est dépendante du temps, alors que le décalage est fixe. La figure 6.1 illustre ces deux phénomènes de manière schématique.

L'asynchronisation entre des signaux apparaît lorsque les microphones ne partagent pas la même carte d'acquisition, ce qui est le cas des différents nœuds d'une AAAH. Elle est due au fait que les cartes d'acquisitions ont des implémentations matérielles et logicielles différentes qui entraînent des dérives et décalages d'horloge (Ceolini et al., 2020; Schmalenstroeer et al., 2015). Dans notre cadre d'étude, nous supposerons que les signaux d'un même nœud sont synchrones, puisqu'ils partagent la même carte d'acquisition. En revanche, deux signaux de deux nœuds différents ne seront pas synchrones. L'asynchronisation peut avoir une forte influence sur le rehaussement de la parole, en particulier sur la formation de voies, qui repose sur les relations temporelles des signaux, par exemple pour estimer la direction d'arrivée de la source. L'asynchronisation perturbe ces structures temporelles et dégrade les performances de rehaussement de la parole. L'influence de l'asynchronisation sur le rehaussement de la parole a été largement étudiée (Lienhart et al., 2003; Cherkassky et al., 2015; Zeng and Hendriks, 2015; Schmalenstroeer and Haeb-Umbach, 2018b) et des solutions proposées pour resynchroniser les signaux sont présentées dans la section suivante.

6.1.2. Solutions à l'asynchronisation de signaux

Deux approches se distinguent pour resynchroniser les signaux. La première se base sur l'échange de signaux spécifiques à la resynchronisation, qui peuvent être soit des signaux de calibration (Lienhart et al., 2003; Wehr et al., 2004), soit des tampons temporels (Schenato and Fiorentin, 2011; Chaudhari, 2011; Schmalenstroeer et al., 2015; Ceolini et al., 2020). Les autres approches sont dites *aveugles*, car elles ne disposent que des signaux mesurés par les microphones pour compenser l'asynchronisation. Pour cela, deux grandeurs peuvent être exploitées. La première

est la fonction de cohérence, dont la phase varie linéairement dans le temps sous l'effet d'une dérive d'horloge. En modélisant cette variation linéaire à partir de la fonction de cohérence de deux signaux, il est alors possible de retrouver la dérive d'horloge (Markovich-Golan et al., 2012a; Bahari et al., 2015; Cherkassky et al., 2015; Schmalenstroeer et al., 2017). L'autre grandeur est la corrélation entre les signaux mesurés, qui est maximale lorsque les signaux partagent la même FE (Miyabe et al., 2013; Wang and Doclo, 2016; Cherkassky and Gannot, 2017; Chinaev et al., 2021). Une fois que le décalage ou la dérive d'horloge d'échantillonnage a pu être estimé à l'aide d'une de ces méthodes, les signaux peuvent être ré-échantillonnés. Plusieurs méthodes ont également été proposées pour optimiser cette étape (Markovich-Golan et al., 2012a; Schmalenstroeer and Haeb-Umbach, 2018a).

Les limites de ces méthodes sont qu'elles rajoutent plusieurs étapes de calcul dans le traitement du signal, d'abord pour estimer les paramètres d'asynchronisation, puis pour ré-échantillonner les signaux. Tout cela alourdit l'ensemble du traitement. Il existe certains algorithmes de rehaussement de la parole dans des antennes de microphones asynchrones qui n'appliquent aucune solution de re-synchronisation, sans que cela n'impacte de manière réellement négative les résultats finaux (Chiba et al., 2014; Corey and Singer, 2018). Plutôt que de resynchroniser les signaux, il est donc possible de concevoir des algorithmes résilients à l'asynchronisation pour effectuer du rehaussement de la parole avec des signaux asynchrones. Afin de simplifier notre système, nous décidons de suivre une logique similaire dans cette section. Nous y étudions l'impact de l'asynchronisation sur Tango et nous proposons une solution pour la pallier sans qu'il soit nécessaire d'estimer les décalages et dérives d'horloge.

6.1.3. Contributions du chapitre

La contribution de cette partie de la thèse consiste à évaluer l'impact de l'asynchronisation sur les performances de Tango, et plus particulièrement sur les performances des réseaux de neurones (RN) qu'il fait intervenir. Bien que, comme présenté auparavant, de nombreux travaux fassent état de l'impact de l'asynchronisation des antennes acoustiques sur le rehaussement de la parole, nous n'avons connaissance d'aucun travail sur l'impact de l'asynchronisation des antennes acoustiques sur les performances de RN. C'est ce que nous proposons d'étudier dans la suite de ce chapitre.

Notre contribution porte sur deux points. Le premier point est l'évaluation empirique de l'impact de l'asynchronisation sur Tango, en distinguant l'impact sur la formation de voies de l'impact sur la prédiction des masques temps-fréquence (TF) par les réseaux de neurones multinœuds (RNMuN). Le second point de notre contribution est que nous proposons une solution pour réduire voire supprimer l'impact négatif de l'asynchronisation sur les performances de Tango. Comme dans le chapitre 5, cette solution se base sur l'introduction d'un mécanisme d'attention dont le but est de réaligner implicitement les signaux dans le plan de la transformée de Fourier à court terme (TFCT). L'asynchronisation ayant deux facteurs, le décalage et la dérive d'horloge, nous étudions séparément ces deux types d'asynchronisation dans nos expériences.

Le reste de ce chapitre est organisé de la manière suivante. Une première étude en section 6.2 évalue l'impact de l'asynchronisation sur le calcul des métriques seules, indépendamment du traitement appliqué sur les signaux. La section 6.3 fait état de l'impact de l'asynchronisation sur les performances de Tango, en analysant l'impact sur la formation de voies seule d'une part, et sur la prédiction des masques TF par les RNMuN d'autre part. Dans la section 6.4, nous proposons deux solutions pour pallier l'asynchronisation et obtenir de bons résultats de rehaussement de la parole malgré elle. La section 6.5 conclut ce chapitre.

6.2. Etude préliminaire de l'asynchronisation sur les métriques

Afin de pouvoir mieux interpréter les métriques dans la suite des expériences, nous évaluons dans un premier temps l'impact de l'asynchronisation sur les métriques.

6.2.1. Cas d'un décalage d'horloge

La simulation d'un décalage d'horloge est effectuée en ajoutant des 0 au début (resp. à la fin) d'un signal, représentant un retard (resp. une avance) d'horloge. Pour cette expérience, un signal est décalé d'un nombre différent d'échantillons, et les métriques de rehaussement de la parole sont calculées entre le signal d'origine et sa version décalée. Puisque le contenu du signal est absolument identique dans les deux versions, les métriques devraient être maximales indépendamment du décalage. Néanmoins, les métriques comme le STOI, qui comparent la forme d'onde d'une référence à celle du signal estimé, peuvent être sensibles à ce décalage et exprimer artificiellement des distorsions.

100 signaux sont aléatoirement pris dans le jeu d'évaluation *aléatoire* du corpus Disco (cf. section 3.4) pour calculer ces métriques. Ce sont des signaux de parole propre convoluée par la réponse impulsionnelle (RI) de la salle. La moyenne des SDR et STOI mesurés entre ces 100 paires est représentée en figure 6.2. Des SIR, SAR et SDR, seul le SDR est représenté car le SIR n'est pas du tout impacté par le décalage des signaux, si bien que sa valeur est toujours infinie. Le SDR est donc égal au SAR.

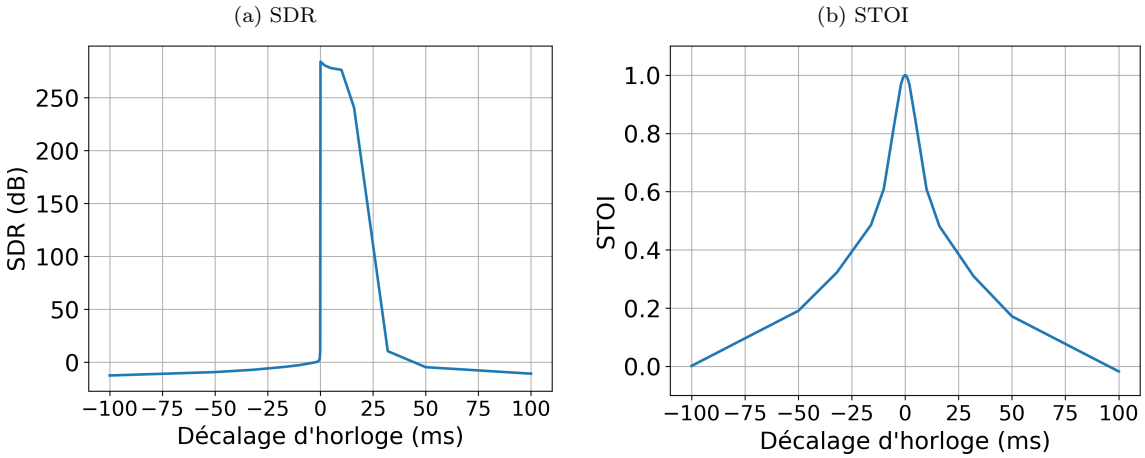


Figure 6.2. – Métriques de rehaussement de la parole calculées lorsque le signal estimé est une version décalée du signal de référence. Les résultats sont reportés pour une moyenne de 100 signaux du corpus Disco.

On observe deux phénomènes à partir des résultats portés sur la figure 6.2. Le premier est que le SDR n'a pas de comportement symétrique lorsque le décalage est positif ou négatif. Pourtant, les distorsions ne sont pas plus fortes lorsqu'un signal est avancé ou retardé. Le phénomène observé sur la figure 6.2(a) tient au fait que pour calculer le SDR, un filtre à réponse finie est appliqué sur la référence, de telle sorte que la corrélation soit maximale entre le signal ainsi filtré et le signal estimé (le signal décalé dans notre cas). Si un filtre à réponse finie peut compenser des retards de signaux, ce qui est le cas avec des décalages d'horloge positifs, il ne peut pas compenser le fait que le signal de référence soit en avance, ce qui arrive avec des décalages d'horloge négatifs [1]. Dans la suite, nous ne considérerons que des décalages d'horloge positifs, ce qui ne change rien à

1. On voit ici un avantage et une limite d'appliquer un tel filtre à réponse finie sur le signal de référence.

τ (ppm)	FE (Hz)	Durée en secondes pour décaler de 256 échantillons
5	16000,08	3200
50	16000,8	320
100	16001,6	160

TABLEAU 6.1. – Équivalences entre la dérive d'horloge et la durée nécessaire pour qu'une dérive donnée conduise à un décalage d'une trame TFCT.

la portée de notre analyse étant donnée la symétrie du phénomène de décalage. Cette symétrie est d'ailleurs bien exprimée par le STOI. De plus, comme notre analyse porte sur un décalage d'horloge relatif à un des nœuds de l'AAAH (et non à une référence absolue), il suffit de prendre le nœud le plus en avance comme référence pour n'avoir plus que des décalages positifs.

La deuxième observation est que les métriques sont sensibles au décalage d'horloge. Le STOI perd 20% de sa valeur maximale dès 6 ms de décalage d'horloge, soit 96 échantillons à 16 kHz. Le SDR est quant à lui plus résilient, mais on observe une chute importante de sa valeur lorsque le décalage dépasse les 16 ms, soit 256 échantillons. Il est intéressant de noter que cette chute intervient lorsque le décalage correspond à une trame de TFCT.

6.2.2. Cas d'une dérive d'horloge

Une dérive d'horloge est simulée par le ré-échantillonnage des signaux d'un nœud j à une fréquence $f_{s_j} = (1 + \epsilon_j) f_{s_k}$, avec le nœud k considéré comme nœud de référence. Différentes valeurs de ϵ_j seront considérées, en parties par million (ppm). Le tableau 6.1 montre l'équivalence entre les valeurs de dérive d'horloge en ppm et celles en Hz, ainsi que la durée nécessaire pour qu'une dérive donnée conduise à un décalage de 256 échantillons (soit une trame TFCT).

Le fait que l'ensemble du signal soit ré-échantillonné traduit la stabilité de la dérive d'horloge au cours du temps. La même expérience qu'en 6.2.1 est répétée : 100 signaux sont aléatoirement pris dans le jeu d'évaluation du corpus DISCO. Chacun de ces signaux est ré-échantillonné avec différentes dérives d'horloge comprises entre -200 ppm et 200 ppm. Les métriques mesurées entre le signal initial et le signal asynchrone sont représentées en figure 6.3.

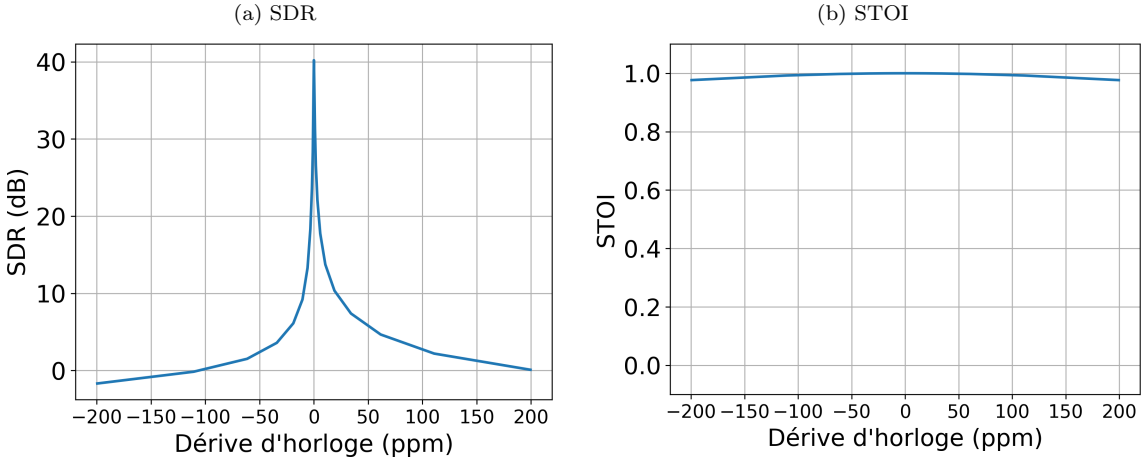


FIGURE 6.3. – Métriques de rehaussement de la parole calculées lorsque le signal estimé dérive par rapport au signal de référence. Les résultats sont reportés pour une moyenne de 100 signaux du corpus DISCO.

Dans cette expérience, comme dans la précédente, le SIR est infini quelle que soit la dérive d'horloge, si bien que le SDR est égal au SAR. On observe que, comme au cours de l'expérience précédente, le SDR est sensible à une dérive d'horloge, bien que dans ce cas, le fait que la dérive soit négative ou positive a moins d'influence sur le calcul de la métrique. Il semble toutefois que des dérives négatives (c'est-à-dire des ré-échantillonnages à des FE plus faibles que la FE de référence) aient un impact plus important sur la métrique. Comme avec le décalage d'horloge, ce phénomène pourrait s'expliquer par le fait qu'un signal ré-échantillonné à une FE plus faible semble en avance par rapport à sa version initiale. En effet, en reprenant l'équation (6.4), on a, avec $\tau_2 = 0$:

$$t_2 = t_1 \overbrace{\underbrace{-\, n \cdot \epsilon_2 T_{s_1}}_{<0}}^{>0} \tag{6.5}$$

$$t_2 > t_1 \,. \tag{6.6}$$

Comme argumenté précédemment, l'implémentation du SDR ne permet pas de compenser l'avance d'un signal estimé sur sa référence, ce qui explique que cette métrique soit plus sensible aux dérives d'horloge négatives qu'aux dérives d'horloge positives.

Cette expérience montre également que l'impact de la dérive d'horloge sur le SDR apparaît même pour de faibles valeurs de dérive d'horloge. Cela s'explique en partie par le simple fait que le signal soit ré-échantillonné, puisque même avec une dérive de 0 ppm, l'étape de ré-échantillonnage amène un SDR de quelques 40 dB alors que le SDR entre deux signaux identiques devrait être infini (au-delà de 200 dB dans des simulations, cf. figure 6.2(a)). Cependant, cela n'explique pas toute la sensibilité de la métrique à la dérive d'horloge, car le SDR décroît lorsque la dérive augmente. Afin de dissocier les effets du ré-échantillonnage et de la dérive d'horloge sur le calcul de la métrique, les mêmes mesures sont effectuées que précédemment, mais où la référence utilisée est le signal convolué, ré-échantilloné à 16000 Hz (donc pour une dérive nulle). Les résultats correspondants sont représentés en figure 6.4.

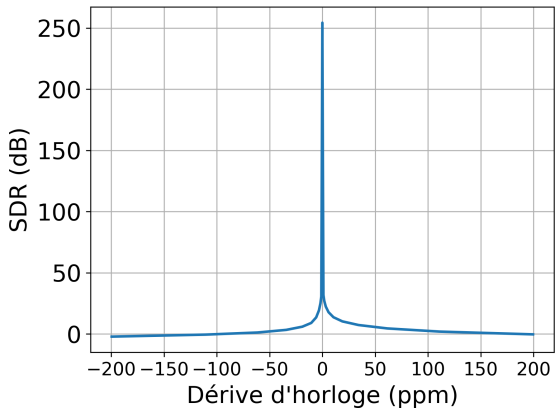


FIGURE 6.4. – SDR calculé lorsque le signal estimé dérive par rapport au signal de référence, qui est ré-échantillonné avec une dérive nulle. Les résultats sont reportés pour une moyenne de 100 signaux du corpus DISCO.

Mis à part pour une dérive nulle, les valeurs du SDR sont exactement les mêmes dans ce cas de figure. La sensibilité du SDR est donc liée à la dérive d'horloge, et non au ré-échantillonnage des signaux.

Le STOI en revanche est très stable en fonction de la dérive d'horloge. Même pour de fortes valeurs de dérive, il reste très proche de 1, sa valeur maximale. Cette métrique est donc robuste

aux ré-échantillonnages, ce qui s'explique par le fait que les signaux sont tous ré-échantillonnés à la fréquence de 10 kHz lors du calcul de la métrique (Taal et al., 2010). Pour les mêmes raisons que pour le décalage d'horloge, les dérives d'horloge dans la suite seront prises toujours positives, quitte à considérer le nœud avec la plus faible FE comme nœud de référence.

Pour éviter que l'asynchronisation ne fasse baisser artificiellement les métriques sans pour autant que cela traduise une réelle baisse de la qualité ou de l'intelligibilité des signaux, il convient dans la suite de ce chapitre de considérer comme références les signaux *après* l'ajout de l'asynchronisation. Puisque nous simulons l'asynchronisation au niveau de chaque nœud après la convolution avec les RI de la pièce, nous prendrons comme référence les signaux convolués qui ne sont donc pas décalés par rapport au signal estimé.

6.3. Impact de l'asynchronisation sur Tango

Dans cette section, nous étudions l'impact de l'asynchronisation des signaux sur les performances de Tango. Afin de dissocier les effets de cette asynchronisation sur la performance des RN de ses effets sur les formateurs de voies, nous effectuons cette évaluation en deux temps. Dans un premier temps, l'impact de l'asynchronisation est évalué lorsque les masques TF nécessaires dans le processus de Tango sont des masques oracles. Cette analyse est présentée en section 6.3.1. Dans un second temps, les masques oracles sont remplacés par les masques prédits par les RN. Cette analyse est présentée en section 6.3.2. Dans chacune des analyses, nous évaluons les effets d'un décalage d'horloge tout comme ceux d'une dérive d'horloge.

Dans la suite de ce chapitre, nous supposons que tous les microphones d'un même nœud sont synchronisés, car ils partagent la même carte d'acquisition, mais que les différents nœuds d'une AAAH ne sont pas synchronisés entre eux. Dans chaque configuration d'entraînement ou d'évaluation, un nœud est arbitrairement choisi comme nœud de référence : sa FE reste inchangée (à 16 kHz) et son décalage d'horloge est fixé à 0. Les trois autres nœuds subissent une dérive positive et un décalage d'horloge positif.

Les résultats présentés dans la suite de ce chapitre seront toujours la moyenne sur l'ensemble du jeu d'évaluation des métriques mesurées au niveau du nœud de chaque configuration avec le meilleur $\mathrm{SIR}_{\mathrm{img}}$ en sortie de filtrage. Suite à l'étude préliminaire de la section 6.2, seules les références convoluées seront considérées pour calculer les métriques de rehaussement de la parole, si bien que les métriques reportées seront le $\Delta\mathrm{SIR}_{\mathrm{img}}$, le $\mathrm{SAR}_{\mathrm{img}}$ et le $\mathrm{STOI}_{\mathrm{img}}$.

6.3.1. Impact sur la formation de voies

Dans ce premier jeu d'expériences, les masques TF utilisés dans Tango sont les masques oracles. Cela permet d'estimer à quel point le filtre spatial est impacté par les asynchronisations, indépendamment des performances des RN.

Pour un nœud k, le masque TF oracle est calculé d'après l'équation (3.13), rappelée ici :

$$m_{\mathrm{s},k1}(t,f) = \frac{|s_{k,1}(t,f)|}{|s_{k,1}(t,f)| + |n_{k,1}(t,f)|}\,. \tag{6.7}$$

$s_{k,1}$ et $n_{k,1}$ sont respectivement les composantes de parole et de bruit dans le mélange du premier microphone du nœud k. Comme ces signaux contiennent également un décalage d'horloge par rapport au nœud de référence, le masque TF oracle est bien aligné avec les signaux $s_{k,1}$ et $n_{k,1}$ (également nécessaire au calcul des métriques), ainsi qu'avec le mélange.

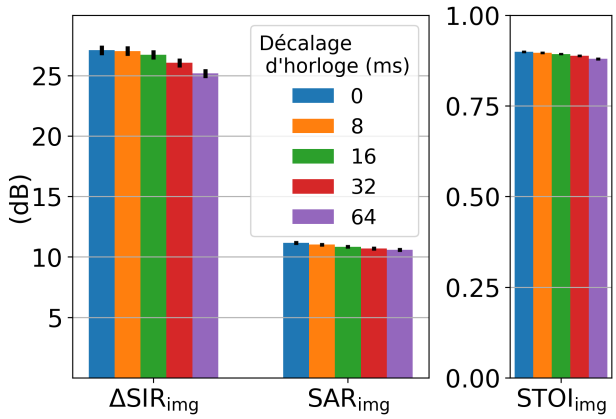


FIGURE 6.5. – Résultats de rehaussement de la parole de Tango avec des masques TF oracles lorsque les signaux de nœuds différents de l'AAAH sont décalés d'un nombre variable d'échantillons.

6.3.1.1. Cas d'un décalage d'horloge

Dans cette section, les données d'évaluation sont issues du jeu d'évaluation *aléatoire* du corpus DISCO (cf. section 3.4). 5 conditions d'évaluation sont comparées :

Cas 1 : Les signaux ne sont pas décalés. Il s'agit des mêmes signaux que ceux considérés dans les expériences des chapitres précédents.

Cas 2 : Les signaux d'un même nœud sont décalés d'une durée prise de manière uniformément aléatoire entre 0 ms et 8 ms. Ceci est fait pour les quatre nœuds de chaque configuration simulée, c'est-à-dire que le décalage d'horloge est constant et égal pour tous les signaux d'un même nœud, mais différent pour les signaux de nœuds différents.

Cas 3 : Les signaux d'un même nœud sont décalés d'une durée prise aléatoirement entre 0 ms et 16 ms.

Cas 4 : Les signaux d'un même nœud sont décalés d'une durée prise aléatoirement entre 0 ms et 64 ms.

Les résultats de rehaussement de la parole de Tango dans ces conditions d'évaluation sont rapportés en figure 6.5.

La dégradation des performances liées au décalage d'horloge est très limitée lorsque les masques sont oracles. On observe certes une diminution du $\Delta\text{SIR}_{\text{img}}$ lorsque les décalages d'horloge sont supérieurs à une trame (16 ms), mais cette diminution est faible, le $\Delta\text{SIR}_{\text{img}}$ restant au-dessus de 25 dB quel que soit le décalage d'horloge d'échantillonnage. La diminution des performances en termes de SAR_{img} et STOI_{img} est quant à elle à peine significative.

6.3.1.2. Cas d'une dérive d'horloge

La dérive d'horloge est étudiée en ré-échantillonnant l'ensemble d'évaluation à différentes FE. Six conditions d'évaluation sont comparées :

Cas 1 : Les signaux sont ré-échantillonnés à la FE de 16 kHz. Bien que cela ne devrait pas les modifier, on a vu que le ré-échantillonnage d'un signal à sa propre FE impactait les métriques.

Cas 2 : Les signaux sont ré-échantillonnés avec une dérive ϵ prise de manière uniformément aléatoire entre 0 ppm et 5 ppm : $\epsilon \in [0; 5]$ ppm.

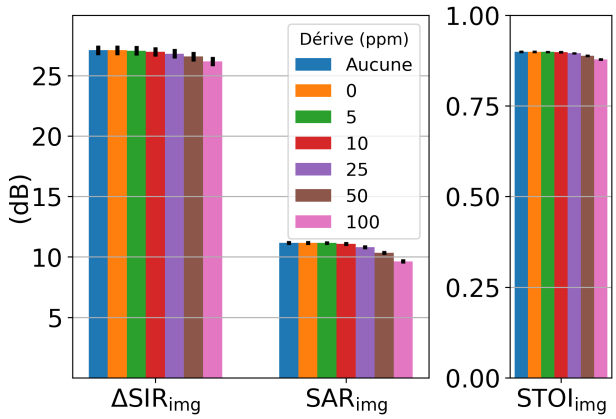


FIGURE 6.6. – Résultats de rehaussement de la parole de Tango avec des masques TF oracles lorsque les signaux de nœuds différents de l'AAAH ont une dérive variant de -100 ppm à 100 ppm.

Cas 3 : Les signaux sont ré-échantillonnés avec une dérive $\epsilon \in [0; 10]$ ppm.

Cas 4 : Les signaux sont ré-échantillonnés avec une dérive $\epsilon \in [0; 25]$ ppm.

Cas 5 : Les signaux sont ré-échantillonnés avec une dérive $\epsilon \in [0; 50]$ ppm.

Cas 6 : Les signaux sont ré-échantillonnés avec une dérive $\epsilon \in [0; 100]$ ppm.

Les résultats obtenus avec des masques TF oracles dans chacune de ces configurations d'évaluation sont représentés en figure 6.6. Nous y portons également les résultats obtenus avec les signaux non ré-échantillonnés, afin d'estimer si le ré-échantillonnage avec une dérive de 0 ppm impacte également les métriques à la suite du rehaussement de la parole. Le comportement de Tango avec des masques TF oracles en présence d'une dérive d'horloge est sensiblement le même qu'en présence d'un décalage d'horloge. Il semble toutefois que la dérive impacte le SAR_{img} plus que le $\Delta\text{SIR}_{\text{img}}$. Néanmoins, là encore, la baisse de performances est contenue, puisque le SAR_{img} diminue d'à peine 1,5 dB lorsque la dérive vaut 100 ppm. Enfin, notons que le ré-échantillonnage à la même FE (pour une dérive de 0 ppm) n'impacte pas la formation de voies, puisque les différences entre les barres bleue et orange (aucune dérive et une dérive de 0 ppm) ne sont jamais significatives. Ceci permet de nuancer les résultats préliminaires observés en section 6.2.

En conclusion de cette section, nous pouvons dire que l'impact de l'asynchronisation des signaux sur la formation de voies avec des masques TF oracles est, sinon négligeable, du moins très limité. Le décalage d'horloge ne conduit qu'à une baisse de moins de 2 dB du $\Delta\text{SIR}_{\text{img}}$ et la dérive d'horloge à une baisse de 1,5 dB du SAR_{img}. Nous proposons maintenant d'évaluer les performances de Tango lorsque les masques TF sont prédits par des RN ayant en entrée des signaux asynchronisés.

6.3.2. Impact sur les réseaux de neurones

Dans la suite des expériences, les masques TF nécessaires au calcul des formateurs de voies sont estimés par des RN aux deux étapes de filtrage de Tango. Etant donné que tous les signaux d'un même nœud sont synchrones, le filtrage local à la première étape de Tango n'est pas influencé par l'asynchronisation des signaux, qui n'impacte que le filtrage à la seconde étape de Tango. Nous étudions dans cette partie du chapitre l'impact de cette asynchronisation sur la prédiction des masques TF par les RNMuN, et donc sur les performances globales de Tango.

Les RN utilisés pour prédire les masques TF sont les mêmes que dans les sections 3.6, 4.3 ou 4.4 ; il s'agit des RN convolutionnels récurrents (CRNN) entraînés avec le corpus d'entraînement

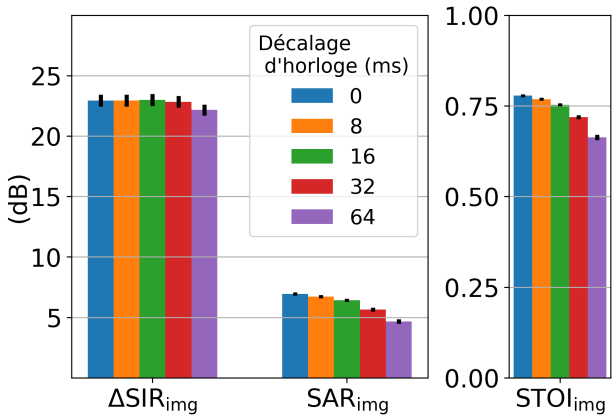


FIGURE 6.7. – Résultats de rehaussement de la parole de Tango avec des masques TF prédits par des RN lorsque les signaux de nœuds différents de l'AAAH sont décalés d'un nombre variable d'échantillons.

de la configuration *aléatoire* de DISCO. Les signaux compressés utilisés par le RNMuN sont les estimations de la parole uniquement car les expériences de la section précédente ont montré qu'il était difficile de sélectionner le signal le plus utile entre l'estimation compressée du bruit et celle de la parole. Afin de se concentrer sur la problématique de l'asynchronisation des signaux et de limiter la consommation en bande passante, l'estimation compressée du bruit ne sera plus utilisée pour prédire les masques TF.

6.3.2.1. Cas d'un décalage d'horloge

Notre système Tango est évalué dans les mêmes conditions que dans la section 6.3.1.1, en présence d'un décalage d'horloge d'importance variable. Les résultats sont représentés en figure 6.7.

Les performances de Tango avec des masques TF prédits plutôt qu'oracles sont plus sensibles au décalage d'horloge. Si le $\Delta\text{SIR}_{\text{img}}$ est très stable lorsque ce décalage augmente, ne diminuant jamais de manière significative, le SAR_{img} et le STOI_{img} perdent respectivement 33% et 15% de leur valeur lorsque le décalage d'horloge maximal passe de 0 ms à 64 ms. Au vu des résultats de la section précédente, cette baisse de performance est à attribuer aux RN puisque les formateurs de voies sont peu impactés par l'asynchronisation. Il reste difficile de déterminer si les masques TF prédits par les RN sont moins précis à cause du mauvais alignement des signaux à leur entrée, ou s'ils sont simplement décalés dans le temps sans pour autant être moins précis. En particulier, étant donné que le réseau de neurones mono-nœud (RNMoN) à la première étape ne dépend que du mélange local, le décalage ne l'impacte pas, puisqu'il n'est pas perceptible sur un seul canal. C'est donc le RNMuN qui est moins performant lorsque les quatre canaux à son entrée ne sont pas alignés dans le temps. Le SAR_{img} et le STOI_{img} caractérisant surtout la distorsion de la parole et son intelligibilité, leur baisse montre que le RNMuN a tendance à prédire des valeurs de masques plus faibles que lorsque les signaux ne sont pas décalés, ce qui ne diminue pas la réduction de bruit, mais ce qui réduit la parole et augmente sa distorsion.

Notons toutefois que pour des décalages modérés, de l'ordre d'une trame TFCT (16 ms), les performances restent bonnes quelle que soit la métrique considérée.

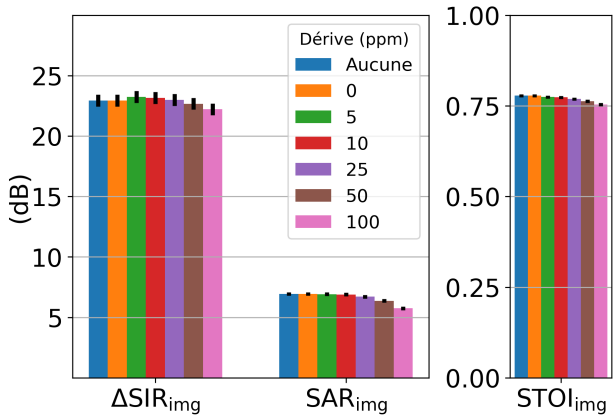


FIGURE 6.8. – Résultats de rehaussement de la parole de Tango avec des masques TF prédits par des RN lorsque les signaux de nœuds différents de l'AAAH subissent une dérive d'horloge.

6.3.2.2. Cas d'une dérive d'horloge

Le système Tango est évalué dans les mêmes conditions que dans la section 6.3.1.2, en présence d'une dérive d'horloge variant de 0 ppm à 100 ppm. Les résultats sont représentés en figure 6.8. Les performances de Tango avec des masques prédits sont plus faibles qu'avec des masques oracles, mais la baisse des performances liées à la dérive d'horloge suit en tous points celles de la formation de voies observée sur la figure 6.6. D'après cette expérience, les RNMuN n'apportent pas de dégradation supplémentaire des performances. L'explication la plus probable pour cela est que les signaux, de 10 s, sont trop courts pour que la dérive entraîne un décalage suffisamment important des signaux et pour que cela se traduise par des masques de moins bonne qualité. Néanmoins, les valeurs élevées de dérive (50 ppm, 100 ppm) permettent d'« accélérer » le décalage entre les signaux, et même pour ces valeurs, la baisse de performances n'est pas amplifiée par l'utilisation des RN. Cette résilience est sans doute aussi à attribuer au traitement effectué dans le domaine de la TFCT, où l'asynchronisation se traduit surtout par des changements de la phase. Les RN ne prédisant les masques qu'à partir de l'amplitude des signaux, leurs performances sont moins sensibles à l'asynchronisation. Cela veut dire qu'une resynchronisation des signaux, même grossière, à des intervalles espacés de plusieurs dizaines de secondes permettent de limiter l'impact d'une dérive d'horloge sur les performances de rehaussement de Tango.

Dans nos conditions d'expérimentation, étant donné que la dérive d'horloge a un impact très faible sur les performances de Tango, nous ne considérerons dans la suite que des décalages d'horloge comme formes d'asynchronisation. Cela peut paraître réducteur, mais notons que sur des intervalles de temps suffisamment courts, une dérive d'horloge ne se traduit pas autrement que par un décalage (constant) des signaux. Par exemple, une dérive d'horloge de 20 ppm à 16 kHz introduit un échantillon supplémentaire dans le signal toutes les 3 s. Observer ce signal entre la 3ème et la 6ème seconde revient donc à observer un signal de 3 s décalé d'un échantillon. On peut donc compenser la dérive d'horloge en compensant le décalage d'horloge sur des intervalles de temps assez courts. Les fenêtres fournies aux RN ne durant que 336 ms, l'hypothèse des intervalles courts semble valide.

Dans la section suivante, nous proposons une solution pour compenser la baisse de performances intervenant suite à un décalage d'horloge.

6.4. Solution proposée pour pallier l'asynchronisation des horloges

Nous proposons une solution qui ne nécessite pas de déterminer le décalage des horloges et de réaligner les signaux. Nous avons en effet vu que l'asynchronisation avait un impact limité sur la formation de voies, si bien que la baisse des performances observée en section 6.3.2 peut être essentiellement contenue par un travail sur les RN uniquement. Etant donné que le décalage d'horloge n'est perceptible que lorsque plusieurs canaux venant de nœuds différents sont vus par le RN, le RNMoN de la première étape de filtrage de Tango n'est pas modifié, et le modèle utilisé dans la suite des expériences est le même que dans les sections précédentes (par exemple dans toutes les expériences du chapitre 4). Seul le RNMuN sera donc modifié. Nous proposons deux méthodes. La première est d'entraîner le RNMuN dans des conditions similaires aux conditions d'évaluation, à savoir de l'entraîner avec des canaux non synchronisés. La seconde méthode est d'utiliser un mécanisme d'attention temporel afin d'implicitement réaligner les signaux pour faciliter la tâche du RNMuN.

6.4.1. Entraînement du réseau de neurones multinœud dans des conditions similaires aux conditions d'évaluation

Nous proposons dans un premier temps d'entraîner le RNMuN avec les signaux asynchronisés. Pour cela, les signaux d'un même nœud dans le jeu d'entraînement *aléatoire* du corpus DISCO sont décalés d'un nombre d'échantillons compris aléatoirement entre 0 ms et 32 ms. Ceci est appliqué à tous les nœuds de toutes les configurations d'entraînement et de validation. L'architecture du CRNN reste la même que celles étudiées auparavant, et les signaux compressés utilisés par le RNMuN sont les estimations compressées de la parole. Les résultats de rehaussement de la parole obtenus avec ces RN pour différentes valeurs maximales de décalages d'horloge sont représentés en figure 6.9. Les résultats obtenus avec le RNMuN entraîné sur les données sans décalage d'horloge y sont également rapportés afin de faciliter la comparaison avec les résultats de la section précédente.

Comme déjà présenté en section 5.3.2, entraîner le RN dans des conditions d'entraînement qui correspondent aux conditions d'évaluation permet d'améliorer les performances sur le jeu d'évaluation. Dans ce cas de figure, utiliser le RNMuN entraîné dans des conditions d'entraînement et de validation similaires ne permet pas d'obtenir un rehaussement de la parole entièrement résilient à l'asynchronisation : les performances lorsque les signaux sont décalés au maximum de 2 et 4 trames (32 ms et 64 ms respectivement) sont plus faibles que lorsque le décalage d'horloge est inférieur ou égal à une trame (0 ms, 1 ms, 16 ms). Cependant, cela permet de largement dépasser les performances obtenues avec le RNMuN qui n'est pas entraîné dans des conditions similaires aux conditions d'évaluation, en particulier pour des décalages d'horloge élevés. La différence entre les deux méthodes est significative en termes de SAR_{img} et de STOI_{img} dès que le décalage d'horloge dépasse 16 ms, et il est intéressant de noter que le SAR_{img} augmente de plus d'1 dB lorsque le décalage d'horloge maximal entre les signaux est de 64 ms.

6.4.2. Utilisation d'un mécanisme d'attention temporelle

6.4.2.1. Méthode

De la même manière que dans le chapitre précédent, nous proposons d'utiliser un mécanisme d'attention afin d'améliorer les performances de Tango avec les RN et de le rendre plus résilient à un décalage d'horloge d'échantillonnage. Cependant, dans cette partie, nous ne chercherons pas à distinguer la pertinence relative des canaux entre eux. Nous ne choisissons donc pas de mécanisme

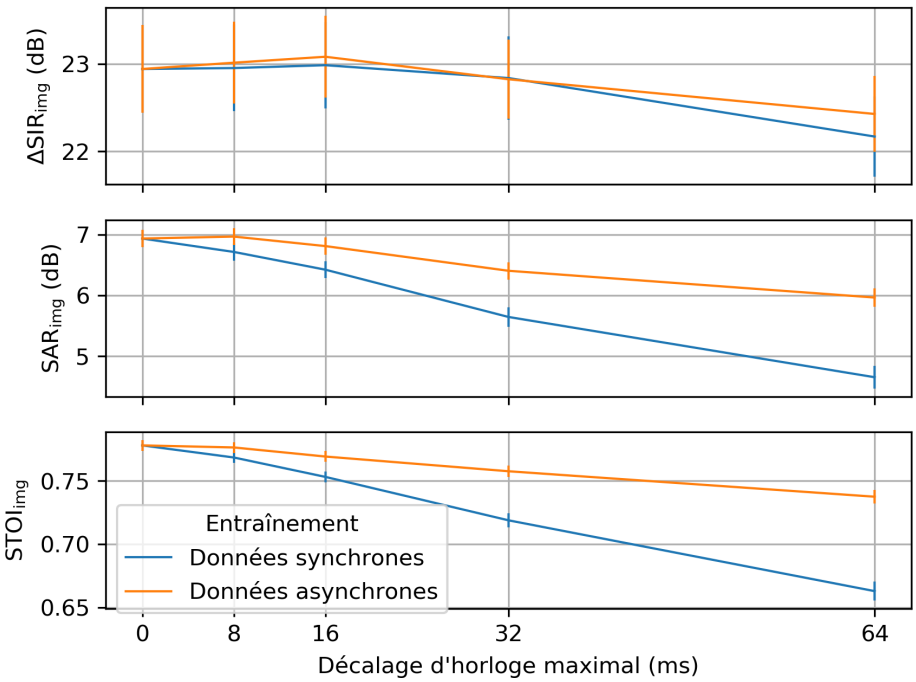


FIGURE 6.9. – Résultats de rehaussement de la parole de Tango lorsque les signaux de l'AAAH sont décalés d'un nombre variable d'échantillons. Les masques TF sont prédits par deux RNMuN entraînés dans des conditions différentes.

d'attention qui pondère l'entrée par des poids associés à la pertinence des canaux. Nous choisissons plutôt un mécanisme qui pourrait réaligner implicitement les canaux afin de rendre plus évidente la correspondance entre les canaux. De tels mécanismes d'attention temporelle ont déjà été proposés dans des travaux précédents. Nous pensons en particulier aux travaux de Schulze-Forster et al. (2021), qui font usage d'un mécanisme d'alignement de séquences (Bahdanau et al., 2014; Luong et al., 2015) dans le contexte de la séparation de voix chantées.

Soient $\mathbf{G}$ et $\mathbf{H}$ deux matrices de dimensions $T \times F$. Soit $\mathbf{W}$ une matrice de poids apprenable de dimensions $F \times F$. L'entraînement du module d'attention consiste à apprendre des scores de correspondance $\tilde{\mathbf{S}} \in \mathbb{R}^{T \times T}$ entre $\mathbf{g}_i$ et $\mathbf{h}_j$, respectivement les $i^{\text{ème}}$ et $j^{\text{ème}}$ colonnes de $\mathbf{G}$ et $\mathbf{H}$:

$$\tilde{s}_{j,i} = \mathbf{g}_i^T \mathbf{W} \mathbf{h}_j \,, \tag{6.8}$$

où $\tilde{s}_{j,i}$ est l'élément de $\tilde{\mathbf{S}}$ déterminé par les indices i et j. Une opération de $softmax$, identique à l'équation (5.6), est appliquée sur les lignes de $\tilde{\mathbf{S}}$ afin d'obtenir la matrice dite de correspondance $\mathbf{S}$:

$$\mathbf{S} = softmax(\tilde{\mathbf{S}}) \tag{6.9}$$

L'idée de ce mécanisme est que $\mathbf{S}$ doit contenir la probabilité que la colonne i de $\mathbf{G}$ soit alignée avec la colonne j de $\mathbf{H}$. Ces probabilités sont multipliées avec les colonnes d'entrée de $\mathbf{G}$, si bien que la $j^{\text{ème}}$ colonne du tenseur de sortie du mécanisme est :

$$\mathbf{c}_j = \sum_{i=1}^{T} s_{j,i} \mathbf{g}_i \,. \tag{6.10}$$

La matrice de sortie $\mathbf{C}$ est de dimensions $T \times F$. Elle est concaténée avec la matrice d'entrée $\mathbf{H}$ sur la dernière dimension pour obtenir une nouvelle matrice de dimensions $T \times 2F$. Avec ces

notations, la matrice $\mathbf{G}$ est la matrice dite de référence, à laquelle on compare la matrice d'entrée $\mathbf{H}$.

Dans notre contexte, le tenseur à l'entrée du RNMuN est constitué de C canaux de dimensions $T \times F$. Nous considérerons le premier canal (qui est le canal du microphone de référence) comme la matrice de référence $\mathbf{G}$ du mécanisme d'attention. Les C canaux (canal de référence inclus) sont comparés individuellement à ce canal, comme si l'on cherchait à aligner temporellement tous les canaux avec le canal du microphone de référence. C canaux sont donc obtenus en sortie du mécanisme d'attention. Un nouveau RNMuN est donc entraîné dans des conditions similaires aux conditions d'évaluation de ce chapitre. Chaque canal d'entrée est concaténé avec sa sortie correspondante du mécanisme d'attention. Le tenseur résultant est donc de dimensions $C \times T \times 2F$. Bien que le premier canal n'ait pas réellement besoin d'être comparé à lui-même, il est conservé dans le tenseur de sortie du mécanisme d'attention afin que l'information du mélange local soit tout de même fournie au RN. En sortie du mécanisme, le tenseur passe par un CRNN identique aux CRNN multinœuds utilisés jusqu'alors à un détail près : puisque le tenseur d'entrée a une dimension doublée le long des fréquences, la taille du noyau de la dernière opération de regroupement par sélection du maximum (*MaxPool*, cf section 3.3.1) est doublée sur la dernière dimension afin que la couche récurrente du CRNN conserve les mêmes propriétés. Etant donné que les *MaxPool* sont des fonctions non apprenables, ce CRNN a le même nombre de paramètres que ceux précédemment utilisés, outre les $F \times F$ paramètres de la matrice $\mathbf{W}$. La figure 6.10 représente sous forme graphique le mécanisme d'attention temporelle utilisée dans ce chapitre.

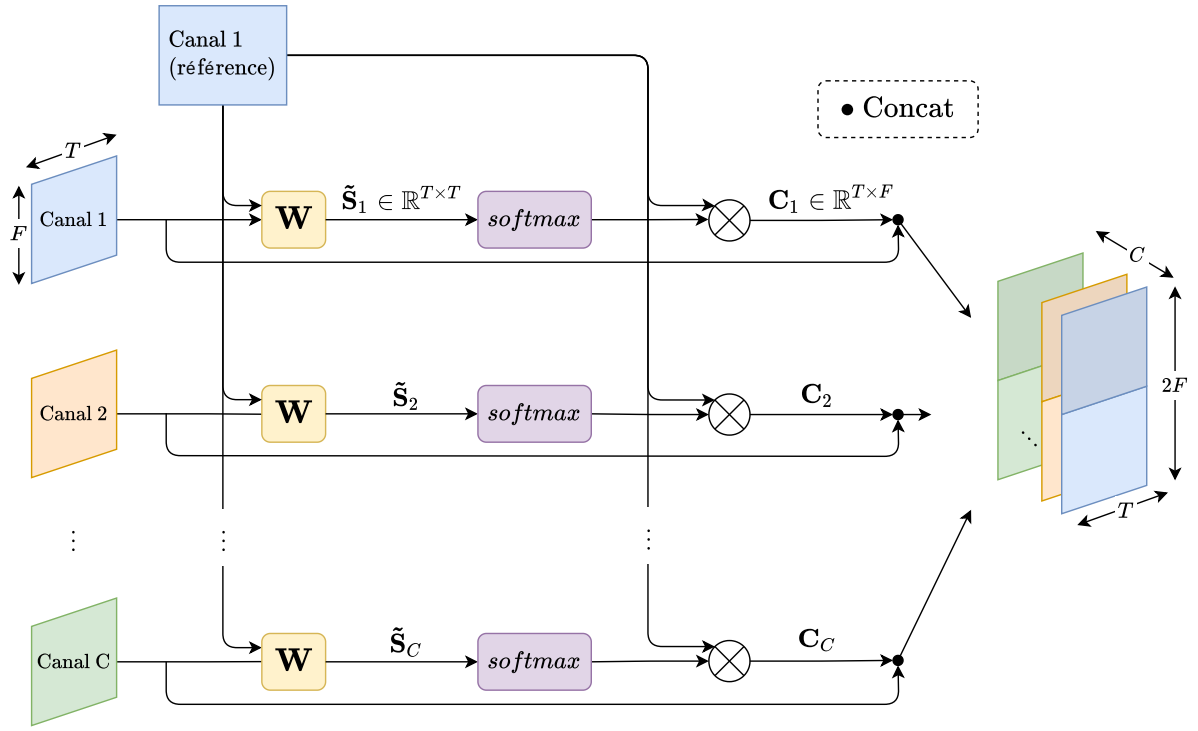


FIGURE 6.10. – Représentation graphique du mécanisme d'attention temporelle utilisée pour l'alignement temporel des canaux asynchrones.

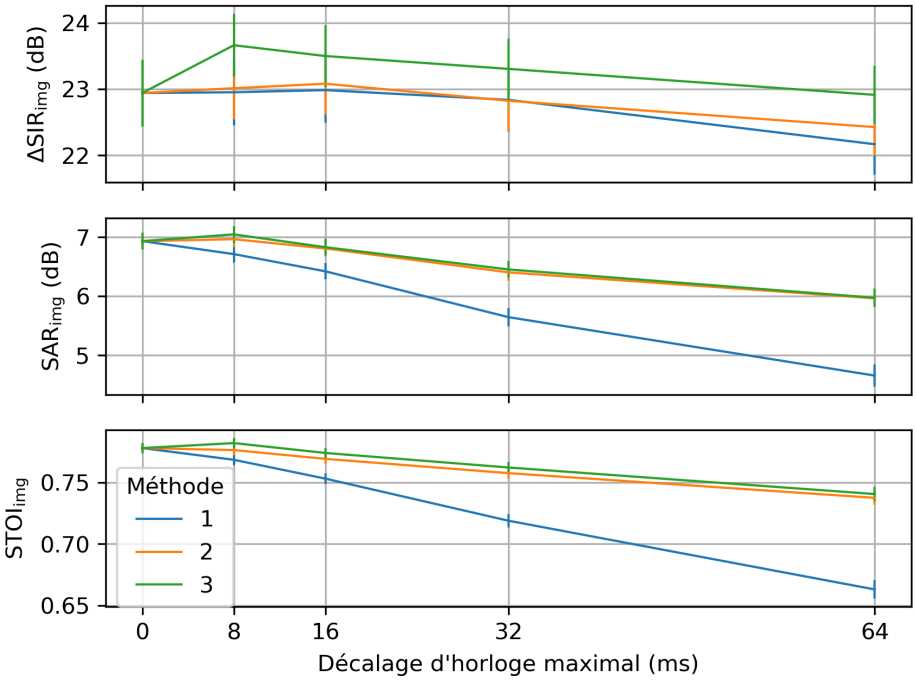


FIGURE 6.11. – Résultats de rehaussement de la parole de Tango lorsque les signaux de l'AAAH sont décalés d'un nombre variable d'échantillons. Les masques TF sont prédits par différents RNMuN. La méthode 1 fait référence au cas où le RNMuN est entraîné avec des données synchrones. La méthode 2 fait référence au cas où le RNMuN est entraîné avec des données asynchrones. La méthode 3 fait référence au cas où le RNMuN contient un mécanisme d'attention temporelle et est entraîné avec des données asynchrones.

6.4.2.2. Résultats

Les résultats obtenus avec ce nouveau RNMuN sont représentés en figure 6.11 où sont également rappelés les résultats obtenus avec les deux autres méthodes de la section précédente. Sur la figure 6.11, la méthode 1 fait référence au cas où le RNMuN n'est pas entraîné dans les conditions d'évaluation. La méthode 2 fait référence au cas où le RNMuN est entraîné dans les conditions d'évaluation. La méthode 3 fait référence au cas où le RNMuN est le réseau précédemment décrit, avec le mécanisme d'attention temporelle.

Les performances obtenues avec le nouveau RNMuN ne se distinguent guère des performances obtenues avec le CRNN entraîné dans les conditions similaires aux conditions d'évaluation, sauf en termes de $\Delta\text{SIR}_{\text{img}}$ où elles sont presque constamment supérieures d'un demi décibel, quoique de manière non significative. Il semble que ce mécanisme permette d'augmenter la réduction du bruit sans augmenter la distorsion de la parole ni diminuer l'intelligibilité de la parole.

Par ailleurs, ce mécanisme révèle un autre avantage, similaire à celui observé avec le mécanisme d'auto-attention en section 5.3.3. En effet, nous avons dit que la matrice $\mathbf{S}$ contenait en quelque sorte la probabilité que les colonnes de $\mathbf{G}$ soient alignées avec celles de $\mathbf{H}$. Dans le contexte de nos canaux, cela reviendrait à représenter la probabilité que les trames du canal d'un signal compressé soient alignées avec les trames du canal de référence.

Afin de vérifier ce phénomène, une salle du jeu d'évaluation est prise aléatoirement, et les signaux des nœuds 2, 3 et 4 sont décalés d'un nombre aléatoire d'échantillons pris entre -128 ms et 128 ms. Nous prenons une valeur absolue élevée de 128 ms et considérons des décalages négatifs afin de rendre plus évident le comportement du mécanisme d'attention. Sur chaque nœud, le RNMuN prédit le masque TF à partir d'un tenseur de quatre canaux. Quatre matrices $\mathbf{S}$ sont

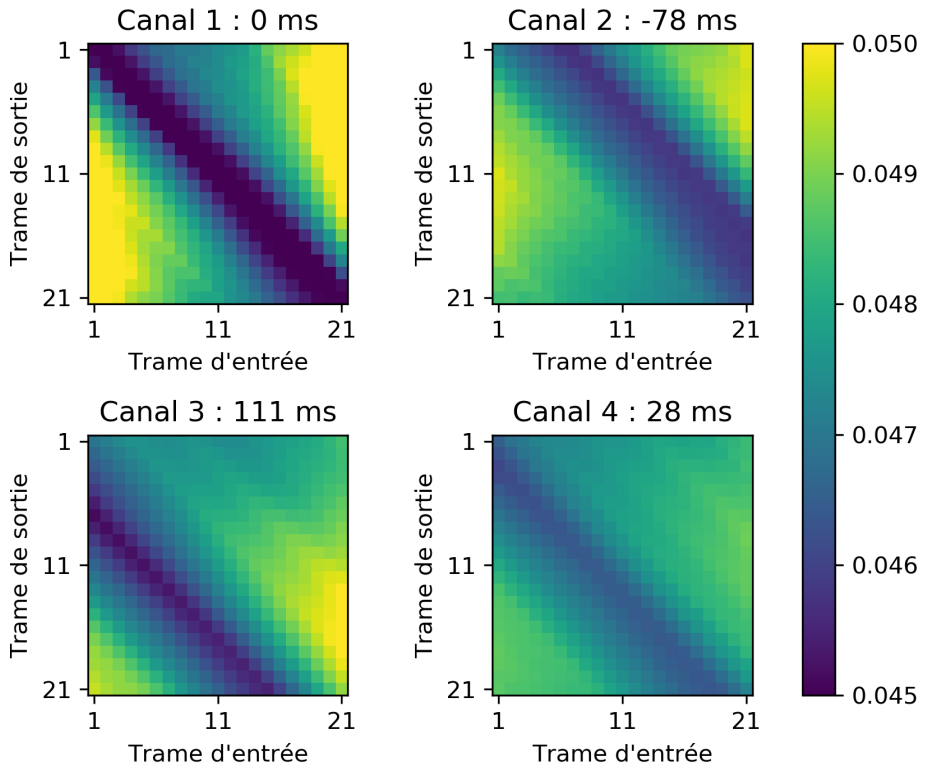


FIGURE 6.12. – Valeurs des matrices de correspondance $\mathbf{S}$ d'un mécanisme d'attention temporel d'un nœud recevant des signaux dont les horloges sont décalées dans le temps.

donc estimées sur chaque nœud. La figure 6.12 représente la moyenne sur le temps des matrices $\mathbf{S}$ estimées sur le nœud 1. Chaque sous-figure k représente la matrice de correspondance entre le canal k (qui est un signal compressé si $k > 1$) et le canal de référence ($k = 1$). Les titres de chaque sous-figure k indiquent également le temps de décalage entre le canal k et le canal de référence.

Il est intéressant de noter le caractère diagonal des matrices représentées. Par exemple au niveau du canal 2, les $5^{\text{ème}}$ et $6^{\text{ème}}$ diagonales supérieures contiennent des valeurs plus faibles que les autres diagonales de la matrice. Or le signal compressé de ce canal est décalé de -78 ms par rapport au canal de référence ; il devance donc le canal de référence d'environ 5 trames. La matrice $\mathbf{S}$ du canal 2 semble justement indiquer une corrélation entre la colonne $i + 5$ du canal de référence et la colonne i du canal 2 car les termes $\{s_{i,i+5}\}_{i=1..16}$ sont plus faibles que les autres termes de la matrice. Les termes de ces diagonales étant inférieurs au reste de la matrice, il conviendrait sans doute de parler de corrélation négative. De même, au niveau du canal 3, les $6^{\text{ème}}$ et $7^{\text{ème}}$ diagonales inférieures contiennent des valeurs plus faibles que les autres diagonales de la matrice. Le signal compressé du canal 3 est décalé de 111 ms par rapport au canal de référence ; il retarde donc d'environ 7 trames par rapport au canal de référence, et la matrice $\mathbf{S}$ semble bien indiquer une corrélation (négative) entre la colonne i du canal de référence et la colonne $i + 7$ du canal 3. La même réflexion peut expliquer les termes de la deuxième diagonale inférieure du quatrième canal, en retard d'environ deux trames par rapport au canal de référence. Quant au canal de référence, il est aligné avec lui-même, ce que traduit la diagonale de la première matrice représentée en figure 6.12.

Cette expérience illustre qu'il est possible, avec ce mécanisme d'attention, d'interpréter la valeur des poids. L'intuition initiale que les poids expriment la probabilité que les trames des

signaux compressés soient alignées avec celles du signal de référence n'est pas vraiment entérinée, car les valeurs des matrices semblent plutôt négativement corrélées à cette probabilité. Néanmoins, il est possible, par l'analyse des matrices, de déterminer l'ordre de grandeur du décalage d'horloge entre les signaux. C'est une information importante que le RNMuN livre en prédisant les masques, et qui a été apprise de manière non-supervisée : à aucun moment au cours de l'entraînement, la valeur des décalages d'horloge entre les signaux n'a été fournie au réseau. Il est donc intéressant de constater que le réseau a pu retrouver cette information cachée et qu'il semble l'utiliser pour améliorer ses performances.

Terminons toutefois sur une réserve : l'intervalle des valeurs des matrices $\mathbf{S}$ est de faible amplitude et il est possible que les matrices de correspondance n'aient qu'un impact très limité sur les performances du RNMuN. Les résultats supérieurs de ce dernier pourraient s'expliquer de la même manière qu'en section 5.3.2 : il est possible que la présence d'un module supplémentaire dans l'architecture du RNMuN permette au modèle dans son ensemble d'être plus performant, par exemple par un meilleur apprentissage, indépendamment des valeurs de sortie de ce module à l'inférence.

6.5. Conclusion

Deux types d'asynchronisation ont été considérés dans ce chapitre, le décalage d'horloge et la dérive d'horloge. Leur impact sur les performances de Tango a été quantifié. Les effets de l'asynchronisation sur la formation de voies seule, et sur la prédiction des masques par les RN de Tango ont été dissociés. La conclusion de cette étude est que la dérive d'horloge a très peu d'impact sur Tango mais que le décalage d'horloge est plus néfaste aux performances de Tango. Le décalage d'horloge a plus d'impact sur la prédiction de masques par des RN que sur la formation de voies. Les performances des RNMuN en présence de décalages d'horloge ont pu être améliorées par l'entraînement des RNMuN dans des conditions similaires aux conditions d'évaluation, et par l'introduction d'un mécanisme d'attention temporelle. Ce mécanisme d'attention semble en outre présenter une corrélation entre les valeurs des matrices qui le composent et la valeur du décalage d'horloge entre les nœuds de l'AAAH. Cela permet de retrouver un ordre de grandeur du décalage d'horloge entre les canaux d'un nœud. Les conclusions et résultats principaux sont rassemblés dans le tableau 6.2.

Au vu de ces résultats, il serait intéressant de modifier le mécanisme d'attention utilisé afin de mieux exploiter l'information des matrices de correspondance qu'il fait intervenir. On pourrait par exemple concaténer les sorties du mécanisme d'attention non pas sur l'axe des fréquences, mais sur celui des canaux, afin de mieux mettre en évidence l'alignement des canaux en entrée des couches convolutives. Par ailleurs, il est probable que les effets de l'asynchronisation sur les performances de Tango en général, et sur celles des RN qui y opèrent en particulier, soient limités par le fait que les opérations sont effectuées dans le domaine de la TFCT. Des RN considérant la forme d'onde des signaux ou intégrant l'information de la phase, seraient probablement plus sensibles aux asynchronisations. Il serait intéressant d'étudier le compromis entre les performances des RN et leur résilience aux asynchronisations.

Section	Points-clé
6.2	• L'asynchronisation a un fort impact sur les métriques. • Il convient de considérer les signaux convolués pour calculer les métriques.
6.3.1	• Le décalage d'horloge a un faible impact sur la formation de voies. • La dérive d'horloge a un faible impact sur la formation de voies.
6.3.2	Le décalage d'horloge impacte plus fortement la distorsion de la parole lorsque des RN servent à prédire les masques TF.
6.3.2.2	La dérive d'horloge a un très faible impact sur la prédiction des masques TF par les RNMuN.
6.4.1	Entraîner le RNMuN de la seconde étape de Tango avec des décalages d'horloge le rend plus résilient à ce type d'asynchronisation.
6.4.2	• Utiliser un mécanisme d'attention temporelle permet de mieux réduire le bruit sans augmenter la distorsion de la parole. • Les matrices de correspondance du mécanisme d'attention, bien que de faibles valeurs relatives, sont interprétables et permettent de retrouver un ordre de grandeur du décalage d'horloge entre les canaux.

TABLEAU 6.2. – Points-clé à retenir du chapitre 6.

7. Exploitation de l'information spatiale enregistrée par une antenne acoustique ad-hoc pour la séparation de sources

Dans ce chapitre, nous montrons que le fonctionnement de Tango peut servir à séparer des sources de parole interférentes. Nous montrons que ce système distribué exploite efficacement l'information spatiale enregistrée par tous les nœuds de l'antenne afin de fournir au niveau de chaque nœud d'une antenne acoustique ad-hoc (AAAH) une estimation de chaque source active. Après avoir montré l'intérêt de Tango pour la séparation de sources dans une situation typique de réunion, nous étudions sa résilience à un nombre variable de sources.

7.1. Présentation du contexte

La séparation de sources de parole constitue un problème difficile du fait que les signaux interférents sont de nature similaire et se superposent plus dans le plan temps-fréquence (TF). Les premières solutions de séparation de sources de parole se sont basées sur des principes d'analyse de sources auditives (Bregman, 1994). Par exemple, dans le plan TF, les représentations du signal d'une même source peuvent être regroupées avant d'être transformées dans le domaine temporel (Hu and Wang, 2012). Une autre solution est de supposer que la représentation dans le plan TF de la parole est une matrice creuse et que chaque point TF n'est dominé que par une seule source. Masquer le mélange dans le plan TF par un masque associé à chaque source permet de dissocier les différentes sources (Yilmaz and Rickard, 2004). La décomposition du mélange par la factorisation en matrices positives (Lee and Seung, 1999) peut également servir à séparer différentes sources (Févotte et al., 2009; Ozerov and Févotte, 2009), mais montre de meilleures performances pour la séparation de sources musicales que pour la séparation de sources de parole.

Les récentes avancées dans le domaine de la séparation de sources ont été réalisées à l'aide de réseaux de neurones (RN). Leur forte capacité de modélisation, couplée à la logique de regroupement (Hershey et al., 2016; Liu and Wang, 2019; Zeghidour and Grangier, 2020), de masquage TF (Luo and Mesgarani, 2019; Subakan et al., 2021) ou de factorisation en matrices positives (Le Roux et al., 2015), a permis de grandes améliorations. Néanmoins, la tendance générale indique que les meilleures performances sont obtenues avec des RN comportant plusieurs millions de paramètres (Subakan et al., 2021). Ces réseaux complexes sont sensibles à l'hyper-paramétrisation, gourmands en énergie et en puissance de calcul, donc inutilisables dans les appareils d'une AAAH. Dans le contexte des AAAH, il convient de trouver une solution qui, si elle fait intervenir des RN, utilise des RN les plus simples possibles.

Dans cette section, nous nous concentrons sur un exemple typique de mélange de parole, à savoir une réunion où N personnes, assises autour d'une table circulaire, parlent en même temps. Comme souvent en réunion, chaque personne a un appareil devant elle, par exemple son téléphone ou son ordinateur. Une telle situation est illustrée avec $N = 3$ en figure 7.1.

A partir de ces connaissances, nous proposons d'appliquer Tango pour séparer les sources de parole, avec la même architecture de RN qu'utilisée jusqu'alors. Comme vu en section 3.6, Tango

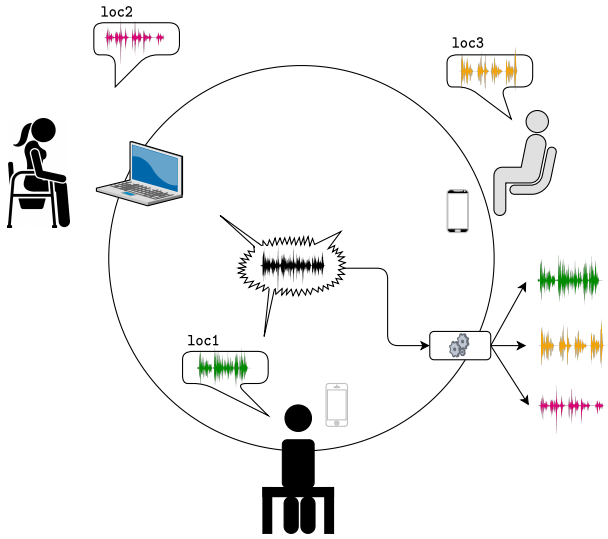


FIGURE 7.1. – Représentation du contexte du chapitre 7

est une solution peu coûteuse en puissance de calcul comparée aux alternatives de l'état de l'art, ce qui la rend attractive pour les applications similaires à la situation de réunion considérée ici. De par la connaissance de la configuration spatiale, l'information véhiculée entre les nœuds au cours des deux étapes de filtrage peut être efficacement exploitée, ce que nous décrivons dans la section suivante.

7.2. Exploitation de l'information a priori

En plus de connaître la configuration spatiale décrite précédemment, nous supposons que toutes les sources ont approximativement la même puissance moyenne. Ainsi, chaque nœud, plus proche de la source devant laquelle il est placé, enregistre un mélange dont une source est dominante [1]. Cette source est la source devant laquelle il est placé, et sera pour chaque nœud la source à estimer. Par exemple, dans la figure 7.1, l'ordinateur est chargé de restituer la parole de la locutrice 2. L'utilisation de Tango dans le contexte de cette section est représenté en figure 7.2. Deux conséquences découlent de cette configuration.

La première conséquence est qu'à l'échelle de chaque nœud, le problème de séparation de sources est transformé en problème de rehaussement de la parole : la source dominante est la source cible du nœud et la somme des sources interférentes constitue le bruit. A chaque étape de filtrage de Tango, on peut donc déterminer le masque oracle nécessaire au calcul du formateur de voies du nœud k par l'équation suivante :

$$m_k = \frac{|x_{k,1}|}{|x_{k,1}| + |n_{k,1}|} \tag{7.1}$$

où $x_{k,1}$ est la $k^{\text{ème}}$ source (dominante au nœud k) enregistrée par le microphone de référence et $n_{k,1} = \sum_{j \neq k} x_{j,1}$ est le bruit, à savoir la somme des autres sources enregistrées par le même microphone de référence. La séparation de N sources est ainsi distribuée de manière très concrète en N tâches de rehaussement de la parole, où la source k est estimée par l'appareil k.

1. La source est dominante dans le sens où son niveau sonore est supérieur à celui de chacune des autres sources prises séparément, mais elle ne domine pas forcément la somme des autres sources interférentes.

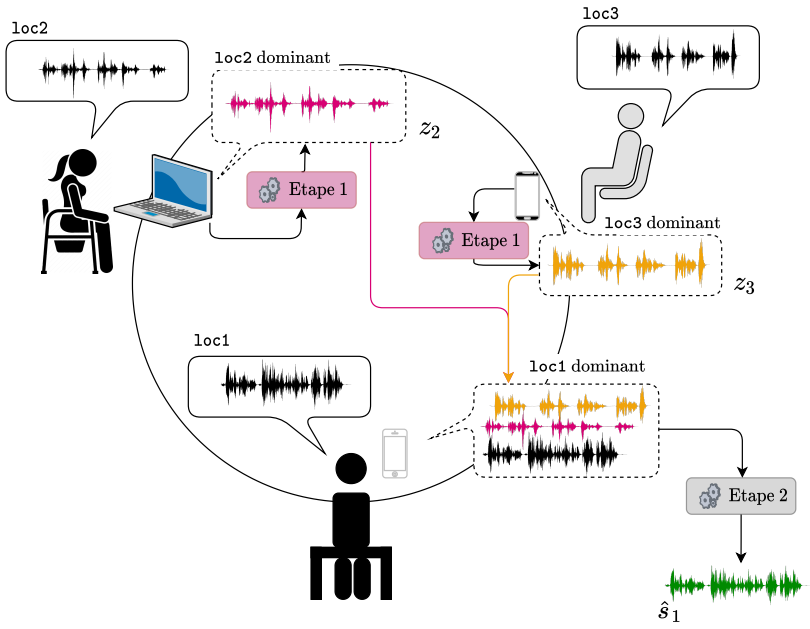


FIGURE 7.2. – Représentation schématique de Tango dans le contexte du chapitre 7. « `loc` » fait référence à « locuteur » ou « locutrice ».

La seconde conséquence est que si le nœud j estime la source dominante devant laquelle il est placé, le nœud k recevra à la seconde étape de filtrage autant de signaux compressés z_j qu'il y a de sources interférentes. A l'échelle d'un nœud k, les signaux compressés ne véhiculent plus l'estimation de la source cible, mais celle des sources interférentes. Comme vu dans la section 4.5.4, cela peut être avantageux lorsque les interférences sont assez faibles au niveau du nœud récepteur, comme dans notre cas de figure. Notons que cela se fait de manière automatique, sans que les RN n'aient besoin de savoir quelle source, parmi celles qu'ils observent, est la source cible. L'entraînement permet de les conditionner à toujours estimer la source dominante. C'est la connaissance *a priori* du scénario qui permet ainsi d'optimiser les échanges de signaux entre nœuds sans qu'on n'ait besoin de déterminer quel type de signal envoyer depuis quel nœud.

7.3. Corpus d'évaluation

Un corpus assez similaire au corpus *réunion* (section 3.4.2) a été simulé pour évaluer Tango dans le contexte de séparation de sources. Trois sous-corpus, contenant deux, trois et quatre sources chacun, sont créés à l'aide de la boîte à outils `Pyroomacoustics` (Scheibler et al., 2018). Chaque sous-corpus contient 10000 configurations d'entraînement, 1000 de validation et 1000 d'évaluation. Dans chacune des configurations, une salle rectangulaire est simulée. Sa longueur est prise aléatoirement entre 3 m et 7 m. Sa largeur est prise aléatoirement entre 3 m et 5 m. Sa hauteur est prise aléatoirement entre 2,5 m et 3 m. Une table circulaire est imaginée dans la pièce, mais sa réflexion n'est pas simulée. Son rayon est pris aléatoirement entre 0,3 m et 2,5 m. Les sources sont placées régulièrement autour de la table. Cela signifie que l'angle entre le centre de la table et deux sources adjacentes est le même pour toutes les sources adjacentes d'une configuration spatiale. Les sources sont placées à une distance comprise entre 0 cm et 50 cm du bord de la table et à une hauteur comprise entre 1,15 m et 1,80 m, comme si les personnes étaient assises ou debout près de la table. Elles sont placées à au moins 50 cm des murs de la salle. Les niveaux sonores des signaux sources (c'est-à-dire non convolués) sont égaux.

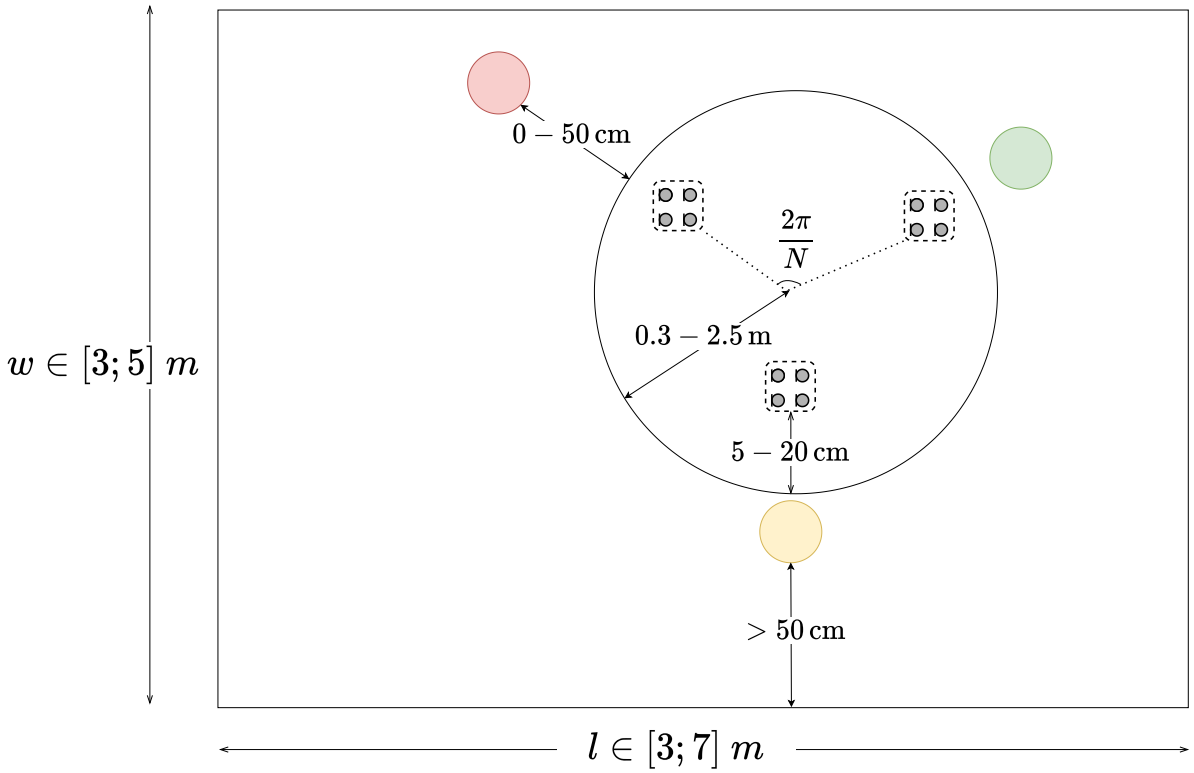


FIGURE 7.3. – Représentation graphique des caractéristiques principales des salles du corpus MEETIT. N est le nombre de sources et nœuds (ici égal à 3).

Autant de nœuds que de sources sont posés sur la table, en face d'une source, à une distance prise aléatoirement entre 5 cm et 20 cm. Chaque nœud comporte 4 microphones. Leur hauteur est égale dans une configuration donnée, prise aléatoirement entre 0,8 m et 0,9 m, comme s'ils étaient posés sur la même table. Le temps de réverbération est pris aléatoirement entre 150 ms et 400 ms. La figure 7.3 représente les caractéristiques principales des salles de ce corpus appelé MEETIT (MEETing InTerferences).

Puisque les sources ont la même puissance sonore et que les nœuds sont presque tous à la même distance de leur source correspondante, le SIR mesuré à chaque nœud dépend du rayon de la table. La figure 7.4 présente le SIR mesuré sur chaque nœud de chaque configuration de validation. La référence pour la source cible est la source réverbérée dominante ; la référence pour les interférences est la somme des autres sources réverbérées (ou de l'unique autre source quand $N = 2$). On observe sur la figure 7.4 que le rayon minimal de la table augmente lorsque N augmente. Cela est lié au fait que le code pour créer les pièces garantit une approximative équi-répartition des SIR. Cela empêche automatiquement la simulation de tables trop petites, sans quoi trop de signaux à faibles SIRs seraient créés. Ce phénomène un peu forcé est en fait tout-à-fait compatible avec la réalité, où il est peu courant que quatre personnes s'assoient autour d'une table de faible rayon [2].

Les signaux de parole des jeux d'entraînement, de validation et d'évaluation sont respectivement issus des sous-dossiers `train-clean-100`, `dev-clean` et `test-clean` de LibriSpeech (Panayotov et al., 2015). Pour un mélange, les différents signaux de parole se superposent entièrement dans le temps : toutes les sources sont actives tout au long du signal.

2. Le ministre de la Santé verrait d'ailleurs cela d'un très mauvais œil.

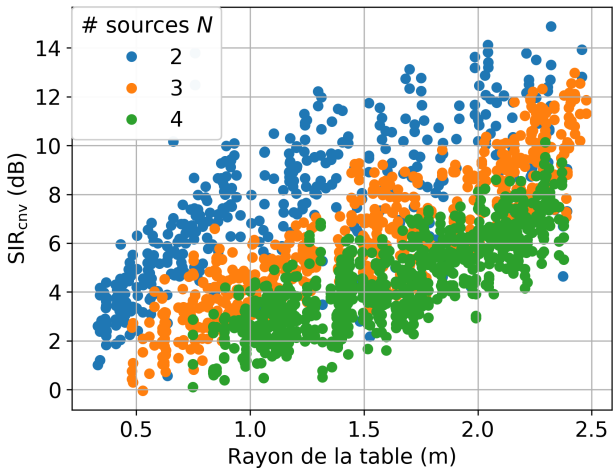


FIGURE 7.4. – Représentation des SIR mesurés à chaque nœud du jeu de validation en fonction du rayon de la table et du nombre de sources.

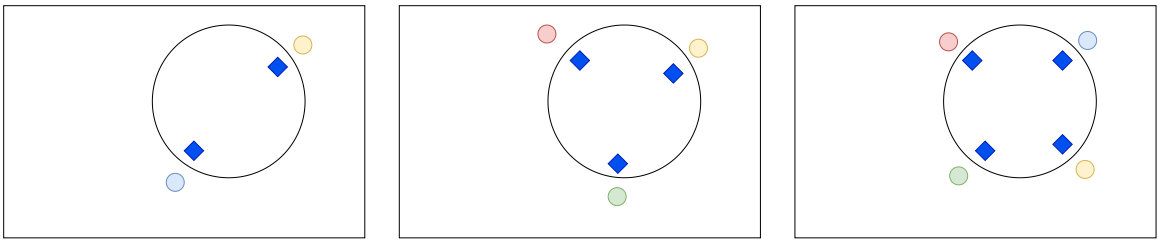
Les signaux sont tous échantillonnés à 16 kHz. La transformée de Fourier à court terme (TFCT) des signaux est calculée à l'aide de fenêtres de Hann de 32 ms se recouvrant sur 16 ms. 30 heures d'entraînement, 3 h de validation et 3 h d'évaluation ont ainsi été simulées. Le code Python pour créer ce corpus est disponible en ligne [3].

7.4. Résultats de séparation de sources avec Tango

Tango est utilisé pour séparer les N sources actives. Les mêmes architectures de RN convolutionnel récurrent (CRNN) que celles présentées en section 3.3 servent à prédire les masques TF nécessaires au calcul des filtres de Wiener multicanaux (MWFs). Trois types de CRNN multi-nœuds sont entraînés, un par nombre de sources. Cette section est divisée en trois parties. Dans la première partie, les résultats de Tango pour la séparation de sources sont présentés dans le cas où il y a autant de nœuds que de sources ; c'est le cas dit « équilibré ». Dans une seconde partie, des cas dits « surdéterminés » sont étudiés, et qui représentent les cas où sont présents plus de nœuds que de sources. Dans une dernière partie, des cas dits « sous-déterminés » sont étudiés, et qui représentent les cas où sont présents moins de nœuds que de sources.

Les métriques à considérer pour quantifier les résultats de séparation de sources sont un long sujet de discussion (Vincent et al., 2006; Le Roux et al., 2019; Drude et al., 2019). Dans cette section, nous utiliserons la métrique la plus couramment utilisée, à savoir le rapport signal-à-distortion à invariance d'échelle (SI-SDR : *scale-invariant signal to distortion ratio*) (Le Roux et al., 2019). Il rassemble en une valeur le rapport signal-à-interférences à invariance d'échelle (SI-SIR : *scale-invariant signal to interference ratio*) et le rapport signal-à-artéfacts à invariance d'échelle (SI-SAR : *scale-invariant signal to artifacts ratio*), ce qui lui confère l'avantage d'être concis. Contrairement aux métriques de SIR et de SAR vues jusqu'à maintenant, le SI-SIR, le SI-SAR et le SI-SDR n'autorisent pas qu'un filtre à réponse finie soit appliqué aux cibles pour calculer les différentes composantes des signaux estimés. Ces métriques n'autorisent qu'un changement d'échelle entre la source propre ciblée et la référence prise pour calculer les composantes des signaux estimés. Pour ne pas pénaliser les effets de la déréverbération, les signaux cibles pour calculer le SI-SDR sont les images convoluées des signaux. Un SI-SDR plus élevé correspond à une meilleure performance. Une description plus détaillée des SDR, SAR, SIR et leurs variantes

3. https://github.com/nfurnon/disco/tree/master/dataset_generation/gen_meetit



(a) Avec $N = 2$ sources (b) Avec $N = 3$ sources (c) Avec $N = 4$ sources

FIGURE 7.5. – Représentations 2D (vue du dessus) de configurations « équilibrées » du corpus.

à invariance d'échelle est proposée en Annexe A.

Les résultats de Tango sont comparés pour trois variantes. La première utilise les masques oracles pour calculer les formateurs de voies aux deux étapes de filtrage de Tango. Elle est notée « MRI ». Elle donne les meilleurs résultats que Tango puisse atteindre. La deuxième méthode est le système de référence, lorsque les masques sont prédits par des réseaux de neurones mononœuds (RNMoN) à la première étape et par des réseaux de neurones multinœuds (RNMuN) à la seconde étape de filtrage. Elle est notée « RN_N » car les RN à la seconde étape de filtrage disposent de N signaux pour prédire les masques. La troisième méthode est Tango lorsque les RNMoN sont utilisés aux deux étapes de filtrage. Elle est notée « RN_1 » car les RN à la seconde étape de filtrage ne disposent que du mélange du microphone de référence pour prédire les masques. La dernière méthode consiste simplement en la première étape de Tango. Puisqu'il s'agit en fait d'un MWF appliqué à tous les nœuds indépendamment les uns des autres, elle est notée « MWF ».

La différence entre les méthodes permet de juger si les différents résultats sont plutôt à attribuer à une moins bonne prédiction des masques TF ou à des formateurs de voies moins performants, malgré des masques TF bien prédits.

Comme toutes les sources doivent être estimées, et qu'une source est estimée à chaque nœud, les résultats représentés dans cette section correspondent à la moyenne sur les 1000 configurations d'évaluation et sur tous les nœuds de chaque configuration.

7.4.1. Cas équilibrés

Les cas équilibrés rassemblent les situations où il y a autant de sources que de nœuds. Ces cas sont schématisés pour $N = K = 2$, $N = K = 3$ et $N = K = 4$ en figure 7.5.

Les résultats de Tango sur le jeu d'évaluation sont présentés en figure 7.6.

Tout d'abord, bien que les méthodes MWF et RN_1 conduisent à des résultats proches, la différence entre ces deux méthodes croît lorsque le nombre de nœuds et de sources augmente, et aboutit à une performance significativement meilleure avec RN_1 qu'avec MWF. Cela montre que les signaux compressés convoient une information utile pour le filtrage, et que le formateur de voies, même avec des masques identiques aux deux étapes, est plus performant lorsque les signaux compressés sont utilisés. C'est un résultat intéressant car dans ces configurations, les signaux compressés reçus au nœud k ne sont pas les estimations du signal cible du nœud k. Malgré cela, ils permettent d'augmenter les résultats de séparation de source.

Une plus grande différence est cependant apportée lorsque les signaux compressés sont également utilisés par les RN pour prédire les masques TF. En effet, les performances obtenues avec RN_N dépassent d'entre 0,5 dB et 1,7 dB les performances obtenues avec RN_1 et MWF, et ce de manière significative quel que soit le nombre de sources et de nœuds. Cela montre que l'intérêt d'envoyer les signaux compressés réside surtout dans leur utilisation par les RN. Ici encore, c'est

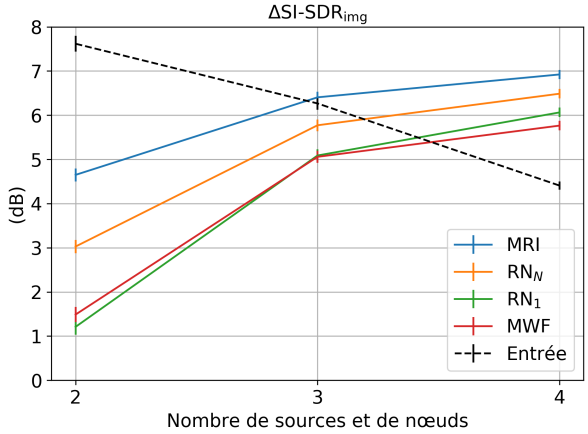
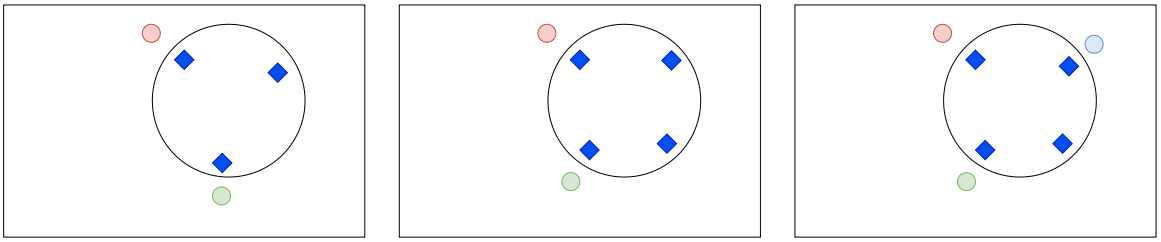


FIGURE 7.6. – Résultats de séparation de sources de Tango dans les cas équilibrés du jeu d'évaluation.



(a) Avec $N = 2$ sources et $K = 3$ nœuds (b) Avec $N = 2$ sources et $K = 4$ nœuds (c) Avec $N = 3$ sources et $K = 4$ nœuds

FIGURE 7.7. – Représentations 2D (vue du dessus) de configurations « surdéterminées » du corpus.

d'autant plus intéressant qu'il s'agit des estimations compressées des interférences pour le nœud récepteur. Ils apportent donc une information riche car non redondante et peu disponible au niveau du nœud récepteur.

Pour terminer, notons que le ΔSI-SDR$_{\text{img}}$ croît lorsque le nombre de sources et de nœuds augmente, et est inférieur de 0,5 dB seulement avec RN$_N$ qu'avec des masques oracles. Comme le montre le ΔSI-SDR$_{\text{img}}$ d'entrée (courbe pointillée), la tâche est plus difficile lorsque le nombre de sources et de nœuds augmente, mais la performance absolue reste constante, d'où la croissance du ΔSI-SDR$_{\text{img}}$.

7.4.2. Cas surdéterminés

Les cas surdéterminés représentent les situations où une personne quitte la salle au cours de la réunion. C'est une situation qui peut arriver fréquemment et qui a pour conséquence qu'il reste plus de nœuds dans la salle qu'il n'y a de sources. L'impact de ces situations est étudié dans cette section.

Puisque les mélanges du corpus sont obtenus en sommant (après réverbération) les signaux images, il est possible de reproduire de tels cas surdéterminés dans les salles comptant trois et quatre sources (donc trois et quatre nœuds), mais en sommant un nombre inférieur de sources qu'il n'y a de nœuds dans la pièce. Les trois cas possibles correspondants sont représentés en figure 7.7. Précisons deux choses à ce stade.

Tout d'abord, si la source k est celle qui a quitté la pièce, le masque oracle n'est plus défini pour le nœud k, puisqu'il n'y a plus de source dominante : le signal $x_{k,1}$ de l'équation (7.1)

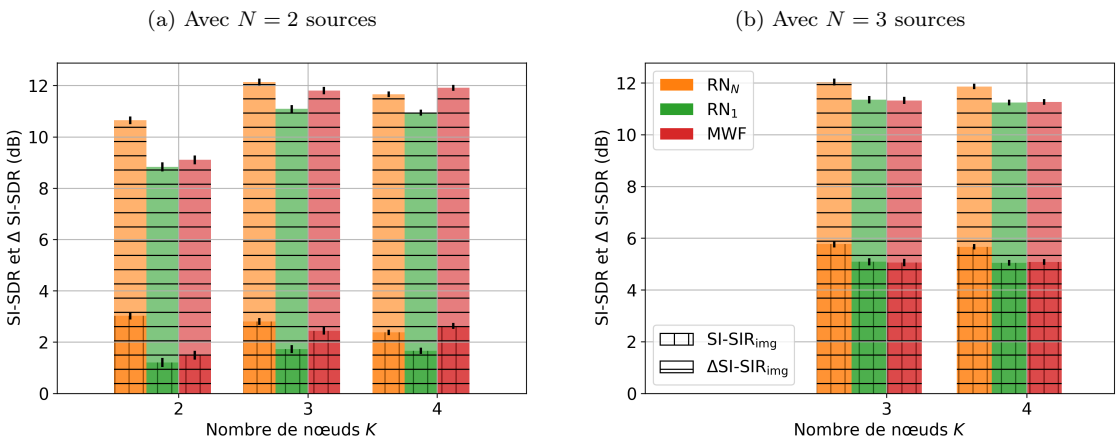


FIGURE 7.8. – Résultats de séparation de sources de Tango en conditions surdéterminées.

n'existe plus. A cause de cette ambiguïté, les résultats oracles ne seront pas rapportés dans les cas surdéterminés. Avec les RN, cette ambiguïté n'a pas d'influence, car les RN estiment les masques à partir des signaux en entrée. Le fait qu'il n'y ait pas de source dominante dans les mélanges locaux ne les empêche pas de prédire un masque TF.

Par ailleurs, les RN utilisés dans ces configurations ont été entraînés dans les cas équilibrés. Dans un cas à deux sources et trois nœuds par exemple, le RNMuN à la seconde étape de filtrage est le RNMuN entraîné avec *trois* sources et trois nœuds. Les RN n'ont pas été ré-entraînés dans des situations où plus de nœuds sont présents que de sources, si bien qu'ils ne sont pas habitués à ne pas avoir de source dominante dans les mélanges observés.

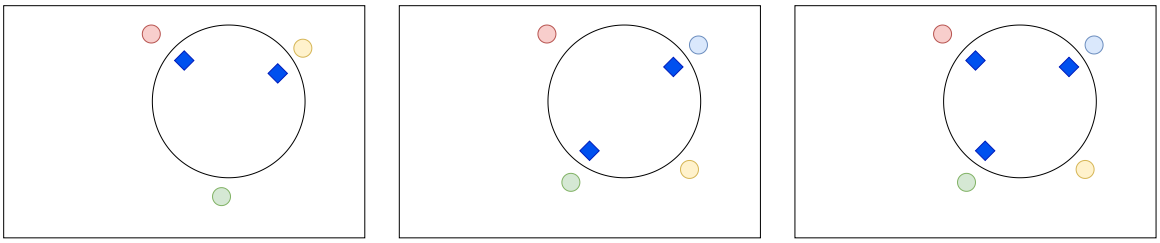
Les résultats de Tango sur chacun de ces cas surdéterminés sont représentés en figure 7.8, où les résultats des cas équilibrés avec $N = K = 2$ et $N = K = 3$ sont rappelés.

Dans ces configurations encore, les résultats obtenus avec les RNMuN sont supérieurs aux deux autre méthodes dans presque tous les cas de figure. Il est profitable à la séparation de source d'envoyer les signaux compressés entre les deux étapes de filtrage de Tango. Toutefois, cela semble profiter surtout à l'estimation des masques TF car la méthode MWF surpasse la méthode RN_1 avec $N = 2$ et l'égale avec $N = 3$.

Néanmoins, on peut observer avec $N = 2, K = 4$ que la méthode MWF égale les performances obtenues avec les RNMuN. De fait, nous avons remarqué que lorsqu'une source manque dans la pièce (disons la source k), le RNMoN placé sur le nœud k prédit un masque presque partout égal à 0 : comme aucune source ne domine, le RN estime qu'il ne voit que du bruit et prédit donc un masque presque nul. La conséquence est que le signal compressé envoyé par le nœud k est lui aussi presque nul, et vient d'une part réduire les performances du formateurs de voies, et d'autre part diminuer la performance des RNMuN des nœuds $j \neq k$. Ce phénomène est compensé lorsque suffisamment de signaux compressés non-nuls sont reçus ($N = 2, K = 3$ et $N = 3, K = 4$) mais prédomine lorsque trop de sources sont manquantes.

7.4.3. Cas sous-déterminés

Les cas sous-déterminés représentent les situations où plus de sources sont présentes que de nœuds. Cela peut arriver par exemple si un des appareils s'éteint par manque de batterie ou si une personne supplémentaire entre dans la salle. Cette fois-ci, tous les masques oracles sont bien définis, et les résultats correspondants sont également rapportés. Les configurations surdéterminées peuvent se retrouver dans les sous-corpus avec trois sources (en ne considérant que



(a) Avec $N = 3$ sources et $K = 2$ nœuds (b) Avec $N = 4$ sources et $K = 2$ nœuds (c) Avec $N = 4$ sources et $K = 3$ nœuds

FIGURE 7.9. – Représentations 2D (vue du dessus) de configurations « sous-déterminées » du corpus.

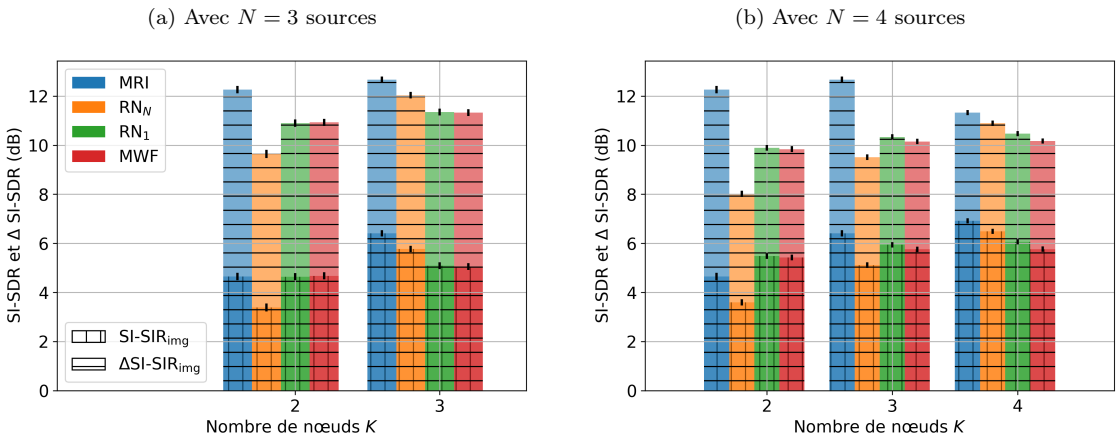


FIGURE 7.10. – Résultats de séparation de sources de Tango en conditions sous-déterminées.

deux des nœuds disponibles), et avec quatre sources (en ne considérant que deux ou trois des nœuds disponibles). Les nœuds considérés sont pris aléatoirement parmi ceux disponibles. Les RNMuN utilisés sont ceux entraînés dans les conditions équilibrées. Par exemple dans un cas avec deux nœuds et trois sources, le RNMuN est celui entraîné avec le sous-corpus de deux nœuds et *deux* sources. La figure 7.9 schématise une instance possible de chacune des configurations. Les résultats de Tango sont présentés en figure 7.10.

Dans ces conditions d'évaluation, les performances avec les RNMuN sont significativement moins bonnes que celles des méthodes RN_1 et MWF, et ce même s'il n'y a qu'une seule source de plus que de nœuds. Puisque la méthode RN_1 conduit à des résultats aussi bons, voire meilleurs, que la méthode MWF, cela signifie que la chute de performances avec la méthode RN_N est à attribuer à une mauvaise prédiction des masques TF par les RNMuN. Il semble que les RNMuN dépendent beaucoup des signaux compressés pour prédire les masques, et que lorsque toute l'information des interférences n'est pas apportée, le masque est mal prédit. Une solution pour pallier cette difficulté serait de ré-entraîner les RN dans des conditions sous-déterminées, afin de les aider à prédire les masques de la source dominante, même lorsque peu d'information relative aux interférences est disponible au travers des signaux compressés.

7.5. Conclusion

Dans cette section, nous avons montré que Tango pouvait être utilisé dans un contexte de séparation de sources et que la logique d'envoyer des signaux compressés pouvait être judicieusement exploitée pour faire transiter des estimations de toutes les sources au niveau de tous les nœuds, et ce par un algorithme distribué. Nous avons montré que cette information spatiale permettait d'obtenir des performance proches des performances oracles dans un contexte typique de réunion. Nous avons également analysé la résilience de Tango face à un nombre variable de sources, et si ce système est assez performant lorsque plus de nœuds sont présents que de sources, cela n'est plus vrai lorsqu'il y a moins de nœuds que de sources. Nos expériences ont montré la sensibilité des RN aux signaux compressés échangés entre les nœuds.

Il serait intéressant de combiner l'approche de ce chapitre pour la séparation de sources avec les solutions des chapitres précédents, basées sur l'introduction d'un mécanisme d'attention, afin d'augmenter la résilience des RN dans des conditions surdéterminées et sous-déterminées. Enfin, étant donné que la logique de Tango permet d'obtenir de bons résultats dans une grande variété de conditions acoustiques, tant pour le rehaussement de la parole que pour la séparation de sources, il serait intéressant d'élargir le spectre d'utilisation de notre système à d'autres domaines. Nous pensons notamment à la détection d'événements sonores, qui gagnerait probablement à exploiter la diversité des enregistrements que fournit une AAAH.

Le tableau 7.1 rappelle les points essentiels de ce chapitre.

Section	Points-clé
7.3	Présentation du corpus MEETIT.
7.4.1	• La connaissance *a priori* de la configuration spatiale permet d'utiliser efficacement Tango pour de la séparation de sources. • Les signaux compressés apportent une estimation de chacune des interférences dans le scénario MEETIT. • Cette information permet d'améliorer les performances des RNMuN.
7.4.2	Les performances de Tango sont stables en conditions surdéterminées, sauf si deux sources ou plus manquent par rapport au nombre de nœuds.
7.4.3	• Les RNMuN de Tango dépendent beaucoup des signaux compressés. • Tango n'est pas résilient aux conditions sous-déterminées.

TABLEAU 7.1. – Points-clé à retenir du chapitre 7.

Quatrième partie

Conclusion

8. Conclusion et perspectives

8.1. Conclusion

Cette thèse est dédiée au rehaussement de la parole dans les antennes acoustiques ad-hoc (AAAH). Nous avons proposé plusieurs systèmes basés sur l'apprentissage profond capables d'exploiter de manière distribuée l'information spatiale enregistrée par une AAAH. Nous avons considéré de nombreux cas d'applications pratiques, propres aux AAAH, évalué nos systèmes sur des données réalistes et réelles, et proposé des solutions pour l'utilisation de réseaux de neurones (RN) dans les conditions d'utilisation des AAAH.

Le **chapitre 3** a présenté l'une de nos principales contributions, appelée Tango. Tango est un système distribué de rehaussement de la parole. Basé sur l'apprentissage profond, Tango utilise des RN pour prédire des masques temps-fréquence (TF) eux-mêmes utilisés pour estimer les paramètres de filtres de Wiener multicanaux. Tango procède en deux temps. Dans un premier temps, un filtrage local est effectué sur chaque nœud d'une AAAH afin d'estimer des signaux dits compressés. Dans un second temps, ces signaux compressés sont échangés entre les nœuds avant qu'un second filtrage multicanal soit appliqué. Les signaux compressés sont utilisés à la fois pour la formation de voies et pour la prédiction des masques TF requis par la formation de voies de la seconde étape de filtrage. Nous évaluons Tango sur un premier corpus simple qui permet de quantifier ses performances dans un scénario maîtrisé. Nous montrons que l'usage de masques TF prédits par un RN permet d'égaler les performances obtenues avec un détecteur d'activité vocale (DAV) oracle. Nous montrons que les couches de convolution en entrée d'un RN permettent d'exploiter l'information spatiale apportée par les signaux compressés tout en réduisant le nombre de paramètres du RN. Enfin, nous montrons qu'utiliser les signaux compressés permet de mieux prédire les masques TF par des RN, et mettons ainsi en évidence l'intérêt de travailler avec des antennes de microphones distribuées. Dans une seconde partie du chapitre, nous présentons le corpus Disco, un corpus de signaux enregistrés par des antennes de microphones distribuées, que nous avons créé dans le cadre de cette thèse. Des évaluations sur ce corpus ont montré que Tango a des performances de rehaussement de la parole comparables à celles d'une solution de l'état de l'art, tout en ayant une complexité algorithmique moindre : il comporte 3,6 fois moins de paramètres apprenables et a une vitesse d'inférence 2,4 fois plus élevée. Il est également plus flexible car il permet d'ajuster la réduction de bruit par rapport à la distorsion de la parole.

Dans le **chapitre 4**, Tango a été évalué sur un grand nombre de scénarios du corpus Disco. Nous montrons qu'il généralise sur une grande variété de configurations spatiales et acoustiques. Nous montrons que Tango est résilient à une grande variabilité de réverbération, de rapports signal-à-bruit (RSB) et à la présence de bruit diffus, même lorsque certaines conditions d'évaluation des RN s'éloignent des conditions d'entraînement. Nous pointons toutefois les limites de Tango lorsqu'il est évalué sur des bruits qui ne sont pas présents dans les mélanges d'entraînement, et lorsqu'il est évalué sur des données réelles. La dernière partie de ce chapitre met en évidence comment exploiter l'information spatiale enregistrée par tous les nœuds d'une AAAH. Nous y montrons que les signaux compressés sont un moyen efficace pour propager l'information

spatiale et qu'ils permettent d'améliorer les performances de rehaussement de la parole de Tango. Nous montrons également que les nœuds d'une même antenne coopèrent et que les estimations compressées de la parole constituent une information également utile au niveau des nœuds les plus proches de la source de parole.

Le **chapitre 5** s'intéresse aux cas d'utilisation où le nombre de nœuds contenus dans une AAAH varie, comme cela peut être couramment le cas en pratique. Au cours d'une première étude, nous montrons que l'absence d'une partie des signaux compressés ne dégrade que légèrement les performances de la formation de voies. En revanche, nous montrons qu'elle est très préjudiciable aux performances des RN s'ils ne sont pas entraînés dans des conditions où certains canaux viennent à manquer. Nous proposons une solution pour rendre les RN utilisés dans Tango plus résilients et généraliser les performances de ce système aux cas d'utilisation où certains nœuds sont déconnectés du reste d'une AAAH. Notre solution se base sur l'introduction de mécanismes d'attention. Nous comparons les performances de deux mécanismes, un mécanisme « compresser et stimuler » (SE) et un mécanisme d'auto-attention. Une étude de dissociation des effets du mécanisme d'attention SE indique que l'amélioration des performances obtenues avec un mécanisme SE sont probablement liées au fait que ce mécanisme favorise l'entraînement du RN entier. Les poids du mécanisme SE, peu interprétables, semblent ne pas avoir de grande influence sur la prédiction des RN. Nous terminons le chapitre avec une évaluation qualitative qui met en évidence le fait que les poids du mécanisme d'auto-attention sont interprétables.

Dans le **chapitre 6**, nous nous intéressons à l'impact de l'asynchronisation des nœuds d'une AAAH sur le rehaussement de la parole. Nous proposons une extension de Tango pour le rendre opérationnel lorsque les signaux de différents nœuds d'une même AAAH ne sont pas synchronisés à cause d'un décalage et d'une dérive d'horloge. Dans une première étude, nous montrons que l'asynchronisation a un fort effet sur le calcul des métriques. Puis nous étudions l'influence de l'asynchronisation sur la formation de voies. L'influence d'un décalage d'horloge est significatif mais celle d'une dérive d'horloge est plus limitée. Nous étudions également l'impact de l'asynchronisation sur les performances des RN. Le décalage d'horloge fait augmenter la distorsion de la parole mais influence peu sur la réduction de bruit. La dérive d'horloge a très peu d'impact sur les RN. Dans la dernière partie de ce chapitre, nous nous concentrons sur l'utilisation de Tango en présence d'un décalage d'horloge, et nous proposons une solution pour limiter les conséquences d'un tel décalage. Notre solution se base sur l'introduction d'un autre mécanisme d'attention temporelle, qui cherche implicitement la corrélation temporelle des canaux asynchronisés. Cette solution rend Tango plus résilient à l'asynchronisation et conduit à une meilleure réduction de bruit sans augmenter la distorsion de la parole. Nous montrons également comment les poids cachés du mécanisme d'attention peuvent être liés à l'asynchronisation et utilisés pour retrouver un ordre de grandeur du décalage d'horloge entre les différents nœuds d'une AAAH.

Le **chapitre 7** est consacré à l'adaptation de Tango dans le contexte de la séparation de sources de parole interférentes. Nous montrons que notre système peut être utilisé dans le contexte d'une réunion au cours de laquelle plusieurs personnes parlent en même temps. Nous présentons un corpus de données créé à cet effet, où nous simulons des mélanges de deux, trois et quatre sources simultanées, assises autour d'une table circulaire. A partir de la connaissance de la configuration spatiale, nous montrons que le problème global de séparation de sources peut se transformer en plusieurs problèmes locaux de rehaussement de la parole au niveau de chaque nœud. Les signaux compressés permettent de convoyer une estimation de chaque source interférence au niveau de chaque nœud, ce qui permet à chaque nœud de disposer d'une vue complète de la

scène acoustique, et donc de mieux effectuer son rehaussement de la parole local. Cela met en évidence l'intérêt de considérer des AAAH et d'exploiter toute l'information spatiale qu'elles enregistrent. Au cours d'une évaluation avec un nombre variable de sources, nous montrons que les performances de Tango sont faibles en conditions sous-déterminées, c'est-à-dire lorsque plus de sources sont présentes que de nœuds, mais que Tango est résilient aux conditions surdéterminées.

8.2. Perspectives à court terme

Nous distinguons plusieurs limites ou modification possibles des systèmes que nous avons présentés au cours de cette thèse. Elles conduisent à quatre axes d'améliorations que nous présentons dans cette section.

8.2.1. Amélioration des performances des réseaux de neurones

Notre recherche s'étant concentrée sur le développement d'un système de rehaussement de la parole distribué, adapté aux AAAH, nous avons décidé de ne pas chercher à optimiser à tout prix les performances des RN, car cela aurait nécessité de longues expériences comparatives et aurait conduit à des architectures de RN complexes que nous voulions justement éviter. Il est néanmoins probable qu'une fonction de coût calculée sur la forme d'onde du signal finalement rehaussé par l'AAAH, plutôt que sur le masque prédit par les RN, permettrait d'améliorer la qualité ou l'intelligibilité de la parole du signal restitué (Hu et al., 2020; Kolbæk et al., 2020). On pourrait également modifier les données en entrée des RN afin de fournir l'information de la phase aux RN, dont il a été montré qu'elle apportait une information complémentaire à l'amplitude exploitable (Paliwal et al., 2011; Pariente et al., 2020b). Enfin, il est probable que d'autres architectures de RN auraient conduit à de meilleures performances. Certains changements pourraient même se faire sans augmentation du nombre de paramètres, comme l'utilisation de convolutions dilatées (Tan and Wang, 2018) et de connections résiduelles (Subakan et al., 2021) ou le remplacement des couches récurrentes par des mécanismes d'attention à têtes multiples (Pandey and Wang, 2021).

8.2.2. Optimisation de l'information échangée entre nœuds

Nous avons montré au cours de nos différentes expériences que l'information spatiale enregistrée par différents nœuds d'une AAAH était utile non seulement à la formation de voies, mais aussi à la prédiction de masques TF. Nous avons également montré qu'au niveau de certains nœuds, l'information convoyée par une estimation de la source de bruit pouvait être plus utile que celle convoyée par une estimation de la source de parole. Néanmoins, il nous a été difficile d'exploiter efficacement ce résultat et d'en profiter pour réduire la consommation en bande passante. Il serait intéressant d'explorer cette voie plus en détail, en concevant dans un premier temps une solution qui sache déterminer quelle estimation, de la parole ou du bruit, doit être échangée entre les nœuds, à partir de ce que chaque nœud voit. Dans un second temps, il serait intéressant de ne plus envoyer entre les nœuds des signaux explicitement issus d'une formation de voies, mais de transformer notre système en système de bout-en-bout dans lequel l'information échangée entre les nœuds serait la plus informative possible pour chaque nœud.

8.2.3. Application dans des antennes à topologies non contraintes

Dans cette thèse, nous avons montré que les performances de notre système sont indépendantes de la géométrie de l'antenne, et nous avons proposé une solution qui généralise avec un nombre

variable de nœuds. Toutefois, ces résultats restent limités par la topologie de l'antenne que nous avons toujours considérée comme toute-connectée. Ceci constitue une contrainte forte qui ne peut pas toujours être respectée et qui conduit à une consommation en bande passante élevée (Bertrand, 2011, Chapitre 1). Dans notre recherche de flexibilité, de généralisation à tous types de configurations et de faible consommation énergétique, il serait nécessaire de concevoir des algorithmes qui puissent opérer dans des antennes de topologie arbitraire. Cela pourrait passer par l'adaptation de nos systèmes aux solutions proposées par Szurley et al. (2016), qui montrent qu'il est possible d'utiliser la même logique de signaux compressés dans des antennes de topologie arbitraire. Une des limites principales de ces algorithmes est leur lente convergence, que les techniques d'apprentissage profond pourraient peut-être accélérer, comme cela a été fait par exemple par Carbajal et al. (2020).

8.2.4. Guidage de l'apprentissage par le format des données d'entrée

Afin d'optimiser l'apprentissage des RN, on pourrait s'inspirer de certaines connaissances en psychoacoustique afin de simplifier le rehaussement de la parole et de guider l'apprentissage des RN. Cela a déjà été fait il y a plusieurs années, lorsque les RN n'étaient pas aussi puissants qu'aujourd'hui, et où les données d'entrée représentaient de manière plus explicite l'information nécessaire aux RN pour traiter les signaux de parole (Wang et al., 2012), certaines données étant explicitement inspirées du système neurobiologique humain (Wang and Brown, 1999).

Toutefois, le reproche qui peut être fait à ces représentations est qu'elles sont « systématiques » (Rouat et al., 2004), c'est-à-dire qu'elles sont indépendantes du contexte (spectro-temporel) du signal. Or il est connu que la perception des sons par l'oreille humaine dépend du contexte (Moore, 2012, 6$^{\text{ème}}$ édition).

On pourrait pour cela représenter les signaux sous forme d'impulsions, selon une proposition de Rouat et al. (2004) qui considère que les nerfs du système auditif humain émettent des signaux sous forme de pics dont l'ordre d'émission est porteur de l'information. L'idée de remplacer les représentations classiques par des pics asynchrones a été reprise récemment pour encoder des signaux de parole avec des RN à impulsions (Pan et al., 2018, 2020), puis combinée à l'utilisation de cartes auto-organisatrices pour la classification des sons (Wu et al., 2018) ou pour la reconnaissance de la parole (Wu et al., 2020). Les RN à impulsions semblent prometteurs car ils ont une puissance de modélisation égale aux RN classiques pour un nombre de paramètres inférieur (Maass, 1997; Tavanaei et al., 2019). Les avantages des RN à impulsions, longtemps inexploitables à cause de la difficile implémentation logicielle et matérielle des architectures associées, pourraient aujourd'hui être efficacement mis à profits pour le rehaussement de la parole. Ils semblent en effet adaptés pour imiter le système perceptif humain et pour encoder les événements temporels déterminants à la perception de la parole, comme les débuts de phonèmes.

8.3. Applications en conditions réelles

8.3.1. Généralisation des performances sur données réelles

L'évaluation en partie 4.4.1 a donné des résultats peu convaincants au sujet de la généralisation des RN sur ce type de données. C'est un problème récurrent dans le domaine du rehaussement de la parole assisté par l'apprentissage profond, où Ceolini et al. (2020) ont montré que les (bonnes) performances obtenues sur des données simulées occultent des résultats souvent moins bons sur des données réelles. Afin d'améliorer la généralisation des performances sur données réelles, nous proposons d'intégrer plus de données réelles dans les corpus d'apprentissage et d'adopter une démarche semi-supervisée (Turpault et al., 2019; Sekiguchi et al., 2019; Leglaive et al., 2019;

Xia et al., 2021) voire non supervisée (Wisdom et al., 2020) afin de pallier l'absence de vérité terrain. On pourrait également concevoir des méthodes d'apprentissage de RN qui minimisent le risque de sur-spécialisation des RN sur les données d'entraînement. Une meilleure normalisation des données en entrée (Yoshioka et al., 2018) et l'augmentation des données (Braun et al., 2021) en sont deux exemples. Enfin, il serait intéressant de recourir à des techniques d'adaptation de domaine (Liao et al., 2019; Cornell et al., 2020) pour rendre les RN plus résilients à des données non vues à l'entraînement.

8.3.2. Développement d'une solution en temps réel

De nombreuses applications réelles de notre système nécessiteraient qu'il fonctionne en temps réel, ce qui a trois implications. La première est que le traitement par blocs, présenté en section 3.2.1, doit être remplacé par une alternative dite *en ligne* qui traite les signaux au fur et à mesure de leur enregistrement. Pour cela, plutôt que d'estimer les matrices d'autocorrélation spatiale (MAS) de la parole et du bruit par moyenne sur des blocs de signaux, il est possible de les calculer à l'aide d'une moyenne glissante ou exponentielle. La deuxième implication est que le temps de latence doit être réduit au minimum. Dans le cas où l'on effectue les traitements dans le domaine de la transformée de Fourier à court terme (TFCT), la latence minimale est la durée du pas d'avancement. Il est toutefois envisageable de considérer des approches de bout-en-bout qui travaillent directement sur la forme d'onde, et dont les fenêtres d'analyse peuvent être aussi courtes que 2 ms (Pariente et al., 2020b), mais reposant en général sur des architectures complexes. La troisième implication est que les calculs doivent être réduits de telle sorte que toutes les opérations puissent être effectuées par des appareils de puissance limitée dans le temps défini par la latence. Aujourd'hui, les solutions reposant sur de faibles latences demandent de grandes puissances de calcul qu'on ne peut pas encore effectuer sur de tels appareils.

8.3.3. Évaluation des performances adaptée aux conditions réelles de fonctionnement

Prendre en compte la subjectivité des personnes à qui est destiné l'usage des solutions de rehaussement de la parole est une problématique encore peu abordée. Cela est particulièrement important dans le cas où le rehaussement de la parole s'adresse à des personnes mal-entendantes, dont les pertes auditives introduisent des artéfacts auxquels les solutions de rehaussement de la parole devraient s'adapter pour un traitement optimal. L'index de perception de la parole dans les aides auditives (HASPI, *Hearing-Aid Speech Perception Index*) de Kates and Arehart (2021) va dans ce sens et montre qu'un travail sur les métriques semble nécessaire afin de considérer ces différents aspects. Un challenge a été récemment créé pour inciter à la proposition de solutions à ce défi [1].

Enfin, dans le même ordre d'esprit, les solutions de rehaussement de la parole gagneraient peut-être en simplicité si elles ne considéraient que les traitements qui apportent des changements perceptibles pour les utilisateurs finaux. Il conviendrait donc d'intégrer les connaissances que l'on a sur le système auditif humain pour orienter les résultats vers des performances plus adaptées à l'audition humaine. Cela avait par exemple été proposé par Lagrange and Marchand (2001) qui ont pris en compte des phénomènes psychoacoustiques pour accélérer la synthèse de sons sans réduire les performances de leur système.

1. The 1st Clarity Prediction Challenge. https://claritychallenge.github.io/clarity_CPC1_doc/

8.4. Ouvertures à d'autres applications

Nous voyons de nombreuses perspectives qui pourraient faire suite aux travaux de cette thèse. Nous en citons ici deux principales, à savoir l'extension de nos recherches à la détection d'événements sonores et leur application dans le domaine des aides auditives.

8.4.1. Application à la détection d'événements sonores

Du fait de la large étendue spatiale des AAAH, elles offrent un fort potentiel pour la détection d'événements sonores. On pourrait utiliser les différents appareils répartis dans nos habitations pour la détection d'événements domestiques, en exploitant la diversité des enregistrements faits dans différentes pièces de l'habitation (Ebbers et al., 2021). Dans un environnement urbain, les techniques de traitement de signaux d'une AAAH peuvent également être mises à profit pour estimer ce qu'il se passe dans une ville (état de la circulation, accident, intervention de police, etc.) en installant des microphones à des endroits stratégiques (Mydlarz et al., 2017; Picaut et al., 2020). Toujours en environnement extérieur, on pourrait élargir le spectre des applications de nos recherches au contexte des antennes de capteurs bioacoustiques et à la détection d'événements biologiques (Lostanlen et al., 2019).

8.4.2. Utilisation dans un contexte multitâche

On peut envisager de restituer le rehaussement de la parole sous différentes formes. Par exemple, il a été montré que la stimulation électro-haptique (en fournissant l'enveloppe du signal débruité par vibrations sur une partie du corps) peut aider des utilisateurs d'implants cochléaires à mieux comprendre le contenu d'un signal de parole (Fletcher et al., 2020). Les aides auditives à conduction osseuse sont d'autres solutions pour restituer un signal débruité sous une autre forme que celle d'un signal audio (Ellsperman et al., 2021). Bien que ces différents appareils (prothèses haptiques, aides auditives à conduction osseuse, etc.) ne soient pas des capteurs, mais des émetteurs, il est possible de les intégrer à une AAAH, par exemple en combinaison avec des aides auditives classiques. Le potentiel des AAAH pourrait également être mieux exploité si l'on profitait des écrans de certains de leurs appareils. On pourrait par exemple représenter le visage de la personne qui parle ou transcrire ce qu'elle dit, afin de faciliter sa compréhension par les utilisateurs des AAAH (McGurk and MacDonald, 1976; Golumbic et al., 2013). Dans ce contexte, le rehaussement de la parole gagnerait à faire coopérer ces différents appareils dans une approche multitâche en optimisant le signal que chacun restitue en fonction de ce qui est enregistré.

8.4.3. Utilisation dans un contexte multimodal

D'autres capteurs que des microphones peuvent appartenir à une AAAH. Intégrer leurs signaux à une approche multimodale ou multivue (Xu et al., 2013; Li et al., 2018) du rehaussement de la parole pourrait exploiter toutes les informations captées par une AAAH et améliorer les performances du rehaussement de la parole.

Tout d'abord, les caméras des téléphones ou des ordinateurs d'une AAAH peuvent apporter une information visuelle supplémentaire. Des recherches ont déjà montré que cette information peut servir à estimer l'angle d'arrivée de la source (Tan et al., 2020) ou des attributs de la parole (Sadeghi et al., 2020; Carbajal et al., 2021) qui permettent d'améliorer les performances du rehaussement de la parole.

Par ailleurs, l'utilisation de signaux électro-encéphalographiques devient de plus en plus envisageable dans le contexte de la séparation de sources, d'une part parce que la recherche propose des solutions encourageantes dans ce sens (Cantisani et al., 2019; Geirnaert et al., 2021), d'autre part parce que du matériel est développé pour obtenir de tels signaux dans un contexte quotidien (An et al., 2021). Le grand avantage des électro-encéphalogrammes est qu'ils contiennent une information *postérieure* au traitement par le cerveau humain. Par exemple ils peuvent indiquer quelle source, parmi plusieurs sources interférentes, est la source d'intérêt, ce qu'il est impossible de savoir *a priori*.

Cinquième partie

Annexes

A. Métriques d'évaluation des performances en séparation aveugle de sources audio

On considère une scène acoustique avec N sources $\{s_i\}_{i=1..N}$. On suppose que toutes les sources sonores sont ponctuelles, donc qu'il n'y a pas de bruit ambiant. Soit y le signal enregistré par un microphone lorsque ces N sources sont actives. On a :

$$y = \sum_{n=1}^{N} h_n * s_n \tag{A.1}$$

$$= \sum_{n=1}^{N} c_n \,, \tag{A.2}$$

où $c_n = h_n * s_n$ est la convolution de la source s_n avec la réponse impulsionnelle (RI) déterminée par les positions de la source n et du microphone.

Dans la suite de cette annexe, on considérera la source n comme la source cible et les autres sources $\{s_j\}_{j\neq n}$ seront dites interférentes. Soit x_n l'estimation de la source n par un algorithme de séparation de sources. Vincent et al. (2006) proposent de décomposer le signal x_n comme la somme de quatre termes :

$$x_n = x_{\text{cible}} + e_{\text{inter}} + e_{\text{bruit}} + e_{\text{arté}} \,, \tag{A.3}$$

avec

- x_{cible} la composante de s_n modifiée par une déformation acceptable.
- e_{inter} la composante résiduelle des autres sources interférentes $\{s_j\}_{j\neq n}$.
- e_{bruit} la composante résiduelle du bruit de microphones.
- $e_{\text{arté}}$ le terme des artéfacts introduits par le traitement du signal.

Ces composantes permettent de calculer les trois métriques suivantes :

- le rapport source-à-distorsion (SDR : *source to distortion ratio*)

$$\text{SDR} = 10\log_{10}\frac{||x_{\text{cible}}||^2}{||e_{\text{inter}} + e_{\text{arté}} + e_{\text{bruit}}||^2} \tag{A.4}$$

- le rapport source-à-interférences (SIR : *source to interferences ratio*)

$$\text{SIR} = 10\log_{10}\frac{||x_{\text{cible}}||^2}{||e_{\text{inter}}||^2} \tag{A.5}$$

- le rapport sources-à-artéfacts (SAR : *sources to artifacts ratio*)

$$\text{SAR} = 10\log_{10}\frac{||x_{\text{cible}} + e_{\text{inter}} + e_{\text{bruit}}||^2}{||e_{\text{arté}}||^2} \tag{A.6}$$

Pour estimer x_{cible}, différentes distorsions peuvent être appliquées à s_n sans que l'utilisateur de l'algorithme considère qu'elles impactent négativement son traitement. Dans le cas de notre

thèse, la distorsion amenée par la réverbération de la pièce est une distorsion que l'on ne cherche pas à annuler, si bien que l'on peut avoir $x_{\text{cible}} = c_n$.

En pratique, une librairie largement utilisée pour calculer les métriques décrites [1] cherche un filtre à réponse finie (FIR) de 512 points g de telle sorte à minimiser la distance entre le signal estimé et le signal de référence convolué par ce FIR. En termes mathématiques, cela reviendrait à considérer une autre référence cible :

$$x_{\text{cible}} = g * s_n \qquad \text{où} \qquad g = \underset{g'}{\text{argmin}}\{||x_n - g' * s_n||^2\}\,. \tag{A.7}$$

Outre la complexité algorithmique requise pour trouver ce FIR, cette méthode de calcul de métriques a deux inconvénients. Le premier inconvénient est qu'elle augmente artificiellement les métriques puisque le FIR g maximise justement la ressemblance entre le signal estimé x_n et la composante en signal de référence. Comme montré par Le Roux et al. (2019), le FIR g peut reproduire des distorsions qui ne sont pas du tout celles de la réverbération et occulter des distorsions indésirables apportées par l'algorithme de séparation de sources. Le deuxième inconvénient est que deux algorithmes de séparation de sources ne seront pas comparés de manière équitable, puisqu'un nouveau FIR sera calculé pour le signal estimé par chacun de ces algorithmes. Implicitement, ils ne seront donc pas comparés à la même référence.

En partant de ces constats, Le Roux et al. (2019) ont proposé une nouvelle métrique, le rapport signal-à-distortion à invariance d'échelle (SI-SDR : *scale-invariant signal to distortion ratio*) qui autorise seulement qu'un changement d'échelle soit appliqué à la référence s_n. Ils définissent le SI-SDR de la manière suivante :

$$\text{SI-SDR} = 10\log_{10}\frac{||\alpha s_n||^2}{||\alpha s_n - x_n\ ||^2}\,, \tag{A.8}$$

avec

$$\alpha = \underset{\alpha'}{\text{argmin}}\{||x_n - \alpha s_n||^2\} \tag{A.9}$$

Le facteur d'échelle qui vérifie l'équation (A.9) est :

$$\alpha = \frac{x_n^T s_n}{||s_n||^2}\,. \tag{A.10}$$

1. https://pypi.org/project/mir_eval/

Bibliographie

Affes, S. and Grenier, Y. (1997). A signal subspace tracking algorithm for microphone array processing of speech. *IEEE Transactions on Speech and Audio Processing*, 5(5) :425–437.

Agnew, J. and Thornton, J. M. (2000). Just noticeable and objectionable group delays in digital hearing aids. *Journal of the American Academy of Audiology*, 11(6).

Amini, J., Hendriks, R. C., Heusdens, R., Guo, M., and Jensen, J. (2019). Rate-constrained noise reduction in wireless acoustic sensor networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28 :1–12.

An, W. W., Shinn-Cunningham, B., Gamper, H., Emmanouilidou, D., Johnston, D., Jalobeanu, M., Cutrell, E., Wilson, A., Chiang, K.-J., and Tashev, I. (2021). Decoding Music Attention from "EEG Headphones" : A User-Friendly Auditory Brain-Computer Interface. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 985–989. IEEE.

Andén, J., Lostanlen, V., and Mallat, S. (2015). Joint time-frequency scattering for audio classification. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

Andersen, K. T. and Moonen, M. (2017). Robust speech-distortion weighted interframe wiener filters for single-channel noise reduction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(1) :97–107.

Anthony, L. F. W., Kanding, B., and Selvan, R. (2020). Carbontracker : Tracking and predicting the carbon footprint of training deep learning models. *arXiv preprint arXiv :2007.03051*.

Araki, S., Hayashi, T., Delcroix, M., Fujimoto, M., Takeda, K., and Nakatani, T. (2015). Exploring multi-channel features for denoising-autoencoder-based speech enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 116–120. IEEE.

Araki, S., Ono, N., Kinoshita, K., and Delcroix, M. (2018). Comparison of reference microphone selection algorithms for distributed microphone array based speech enhancement in meeting recognition scenarios. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 316–320. IEEE.

Avargel, Y. and Cohen, I. (2007). Adaptive system identification in the short-time fourier transform domain using cross-multiplicative transfer function approximation. *IEEE transactions on audio, speech, and language processing*, 16(1) :162–173.

Badeau, R. (2011). Gaussian modeling of mixtures of non-stationary signals in the time-frequency domain (hr-nmf). In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 253–256. IEEE.

Badue, C., Guidolini, R., Carneiro, R. V., Azevedo, P., Cardoso, V. B., Forechi, A., Jesus, L., Berriel, R., Paixao, T. M., Mutz, F., et al. (2021). Self-driving cars : A survey. *Expert Systems with Applications*, 165 :113816.

Bahari, M. H., Bertrand, A., and Moonen, M. (2015). Blind sampling rate offset estimation based on coherence drift in wireless acoustic sensor networks. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 2281–2285. IEEE.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv :1409.0473*.

Balan, R. and Rosca, J. (2002). Microphone array speech enhancement by bayesian estimation of spectral amplitude and phase. In *Sensor Array and Multichannel Signal Processing Workshop Proceedings, 2002*, pages 209–213. IEEE.

Barker, J., Marxer, R., Vincent, E., and Watanabe, S. (2015). The third 'CHiME' speech separation and recognition challenge : Dataset, task and baselines. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 504–511. IEEE.

Baugé, C., Lagrange, M., Andén, J., and Mallat, S. (2013). Representing environmental sounds using the separable scattering transform. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8667–8671. IEEE.

Benesty, J. and Huang, Y. (2011). A single-channel noise reduction mvdr filter. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 273–276.

Berouti, M., Schwartz, R., and Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. In *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 208–211. IEEE.

Bertrand, A. (2011). *Signal Processing Algorithms for Wireless Acoustic Sensor Networks (Signaalverwerkingsalgoritmes voor draadloze akoestische sensornetwerken)*. PhD thesis, Katholieke Universiteit Leuven.

Bertrand, A. and Moonen, M. (2009). Robust distributed noise reduction in hearing aids with external acoustic sensor nodes. *EURASIP Journal on Advances in Signal Processing*, 2009 :1–14.

Bertrand, A. and Moonen, M. (2010a). Distributed adaptive node-specific signal estimation in fully connected sensor networks — Part I : Sequential node updating.

Bertrand, A. and Moonen, M. (2010b). Distributed adaptive node-specific signal estimation in fully connected sensor networks — Part II : Simultaneous and asynchronous node updating.

Bertrand, A. and Moonen, M. (2010). Efficient sensor subset selection and link failure response for linear mmse signal estimation in wireless sensor networks. In *2010 18th European Signal Processing Conference*, pages 1092–1096. IEEE.

Bertrand, A. and Moonen, M. (2011a). Consensus-based distributed total least squares estimation in ad hoc wireless sensor networks. *IEEE Transactions on Signal Processing*, 59(5) :2320–2330.

Bertrand, A. and Moonen, M. (2011b). Distributed node-specific LCMV beamforming in wireless sensor networks. *IEEE Transactions on Signal Processing*, 60(1) :233–246.

Bertrand, A. and Moonen, M. (2013). Distributed LCMV beamforming in a wireless sensor network with single-channel per-node signal transmission. *IEEE Transactions on Signal Processing*, 61(13) :3447–3459.

Bhat, G. S., Shankar, N., Reddy, C. K., and Panahi, I. M. (2019). A real-time convolutional neural network based speech enhancement for hearing impaired listeners using smartphone. *IEEE Access*, 7 :78421–78433.

Bie, X., Leglaive, S., Alameda-Pineda, X., and Girin, L. (2021). Unsupervised speech enhancement using dynamical variational auto-encoders. *arXiv preprint arXiv :2106.12271*.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, 27(2) :113–120.

Boyd, S., Ghosh, A., Prabhakar, B., and Shah, D. (2006). Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6) :2508–2530.

Braun, S., Gamper, H., Reddy, C. K., and Tashev, I. (2021). Towards efficient models for real-time deep noise suppression. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 656–660. IEEE.

Breed, B. R. and Strauss, J. (2002). A short proof of the equivalence of lcmv and gsc beamforming. *IEEE Signal Processing Letters*, 9(6) :168–169.

Bregman, A. S. (1994). *Auditory scene analysis : The perceptual organization of sound*. MIT press.

Bryson, A. and Johansen, D. (1965). Linear filtering for time-varying systems using measurements containing colored noise. *IEEE Transactions on Automatic Control*, 10(1) :4–10.

Buckley, K. and Griffiths, L. (1986). An adaptive generalized sidelobe canceller with derivative constraints. *IEEE Transactions on antennas and propagation*, 34(3) :311–319.

Cantisani, G., Essid, S., and Richard, G. (2019). Eeg-based decoding of auditory attention to a target instrument in polyphonic music. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 80–84. IEEE.

Capon, J. (1969). High-resolution frequency-wavenumber spectrum analysis. *Proceedings of the IEEE*, 57(8) :1408–1418.

Carbajal, G., Richter, J., and Gerkmann, T. (2021). Disentanglement learning for variational autoencoders applied to audio-visual speech enhancement. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE.

Carbajal, G., Serizel, R., Vincent, E., and Humbert, E. (2020). Joint NN-Supported Multichannel Reduction of Acoustic Echo, Reverberation and Noise. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28 :2158–2173.

Cardoso, J.-F. and Souloumiac, A. (1994). Blind beamforming for non-gaussian signals. volume 140, pages 362 – 370.

Cardoso, J.-F. and Souloumiac, A. (1996). Jacobi angles for simultaneous diagonalization. *SIAM journal on matrix analysis and applications*, 17(1) :161–164.

Casebeer, J., Kaikaus, J., and Smaragdis, P. (2021). Communication-cost aware microphone selection for neural speech enhancement with ad-hoc microphone arrays. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8438–8442. IEEE.

Casebeer, J., Luc, B., and Smaragdis, P. (2018). Multi-view networks for denoising of arbitrary numbers of channels. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 496–500. IEEE.

Ceolini, E., Kiselev, I., and Liu, S.-C. (2020). Evaluating multi-channel multi-device speech separation algorithms in the wild : a hardware-software solution. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28 :1428–1439.

Chakrabarty, S. and Habets, E. A. P. (2019). Time-Frequency Masking Based Online Multi-Channel Speech Enhancement With Convolutional Recurrent Neural Networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(4) :1–1.

Chaudhari, Q. M. (2011). A simple and robust clock synchronization scheme. *IEEE transactions on communications*, 60(2) :328–332.

Chen, F., Hazrati, O., and Loizou, P. C. (2013). Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure. *Biomedical signal processing and control*, 8(3) :311–314.

Chen, M., Liu, Z., He, L.-W., Chou, P., and Zhang, Z. (2007). Energy-based position estimation of microphones and speakers for ad hoc microphone arrays. In *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 22–25.

Chen, S., Wu, Y., Chen, Z., Wu, J., Yoshioka, T., Liu, S., Li, J., and Yu, X. (2021). Ultra Fast Speech Separation Model with Teacher Student Learning. In *Proc. Interspeech 2021*, pages 3026–3030.

Cherkassky, D. and Gannot, S. (2017). Blind synchronization in wireless acoustic sensor networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(3) :651–661.

Cherkassky, D., Markovich-Golan, S., and Gannot, S. (2015). Performance analysis of MVDR beamformer in WASN with sampling rate offsets and blind synchronization. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 245–249. IEEE.

Chiba, H., Ono, N., Miyabe, S., Takahashi, Y., Yamada, T., and Makino, S. (2014). Amplitude-based speech enhancement with nonnegative matrix factorization for asynchronous distributed recording. In *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 203–207. IEEE.

Chinaev, A., Thüne, P., and Enzner, G. (2021). Double-cross-correlation processing for blind sampling-rate and time-offset estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29 :1881–1896.

Cho, K., Courville, A., and Bengio, Y. (2015). Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*, 17(11) :1875–1886.

Cho, Y.-C., Choi, S., and Bang, S.-Y. (2003). Non-negative component parts of sound for classification. In *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (IEEE Cat. No. 03EX795)*, pages 633–636. IEEE.

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, R. M. and Singer, A. C. (2018). Speech separation using partially asynchronous microphone arrays without resampling. In *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–9. IEEE.

Corey, R. M., Tsuda, N., and Singer, A. C. (2019). Acoustic impulse responses for wearable audio devices. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 216–220. IEEE.

Cornelis, B., Moonen, M., and Wouters, J. (2010). Performance analysis of multichannel wiener filter-based noise reduction in hearing aids under second order statistics estimation errors. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5) :1368–1381.

Cornell, S., Olvera, M., Pariente, M., Pepe, G., Principi, E., Gabrielli, L., and Squartini, S. (2020). Domain-adversarial training and trainable parallel front-end for the dcase 2020 task 4 sound event detection challenge. In *DCASE 2020-5th Workshop on Detection and Classification of Acoustic Scenes and Events*.

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, H., Zeskind, R., and Owen, M. (1987). Robust adaptive beamforming. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(10) :1365–1376.

Cui, Z. and Bao, C. (2021). Power exponent based weighting criterion for DNN-based mask approximation in speech enhancement. *IEEE Signal Processing Letters*, 28 :618–622.

Dai, Y., Zhu, C., Shan, X., Cheng, Z., and Zou, B. (2019). A survey on intelligent screening for diabetic retinopathy. *Chinese Medical Sciences Journal*, 34(2) :120–132.

Dash, T. K., Solanki, S. S., and Panda, G. (2021). Multi-objective approach to speech enhancement using tunable q-factor-based wavelet transform and ann techniques. *Circuits, Systems, and Signal Processing*, pages 1–31.

Delcroix, M., Hikichi, T., and Miyoshi, M. (2007). Precise dereverberation using multichannel linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2) :430–440.

Dendrinos, M., Bakamidis, S., and Carayannis, G. (1991). Speech enhancement from noise : A regenerative approach. *Speech Communication*, 10(1) :45–57.

Deng, L., Li, G., Han, S., Shi, L., and Xie, Y. (2020). Model compression and hardware acceleration for neural networks : A comprehensive survey. *Proceedings of the IEEE*, 108(4) :485–532.

Dinkel, H., Wang, S., Xu, X., Wu, M., and Yu, K. (2021). Voice activity detection in the wild : A data-driven approach using teacher-student training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29 :1542–1555.

Doclo, S., Gannot, S., Moonen, M., Spriet, A., Haykin, S., and Liu, K. R. (2010a). Acoustic beamforming for hearing aid applications. *Handbook on array processing and sensor networks*, pages 269–302.

Doclo, S., Klasen, T. J., Van den Bogaert, T., Wouters, J., and Moonen, M. (2006). Theoretical analysis of binaural cue preservation using multi-channel wiener filtering and interaural transfer functions. In *Proc. Int. Workshop Acoust. Echo Noise Control (IWAENC)*, pages 1–4.

Doclo, S., Lawin-Ore, T. C., and Rohdenburg, T. (2010b). Rate-constrained binaural MWF-based noise reduction algorithms. *signal*, 1 :2.

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.

Doclo, S., Moonen, M., Van den Bogaert, T., and Wouters, J. (2009). Reduced-bandwidth and distributed mwf-based noise reduction algorithms for binaural hearing aids. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1) :38–51.

Doclo, S., Spriet, A., Wouters, J., and Moonen, M. (2007). Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction. *Speech Communication*, 49(7-8) :636–656.

Drude, L., Heitkaemper, J., Boeddeker, C., and Haeb-Umbach, R. (2019). SMS-WSJ : Database, performance measures, and baseline recipe for multi-channel source separation and recognition. *arXiv preprint arXiv :1910.13934*.

Drude, L., Heymann, J., Schwarz, A., and Valin, J.-M. (2021). Multi-channel opus compression for far-field automatic speech recognition with a fixed bitrate budget. *arXiv preprint arXiv :2106.07994*.

Défossez, A., Synnaeve, G., and Adi, Y. (2020). Real Time Speech Enhancement in the Waveform Domain. In *Proc. Interspeech 2020*, pages 3291–3295.

Ebbers, J., Keyser, M. C., and Haeb-Umbach, R. (2021). Adapting sound recognition to a new environment via self-training. In *29th European Signal Processing Conference (EUSIPCO)*.

Ellsperman, S. E., Nairn, E. M., and Stucken, E. Z. (2021). Review of bone conduction hearing devices. *Audiology Research*, 11(2) :207–219.

Elson, J. and Römer, K. (2003). Wireless sensor networks : A new regime for time synchronization. *ACM SIGCOMM Computer Communication Review*, 33(1) :149–154.

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, 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.

Ephraim, Y. and Van Trees, H. L. (1995). A signal subspace approach for speech enhancement. *IEEE Transactions on speech and audio processing*, 3(4) :251–266.

Er, M. and Cantoni, A. (1983). Derivative constraints for broad-band element space antenna array processors. *IEEE transactions on acoustics, speech, and signal processing*, 31(6) :1378–1393.

Erdogan, H., Hershey, J. R., Watanabe, S., and Le Roux, J. (2015). Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 708–712. IEEE.

Erdogan, H., Hershey, J. R., Watanabe, S., Mandel, M. I., and Le Roux, J. (2016). Improved MVDR beamforming using single-channel mask prediction networks. In *Interspeech*, pages 1981–1985.

Erkelens, J. S., Hendriks, R. C., Heusdens, R., and Jensen, J. (2007). Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(6) :1741–1752.

Falk, T. H., Zheng, C., and Chan, W.-Y. (2010). A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7) :1766–1774.

Fedorov, I., Stamenovic, M., Jensen, C., Yang, L.-C., Mandell, A., Gan, Y., Mattina, M., and Whatmough, P. N. (2020). Tinylstms : Efficient neural speech enhancement for hearing aids. *arXiv preprint arXiv :2005.11138*.

Févotte, C., Bertin, N., and Durrieu, J.-L. (2009). Nonnegative matrix factorization with the itakura-saito divergence : With application to music analysis. *Neural computation*, 21(3) :793–830.

Flanagan, J., Johnston, J., Zahn, R., and Elko, G. (1985). Computer-steered microphone arrays for sound transduction in large rooms. *The Journal of the Acoustical Society of America*, 78(5) :1508–1518.

Flandrin, P., Goncalves, P., and Rilling, G. (2004). Detrending and denoising with empirical mode decompositions. In *2004 12th European Signal Processing Conference*, pages 1581–1584. IEEE.

Fletcher, M. D., Song, H., and Perry, S. W. (2020). Electro-haptic stimulation enhances speech recognition in spatially separated noise for cochlear implant users. *Scientific Reports*, 10(1) :1–8.

Font, F., Roma, G., and Serra, X. (2013). Freesound technical demo. *ACM International Conference on Multimedia (MM'13)*, pages 411–412.

Fraley, J. B. and Cannady, J. (2017). The promise of machine learning in cybersecurity. In *SoutheastCon 2017*, pages 1–6. IEEE.

Frost, O. L. (1972). An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE*, 60(8) :926–935.

Fu, S.-W., Hu, T.-y., Tsao, Y., and Lu, X. (2017). Complex spectrogram enhancement by convolutional neural network with multi-metrics learning. In *2017 IEEE 27th international workshop on machine learning for signal processing (MLSP)*, pages 1–6. IEEE.

Fujimura, T., Koizumi, Y., Yatabe, K., and Miyazaki, R. (2021). Noisy-target training : A training strategy for dnn-based speech enhancement without clean speech. *arXiv preprint arXiv :2101.08625*.

Gamper, H., Reddy, C. K., Cutler, R., Tashev, I. J., and Gehrke, J. (2019). Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 85–89. IEEE.

Gaubitch, N. D., Kleijn, W. B., and Heusdens, R. (2013). Auto-localization in ad-hoc microphone arrays. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 106–110. IEEE.

Gaubitch, N. D., Kleijn, W. B., and Heusdens, R. (2014). Calibration of distributed sound acquisition systems using toa measurements from a moving acoustic source. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7455–7459. IEEE.

Gburrek, T., Schmalenstroeer, J., Brendel, A., Kellermann, W., and Haeb-Umbach, R. (2021a). Deep neural network based distance estimation for geometry calibration in acoustic sensor networks. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 196–200. IEEE.

Gburrek, T., Schmalenstroeer, J., and Haeb-Umbach, R. (2021b). Geometry calibration in wireless acoustic sensor networks utilizing doa and distance information. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1) :1–17.

Geirnaert, S., Vandecappelle, S., Alickovic, E., de Cheveigne, A., Lalor, E., Meyer, B. T., Miran, S., Francart, T., and Bertrand, A. (2021). Electroencephalography-based auditory attention decoding : Toward neurosteered hearing devices. *IEEE Signal Processing Magazine*, 38(4) :89–102.

Gentet, E., David, B., Denjean, S., Richard, G., and Roussarie, V. (2020). Speech intelligibility enhancement by equalization for in-car applications. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6934–6938. IEEE.

Gerkmann, T. and Hendriks, R. C. (2011). Unbiased mmse-based noise power estimation with low complexity and low tracking delay. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4) :1383–1393.

Giri, R., Isik, U., and Krishnaswamy, A. (2019). Attention wave-u-net for speech enhancement. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 249–253. IEEE.

Giri, R., Venkataramani, S., Valin, J.-M., Isik, U., and Krishnaswamy, A. (2021). Personalized percepnet : Real-time, low-complexity target voice separation and enhancement. *arXiv preprint arXiv :2106.04129*.

Golumbic, E. Z., Cogan, G. B., Schroeder, C. E., and Poeppel, D. (2013). Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". *Journal of Neuroscience*, 33(4) :1417–1426.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. (2014). Generative adversarial networks. *ArXiv*, abs/1406.2661.

Griffiths, L. and Jim, C. (1982). An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on antennas and propagation*, 30(1) :27–34.

Guo, X., Meng, R., Zheng, C., and Li, X. (2021a). Spatial-temporal correlation based signals gathering in wasns. In *Audio Engineering Society Convention 150*. Audio Engineering Society.

Guo, X., Yuan, M., Ke, Y., Zheng, C., and Li, X. (2021b). Distributed node-specific block-diagonal LCMV beamforming in wireless acoustic sensor networks. *Signal Processing*, 185 :108085.

Haigh, J. and Mason, J. (1993). Robust voice activity detection using cepstral features. In *Proceedings of TENCon'93. IEEE Region 10 International Conference on Computers, Communications and Automation*, volume 3, pages 321–324. IEEE.

Hao, X., Shan, C., Xu, Y., Sun, S., and Xie, L. (2019). An attention-based neural network approach for single channel speech enhancement. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6895–6899. IEEE.

Hao, X., Su, X., Wang, Z., Zhang, Q., Xu, H., and Gao, G. (2020). SNR-Based Teachers-Student Technique For Speech Enhancement. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Haruta, C. and Ono, N. (2021). A low-computational dnn-based speech enhancement for hearing aids based on element selection. In *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE.

Hassani, A., Bertrand, A., and Moonen, M. (2015). GEVD-based low-rank approximation for distributed adaptive node-specific signal estimation in wireless sensor networks. *IEEE Transactions on Signal Processing*, 64(10) :2557–2572.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., and Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248) :1–43.

Hendriks, R. C., Heusdens, R., and Jensen, J. (2010). Mmse based noise psd tracking with low complexity. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4266–4269. IEEE.

Hennecke, M., Plotz, T., Fink, G. A., Schmalenstroeer, J., and Hab-Umbach, R. (2009). A hierarchical approach to unsupervised shape calibration of microphone array networks. In *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 257–260. IEEE.

Hershey, J. R., Chen, Z., Le Roux, J., and Watanabe, S. (2016). Deep clustering : Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE.

Heusdens, R., Zhang, G., Hendriks, R. C., Zeng, Y., and Kleijn, W. B. (2012). Distributed MVDR beamforming for (wireless) microphone networks using message passing. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–4.

Heymann, J., Drude, L., and Haeb-Umbach, R. (2016). Neural network based spectral mask estimation for acoustic beamforming. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2016-May, pages 196–200.

Higuchi, T., Kinoshita, K., Delcroix, M., and Nakatani, T. (2017). Adversarial training for data-driven speech enhancement without parallel corpus. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 40–47. IEEE.

Ho, M. T., Lee, J., Lee, B.-K., Yi, D. H., and Kang, H.-G. (2020). A cross-channel attention-based wave-u-net for multi-channel speech enhancement. In *INTERSPEECH*, pages 4049–4053.

Hu, D., Chen, Z., and Yin, F. (2019). Frequency response calibration using multi-channel wiener filters for microphone arrays. *IEEE Sensors Journal*, 19(17) :7507–7514.

Hu, G. and Wang, D. (2004). Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on neural networks*, 15(5) :1135–1150.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Hu, K. and Wang, D. (2012). An unsupervised approach to cochannel speech separation. *IEEE Transactions on audio, speech, and language processing*, 21(1) :122–131.

Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Zhang, B., and Xie, L. (2020). DC-CRN : Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement. In *Proc. Interspeech 2020*, pages 2472–2476.

Huang, H., Zhao, L., Chen, J., and Benesty, J. (2014). A minimum variance distortionless response filter based on the bifrequency spectrum for single-channel noise reduction. *Digital Signal Processing*, 33 :169–179.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N.-C., Tung, C. C., and Liu, H. H. (1998). The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A : mathematical, physical and engineering sciences*, 454(1971) :903–995.

Hummersone, C., Stokes, T., and Brookes, T. (2014). On the ideal ratio mask as the goal of computational auditory scene analysis. In *Blind source separation*, pages 349–368. Springer.

ITU, I. (2002). Méthodologie d'évaluation subjective de la qualité des images de télévision.

Jensen, J. and Hendriks, R. C. (2011). Spectral magnitude minimum mean-square error binary masks for dft based speech enhancement. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4736–4739. IEEE.

Jensen, S. H., Hansen, P. C., Hansen, S. D., and Sorensen, J. A. (1995). Reduction of broad-band noise in speech by truncated qsvd. *IEEE Transactions on speech and audio processing*, 3(6) :439–448.

Jiang, Y., Wang, D., Liu, R., and Feng, Z. (2014). Binaural classification for reverberant speech segregation using deep neural networks. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 22(12) :2112–2121.

Kadıoğlu, B., Horgan, M., Liu, X., Pons, J., Darcy, D., and Kumar, V. (2020). An empirical study of conv-tasnet. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7264–7268. IEEE.

Kates, J. M. and Arehart, K. H. (2021). The hearing-aid speech perception index (haspi) version 2. *Speech Communication*, 131 :35–46.

Kinoshita, K., Delcroix, M., Nakatani, T., and Miyoshi, M. (2009). Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. *IEEE transactions on audio, speech, and language processing*, 17(4) :534–545.

Koizumi, Y., Yatabe, K., Delcroix, M., Masuyama, Y., and Takeuchi, D. (2020). Speech enhancement using self-adaptation and multi-head self-attention. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 181–185. IEEE.

Kolbæk, M., Tan, Z.-H., and Jensen, J. (2016). Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1) :153–167.

Kolbæk, M., Tan, Z.-H., Jensen, S. H., and Jensen, J. (2020). On loss functions for supervised monaural time-domain speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28 :825–838.

Kolbæk, M., Yu, D., Tan, Z.-H., and Jensen, J. (2017). Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10) :1901–1913.

Koutrouvelis, A. I., Sherson, T. W., Heusdens, R., and Hendriks, R. C. (2018). A low-cost robust distributed linearly constrained beamformer for wireless acoustic sensor networks with arbitrary topology. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(8) :1434–1448.

Kowalski, M., Vincent, E., and Gribonval, R. (2010). Beyond the narrowband approximation : Wideband convex methods for under-determined reverberant audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7) :1818–1829.

Koyama, Y., Vuong, T., Uhlich, S., and Raj, B. (2020). Exploring the best loss function for dnn-based low-latency speech enhancement with temporal convolutional networks. *arXiv preprint arXiv :2005.11611*.

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. (2019). Quantifying the carbon emissions of machine learning. *Workshop on Tackling Climate Change with Machine Learning at NeurIPS 2019*.

Lagrange, M. and Marchand, S. (2001). Real-time additive synthesis of sound by taking advantage of psychoacoustics. In *Proceedings of the Digital Audio Effects (DAFx01) Conference*, pages 249–258.

Lan, T., Lyu, Y., Hui, G., Mokhosi, R., Li, S., and Liu, Q. (2020a). Redundant convolutional network with attention mechanism for monaural speech enhancement. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6654–6658. IEEE.

Lan, T., Lyu, Y., Ye, W., Hui, G., Xu, Z., and Liu, Q. (2020b). Combining multi-perspective attention mechanism with convolutional networks for monaural speech enhancement. *IEEE Access*, 8 :78979–78991.

Lawin-Ore, T. C. and Doclo, S. (2012). Reference microphone selection for mwf-based noise reduction using distributed microphone arrays. In *Speech Communication ; 10. ITG Symposium*, pages 1–4. VDE.

Le Roux, J., Hershey, J. R., and Weninger, F. (2015). Deep NMF for speech separation. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 66–70. IEEE.

Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J. R. (2019). SDR – half-baked or well done ? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 626–630. IEEE.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755) :788–791.

Leglaive, S., Girin, L., and Horaud, R. (2018). A variance modeling framework based on variational autoencoders for speech enhancement. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

Leglaive, S., Girin, L., and Horaud, R. (2019). Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 101–105. IEEE.

Li, G., Liang, S., Nie, S., Liu, W., and Yang, Z. (2021). Deep neural network-based generalized sidelobe canceller for dual-channel far-field speech recognition. *Neural Networks*, 141 :225–237.

Li, H., Zhang, X., and Gao, G. (2020a). Beamformed feature for learning-based dual-channel speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4722–4726. IEEE.

Li, X., Girin, L., Gannot, S., and Horaud, R. (2016). Non-stationary noise power spectral density estimation based on regional statistics. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 181–185. IEEE.

Li, X. and Horaud, R. (2019). Multichannel speech enhancement based on time-frequency masking using subband long short-term memory. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 298–302. IEEE.

Li, Y., Yang, M., and Zhang, Z. (2018). A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering*, 31(10) :1863–1883.

Li, Z., Yiu, K. F. C., Dai, Y.-H., and Nordholm, S. (2020b). Distributed LCMV beamformer design by randomly permuted ADMM. *Digital Signal Processing*, 106 :102820.

Liang, R., Kong, F., Xie, Y., Tang, G., and Cheng, J. (2020). Real-time speech enhancement algorithm based on attention LSTM. *IEEE Access*, 8 :48464–48476.

Liang, S., Liu, W., Jiang, W., and Xue, W. (2014). The analysis of the simplification from the ideal ratio to binary mask in signal-to-noise ratio sense. *Speech Communication*, 59 :22–30.

Liao, C.-F., Tsao, Y., Lee, H.-Y., and Wang, H.-M. (2019). Noise Adaptive Speech Enhancement Using Domain Adversarial Training. In *Proc. Interspeech 2019*, pages 3148–3152.

Lienhart, R., Kozintsev, I., Wehr, S., and Yeung, M. (2003). On the importance of exact synchronization for distributed audio signal processing. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 4, pages IV–840. IEEE.

Lim, J. (1978). Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(5) :471–472.

Liu, C.-L., Fu, S.-W., Li, Y.-J., Huang, J.-W., Wang, H.-M., and Tsao, Y. (2020). Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28 :1888–1900.

Liu, Y. and Wang, D. (2019). Divide and conquer : A deep casa approach to talker-independent monaural speaker separation. *IEEE/ACM Transactions on audio, speech, and language processing*, 27(12) :2092–2102.

Liu, Z., Zhang, Z., He, L.-W., and Chou, P. (2007). Energy-based sound source localization and gain normalization for ad hoc microphone arrays. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 2, pages II–761. IEEE.

Liutkus, A. and Badeau, R. (2015). Generalized wiener filtering with fractional power spectrograms. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 266–270. IEEE.

Llombart, J., Ribas, D., Miguel, A., Vicente, L., Ortega, A., and Lleida, E. (2021). Progressive loss functions for speech enhancement with deep neural networks. *EURASIP Journal on Audio, Speech, and Music Processing*, 2021(1) :1–16.

Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. In *In International Symposium on Music Information Retrieval*. Citeseer.

Loizou, P. C. (2007). *Speech enhancement : theory and practice*. CRC press.

Lopes, C. G. and Sayed, A. H. (2008). Diffusion least-mean squares over adaptive networks : Formulation and performance analysis. *IEEE Transactions on Signal Processing*, 56(7) :3122–3136.

Lorenzi, C., Berthommier, F., Apoux, F., and Bacri, N. (1999). Effects of envelope expansion on speech recognition. *Hearing research*, 136(1-2) :131–138.

Lostanlen, V., Salamon, J., Farnsworth, A., Kelling, S., and Bello, J. P. (2019). Robust sound event detection in bioacoustic sensor networks. *PloS one*, 14(10) :e0214168.

Lotter, T. and Vary, P. (2005). Speech enhancement by map spectral amplitude estimation using a super-gaussian speech model. *EURASIP Journal on Advances in Signal Processing*, 2005(7) :1–17.

Lottick, K., Susai, S., Friedler, S. A., and Wilson, J. P. (2019). Energy usage reports : Environmental awareness as part of algorithmic accountability. *Workshop on Tackling Climate Change with Machine Learning at NeurIPS 2019*.

Luo, Y., Chen, Z., Mesgarani, N., and Yoshioka, T. (2020a). End-to-end microphone permutation and number invariant multi-channel speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6394–6398. IEEE.

Luo, Y., Chen, Z., and Yoshioka, T. (2020b). Dual-path RNN : efficient long sequence modeling for time-domain single-channel speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 46–50. IEEE.

Luo, Y., Han, C., and Mesgarani, N. (2021). Ultra-lightweight speech separation via group communication. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16–20.

Luo, Y., Han, C., Mesgarani, N., Ceolini, E., and Liu, S.-C. (2019). FaSNet : Low-latency adaptive beamforming for multi-microphone audio processing. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 260–267. IEEE.

Luo, Y. and Mesgarani, N. (2019). Conv-tasnet : Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8) :1256–1266.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *Conference on Empirical Methods in Natural Language Processing*.

Maass, W. (1997). Networks of spiking neurons : the third generation of neural network models. *Neural networks*, 10(9) :1659–1671.

Macartney, C. and Weyde, T. (2018). Improved speech enhancement with the Wave-U-Net. *arXiv preprint arXiv :1811.11307*.

Maciejewski, M., Sell, G., Garcia-Perera, L. P., Watanabe, S., and Khudanpur, S. (2018). Building corpora for single-channel speech separation across multiple domains. *arXiv preprint arXiv :1811.02641*.

Madhu, N. and Martin, R. (2011). Low-complexity, robust algorithm for sensor anomaly detection and self-calibration of microphone arrays. *IET signal processing*, 5(1) :97–103.

Malik, M., Malik, M. K., Mehmood, K., and Makhdoom, I. (2021). Automatic speech recognition : a survey. *Multimedia Tools and Applications*, 80(6) :9411–9457.

Markovich-Golan, S., Bertrand, A., Moonen, M., and Gannot, S. (2015). Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks. *Signal Processing*, 107 :4–20.

Markovich-Golan, S., Gannot, S., and Cohen, I. (2012a). Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming. In *IWAENC 2012 ; International Workshop on Acoustic Signal Enhancement*, pages 1–4. VDE.

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. IEEE.

Martin, R. (1994). Spectral subtraction based on minimum statistics. *1994 European Signal Processing Conference (EUSIPCO)*, 6(8).

Martin, R. (2002). Speech enhancement using mmse short time spectral estimation with gamma distributed speech priors. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–253. IEEE.

Martín-Doñas, J. M., Gomez, A. M., López-Espejo, I., and Peinado, A. M. (2017). Dual-channel dnn-based speech enhancement for smartphones. In *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE.

Martín-Doñas, J. M., Jensen, J., Tan, Z.-H., Gomez, A. M., and Peinado, A. M. (2020). Online Multichannel Speech Enhancement Based on Recursive EM and DNN-Based Speech Presence Estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28 :3080–3094.

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.

McCowan, I., Lincoln, M., and Himawan, I. (2008). Microphone array shape calibration in diffuse noise fields. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(3) :666–670.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588) :746–748.

Mesaros, A., Heittola, T., and Virtanen, T. (2016). TUT database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1128–1132. IEEE.

Meyer, J. (2001). Beamforming for a circular microphone array mounted on spherically shaped objects. *The Journal of the Acoustical Society of America*, 109(1) :185–193.

Meyer, J. and Simmer, K. U. (1997). Multi-channel speech enhancement in a car environment using wiener filtering and spectral subtraction. In *1997 IEEE international conference on acoustics, speech, and signal processing*, volume 2, pages 1167–1170. IEEE.

Mishra, R., Gupta, H. P., and Dutta, T. (2020). A survey on deep neural network compression : Challenges, overview, and solutions. *arXiv preprint arXiv :2010.03954*.

Miyabe, S., Ono, N., and Makino, S. (2013). Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in stft domain. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 674–678. IEEE.

Mohammadiha, N., Martin, R., and Leijon, A. (2013). Spectral domain speech enhancement using hmm state-dependent super-gaussian priors. *IEEE Signal Processing Letters*, 20(3) :253–256.

Moore, B. C. (2012). *An introduction to the psychology of hearing*. Brill.

Moore, B. C. and Glasberg, B. R. (1996). A revision of zwicker's loudness model. *Acta Acustica united with Acustica*, 82(2) :335–345.

Musluoglu, C. A. and Bertrand, A. (2021). Distributed adaptive trace ratio optimization in wireless sensor networks. *IEEE Transactions on Signal Processing*.

Mydlarz, C., Shamoon, C., and Bello, J. P. (2017). Noise monitoring and enforcement in new york city using a remote acoustic sensor network. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, volume 255, pages 5509–5520. Institute of Noise Control Engineering.

Naithani, G., Barker, T., Parascandolo, G., Bramsl, L., Pontoppidan, N. H., Virtanen, T., et al. (2017). Low latency sound source separation using convolutional recurrent neural networks. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 71–75. IEEE.

Nakano, M., Le Roux, J., Kameoka, H., Kitano, Y., Ono, N., and Sagayama, S. (2010). Non-negative matrix factorization with markov-chained bases for modeling time-varying patterns in music spectrograms. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 149–156. Springer.

Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., and Juang, B.-H. (2010). Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7) :1717–1731.

Narayanan, A. and Wang, D. (2013). Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Naylor, P. A. and Gaubitch, N. D. (2010). *Speech dereverberation*. Springer Science & Business Media.

Neri, J., Badeau, R., and Depalle, P. (2021). Unsupervised blind source separation with variational auto-encoders. In *29th European Signal Processing Conference (EUSIPCO 2021)*.

Nian, Z., Tu, Y.-H., Du, J., and Lee, C.-H. (2021). A progressive learning approach to adaptive noise and speech estimation for speech enhancement and noisy speech recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6913–6917.

Nicolson, A. and Paliwal, K. K. (2019). Deep learning for minimum mean-square error approaches to speech enhancement. *Speech Communication*, 111 :44–55.

Nicolson, A. and Paliwal, K. K. (2020). Masked multi-head self-attention for causal speech enhancement. *Speech Communication*, 125 :80–96.

Ning, Y., He, S., Wu, Z., Xing, C., and Zhang, L.-J. (2019). A review of deep learning based speech synthesis. *Applied Sciences*, 9(19) :4050.

Nugraha, A. A., Liutkus, A., and Vincent, E. (2016). Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9) :1652–1664.

Oak, P. and Kellermann, W. (2005). A calibration algorithm for robust generalized sidelobe cancelling beamformers. In *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 97–100. Citeseer.

Ochiai, T., Delcroix, M., Ikeshita, R., Kinoshita, K., Nakatani, T., and Araki, S. (2020). Beamtasnet : Time-domain audio separation network meets frequency-domain beamformer. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6384–6388. IEEE.

O'Connor, M. and Kleijn, W. B. (2014). Diffusion-based distributed MVDR beamformer. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 810–814. IEEE.

Ozerov, A. and Févotte, C. (2009). Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3) :550–563.

Paliwal, K., Wójcicki, K., and Shannon, B. (2011). The importance of phase in speech enhancement. *speech communication*, 53(4) :465–494.

Pan, Z., Chua, Y., Wu, J., Zhang, M., Li, H., and Ambikairajah, E. (2020). An efficient and perceptually motivated auditory neural encoding and decoding algorithm for spiking neural networks. *Frontiers in neuroscience*, 13 :1420.

Pan, Z., Li, H., Wu, J., and Chua, Y. (2018). An event-based cochlear filter temporal encoding scheme for speech signals. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech : an ASR corpus based on public domain audio books. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Pandey, A. and Wang, D. (2019a). Exploring deep complex networks for complex spectrogram enhancement. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6885–6889. IEEE.

Pandey, A. and Wang, D. (2019b). TCNN : Temporal convolutional neural network for real-time speech enhancement in the time domain. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6875–6879. IEEE.

Pandey, A. and Wang, D. (2021). Self-attending RNN for speech enhancement to improve cross-corpus generalization. *arXiv preprint arXiv :2105.12831*.

Pariente, M., Cornell, S., Cosentino, J., Sivasankaran, S., Tzinis, E., Heitkaemper, J., Olvera, M., Stöter, F.-R., Hu, M., Martín-Doñas, J. M., Ditter, D., Frank, A., Deleforge, A., and Vincent, E. (2020a). Asteroid : The PyTorch-Based Audio Source Separation Toolkit for Researchers. In *Proc. Interspeech 2020*, pages 2637–2641.

Pariente, M., Cornell, S., Deleforge, A., and Vincent, E. (2020b). Filterbank design for end-to-end speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6364–6368. IEEE.

Pariente, M., Deleforge, A., and Vincent, E. (2019). A Statistically Principled and Computationally Efficient Approach to Speech Enhancement Using Variational Autoencoders. In *Proc. Interspeech 2019*, pages 3158–3162.

Park, S. R. and Lee, J. (2016). A fully convolutional neural network for speech enhancement. *arXiv preprint arXiv :1609.07132*.

Parry, R. M. and Essa, I. (2007). Incorporating phase information for source separation via spectrogram factorization. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 2, pages II–661. IEEE.

Pascual, S., Bonafonte, A., and Serra, J. (2017). Segan : Speech enhancement generative adversarial network. *arXiv preprint arXiv :1703.09452*.

Perotin, L., Serizel, R., Vincent, E., and Guérin, A. (2018). Multichannel speech separation with recurrent neural networks from high-order ambisonics recordings. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 36–40.

Perotin, L., Serizel, R., Vincent, E., and Guerin, A. (2019). Crnn-based multiple doa estimation using acoustic intensity features for ambisonics recordings. *IEEE Journal of Selected Topics in Signal Processing*, 13(1) :22–33.

Pfeifenberger, L., Zöhrer, M., and Pernkopf, F. (2019). Deep complex-valued neural beamformers. pages 2902–2906.

Picaut, J., Can, A., Fortin, N., Ardouin, J., and Lagrange, M. (2020). Low-cost sensors for urban noise monitoring networks—a literature review. *Sensors*, 20(8).

Plinge, A., Jacob, F., Haeb-Umbach, R., and Fink, G. A. (2016). Acoustic microphone geometry calibration : An overview and experimental evaluation of state-of-the-art algorithms. *IEEE Signal Processing Magazine*, 33(4) :14–29.

Plomp, R. and Mimpen, A. (1979). Speech-reception threshold for sentences as a function of age and noise level. *The Journal of the Acoustical Society of America*, 66(5) :1333–1342.

Qian, K., Zhang, Y., Chang, S., Yang, X., Florencio, D., and Hasegawa-Johnson, M. (2018). Deep learning based speech beamforming. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5389–5393. IEEE.

Qin, S. and Jiang, T. (2018). Improved wasserstein conditional generative adversarial network speech enhancement. *EURASIP Journal on Wireless Communications and Networking*, 2018(1) :1–10.

Rafaely, B., Peled, Y., Agmon, M., Khaykin, D., and Fisher, E. (2010). Spherical microphone array beamforming. *Speech Processing in Modern Communication*, pages 281–305.

Rajan, R. T. and van der Veen, A.-J. (2011). Joint ranging and clock synchronization for a wireless network. In *2011 4th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 297–300. IEEE.

Ranjbaryan, R. and Abutalebi, H. R. (2020). Distributed speech presence probability estimator in fully connected wireless acoustic sensor networks. *Circuits, Systems, and Signal Processing*, 39 :6121–6141.

Ranjbaryan, R. and Abutalebi, H. R. (2021). Multiframe maximum a posteriori estimators for single-microphone speech enhancement. *IET Signal Processing*.

Ranjbaryan, R., Doclo, S., and Abutalebi, H. R. (2018). Distributed MAP estimators for noise reduction in fully connected wireless acoustic sensor networks. In *Speech Communication ; 13th ITG-Symposium*, pages 1–5. VDE.

Raykar, V. C., Kozintsev, I. V., and Lienhart, R. (2004). Position calibration of microphones and loudspeakers in distributed computing platforms. *IEEE transactions on Speech and Audio Processing*, 13(1) :70–83.

Reddy, C. K., Beyrami, E., Dubey, H., Gopal, V., Cheng, R., Cutler, R., Matusevych, S., Aichner, R., Aazami, A., Braun, S., et al. (2020). The interspeech 2020 deep noise suppression challenge : Datasets, subjective speech quality and testing framework. *arXiv preprint arXiv :2001.08662*.

Reddy, C. K., Gopal, V., and Cutler, R. (2021). Dnsmos : A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6493–6497. IEEE.

Rethage, D., Pons, J., and Serra, X. (2018). A wavenet for speech denoising. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5069–5073. IEEE.

Rouat, J., Pichevar, R., and Loiselle, S. (2004). Perceptive, non-linear speech processing and spiking neural networks. In *International School on Neural Networks, Initiated by IIASS and EMFCSC*, pages 317–337. Springer.

Roy, A. G., Navab, N., and Wachinger, C. (2018). Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In *International conference on medical image computing and computer-assisted intervention*, pages 421–429. Springer.

Roy, S. K., Nicolson, A., and Paliwal, K. K. (2021). Deeplpc : A deep learning approach to augmented kalman filter-based single-channel speech enhancement. *IEEE Access*, 9 :64524–64538.

Sachar, J. M., Silverman, H. F., and Patterson, W. R. (2004). Microphone position and gain calibration for a large-aperture microphone array. *IEEE Transactions on Speech and Audio Processing*, 13(1) :42–52.

Sadeghi, M., Leglaive, S., Alameda-Pineda, X., Girin, L., and Horaud, R. (2020). Audio-visual speech enhancement using conditional variational auto-encoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28 :1788–1800.

Sainath, T. N., Weiss, R. J., Wilson, K. W., Li, B., Narayanan, A., Variani, E., Bacchiani, M., Shafran, I., Senior, A., Chin, K., Misra, A., and Kim, C. (2017). Multichannel signal processing with deep neural networks for automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(5) :965–979.

Saito, K., Uhlich, S., Fabbro, G., and Mitsufuji, Y. (2021). Training speech enhancement systems with noisy speech datasets. *arXiv preprint arXiv :2105.12315*.

Santos, J. F., Cosentino, S., Hazrati, O., Loizou, P. C., and Falk, T. H. (2013). Objective speech intelligibility measurement for cochlear implant users in complex listening environments. *Speech Communication*, 55(7) :815–824.

Scalart, P. et al. (1996). Speech enhancement based on a priori signal to noise estimation. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 2, pages 629–632. IEEE.

Scheibler, R., Bezzam, E., and Dokmanic, I. (2018). Pyroomacoustics : A python package for audio room simulation and array processing algorithms. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Schenato, L. and Fiorentin, F. (2011). Average TimeSynch : A consensus-based protocol for clock synchronization in wireless sensor networks. *Automatica*, 47(9) :1878–1886.

Schmalenstroeer, J. and Haeb-Umbach, R. (2018a). Efficient sampling rate offset compensation-an overlap-save based approach. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 499–503. IEEE.

Schmalenstroeer, J. and Haeb-Umbach, R. (2018b). Insights into the interplay of sampling rate offsets and MVDR beamforming. In *Speech Communication ; 13th ITG-Symposium*, pages 1–5. VDE.

Schmalenstroeer, J., Heymann, J., Drude, L., Boeddecker, C., and Haeb-Umbach, R. (2017). Multi-stage coherence drift based sampling rate synchronization for acoustic beamforming. In *2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE.

Schmalenstroeer, J., Jebramcik, P., and Haeb-Umbach, R. (2015). A combined hardware–software approach for acoustic sensor network synchronization. *Signal Processing*, 107 :171–184.

Schulze-Forster, K., Doire, C. S. J., Richard, G., and Badeau, R. (2021). Phoneme level lyrics alignment and text-informed singing voice separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Sekiguchi, K., Bando, Y., Nugraha, A. A., Yoshii, K., and Kawahara, T. (2019). Semi-supervised multichannel speech enhancement with a deep speech prior. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12) :2197–2212.

Seltzer, M. L., Yu, D., and Wang, Y. (2013). An investigation of deep neural networks for noise robust speech recognition. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 7398–7402. IEEE.

Seok, J. W. and Bae, K. S. (1997). Speech enhancement with reduction of noise components in the wavelet domain. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1323–1326. IEEE.

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.

Serrà, J., Pons, J., and Pascual, S. (2021). Sesqa : semi-supervised learning for speech quality assessment. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 381–385. IEEE.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3) :379–423.

Sharma, S., Sharma, A., Malhotra, R., and Rattan, P. (2021). Voice activity detection using windowing and updated k-means clustering algorithm. In *2021 2nd International Conference on Intelligent Engineering and Management (ICIEM)*, pages 114–118. IEEE.

Sherson, T., Kleijn, W. B., and Heusdens, R. (2016). A distributed algorithm for robust LCMV beamforming. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 101–105. IEEE.

Shi, Y., Huang, Q., and Hain, T. (2020). Robust speaker recognition using speech enhancement and attention model. *arXiv preprint arXiv :2001.05031*.

Simmer, K. U., Bitzer, J., and Marro, C. (2001). Post-filtering techniques. In *Microphone arrays*, pages 39–60. Springer.

Souden, M., Araki, S., Kinoshita, K., Nakatani, T., and Sawada, H. (2013). A multichannel MMSE-based framework for speech source separation and noise reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(9) :1913–1928.

Souden, M., Benesty, J., and Affes, S. (2010). On optimal frequency-domain multichannel linear filtering for noise reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2) :260–276.

Souden, M., Chen, J., Benesty, J., and Affes, S. (2011). An integrated solution for online multichannel noise tracking and reduction. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7) :2159–2169.

Srinivasan, S. and Den Brinker, A. C. (2009). Analyzing rate-constrained beamforming schemes in wireless binaural hearing aids. In *2009 17th European Signal Processing Conference*, pages 1854–1858. IEEE.

Srinivasan, S., Roman, N., and Wang, D. (2006). Binary and ratio time-frequency masks for robust speech recognition. *Speech Communication*, 48(11) :1486–1501.

Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3) :185–190.

Stöter, F.-R., Liutkus, A., and Ito, N. (2018). The 2018 signal separation evaluation campaign. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 293–305. Springer.

Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., and Zhong, J. (2021). Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21–25. IEEE.

Szurley, J., Bertrand, A., Moerman, I., and Moonen, M. (2013). Network topology selection for distributed speech enhancement in wireless acoustic sensor networks. In *21st European Signal Processing Conference (EUSIPCO 2013)*, pages 1–5. IEEE.

Szurley, J., Bertrand, A., and Moonen, M. (2016). Topology-independent distributed adaptive node-specific signal estimation in wireless sensor networks. *IEEE Transactions on Signal and Information Processing over Networks*, 3(1) :130–144.

Szurley, J., Bertrand, A., Moonen, M., Ruckebusch, P., and Moerman, I. (2011). Utility based cross-layer collaboration for speech enhancement in wireless acoustic sensor networks. In *2011 19th European Signal Processing Conference*, pages 235–239. IEEE.

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. IEEE.

Taghizadeh, M. J., Parhizkar, R., Garner, P. N., Bourlard, H., and Asaei, A. (2015). Ad hoc microphone array calibration : Euclidean distance matrix completion algorithm and theoretical guarantees. *Signal Processing*, 107 :123–140.

Taherian, H., Wang, Z.-Q., Chang, J., and Wang, D. (2020). Robust speaker recognition based on single-channel and multi-channel speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28 :1293–1302.

Takeuchi, D., Yatabe, K., Koizumi, Y., Oikawa, Y., and Harada, N. (2020). Invertible DNN-based nonlinear time-frequency transform for speech enhancement. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6644–6648. IEEE.

Tammen, M. and Doclo, S. (2021). Deep multi-frame MVDR filtering for single-microphone speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8443–8447. IEEE.

Tamura, S. and Waibel, A. (1988). Noise reduction using connectionist models. In *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, pages 553–556. IEEE.

Tan, K. and Wang, D. (2018). A convolutional recurrent neural network for real-time speech enhancement. In *Interspeech*, pages 3229–3233.

Tan, K. and Wang, D. (2019). Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28 :380–390.

Tan, K. and Wang, D. (2021). Compressing deep neural networks for efficient speech enhancement. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8358–8362. IEEE.

Tan, K., Xu, Y., Zhang, S.-X., Yu, M., and Yu, D. (2020). Audio-visual speech separation and dereverberation with a two-stage multimodal network. *IEEE Journal of Selected Topics in Signal Processing*, 14(3) :542–553.

Tan, K., Zhang, X., and Wang, D. (2021a). Deep learning based real-time speech enhancement for dual-microphone mobile phones. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29 :1853–1863.

Tan, K., Zhang, X., and Wang, D. (2021b). Real-time speech enhancement for mobile communication based on dual-channel complex spectral mapping. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6134–6138. IEEE.

Tan, X., Qin, T., Soong, F., and Liu, T.-Y. (2021c). A survey on neural speech synthesis. *arXiv preprint arXiv :2106.15561*.

Tanyer, S. G. and Ozer, H. (2000). Voice activity detection in nonstationary noise. *IEEE Transactions on speech and audio processing*, 8(4) :478–482.

Tavakoli, V. M., Jensen, J. R., Heusdens, R., Benesty, J., and Christensen, M. G. (2016). Ad-hoc microphone array beamforming using the primal-dual method of multipliers. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1088–1092. IEEE.

Tavakoli, V. M., Jensen, J. R., Heusdens, R., Benesty, J., and Christensen, M. G. (2017). Distributed max-SINR speech enhancement with ad hoc microphone arrays. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155. IEEE.

Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T., and Maida, A. (2019). Deep learning in spiking neural networks. *Neural Networks*, 111 :47–63.

Taylor, T. (1952). A synthesis method for circular and cylindrical antennas composed of discrete elements. *Transactions of the IRE Professional Group on Antennas and Propagation*, pages 251–261.

Togami, M. (2019). Multi-channel itakura saito distance minimization with deep neural network. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 536–540. IEEE.

Tolooshams, B., Giri, R., Song, A. H., Isik, U., and Krishnaswamy, A. (2020). Channel-attention dense u-net for multichannel speech enhancement. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 836–840. IEEE.

Tribolet, J., Noll, P., McDermott, B., and Crochiere, R. (1978). A study of complexity and quality of speech waveform coders. In *ICASSP'78. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 586–590. IEEE.

Tu, Y.-H., Du, J., and Lee, C.-H. (2019). Speech enhancement based on teacher–student deep learning using improved speech presence probability for noise-robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12) :2080–2091.

Turpault, N., Serizel, R., and Vincent, E. (2019). Semi-supervised triplet loss based learning of ambient audio embeddings. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 760–764. IEEE.

Tzinis, E., Wang, Z., and Smaragdis, P. (2020). Sudo rm-rf : Efficient networks for universal audio source separation. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

Vadwala, A. Y., Suthar, K. A., Karmakar, Y. A., and Pandya, N. (2017). Survey paper on different speech recognition algorithm : challenges and techniques. *Int J Comput Appl*, 175(1) :31–36.

Valin, J.-M. (2018). A hybrid DSP/deep learning approach to real-time full-band speech enhancement. In *2018 IEEE 20th international workshop on multimedia signal processing (MMSP)*, pages 1–5. IEEE.

Valin, J.-M., Isik, U., Phansalkar, N., Giri, R., Helwani, K., and Krishnaswamy, A. (2020). A Perceptually-Motivated Approach for Low-Complexity, Real-Time Enhancement of Fullband Speech. In *Proc. Interspeech 2020*, pages 2482–2486.

Van Compernolle, D., Ma, W., Xie, F., and Van Diest, M. (1990). Speech recognition in noisy environments with the aid of microphone arrays. *Speech Communication*, 9(5-6) :433–442.

Van Trees, H. L. (2004). *Detection, estimation, and modulation theory, part I : detection, estimation, and linear modulation theory.* John Wiley & Sons.

Van Veen, B. D. and Buckley, K. M. (1988). Beamforming : A versatile approach to spatial filtering. *IEEE ASSP magazine*, 5(2) :4–24.

Varga, A. and Moore, R. K. (1990). Hidden markov model decomposition of speech and noise. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 845–848. IEEE.

Vaseghi, S. V. (2008). *Advanced digital signal processing and noise reduction.* John Wiley & Sons.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vincent, E., Gribonval, R., and Févotte, C. (2006). Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4) :1462–1469.

Vincent, E., Virtanen, T., and Gannot, S. (2018). *Audio source separation and speech enhancement*. John Wiley & Sons.

Virag, N. (1999). Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Transactions on speech and audio processing*, 7(2) :126–137.

Virtanen, T. (2007). Monaural sound source separation by perceptually weighted non-negative matrix factorization. *Tampere University of Technology, Tech. Rep.*

Wang, D. (2005). On ideal binary mask as the computational goal of auditory scene analysis. In *Speech separation by humans and machines*, pages 181–197. Springer.

Wang, D., Chen, Z., and Yoshioka, T. (2020). Neural Speech Separation Using Spatially Distributed Microphones. In *Proc. Interspeech 2020*, pages 339–343.

Wang, D. L. and Brown, G. J. (1999). Separation of speech from interfering sounds based on oscillatory correlation. *IEEE transactions on neural networks*, 10(3) :684–697.

Wang, L. and Doclo, S. (2016). Correlation maximization-based sampling rate offset estimation for distributed microphone arrays. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3) :571–582.

Wang, L., Zhu, J., and Kodrasi, I. (2021a). Multi-task single channel speech enhancement using speech presence probability as a secondary task training target. *arXiv preprint arXiv :2011.07547*.

Wang, R., Chen, Z., and Yin, F. (2021b). Distributed frequency response calibration based on consensus strategy in microphone array. *IEEE Transactions on Instrumentation and Measurement*, 70 :1–12.

Wang, Y., Han, K., and Wang, D. (2012). Exploring monaural features for classification-based speech segregation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(2) :270–279.

Wang, Z.-Q. and Wang, D. (2018). Mask weighted STFT ratios for relative transfer function estimation and its application to robust ASR. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5619–5623. IEEE.

Warsitz, E. and Haeb-Umbach, R. (2007). Blind acoustic beamforming based on generalized eigenvalue decomposition. *IEEE Transactions on audio, speech, and language processing*, 15(5) :1529–1539.

Wehr, S., Kozintsev, I., Lienhart, R., and Kellermann, W. (2004). Synchronization of acoustic sensors for distributed ad-hoc audio networks and its use for blind source separation. In *IEEE Sixth International Symposium on Multimedia Software Engineering*, pages 18–25. IEEE.

Weninger, F., Hershey, J. R., Le Roux, J., and Schuller, B. (2014). Discriminatively trained recurrent neural networks for single-channel speech separation. In *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 577–581. IEEE.

Westerlund, N., Dahl, M., and Claesson, I. (2005). Speech enhancement for personal communication using an adaptive gain equalizer. *Signal processing*, 85(6) :1089–1101.

Widrow, B., Mantey, P., Griffiths, L., and Goode, B. (1967). Adaptive antenna systems. *Proceedings of the IEEE*, 55(12) :2143–2159.

Williamson, D. S., Wang, Y., and Wang, D. (2016). Complex ratio masking for joint enhancement of magnitude and phase. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224. IEEE.

Wisdom, S., Tzinis, E., Erdogan, H., Weiss, R., Wilson, K., and Hershey, J. (2020). Unsupervised sound separation using mixture invariant training. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3846–3857. Curran Associates, Inc.

Wolfe, J., Neumann, S., Schafer, E., Towler, W., Miller, S., Dunn, A., Jones, C., and Nelson, J. (2021). Evaluation of a dual adaptive remote microphone system. *Journal of Educational, Pediatric & (Re)Habilitative Audiology*, (25).

Wolfe, P. J. and Godsill, S. J. (2003). Efficient alternatives to the ephraim and malah suppression rule for audio signal enhancement. *EURASIP Journal on Advances in Signal Processing*, 2003(10) :1–9.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). CBAM : Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

Woodbury, M. A. (1950). *Inverting modified matrices*. Statistical Research Group.

Wu, J., Chua, Y., Zhang, M., Li, H., and Tan, K. C. (2018). A spiking neural network framework for robust sound classification. *Frontiers in neuroscience*, 12 :836.

Wu, J., Yılmaz, E., Zhang, M., Li, H., and Tan, K. C. (2020). Deep spiking neural networks for large vocabulary automatic speech recognition. *Frontiers in neuroscience*, 14 :199.

Xia, S., Li, H., and Zhang, X. (2017). Using optimal ratio mask as training target for supervised speech separation. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 163–166. IEEE.

Xia, W. and Koishida, K. (2019). Sound Event Detection in Multichannel Audio Using Convolutional Time-Frequency-Channel Squeeze and Excitation. In *Proc. Interspeech 2019*, pages 3629–3633.

Xia, Y., Xu, B., and Kumar, A. (2021). Incorporating real-world noisy speech in neural-network-based speech enhancement systems. *arXiv preprint arXiv :2109.05172*.

Xie, F. and Van Compernolle, D. (1996). Speech enhancement by spectral magnitude estimation—a unifying approach. *Speech Communication*, 19(2) :89–104.

Xu, C., Tao, D., and Xu, C. (2013). A survey on multi-view learning. *arXiv preprint arXiv :1304.5634*.

Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H. (2014). A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1) :7–19.

Xue, C., Huang, W., Chen, W., and Feng, J. (2021). Real-Time Multi-Channel Speech Enhancement Based on Neural Network Masking with Attention Model. In *Proc. Interspeech 2021*, pages 1862–1866.

Yilmaz, O. and Rickard, S. (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on signal processing*, 52(7) :1830–1847.

Yoshioka, T., Erdogan, H., Chen, Z., and Alleva, F. (2018). Multi-microphone neural speech separation for far-field multi-talker speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5739–5743.

Yoshioka, T., Nakatani, T., and Miyoshi, M. (2009). Integrated speech enhancement method using noise suppression and dereverberation. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(2) :231–246.

Zão, L., Coelho, R., and Flandrin, P. (2014). Speech enhancement with emd and hurst-based mode selection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(5) :899–911.

Zeghidour, N. and Grangier, D. (2020). Wavesplit : End-to-end speech separation by speaker clustering. *arXiv preprint arXiv :2002.08933*.

Zeng, Y. and Hendriks, R. C. (2014). Distributed delay and sum beamformer for speech enhancement via randomized gossip. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1) :260–273.

Zeng, Y. and Hendriks, R. C. (2015). Distributed estimation of the inverse of the correlation matrix for privacy preserving beamforming. *Signal Processing*, 107 :109–122.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2019). Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR.

Zhang, J., Chen, H., Dai, L.-R., and Hendriks, R. C. (2020). A study on reference microphone selection for multi-microphone speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29 :671–683.

Zhang, J., Heusdens, R., and Hendriks, R. C. (2018). Rate-distributed spatial filtering based noise reduction in wireless acoustic sensor networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11) :2015–2026.

Zhang, K., Li, X. R., Zhang, P., Li, H., et al. (2003). Optimal linear estimation fusion—part vi : Sensor data compression. In *Proc. Int. Conf. Information Fusion*, volume 23, page 221. Citeseer.

Zhang, S. and Li, X. (2021). Microphone array generalization for multichannel narrowband deep speech enhancement. *arXiv preprint arXiv :2107.12601*.

Zhang, X.-L. and Wu, J. (2012). Deep belief networks based voice activity detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(4) :697–710.

Zhao, S., Sugitani, Y., and Miyajima, T. (2021). Distributed minimum variance equalization in wireless sensor networks. *Nonlinear Theory and Its Applications, IEICE*, 12(3) :442–452.

Zhao, X., Chen, J., and Sayed, A. H. (2012). Beam coordination via diffusion adaptation over array networks. In *2012 IEEE 13th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 105–109. IEEE.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (frequenzgruppen). *The Journal of the Acoustical Society of America*, 33(2) :248–248.