Apprentissage profond bout-en-bout pour le rehaussement de la parole

Guillaume Carbajal

UNIVERSITÉ DE LORRAINE

Thèse de Doctorat

Guillaume CARBAJAL

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 24 avril 2020
Thèse N° :

Apprentissage profond bout-en-bout pour le rehaussement de la parole

JURY

Rapporteur :	**Jean-François BONASTRE**, Professeur, Avignon Université, Avignon, France
Rapporteur :	**Abdeldjalil AÏSSA EL BEY**, Professeur, IMT Atlantique, Brest, France
Examinatrice :	**Ann SPRIET**, Ingénieure, GOODiX Technology, Leuven, Belgique
Examinateur :	**François CHARPILLET**, Directeur de Recherche, Inria Nancy - Grand Est, France
Examinatrice :	**Nancy BERTIN**, Chargée de Recherche, CNRS, Rennes, France
Directeur de thèse :	**Emmanuel VINCENT**, Directeur de recherche, Inria Nancy - Grand Est, France
Co-directeur de thèse :	**Romain SERIZEL**, Maître de conférences, Université de Lorraine, Nancy, France
Co-encadrant de thèse :	**Éric HUMBERT**, Ingénieur de recherche, Invoxia, Issy-les-Moulineaux, France

Résumé

Cette thèse s'insère dans le développement des systèmes de télécommunication mains-libres, en particulier avec des enceintes intelligentes en environnement domestique. L'utilisateur interagit avec un correspondant distant en étant généralement situé à quelques mètres de ce type de système. Les microphones sont susceptibles de capter des sons de l'environnement qui se mêlent à la voix de l'utilisateur, comme le bruit ambiant, l'écho acoustique et la réverbération. Ces types de distorsions peuvent gêner fortement l'écoute et la compréhension de la conversation par le correspondant distant, et il est donc nécessaire de les réduire. Des méthodes de filtrage existent pour réduire individuellement chacun de ces types de distorsion sonore, et leur réduction simultanée implique de combiner ces méthodes. Toutefois, celles-ci interagissent entre elles, et leurs interactions peuvent dégrader de la voix de l'utilisateur. Il est donc nécessaire d'optimiser conjointement ces méthodes.

En premier lieu, nous présentons une approche de réduction de l'écho acoustique combinant un filtre d'annulation d'écho avec un post-filtre de suppression d'écho résiduel conçu de manière à s'adapter à différents modes de fonctionnement du filtre d'annulation. Pour cela, nous proposons d'estimer les coefficients du post-filtre en utilisant les spectres à court terme de plusieurs signaux observés, dont le signal estimé par le filtre d'annulation, en entrée d'un réseau de neurones. Nous montrons que cette approche améliore la performance et la robustesse du post-filtre en matière de réduction d'écho, tout en limitant la dégradation de la parole de l'utilisateur, sur plusieurs scénarios dans des conditions réelles.

En second lieu, nous décrivons une approche conjointe de réduction multicanale de l'écho, de la réverbération et du bruit. Nous proposons de modéliser simultanément la parole cible et les signaux résiduels après annulation d'écho et déréverbération dans un cadre probabiliste et de représenter conjointement leurs spectres à court terme à l'aide d'un réseau de neurones récurrent. Nous intégrons cette modélisation dans un algorithme de montée par blocs de coordonnées pour mettre à jour les filtres d'annulation d'écho et de déréverbération, ainsi que le post-filtre de suppression des signaux résiduels. Nous évaluons notre approche sur des enregistrements réels dans différentes conditions. Nous montrons qu'elle améliore la qualité de la parole ainsi que la réduction de l'écho, de la réverbération et du bruit, par rapport à une approche optimisant séparément les méthodes de filtrage et une autre approche de réduction conjointe.

En dernier lieu, nous formulons une version en ligne de notre approche adaptée aux situations où les conditions acoustiques varient dans le temps. Nous évaluons la qualité perceptuelle sur des exemples réels où l'utilisateur se déplace durant la conversation.

Abstract

This PhD falls within the development of hands-free telecommunication systems, more specifically smart speakers in domestic environments. The user interacts with another speaker at a far-end point and can be typically a few meters away from this kind of system. The microphones are likely to capture sounds of the environment which are added to the user's voice, such background noise, acoustic echo and reverberation. These types of distortion degrade speech quality, intelligibility and listening comfort for the far-end speaker, and must be reduced. Filtering methods can reduce individually each of these types of distortion. Reducing all of them implies combining the corresponding filtering methods. As these methods interact with each other which can deteriorate the user's speech, they must be jointly optimized.

First of all, we introduce an acoustic echo reduction approach which combines an echo cancellation filter with a residual echo postfilter designed to adapt to the echo cancellation filter. To do so, we propose to estimate the postfilter coefficients using the short term spectra of multiple known signals, including the output of the echo cancellation filter, as inputs to a neural network. We show that this approach improves the performance and the robustness of the postfilter in terms of echo reduction, while limiting speech degradation, on several scenarios in real conditions.

Secondly, we describe a joint approach for multichannel reduction of echo, reverberation and noise. We propose to simultaneously model the target speech and undesired residual signals after echo cancellation and dereveberation in a probabilistic framework, and to jointly represent their short-term spectra by means of a recurrent neural network. We develop a block-coordinate ascent algorithm to update the echo cancellation and dereverberation filters, as well as the postfilter that reduces the undesired residual signals. We evaluate our approach on real recordings in different conditions. We show that it improves speech quality and reduction of echo, reverberation and noise compared to a cascade of individual filtering methods and another joint reduction approach.

Finally, we present an online version of our approach which is suitable for time-varying acoustic conditions. We evaluate the perceptual quality achieved on real examples where the user moves during the conversation.

Remerciements

Je tiens à remercier tout d'abord mes directeurs de thèse, Emmanuel et Romain. En plus des connaissances techniques que vous m'avez apportées, vous êtes toujours restés disponibles pour m'aider à chaque fois que j'étais bloqué durant ma thèse. Je vous en suis très reconnaissant. Cela a représenté un soutien important qui m'a permis de rester concentré sur mes objectifs et de me dépasser. J'espère que l'on pourra faire un jour ce pot de thèse qui n'a pas pu avoir lieu à cause du coronavirus.

Je remercie ensuite la société Invoxia, et plus particulièrement Éric, pour la flexibilité et la liberté dont j'ai bénéficié pour les différents choix dans cette thèse.

Je remercie tous les gens avec qui j'ai travaillé quotidiennement ou durant certaines périodes de ma thèse. En particulier, Adam et Raphaël avec qui j'ai commencé la thèse, qui ont subi mes changements d'humeur permanents et mes réflexions sur le cas oracle. Aussi les personnes de Multispeech : Lauréline, Baldwin, Mathieu, Sunit, les deux Nicolas, Manu, Adrien, Élodie et Antoine. Vous avez vraiment rendu mon quotidien agréable, surtout dans les moments difficiles de cette thèse. S'ajoutent à eux, les doctorants et personnes liées à la recherche en dehors de Multispeech : Morgan, Vesna, Robin, Pierre et ce mofo de Gabor. Je tiens particulièrement à remercier Ngoc Duong que j'ai rencontré à ICASSP 2018, et qui m'a aidé à obtenir l'idée principale de ma thèse.

Je remercie mes amis non-chercheurs que j'ai régulièrement côtoyés tout au long de ma thèse. Vous m'avez apporté un énorme soutien moral. Je pense à la team des Ponts avec qui j'ai pu passer certaines vacances et quelques festivals de déglingue : Paul, Thibault, Igor, Sam, Thomas, Moukak et Nono. Je pense aussi à la bande de potes dromadaires/historiens, avec qui j'ai notamment passé un bon mariage avant de déposer le manuscrit : Gautier, Agathe, Pierre-Alain, Pauline, les deux Mathieu, Keyvan, Lauriane, Baptiste, les deux Nathalie, Bastien, Kevin, Caro, Clément, Maryline et Benoît. Je n'oublie pas la team Dimensions, mes anciens colocs, le collectif Colapso, Delphine et Camille, qui sont maintenant tous dispersés un peu partout.

Je tiens à remercier ma soeur Mélanie et mes parents, qui m'ont toujours soutenu pendant tout le long, et notamment pendant le confinement où j'ai dû préparer ma défense de thèse.

Beaucoup de thésards et thésardes remercient leur mec/meuf dans le dernier paragraphe des remerciements. Moi je termine cette thèse en bon célibataire, donc je vais simplement remercier l'amour de manière générale. Je verrai ce que l'avenir me réserve, et je vais m'arrêter là.

Note : ce manuscrit est rédigé en écriture alternée. Il sera indifféremment fait référence aux locuteurs ou aux locutrices, sans que cela n'ait de rapport avec le contenu scientifique lui-même.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

CNN convolutional neural network
DNN deep neural network
DSP densité spectrale de puissance
dB décibel
EM espérance-maximisation
ELR rapport précoce-à-tardif
ERLE echo return loss enhancement
FSP filtre sensible à la phase
LSTM long short-term memory
MCS matrice de covariance spatiale
MIA masque idéal d'amplitude
MIR masque idéal de rapport
MWF multichannel Wiener filter
MV maximum de vraisemblance
NMF non-negative matrix factorization
RNN recurrent neural network
SI-SAR rapport signal-à-artefacts invariant à l'échelle
SI-SDR rapport signal-à-distorsion invariant à l'échelle
SER rapport signal-à-écho
SNR rapport signal-à-bruit
$\mathbf{T}_{60}$ temps de réverbération à 60 dB
TFCT transformée de Fourier à court terme
WPE weighted prediction error

Liste des notations mathématiques

$\|\cdot\|$ norme euclidienne

$\mathbb{E}[\cdot]$ espérance mathématique

$(\cdot)^H$ transposée conjuguée d'une matrice ou d'un vecteur

$\widehat{\mathbb{E}}[\cdot]$ moyenne empirique d'une variable déterministe à un instant donné

$(\cdot)^T$ transposée d'une matrice ou d'un vecteur

$\mathrm{tr}(\cdot)$ trace d'une matrice carrée

$\mathbf{I}$ matrice identité, dont la dimension est déduite d'après le contexte

$\mathbf{I}_M$ matrice identité de dimension M

$\underline{(\cdot)}$ concaténation d'un vecteur ou d'une matrice sur plusieurs observations passées

$\otimes$ produit de Kronecker

1. Introduction

Ce chapitre décrit le contexte scientifique et industriel dans lequel s'inscrit cette thèse, en particulier l'émergence des enceintes intelligentes dans les systèmes de communication téléphonique mains-libres et les difficultés techniques soulevées par les nouveaux assistants personnels en environnement domestique. Nous définissons le cadre d'étude de la thèse et les objectifs fixés dans ce scénario. Enfin, nous détaillons l'organisation de ce document et présentons les principales contributions de la thèse.

1.1. Motivation et cadre

Télécommunications et rehaussement de la parole Lors d'une conversation téléphonique, il est aujourd'hui de plus en plus courant de vouloir communiquer à l'aide de systèmes mains-libres. Toutefois, il est rare que l'on se trouve dans un endroit silencieux. Dans le salon d'une maison, une personne passe l'aspirateur et discute en même temps par téléphone avec son ami. « Salut Paul, où es-tu ? - Guillaume ? Je suis chez toi dans 5 minutes. ». Des employés d'une entreprise à Hong-Kong font une réunion par visioconférence avec la filiale en Suisse tandis qu'une manifestation a lieu à l'extérieur. Une conductrice est coincée dans les bouchons, la fenêtre ouverte, et utilise le système mains-libres de sa voiture pour appeler sa famille. Une enceinte intelligente comme Amazon Echo joue de la musique dans une pièce et l'utilisateur, qui est dans la pièce à côté, demande à l'assistant personnel Alexa de changer de chanson. Tous ces scénarios comportent plusieurs points communs. D'une part, un locuteur local interagit avec un correspondant, humain (conversation téléphonique) ou virtuel (assistant personnel), à l'aide d'un système de télécommunication mains-libres. D'autre part, le locuteur local est situé dans un environnement bruyant à une distance des microphones du système qui peut aller de plusieurs dizaines de centimètres à quelques mètres. Grâce aux capacités de l'audition humaine, le locuteur local parvient à isoler la voix du correspondant provenant du haut-parleur, en utilisant les informations de son environnement. Toutefois, le correspondant n'a pas accès à ces informations pour comprendre ce que lui dit le locuteur local.

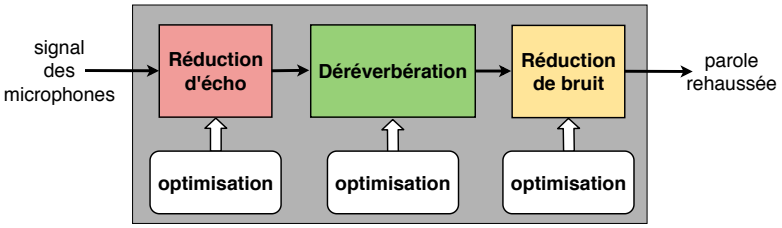
Les interactions entre le locuteur local et son correspondant avec un système mains-libres sont alors caractérisées par deux aspects. D'une part, la parole du locuteur local captée par les microphones peut être faible. D'autre part, les interactions sont susceptibles d'être soumises à plusieurs types de distorsion telles que le bruit ambiant, l'écho acoustique et la réverbération. Le bruit ambiant correspond à l'ensemble des sons qui ne sont pas désirés par le correspondant dans ses interactions avec le locuteur local. Ce peut être le son d'un réfrigérateur ou bien des bribes de conversations de personnes présentes

dans le fond de la salle. L'écho acoustique correspond à un bruit particulier qui provient du système mains-libres : c'est la rétroaction sonore entre le haut-parleur et les microphones. Par conséquent, le correspondant entendra une version retardée de sa propre voix, et l'assistant personnel « entendra » la musique jouée par l'enceinte connectée. La réverbération désigne l'ensemble des réflexions d'un son sur des murs ou des objets. En particulier, la réverbération produit un phénomène de persistance sonore de la parole du locuteur local au cours du temps. Dans une pièce comme une salle de bain, cette persistance peut être longue. Le bruit ambiant, l'écho acoustique et la réverbération dégradent la qualité et l'intelligibilité de la voix du locuteur local, ainsi que le confort d'écoute. Ces distorsions font aussi décroître les performances en reconnaissance automatique de la parole des assistants personnels. Le rehaussement de la parole a pour but de réduire le bruit ambiant, l'écho acoustique et la réverbération afin d'améliorer l'intelligibilité de la parole pour le correspondant et/ou sa transcription par un système de reconnaissance automatique [Loizou, 2007].

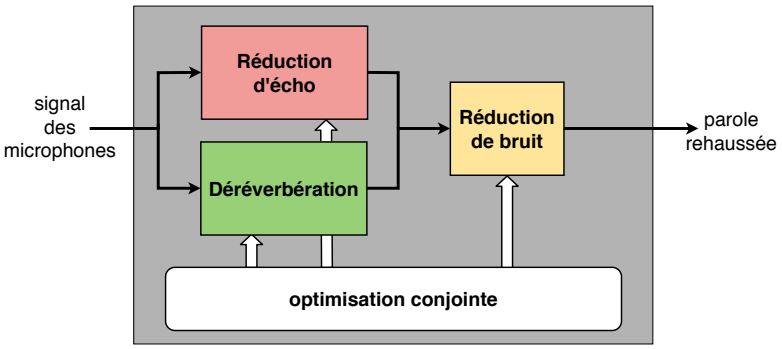
Rehaussement de la parole et approche bout-en-bout Dans les scénarios du quotidien, le bruit ambiant, l'écho acoustique et la réverbération peuvent se produire de manière simultanée. Il est possible de les réduire en disposant à la chaîne, ou en *cascade*, les modules conçus pour réduire indépendamment chaque type de distorsion. Différents ordres de traitement sont possibles. La figure 1.1a illustre la structure de modules de rehaussement de la parole utilisée dans l'enceinte intelligente HomePod (Apple). Toutefois, si chacun de ces modules est réglé indépendamment des autres, cela peut être sous-optimal et peut même produire des dégradations sur la parole du locuteur local, notamment dans le cas où les conditions acoustiques varient au cours du temps. Par exemple, un module de déréverbération placé en amont peut réduire le signal de bruit. Le module de réduction de bruit placé en aval peut alors dégrader la parole du locuteur local.

Pour éviter les phénomènes de ce type, il est nécessaire d'optimiser conjointement les modules de réduction du bruit, de l'écho acoustique et de la réverbération (voir la figure 1.1b). C'est en ce sens que nous définissons le terme *bout-en-bout* dans cette thèse. Dans le sens usuel, le terme *bout-en-bout* désigne les approches basées sur l'apprentissage profond qui estiment la parole rehaussée en définissant implicitement les modules de rehaussement à l'aide de l'apprentissage profond. Dans cette thèse, nous définissons explicitement les modules de rehaussement. Nous utilisons alors le terme *bout-en-bout* pour désigner les approches d'optimisation conjointe de ces modules.

Scénario considéré et objectifs Cette thèse s'inscrit dans un contexte industriel précis : la télécommunication mains-libres en environnement domestique via les enceintes intelligentes. En particulier, nous considérons l'enceinte intelligente Triby développée par la société Invoxia. Cette enceinte est similaire à Amazon Echo, Google Home et HomePod. Nous considérons un scénario courant d'utilisation de Triby (voir la figure 1.2) : il s'agit d'une conversation téléphonique entre un locuteur local et un correspondant distant via Triby dans un environnement domestique bruyant et réverbérant. Puisque nous



(a) Approche en cascade utilisée dans HomePod (Apple) [Audio Software Engineering and Siri Speech Team, 2018].



(b) Approche conjointe de Togami et Kawaguchi [2014]

FIGURE 1.1. – Exemples de réduction en cascade et conjointe de bruit, d'écho et de réverbération. Différents ordres de traitements sont possibles.

sommes seulement intéressés par le rehaussement de la parole du locuteur local, nous supposons que seul le locuteur local utilise l'enceinte Triby. Le correspondant distant utilise quant à lui un téléphone dans un environnement domestique silencieux. Ainsi, la parole distante jouée par l'enceinte Triby ne nécessite pas de rehaussement. En ce qui concerne le déroulement de la conversation, nous considérons que le locuteur local et le correspondant distant parlent l'un après l'autre, avec un temps de parole équivalent. Puisqu'il est courant que les deux locuteurs se coupent régulièrement la parole dans une conversation téléphonique, les deux paroles peuvent se chevaucher.

Le locuteur local est situé à quelques mètres des microphones de Triby, et fait face à l'enceinte intelligente lorsqu'il parle. Compte tenu du bruit ambiant, ainsi que de la distance entre le locuteur local et les microphones, le volume de l'enceinte intelligente est élevée pour que le locuteur local puisse entre la voix du correspondant distant sans difficulté. Par conséquent, l'écho acoustique enregistré par les microphones est fort par rapport à la voix du locuteur local. Par ailleurs, il est généralement peu probable que la position de l'enceinte intelligente change durant la conversation car, comme tout système mains-libres, l'enceinte est conçue afin que l'utilisateur ne soit pas obligé de la déplacer durant la conversation. C'est pourquoi nous considérons un scénario où l'enceinte intelligente reste fixe. En revanche, le locuteur local est susceptible de se déplacer environ toutes les dizaines de secondes durant la conversation dans un périmètre restreint de quelques mètres.

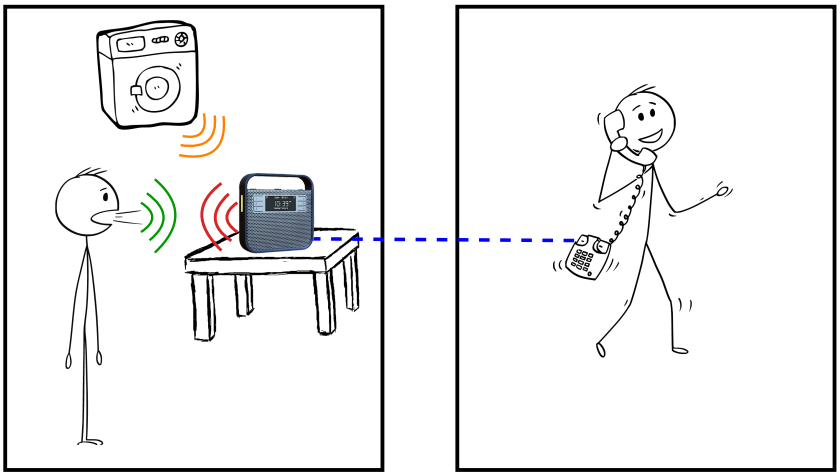


FIGURE 1.2. – Exemple d'une communication téléphonique avec Triby.

Puisque l'objectif est d'améliorer la télécommunication mains-libres avec Triby, les performances finales de l'approche seront mesurées dans le but d'une qualité perceptuelle pour l'auditeur humain, et non pas en vue de la reconnaissance automatique de la parole. Cette qualité perceptuelle sera déterminée à l'aide de métriques objectives et de tests d'écoute. Enfin, il convient d'ajouter que le rehaussement de la parole est réalisé dans l'enceinte Triby, dont les composants électroniques ont des capacités limitées par rapport à des serveurs de calcul hébergés (processeur à un seul coeur, faible mémoire vive). Par conséquent, la complexité algorithmique de l'approche doit rester faible. Les performances finales de l'approche seront donc aussi mesurées en matière de complexité algorithmique.

Pour valider l'approche, nous devons mener des tests en conditions acoustiques domestiques réalistes, voire réelles. Toutefois, il n'existe pas de base de données audio en accès public correspondant au scénario considéré. Nous collections donc des données à partir de simulations d'environnements réverbérants permettant d'obtenir les signaux de bruit et de parole. Nous complétons ces simulations en réalisant des enregistrements de signaux réels qui sont effectués avec l'enceinte intelligente Triby, dont l'écho acoustique, qui est un type de distorsion spécifique au système mains-libres difficile à simuler. L'enceinte Triby possède une antenne de quatre microphones avec une géométrie linéaire, qui nous permet de réaliser un traitement multicanal. Toutefois, on ne suppose aucune connaissance sur la topologie des microphones, ni sur la position des sources sonores. Ainsi, les résultats de la thèse sont potentiellement applicables à tout système de communication mains-libres et non seulement au Triby.

1.2. Outils utilisés

Apprentissage automatique L'apprentissage automatique regroupe les algorithmes permettant d'effectuer une tâche, par exemple, détecter la présence d'un chat sur une image

ou décider de la direction à prendre par un véhicule autonome, à partir de données observées. Pour ce faire, l'algorithme possède un certain nombre de paramètres, qui sont calculés automatiquement durant une phase dite d'apprentissage. Dans le cas de l'apprentissage supervisé, on présente à l'algorithme des données d'entrée (par exemple, des images) pour lesquelles les données de sortie correspondant à la tâche (par exemple, la classe de l'image ou l'angle du volant) sont connues. Ces données forment ensemble un jeu de données dit d'apprentissage ou d'entraînement. Si la sortie estimée est bonne, les paramètres sont inchangés. Autrement, on mesure le niveau d'erreur à partir d'une fonction de coût. Les paramètres de l'algorithme sont alors ajustés automatiquement en optimisant cette fonction de coût.

La performance de l'algorithme est ensuite mesurée sur des données qui n'ont pas été « vues » durant la phase d'apprentissage. Ces données forment un jeu de données dit de test. Plus le nombre de paramètres à calculer est grand, plus la quantité de données nécessaires à l'apprentissage est importante. Les arbres de décision et les méthodes à noyau font partie des méthodes d'apprentissage automatique. Toutefois, leur capacité de représentation est limitée du fait de leur faible nombre de paramètres, qui les rendent incapables de résoudre des tâches complexes comme la reconnaissance de la parole.

Apprentissage profond Les méthodes d'apprentissage par réseaux de neurones font partie des algorithmes d'apprentissage automatique. Les réseaux de neurones possèdent un certain nombre d'unités élémentaires (les « neurones ») qui peuvent s'organiser en groupes appelés « couches ». Quand le nombre de ces couches augmente, on parle alors de réseaux de neurones profonds. Grâce à la combinaison de ces couches, les réseaux de neurones ont une bien meilleure capacité de modélisation de tâches complexes que les méthodes classiques d'apprentissage automatique. Leurs paramètres nécessitent d'être choisis en fonction de la tâche à réaliser. L'organisation des couches, que l'on désigne par le terme d'architecture du réseau, doit être adaptée à la structure des données d'entrée, comme pour une image (structure des pixels adjacents), ou pour une série temporelle (structure des données successives). La fonction de coût doit prendre en compte la manière de mesurer l'erreur d'estimation des données de sortie correspondant à la tâche considérée. L'optimisation du grand nombre de paramètres du réseau durant la phase d'apprentissage nécessite une grande quantité de données. La dernière décennie a vu l'émergence de ce type d'algorithmes grâce aux avancées considérables en puissance de calcul au moyen de cartes graphiques (GPU, *graphics processing unit*) pour la phase d'apprentissage, ainsi que la quantité de données disponibles. Toutefois, l'interprétation des paramètres du réseau de neurones est difficile du fait de leur grand nombre.

Application au rehaussement de la parole Dans le domaine de la séparation de sources et du rehaussement de la parole, les réseaux de neurones ont permis des progrès considérables par rapport aux méthodes classiques de rehaussement, comme les approches basées sur la soustraction spectrale ou le traitement d'antennes. En particulier, la réduction de bruit et la séparation de sources de paroles ont été beaucoup améliorées [Vincent et al., 2018], et les réseaux de neurones ont amené des avancées en réduction d'écho [Lee

et al., 2015] et en déréverbération [Ernst et al., 2018]. Toutefois, ils n'ont jamais été utilisés pour résoudre la réduction conjointe de bruit, d'écho acoustique et de réverbération de *bout-en-bout* (tel que défini dans cette thèse, c'est-à-dire pour l'optimisation conjointe des modules de rehaussement, et non pas dans le sens usuel).

1.3. Contributions et plan du document

Nous avons développé trois contributions à cette thèse. La première est une approche monocanale de réduction d'écho qui corrige le problème lié au manque de convergence du filtre d'annulation d'écho, en estimant un post-filtre à partir d'un réseau de neurones appliqué sur plusieurs signaux différents. La seconde est une approche multicanale de réduction conjointe de bruit, d'écho et de réverbération qui optimisent tous les filtres en modélisant les caractéristiques spectrales de leurs interactions avec un réseau de neurones. La troisième est une extension en ligne de l'approche précédente de réduction conjointe dans le but de son implémentation en temps réel.

Le chapitre 2 décrit le contexte du rehaussement de la parole considéré dans cette thèse et définit la tâche à résoudre. Il expose ensuite les méthodes de réduction individuelle de bruit, d'écho acoustique et de réverbération. Ces méthodes sont utilisées pour présenter d'abord les méthodes de réduction conjointe de deux types de distorsion, puis les méthodes de réduction conjointe de bruit, d'écho et de réverbération. La fin du chapitre détaille les métriques dont nous nous servons pour mesurer la réduction de chaque type de distorsion, ainsi que la qualité du signal.

Le chapitre 3 présente notre méthode de réduction d'écho en monocanal, qui consiste à utiliser un post-filtre pour supprimer l'écho résiduel subsistant après l'application d'un filtre d'annulation d'écho. Dans des scénarios réels, ce filtre n'a pas nécessairement convergé. Par conséquent, l'écho résiduel peut être important, et le post-filtre nécessite d'être adapté en conséquence. Nous proposons d'estimer les coefficients d'un post-filtre d'écho résiduel à l'aide d'un réseau de neurones. Le réseau de neurones utilise plusieurs données d'entrées comprenant le signal après annulation d'écho, le signal source de l'écho, et l'écho estimé par le filtre en amont. Nous proposons un critère d'optimisation du réseau de neurones permettant d'améliorer les performances du post-filtre. Nous évaluons notre méthode sur des enregistrements réels d'écho et de parole acquis dans différentes situations, et nous la comparons à d'autres méthodes de suppression d'écho résiduel.

Le chapitre 4 est consacré à notre méthode de réduction conjointe de bruit, d'écho et de réverbération en multicanal, combinant plusieurs filtres qui doivent être optimisés simultanément. Nous proposons de modéliser la parole cible et les signaux résiduels après annulation d'écho et déréverbération, et d'estimer conjointement leurs caractéristiques spectrales à l'aide d'un réseau de neurones. Nous développons un algorithme de montée par blocs de coordonnées pour optimiser tous les filtres. Nous évaluons notre méthode sur des enregistrements réels de bruit, d'écho et de réverbération acquis dans différentes

situations. Nous la comparons à une combinaison en cascade des approches de réduction individuelle, et une autre méthode de réduction conjointe qui ne modélise pas les caractéristiques spectrales de la parole cible et des signaux résiduels.

Le chapitre 5 étend la méthode de réduction conjointe, présentée au chapitre précédent, à une version en ligne dans le but d'une implémentation en temps réel. Nous modifions l'algorithme de montée par blocs de coordonnées afin d'optimiser les filtres de manière récursive. Nous évaluons la méthode proposée sur les mêmes données que le chapitre précédent, et en particulier dans le le cas où les conditions acoustiques varient au cours du temps. Nous comparons cette méthode à la méthode hors-ligne proposée au chapitre précédent et à une version en ligne de la méthode en cascade utilisée dans le chapitre précédent pour la comparaison à notre méthode.

Le chapitre 6 rappelle les points principaux de ce manuscrit, y apporte une conclusion ainsi que des propositions de poursuite de ce travail de recherche.

1.4. Publications associées à cette thèse

1.4.1. Article de revue

— Carbajal, G., Serizel, R., Vincent, E., et Humbert, É. (2020). Joint NN-supported multichannel reduction of acoustic echo, reverberation and noise. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **(article soumis)**.

1.4.2. Articles de conférence

— Carbajal, G., Serizel, R., Vincent, E., et Humbert, É. (2020). Online DNN-based multichannel reduction of acoustic echo, reverberation and noise. In *IWAENC*, **(article en cours)**.

— Carbajal, G., Serizel, R., Vincent, E., et Humbert, É. (2018). Multiple-input neural network-based residual echo suppression. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 231–235.

2. Contexte et état de l'art

Ce chapitre présente un état de l'art de la réduction conjointe des types de distorsion en rehaussement de la parole. Nous formulons tout d'abord le problème de rehaussement de la parole considéré dans cette thèse. Nous détaillons ensuite les méthodes de réduction individuelle de bruit, d'écho acoustique et de réverbération dont nous nous servons par la suite pour présenter les composantes des solutions proposées. Cette partie nous permet de couvrir les approches de réduction conjointe de deux types de distorsion, et enfin les approches de réduction conjointe de bruit, d'écho acoustique et de réverbération.

2.1. Formulation du problème

Nous considérons un scénario où un locuteur local interagit avec un correspondant distant à l'aide d'un système de télécommunication mains-libres et en présence de bruit ambiant. Dans ce scénario, la parole du locuteur local, que nous désignons par le terme de *parole locale*, est soumise à trois types de distorsion : le bruit ambiant, l'écho acoustique et la réverbération de la salle. Le but est d'améliorer la perception de la parole du locuteur local pour le correspondant distant. Nous ne faisons aucune hypothèse sur la topologie de l'antenne de microphones ou sur la position des sources.

2.1.1. Formulation temporelle

La figure 2.1 illustre le scénario considéré au niveau du système de télécommunication mains-libres. Le mélange $\mathbf{d}(t) = [d_1(t) \dots d_M(t)]^T \in \mathbb{R}^{M\times 1}$ est le signal enregistré par les M microphones à l'instant $t \in \{0 \dots T-1\}$, où l'indice $m \in \{1 \dots M\}$ représente le microphone qui enregistre le signal, $(\cdot)^T$ la transposée d'une matrice ou d'un vecteur, et T le nombre d'échantillons de l'enregistrement. Le signal $\mathbf{d}(t)$ est composé de la somme de la parole locale réverbérée $\mathbf{s}(t) \in \mathbb{R}^{M\times 1}$, du signal de bruit $\mathbf{b}(t) \in \mathbb{R}^{M\times 1}$ et de l'écho $\mathbf{y}(t) \in \mathbb{R}^{M\times 1}$, observés au niveau des microphones :

$$\mathbf{d}(t) = \mathbf{s}(t) + \mathbf{b}(t) + \mathbf{y}(t). \tag{2.1}$$

Les signaux $\mathbf{s}(t)$, $\mathbf{b}(t)$ et $\mathbf{y}(t)$ sont les *images spatiales* de leur source sonore respective, c'est-à-dire la contribution de chacune de ces sources au mélange $\mathbf{d}(t)$. Ils ont des caractéristiques propres qui dépendent de la nature des sources qui les émettent, ainsi que de l'environnement acoustique.

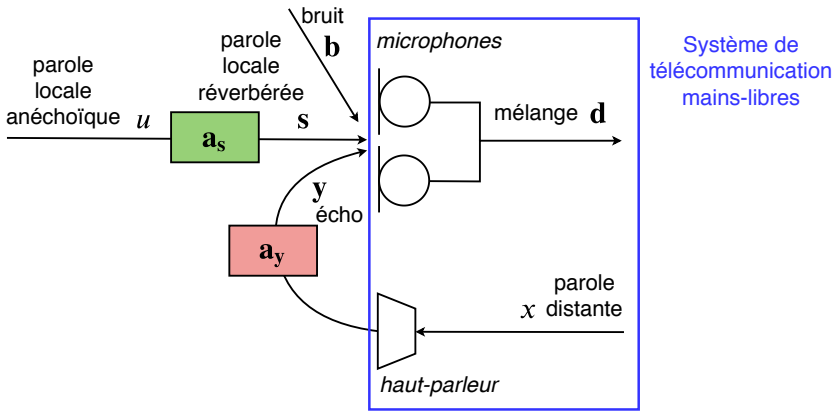


FIGURE 2.1. – Problème de l'écho acoustique, du bruit et de la réverbération.

2.1.1.1. Propagation du son

Un signal $\mathbf{c}(t)$ est le résultat de l'enregistrement du son émis par une source sonore qui se propage dans l'environnement acoustique jusqu'aux microphones. Dans un endroit clos, comme une salle, ce son se réfléchit sur les surfaces et les objets de la salle. La figure 2.2a illustre ce phénomène appelé *réverbération*. En rehaussement de la parole, une source est dite *ponctuelle* lorsqu'elle émet du son depuis un point précis de la salle. Le son émis depuis ce point est modélisé par un signal source monocanal $o(t) \in \mathbb{R}$. Le signal $\mathbf{c}(t)$, dit *réverbéré*, est alors représenté par la convolution linéaire entre le signal source $o(t)$ et une impulsion sonore $\mathbf{a}_c(t-\tau,\tau) \in \mathbb{R}^{M\times 1}$, appelée réponse impulsionnelle de salle (RIR, *room impulse response*) et qui modélise le chemin acoustique entre la source ponctuelle et les microphones :

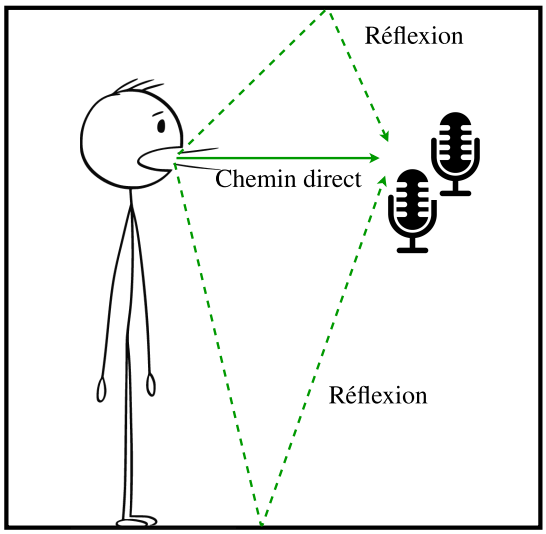
$$\mathbf{c}(t) = \sum_{\tau \geq 0} \mathbf{a}_c(t-\tau,\tau) o(t-\tau). \tag{2.2}$$

La RIR $\mathbf{a}_c(t-\tau,\tau)$ dépend du temps t, en raison des conditions acoustiques qui varient potentiellement dans le temps (par exemple, la source est en mouvement). Les paramètres du processus de propagation du son $\mathbf{a}_c(t-\tau,\tau)$ varient cependant lentement par rapport au signal $u(t)$. Par la suite, on considère que les conditions acoustiques sont invariantes dans le temps :

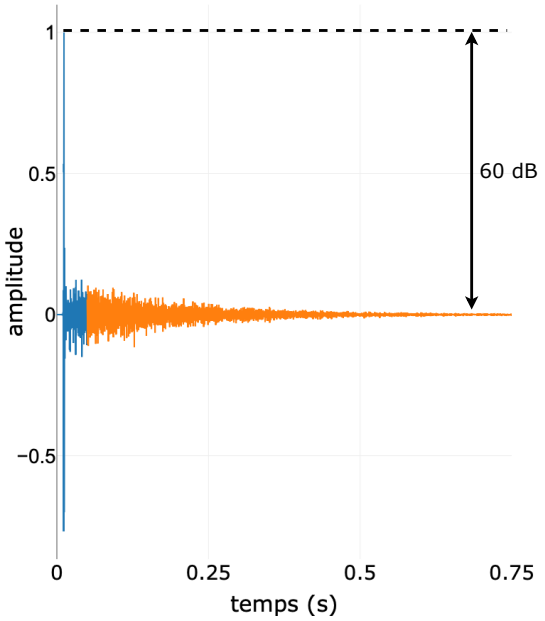
$$\mathbf{a}_c(t-\tau,\tau) = \mathbf{a}_c(\tau). \tag{2.3}$$

La figure 2.2b illustre une mesure de RIR dans une salle. L'intervalle de silence initial, appelé délai de propagation, correspond au temps de propagation de la source vers le microphone en chemin direct. La RIR comporte deux périodes. La première inclut le pic principal, qui correspond au chemin direct, et les réflexions précoces, qui désignent l'ensemble des sons réfléchis arrivant dans un court délai t_e après le chemin direct, appelé *temps de mélange*. La réverbération tardive désigne toutes les réflexions qui arrivent après ce délai.

La réverbération est généralement quantifiée avec les statistiques d'acoustique de salle : le temps de réverbération T_{60} et le le rapport direct sur réverbérant (DRR, *direct-*



(a) Schéma de la réverbération dans une salle.



(b) Exemple de RIR. Le chemin direct et les réflexions précoces sont représentées en bleu. La réverbération tardive en représentée en orange.

to-reverberant ratio) [Naylor et Gaubitch, 2010, Chapitre 2]. Le temps de réverbération T_{60} est défini à l'aide d'une mesure d'énergie en décibels (dB), qui est une échelle liée à la perception de la puissance d'un son par le système auditif. Cette métrique correspond au temps nécessaire pour que l'énergie des réflexions diminue de 60 dB par rapport à l'énergie du chemin direct (voir la figure 2.2b). Toutefois, le temps de réverbération T_{60} ne caractérise pas totalement la réverbération d'un signal. C'est pourquoi l'on utilise aussi le DRR, qui est défini comme le rapport d'énergie du chemin direct sur l'ensemble des réflexions [Naylor et Gaubitch, 2010, Chapitre 2].

2.1.1.2. Parole locale réverbérée

En rehaussement de la parole, le locuteur local est considéré comme une source ponctuelle, qui produit, au niveau de la bouche, un signal source monocanal $u(t) \in \mathbb{R}$, que l'on désigne par le terme de *parole locale anéchoïque*. Dans la suite de la thèse, nous désignerons la parole locale réverbérée $\mathbf{s}(t)$ plus simplement par le terme de *parole locale*, qui sera distinct de la *parole locale anéchoïque* $u(t)$. La parole locale $\mathbf{s}(t)$ est alors représentée par la convolution linéaire entre la parole anéchoïque $u(t)$ et la RIR $\mathbf{a}_s(t-\tau,\tau) \in \mathbb{R}^{M\times 1}$ (voir la figure 2.1) :

$$\mathbf{s}(t) = \sum_{\tau \geq 0} \mathbf{a}_s(\tau) u(t-\tau). \tag{2.4}$$

La parole locale $\mathbf{s}(t)$ se décompose alors de la manière suivante :

$$\mathbf{s}(t) = \underbrace{\sum_{0 \leq \tau \leq t_\mathrm{e}} \mathbf{a}_\mathrm{s}(\tau) u(t-\tau)}_{=\mathbf{s}_\mathrm{e}(t)} + \underbrace{\sum_{\tau > t_\mathrm{e}} \mathbf{a}_s(\tau) u(t-\tau)}_{=\mathbf{s}_\mathrm{l}(t)}, \tag{2.5}$$

où la composante $\mathbf{s}_\mathrm{e}(t)$ correspond au chemin direct et aux réflexions précoces de la parole locale, qu'on désigne par le terme de *composante précoce*, et la composante $\mathbf{s}_\mathrm{l}(t)$ correspond à la réverbération tardive. En particulier, cette composante présente les caractéristiques d'un signal diffus, c'est-à-dire un signal dont l'énergie sonore est uniformément répartie dans l'espace. Le temps de mélange t_e se situe généralement autour de 50 ms.

2.1.1.3. Bruit

Les sons qui ne sont pas de la parole cible sont considérés comme étant du bruit. En général, on considère comme du bruit tous les sons qui ne sont pas de la parole intelligible. Un scénario compliqué correspond au cas où le bruit est aussi de la parole intelligible. Dans ce cas, on parle plutôt d'interférence. Le bruit peut provenir d'une source ponctuelle, comme d'un appareil domestique par exemple. La propagation du bruit peut alors être représentée de la même manière que pour la parole locale. Une source de bruit peut être constituée de plusieurs sources physiques, comme le bavardage d'une dizaines personnes dans un café. Le bruit peut aussi provenir d'une source diffuse, c'est-à-dire que le son émis provient d'une région entière de l'espace, comme de la pluie. Dans ce cas, le processus de propagation est plus difficile à modéliser car le signal source ne peut pas être considéré comme monocanal. Il est toutefois possible de représenter le signal comme une agrégation de sources ponctuelles [Vincent et al., 2018, Chapitre 1]. En général, on représente directement le signal de bruit par $\mathbf{b}(t)$.

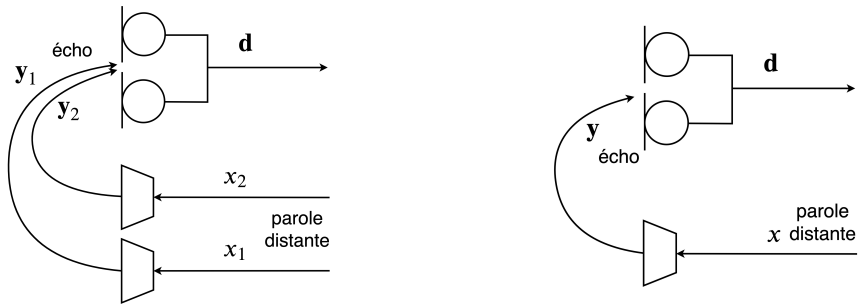
2.1.1.4. Écho

L'écho acoustique $\mathbf{y}(t)$ est un signal non désiré qui provient de la rétroaction acoustique du haut-parleur (source ponctuelle) vers les microphones. En raison de la proximité du haut-parleur et des microphones, l'écho est généralement beaucoup plus puissant que les autres signaux. Bien que l'écho acoustique ne soit pas nécessairement de la parole, il coïncide ici à la parole du correspondant distant. Contrairement à la parole locale, cette parole non désirée est une version déformée de manière non linéaire du signal source de parole distante $x(t) \in \mathbb{R}$ joué par le haut-parleur. La composante non-linéaire de l'écho $\mathbf{y}(t)$ est causée par les réponses non-linéaires du haut-parleur et des microphones, les vibrations de l'enceinte et les effets de coupure dus à l'amplification. De même que pour la parole locale $\mathbf{s}(t)$, la composante linéaire de l'écho $\mathbf{y}(t)$ peut être représentée par la convolution linéaire entre le signal $x(t)$ et la RIR de l'écho $\mathbf{a}_y(\tau) \in \mathbb{R}^{M \times 1}$, aussi appelée *chemin d'écho* (voir la figure 2.1) :

$$\mathbf{y}(t) \approx \sum_{\tau \geq 0} \mathbf{a}_\mathrm{y}(\tau) x(t-\tau). \tag{2.6}$$

Dans la littérature liée à l'écho acoustique, le caractère multicanal désigne généralement un contexte stéréophonique, c'est-à-dire avec deux haut-parleurs : la parole distante est alors un signal multicanal $\mathbf{x}(t) = [x_1(t)\ x_2(t)]^T$, et chaque haut-parleur joue un canal différent de ce signal [Sondhi et al., 1995] (voir la figure 2.3a). L'écho $\mathbf{y}(t)$ n'est plus représenté comme dans (2.6), mais par la somme de deux convolutions d'une RIR différente avec chacun des canaux de $\mathbf{x}(t)$. Les systèmes stéréophoniques sont notamment utilisés dans les systèmes de visioconférence, où l'on cherche à obtenir un réalisme spatial entre l'audio et la vidéo. Cela se révèle important lorsque le locuteur local cherche à mieux identifier qui parle sur la vidéo parmi un groupe de correspondants distants.

Dans le cas des communications téléphoniques plus classiques, comme avec le Triby, la parole distante a généralement subi des traitements antérieurs, et arrive au système mains-libres sous forme monocanale $x(t)$. De plus, les haut-parleurs des systèmes mains-libres sont suffisamment proches pour être considérés comme une seule source. Comme ces systèmes sont généralement dotés de plusieurs microphones, le caractère multicanal désignera ici un contexte avec plusieurs microphones (voir la figure 2.3b). En ce qui concerne le Triby, il possède deux haut-parleurs, séparés d'environ 8 cm, et quatre microphones.



(a) Défini dans la littérature.

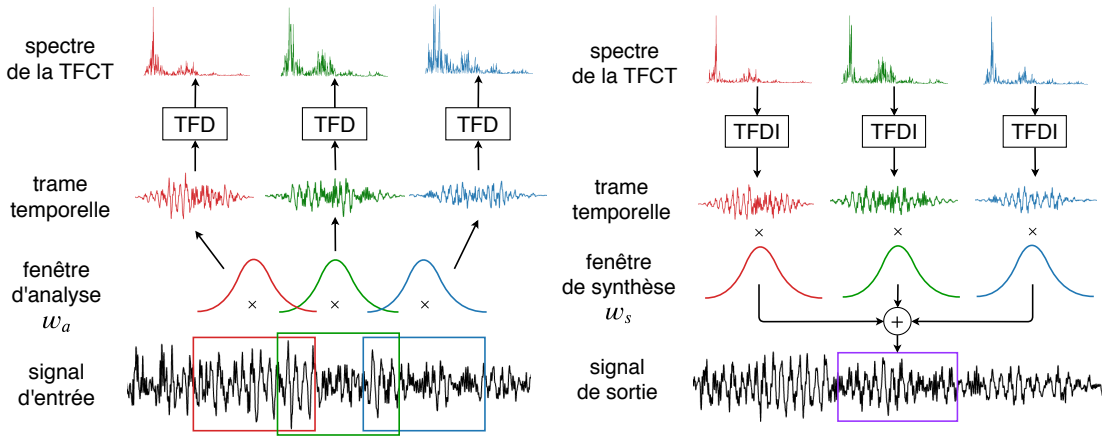
(b) Défini dans cette thèse.

FIGURE 2.3. – Définition du terme « multicanal » dans la réduction d'écho acoustique.

2.1.2. Représentation temps-fréquence

La représentation temporelle des signaux n'est pas suffisamment informative pour déterminer leurs propriétés. En général, les méthodes de rehaussement opèrent dans le domaine temps-fréquence, où l'on peut représenter à la fois les caractéristiques spectrales et temporelles des signaux. La représentation la plus utilisée est la transformation de Fourier à court terme (TFCT) qui modélise l'amplitude et la phase du signal [Allen, 1977]. D'autres représentations temps-fréquence utilisent différentes échelles de fréquences non-linéaires basées sur le système auditif humain, comme l'échelle Mel [Stevens et al., 1937] et de l'échelle de la bande passante rectangulaire équivalente (ERB, *equivalent rectangular bandwith*). Dans cette thèse, nous n'utilisons que la TFCT et son inverse. Une phase d'analyse convertit le signal temporel dans le domaine temps-fréquence. Puis, une phase

de synthèse reconstruit le signal rehaussé dans le domaine temporel à partir de sa représentation temps-fréquence estimée. La figure 2.4 résume ces processus. Nous considérons le cas monophonique ($M = 1$) et l'indice m du microphone est omis par souci de clarté. Dans le cas où $M > 1$, les phases d'analyse et de synthèse sont appliquées sur chaque canal m.



(a) Phase d'analyse.

(b) Phase de synthèse.

FIGURE 2.4. – Schéma représentant les phases d'analyse et de synthèse d'un signal temporel. Il convient de noter qu'ici le pas de chevauchement des fenêtres de la phase d'analyse ne satisfait pas les conditions de reconstruction parfaite.

Phase d'analyse Cette phase consiste à d'abord segmenter le signal temporel en trames temporelles qui se chevauchent. T_{TFCT} désigne la taille des trames et P le pas d'avancement entre chaque trame consécutive. Le chevauchement des trames est nécessaire à la reconstruction exacte du signal. Chaque trame est ensuite multipliée par une fenêtre. Ainsi, à la trame $n \in \{0, \dots, N-1\}$, avec $N \in \mathbb{N}$, le signal du mélange $d(n,t)$ est défini par

$$d(n,t) = d(t + nP)w_a(t), \qquad t \in \{0, \dots, T_{\text{TFCT}} - 1\}, \tag{2.7}$$

où $w_a(t)$ est la *fenêtre d'analyse* et N le nombre de trames de l'enregistrement. On utilise généralement une fenêtre dont la transformée de Fourier est restreinte sur l'intervalle $\{0, \dots, T_{\text{TFCT}} - 1\}$, ce qui permet une meilleure séparation des composantes spectrales. Différentes fenêtres, comme la fenêtre de Hamming, sont possibles et se traduisent par différentes résolutions de trames T_{TFCT}. Après cette étape de fenêtrage, on applique la *transformée de Fourier discrète* sur chacune des trames pour obtenir les coefficients complexes de la TFCT à la trame n et à la bande de fréquence f

$$d(n,f) = \sum_{t=0}^{T_{\text{TFCT}}-1} d(n,t) \exp(-2i\pi ft/F), \qquad f \in \{0, \dots, F-1\} \tag{2.8}$$

où F est le nombre de bandes de fréquences. Très souvent, on choisit $F = T_{\text{TFCT}}$. La *transformation de Fourier discrète* (TFD) crée F coefficients complexes. Comme le signal analysé a des valeurs réelles dans le domaine temporel, les coefficients des bandes de fréquence $f \in \{1, \ldots, F/2 - 1\}$ sont les conjugués de ceux des bandes $f \in \{F/2 + 1, \ldots, F - 1\}$. Ainsi, seuls $F/2 + 1$ coefficients sont utiles pour la représentation temps-fréquence. Le spectre complexe de la TFCT permet de manipuler séparément l'amplitude $|d(n, f)|$ et la phase $\theta_d(n, f)$. Le filtrage du signal $d(n, f)$, détaillé dans les sections suivantes, produit ensuite l'estimation d'un signal $\widehat{c}(n, f)$.

Phase de synthèse Cette phase consiste à appliquer d'abord la *transformation de Fourier discrète inverse* (TFDI) pour convertir les trames $\widehat{c}(n, f)$ dans le domaine temporel. Ainsi, à la trame n, le signal estimé $\widehat{c}(n, t)$ est obtenu comme :

$$\widehat{c}(n, t) = \sum_{f=0}^{F-1} \widehat{c}(n, f) \exp(+2i\pi ft/F), \qquad t \in \{0, \ldots, T_{\text{TFCT}} - 1\}. \tag{2.9}$$

La procédure d'*overlap-add* consiste alors à sommer les trames du signal $\widehat{s}_{\text{e}}(n, t)$ pour reconstruire le signal estimé $\widehat{s}_{\text{e}}(t)$ dans le domaine temporel [Allen, 1977]. Toutefois, lorsque le filtrage appliqué sur les trames de la TFCT est non-linéaire, des discontinuités entre les trames peuvent apparaître au niveau des bords, ce qui produit des artefacts dans le signal temporel $\widehat{s}_{\text{e}}(t)$. Pour éviter ce phénomène, chaque trame du signal $\widehat{s}_{\text{e}}(n, t)$ est multipliée par une fenêtre dite de synthèse $w_s(t)$ avant que les trames $\widehat{s}_{\text{e}}(n, t)$ ne soient sommées ensemble

$$\widehat{c}(t) = \sum_{n=-\infty}^{+\infty} \widehat{c}(n, t - nH) w_s(t - nP), \qquad t \in \{0, \ldots, T_{\text{TFCT}} - 1\}, \tag{2.10}$$

où $w_s(t)$ est la *fenêtre de synthèse*. Cette procédure est appelée *weighted overlap-add* [Crochiere, 1980].

La fenêtre de synthèse doit cependant être choisie pour satisfaire la propriété de reconstruction parfaite du signal : dans le cas où aucun filtrage n'a été appliqué sur le signal $d(n, f)$, la phase de synthèse doit redonner le même signal temporel $d(t)$ avant la phase d'analyse. Cette propriété est satisfaite si les fenêtres d'analyse w_a et de synthèse w_s vérifient :

$$\sum_{n=-\infty}^{+\infty} w_a(t - nP) w_s(t - nP) = 1. \tag{2.11}$$

Résolution en temps et en fréquence La résolution de la TFCT dépend surtout de la taille des trames T_{TFCT}. La résolution temporelle est inversement proportionnelle à la résolution fréquentielle. Un compromis entre les deux est donc nécessaire pour le rehaussement de la parole cible. Ce compromis est choisi en fonction de la stationnarité des signaux. Le spectre de la TFCT d'un signal *stationnaire* ne varie pas ou peu pendant la période de temps de traitement, comme le bruit d'un ventilateur, tandis que le

spectre de la TFCT d'un signal non-stationnaire évolue continuellement dans le temps, comme la parole. Certains sons peuvent être quasi-périodiques, comme les voyelles, ou bien ressembler plutôt à des impulsions. De plus, la parole est un signal dit intermittent, car il existe des silences entre les mots. Une conversation est généralement constituée de plus de 50% de pauses. Ainsi, en rehaussement de la parole, la taille des trames est un enjeu critique. Si T_{TFCT} est plus grand que le temps d'évolution des caractéristiques des signaux, les caractéristiques spectrales correspondantes seront mal estimées. Si T_{TFCT} est trop petit, le nombre de bandes de fréquence sera insuffisant pour séparer les composantes spectrales des sources entre elles. En général, pour obtenir une représentation parcimonieuse adéquate de la parole, la taille des trames doit être de l'ordre de 50 ms (soit $T_{\text{TFCT}} = 1\,024$ points avec une fréquence d'échantillonnage $f_s = 16$ kHz) [Vincent et al., 2018, Chapitre 2]. Un exemple de spectrogramme de parole est illustré sur la figure 2.5.

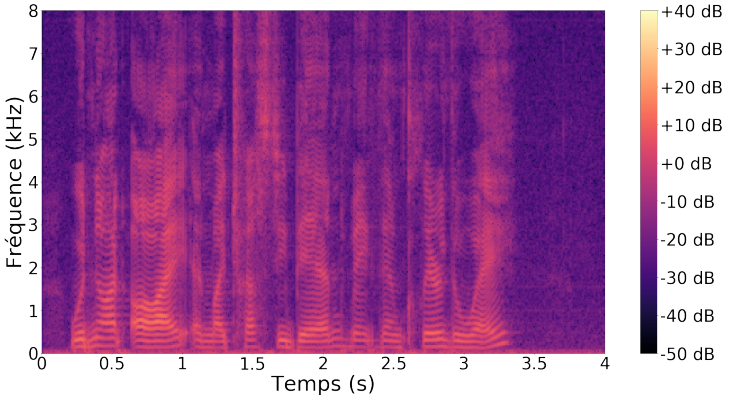


FIGURE 2.5. – Exemple de spectrogramme de la parole.

Processus de propagation en temps-fréquence La convolution temporelle du processus de propagation de la parole locale $s(t)$ dans (2.4) peut s'exprimer de manière exacte dans le domaine temps-fréquence par [Vincent et al., 2018, Chapitre 2] :

$$s(n,f) = \sum_{f'=0}^{F-1} \sum_{k\geq 0} a_s(k,f',f)u(n-k,f'), \tag{2.12}$$

où k et f' sont respectivement les indices associés au délai τ, exprimé en trame, et à la bande de fréquence. La réverbération introduit une corrélation temporelle entre les trames successives de la parole locale $s(n,f)$.

2.1.3. Positionnement de la thèse

Ainsi, d'après les parties précédentes, le signal $\mathbf{d}(n,f)$ observé aux microphones peut s'exprimer dans le domaine temps-fréquence de la manière suivante :

$$\mathbf{d}(n,f) = \mathbf{s}_{\text{e}}(n,f) + \mathbf{s}_{\text{l}}(n,f) + \mathbf{b}(n,f) + \mathbf{y}(n,f). \tag{2.13}$$

Dans des scénarios réels, la réverbération tardive $\mathbf{s}_{\mathrm{l}}(n,f)$, le bruit $\mathbf{b}(n,f)$ et l'écho $\mathbf{y}(n,f)$ peuvent être présents simultanément. Le but de cette thèse est d'extraire les M canaux du signal désiré $\mathbf{s}_{\mathrm{e}}(n,f)$ tout en limitant les dégradations sur ce signal. La réduction simultanée de ces trois types de distorsion implique l'utilisation d'un système de rehaussement combinant des méthodes individuelles de filtrage du bruit, de l'écho et de la réverbération.

Dans la suite de ce chapitre, nous présentons les méthodes de filtrage dans des scénarios à un, deux et trois types de distorsion, où le nombre de signaux et composantes de signaux peuvent varier. Pour cette raison, la notation $\mathbf{c}(n,f)$ désignera indifféremment un signal ou la composante d'un signal.

2.2. Réduction d'un type de distorsion

Dans cette partie, nous considérons un scénario où un seul type de distorsion est présent. Nous présentons d'abord la réduction de bruit, puis la réduction d'écho acoustique, et enfin la déréverbération.

2.2.1. Réduction de bruit

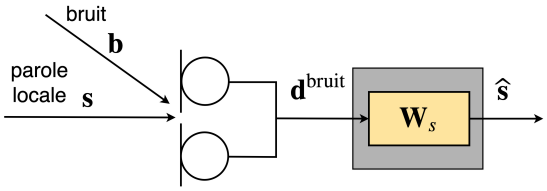


FIGURE 2.6. – Problème de la réduction de bruit.

Le problème de la réduction de bruit est illustré sur la figure 2.6. Le signal $\mathbf{d}^{\mathrm{bruit}}(n,f) \in \mathbb{C}^{M\times 1}$ observé aux microphones est la somme de la parole locale $\mathbf{s}(n,f) \in \mathbb{C}^{M\times 1}$ et du signal de bruit $\mathbf{b}(n,f) \in \mathbb{C}^{M\times 1}$:

$$\mathbf{d}^{\mathrm{bruit}}(n,f) = \mathbf{s}(n,f) + \mathbf{b}(n,f). \tag{2.14}$$

La réduction de bruit consiste à extraire la parole locale $\mathbf{s}(n,f)$ en appliquant un filtre causal sur le signal de mélange $\mathbf{d}^{\mathrm{bruit}}(n,f)$ [Vincent et al., 2018, Chapitre 2] :

$$\widehat{\mathbf{s}}(n,f) = \sum_{f'=0}^{F-1} \sum_{k} \mathbf{W}_s(k,n,f',f)\mathbf{d}^{\mathrm{bruit}}(n-k,f'), \tag{2.15}$$

où $\mathbf{W}_s(k,n,f',f) \in \mathbb{C}^{M\times M}$ est la matrice carrée de dimension M correspondant au k-ième délai du filtre causal. Pour réduire la complexité du filtrage, il est possible de faire l'*approximation en bande étroite*. Cela consiste à supposer que les bandes de fréquence peuvent être traitées séparément par le filtre, c'est-à-dire que les coefficients du filtre

$\mathbf{W}_s(n', n, f', f) = \mathbf{0}$ pour $f' \neq f$, et que la taille du filtre dans le domaine temporel est très inférieure à la taille des trames T_{TFCT}. La convolution en trame et en fréquence dans (2.15) se ramène à une simple multiplication dans le domaine temps-fréquence :

$$\widehat{\mathbf{s}}(n, f) \approx \mathbf{W}_s(n, f)\mathbf{d}^{\text{bruit}}(n, f). \tag{2.16}$$

On parle de *filtre court* pour désigner $\mathbf{W}_s(n, f)$ de manière générale, de *masque temps-fréquence* dans le cas monocanal, et de *suppression* pour désigner le filtrage par filtre court. L'approximation en bande étroite est généralement utilisée même dans le cas où la taille du filtre causal dans le domaine temporel est égal ou supérieure à la taille des trames T_{TFCT}. Bien qu'elle génère des artefacts à cause des convolutions cycliques, la forme décroissante des fenêtres d'analyse et de synthèse aux bords des trames réduit ces effets non désirés. Pour être robuste aux fluctuations du spectre des signaux et des conditions acoustiques, le filtre court $\mathbf{W}_s(n, f)$ est généralement conçu pour pouvoir varier rapidement dans le temps : le filtre est alors qualifié de *non-linéaire*.

2.2.1.1. Filtrage spatial

Si l'on connaît la position du locuteur local, on peut construire un filtre spatial fixe $\mathbf{W}_s^{\text{FBF}}(f)$, c'est-à-dire invariant dans le temps, qui se « focalise » sur celui-ci. Le filtre $\mathbf{W}_s^{\text{FBF}}(f)$ est construit en retardant les signaux de certains microphones pour que la parole locale atteigne tous les microphones au même moment, et en faisant la moyenne des signaux ainsi obtenus. La parole locale présente dans le signal de sortie est alors plus forte que sur chaque microphone séparément. Toutefois, les filtres spatiaux fixes ne permettent de réduire que partiellement le signal de bruit $\mathbf{b}(n, f)$. Les approches de filtrage spatial adaptatif visent à pallier ce problème. L'annulateur du lobe latéral généralisé (GSC, *generalized sidelobe canceller*) est une approche de filtrage spatial adaptatif couramment utilisée. La figure 2.7 illustre la structure du GSC. Cette approche estime la parole locale $\mathbf{s}(n, f)$ à l'aide d'un critère des moindres carrés entre le signal de sortie d'un filtre spatial fixe $\mathbf{W}_s^{\text{FBF}}(f)$ focalisé sur le locuteur local, et le signal de sortie d'un filtre spatial fixe $\mathbf{B}(f)$, appelé *matrice de blocage*, qui «bloque» les signaux arrivant dans la direction du locuteur local [Frost, 1972; Widrow et al., 1975; Griffiths et Jim, 1982]. D'autres variantes du GSC ont été proposées afin de limiter la dégradation de la parole locale [Affes et Grenier, 1997; Hoshuyama et al., 1999; Gannot et al., 2001]. Toutefois ces méthodes fonctionnent mal en environnement réverbérant.

2.2.1.2. Filtre de Wiener

Une autre manière d'estimer de la parole locale $\mathbf{s}(n, f)$ consiste à calculer un filtre adaptatif $\mathbf{W}_s(n, f)$ optimal selon un certain critère. Un critère fréquemment utilisé pour déterminer $\mathbf{W}_s(n, f)$ est l'erreur quadratique moyenne minimale (MMSE, en anglais *minimum mean square error*)

$$\min_{\mathbf{W}_s(n,f)} \mathbb{E}\Big[\Big\| \mathbf{s}(n, f) - \mathbf{W}_s(n, f)\mathbf{d}^{\text{bruit}}(n, f) \Big\|^2 \Big], \tag{2.17}$$

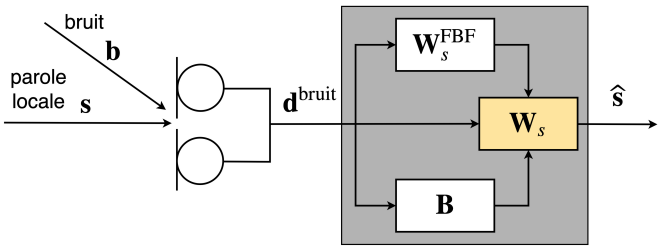


FIGURE 2.7. – Schéma du GSC pour la réduction de bruit.

où $\|\cdot\|^2$ désigne la norme euclidienne, et $\mathbb{E}[\cdot]$ l'espérance mathématique. Généralement, on suppose que le signal de bruit $\mathbf{b}(n,f)$ et la parole locale $\mathbf{s}(n,f)$ sont indépendants. Le problème de minimisation se ramène à

$$\min_{\mathbf{W}_s(n,f)} \mathbb{E}\Big[\,\|\mathbf{s}(n,f) - \mathbf{W}_s(n,f)\mathbf{s}(n,f)\|^2\,\Big] + \mathbb{E}\Big[\,\|\mathbf{W}_s(n,f)\mathbf{b}(n,f)\|^2\,\Big]. \tag{2.18}$$

La solution de ce problème d'optimisation est appelée filtre de Wiener :

$$\mathbf{W}_s^{\text{WF}}(n,f) = \mathbf{\Sigma}_s(n,f)\Big(\mathbf{\Sigma}_s(n,f) + \mathbf{\Sigma}_b(n,f)\Big)^{-1}, \tag{2.19}$$

où

$$\mathbf{\Sigma}_c(n,f) = \mathbb{E}\Big[\mathbf{c}(n,f)\mathbf{c}(n,f)^H\Big] \tag{2.20}$$

désigne la covariance de la source $\mathbf{c} = \mathbf{s}$ ou $\mathbf{b}$, et $(\cdot)^H$ la transposée conjuguée. La covariance contient à la fois les caractéristiques spectrales de la source $\mathbf{c}$, c'est-à-dire la distribution de son énergie en temps et en fréquence, et ses caractéristiques spatiales, c'est-à-dire les informations sur son processus de propagation dans la salle.

Le filtre de Wiener $\mathbf{W}_s^{\text{WF}}(n,f)$ opère principalement sur le spectre en amplitude du signal $|\mathbf{d}^{\text{bruit}}(n,f)|$ et néglige la phase du signal du bruit $\theta_b(n,f)$ dans la phase du mélange $\theta_d(n,f)$. Il introduit alors des artefacts dans la parole locale estimée $\mathbf{s}(n,f)$, qui peuvent être importants pour des rapports signal-à-bruit (SNR, *signal-to-noise ratio*) faibles [Oppenheim et Lim, 1981]. D'autres variantes du filtre de Wiener contrôlent le compromis entre la réduction de bruit et l'introduction d'artefacts [Benesty et al., 2005, Chapitre 9], ou prennent en compte à la fois l'amplitude et la phase des signaux pour l'estimation du filtre court $\mathbf{W}_s(n,f)$ [Williamson et al., 2016].

Le filtre $\mathbf{W}_s(n,f)$ nécessite un estimateur des covariances de la parole locale $\mathbf{\Sigma}_s(n,f)$ et du bruit $\mathbf{\Sigma}_b(n,f)$. Nous détaillons tout d'abord les méthodes d'estimation des covariances par soustraction spectrale, puis celles par séparation de sources, et enfin les méthodes de réduction de bruit par apprentissage profond.

2.2.1.3. Estimation des covariances par soustraction spectrale

Les méthodes d'estimation des covariances par soustraction spectrale modélisent la covariance du bruit $\mathbf{\Sigma}_b(n,f)$ et déduisent la covariance de la parole locale $\mathbf{\Sigma}_s(n,f)$ par

soustraction spectrale :

$$\mathbf{\Sigma}_s(n,f) = \widehat{\mathbb{E}}\Big[\mathbf{d}^{\text{bruit}}(n,f)\mathbf{d}^{\text{bruit}}(n,f)^H\Big] - \mathbf{\Sigma}_b(n,f), \tag{2.21}$$

où $\widehat{\mathbb{E}}$ désigne la moyenne empirique d'une variable déterministe à un instant donné. Par la suite, on désigne le terme $\widehat{\mathbb{E}}\Big[\mathbf{d}^{\text{bruit}}(n,f)\mathbf{d}^{\text{bruit}}(n,f)^H\Big]$ comme la « covariance » de la variable déterministe $\mathbf{d}^{\text{bruit}}(n,f)$, bien qu'elle n'ait la même signification que pour la variable aléatoire dans (2.20).

Historiquement, les approches de soustraction spectrale font partie des premières méthodes de réduction de bruit [Boll, 1979; McAulay et Malpass, 1980]. Elles supposent que le signal du bruit $\mathbf{b}(n,f)$ est plus stationnaire que la parole locale $\mathbf{s}(n,f)$. La covariance du bruit $\mathbf{\Sigma}_b(n,f)$ peut alors être estimée pendant les périodes d'absence de la parole, où $\mathbf{d}^{\text{bruit}}(n,f) = \mathbf{b}(n,f)$. Pour déterminer les périodes d'absence et de présence de la parole, une technique simple consiste à utiliser une méthode de classification appelée détecteur d'activité vocale (VAD, *voice activity detector*) [Boll, 1979]. L'inconvénient du VAD est que la covariance du bruit $\mathbf{\Sigma}_b(n,f)$ n'est pas estimée durant les périodes de présence de la parole. La covariance du bruit $\mathbf{\Sigma}_b(n,f)$ est alors mal estimée, notamment si les caractéristiques du bruit $\mathbf{b}(n,f)$ varient dans le temps. L'idée consiste à introduire une probabilité de présence de la parole $p(n,f)$ (SPP, *speech presence probability*) pour obtenir une décision douce et permet de pallier ce problème [McAulay et Malpass, 1980; Ephraim et Malah, 1984; Martin, 2001; Cohen, 2003; Souden et al., 2010]. Toutefois, les méthodes de soustraction spectrale produisent une mauvaise estimation de la covariance de la parole locale $\mathbf{\Sigma}_s(n,f)$ avec (2.21), qui n'est en particulier pas valable pour des SNRs faibles.

2.2.1.4. Estimation des covariances par séparation de sources

Les méthodes de séparation de sources modélisent à la fois la covariance du bruit $\mathbf{\Sigma}_b(n,f)$ et celle de la parole locale $\mathbf{\Sigma}_s(n,f)$, en optimisant le critère du maximum de vraisemblance (MV). Ces méthodes utilisent une approche probabiliste qui consiste à faire non pas l'hypothèse sur la relation de filtrage dans (2.16) mais plutôt une hypothèse sur la distribution des signaux. Très souvent, on suppose que les signaux sont des variables gaussiennes de moyenne nulle et de matrice de variance $\mathbf{\Sigma}_c(n,f)$:

$$\mathbf{c}(n,f) \sim \mathcal{N}\big(\mathbf{0}, \mathbf{\Sigma}_c(n,f)\big). \tag{2.22}$$

Dans le cas de la réduction de bruit, le mélange $\mathbf{d}^{\text{bruit}}(n,f)$ est alors une variable gaussienne de moyenne nulle et covariance $\mathbf{\Sigma}_s(n,f) + \mathbf{\Sigma}_b(n,f)$:

$$\mathbf{d}^{\text{bruit}}(n,f) \sim \mathcal{N}\Big(\mathbf{0}, \mathbf{\Sigma}_s(n,f) + \mathbf{\Sigma}_b(n,f)\Big). \tag{2.23}$$

Les matrices de covariance $\mathbf{\Sigma}_s(n,f)$ et $\mathbf{\Sigma}_b(n,f)$ sont obtenues en maximisant la vraisemblance du mélange $\mathbf{d}^{\text{bruit}}(n,f)$. Puis, la parole locale $\mathbf{s}(n,f)$ est estimée au sens du

MMSE par filtrage de Wiener :

$$\widehat{\mathbf{s}}(n,f) = \mathbb{E}[\mathbf{s}(n,f)|\mathbf{d}^{\text{bruit}}(n,f)] \tag{2.24}$$

$$= \underbrace{\boldsymbol{\Sigma}_s(n,f)\Big(\boldsymbol{\Sigma}_s(n,f)+\boldsymbol{\Sigma}_b(n,f)\Big)^{-1}}_{=\mathbf{W}_s^{\text{WF}}(n,f)}\mathbf{d}^{\text{bruit}}(n,f). \tag{2.25}$$

Modèle de la covariance Sous l'approximation en bande étroite, le processus de propagation de la parole locale $\mathbf{s}(n,f)$ dans (2.12) se ramène, comme dans (2.16), à une simple opération sur la trame n :

$$\mathbf{s}(n,f) = \mathbf{a}_s(f)u(n,f), \tag{2.26}$$

où $\mathbf{a}_s(f) \in \mathbb{C}^{M\times 1}$. La covariance de la parole locale $\boldsymbol{\Sigma}_s(n,f)$ s'exprime donc par :

$$\boldsymbol{\Sigma}_s(n,f) = v_s(n,f)\mathbf{R}_s(f), \tag{2.27}$$

où le scalaire $v_s(n,f) = \mathbb{E}[|u(n,f)|^2] \in \mathbb{R}_+$ est la *densité spectrale de puissance* (DSP), qui contient les paramètres spectraux de la parole locale, et la matrice de rang 1 $\mathbf{R}_s(f) = \mathbf{a}_s(f)\mathbf{a}_s(f)^H \in \mathbb{C}^{M\times M}$ est la *matrice de covariance spatiale* (MCS), qui contient les paramètres spatiaux, c'est-à-dire les paramètres sur le processus de propagation de la parole locale.

Toutefois, la formulation de la RIR de la parole locale sous l'approximation en bande étroite dans (2.26) est une représentation imparfaite du processus de propagation dans des milieux très réverbérants, où la RIR $\mathbf{a}_s(\tau)$ est beaucoup plus longue que la taille des trames T_{TFCT}. Le *modèle gaussien local* permet de conserver l'approximation en bande étroite pour appliquer le filtre court $\mathbf{W}_s(n,f)$ en supposant la MCS $\mathbf{R}_s(f)$, non plus comme étant de rang 1 mais comme étant de rang plein [Duong et al., 2010]. Ce modèle est aussi appliqué au bruit $\mathbf{b}(n,f)$, ce qui permet de modéliser la covariance d'un bruit diffus, où l'énergie sonore est uniformément répartie dans l'espace.

Estimation des paramètres Les DSPs $v_c(n,f)$ et les MCSs $\mathbf{R}_c(f)$ sont estimées conjointement selon le critère du MV à l'aide d'un algorithme d'espérance-maximisation (EM). À l'étape E, les signaux $\mathbf{c}(n,f)$ sont estimés par filtrage de Wiener

$$\widehat{\mathbf{c}}(n,f) = \mathbf{W}_c^{\text{WF}}(n,f)\mathbf{d}^{\text{bruit}}(n,f), \tag{2.28}$$

où $\mathbf{W}_c^{\text{WF}}(n,f)$ représente le filtre de Wiener pour la source $\mathbf{c}$, et les moments d'ordre 2 non centrés $\widehat{\boldsymbol{\Sigma}}_c(n,f)$ sont estimés de la manière suivante :

$$\widehat{\boldsymbol{\Sigma}}_c(n,f) = \widehat{\mathbf{c}}(n,f)\widehat{\mathbf{c}}^H(n,f) + \Big(\mathbf{I} - \mathbf{W}_c^{\text{WF}}(n,f)\Big)\, v_c(n,f)\mathbf{R}_c(f), \tag{2.29}$$

où $\mathbf{I}$ représente la matrice identité, dont la dimension est égale à M dans ce contexte particulier. A l'étape M, les DSPs et MCSs sont mises à jour de la manière suivante :

$$v_c(n,f) = \frac{1}{M}\text{tr}\left(\mathbf{R}_c^{-1}(f)\widehat{\boldsymbol{\Sigma}}_c(n,f)\right), \tag{2.30}$$

$$\mathbf{R}_c(f) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{v_c(n,f)}\widehat{\boldsymbol{\Sigma}}_c(n,f), \tag{2.31}$$

où tr(·) représente la trace d'une matrice. L'algorithme EM fait partie des méthodes d'estimation dites *hors ligne*, c'est-à-dire qu'elles utilisent à la fois les observations passées et futures des signaux. Dans le cas où les conditions acoustiques varient dans le temps, par exemple lorsque les sources se déplacent, l'estimation des MCSs dans (2.31) ne permet pas de prendre en compte le fait que les MCSs $\mathbf{R}_c(n, f)$ varient dans le temps. Il existe des versions incrémentales de l'algorithme EM qui procèdent à un traitement en temps réel et sont adaptées aux situations où les conditions acoustiques varient au cours du temps [Togami, 2011; Simon et Vincent, 2012]. Toutefois, l'algorithme EM exact opère indépendamment sur chaque bande de fréquence f ce qui conduit à une ambiguïté de permutation des composantes de la parole et du bruit à chaque bande de fréquence f [Duong et al., 2010].

Cette ambiguïté peut être résolue avec des méthodes de post-traitement basées sur la corrélation des spectres entre les bandes de fréquence [Sawada et al., 2004], sur des contraintes de parcimonie sur les DSPs $v_c(n, f)$ [Yılmaz et Rickard, 2004; Winter et al., 2007], ou sur des modèles spectraux, comme la factorisation matricielle positive (NMF, en anglais *nonnegative matrix factorization*) [Lee et Seung, 1999]. Il est généralement préférable d'introduire des modèles spectraux dans un cadre d'estimation conjointe. La NMF, qui est l'une des méthodes les plus courantes en séparation de sources, consiste à approximer le spectrogramme d'une source $\mathbf{c}$ par le produit de deux matrices à valeurs positives [Smaragdis, 2007; Févotte et al., 2009; Ozerov et Fevotte, 2010; Ozerov et al., 2012; Sawada et al., 2013; Weninger et al., 2014b]. D'autres modèles spectraux ont aussi été proposés, tels que les modèles de mélanges gaussiens [Attias, 2003], l'analyse en composantes indépendantes [Buchner et al., 2005], les modèles de Markov cachés [Higuchi et Kameoka, 2015], les modèles de noyaux additifs [Liutkus et al., 2014], ainsi que des distributions *a priori* des DSPs $v_c(n, f)$ [Duong et al., 2011; Leglaive et al., 2016].

2.2.1.5. Apprentissage profond pour la réduction de bruit

Les méthodes basées sur l'apprentissage profond ont montré une meilleure capacité d'estimation de la covariance du bruit $\mathbf{\Sigma}_b(n, f)$ et de celle de la parole locale $\mathbf{\Sigma}_s(n, f)$ par rapport aux méthodes classiques de soustraction spectale et de séparation de sources. Nous rappelons tout d'abord très brièvement le fonctionnement général des réseaux de neurones, puis nous détaillons l'ensemble des approches basées sur l'apprentissage profond en réduction de bruit.

Réseaux de neurones profonds Un neurone artificiel est un modèle mathématique inspiré du fonctionnement des neurones biologiques, et qui effectue une opération élémentaire [McCulloch et Pitts, 1943]. Il calcule d'abord une somme pondérée des entrées et applique au résultat obtenu une fonction g dite d'activation :

$$o_j = g\Big(\mathbf{w}_j^T \mathbf{i} + b_j\Big), \tag{2.32}$$

où le vecteur $\mathbf{i}$ représente l'ensemble des entrées, et le vecteur $\mathbf{w}_j$ et le scalaire b_j les paramètres du neurone. Les deux vecteurs $\mathbf{i}$ et $\mathbf{w}_j$ sont de même dimension. La fonction

d'activation g a pour but d'introduire une non-linéarité entre les entrées $\mathbf{i}$ et la sortie o_j. En pratique, on choisit des fonctions qui ont un effet de seuillage, comme les fonctions sigmoïde, tangente hyperbolique ou ReLU (en anglais *rectified linear unit*). Le schéma d'un neurone est représenté sur la figure 2.8.

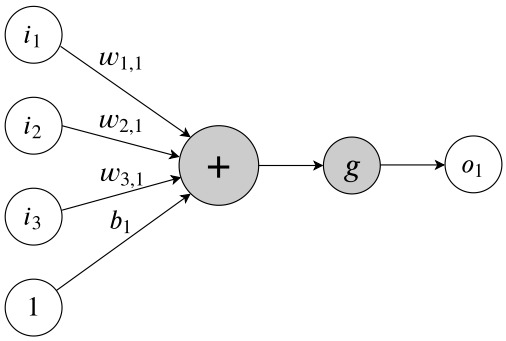


FIGURE 2.8. – Schéma d'un neurone.

On construit de cette manière un ensemble de J sorties o_j qu'on désigne par couche. L'idée consiste ensuite à connecter plusieurs couches à la suite pour combiner les non-linéarités et approcher une relation complexe entre les entrées et les cibles qu'on souhaite estimer. On désigne la dernière couche par le terme de *couche de sortie*, et toutes autres par *couches cachées*. Un réseau de neurones est qualifié de « profond » lorsqu'il possède plusieurs couches cachées. La figure 2.9 illustre un exemple de réseau de neurones.

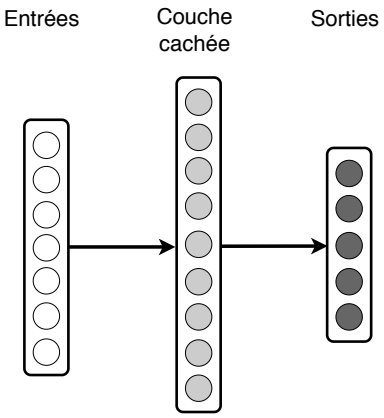


FIGURE 2.9. – Exemple de réseau de neurones à une couche cachée.

Les paramètres du réseau sont estimés, durant la phase d'apprentissage, en minimisant une fonction de coût entre les sorties du réseau et les cibles. Une des fonctions de coût les plus utilisés pour la régression est l'erreur quadratique moyenne. En général, comme l'ensemble d'apprentissage est très grand, on utilise des algorithmes de type descente de gradient stochastique, qui calculent le gradient sur des petits sous-ensembles de données appelés *mini-lots*. Il existe de nombreuses architectures de réseaux de neurones. Les

plus communes sont le perceptron multicouches (MLP, *multilayer perceptron*), le réseau de neurones récurrent (RNN, *recurrent neural network*) avec sa variante *long short-term memory* (LSTM), et le réseau de neurones convolutif (CNN, *convolutional neural network*).

Estimation des covariances par apprentissage profond Récemment, l'utilisation des réseaux de neurones s'est imposé comme l'état de l'art en réduction de bruit. Un premier type de méthodes, résumé dans le tableau 2.1, utilise les réseaux de neurones pour estimer les covariances de la parole locale $\mathbf{\Sigma}_s(n, f)$ et du bruit $\mathbf{\Sigma}_b(n, f)$. Inspirés des méthodes de soustraction spectrale (voir la partie 2.2.1.3), certains travaux se sont intéressés à l'estimation de la SPP en monocanal [Xia et Bao, 2013] et en multicanal [Heymann et al., 2016; Erdogan et al., 2016; Ochiai et al., 2017; Wang et al., 2018a]. Toutefois, ces méthodes souffrent de l'approximation en bande étroite en multicanal. D'autres travaux ont abordé le problème du point de vue de la séparation de sources (voir la partie 2.2.1.4), en estimant directement les DSPs de la parole locale $v_s(n, f)$ et du signal de bruit $v_b(n, f)$ [Araki et Nakatani, 2011; Huang et al., 2015; Nugraha et al., 2016a; Seki et al., 2018; Leglaive et al., 2019], ou en estimant certains paramètres de la NMF [Kang et al., 2015; Le Roux et al., 2015].

D'autres travaux utilisent les DNNs pour construire une représentation des spectres du bruit et de la parole locale dans un espace de grande dimension, puis à regrouper les points similaires dans cet espace à l'aide d'un algorithme de partitionnement, comme les k-moyennes, pour estimer le filtre $\mathbf{W}_s(n, f)$. De manière similaire à la NMF, ce type de méthodes détermine cette représentation des spectres des signaux en minimisant un critère de type MMSE lié à la reconstruction des spectres de la parole et du bruit [Hershey et al., 2016; Wang et al., 2018c; Chen et al., 2017]. Ces méthodes souffrent du problème d'ambiguïté de permutation, qui peut être résolu en modifiant le critère d'apprentissage [Yu et al., 2017; Luo et al., 2018]. Toutefois, cette catégorie de méthodes a une complexité plus élevée que la première catégorie à cause des étapes de représentation des spectres des signaux et de partitionnement, notamment dans le cas multicanal [Wang et al., 2018b; Drude et al., 2018].

Estimation directe par apprentissage profond Un deuxième type de méthodes, résumé dans le tableau 2.2, estime directement soit le filtre $\mathbf{W}_s(n, f)$, soit la parole locale $\mathbf{s}(n, f)$. En monocanal, les méthodes qui estiment directement le filtre $\mathbf{W}_s(n, f)$ peuvent utiliser d'autres critères que le filtre de Wiener, comme le masque idéal de rapport (MIR) et le masque idéal d'amplitude (MIA), qui utilisent uniquement l'amplitude des signaux [Weninger et al., 2014a; Wang et al., 2014], ou comme le filtre sensible à la phase (FSP) et le masque complexe [Erdogan et al., 2015; Williamson et al., 2016], qui utilisent également la phase des signaux. En particulier, ces derniers ont montré de meilleures performances que les filtres utilisant uniquement l'amplitude des signaux. En multicanal, quelques travaux se sont aussi intéressés à la prédiction du filtre $\mathbf{W}_s(n, f)$, mais présentent de mauvais résultats par rapport au premier type de méthodes basées sur l'apprentissage profond [Xiao et al., 2016; Sainath et al., 2017].

		SPP	**DSP**	**Représentation + partitionnement**
Mono-canal	Références	[Xia et Bao, 2013]	[Huang et al., 2015] [Kang et al., 2015] [Le Roux et al., 2015] [Leglaive et al., 2019]	[Hershey et al., 2016] [Chen et al., 2017] [Yu et al., 2017] [Wang et al., 2018c] [Luo et al., 2018]
	Limitation	Estimation indirecte		Complexité élevée
	Inspiration pour la thèse	Non		
Multi-canal	Références	[Heymann et al., 2016] [Erdogan et al., 2016] [Ochiai et al., 2017] [Wang et al., 2018a]	[Araki et Nakatani, 2011] [Nugraha et al., 2016a] [Seki et al., 2018]	[Wang et al., 2018b] [Drude et al., 2018]
	Limitation	Approx. bande étroite	-	Complexité élevée
	Inspiration pour la thèse	Non	Oui	Non

TABLEAU 2.1. – Méthodes de réduction de bruit basées sur l'estimation des covariances par apprentissage profond.

Enfin, d'autres travaux estiment le signal de la parole locale $\mathbf{s}(n,f)$ en modélisant directement le spectre en amplitude $|\mathbf{s}(n,f)|$ par un DNN et en multipliant ce spectre en amplitude par la phase du mélange $\theta_d(n,f)$ [Lu et al., 2013; Park et Lee, 2017; Tan et Wang, 2018]. Pour inclure l'estimation de la phase, des méthodes bout-en-bout estiment le signal de la parole dans le domaine temporel directement dans le cas monocanal [Grais et al., 2018; Rethage et al., 2018; Luo et Mesgarani, 2019; Shi et al., 2019]. Toutefois, le cas multicanal n'a pas été traité.

2.2.2. Réduction d'écho acoustique

Le problème de la réduction d'écho est illustré sur la figure 2.10. Le signal $\mathbf{d}^{\text{echo}}(n,f)$ observé aux microphones est la somme de la parole locale $\mathbf{s}(n,f)$ et de l'écho acoustique $\mathbf{y}(n,f) \in \mathbb{C}^{M\times 1}$:

$$\mathbf{d}^{\text{echo}}(n,f) = \mathbf{s}(n,f) + \mathbf{y}(n,f). \tag{2.33}$$

La réduction d'écho consiste à extraire la parole locale $\mathbf{s}(n,f)$ en soustrayant tout d'abord une estimation de l'écho $\widehat{\mathbf{y}}(n,f)$ au signal du mélange $\mathbf{d}^{\text{echo}}(n,f)$ [Hänsler et Schmidt, 2004, Chapitre 9] :

$$\mathbf{e}^{\text{echo}}(n,f) = \mathbf{d}^{\text{echo}}(n,f) - \widehat{\mathbf{y}}(n,f), \tag{2.34}$$

où $\widehat{\mathbf{y}}(n,f)$ est obtenu en appliquant un filtre causal $\mathcal{H}(f)$ sur la parole distante $x(n,f)$, qui est le signal source connu de l'écho $\mathbf{y}(n,f)$. Le filtre $\mathcal{H}(f)$ est censé reproduire le chemin d'écho $\mathbf{a}_y(k,f',f) \in \mathbb{C}^{M\times 1}$ comme dans (2.6). Très souvent dans la littérature,

		Filtre	**Signal**
Monocanal	Références	[Weninger et al., 2014a] [Wang et al., 2014] [Erdogan et al., 2015] [Williamson et al., 2016]	[Lu et al., 2013] [Park et Lee, 2017] [Tan et Wang, 2018] [Grais et al., 2018] [Rethage et al., 2018] [Luo et Mesgarani, 2019] [Shi et al., 2019]
	Limitation	Invariance à l'échelle des signaux du mélange	Boîte noire
	Inspiration pour la thèse	Oui	Non
Multicanal	Références	[Xiao et al., 2016] [Sainath et al., 2017]	-
	Limitation	Réduction de bruit médiocre	-
	Inspiration pour la thèse	Non	

TABLEAU 2.2. – Méthodes de réduction de bruit basées sur l'estimation directe du filtre ou de la parole locale par apprentissage profond.

cette opération de filtrage est exprimée dans le domaine temporel [Hänsler et Schmidt, 2004, Chapitre 9]. Toutefois, pour réduire la complexité du filtrage dans le domaine temps-fréquence, il est possible de faire l'*approximation en sous-bande*, qui consiste à supposer que les bandes de fréquence peuvent être traitées séparément par le filtre $\mathcal{H}(f)$, c'est-à-dire que les coefficients du filtre $\mathbf{h}(k, f', f)$ sont égaux à zéro pour $f' \neq f$. La convolution en trame et en fréquence comme dans (2.15) se ramène à une convolution en trame [Hänsler et Schmidt, 2004, Chapitre 9] :

$$\widehat{\mathbf{y}}(n, f) \approx \sum_{k=0}^{K} \mathbf{h}(k, f) x(n-k, f), \tag{2.35}$$

où $\mathbf{h}(k, f) \in \mathbb{C}^{M \times 1}$ est le vecteur de dimension M correspondant au k-ième délai du filtre $\mathcal{H}(f)$, et K est la longueur du filtre qui, idéalement, doit être au moins aussi grande que la longueur du chemin d'écho $\mathbf{a}_y(k, f', f)$. On parle de filtre *long* pour désigner $\mathcal{H}(f) = [\mathbf{h}(0, f) \dots \mathbf{h}(K-1, f)] \in \mathbb{C}^{M \times K}$ de manière générale, et d'*annulation* pour désigner le filtrage par filtre long. Contrairement à la réduction de bruit, l'approximation en bande étroite ne peut pas être utilisée comme dans (2.16) pour le filtre $\mathcal{H}(f)$ car le chemin d'écho $\mathbf{a}_y(k, f', f)$ est beaucoup plus long que la taille des trames T_{TFCT}. Comme les conditions acoustiques varient lentement par rapport aux spectres des signaux, le filtre long $\mathcal{H}(f)$ est qualifié de *linéaire*. Ce filtre n'introduit que très peu d'artefacts dans la parole locale $\mathbf{s}(n, f)$ car il varie lentement au cours du temps.

En pratique, le signal après annulation d'écho $\mathbf{e}^{\text{echo}}(n, f)$ n'est pas égal à la parole locale $\mathbf{s}(n, f)$ pour trois raisons. Premièrement, ceci est dû à l'erreur d'estimation du filtre

$\mathcal{H}(f)$. Deuxièmement, la longueur du filtre $\mathcal{H}(f)$ en pratique est souvent plus petite que celle du chemin d'écho $\mathbf{a}_y(k, f', f)$. Dès lors, une composante de l'écho $\mathbf{y}(n, f)$ subsiste même après un filtrage idéal, qui est parfois désigné par le terme d'*écho résiduel tardif* $\mathbf{y}_l(n, f)$ [Habets et al., 2008b]. Enfin, des non-linéarités sont présentes dans l'écho $\mathbf{y}(n, f)$ (voir la partie 2.1.1.4), et ne peuvent pas être modélisées par (2.35). Par conséquent, un écho résiduel $\mathbf{z}(n, f)$ subsiste et s'exprime de la manière suivante [Hänsler et Schmidt, 2004, Chapitre 10] :

$$\mathbf{e}^{\text{echo}}(n, f) - \mathbf{s}(n, f) = \underbrace{\mathbf{y}(n, f) - \widehat{\mathbf{y}}(n, f)}_{=\mathbf{z}(n,f)}. \tag{2.36}$$

La parole locale $\mathbf{s}(n, f)$ est extraite en appliquant un post-filtre court $\mathbf{W}_s(n, f)$ sur le signal $\mathbf{e}^{\text{echo}}(n, f)$, comme pour la suppression de bruit [Hänsler et Schmidt, 2004, Chapitre 10] :

$$\widehat{\mathbf{s}}(n, f) = \mathbf{W}_s(n, f)\mathbf{e}^{\text{echo}}(n, f). \tag{2.37}$$

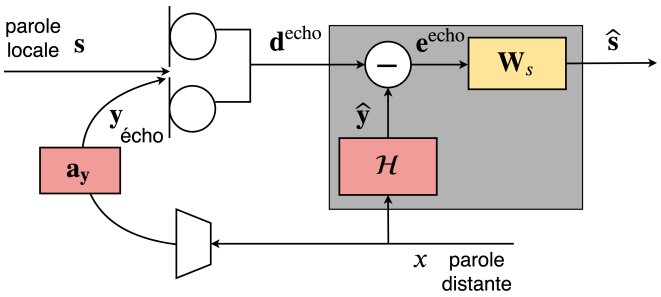


FIGURE 2.10. – Problème de réduction de l'écho acoustique.

Nous présentons tout d'abord les méthodes d'annulation d'écho, qui estiment le filtre $\mathcal{H}(f)$, puis les méthodes de suppression d'écho résiduel, qui estiment le post-filtre $\mathbf{W}_s(n, f)$, et enfin les méthodes alternatives de réduction d'écho.

2.2.2.1. Annulation d'écho

Critère de détermination du filtre De même que pour la réduction de bruit, le critère fréquemment utilisé pour déterminer le filtre $\mathcal{H}(f)$ est le critère MMSE :

$$\min_{\mathcal{H}(f)} \sum_{n=0}^{N-1} \left\| \mathbf{y}(n, f) - \sum_{k=0}^{K} \mathbf{h}(k, f)x(n-k, f) \right\|^2. \tag{2.38}$$

On suppose que la parole locale $\mathbf{s}(n, f)$ et l'écho $\mathbf{y}(n, f)$ sont indépendants. Comme la parole distante $x(n, f)$ est connue, le problème de minimisation dans (2.38) est donc équivalent à :

$$\min_{\mathcal{H}(f)} \sum_{n=0}^{N-1} \left\| \mathbf{e}^{\text{echo}}(n, f) \right\|^2. \tag{2.39}$$

Le problème d'optimisation multicanale dans (2.39) peut se décomposer en un problème d'optimisation séparé sur chaque canal m. Par conséquent, dans la suite de cette sous-partie, nous considérons dans le cas monophonique ($M = 1$) et l'indice m du microphone est omis par souci de clarté. La solution de ce problème d'optimisation est alors :

$$\mathcal{H}^{\mathrm{WF}}(f) = \frac{\sum_{n=0}^{N-1} \|y(n,f)\underline{\mathbf{x}}(n,f)\|^2}{\sum_{n=0}^{N-1} \|\underline{\mathbf{x}}(n,f)\|^2}, \tag{2.40}$$

où $\underline{\mathbf{x}}(n,f) = [x(n,f)\dots x(n-K+1,f)] \in \mathbb{C}^{1\times K}$. La notation soulignée dans $\underline{\mathbf{x}}(n,f)$ désigne la concaténation des K précédentes trames de la parole distante $x(n,f)$.

Filtrage adaptatif Pour réduire la complexité de (2.40) et ajuster le filtre $\mathcal{H}(f)$ en cas de variations du chemin d'écho $a_y(k,f',f)$ au cours du temps, le filtre $\mathcal{H}(f)$ est estimé à l'aide d'un algorithme adaptatif basé sur le critère MMSE, qui ajuste le filtre $\mathcal{H}(f)$ de manière incrémentale par descente de gradient stochastique [Haykin, 2002, Chapitre 9]. Un des types d'algorithmes les plus utilisés est celui des moindres carrés normalisés (NLMS, *normalized least mean squares*) [Hänsler et Schmidt, 2004, Chapitre 7] :

$$\mathcal{H}(n+1,f) = \mathcal{H}(n,f) - \mu(n,f)\boldsymbol{\nabla}_{\mathcal{H}} \left|e^{\mathrm{echo}}(n,f)\right|^2, \tag{2.41}$$

où

$$\boldsymbol{\nabla}_{\mathcal{H}} \left|e^{\mathrm{echo}}(n,f)\right|^2 \approx \frac{e^{\mathrm{echo}}(n,f)\underline{\mathbf{x}}(n,f)}{\|\underline{\mathbf{x}}(n,f)\|^2}, \tag{2.42}$$

et $\mu(n,f)$ est un pas d'adaptation variant dans le temps qui permet de contrôler les vitesses d'adaptation et de convergence du filtre $\mathcal{H}(f)$. Il existe d'autres types d'algorithmes adaptatifs basés sur le critère MMSE, comme la projection affine, les moindres carrés récursifs (RLS, *recursive least squares*) et le filtre de Kalman, qui modélisent une «mémoire» des signaux passés pour la mise à jour du filtre $\mathcal{H}(f)$ [Hänsler et Schmidt, 2004, Chapitre 7].

Si le pas d'adaptation $\mu(n,f)$ est trop petit, la convergence du filtre $\mathcal{H}(f)$ sera lente. Si le pas d'adaptation $\mu(n,f)$ est trop grand, le filtre $\mathcal{H}(f)$ ne peut pas être ajusté aux variations brusques de la parole distante $x(n,f)$ ou du chemin d'écho $a_y(k,f',f)$ au cours du temps. Le pas d'adaptation optimal est le suivant [Mader et al., 2000] :

$$\mu^{\mathrm{opt}}(n,f) = \frac{\Sigma_z(n,f)}{\widehat{\mathbb{E}}\left[|e^{\mathrm{echo}}(n,f)|^2\right]}, \tag{2.43}$$

où $\Sigma_z(n,f) = \mathbb{E}[|z(n,f)|^2]$ est la covariance de l'écho résiduel en monocanal. Lorsque le locuteur local est actif, l'écho résiduel $z(n,f)$ n'est plus observé, et la covariance de l'écho résiduel $\Sigma_z(n,f)$ nécessite d'être estimée. Le principal problème pour obtenir le pas d'adaptation optimal $\mu^{\mathrm{opt}}(n,f)$ est donc dû à la présence de la parole locale $s(n,f)$, qui joue le rôle de signal perturbant l'estimation de $\Sigma_z(n,f)$. De plus, le caractère perturbateur de la parole locale $s(n,f)$ devient important au fur et à mesure que le filtre $\mathcal{H}(f)$ converge, et que l'énergie de l'écho résiduel $z(n,f)$ diminue par rapport à la parole locale $s(n,f)$.

Contrôle de l'adaptation du filtre En général, pour estimer $\mu^{\text{opt}}(n,f)$, les méthodes d'annulation d'écho ne modélisent que la composante de $\Sigma_z(n,f)$ due à l'erreur d'estimation du filtre $\mathcal{H}(f)$, et négligent les composantes de l'écho résiduel dues aux non-linéarités et à l'écho résiduel tardif $\mathbf{y}_{\text{l}}(n,f)$.

Un premier type de solution consiste à estimer $\mu^{\text{opt}}(n,f)$ en utilisant un détecteur de double parole (DTD, *double-talk detector*) qui est une méthode de classification similaire à un VAD, permettant de détecter la présence simultanée de la parole locale $s(n,f)$ et de l'écho $y(n,f)$ [Sondhi, 1967]. Lorsqu'une période de double parole est détectée, le pas d'adaptation $\mu(n,f)$ est réduit ou fixé à zéro pour ralentir ou figer l'adaptation du filtre $\mathcal{H}(f)$ [Mader et al., 2000]. Toutefois, ces méthodes de classification se basent sur un seuil de détection qui effectue un compromis entre *faux positifs* (double parole détectée à tort) et *faux négatifs* (double parole non détectée à tort). Ceci entraîne un délai entre la détection de la double parole et l'adaptation du filtre $\mathcal{H}(f)$, ce qui peut perturber considérablement l'estimation du filtre $\mathcal{H}(f)$.

Un deuxième type de solution consiste à estimer $\mu^{\text{opt}}(n,f)$ avec une méthode de régression appliquée à la covariance de la parole distante $\widehat{\mathbb{E}}[|x(n,f)|^2]$ [Paleologu et al., 2015]. Ce type de solution est généralement plus robuste aux périodes de double parole que les solutions utilisant un DTD. Plusieurs méthodes ont été proposées, utilisant soit un modèle linéaire [Myllylä, 2006; Valin, 2007; Luis Valero et Habets, 2017], soit un modèle non-linéaire [Paleologu et al., 2008; Wada et Juang, 2012].

Enfin, un troisième de type de solution utilise le filtrage de Kalman pour l'adaptation du filtre $\mathcal{H}(f)$ plutôt que l'algorithme NLMS [Paleologu et al., 2013]. Enzner et Vary [2006] ont proposé pour la première fois cette modélisation équivalente à la mise à jour du filtre $\mathcal{H}(f)$ dans (2.41), et ne nécessitant aucun pas d'adaptation $\mu(n,f)$. Paleologu et al. [2015] ont étudié l'équivalence entre l'algorithme NLMS et le filtrage de Kalman pour l'annulation d'écho.

2.2.2.2. Suppression d'écho résiduel

Critère de détermination du post-filtre Comme en réduction de bruit, on utilise fréquemment le filtre de Wiener pour déterminer le post-filtre $\mathbf{W}_s(n,f)$ [Hänsler et Schmidt, 2004, Chapitre 10] :

$$\mathbf{W}_s(n,f) = \mathbf{\Sigma}_s(n,f)\Big(\mathbf{\Sigma}_s(n,f) + \mathbf{\Sigma}_z(n,f)\Big)^{-1}, \tag{2.44}$$

où $\mathbf{\Sigma}_z(n,f)$ est la covariance de l'écho résiduel, définie comme dans (2.20). Le post-filtre $\mathbf{W}_s(n,f)$ introduit des artefacts dans la parole locale estimée $\widehat{\mathbf{s}}(n,f)$ (voir la partie 2.2.1.2).

Soustraction spectrale Généralement, le post-filtre $\mathbf{W}_s(n,f)$ est estimé à l'aide de méthodes de soustraction spectrale [Hänsler et Schmidt, 2004, Chapitre 10]. De même qu'en réduction de bruit (voir la partie 2.2.1.3), ces méthodes modélisent la covariance de l'écho résiduel $\mathbf{\Sigma}_z(n,f)$ et déduisent la covariance de la parole locale $\mathbf{\Sigma}_s(n,f)$ par

soustraction spectrale :

$$\mathbf{\Sigma}_s(n,f) = \widehat{\mathbb{E}}\Big[\mathbf{e}^{\text{echo}}(n,f)\mathbf{e}^{\text{echo}}(n,f)^H\Big] - \mathbf{\Sigma}_z(n,f). \tag{2.45}$$

Contrairement à la réduction de bruit, ces méthodes peuvent modéliser la covariance de l'écho résiduel $\mathbf{\Sigma}_z(n,f)$ en appliquant une transformation à un signal qui ne contient pas la parole locale $\mathbf{s}(n,f)$. Généralement, cette transformation est réalisée soit sur la parole distante $x(n,f)$ [Gustafsson et al., 2002; Chhetri et al., 2005; Lee et Kim, 2007; Bendersky et al., 2008; Schwarz et al., 2013; Valero et al., 2014], soit sur l'écho estimé $\widehat{\mathbf{y}}(n,f)$ [Turbin et al., 1997; Gustafsson et al., 2002; Hoshuyama et Sugiyama, 2006; Habets et al., 2008b]. Parmi ces méthodes, certaines incluent une représentation explicite des non-linéarités avec un modèle polynomial [Chhetri et al., 2005; Bendersky et al., 2008] ou un DNN [Schwarz et al., 2013]. D'autres modélisent la composante due à l'écho résiduel tardif $\mathbf{y}_\text{l}(n,f)$, mais en négligeant les non-linéarités [Habets et al., 2008b; Valero et al., 2014]. Wung et al. [2011] utilisent plusieurs signaux d'entrée, à la fois le mélange $\mathbf{d}(n,f)$ et l'écho estimé $\widehat{\mathbf{y}}(n,f)$, sans toutefois modéliser explicitement les différentes composantes de l'écho résiduel $\mathbf{z}(n,f)$. Cependant, les méthodes de soustraction spectrale produisent une mauvaise estimation de la covariance de la parole locale $\mathbf{\Sigma}_s(n,f)$ avec (2.45), en particulier pour des rapports signal-à-écho (SER, *signal-to-echo ratio*) faibles.

Récemment, Lee et al. [2015] ont proposé une méthode qui estime directement le post-filtre $\mathbf{W}_s(n,f)$ avec un DNN utilisant en entrée les signaux $\mathbf{e}^{\text{echo}}(n,f)$ et $x(n,f)$, et qui montre une meilleure performance que les méthodes de soustraction spectrale. Toutefois, la méthode de Lee et al. [2015] et les méthodes de soustraction spectrale estiment le post-filtre $\mathbf{W}_s(n,f)$ séparément du filtre $\mathcal{H}(f)$. En particulier, elles supposent que le filtre $\mathcal{H}(f)$ a déjà convergé, ce qui n'est pas le cas dans des scénarios réels. Ainsi, elles négligent la composante de l'écho résiduel due à l'erreur d'estimation du filtre $\mathcal{H}(f)$, qui peut être importante par rapport aux non-linéarités et à la composante $\mathbf{y}_\text{l}(n,f)$ durant les périodes où le filtre $\mathcal{H}(f)$ n'a pas convergé.

Estimation conjointe du filtre et du post-filtre Pour une meilleure robustesse en performances durant les périodes où le filtre $\mathcal{H}(f)$ n'a pas convergé, certaines méthodes optimisent conjointement le filtre $\mathcal{H}(f)$ et le post-filtre $\mathbf{W}_s(n,f)$. Pour cela, ces méthodes se servent de l'estimation de la covariance de l'écho résiduel $\mathbf{\Sigma}_z(n,f)$ utilisée pour l'adaptation du filtre $\mathcal{H}(n,f)$ pour estimer le post-filtre $\mathbf{W}_s(n,f)$ par soustraction spectrale [Enzner et Vary, 2006; Myllylä, 2006]. Togami et Hori [2011] ont proposé une méthode inspirée de la séparation de sources pour optimiser conjointement le filtre $\mathcal{H}(f)$ et le post-filtre $\mathbf{W}_s(n,f)$. Ils modélisent la parole locale $\mathbf{s}(n,f)$ par une variable gaussienne de moyenne nulle comme dans (2.27), et l'écho résiduel $\mathbf{z}(n,f)$ par une variable gaussienne de moyenne non nulle. Ils optimisent les paramètres de ce modèle selon le critère du MV. Toutefois, ces méthodes n'imposent pas de contrainte sur la covariance de la parole locale $\mathbf{\Sigma}_s(n,f)$, et ne modélisent pas les composantes de l'écho résiduel $\mathbf{z}(n,f)$ dues à l'écho résiduel tardif $\mathbf{y}_\text{l}(n,f)$ et à la présence de non-linéarités.

2.2.2.3. Méthodes alternatives

En sus des méthodes présentées ci-dessus, il existe deux types de méthodes alternatives pour la réduction d'écho : la suppression d'écho et l'annulation d'écho non-linéaire.

Suppression d'écho Ce type de méthodes consiste à appliquer un filtre de suppression d'écho $\mathbf{W}_s(n,f)$ sur le signal de mélange $\mathbf{d}^{\text{echo}}(n,f)$, à la place du filtre d'annulation d'écho $\mathcal{H}(f)$:

$$\widehat{\mathbf{s}}(n,f) = \mathbf{W}_s(n,f)\mathbf{d}^{\text{echo}}(n,f). \tag{2.46}$$

Ces méthodes ont une complexité plus faible et une vitesse de convergence supérieure par rapport à la combinaison d'un filtre d'annulation d'écho et d'un post-filtre de suppression d'écho résiduel. Le filtre de Wiener s'exprime de la manière suivante :

$$\mathbf{W}_s^{\text{WF}}(n,f) = \mathbf{\Sigma}_s(n,f)\Big(\mathbf{\Sigma}_s(n,f) + \mathbf{\Sigma}_y(n,f)\Big)^{-1}, \tag{2.47}$$

où $\mathbf{\Sigma}_y(n,f)$ est la covariance de l'écho, définie comme dans (2.20). Comme pour la suppression d'écho résiduel, la plupart de ces méthodes se basent sur la soustraction spectrale. La covariance de l'écho $\mathbf{\Sigma}_y(n,f)$ peut être estimée à l'aide d'algorithmes adaptatifs [Avendano, 2001; Faller et Chen, 2005; Favrot et al., 2012] ou de méthodes basées sur la SPP [Hoshuyama et Sugiyama, 2006; Park et Chang, 2009; Tong et Gu, 2016; Huang et al., 2016]. Parmi elles, certaines approches incluent un modèle pour les composantes de l'écho $\mathbf{y}(n,f)$, comme l'écho résiduel tardif $\mathbf{y}_{\text{l}}(n,f)$ [Favrot et al., 2012] et les non-linéarités [Hoshuyama et Sugiyama, 2006]. Récemment, Madrid Portillo [2017] a proposé une méthode d'estimation directe du filtre $\mathbf{W}_s(n,f)$ par DNN, qui permet de modéliser les non-linéarités présentes dans l'écho $\mathbf{y}(n,f)$ et qui est plus robuste que les méthodes de soustraction spectrale.

Toutefois, ces méthodes présentent deux limites par rapport à la combinaison d'un filtre d'annulation d'écho et d'un post-filtre de suppression d'écho résiduel. D'une part, comme le SER dans le mélange $\mathbf{d}^{\text{echo}}(n,f)$ est beaucoup plus faible que le SER dans le signal $\mathbf{e}^{\text{echo}}(n,f)$, le filtre de suppression d'écho $\mathbf{W}_s(n,f)$ introduit beaucoup plus d'artefacts dans la parole locale estimée $\widehat{\mathbf{s}}(n,f)$ (voir la partie 2.2.1.2). D'autre part, un écho résiduel subsiste tout de même après la suppression d'écho, et une méthode de suppression d'écho résiduel doit être appliquée [Hänsler et Schmidt, 2004, Chapitre 10]. Ces méthodes sont similaires aux méthodes de suppression d'écho résiduel après annulation d'écho.

Annulation d'écho non-linéaire Ce type de méthodes consiste à appliquer un filtre d'annulation d'écho $\mathcal{H}(f)$ qui modélise aussi les non-linéarités présentes dans l'écho $\mathbf{y}(n,f)$. Guérin et al. [2003] ont proposé d'étendre la relation polynomiale d'ordre 1 de l'annulation d'écho dans (2.35) par rapport à la parole distante $x(n,f)$ à un polynôme d'ordre plus élevé. Toutefois, la complexité de l'adaptation du filtre $\mathcal{H}(f)$ est trop élevée pour que cette méthode soit utilisée. Pour diminuer la complexité de l'adaptation, d'autres méthodes se basent sur le filtrage de Hammerstein, qui consiste à appliquer

une fonction non-linéaire sur la parole distante $x(n,f)$ avant d'appliquer un filtre linéaire d'annulation écho $\mathcal{H}(f)$ [Ngia et Sjobert, 1998; Costa et al., 2003; Scarpiniti et al., 2011]. Cette fonction non-linéaire peut être un polynôme [Ngia et Sjobert, 1998; Stenger et Kellermann, 2000; Costa et al., 2003; Kuech et al., 2005], une fonction sigmoïde [Scarpiniti et al., 2011], ou un DNN [Birkett et Goubran, 1995; Malek et Koldovský, 2016]. Cependant, les paramètres de la fonction non-linéaire et du filtre sont optimisés séparément. Pour les optimiser conjointement, d'autres méthodes se basent sur le filtrage de Kalman non-linéaire [Malik et Enzner, 2012; Huemmer et al., 2014b,a]. Toutefois, ces méthodes ont une complexité plus élevée que la combinaison d'un filtre d'annulation linéaire d'écho et d'un post-filtre de suppression d'écho résiduel. De plus, elles ne permettent pas de réduire l'écho résiduel tardif $\mathbf{y}_{\mathrm{l}}(n,f)$ à cause de la longueur du filtre $\mathcal{H}(f)$ plus petite que le chemin d'écho $\mathbf{a}_y(k,f',f)$.

2.2.3. Déréverbération

Le problème de déréverbération concerne la situation où le signal de mélange $\mathbf{d}^{\mathrm{rev}}(n,f)$ contient uniquement la parole locale, qui est composée de la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$ et de la réverbération tardive $\mathbf{s}_{\mathrm{l}}(n,f)$ (voir la partie 2.1.1.2) :

$$\mathbf{d}^{\mathrm{rev}}(n,f) = \mathbf{s}(n,f) = \mathbf{s}_{\mathrm{e}}(n,f) + \mathbf{s}_{\mathrm{l}}(n,f). \tag{2.48}$$

Nous nous focalisons sur les méthodes de déréverbération qui extraient la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$, sans aucune connaissance sur la RIR de la parole locale $\mathbf{a}_s(k,f',f)$. Nous détaillons tout d'abord les méthodes de suppression de réverbération, puis nous décrivons les méthodes de filtrage inverse, et enfin nous présentons succinctement les méthodes alternatives de déréverbération.

2.2.3.1. Suppression de réverbération

Le problème de la suppression de réverbération est illustré sur la figure 2.11. Il consiste à extraire la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$ en appliquant un filtre court $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$ sur le signal de mélange $\mathbf{s}(n,f)$, comme pour la suppression de bruit [Kodrasi et Doclo, 2018a] :

$$\widehat{\mathbf{s}}_{\mathrm{e}}(n,f) = \mathbf{W}_{s_{\mathrm{e}}}(n,f)\mathbf{s}(n,f). \tag{2.49}$$

Ces méthodes supposent que la réverbération tardive $\mathbf{s}_{\mathrm{l}}(n,f)$ est indépendante de la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$. Bien que cette hypothèse soit discutable, elle permet de formuler le filtre de Wiener, ou ses variantes, pour le filtre $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$:

$$\mathbf{W}_{s_{\mathrm{e}}}^{\mathrm{WF}}(n,f) = \mathbf{\Sigma}_{s_{\mathrm{e}}}(n,f)\Big(\mathbf{\Sigma}_{s_{\mathrm{e}}}(n,f) + \mathbf{\Sigma}_{s_{\mathrm{l}}}(n,f)\Big)^{-1}. \tag{2.50}$$

Le filtre $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$ nécessite un estimateur des covariances de la composante précoce $\mathbf{\Sigma}_{s_{\mathrm{e}}}(n,f)$ et de la réverbération tardive $\mathbf{\Sigma}_{s_{\mathrm{l}}}(n,f)$.

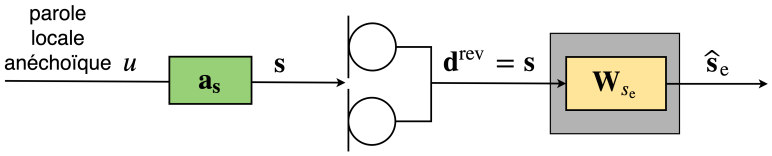


FIGURE 2.11. – Suppression de révebération.

Modélisation temporelle Les méthodes de modélisation temporelle exploitent la structure temporelle de la réverbération, qui introduit une forte corrélation entre les trames successives de la parole locale $\mathbf{s}(n,f)$, pour modéliser la covariance de la réverbération tardive $\mathbf{\Sigma}_{s_l}(n,f)$ [Naylor et Gaubitch, 2010, Chapitre 3]. Elles déduisent ensuite la covariance de la composante précoce $\mathbf{\Sigma}_{s_e}(n,f)$ par soustraction spectrale :

$$\mathbf{\Sigma}_{s_e}(n,f) = \widehat{\mathbb{E}}\left[\mathbf{s}(n,f)\mathbf{s}(n,f)^H\right] - \mathbf{\Sigma}_{s_l}(n,f). \tag{2.51}$$

Pour déterminer la covariance de la réverbération tardive $\mathbf{\Sigma}_{s_l}(n,f)$, certaines méthodes supposent la décroissance exponentielle de la RIR de la parole locale $\mathbf{a}_s(k,f',f)$ au cours du temps [Polack, 1988]. Par conséquent, ces méthodes modélisent la covariance de la réverbération tardive $\mathbf{\Sigma}_{s_l}(n,f)$ comme une version atténuée de la covariance de la parole locale $\mathbf{\Sigma}_s(n,f)$:

$$\mathbf{\Sigma}_{s_l}(n,f) = \mathbf{\Sigma}_s(n - N_{\text{rev}}, f)\exp(-\beta N_{\text{rev}}), \tag{2.52}$$

où β est un coefficient qui décrit la décroissance exponentielle de l'énergie de la parole locale, et N_{rev} est un délai associé à cette décroissance. Le coefficient β est estimé à partir des statistiques d'acoustique de la salle [Lebart et al., 2001; Habets, 2005; Habets et al., 2008a, 2009]. D'autres méthodes estiment la covariance de la réverbération tardive $\mathbf{\Sigma}_{s_l}(n,f)$ en modélisant la RIR de la parole locale $\mathbf{a}_s(k,f',f)$ par un filtre long, dont les coefficients sont déterminés à partir des trames précédentes de la parole locale $\mathbf{s}(n,f)$ [Erkelens et Heusdens, 2010; Braun et al., 2016]. Plus récemment, Kodrasi et Bourlard [2018] ont modélisé la covariance de la réverbération tardive $\mathbf{\Sigma}_{s_l}(n,f)$ à l'aide d'un DNN appliqué sur les trames précédentes de la parole locale $\mathbf{s}(n,f)$. Toutefois, la soustraction spectrale dans (2.51) produit une mauvaise estimation de la covariance de la composante précoce $\mathbf{\Sigma}_{s_e}(n,f)$, notamment dans des environnements très réverbérants où le DRR est faible.

Filtrage spatial adaptatif Reprenant la structure du GSC en réduction de bruit (voir la figure 2.7), Habets et Gannot [2007] ont proposé une méthode utilisant deux filtres spatiaux fixes pour estimer les covariances de la composante précoce $\mathbf{\Sigma}_{s_e}(n,f)$ et de la réverbération tardive $\mathbf{\Sigma}_{s_l}(n,f)$. Pour estimer $\mathbf{\Sigma}_{s_e}(n,f)$, le premier filtre se focalise dans la direction du locuteur local pour réduire l'énergie des signaux provenant des autres directions, qui sont censés correspondre aux réflexions de la parole locale $\mathbf{s}(n,f)$. Pour estimer $\mathbf{\Sigma}_{s_l}(n,f)$, une matrice de blocage $\mathbf{B}(f)$ vient au contraire «bloquer» les signaux provenant de la direction du locuteur local. Toutefois, comme cette méthode ne modélise

pas les propriétés spatiales des deux composantes, elle produit de mauvaises estimations des deux covariances.

Pour représenter les propriétés des deux composantes $\mathbf{s}_{\mathrm{e}}(n,f)$ et $\mathbf{s}_{\mathrm{l}}(n,f)$, les méthodes de filtrage spatial adaptatif utilisent le fait que les caractéristiques spatiales de la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$ correspondent à un signal localisé, et que celles de la réverbération tardive $\mathbf{s}_{\mathrm{l}}(n,f)$ à un signal diffus (voir la partie 2.1.1.2). Le but est alors d'estimer les DSPs de la composante précoce $v_{s_{\mathrm{e}}}(n,f)$ et de la réverbération tardive $v_{s_{\mathrm{l}}}(n,f)$ [Braun et al., 2018].

Un premier type de méthodes estime conjointement les DSPs $v_{s_{\mathrm{e}}}(n,f)$ et $v_{s_{\mathrm{l}}}(n,f)$ sans filtre spatial fixe en amont du filtre $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$ en utilisant un critère MMSE [Schwartz et al., 2016b], le critère du MV [Kuklasiński et al., 2014; Schwartz et al., 2016c], ou une décomposition en valeurs propres [Kodrasi et Doclo, 2017]. Un second type de méthode utilise une matrice de blocage $\mathbf{B}(f)$ en amont du filtre $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$, comme la méthode de Habets et Gannot [2007], et estime conjointement les DSPs $v_{s_{\mathrm{e}}}(n,f)$ et $v_{s_{\mathrm{l}}}(n,f)$ avec le critère du MV [Braun et Habets, 2013; Kuklasiński et al., 2016; Schwartz et al., 2015b; Braun et Habets, 2015]. Toutefois, ces deux types de méthodes reposent sur la connaissance de la position du locuteur local, qui peut être mal estimée dans des environnements très réverbérants [Brandstein et Ward, 2001, Chapitre 8].

Apprentissage profond Plus récemment, Mack et al. [2018] ont estimé directement le filtre $\mathbf{W}_{\mathbf{s}_{\mathrm{e}}}(n,f)$ en monocanal, en utilisant le critère MIA (voir la partie 2.2.1.5). Cette méthode obtient de meilleures performances que des méthodes de modélisation temporelle. D'autres méthodes estiment directement la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$ en modélisant son spectre en amplitude $|\mathbf{s}_{\mathrm{e}}(n,f)|$ et en multipliant ce spectre en amplitude par la phase du mélange $\theta_s(n,f)$ [Ernst et al., 2018; Wu et al., 2017]. Ces méthodes ont montré de meilleurs résultats que des méthodes de filtrage spatial adaptatif.

Les méthodes de suppression de réverbération sont robustes aux fluctuations de la RIR de la parole locale $\mathbf{a}_s(k,f',f)$ [Braun et al., 2018]. Toutefois, comme ces méthodes supposent que la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$ et la réverbération tardive $\mathbf{s}_{\mathrm{l}}(n,f)$ sont indépendantes, la modélisation du processus de propagation de la parole locale $\mathbf{s}(n,f)$ est mauvaise. Ces méthodes ne peuvent restituer correctement la phase de la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$. Par conséquent, elles introduisent des artefacts dans la parole cible estimée, qui peuvent être importants dans des environnements très réverbérants où le DRR est faible. De plus, ces méthodes reposent sur l'approximation de la bande étroite, qui ne modélise pas correctement le processus de réverbération.

2.2.3.2. Filtrage inverse

Le problème de la déréverbération par filtrage inverse est illustré sur la figure 2.12. Ces méthodes ne font pas l'hypothèse d'indépendance entre la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$ et la réverbération tardive $\mathbf{s}_{\mathrm{l}}(n,f)$. Elles consistent à extraire la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$ en soustrayant une estimation de la réverbération tardive $\widehat{\mathbf{s}}_{\mathrm{l}}(n,f)$ au signal de

mélange $\mathbf{s}(n,f)$ [Naylor et Gaubitch, 2010, Chapitre 9] :

$$\mathbf{r}^{\text{rev}}(n,f) = \mathbf{s}(n,f) - \widehat{\mathbf{s}}_{\text{l}}(n,f), \tag{2.53}$$

où $\widehat{\mathbf{s}}_{\text{l}}(n,f)$ est obtenu en appliquant un filtre linéaire long, modélisant la corrélation temporelle entre les trames successives de la parole locale $\mathbf{s}(n,f)$ introduite par la réverbération, sur le mélange $\mathbf{s}(n-\Delta,f)$. On désigne cette opération par le terme de *déréverbération linéaire*. Un délai Δ est introduit pour éviter la dégradation de la composante précoce $\mathbf{s}_{\text{e}}(n,f)$:

$$\widehat{\mathbf{s}}_{\text{l}}(n,f) = \sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\mathbf{s}(n-l,f), \tag{2.54}$$

où $\mathbf{G}(l,f) \in \mathbb{C}^{M\times M}$ est la matrice carrée de dimension M correspondant au l-ième délai du filtre long $\mathcal{G}(f) = [\mathbf{G}(\Delta,f)\dots\mathbf{G}(\Delta+L-1,f)] \in \mathbb{C}^{M\times ML}$, et L est la longueur du filtre. Pour pouvoir modéliser parfaitement la réverbération tardive $\mathbf{s}_{\text{l}}(n,f)$, la longueur L doit être, idéalement, aussi grande que la longueur de la RIR de la parole locale $\mathbf{a}_s(k,f',f)$. Ce filtre introduit peu d'artefacts dans la composante précoce $\mathbf{s}_{\text{e}}(n,f)$ car son application sur les trames successives du mélange $\mathbf{s}(n-\Delta,f)$ permet de modéliser à la fois le spectre en amplitude et la phase de la réverbération tardive $\mathbf{s}_{\text{l}}(n,f)$. Il convient de noter que l'estimation de la réverbération tardive $\widehat{\mathbf{s}}_{\text{l}}(n,f)$ par le filtre $\mathcal{G}(f)$ peut aussi être utilisée pour la suppression de réverbération [Kinoshita et al., 2009; Braun et Habets, 2016].

En pratique, le signal après déréverbération linéaire $\mathbf{r}^{\text{rev}}(n,f)$ n'est pas égal à la composante précoce $\mathbf{s}_{\text{e}}(n,f)$ pour deux raisons : l'erreur d'estimation du filtre $\mathcal{G}(f)$, notamment à cause des variations de la RIR de la parole locale $\mathbf{a}_s(k,f',f)$, et la longueur du filtre $\mathcal{G}(f)$, qui est plus petite que celle de la RIR de la parole locale $\mathbf{a}_s(k,f',f)$. Par conséquent, une réverbération résiduelle $\mathbf{s}_{\text{r}}(n,f)$ subsiste et s'exprime de la manière suivante [Furuya et Kataoka, 2007] :

$$\mathbf{r}^{\text{rev}}(n,f) - \mathbf{s}_{\text{e}}(n,f) = \underbrace{\mathbf{s}_{\text{l}}(n,f) - \widehat{\mathbf{s}}_{\text{l}}(n,f)}_{=\mathbf{s}_{\text{r}}(n,f)}. \tag{2.55}$$

La composante précoce $\mathbf{s}_{\text{e}}(n,f)$ est extraite en appliquant un post-filtre court $\mathbf{W}_{s_{\text{e}}}(n,f)$ sur le signal $\mathbf{r}^{\text{rev}}(n,f)$, comme pour la suppression d'écho résiduel (voir la figure 2.12) :

$$\widehat{\mathbf{s}}_{\text{e}}(n,f) = \mathbf{W}_{s_{\text{e}}}(n,f)\mathbf{r}^{\text{rev}}(n,f). \tag{2.56}$$

Nous présentons tout d'abord les méthodes de déréverbération linéaire, qui estiment le filtre $\mathcal{G}(f)$ à l'aide de la méthode *weighted prediction error* (WPE). Nous détaillons ensuite les méthodes de suppression de réverbération résiduelle, qui estiment le post-filtre $\mathbf{W}_{s_{\text{e}}}(n,f)$.

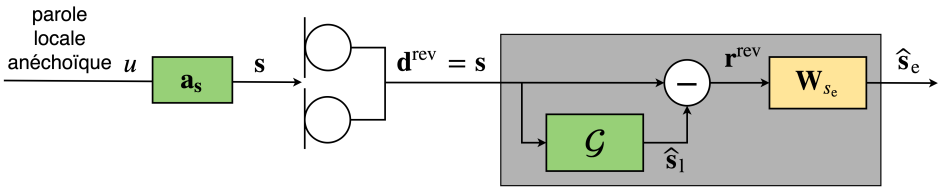


FIGURE 2.12. – Déréverbération par filtrage inverse.

Déréverbération linéaire par WPE La méthode WPE consiste à estimer le filtre $\mathcal{G}(f)$ en modélisant les signaux présents dans le signal $\mathbf{r}^{\text{rev}}(n,f)$. Comme la composante précoce $\mathbf{s}_{\text{e}}(n,f)$ n'est pas observée, elle est modélisée par une variable gaussienne de moyenne nulle, comme dans (2.27) en séparation de sources [Nakatani et al., 2008, 2010; Yoshioka et Nakatani, 2012]. En ce qui concerne la réverbération résiduelle $\mathbf{s}_{\text{r}}(n,f)$, elle est considérée comme étant égale à zéro :

$$\mathbf{s}_{\text{r}}(n,f) = \mathbf{0}. \tag{2.57}$$

Le filtre $\mathcal{G}(f)$ et la covariance de la composante précoce $\mathbf{\Sigma}_{s_{\text{e}}}(n,f)$ sont estimés en optimisant le critère du MV. Comme il n'y a pas de solution analytique, une procédure itérative alternant entre le calcul du filtre $\mathcal{G}(f)$ et le calcul de la covariance $\mathbf{\Sigma}_{s_{\text{e}}}(n,f)$ est appliquée. Toutefois, aucune contrainte n'est appliquée sur la DSP $v_{s_{\text{e}}}(n,f)$ ou la MCS $\mathbf{R}_{s_{\text{e}}}(f)$. Par conséquent, la déréverbération de la parole locale $\mathbf{s}(n,f)$ est limitée.

D'autres méthodes améliorent la déréverbération linéaire en modélisant la DSP $v_{s_{\text{e}}}(n,f)$ avec du codage prédictif linéaire [Yoshioka et al., 2011], des distributions *a priori* parcimonieuses [Jukić et al., 2015], une NMF [Kagami et al., 2018], ou un DNN [Kinoshita et al., 2017].

Suppression de réverbération résiduelle Comme en suppression de réverbération, il est possible d'utiliser le filtre de Wiener, ou ses variantes, pour déterminer le post-filtre $\mathbf{W}_{s_{\text{e}}}(n,f)$:

$$\mathbf{W}_{s_{\text{e}}}(n,f) = \mathbf{\Sigma}_{s_{\text{e}}}(n,f)\Big(\mathbf{\Sigma}_{s_{\text{e}}}(n,f) + \mathbf{\Sigma}_{s_{\text{r}}}(n,f)\Big)^{-1}. \tag{2.58}$$

De même qu'en suppression de réverbération, le post-filtre $\mathbf{W}_{s_{\text{e}}}(n,f)$ introduit des artefacts dans la composante précoce estimée $\widehat{\mathbf{s}}_{\text{e}}(n,f)$, qui sont toutefois moins importants puisque le DRR du signal $\mathbf{r}^{\text{rev}}(n,f)$ est plus élevé que dans le mélange $\mathbf{s}(n,f)$.

Les covariances de la composante précoce $\mathbf{\Sigma}_{s_{\text{e}}}(n,f)$ et de la réverbération résiduelle $\mathbf{\Sigma}_{s_{\text{r}}}(n,f)$ sont estimées à l'aide de méthodes de suppression de réverbération. En particulier, Furuya et Kataoka [2007] utilisent une méthode de déréverbération linéaire sur le signal $\mathbf{r}(n,f)$, et Cohen et al. [2017] utilisent une méthode de modélisation temporelle comme dans (2.52). Toutefois, ces méthodes estiment le post-filtre $\mathbf{W}_{s_{\text{e}}}(n,f)$ séparément du filtre $\mathcal{G}(f)$. En particulier, la réverbération résiduelle $\mathbf{s}_{\text{r}}(n,f)$ n'est pas modélisée dans l'estimation du filtre $\mathcal{G}(f)$. Cette dernière peut être importante lorsque le filtre $\mathcal{G}(f)$ n'a pas convergé, comme par exemple lorsque la RIR de la parole locale $\mathbf{a}_s(k,f'f)$ varie au cours du temps.

Togami et al. [2013] ont proposé une méthode de séparation de sources pour optimiser conjointement le filtre $\mathcal{G}(f)$ et le post-filtre $\mathbf{W}_{s_\mathrm{e}}(n,f)$. Ils modélisent la composante précoce $\mathbf{s}_\mathrm{e}(n,f)$ par une variable gaussienne de moyenne nulle comme dans (2.27), la réverbération résiduelle $\mathbf{s}_\mathrm{r}(n,f)$ par une variable gaussienne de moyenne non nulle, et utilisent un algorithme EM pour optimiser les paramètres de ce modèle selon le critère du MV. Toutefois, cette méthode n'impose pas de contrainte sur la covariance de la composante précoce $\mathbf{\Sigma}_{s_\mathrm{e}}(n,f)$, et ne modélise pas la composante de la réverbération résiduelle $\mathbf{s}_\mathrm{r}(n,f)$ due à la longueur du filtre $\mathcal{G}(f)$ plus petite que celle de la RIR de la parole locale $\mathbf{a}_s(k,f',f)$.

2.2.3.3. Méthodes alternatives

Il convient d'ajouter qu'il existe deux types de méthodes alternatives pour la déréverbération. Le premier type de méthodes alternatives extraient la parole anéchoïque $u(n,f)$ [Schmid et al., 2012; Jukić et al., 2014; Schwartz et al., 2015a; Mohammadiha et al., 2015]. Toutefois, ce signal a une moins bonne qualité et intelligibilité que la composante précoce $\mathbf{s}_\mathrm{e}(n,f)$ (voir la partie 2.1.1.2). Le deuxième type de méthodes alternatives les méthodes d'égalisation partielle de canaux acoustiques, qui nécessitent au préalable l'estimation de la RIR de la parole locale $\mathbf{a}_s(k,f',f)$ [Miyoshi et Kaneda, 1988; Kallinger et Mertins, 2006; Zhang et al., 2010; Lim et al., 2014; Kodrasi et al., 2013, 2014]. Cependant, ces méthodes sont sensibles à l'erreur d'estimation et aux fluctuations de la RIR de la parole locale $\mathbf{a}_s(k,f',f)$.

2.2.4. Choix des méthodes

Dans les parties 2.2.1, 2.2.2 et 2.2.3, nous avons présenté les méthodes de réduction individuelle du bruit, de l'écho et de la réverbération, dans le but de sélectionner celles sur lesquelles nous nous baserons pour concevoir notre propre méthode dans cette thèse. En réduction de bruit, nous retenons donc, pour le cas monocanal, les méthodes d'estimation directe du filtre $\mathbf{W}_s(n,f)$ par apprentissage profond, et pour le cas multicanal, les méthodes de séparation de sources qui estiment les DSPs $v_c(n,f)$ par apprentissage profond (voir la partie 2.2.1.5). En réduction d'écho, nous retenons la combinaison de l'annulation d'écho (voir la partie 2.2.2.1) avec la suppression d'écho résiduel (voir la partie 2.2.2.2), car elle permet un compromis entre l'introduction d'artefacts dans la parole locale estimée $\widehat{\mathbf{s}}(n,f)$ et la complexité algorithmique. En déréverbération, nous retenons les méthodes de filtrage inverse (voir la partie 2.2.3.2), car elles introduisent moins d'artefacts que les méthodes de suppression de réverbération, et n'ont pas les désavantages des méthodes alternatives de déréverbération (voir la partie 2.2.3.3).

Dans la partie suivante, nous présentons les problèmes et les méthodes de réduction conjointe de deux et trois types de distorsion.

2.3. Réduction conjointe de deux et trois types de distorsion

Dans cette partie, nous considérons un scénario où deux types de distorsion sont présents. Nous considérons ensuite le scénario où la réduction de bruit, d'écho et de réverbération est traitée conjointement. Nous présentons les différentes méthodes de réduction conjointe, qui impliquent de modéliser l'interaction entre les types de distorsion.

2.3.1. Réduction conjointe de bruit et d'écho

On considère pour commencer que le mélange $\mathbf{d}(n,f)$ est la somme de la parole locale $\mathbf{s}(n,f)$, du signal de bruit $\mathbf{b}(n,f)$ et de l'écho acoustique $\mathbf{y}(n,f)$:

$$\mathbf{d}(n,f) = \mathbf{s}(n,f) + \mathbf{b}(n,f) + \mathbf{y}(n,f). \tag{2.59}$$

La réduction conjointe de bruit et d'écho consiste à extraire la parole locale $\mathbf{s}(n,f)$ du mélange $\mathbf{d}(n,f)$ en combinant les méthodes individuelles de réduction de ces deux types de distorsion. Nous détaillons un premier type de méthodes combinant un filtre d'annulation d'écho avec un post-filtre de suppression conjointe de bruit et d'écho résiduel. Nous présentons ensuite les méthodes de suppression conjointe d'écho et de bruit.

2.3.1.1. Annulation d'écho et suppression d'écho résiduel et de bruit

La figure 2.13 illustre le problème de réduction d'écho par annulation d'écho suivi de suppression conjointe d'écho résiduel et de bruit. Pour extraire la parole locale $\mathbf{s}(n,f)$, ce type de méthode applique tout d'abord une annulation d'écho [Hänsler et Schmidt, 2004, Chapitre 9] :

$$\mathbf{e}^{\mathrm{b+e}}(n,f) = \mathbf{d}(n,f) - \widehat{\mathbf{y}}(n,f), \tag{2.60}$$

où $\widehat{\mathbf{y}}(n,f)$ est obtenu à l'aide du filtre $\mathcal{H}(f)$ comme dans (2.35). Le filtre $\mathcal{H}(n,f)$ est généralement estimé de manière adaptative comme dans le cas de la réduction individuelle d'écho (voir la partie 2.2.2.1), où la parole locale $\mathbf{s}(n,f)$ et le signal de bruit $\mathbf{b}(n,f)$ sont considérés comme une seule source $\mathbf{c}_{s+b}(n,f) = \mathbf{s}(n,f) + \mathbf{b}(n,f)$ perturbant l'adaptation [Hänsler et Schmidt, 2004, Chapitre 9]. Le signal après annulation d'écho $\mathbf{e}^{\mathrm{b+é}}(n,f)$ contient la parole locale $\mathbf{s}(n,f)$, le signal de bruit $\mathbf{b}(n,f)$ et l'écho résiduel $\mathbf{z}(n,f)$:

$$\mathbf{e}^{\mathrm{b+e}}(n,f) - \mathbf{s}(n,f) = \mathbf{b}(n,f) + \underbrace{\mathbf{y}(n,f) - \widehat{\mathbf{y}}(n,f)}_{=\mathbf{z}(n,f)}. \tag{2.61}$$

La parole locale $\mathbf{s}(n,f)$ est ensuite extraite en appliquant un post-filtre court $\mathbf{W}_s(n,f)$ sur le signal $\mathbf{e}^{\mathrm{b+é}}(n,f)$ [Hänsler et Schmidt, 2004, Chapitre 10] :

$$\widehat{\mathbf{s}}(n,f) = \mathbf{W}_s(n,f)\mathbf{e}^{\mathrm{b+e}}(n,f). \tag{2.62}$$

Nous présentons les méthodes d'estimation du post-filtre de suppression conjointe de bruit et d'écho résiduel $\mathbf{W}_s(n,f)$ dans le paragraphe suivant.

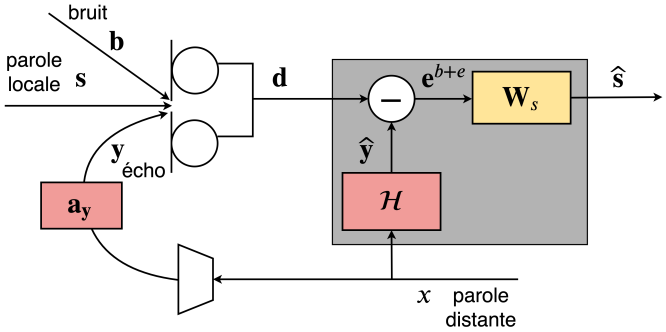


FIGURE 2.13. – Annulation d'écho et suppression d'écho résiduel et de bruit.

Suppression conjointe d'écho résiduel et de bruit Le filtre de Wiener est généralement utilisé comme critère pour estimer le post-filtre de suppression conjointe de bruit et d'écho résidul $\mathbf{W}_s(n,f)$ [Hänsler et Schmidt, 2004, Chapitre 10] :

$$\mathbf{W}_s(n,f) = \mathbf{\Sigma}_s(n,f)\Big(\mathbf{\Sigma}_s(n,f) + \mathbf{\Sigma}_b(n,f) + \mathbf{\Sigma}_z(n,f)\Big)^{-1} \tag{2.63}$$

Généralement, la covariance du bruit $\mathbf{\Sigma}_b(n,f)$ est estimée à l'aide d'une méthode de soustraction spectrale basée sur la SPP (voir la partie 2.2.1.3), celle de l'écho résiduel $\mathbf{\Sigma}_z(n,f)$ est estimée à partir d'une transformation réalisée sur un ou plusieurs signaux comme en suppression d'écho résiduel (voir la partie 2.2.2.2), et celle de la parole locale $\mathbf{\Sigma}_s(n,f)$ est déduite par soustraction spectrale [Le Bouquin-Jeannès et al., 2001; Gustafsson et al., 2002; Park et al., 2002; Luis Valero et Habets, 2019] :

$$\mathbf{\Sigma}_s(n,f) = \widehat{\mathbb{E}}\left[\mathbf{e}^{\mathrm{b+e}}(n,f)\mathbf{e}^{\mathrm{b+e}}(n,f)^H\right] - \mathbf{\Sigma}_b(n,f) - \mathbf{\Sigma}_z(n,f). \tag{2.64}$$

Doclo et al. [2000] ont utilisé une méthode de décomposition en valeurs singulières du signal $\mathbf{e}(n,f)$ pour estimer $\mathbf{\Sigma}_b(n,f)$, $\mathbf{\Sigma}_z(n,f)$ et $\mathbf{\Sigma}_s(n,f)$. Toutefois, toutes ces approches négligent la composante de l'écho résiduel $\mathbf{z}(n,f)$ due à l'erreur d'estimation du filtre $\mathcal{H}(f)$. En effet, elles supposent que le filtre d'annulation d'écho $\mathcal{H}(n,f)$ a déjà convergé, ce qui n'est pas le cas dans des scénarios réels. Ceci revient alors à estimer le post-filtre $\mathbf{W}_s(n,f)$ séparément du filtre $\mathcal{H}(f)$.

Togami et al. [2014] ont proposé une méthode pour estimer conjointement le filtre $\mathcal{H}(n,f)$ et le post-filtre $\mathbf{W}_s(n,f)$. Ils modélisent le filtre $\mathcal{H}(f)$ par filtrage de Kalman (voir la partie 2.2.2.2), et la parole locale $\mathbf{s}(n,f)$ et le signal de bruit $\mathbf{b}(n,f)$ par des variables gaussiennes comme en séparation de sources (voir la partie 2.2.1.4). Ils optimisent conjointement les paramètres de ce modèle selon le critère du MV. Toutefois, cette méthode n'impose pas de contrainte sur les covariances $\mathbf{\Sigma}_c(n,f)$, et ne modélise pas toutes les composantes de l'écho résiduel $\mathbf{z}(n,f)$.

Structure alternative Un autre combinaison possible des filtres consiste à placer un filtre court $\mathbf{W}_s(n,f)$ pour réduire le bruit, avant le filtre $\mathcal{H}(f)$ (voir la figure 2.14)

[Martin et Vary, 1996; Doclo et al., 2000; Kellermann, 2001]. Le filtre $\mathbf{W}_s(n,f)$ réduit alors partiellement l'écho $\mathbf{y}(n,f)$, et un écho résiduel $\mathbf{z}(n,f)$ subsiste, qui est fortement non-linéaire en raison des variations rapides de ce filtre. La relation entre l'écho résiduel $\mathbf{z}(n,f)$ et la parole locale $x(n,f)$ ne peut alors pas être modélisée par un filtre long $\mathcal{H}(f)$ qui varie lentement au cours du temps. Pour résoudre ce problème, des méthodes de soustraction spectrale basées sur un filtrage spatial adaptatif ont été proposées pour estimer conjointement le filtre $\mathbf{W}_s(n,f)$ et un filtre long $\mathcal{H}(n,f)$ qui varie rapidement au cours du temps [Rombouts et Moonen, 2005; Herbordtt et al., 2005; Reuven et al., 2007]. Toutefois, ces méthodes reposent sur la connaissance de la position du locuteur local, qui est difficile à estimer en présence de bruit et d'écho.

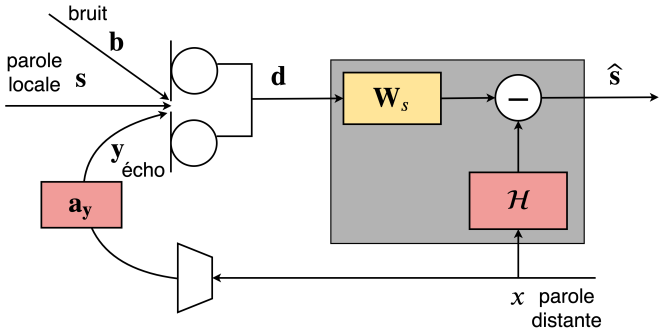


FIGURE 2.14. – Suppression d'écho et de bruit avant annulation d'écho.

2.3.1.2. Suppression conjointe d'écho et de bruit

Le problème de la suppression conjointe de bruit et de réverbération est illustré sur la figure 2.15. Pour extraire la parole locale $\mathbf{s}(n,f)$, ce type de méthode applique un filtre court $\mathbf{W}_s(n,f)$ sur le mélange $\mathbf{d}(n,f)$ pour extraire la parole locale $\widehat{\mathbf{s}}(n,f)$:

$$\widehat{\mathbf{s}}(n,f) = \mathbf{W}_s(n,f)\mathbf{d}(n,f). \tag{2.65}$$

Le filtre de Wiener s'exprime de la manière suivante :

$$\mathbf{W}_s(n,f) = \mathbf{\Sigma}_s(n,f)\Big(\mathbf{\Sigma}_s(n,f) + \mathbf{\Sigma}_b(n,f) + \mathbf{\Sigma}_y(n,f)\Big)^{-1}. \tag{2.66}$$

Comme pour la suppression d'écho, Park et Chang [2012] estiment les trois covariances $\mathbf{\Sigma}_s(n,f)$, $\mathbf{\Sigma}_b(n,f)$ et $\mathbf{\Sigma}_y(n,f)$ à l'aide d'une méthode de soustraction spectrale basée sur la SPP. Seo et al. [2018] utilisent un DNN pour améliorer l'estimation des paramètres de cette méthode. Des méthodes d'estimation directe du filtre $\mathbf{W}_s(n,f)$ ont aussi été proposées [Zhang et Wang, 2018; Seo et al., 2018].

2.3.2. Réduction conjointe de bruit et de réverbération

On considère maintenant le mélange $\mathbf{d}^{\mathrm{b+r}}(n,f)$ qui est la somme de la parole locale $\mathbf{s}(n,f)$ et du signal de bruit $\mathbf{b}(n,f)$. D'après (2.5), le mélange s'exprime de la manière

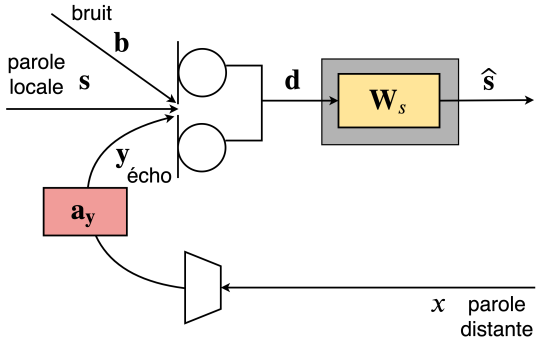


FIGURE 2.15. – Suppression conjointe d'écho et de bruit.

suivante :

$$\mathbf{d}^{\mathrm{b+r}}(n,f) = \mathbf{s}_{\mathrm{e}}(n,f) + \mathbf{s}_{\mathrm{l}}(n,f) + \mathbf{b}(n,f). \tag{2.67}$$

Nous détaillons tout d'abord les méthodes de suppression conjointe de bruit et de réverbération, puis les méthodes basées sur le filtrage inverse.

2.3.2.1. Suppression conjointe de bruit et réverbération

Le problème de la suppression conjointe de bruit et de réverbération est illustré sur la figure 2.16. Pour extraire la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$, ce type de méthodes applique un filtre court $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$ sur le mélange $\mathbf{d}^{\mathrm{b+r}}(n,f)$:

$$\widehat{\mathbf{s}}_{\mathrm{e}}(n,f) = \mathbf{W}_{s_{\mathrm{e}}}(n,f)\mathbf{d}^{\mathrm{b+r}}(n,f). \tag{2.68}$$

Le filtre de Wiener s'exprime de la manière suivante :

$$\mathbf{W}^{\mathrm{WF}}_{s_{\mathrm{e}}}(n,f) = \mathbf{\Sigma}_{s_{\mathrm{e}}}(n,f)\Big(\mathbf{\Sigma}_{s_{\mathrm{e}}}(n,f) + \mathbf{\Sigma}_{s_{\mathrm{l}}}(n,f) + \mathbf{\Sigma}_{b}(n,f)\Big)^{-1}. \tag{2.69}$$

Comme l'approximation en bande étroite n'est pas valable pour la réverbération tardive $\mathbf{s}_{\mathrm{l}}(n,f)$, la SCM associée à la réverbération tardive $\mathbf{s}_{\mathrm{l}}(n,f)$ est modélisée par une matrice diffuse. Les paramètres de la réverbération tardive $\mathbf{s}_{\mathrm{l}}(n,f)$ et du bruit $\mathbf{b}(n,f)$ nécessitent alors d'être estimés.

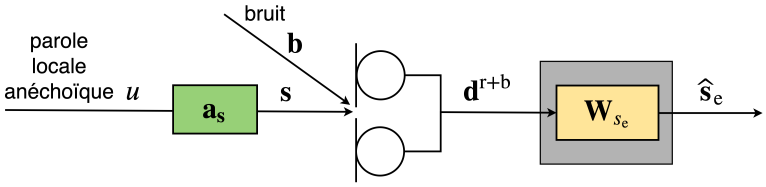


FIGURE 2.16. – Suppression conjointe de bruit et de révebération.

Modélisation temporelle de $\mathbf{s_l}(n,f)$ Löllmann et Vary [2009] estiment les covariances du bruit $\mathbf{\Sigma}_b(n,f)$ et de la parole locale $\mathbf{\Sigma}_s(n,f)$ par une méthode de soustraction spectrale basée sur la SPP (voir la partie 2.2.1.3). Ils utilisent ensuite une méthode de modélisation temporelle pour estimer la covariance de la réverbération tardive $\mathbf{\Sigma}_{s_\mathrm{l}}(n,f)$ à partir de celle de la parole locale $\mathbf{\Sigma}_s(n,f)$ (voir la partie 2.2.3.1). Ils déduisent enfin celle de la composante précoce $\mathbf{\Sigma}_{s_\mathrm{e}}(n,f)$ par soustraction spectrale :

$$\mathbf{\Sigma}_{s_\mathrm{e}}(n,f) = \widehat{\mathbb{E}}\left[\mathbf{d}^{\mathrm{b+r}}(n,f)\mathbf{d}^{\mathrm{b+r}}(n,f)^H\right] - \mathbf{\Sigma}_b(n,f) - \mathbf{\Sigma}_{s_\mathrm{l}}(n,f). \tag{2.70}$$

D'autres méthodes estiment $\mathbf{\Sigma}_{s_\mathrm{l}}(n,f)$ en modélisant la RIR de la parole locale $\mathbf{a}_s(k,f',f)$ par un filtre long [Erkelens et Heusdens, 2010; Braun et al., 2016], comme en déréverbération linéaire (voir la partie 2.2.3.2). Toutefois, l'utilisation de deux opérations de soustraction spectrale conduit à une mauvaise estimation des covariances de la composante précoce $\mathbf{\Sigma}_{s_\mathrm{e}}(n,f)$ et de la réverbération tardive $\mathbf{\Sigma}_{s_\mathrm{l}}(n,f)$, notamment dans les situations où les SNR et DRR sont faibles.

Filtrage spatial adaptatif Ces méthodes estiment tout d'abord les covariances du bruit $\mathbf{\Sigma}_b(n,f)$ et de la parole locale $\mathbf{\Sigma}_s(n,f)$ par une méthode de soustraction spectrale basée sur la SPP (voir la partie 2.2.1.3). De même que les méthodes de suppression de déréverbération par filtrage adaptatif, ces méthodes estiment ensuite les DSPs de la composante précoce $v_{s_\mathrm{e}}(n,f)$ et de la réverbération tardive $v_{s_\mathrm{l}}(n,f)$ par filtrage spatial adaptatif (voir la partie 2.2.3.1). Ces estimations sont réalisées soit sans filtre spatial fixe en amont du filtre $\mathbf{W}_{s_\mathrm{e}}(n,f)$ [Schwartz et al., 2016b,a,c; Kodrasi et Doclo, 2017], soit avec une matrice de blocage $\mathbf{B}(f)$ en amont du filtre $\mathbf{W}_{s_\mathrm{e}}(n,f)$ [Braun et Habets, 2013; Schwartz et al., 2015b; Kuklasiński et al., 2016; Schwartz et al., 2015c; Braun et Habets, 2015; Kodrasi et Doclo, 2018b]. Toutefois, ces méthodes reposent sur la connaissance de la position du locuteur local, qui peut être mal estimée dans des environnements à la fois très réverbérants et en présence de bruit.

Apprentissage profond Récemment, des méthodes d'estimation par apprentissage profond ont été proposées en monocanal [Zhao et al., 2017; Williamson et Wang, 2017]. Contrairement aux deux types de méthodes précédents, ces méthodes prennent en compte la phase $\theta_{s_\mathrm{e}}(n,f)$ pour estimer la composante précoce $\mathbf{s}_\mathrm{e}(n,f)$. Ainsi, Williamson et Wang [2017] estiment directement le filtre $\mathbf{W}_{s_\mathrm{e}}(n,f)$ en utilisant le critère du masque complexe, et Zhao et al. [2017] estiment directement la composante précoce dans le domaine temporel $\theta_{s_\mathrm{e}}(t)$. Toutefois, ces méthodes ne sont pas comparées à d'autres méthodes de réduction conjointe de bruit et de réverbération.

Il convient de noter que les méthodes de suppression conjointe de bruit et de réverbération présentent les mêmes limites que les méthodes de suppression de réverbération (voir la partie 2.2.3.1).

2.3.2.2. Filtrage inverse

Le problème de la réduction conjointe de bruit et de réverbération par filtrage inverse est illusté sur la figure 2.17. Pour extraire la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$, ce type de méthode applique tout d'abord une déréverbération linéaire sur le mélange $\mathbf{d}^{\mathrm{b+r}}(n,f)$ à l'aide du filtre $\mathcal{G}(f)$ comme dans (2.54) :

$$\mathbf{r}^{\mathrm{b+r}}(n,f) = \mathbf{d}^{\mathrm{b+r}}(n,f) - \underbrace{\sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\mathbf{d}^{\mathrm{b+r}}(n-l,f)}_{=\widehat{\mathbf{d}}_{\mathrm{l}}^{\mathrm{b+r}}(n,f)} \tag{2.71}$$

À cause des raisons évoquées dans la partie 2.2.3.2 et de la présence du signal de bruit $\mathbf{b}(n,f)$, des signaux résiduels non désirés subsistent et s'expriment de la manière suivante :

$$\mathbf{r}^{\mathrm{b+r}}(n,f) - \mathbf{s}_{\mathrm{e}}(n,f) = \underbrace{\mathbf{s}_{\mathrm{l}}(n,f) - \widehat{\mathbf{d}}_{\mathrm{l},s}^{\mathrm{b+r}}(n,f)}_{=\mathbf{s}_{\mathrm{r}}(n,f)} + \underbrace{\mathbf{b}(n,f) - \widehat{\mathbf{d}}_{\mathrm{l},b}^{\mathrm{b+r}}(n,f)}_{=\mathbf{b}_{\mathrm{r}}(n,f)}, \tag{2.72}$$

où les signaux $\widehat{\mathbf{d}}_{\mathrm{l},s}^{\mathrm{b+r}}(n,f)$ et $\widehat{\mathbf{d}}_{\mathrm{l},b}^{\mathrm{b+r}}(n,f)$ sont les composantes latentes de la réverbération tardive estimée $\widehat{\mathbf{d}}_{\mathrm{l}}^{\mathrm{b+r}}(n,f)$ qui résultent de (2.71) comme suit

$$\widehat{\mathbf{d}}_{\mathrm{l}}^{\mathrm{b+r}}(n,f) = \sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\Big(\mathbf{s}(n-l,f) + \mathbf{b}(n-l,f)\Big) \tag{2.73}$$

$$= \underbrace{\sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\mathbf{s}(n-l,f)}_{=\widehat{\mathbf{d}}_{\mathrm{l},s}^{\mathrm{b+r}}(n,f)} + \underbrace{\sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\mathbf{b}(n-l,f)}_{=\widehat{\mathbf{d}}_{\mathrm{l},b}^{\mathrm{b+r}}(n,f)}, \tag{2.74}$$

et $\mathbf{b}_{\mathrm{r}}(n,f)$ est le signal de bruit *déréverbéré*. Le terme *déréverbéré* signifie « après application du filtre $\mathcal{G}(f)$ ». La composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$ est ensuite extraite en appliquant un post-filtre court $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$ sur le signal $\mathbf{r}^{\mathrm{b+r}}(n,f)$:

$$\widehat{\mathbf{s}}_{\mathrm{e}}(n,f) = \mathbf{W}_{s_{\mathrm{e}}}(n,f)\mathbf{r}^{\mathrm{b+r}}(n,f). \tag{2.75}$$

Nous présentons les méthodes de déréverbération linéaire, qui estiment le filtre $\mathcal{G}(f)$ à l'aide d'une extension de la méthode WPE. Nous détaillons ensuite les méthodes de suppression conjointe de bruit et de réverbération résiduelle, qui estiment le post-filtre $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$.

Estimation des filtres Yoshioka et al. [2009a] ont proposé une extension de la méthode WPE pour estimer le filtre $\mathcal{G}(f)$. Ils modélisent la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$ et le signal de bruit déréverbéré $\mathbf{b}_{\mathrm{r}}(n,f)$ par des variables gaussiennes de moyenne nulle comme dans (2.27) en séparation de sources (voir la partie 2.2.1.4). Ils considèrent la réverbération résiduelle $\mathbf{s}_{\mathrm{r}}(n,f)$ égale à zéro comme dans (2.57). Le filtre $\mathcal{G}(f)$ et les covariances

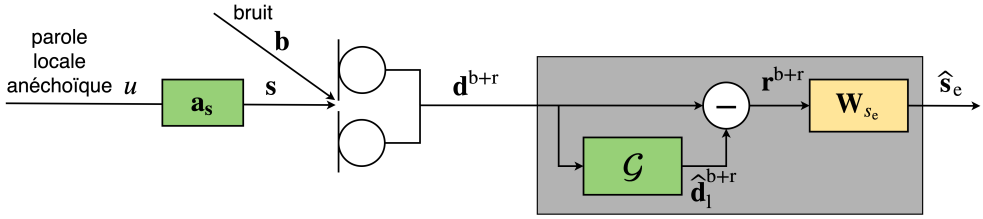


FIGURE 2.17. – Réduction conjointe de bruit et de réverbération par filtrage inverse.

de la composante précoce $\mathbf{\Sigma}_{s_\mathrm{e}}(n,f)$ et du bruit déréverbéré $\mathbf{\Sigma}_{b_\mathrm{r}}(n,f)$ sont estimés en optimisant le critère du MV. Comme il n'y a pas de solution analytique, une procédure itérative similaire à la méthode WPE en déréverbération linéaire est appliquée (voir la partie 2.2.3.2). La DSP $v_{s_\mathrm{e}}(n,f)$ est modélisée par codage prédictif linéaire. Toutefois, aucune contrainte n'est appliquée sur la covariance du bruit déréverbéré $\mathbf{\Sigma}_{b_\mathrm{r}}(n,f)$. Par conséquent, la déréverbération de la parole locale $\mathbf{s}(n,f)$ est limitée. De plus, cette méthode n'estime pas le post-filtre $\mathbf{W}_{s_\mathrm{e}}(n,f)$, ce qui empêche notamment la réduction du bruit déréverbéré $\mathbf{b}_\mathrm{r}(n,f)$.

Il est possible d'estimer le post-filtre $\mathbf{W}_{s_\mathrm{e}}(n,f)$ à partir des covariances $\mathbf{\Sigma}_{s_\mathrm{e}}(n,f)$ et $\mathbf{\Sigma}_{b_\mathrm{r}}(n,f)$ en utilisant le filtre de Wiener (ou ses variantes) :

$$\mathbf{W}_{s_\mathrm{e}}^{\mathrm{WF}}(n,f) = \mathbf{\Sigma}_{s_\mathrm{e}}(n,f)\Big(\mathbf{\Sigma}_{s_\mathrm{e}}(n,f) + \mathbf{\Sigma}_{b_\mathrm{r}}(n,f)\Big)^{-1}. \tag{2.76}$$

En optimisant le même critère du MV que Yoshioka et al. [2009a], certaines méthodes estiment conjointement le filtre $\mathcal{G}(f)$ et le post-filtre $\mathbf{W}_{s_\mathrm{e}}(n,f)$, tout en imposant des contraintes sur les covariances [Yoshioka et al., 2011; Ito et al., 2014; Kagami et al., 2018; Dietzen et al., 2018; Nakatani et Kinoshita, 2019]. Ainsi, ces méthodes estiment les DSPs $v_{s_\mathrm{e}}(n,f)$ et $v_{b_\mathrm{r}}(n,f)$ par codage prédictif linéaire [Yoshioka et al., 2011], une méthode d'analyse en composantes indépendantes [Ito et al., 2014], et une méthode de NMF [Kagami et al., 2018]. Récemment, d'autres méthodes imposent plutôt des contraintes sur la MCS de la composante précoce $\mathbf{R}_{s_\mathrm{e}}(f)$, ce qui nécessite de connaître par avance la position du locuteur local [Dietzen et al., 2018; Nakatani et Kinoshita, 2019].

Toutefois, aucune de ces méthodes ne modélise la réverbération résiduelle $\mathbf{s}_\mathrm{r}(n,f)$. Par conséquent, la déréverbération et la réduction de bruit sont limitées.

2.3.3. Réduction conjointe d'écho, de bruit et de réverbération

En pratique, seuls les travaux de Togami [2015], Togami et Kawaguchi [2012], et Takeda et al. [2009] se sont consacrés à la réduction conjointe d'écho et de réverbération par rapport à la réduction conjointe d'écho et de bruit, et la réduction conjointe de réverbération et de bruit. Nous considérons donc directement le scénario où l'écho, le bruit et la réverbération sont présents simultanément (voir la figure 2.1). Le mélange $\mathbf{d}(n,f)$ s'exprime comme dans (2.13) :

$$\mathbf{d}(n,f) = \mathbf{s}_\mathrm{e}(n,f) + \mathbf{s}_\mathrm{l}(n,f) + \mathbf{y}(n,f) + \mathbf{b}(n,f). \tag{2.77}$$

À notre connaissance, seul deux travaux se sont intéressés à la réduction conjointe d'écho, de bruit et de réverbération [Habets et al., 2008b; Togami et Kawaguchi, 2014].

2.3.3.1. Méthode de Habets et al. [2008b]

La figure 2.18 illustre la méthode de Habets et al. [2008b]. Pour extraire la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$, cette méthode applique tout d'abord une annulation d'écho pour réduire l'écho $\mathbf{y}(n,f)$:

$$\mathbf{e}(n,f) = \mathbf{d}(n,f) - \widehat{\mathbf{y}}(n,f), \tag{2.78}$$

où $\widehat{\mathbf{y}}(n,f)$ est obtenu à l'aide du filtre $\mathcal{H}(f)$ comme dans (2.35). Le signal $\mathbf{e}(n,f)$ contient des signaux résiduels non désirés :

$$\mathbf{e}(n,f) - \mathbf{s}_{\mathrm{e}}(n,f) = \mathbf{s}_{\mathrm{l}}(n,f) + \mathbf{z}(n,f) + \mathbf{b}(n,f). \tag{2.79}$$

Habets et al. [2008b] appliquent ensuite un post-filtre $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$ pour extraire la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$ du signal $\mathbf{e}(n,f)$:

$$\widehat{\mathbf{s}}_{\mathrm{e}}(n,f) = \mathbf{W}_{s_{\mathrm{e}}}(n,f)\mathbf{e}(n,f). \tag{2.80}$$

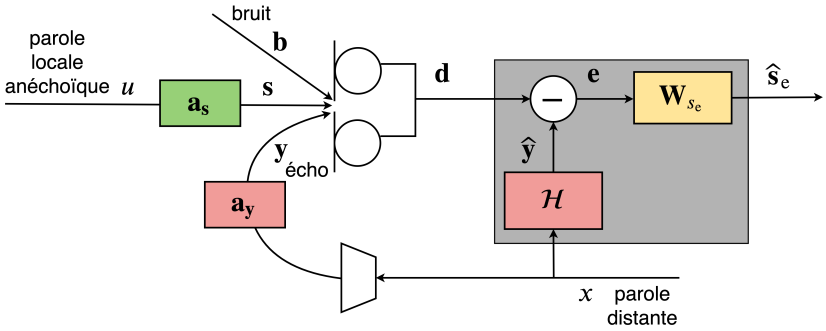


FIGURE 2.18. – Approche de réduction conjointe d'écho, de bruit et de réverbération proposée par Habets et al. [2008b].

Le filtre $\mathcal{H}(f)$ et le post-filtre $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$ sont estimés séparément. Le filtre $\mathcal{H}(f)$ est estimé à partir d'un algorithme de type NLMS (voir la partie 2.2.2.1). Le post-filtre $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$ est estimé en utilisant le filtre de Wiener :

$$\mathbf{W}_{s_{\mathrm{e}}}(n,f) = \mathbf{\Sigma}_{s_{\mathrm{e}}}(n,f)\Big(\mathbf{\Sigma}_{s_{\mathrm{e}}}(n,f) + \mathbf{\Sigma}_{s_{\mathrm{l}}}(n,f) + \mathbf{\Sigma}_{z}(n,f) + \mathbf{\Sigma}_{b}(n,f)\Big)^{-1}. \tag{2.81}$$

La covariance du bruit $\mathbf{\Sigma}_b(n,f)$ est tout d'abord estimée avec une méthode de soustraction spectrale basée sur la SPP (voir la partie 2.2.1.3). La covariance de l'écho résiduel $\mathbf{\Sigma}_z(n,f)$, Habets et al. [2008b] est ensuite estimée en modélisant l'écho résiduel tardif $\mathbf{y}_{\mathrm{l}}(n,f)$ comme dans (2.52) à partir de l'écho estimé $\widehat{\mathbf{y}}(n,f)$. La covariance de la parole locale $\mathbf{\Sigma}_s(n,f)$ est déduite par soustraction spectrale. La covariance de la réverbération

tardive $\mathbf{\Sigma}_{s_\mathrm{l}}(n,f)$ est ensuite estimée comme dans (2.52). La covariance de la composante précoce $\mathbf{\Sigma}_{s_\mathrm{e}}(n,f)$ est enfin déduite par soustraction spectrale.

Toutefois, cette méthode souffre de plusieurs problèmes. Premièrement, elle n'utilise pas de filtre de déréverbération linéaire $\mathcal{G}(f)$ qui permet de réduire la réverbération tardive $\mathbf{s}_\mathrm{l}(n,f)$ sans dégrader la composante précoce $\mathbf{s}_\mathrm{e}(n,f)$. Deuxièmement, elle néglige les composantes de l'écho résiduel $\mathbf{z}(n,f)$ dues aux non-linéarités et à l'estimation du filtre $\mathcal{H}(f)$. Autrement dit, le post-filtre $\mathbf{W}_{s_\mathrm{e}}(n,f)$ est estimé séparément du filtre $\mathcal{H}(f)$. Enfin, l'utilisation de deux opérations de soustraction spectrale produit une mauvaise estimation des covariances de la composante précoce $\mathbf{\Sigma}_{s_\mathrm{e}}(n,f)$ et de la réverbération tardive $\mathbf{\Sigma}_{s_\mathrm{l}}(n,f)$, notamment en cas de faibles SNR et DRR.

2.3.3.2. Méthode de Togami et Kawaguchi [2014]

La figure 2.19 illustre la méthode de Togami et Kawaguchi [2014]. Ils appliquent une annulation d'écho à l'aide du filtre $\mathcal{H}(f)$ comme dans (2.78). En parallèle, ils appliquent une déréverbération linéaire sur le mélange $\mathbf{d}(n,f)$ à l'aide du filtre $\mathcal{G}(f)$ comme dans (2.54). Le signal $\mathbf{r}(n,f)$ qui résulte de ces deux filtrages s'exprime de la manière suivante :

$$\mathbf{r}(n,f) = \mathbf{d}(n,f) - \underbrace{\sum_{k=0}^{K-1} \mathbf{h}(k,f)x(n-k,f)}_{=\widehat{\mathbf{y}}(n,f)} - \underbrace{\sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\mathbf{d}(n-l,f)}_{=\widehat{\mathbf{d}}_\mathrm{l}(n,f)}. \tag{2.82}$$

À cause des raisons évoquées dans les parties précédentes, des signaux résiduels non désirés subsistent dans le signal $\mathbf{r}(n,f)$ et s'expriment de la manière suivante :

$$\mathbf{r}(n,f) - \mathbf{s}_\mathrm{e}(n,f) = \mathbf{z}_\mathrm{e}(n,f) + \widetilde{\mathbf{b}}_\mathrm{r}(n,f) + \mathbf{b}_\mathrm{r}(n,f). \tag{2.83}$$

Les signaux $\mathbf{z}_\mathrm{e}(n,f)$ et $\widetilde{\mathbf{b}}_\mathrm{r}(n,f)$ sont définis de la manière suivante :

$$\mathbf{z}_\mathrm{e}(n,f) = \mathbf{y}_\mathrm{e}(n,f) - \widehat{\mathbf{y}}(n,f), \tag{2.84}$$

$$\widetilde{\mathbf{b}}_\mathrm{r}(n,f) = \mathbf{s}_\mathrm{l}(n,f) - \widehat{\mathbf{d}}_{\mathrm{l},s}(n,f) + \mathbf{y}_\mathrm{l}(n,f) - \widehat{\mathbf{d}}_{\mathrm{l},y}(n,f), \tag{2.85}$$

où le signal $\mathbf{y}_\mathrm{e}(n,f)$ désigne la composante précoce de l'écho $\mathbf{y}(n,f)$, les signaux $\widehat{\mathbf{d}}_{\mathrm{l},s}(n,f)$ et $\widehat{\mathbf{d}}_{\mathrm{l},y}(n,f)$ sont les composantes latentes de $\widehat{\mathbf{d}}_\mathrm{l}(n,f)$ définies de manière similaire à (2.74), et $\mathbf{b}_\mathrm{r}(n,f)$ est le signal de bruit déréverbéré défini comme dans (2.72).

La composante précoce $\mathbf{s}_\mathrm{e}(n,f)$ est extraite en appliquant un post-filtre court $\mathbf{W}_{s_\mathrm{e}}(n,f)$ sur le signal $\mathbf{r}(n,f)$:

$$\widehat{\mathbf{s}}_\mathrm{e}(n,f) = \mathbf{W}_{s_\mathrm{e}}(n,f)\mathbf{r}(n,f). \tag{2.86}$$

Pour estimer les filtres $\mathcal{H}(f)$, $\mathcal{G}(f)$ et $\mathbf{W}_{s_\mathrm{e}}(n,f)$, Togami et Kawaguchi [2014] modélisent les signaux $\mathbf{s}_\mathrm{e}(n,f)$ et $\mathbf{b}_\mathrm{r}(n,f)$ par des variables gaussiennes de moyenne nulle comme dans (2.27), et les signaux $\mathbf{z}_\mathrm{e}(n,f)$ et $\widetilde{\mathbf{b}}_\mathrm{r}(n,f)$ par des variables gaussiennes de moyenne non nulle. Les paramètres de ce modèle sont alors optimisés conjointement selon le critère du MV en utilisant l'algorithme EM. En particulier, Togami et Kawaguchi

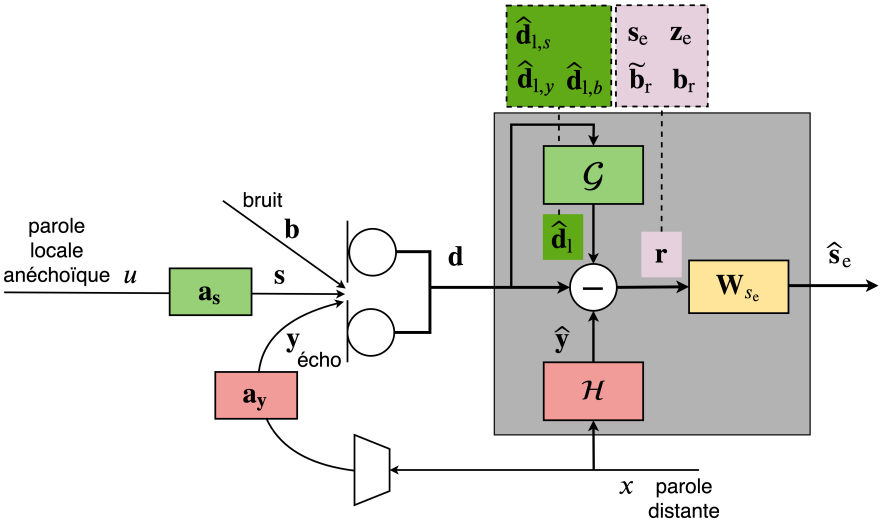


FIGURE 2.19. – Approche proposée par Togami et Kawaguchi [2014]. Les flèches en gras désignent les étapes de filtrage. Les lignes en pointillés désignent les composantes latentes des signaux. Les flèches blanches désignent les mises à jour des filtres.

[2014] déterminent le post-filtre $\mathbf{W}_{s_\mathrm{e}}(n,f)$ à partir du filtre de Wiener :

$$\mathbf{W}_{s_\mathrm{e}}(n,f) = \mathbf{\Sigma}_{s_\mathrm{e}}(n,f)\Big(\mathbf{\Sigma}_{s_\mathrm{e}}(n,f) + \mathbf{\Sigma}_{z_\mathrm{e}}(n,f) + \mathbf{\Sigma}_{\widetilde{b}_\mathrm{r}}(n,f) + \mathbf{\Sigma}_{b_\mathrm{r}}(n,f)\Big)^{-1}. \tag{2.87}$$

Toutefois, cette approche souffre de quatre limites. Premièrement, aucune contrainte n'est imposée sur les DSPs ou MCSs de la composante précoce $\mathbf{s}_\mathrm{e}(n,f)$ et du signal de bruit déréverbéré $\mathbf{b}_\mathrm{r}(n,f)$. Deuxièmement, le signal $\widetilde{\mathbf{b}}_\mathrm{r}(n,f)$ défini dans (4.5) regroupe la composante $\mathbf{s}_\mathrm{l}(n,f) - \widehat{\mathbf{d}}_{\mathrm{l},s}(n,f)$ liée à la parole locale et la composante $\mathbf{y}_\mathrm{l}(n,f) - \widehat{\mathbf{d}}_{\mathrm{l},y}(n,f)$ liée à l'écho. Par conséquent, ces deux composantes vont partager les mêmes paramètres spectraux et spatiaux dans ce modèle, alors qu'en réalité ce n'est pas le cas puisqu'elles correspondent à deux sources différentes. Ces deux limites conduisent à une mauvaise estimation des filtres $\mathcal{H}(f)$, $\mathcal{G}(f)$ et $\mathbf{W}_{s_\mathrm{e}}(n,f)$. Troisièmement, les filtres $\mathcal{H}(f)$ et $\mathcal{G}(f)$ opèrent en parallèle sur le mélange $\mathbf{d}(n,f)$ (voir la figure 2.19). Par conséquent, leurs composantes respectives $\widehat{\mathbf{y}}(n,f)$ et $\widehat{\mathbf{d}}_{\mathrm{l},y}$ sont susceptibles d'interférer entre elles, car elles sont soustraites à des composantes de la même source (l'écho $\mathbf{y}(n,f)$) dans (2.84) et (2.85), respectivement. Enfin, puisque l'écho $\mathbf{y}(n,f)$ est souvent bien plus fort que la parole locale $\mathbf{s}(n,f)$ et le bruit $\mathbf{b}(n,f)$, le filtre $\mathcal{G}(f)$ réduit surtout l'écho résiduel tardif $\mathbf{y}_\mathrm{l}(n,f)$ plutôt que la réverbération tardive de la parole locale $\mathbf{s}_\mathrm{l}(n,f)$.

2.3.4. Choix des méthodes

Dans les parties 2.3.1 et 2.3.2, nous avons présenté les méthodes de réduction conjointe de bruit et d'écho, et de bruit et de réverbératon, respectivement, dans le but d'analyser les stratégies de combinaison de filtres. En réduction conjointe de bruit et d'écho, nous

avons évoqué les méthodes de suppression conjointe de bruit et d'écho (voir la partie 2.3.1.2) car l'estimation du filtre de suppression conjointe de bruit et d'écho est similaire à l'estimation d'un post-filtre de suppression de bruit et d'écho résiduel (voir la partie 2.3.1.1). En réduction conjointe de bruit et de réverbération, nous avons évoqué les méthodes de suppression conjointe de bruit et de revebération (voir la partie 2.3.2.1) car l'estimation du filtre de suppression conjointe de bruit et de réverbération est similaire à l'estimation d'un post-filtre de suppression de bruit et de réverbération (voir la partie 2.3.1.1). Dans la partie 2.3.3, nous avons présenté les méthodes de réduction conjointe de bruit, d'écho et de réverbération, dans le but de sélectionner celle à laquelle nous nous comparons dans cette thèse. Nous retenons la méthode de Togami et Kawaguchi [2014] car c'est la seule qui estime conjointement tous les filtres.

Il convient de remarquer que toutes ces méthodes extraient des caractéristiques de la parole cible et des types de distorsion pour estimer les filtres. En reconnaissance du locuteur, il est possible d'extraire des descripteurs de la parole d'un locuteur cible à l'aide d'une méthode appelée analyse factorielle conjointe (*joint factor analysis*) [Kenny et al., 2007]. Cette méthode suppose qu'il existe une représentation, calculée à partir des *Mel-Frequency Cepstral Coefficients*, de la parole locale $\mathbf{s}(t)$ associé à un locuteur actif, appelée supervecteur. Celui-ci peut s'exprimer comme la somme de composantes, qui proviennent de sous-espaces liés au locuteur et à des paramètres représentant la variabilité d'un enregistrement à l'autre de ce même locuteur, qui sont désignés par le terme *i-vectors* [Dehak et al., 2011]. Puisque cette décomposition est additive dans le domaine des MFCC, elle permet intrinsèquement de séparer les composantes liées au locuteur de celles liées aux déformations convolutives dans le domaine temporel, comme la propagation de la parole du locuteur dans une salle. Toutefois, le signal de bruit $\mathbf{b}(t)$ est une déformation additive dans le domaine temporel, et non une déformation convolutive. Il est possible de prendre en compte dans la représentation en supervecteur un signal de bruit $\mathbf{b}(n, f)$ [Matsui et al., 1996; Wong et Russell, 2001; Ben Kheder et al., 2015], ainsi que la réverbération tardive $\mathbf{s}_\mathrm{l}(n, f)$ [Ben Kheder et al., 2018]. Toutefois, il semble difficile de parvenir à reconstruire la parole cible du locuteur dans le domaine temporel à partir des descripteurs extraits avec l'analyse factorielle conjointe.

2.4. Métriques

En pratique, le rehaussement n'est jamais parfait, et l'estimation de la parole locale $\widehat{\mathbf{s}}(t)$ ou de la composante précoce $\widehat{\mathbf{s}}_\mathrm{e}(t)$ diffèrent du signal cible en raison de :

- la présence d'artefacts, qui correspondent aux déformations (ou dégradations) du signal cible,
- la présence de signaux résiduels provenant du bruit, de l'écho et de la réverbération.

Il est possible d'évaluer la qualité du signal estimé avec des tests d'écoute. Il existe plusieurs méthodologies de test d'écoute. Dans un test MUSHRA [Int. Telecomm. Union (ITU-T) Rec., 2015], les auditeurs doivent donner un score à la qualité perçue de l'estimation. Un test ABX est un test de préférence entre deux signaux [Munson et Gardner, 1950].

Cette qualité est aussi évaluée grâce à des mesures objectives. Parmi celles-ci, nous détaillons un ensemble de métriques qui se basent sur les rapports d'énergie entre les types de distorsion évoqués et la parole cible. Nous présentons ensuite les métriques qui se basent sur un score perceptuel.

2.4.1. Rapports d'énergie

Les métriques basées sur les rapports d'énergie sont mesurées en décibels (dB), qui est une échelle liée à la perception de la puissance d'un son par le système auditif. Ces métriques sont calculées dans le domaine temporel. La qualité de l'estimation est indiquée par des valeurs plus élevées de ces métriques. Nous exprimons ces métriques dans le cas monophonique ($M = 1$) et l'indice m du microphone est omis par souci de clarté. Dans le cas où $M > 1$, nous calculons les métriques séparément sur chaque canal m, puis nous faisons leur moyenne sur les M canaux. La notation c désigne l'ensemble des T échantillons du signal $c(t)$.

Cet ensemble de métriques est basé sur la décomposition d'une parole cible estimée $\widehat{c}_{\text{cible}}$ dans le domaine temporel [Vincent et al., 2006]. Ici, cette parole cible peut être soit la parole locale $c_{\text{cible}} = s$, soit la composante précoce $c_{\text{cible}} = s_{\text{e}}$. Dans le cas de la réduction conjointe de bruit, d'écho et de réverbération, la parole cible estimée $\widehat{c}_{\text{cible}} = \widehat{s}_{\text{e}}$, obtenue avec (2.86), possède cinq composantes :

$$\widehat{c}_{\text{cible}} = c_{\text{cible}}^{\text{post}} + s_{\text{l}}^{\text{post}} + b^{\text{post}} + y^{\text{post}} + c_{\text{cible}}^{\text{art}} \tag{2.88}$$

où $c_{\text{cible}}^{\text{post}}$ est une version de la parole cible $c_{\text{cible}} = s_{\text{e}}$ potentiellement atténuée par le système de rehaussement, $s_{\text{l}}^{\text{post}}$, b^{post} et y^{post} sont les signaux post-résiduels de la réverbération tardive, du bruit et de l'écho, respectivement, et $c_{\text{cible}}^{\text{art}}$ représente les artefacts introduits dans la parole cible $c_{\text{cible}} = s_{\text{e}}$. Nous utilisons l'adjectif *post-résiduel* car le système de rehaussement peut inclure plusieurs filtres de réduction de distorsion. Les composantes $c_{\text{cible}}^{\text{post}}$, $s_{\text{l}}^{\text{post}}$, b^{post} et y^{post} sont calculées de la manière suivante :

$$c^{\text{post}} = \gamma_c \, c, \tag{2.89}$$

où c désigne l'un des quatre signaux c_{cible}, s_{l}, b et y,

$$\gamma_c = \frac{\langle \widehat{c}_{\text{cible}}, c \rangle}{\|c\|^2}, \tag{2.90}$$

$\langle \cdot, \cdot \rangle$ le produit scalaire de deux signaux. Le signal c^{post} est ainsi défini comme la projection de la parole estimée $\widehat{c}_{\text{cible}}$ sur le signal c. La composante liée aux artefacts de la parole cible $c_{\text{cible}}^{\text{art}}$ est alors calculée comme

$$c_{\text{cible}}^{\text{art}} = \widehat{c}_{\text{cible}} - \left(c_{\text{cible}}^{\text{post}} + s_{\text{l}}^{\text{post}} + b^{\text{post}} + y^{\text{post}} \right). \tag{2.91}$$

Il convient de noter que la définition de ces composantes est une extension de la définition de Le Roux et al. [2019] en réduction de bruit à plusieurs types de distorsion. À partir de cette décomposition, nous pouvons définir 6 métriques, qui sont invariantes à l'atténuation.

Réduction de bruit Pour le bruit, on utilise le SNR, qui mesure l'énergie du signal désiré c_{cible} par rapport à celle du signal de bruit b. Le SNR en entrée, c'est-à-dire avant application du système de rehaussement est défini par :

$$\text{SNR}_\text{i} = 10 \log_{10} \frac{\|s\|^2}{\|b\|^2}, \tag{2.92}$$

où $\|c\|^2 = \sum_{t=0}^{T-1} c(t)^2$. Le SNR en sortie, c'est-à-dire après application du système de rehaussement, est défini par :

$$\text{SNR}_\text{o} = 10 \log_{10} \frac{\|c_{\text{cible}}^{\text{post}}\|^2}{\|b^{\text{post}}\|^2}. \tag{2.93}$$

Réduction d'écho Pour l'écho, on utilise le rapport signal-à-écho (SER, *signal-to-echo ratio*) qui mesure l'énergie du signal désiré c_1 par rapport à celle de l'écho y. Le SER en entrée est défini par :

$$\text{SER}_\text{i} = 10 \log_{10} \frac{\|s\|^2}{\|y\|^2}, \tag{2.94}$$

et le SER en sortie, c'est-à-dire après application du système de rehaussement, est défini par :

$$\text{SER}_\text{o} = 10 \log_{10} \frac{\|c_{\text{cible}}^{\text{post}}\|^2}{\|y^{\text{post}}\|^2}. \tag{2.95}$$

On utilise aussi la métrique ERLE (en anglais *echo return loss enhancement*), qui mesure l'atténuation de l'écho par le système de rehaussement [Hänsler et Schmidt, 2004, Chapitre 3] :

$$\text{ERLE} = 10 \log_{10} \frac{\|y\|^2}{\|y^{\text{post}}\|^2}. \tag{2.96}$$

Déréverbération Comme défini à la partie 2.1.1.2, la réverbération est quantifiée de manière précise avec le DRR, qui mesure l'énergie du chemin direct par rapport à celle des réflexions précoces et de la réverbération tardive. En pratique, cette métrique est compliquée à déterminer précisément, car le chemin direct est difficilement distinguable des réflexions précoces. Comme les réflexions précoces renforcent l'intelligibilité de la parole, on utilise plutôt le rapport précoce-à-tardif (ELR, *early-to-late ratio*) dans cette thèse, qui mesure l'énergie de la composante précoce s_e sur la réverbération tardive s_l [Naylor et Gaubitch, 2010, Chapitre 2]. L'ELR en entrée est défini par :

$$\text{ELR}_\text{i} = 10 \log_{10} \frac{\|s_\text{e}\|^2}{\|s_\text{l}\|^2}, \tag{2.97}$$

et l'ELR en sortie est défini par :

$$\text{ELR}_\text{o} = 10 \log_{10} \frac{\|s_\text{e}^{\text{post}}\|^2}{\|s_\text{l}^{\text{post}}\|^2}. \tag{2.98}$$

Artefacts et distorsion globale Pour les artefacts, on utilise le rapport signal-à-artefacts invariant à l'échelle (SI-SAR, *scale-invariant signal-to-artifacts ratio*), défini par

$$\text{SI-SAR} = 10\log_{10}\frac{\|c_{\text{cible}}^{\text{post}}\|^2}{\|c_{\text{cible}}^{\text{art}}\|^2}. \tag{2.99}$$

Pour mesurer l'ensemble des distorsions, on utilise le rapport signal-à-distorsion invariant à l'échelle (SI-SDR, *scale-invariant signal-to-distortion ratio*), défini par :

$$\text{SI-SDR} = 10\log_{10}\frac{\|c_{\text{cible}}^{\text{post}}\|^2}{\|s_1^{\text{post}} + b^{\text{post}} + y^{\text{post}} + c_{\text{cible}}^{\text{art}}\|^2}. \tag{2.100}$$

2.4.2. Scores perceptuels

Un autre type de métriques donne une mesure en termes de score perceptuels. L'évaluation de la qualité vocale perçue (PESQ, *Perceptual Evaluation of Speech Quality*) donne un score de similarité entre le signal estimé et le signal cible [Int. Telecomm. Union (ITU-T) Rec., 2001]. La métrique STOI (*short-time objective intelligibility*) est davantage liée à la perception de la parole [Taal et al., 2010]. Enfin, PEASS regroupe des métriques qui intègrent aussi un aspect perceptuel [Emiya et al., 2011; Vincent, 2012]. Toutefois, nous ne les utilisons pas dans cette thèse car nous allons faire des tests d'écoute.

2.5. Résumé

Ce chapitre définit les trois types de distorsion considérés dans cette thèse pour le rehaussement, qui sont le bruit, l'écho et la réverbération. Il présente les méthodes de réduction individuelle de chaque type de distorsion, dont certaines nécessitent l'optimisation conjointe de deux filtres. Ce chapitre détaille ensuite les méthodes de réduction conjointe de deux types de distorsion. Ces méthodes nécessitent pour la plupart la combinaison de plusieurs filtres utilisés dans les méthodes de réduction individuelle du bruit, de l'écho et de la réverbération. Pour une configuration optimale, ces filtres nécessitent d'être optimisés conjointement, ce qui implique la modélisation des signaux transformés par ces filtres. En particulier, les méthodes basées sur l'apprentissage profond montrent leur efficacité à estimer simultanément les caractéristiques de ces signaux. Nous décrivons ensuite les méthodes de réduction conjointe de bruit, d'écho et de réverbération, dont la méthode de Togami et Kawaguchi [2014] et les limites de cette méthode. Toutes ces méthodes nous permettent d'introduire les filtres utilisés dans cette thèse, et les interactions entre ces filtres à modéliser pour le rehaussement de la parole cible.

3. Suppression d'écho résiduel par réseau de neurones combinée à l'annulation d'écho

Ce chapitre développe notre méthode monocanale de suppression d'écho résiduel. Ce type de méthode consiste à appliquer un post-filtre court visant à supprimer l'écho résiduel subsistant après l'annulation d'écho avec un filtre long. La combinaison de l'annulation d'écho et de la suppression d'écho résiduel fait partie des méthodes de réduction d'écho. Dans des scénarios réels, le filtre d'annulation d'écho n'a pas nécessairement convergé, ce qui amène l'écho résiduel à être potentiellement important. Le post-filtre de suppression d'écho résiduel doit donc être adapté à la fois aux périodes antérieures et postérieures à la convergence du filtre d'annulation d'écho. Nous détaillons l'approche proposée, qui estime les coefficients du post-filtre à l'aide d'un réseau de neurones. Ce réseau de neurones prend en entrée le signal après annulation d'écho, le signal source de l'écho, et l'écho estimé par le filtre d'annulation d'écho. De plus, nous proposons d'inclure l'information de phase dans le critère d'apprentissage du réseau de neurones. Nous étudions les performances du post-filtre dans plusieurs situations. Nous analysons l'impact des signaux d'entrée du réseau de neurones, ainsi que différents critères de détermination du post-filtre. Enfin, nous comparons notre approche à d'autres approches de réduction d'écho résiduel.

3.1. Formulation du problème

Nous avons développé dans la partie 2.2.2 les différentes approches utilisées pour la réduction d'écho. Dans ce travail, nous considérons un système qui combine l'annulation d'écho avec une suppression d'écho résiduel. Nous nous plaçons dans le cas monocanal, c'est-à-dire $M = 1$ microphone. Par souci de clarté, nous omettons l'indice m dans ce chapitre.

Cette approche, illustrée sur la figure 3.1, permet le meilleur compromis entre la réduction d'écho et la dégradation de la parole locale.

Nous rappelons ici brièvement le problème de la combinaison de l'annulation d'écho avec la suppression d'écho résiduel décrit dans la partie 2.2.2.2. Le mélange $d^{\text{echo}}(n, f)$ est la somme de la parole locale $s(n, f)$ et de l'écho acoustique $y(n, f)$:

$$d^{\text{echo}}(n, f) = s(n, f) + y(n, f). \tag{3.1}$$

Après annulation d'écho, le signal $e^{\text{echo}}(n, f)$ est constitué de la parole locale $s(n, f)$ et

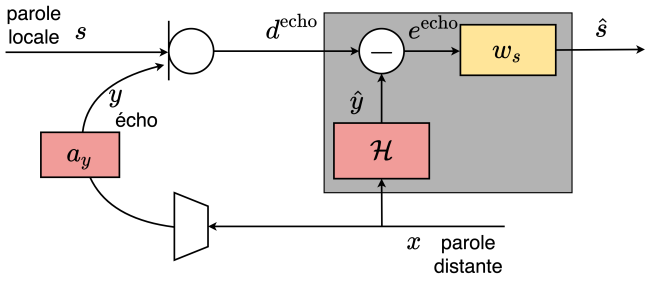


FIGURE 3.1. – Approche générale de réduction d'écho en monocanal combinant un filtre d'annulation d'écho $\mathcal{H}$ avec un post-filtre de suppression d'écho résiduel w_s.

d'un écho résiduel $z(n,f)$ qui subsiste pour plusieurs raisons (voir la partie 2.2.2) :

$$e^{\text{echo}}(n,f) - s(n,f) = \underbrace{y(n,f) - \widehat{y}(n,f)}_{=z(n,f)}, \tag{3.2}$$

où $\widehat{y}(n,f)$ est obtenu avec le filtre d'annulation d'écho $\mathcal{H}(f)$ comme dans (2.35). Pour extraire la parole locale $s(n,f)$, on applique un post-filtre court $w_s(n,f)$ sur le signal $e^{\text{echo}}(n,f)$:

$$\widehat{s}(n,f) = w_s(n,f)e^{\text{echo}}(n,f). \tag{3.3}$$

Le filtre $\mathcal{H}(f)$ est estimé de manière adaptative (voir la partie 2.2.2.1). Pour déterminer le post-filtre $w_s(n,f)$, les méthodes de suppression d'écho résiduel utilisent généralement le filtre de Wiener (voir (2.44)), ce qui nécessitent d'estimer les covariances de la parole locale $\Sigma_s(n,f)$ et de l'écho résiduel $\Sigma_z(n,f)$, où la covariance du signal $c(n,f)$ est définie de la manière suivante en monocanal :

$$\Sigma_c(n,f) = |c(n,f)|^2. \tag{3.4}$$

Parmi elles, les méthodes de soustraction spectrale modélisent la covariance de l'écho résiduel $\Sigma_z(n,f)$ et déduisent la covariance de la parole locale $\Sigma_s(n,f)$ par soustraction spectrale (voir la partie 2.2.2.2). Le filtre de Wiener s'exprime alors comme suit (voir la figure 3.2a) :

$$w_s^{\text{WF}}(n,f) = 1 - \frac{|z(n,f)|^2}{|e^{\text{echo}}(n,f)|^2} \tag{3.5}$$

La covariance de l'écho résiduel $|z(n,f)|^2$ est estimée en appliquant une transformation sur un seul signal, qui est soit la parole distante $x(n,f)$ [Gustafsson et al., 2002; Chhetri et al., 2005; Lee et Kim, 2007; Bendersky et al., 2008; Schwarz et al., 2013; Valero et al., 2014], soit l'écho estimé $\widehat{y}(n,f)$ [Turbin et al., 1997; Gustafsson et al., 2002; Hoshuyama et Sugiyama, 2006; Habets et al., 2008b]. Cette transformation peut inclure une modélisation explicite de la composante de l'écho résiduel $z(n,f)$ due aux non-linéarités [Chhetri et al., 2005; Bendersky et al., 2008; Schwarz et al., 2013] ou à l'écho résiduel tardif $y_{\text{l}}(n,f)$ [Habets et al., 2008b; Valero et al., 2014]. Ces méthodes ne considèrent que

le cas où que le filtre $\mathcal{H}(f)$ a convergé. Elles négligent alors la composante de l'écho résiduel $z(n,f)$ à l'erreur d'estimation du filtre $\mathcal{H}(f)$. Les méthodes d'estimation conjointe du filtre et du post-filtre modélisent cette composante [Enzner et Vary, 2006; Myllylä, 2006]. Toutefois, toutes ces méthodes déduisent la covariance de la parole locale $R_s(n,f)$ par soustraction spectrale, ce qui produit une mauvaise estimation de celle-ci, notamment lorsque le SER dans le signal $e^{\text{echo}}(n,f)$ est faible, dû au manque de convergence du filtre $\mathcal{H}(f)$.

Pour résoudre ce problème d'estimation, Lee et al. [2015] a proposé une méthode qui estime directement le post-filtre $w_s(n,f)$ avec un réseau de neurones (voir la figure 3.2b). Celui-ci prend en entrée deux signaux, qui sont la parole distante $x(n,f)$ et le signal après annulation d'écho $e^{\text{echo}}(n,f)$, pour combiner leurs avantages respectifs. Pour déterminer le post-filtre $w_s(n,f)$, cette méthode utilise le critère MIA [Weninger et al., 2014a] (voir la partie 2.2.1.5) plutôt que le filtre de Wiener :

$$w_s^{\text{MIA}}(n,f) = \frac{|s(n,f)|}{|e^{\text{echo}}(n,f)|}. \tag{3.6}$$

Toutefois, cette méthode n'a pas été évaluée dans le cas où le filtre $\mathcal{H}(f)$ n'a pas convergé.

(a) Soustraction spectrale.

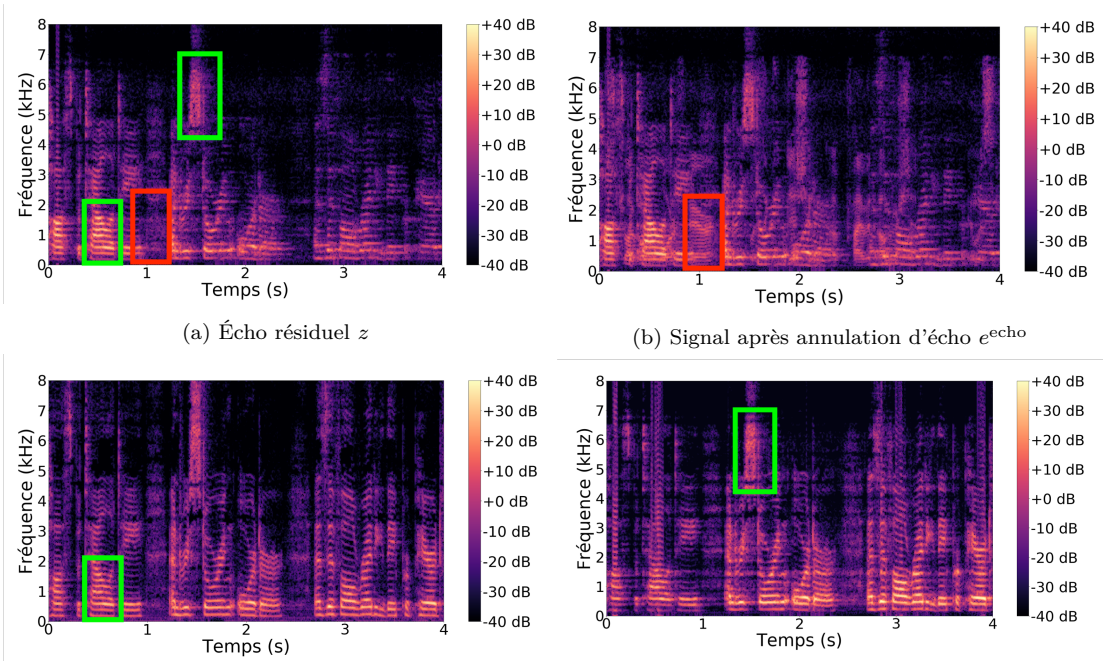
(b) Estimation directe.

FIGURE 3.2. – Méthodes de suppression d'écho résiduel en monocanal.

3.2. Solution proposée

Pour obtenir un post-filtre $w_s(n,f)$ plus performant durant les périodes antérieures et postérieures à la convergence du filtre $\mathcal{H}(f)$, nous proposons une approche estimant directement le post-filtre $w_s(n,f)$, à l'aide d'un réseau de neurones de type MLP. Nous utilisons les signaux $e^{\text{echo}}(n,f)$, $x(n,f)$ et $\widehat{y}(n,f)$ en entrée de ce réseau pour combiner l'information contenue dans chacun de ces signaux (voir la figure 3.3). En particulier, l'écho estimé $\widehat{y}$ possède trois avantages. Premièrement, il contient une information du chemin d'écho $a_y(k,f',f)$, ce qui permet de mieux modéliser la composante de l'écho résiduel $z(n,f)$ due à l'écho résiduel tardif $y_{\text{l}}(n,f)$, en particulier lorsque le filtre $\mathcal{H}(f)$ a convergé. Deuxièmement, l'écho estimé $\widehat{y}(n,f)$ contient le niveau de volume de l'écho

$y(n, f)$, contrairement à la parole distante $x(n, f)$. Enfin, il contient une information sur l'état de convergence du filtre $\mathcal{H}(f)$, c'est-à-dire sur la quantité d'écho réduite par le filtre $\mathcal{H}(f)$.



(a) Écho résiduel z

(b) Signal après annulation d'écho e^{echo}

(c) Parole distante x

(d) Écho estimé $\widehat{y}$

FIGURE 3.3. – Exemple de spectrogrammes des signaux utilisés en entrée du réseau de neurones dans notre approche. Les rectangles verts indiquent les zones qui sont similaires. Les rectangles rouges indiquent les zones qui sont dissemblables.

Pour des niveaux de SER élevés, la phase du signal $e^{\text{echo}}(n, f)$ sur lequel $w_s(n, f)$ est appliqué est relativement proche de la phase de la parole locale $s(n, f)$. Toutefois, lorsque le filtre $\mathcal{H}(f)$ n'a pas convergé, les niveaux de SER sont faibles et la phase de l'écho résiduel $\theta_z(n, f)$ domine dans le signal $e^{\text{echo}}(n, f)$. Pour tenir compte de l'information de phase dans la détermination du post-filtre $w_s(n, f)$, nous proposons donc d'utiliser le critère FSP [Erdogan et al., 2015] (voir la partie 2.2.1.5) :

$$w_s^{\text{FSP}}(n, f) = \frac{|s(n, f)|}{|e^{\text{echo}}(n, f)|} \cos\Big(\theta_s(n, f) - \theta_{e^{\text{echo}}}(n, f)\Big). \tag{3.7}$$

Nous comparons ce critère au critère MIA, défini dans (3.6), et au critère MIR [Wang et al., 2014] (voir la partie 2.2.1.5, et en particulier le tableau 2.2) :

$$w_s^{\text{MIR}}(n, f) = \frac{|s(n, f)|}{\sqrt{|s(n, f)|^2 + |z(n, f)|^2}}. \tag{3.8}$$

Il convient de remarquer le critère FSP correspond au critère MIA pondéré par le cosinus de la différence de phase $\theta_s(n, f) - \theta_e(n, f)$ entre la parole cible $s(n, f)$ et le signal

$e^{\text{echo}}(n, f)$. Nous considérons l'amplitude des signaux $|e^{\text{echo}}(n, f)|$, $|x(n, f)|$ et $|\widehat{y}(n, f)|$ comme entrées du MLP. La figure 3.4 illustre un exemple de la topologie utilisée.

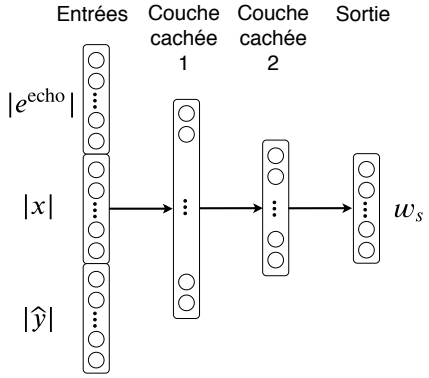


FIGURE 3.4. – Exemple de l'approche proposée basée sur un MLP à deux couches cachées prenant en entrée l'amplitude des signaux $|e^{\text{echo}}(n, f)|$, $|x(n, f)|$ et $|\widehat{y}(n, f)|$.

3.3. Protocole expérimental

3.3.1. Scénarios

Nous considérons la situation où un locuteur local interagit avec un correspondant distant à l'aide de Triby à une distance de 1 m de ce système dans un environnement silencieux. Nous étudions 3 scénarios de 10 s : 1) *parole distante seule*, 2) *parole locale seule*, 3) *parole simultanée* (paroles locale et distante actives simultanément).

3.3.2. Données

3.3.2.1. Description générale

Nous créons trois ensembles de données disjoints pour l'apprentissage, la validation et le test, dont les caractéristiques sont résumées dans le tableau 4.1. Pour chaque ensemble de données, nous avons créé séparément l'écho acoustique $y(t)$ et la parole locale $s(t)$, à partir de signaux originaux de parole. Nous avons enregistré l'écho avec un Triby, l'enceinte intelligente développée par Invoxia, qui est dotée de $M = 4$ microphones. Pour obtenir un signal d'écho monophonique $y(t)$ ($M = 1$), nous avons ensuite appliqué un filtrage spatial fixe. Nous avons soit simulé, soit enregistré avec le Triby (suivi d'un filtrage spatial fixe) la parole locale $s(t)$. Le signal du microphone $d^{\text{echo}}(t)$ a été calculé comme dans (4.1). Ce protocole est nécessaire afin d'obtenir les vérités terrain des signaux pour l'apprentissage et la validation, ce qui n'est pas possible avec des enregistrements réels où ces signaux ne sont pas observés séparément. Les trois ensembles de données correspondent à des conditions acoustiques invariantes dans le temps.

Ensemble de données	**Apprentissage**	**Validation**	**Test**
Signaux y s	enregistrés RIRs a_s mesurées		enregistrés
Salles	1	1	2
# locuteurs	27	5	7
# phrases	629	205	208
Plage de SER (dB)	$[-15, -9]$		

TABLEAU 3.1. – Caractéristiques des trois ensembles de données.

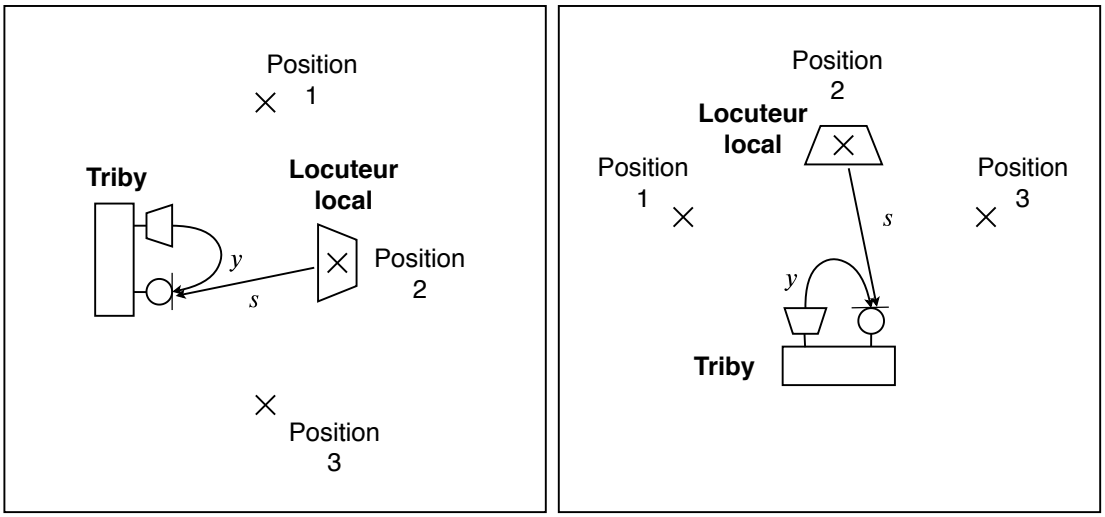
Signaux originaux de parole Les signaux de parole pure proviennent du sous-ensemble « dev-clean » du corpus Librispeech [Panayotov et al., 2015], qui comprend 84 locuteurs lisant des extraits de livres, dont l'ensemble des segments par locuteur font en moyenne 8 min, soit 11, 2 h de livres audio au total. Nous avons sélectionné 39 locuteurs, répartis en sous-ensembles disjoints entre l'apprentissage, la validation et le test. Nous avons considéré chaque locuteur au moins une fois comme locuteur local et correspondant distant. Nous avons réparti les livres audio en segments disjoints de 10 s. Chaque segment a été utilisé pour créer soit la parole locale $s(t)$, soit l'écho $y(t)$.

Enregistrements d'écho réels Dans les systèmes mains-libres, l'écho acoustique $y(t)$ contient des non-linéarités dues à la réponse non-linéaire du haut-parleur et des microphones, aux vibrations de l'enceinte et aux effets d'écrêtage de l'amplification (voir la partie 2.2.2). Pour réaliser les tests dans des conditions réalistes, nous créons l'écho $y(t)$ en enregistrant la rétroaction acoustique du haut-parleur du Triby vers les microphones du Triby. La parole distante $x(t)$ a été jouée à une fréquence d'échantillonnage de 16 kHz par le Triby, et l'écho correspondant a été enregistré par l'antenne linéaire de 4 microphones de ce même Triby.

La distance entre le haut-parleur et l'antenne de microphones du Triby est de 11 cm, et la distance entre les microphones est de 3 cm. Les enregistrements ont été réalisées dans 2 salles de taille et temps de réverbération différents (voir le tableau 3.2). Le Triby a été placé au centre de chaque salle dans 2 positions différentes pour augmenter la diversité des chemins d'écho $a_y(\tau)$ (voir la figure A.3). La parole distante $x(t)$ a été jouée à 3 réglages de volume différents du Triby, pour augmenter la diversité des non-linéarités : plus la parole distante $x(t)$ est forte, plus il y a de non-linéarités dans l'écho $y(t)$. Le niveau de bruit des salles était d'environ 50 dBA, mesuré avec un sonomètre.

Salle	**Dimensions** (m)	$\mathbf{T}_{60}$ (s)
1	$3 \times 3 \times 3$	$0,3$
2	$7 \times 7 \times 3$	$0,5$

TABLEAU 3.2. – Caractéristiques des salles.



(a) Position 1 du Triby.

(b) Position 2 du Triby.

FIGURE 3.5. – Schéma des configurations expérimentales de création des signaux.

3.3.2.2. Ensemble d'apprentissage

Pour l'ensemble d'apprentissage, les enregistrements d'écho $y(t)$ ont été réalisés dans la salle 1 (voir le tableau 3.2). Pour créer la parole locale $s(t)$, nous avons placé un haut-parleur standard sur un demi-cercle de 1 m de rayon, centré sur l'enceinte connectée Triby, dans 3 positions : 0°, 60°et −60°par rapport à la direction dans laquelle joue le Triby (voir la figure A.3). Nous avons mesuré 3 RIRs a_s avec ce haut-parleur, correspondant aux 3 positions considérées, que nous avons ensuite convoluées aux signaux de parole pure $u(t)$ pour simuler la parole locale $s(t)$ comme dans (2.4). Nous avons utilisé 27 locuteurs que nous avons groupés par paire de manière aléatoire pour chacune des phrases. Les niveaux de volume de l'écho enregistré $y(t)$ et de la parole locale $s(t)$ ont été choisis tels que le SER varie de −9 dB à −15 dB. Il convient de souligner que nous avons seulement fait varier le niveau de la parole locale $s(t)$ pour obtenir ces niveaux de SER. Au total, nous avons enregistré 629 signaux d'écho, et nous avons simulé 629 signaux de parole locale, pour chacun des 3 scénarios considérés, soit environ 5, 2 h d'audio (voir le tableau 4.1).

3.3.2.3. Ensemble de validation

Les signaux pour l'ensemble de validation ont été générés de la même manière que pour l'ensemble d'entraînement, en utilisant 5 locuteurs qui ne sont pas dans l'ensemble d'apprentissage. Les enregistrements d'écho $y(t)$ ont été réalisés dans la salle 1, et les mêmes RIRs ont été utilisées pour simuler la parole locale $s(t)$. Les niveaux de volume de l'écho enregistré $y(t)$ et de la parole locale $s(t)$ ont été choisis dans le même intervalle de SER que l'ensemble d'apprentissage. Au total, nous avons enregistré 205 signaux d'écho, et nous avons simulé 205 signaux de parole locale, pour chacun des 3 scénarios

considérés, soit environ 1, 7 h d'audio (voir le tableau 4.1).

3.3.2.4. Ensemble de test

L'ensemble de test a été créé uniquement à partir d'enregistrements réels, en utilisant 7 locuteurs qui ne sont ni dans l'ensemble d'apprentissage, ni dans l'ensemble de validation. L'écho $y(t)$ et la parole locale $s(t)$ ont été enregistrés dans la salle 2 (voir le tableau 3.2) avec un Triby différent que pour les ensembles d'apprentissage et de validation. Les niveaux de volume de l'écho $y(t)$ et de la parole locale $s(t)$ enregistrés ont été choisis dans le même intervalle de SER que l'ensemble d'apprentissage. Nous n'avons considéré que 2 scénarios pour l'ensemble de test : *far-end talk* et *double-talk*. Au total, nous avons enregistré 208 signaux d'écho et de parole locale, pour chacun des 2 scénarios considérés, soit environ 1, 2 h d'audio (voir le tableau 4.1).

3.3.3. Métriques

La parole locale estimée $\widehat{s}$, obtenue avec (3.3), contient trois composantes :

$$\widehat{s} = s^{\text{post}} + z^{\text{post}} + s^{\text{art}}, \tag{3.9}$$

où s^{post} est la parole locale potentiellement atténuée, z^{post} est l'écho post-résiduel qui est idéalement égal à zéro, et s^{art} représente les artefacts introduits dans la parole locale s. Ces composantes sont calculées de la même que dans (2.89)–(2.91), sans la composante de réverbération post-résiduelle $s_{\text{l}}^{\text{post}}$. Nous utilisons les métriques définies dans la partie 2.4.1 pour évaluer la réduction d'écho et la dégradation de la parole locale. Ces métriques sont résumées dans le tableau 3.3.

Pour la réduction d'écho, nous utilisons le SER et l'ERLE. Les artefacts introduits dans la parole locale sont mesurés avec le SI-SAR et l'ensemble des distorsions de la parole locale est mesuré avec le SI-SDR. Les SER, SI-SAR et SI-SDR sont évalués uniquement dans le scénario de *parole simultanée*. L'ERLE est évalué à la fois dans les scénarios de *parole distante seule* et de *parole simultanée*.

Comme les performances peuvent varier selon le scénario et l'état de convergence du filtre $\mathcal{H}(f)$, nous calculons les métriques séparément dans chaque scénario et dans chaque période avant et après convergence du filtre $\mathcal{H}(f)$. En particulier, les métriques dépendent de l'estimation d'un facteur d'échelle γ_c associé au signal c et défini comme dans (2.90) :

$$\gamma_c = \frac{\langle \widehat{s}, c \rangle}{\|c\|^2}. \tag{3.10}$$

Dans chaque scénario de l'ensemble de test (*parole distante seule* et *parole simultanée*), nous supposons que γ_c est fixe pendant la période de convergence du filtre $\mathcal{H}(f)$, et pendant la période après convergence. Toutefois, nous supposons que γ_c peut varier entre ces deux périodes. Nous faisons ensuite la moyenne pondérée pour chaque métrique en fonction de la durée de chaque période sur laquelle sont calculées ces métriques, de la même manière que le calcul du SNR segmenté [Vincent et al., 2006].

Écho	ERLE	$10\log_{10}\frac{\|y\|^2}{\|z^{\text{post}}\|^2}$
	SER	$10\log_{10}\frac{\|s^{\text{post}}\|^2}{\|z^{\text{post}}\|^2}$
Artefacts	SI-SAR	$10\log_{10}\frac{\|s^{\text{post}}\|^2}{\|s^{\text{art}}\|^2}$
Distorsion globale	SI-SDR	$10\log_{10}\frac{\|s^{\text{post}}\|^2}{\|z^{\text{post}}+s^{\text{art}}\|^2}$

TABLEAU 3.3. – Métriques considérées pour la réduction d'écho.

3.3.4. Méthodes de référence

L'annulation d'écho est réalisée avec SpeexDSP [1], qui est une implémentation de l'approche adaptative de Valin [2007] (voir la partie 2.2.2). En ce qui concerne la suppression d'écho résiduel, nous comparons notre méthode avec trois méthodes de référence :

1. la méthode de soustraction spectrale de Schwarz et al. [2013] qui calcule le post-filtre $w_s^{\text{WF}}(n,f)$ comme dans (3.5), en estimant la covariance de l'écho résiduel $|z(n,f)|^2$ à partir d'un MLP appliqué sur l'amplitude de la parole distante $|x(n,f)|$,

2. une méthode de soustraction spectrale basée sur l'estimation conjointe du filtre et du post-filtre, qui calcule le post-filtre $w_s^{\text{WF}}(n,f)$ comme dans (3.5), en modélisant la covariance de l'écho résiduel $|z(n,f)|^2$ à partir d'une transformation linéaire appliquée sur l'amplitude de l'écho estimé $|\widehat{y}(n,f)|$ [Valin, 2007],

3. la méthode de Lee et al. [2015] qui estime directement le post-filtre $w_s^{\text{MIA}}(n,f)$ à partir d'un MLP appliqué sur l'amplitude de la parole distante $|x(n,f)|$ et du signal après annulation d'écho $|e^{\text{echo}}(n,f)|$.

À l'origine, la méthode de Valin [2007] est utilisée pour estimer uniquement le filtre $\mathcal{H}(f)$. Toutefois, nous pouvons l'utiliser comme méthode d'estimation conjointe du filtre $\mathcal{H}(f)$ et du post-filtre $w_s(n,f)$, en utilisant l'estimation de $|z(n,f)|$, dans la mise à jour du filtre $\mathcal{H}(f)$, pour le calcul du post-filtre $w_s^{\text{WF}}(n,f)$ comme dans (3.5).

Il convient de noter que l'approche proposée correspond à l'approche de Lee et al. [2015], avec l'écho estimé $|\widehat{y}(n,f)|$ en entrée supplémentaire du MLP, et le critère FSP pour l'optimisation du réseau de neurones. Afin d'analyser l'influence de ces deux paramètres, nous considérons aussi deux variantes de l'approche de Lee et al. [2015] : la première avec l'écho estimé $|\widehat{y}(n,f)|$ en entrée supplémentaire du MLP, et la seconde avec le critère FSP.

1. https ://github.com/xiph/speexdsp

3.3.5. Réglage des hyperparamètres

Annulation d'écho Pour l'annulation d'écho, nous appliquons SpeexDSP sur le signal du microphone $d^{\text{echo}}(n, f)$ séparément sur chaque scénario (*parole locale seule*, *parole distante seule*, *parole simultanée*). Puisque SpeexDSP utilise des fenêtres de TFCT rectangulaires à pas d'avancement de 50%, nous utilisons une taille de fenêtre de $T_{\text{TFCT}} = 640$ échantillons et un pas d'avancement de $P = 320$ échantillons. Nous fixons la taille du filtre à $0,16$ s dans le domaine temporel, soit $K = 8$ trames. Cette configuration permet un compromis entre réduction d'écho, vitesse de convergence et complexité de l'algorithme. Comme les conditions acoustiques sont ici invariantes dans le temps, nous observons que SpeexDSP converge au bout d'une période 4 s après le début de la phrase, ce qui est en accord avec les observations de Valin [2007]. Il convient de remarquer que durant cette période de 4 s, le filtre $\mathcal{H}(f)$ n'a pas encore convergé. Par conséquent, l'écho résiduel $\mathbf{z}(n, f)$ est plus important avant $t = 4$ s qu'après $t = 4$ s.

Suppression d'écho résiduel Pour la suppression d'écho résiduel, nous utilisons la même taille de fenêtre de TFCT T_{TFCT} et le même pas P que pour l'annulation d'écho, avec une fenêtre de Hanning, ce qui conduit à obtenir $F = 321$ bandes de fréquence. À cause des interférences destructives, il est possible que le rapport $|z(n, f)/|e(n, f)| > 1$ dans (3.5) pour certains points temps-fréquence. Cela signifie alors que la covariance de la parole locale $\Sigma_s(n, f) < 0$. Dans ce cas, nous supposons $\Sigma_s(n, f) = 0$. Cela revient à obtenir une version tronquée du filtre $w_s^{\text{WF}}(n, f)$ dans l'intervalle $[0; 1]$ [Schwarz et al., 2013] :

$$w_s^{\text{WF}}(n, f) = \max\left(1 - \frac{|z(n, f)|^2}{|e(n, f)|^2}, 0\right). \tag{3.11}$$

En ce qui concerne les méthodes d'estimation directe du post-filtre $w_s(n, f)$, nous considérons une version tronquée du critère MIA dans l'intervalle $[0; 1]$ pour compresser la dynamique du post-filtre $w_s(n, f)$ [Lee et al., 2015] :

$$w_s^{\text{MIA}}(n, f) = \min\left(1, \frac{|s(n, f)|}{|e(n, f)|}\right). \tag{3.12}$$

De manière similaire,nous considérons une version tronquée du critère FSP dans l'intervalle $[-1; +1]$:

$$w_s^{\text{FSP}}(n, f) = \max\left(\min\left(1, \frac{|s(n, f)|}{|e(n, f)|}\cos\left(\theta_s(n, f) - \theta_e(n, f)\right)\right), -1\right). \tag{3.13}$$

Réseau de neurones En ce qui concerne le réseau de neurones de notre méthode d'estimation et de celles Schwarz et al. [2013] et de Lee et al. [2015], nous utilisons une architecture MLP à deux couches cachées. Nous choisissons 1 024 neurones par couche cachée avec la fonction tangente hyperbolique comme fonction d'activation. Pour la couche de sortie, nous prenons la fonction sigmoïde pour les critères MIR et MIA, car

sa sortie est bornée dans l'intervalle $[0; 1]$, et la fonction tangente hyperbolique pour le critère FSP, car sa sortie est bornée dans l'intervalle $[-1; +1]$. En ce qui concerne la méthode de Schwarz et al. [2013], nous considérons la même MLP que celui considéré par les auteurs, c'est-à-dire avec 2 couches cachées, et avec la fonction ReLU comme fonction d'activation pour toutes les couches.

3.4. Résultats et discussion

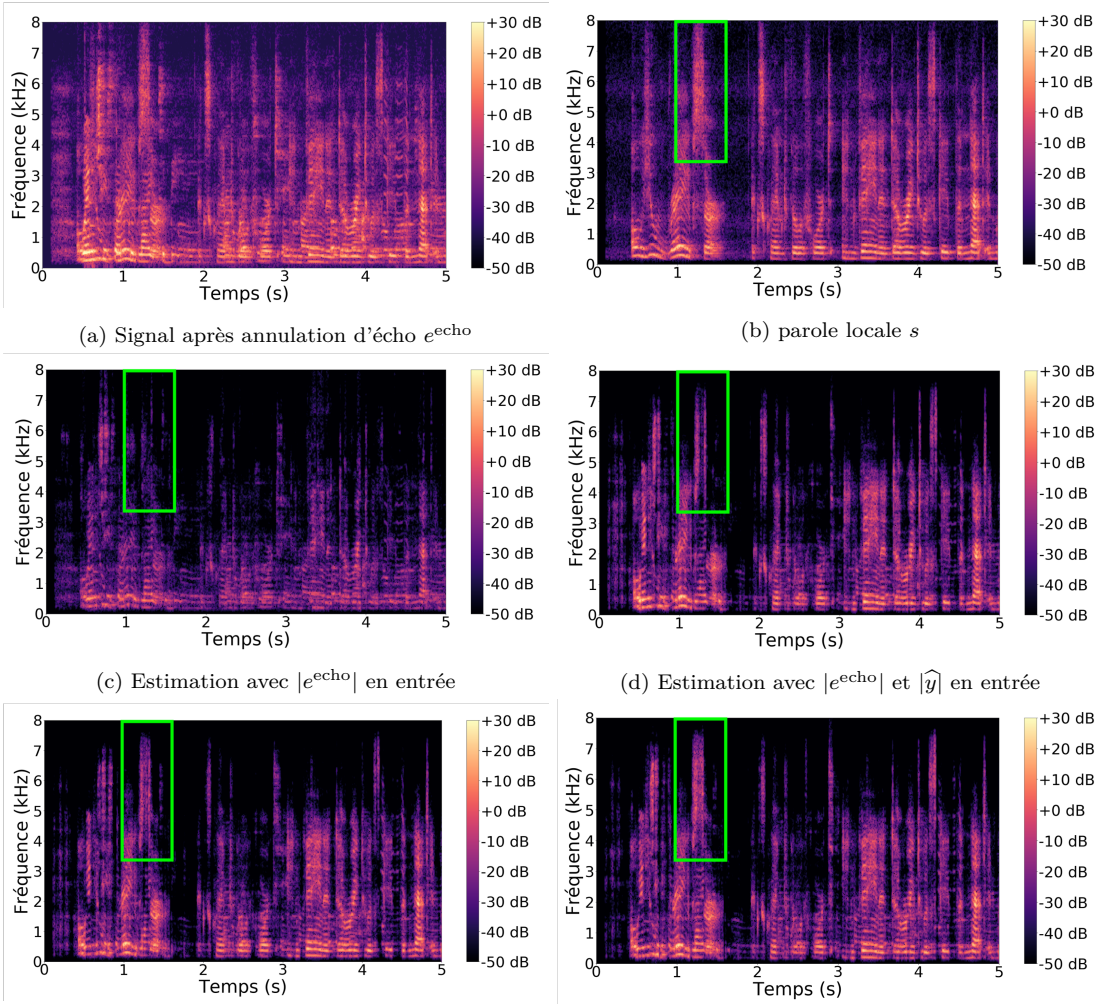
Tout d'abord, nous étudions différentes configurations du post-filtre proposé. En particulier, nous examinons l'impact des signaux $e^{\text{echo}}(n, f)$, $x(n, f)$, $\widehat{y}(n, f)$ utilisés en entrée du post-filtre proposé, ainsi que l'impact du critère (MIR, MIA et FSP) pour l'optimisation du réseau de neurones. Enfin, nous comparons la meilleure configuration du post-filtre proposé aux méthodes de référence.

3.4.1. Signaux d'entrée et type de critère

Signaux d'entrée du réseau de neurones Le tableau 3.4 montre les performances moyennes du post-filtre proposé $w_s(n, f)$ en fonction des signaux $e^{\text{echo}}(n, f)$, $x(n, f)$, $\widehat{y}(n, f)$ en entrée de l'approche proposée. Nous avons fait la moyenne des performances sur tous les critères (MIR, MIA et FSP). L'utilisation des signaux $e^{\text{echo}}(n, f)$ avec $\widehat{y}(n, f)$ et/ou $x(n, f)$ améliore significativement les performances selon toutes les métriques par rapport à l'utilisation du signal $e^{\text{echo}}(n, f)$ seul. La combinaison du signal $e^{\text{echo}}(n, f)$ avec la parole distante $x(n, f)$ améliore significativement le SI-SDR par rapport à la combinaison du signal $e^{\text{echo}}(n, f)$ avec l'écho estimé $\widehat{y}(n, f)$. Comme le SI-SAR de la combinaison $|e^{\text{echo}}(n, f)|, |x(n, f)|$ est significativement supérieur au SI-SAR de la combinaison $|e^{\text{echo}}(n, f)|, |\widehat{y}(n, f)|$, nous en déduisons que la combinaison $|e^{\text{echo}}(n, f)|, |x(n, f)|$ est meilleure que la combinaison $|e^{\text{echo}}(n, f)|, |\widehat{y}(n, f)|$ car elle introduit moins de dégradations dans la parole locale $s(n, f)$.

L'utilisation des trois signaux $e(n, f)$, $x(n, f)$ et $\widehat{y}(n, f)$ permet d'obtenir les meilleures performances selon toutes les métriques. L'amélioration des métriques n'est cependant pas significative par rapport à l'utilisation des deux signaux $e(n, f)$ et $x(n, f)$. Toutefois, nous remarquons que la différence de performances entre les combinaisons $|e(n, f)|$, $|x(n, f)|$, $|\widehat{y}(n, f)|$ et $|e^{\text{echo}}(n, f)|, |\widehat{y}(n, f)|$ est plus significative que la différence entre les combinaisons $|e^{\text{echo}}(n, f)|, |x(n, f)|$ et $|e^{\text{echo}}(n, f)|, |\widehat{y}(n, f)|$, notamment en SER. La parole distante $x(n, f)$ et l'écho estimé $\widehat{y}(n, f)$ contiennent donc des informations complémentaires au signal $e(n, f)$ pour l'estimation du post-filtre $w_s(n, f)$. Nous choisissons la combinaison des trois signaux $e(n, f)$, $x(n, f)$ et $\widehat{y}(n, f)$ pour le post-filtre proposée. La figure 3.6 illustre un exemple de spectrogrammes de la parole locale estimée $\widehat{s}(n, f)$ en fonction des signaux utilisés en entrée.

La figure 3.6 illustre un exemple de spectrogrammes de la parole locale estimée $\widehat{s}(n, f)$ avec le critère FSP en fonction des signaux utilisés en entrée. La figure 3.6 montre les zones d'amélioration de la parole locale estimé $\widehat{s}(n, f)$.



(a) Signal après annulation d'écho e^{echo}

(b) parole locale s

(c) Estimation avec $|e^{\text{echo}}|$ en entrée

(d) Estimation avec $|e^{\text{echo}}|$ et $|\widehat{y}|$ en entrée

(e) Estimation avec $|e^{\text{echo}}|$ et $|x|$ en entrée

(f) Estimation avec $|e^{\text{echo}}|$, $|x|$ et $|\widehat{y}|$ en entrée

FIGURE 3.6. – Exemple de spectrogrammes de la parole locale estimée $\widehat{s}$ avec le critère FSP en fonction des signaux utilisés en entrée de l'approche proposée. Le rectangle vert montre une zone du spectrogramme dont l'estimation est améliorée avec l'ajout de signaux en entrée.

	SI-SDR	**SI-SAR**	**SER**	**ERLE**
$\lvert e^{\text{echo}}\rvert$	$-3,4 \pm 0,5$	$-2,6 \pm 0,5$	$15,7 \pm 1,3$	$43,9 \pm 1,2$
$\lvert e^{\text{echo}}\rvert, \lvert\widehat{y}\rvert$	$1,1 \pm 0,4$	$1,5 \pm 0,4$	$24,8 \pm 1,3$	$48,9 \pm 1,3$
$\lvert e^{\text{echo}}\rvert, \lvert x\rvert$	$3,3 \pm 0,4$	$3,8 \pm 0,4$	$26,3 \pm 1,2$	$49,3 \pm 1,3$
$\lvert e^{\text{echo}}\rvert, \lvert x\rvert, \lvert\widehat{y}\rvert$	$\mathbf{3,8 \pm 0,5}$	$\mathbf{4,2 \pm 0,4}$	$\mathbf{27,6 \pm 1,2}$	$\mathbf{50,3 \pm 1,3}$

TABLEAU 3.4. – Performances moyennes (en dB) du post-filtre proposé en fonction des signaux d'entrée du réseau de neurones.

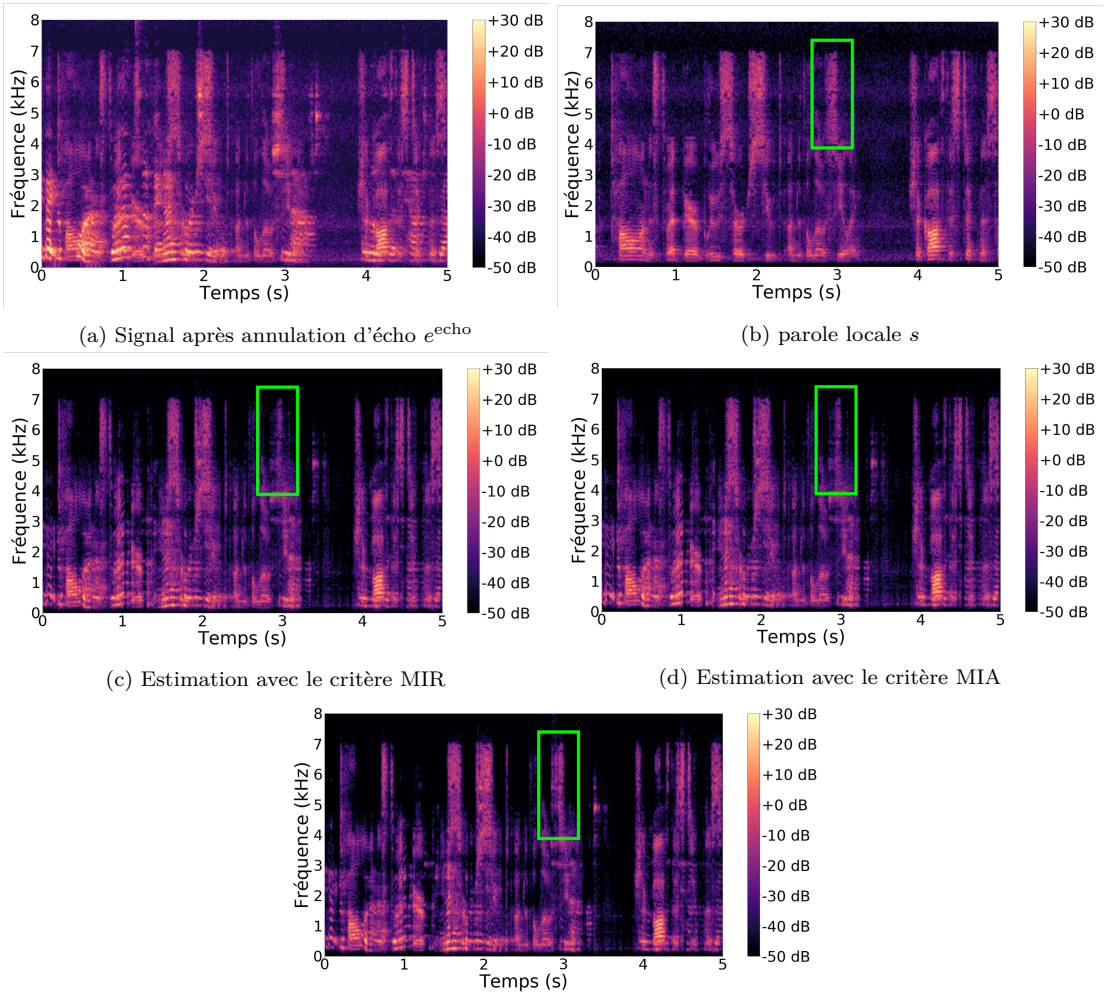
Critère de détermination Le tableau 3.5 montre les performances moyennes de l'approche proposée en fonction du critère d'optimisation du réseau de neurones (MIR, MIA et FSP). Nous avons fait la moyenne des performances sur toutes les combinaisons de signaux d'entrée du réseau de neurones (voir le paragraphe précédent). Les critères MIR et MIA ont des performances similaires en SI-SDR et SER. Le critère FSP permet d'obtenir les meilleures performances selon toutes les métriques, bien que la différence en SI-SDR, SI-SAR et SER ne soit pas significative. L'ERLE est significativement plus élevé pour le critère FSP que pour les critères MIR et MIA. L'information de phase dans le critère de détermination du post-filtre $w_s(n,f)$ permet donc de réduire plus fortement l'écho résiduel $z(n,f)$, tout en évitant d'introduire plus de dégradation dans la parole locale $s(n,f)$. Ceci vient du fait que la phase de la parole estimée $\widehat{s}(n,f)$ avec le critère FSP (voir (3.7)) se rapproche plus de la phase de la parole locale $s(n,f)$ qu'avec les critères MIR et MIA (voir (3.8) et (3.6)).

	SI-SDR	**SI-SAR SER**	**ERLE**	
MIR	$1,1 \pm 0,4$	$1,6 \pm 0,4$	$23,2 \pm 1,3$	$47,2 \pm 1,3$
MIA	$1,0 \pm 0,4$	$1,5 \pm 0,4$	$23,1 \pm 1,2$	$47,1 \pm 1,2$
FSP	$\mathbf{1,6 \pm 0,5}$	$\mathbf{2,0 \pm 0,5}$	$\mathbf{24,6 \pm 1,3}$	$\mathbf{50,0 \pm 1,3}$

TABLEAU 3.5. – Performances moyennes (en dB) du post-filtre proposé en fonction du critère d'optimisation.

Le tableau 3.6 montre les performances de l'approche proposée en fonction du critère d'optimisation du réseau de neurones (MIR, MIA et FSP), avec cette fois les trois signaux $e^{\text{echo}}(n,f)$, $x(n,f)$ et $\widehat{y}(n,f)$ en entrée. Le critère FSP permet d'obtenir les meilleures performances selon toutes les métriques. En particulier, les SI-SDR et SI-SAR sont significativement plus élevés que pour les critères MIR et MIA. Ceci confirme donc l'importance de l'information de phase dans le critère d'optimisation.

La figure 3.7 illustre un exemple de spectrogrammes de la parole locale estimée $\widehat{s}(n,f)$ en fonction des critères de détermination du post-filtre $w_s(n,f)$, en utilisant les les trois signaux $e^{\text{echo}}(n,f)$, $x(n,f)$ et $\widehat{y}(n,f)$ en entrée. La figure 3.7 montre les zones d'amélioration de la parole locale estimé $\widehat{s}(n,f)$.



(a) Signal après annulation d'écho e^{echo}

(b) parole locale s

(c) Estimation avec le critère MIR

(d) Estimation avec le critère MIA

(e) Estimation avec le critère FSP

FIGURE 3.7. – Exemple de spectrogrammes de la parole locale estimée $\widehat{s}$ avec l'amplitude des signaux $|e^{\text{echo}}(n,f)|$, $|x(n,f)|$ et $|\widehat{y}(n,f)|$ en entrée de l'approche proposée en fonction des critères de détermination. Le rectangle vert montre une zone du spectrogramme dont l'estimation est améliorée avec le critère FSP.

	SI-SDR	**SI-SAR**	**SER**	**ERLE**
MIR	$3,4 \pm 0,4$	$3,8 \pm 0,4$	$26,8 \pm 1,2$	$49,7 \pm 1,3$
MIA	$3,5 \pm 0,4$	$3,9 \pm 0,4$	$26,9 \pm 1,2$	$50,0 \pm 1,3$
FSP	$\mathbf{4,5 \pm 0,5}$	$\mathbf{4,8 \pm 0,5}$	$\mathbf{29,0 \pm 1,2}$	$\mathbf{51,1 \pm 1,3}$

TABLEAU 3.6. – Performances (en dB) du post-filtre proposé avec les trois signaux $e^{\text{echo}}(n,f)$, $x(n,f)$ et $\widehat{y}(n,f)$ en entrée en fonction du critère d'optimisation.

3.4.2. Comparaison aux méthodes de référence

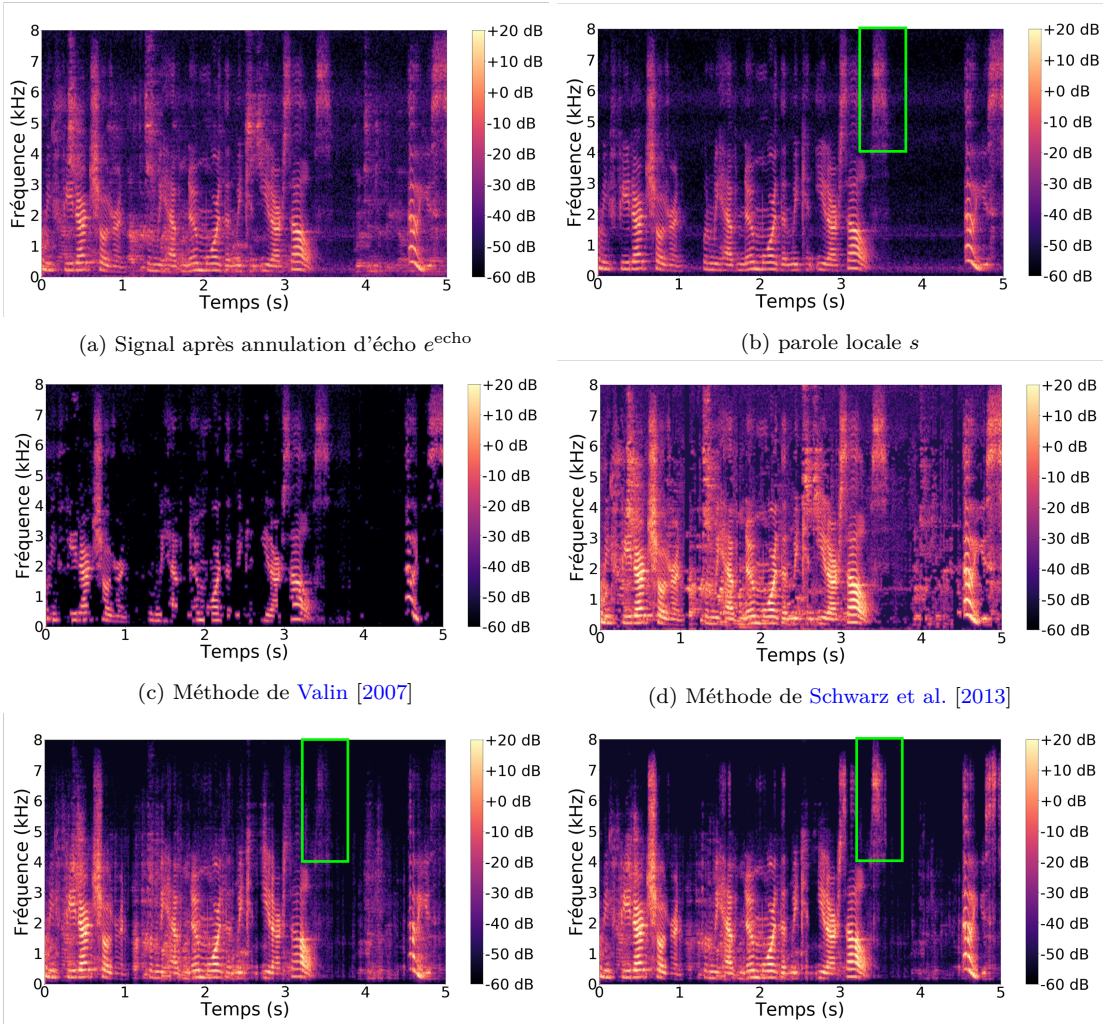
3.4.2.1. Performances moyennes

Nous avons obtenu la meilleure configuration de notre approche en utilisant les signaux $e(n,f)$, $x(n,f)$ et $\widehat{y}(n,f)$ avec le critère FSP (voir le tableau 3.6). Le tableau 3.7 montre la moyenne des performances de la méthode proposée et des méthodes de référence sur tous les scénarios (*parole distante seule* et *parole simultantée*). L'approche proposée obtient le meilleur SI-SDR. Elle dépasse significativement l'approche de Lee et al. [2015] dans toutes les métriques. Nous en déduisons que l'approche proposée est meilleure que celle de Lee et al. [2015] car elle introduit moins de dégradations dans la parole locale $s(n,f)$ (SI-SAR plus élevé), tout en réduisant plus l'écho (ERLE plus élevé). Cette explication est confirmée par un SER plus élévé. Pour comprendre cette différence de performance, nous avons ajouté les deux variantes de l'approche de Lee et al. [2015] au tableau 3.7. Nous remarquons que la différence est plus significativement plus élevée avec le critère FSP plutôt qu'avec le signal d'entrée $\widehat{y}$. Cependant, la combinaison du signal d'entrée $\widehat{y}$ et du critère FSP permet aussi d'améliorer les performances par rapport à leur utilisation séparée.

L'approche proposée dépasse significativement l'approche de Schwarz et al. [2013] en SI-SDR, SER et ERLE. Les performances en SI-SDR s'expliquent du fait d'une meilleure réduction d'écho (ERLE plus élevé) et d'une introduction d'artefacts moins élevée (SI-SAR plus élevé). De plus, l'ERLE de l'approche de Schwarz et al. [2013] est proche de l'ERLE de la méthode sans post-filtre. Nous en déduisons que l'approche de Schwarz et al. [2013] ne réduit que très peu l'écho résiduel.

L'approche proposée dépasse significativement l'approche de Valin [2007] en SI-SDR. Malgré que l'approche de Valin [2007] soit nettement plus agressive en réduction d'écho (ERLE plus élevé), cette différence s'explique du fait que l'approche proposée introduit beaucoup moins de dégradations dans la parole locale $s(n,f)$ (SI-SAR proche de 0dB pour l'approche de Valin [2007]). Le SER de l'approche de Valin [2007]) reste cependant plus élevé que l'approche proposée en raison de la forte réduction d'écho.

La figure 3.8 illustre un exemple de spectrogrammes de la parole locale estimée $\widehat{s}(n,f)$ par la méthode proposée et les différentes méthodes de l'état de l'art. Il est possible de voir que la méthode de Valin [2007] est très agressive pour la réduction d'écho résiduel, que la méthode de Schwarz et al. [2013] réduit peu l'écho résiduel, et que l'approche proposée produit une meilleure estimation de $s(n,f)$ que l'approche de Lee et al. [2015].



(a) Signal après annulation d'écho e^{echo}

(b) parole locale s

(c) Méthode de Valin [2007]

(d) Méthode de Schwarz et al. [2013]

(e) Méthode de Lee et al. [2015]

(f) Méthode proposée

FIGURE 3.8. – Exemple de spectrogrammes de la parole locale estimée $\widehat{s}$ avec les méthodes de référence et la méthode proposée. Le rectangle vert montre une zone du spectrogramme dont l'estimation est améliorée entre l'approche de Lee et al. [2015] et l'approche proposée.

Méthode de référence	**SI-SDR**	**SER**	**SI-SAR**	**ERLE**
$\emptyset$ post-filtre (filtre $\mathcal{H}(f)$ seul)	$0,7 \pm 0,8$	$15,9 \pm 1,5$	$2,5 \pm 0,7$	$31,2 \pm 1,2$
[Valin, 2007]	$0,3 \pm 0,5$	$\mathbf{40,8 \pm 1,1}$	$0,4 \pm 0,4$	$\mathbf{64,2 \pm 1,6}$
[Schwarz et al., 2013]	$2,0 \pm 0,8$	$19,1 \pm 1,5$	$3,7 \pm 0,7$	$36,5 \pm 1,4$
[Lee et al., 2015]	$3,0 \pm 0,4$	$25,1 \pm 1,2$	$3,4 \pm 0,4$	$48,1 \pm 1,2$
+ signal d'entrée $\widehat{y}$	$3,5 \pm 0,4$	$26,9 \pm 1,2$	$3,9 \pm 0,4$	$50,0 \pm 1,3$
+ critère FSP	$4,2 \pm 0,5$	$27,8 \pm 1,2$	$4,5 \pm 0,5$	$50,8 \pm 1,3$
Méthode proposée	$\mathbf{4,5 \pm 0,5}$	$29,0 \pm 1,2$	$\mathbf{4,8 \pm 0,5}$	$51,1 \pm 1,3$

TABLEAU 3.7. – Performances moyennes (en dB) de la méthode proposée et des méthodes de l'état de l'art.

3.4.2.2. Interactions entre le filtre et le post-filtre

Tandis que les résultats ci-dessus concernent les performances moyennes sur tous les scénarios (*parole distante seule* et *parole simultanée*), nous avons besoin d'une analyse plus approfondie des performances pendant les périodes avant et après convergence du filtre d'annulation d'écho $\mathcal{H}(f)$. En particulier, l'écho résiduel $z(n,f)$ est fort avant la convergence du filtre $\mathcal{H}(f)$, ce qui représente des conditions de SER difficiles pour le post-filtre $w_s(n,f)$. Comme nous cherchons à transmettre la parole locale $s(n,f)$, nous ne considérons ici que l'analyse durant les périodes de *parole simultanée*.

Avant convergence du filtre $\mathcal{H}(f)$ La figure 3.9a présente les performances avant la convergence du filtre $\mathcal{H}(f)$. L'approche proposée obtient le meilleur SI-SDR. En particulier, c'est la seule approche qui réussit à avoir un SI-SDR positif. Elle surpasse significativement l'approche de Valin [2007] en SI-SDR et SI-SAR, avec un ERLE équivalent. Les deux approches réduisent donc l'écho résiduel $z(n,f)$ de manière importante. Toutefois, l'approche de soustraction spectrale de Valin [2007] n'estime la covariance de l'écho résiduel $|z(n,f)|^2$ qu'à partir de l'écho estimé $\widehat{y}(n,f)$, qui est mal déterminé avant la convergence du filtre $\mathcal{H}(f)$. Comme la covariance de la parole locale $|s(n,f)|^2$ est déduite par soustraction spectrale, l'approche de Valin [2007] produit une mauvaise estimation de celle-ci, compte tenu du SER qui est faible avant la convergence du filtre $\mathcal{H}(f)$. Par conséquent, le post-filtre $w_s^{\mathrm{WF}}(n,f)$ de l'approche de Valin [2007] introduit beaucoup de dégradations dans la parole locale $s(n,f)$ (SI-SAR négatif).

L'approche proposée fait aussi significativement mieux que l'approche de Schwarz et al. [2013] selon toutes les métriques. En particulier, l'approche de Schwarz et al. [2013] obtient le niveau de métriques le plus faible de toutes les approches. Son SI-SAR équivalent à l'approche de Valin [2007]. L'approche de Schwarz et al. [2013] ne supprime donc que très peu l'écho résiduel $z(n,f)$ (ERLE faible). Ceci s'explique du fait que l'écho résiduel est fort avant la convergence du filtre $\mathcal{H}(f)$, et que cette approche n'utilise en entrée que la parole distante $x(n,f)$, qui ne contient aucune information sur le volume ou le chemin d'écho $a_y(k,f',f)$. De plus, comme la covariance de la parole locale $|s(n,f)|^2$ est déduite par soustraction spectrale, et que le SER est faible avant la convergence

du filtre $\mathcal{H}(f)$, l'approche de Schwarz et al. [2013] produit une mauvaise estimation de la covariance de la parole locale $|s(n,f)|^2$. Par conséquent, le post-filtre $w_s^{\text{WF}}(n,f)$ de l'approche de Valin [2007] introduit beaucoup de dégradations dans la parole locale $s(n,f)$ (SI-SAR négatif).

Enfin, l'approche proposée dépasse significativement l'approche de Lee et al. [2015] en SI-SDR, SER et ERLE, avec un SI-SAR plus élevé, sans que la différence ne soit significative. Pour comprendre cette différence de performance, nous avons ajouté les deux variantes de l'approche de Lee et al. [2015] à la figure 3.9a. Nous remarquons que la différence est significativement plus élevée en SER et ERLE en utilisant soit le critère FSP, soit le signal d'entrée $\widehat{y}$. Avant la convergence du filtre $\mathcal{H}(f)$, le critère FSP permet de réduire plus fortement l'écho résiduel $z(n,f)$, où le niveau de SER est faible, et où la phase de l'écho résiduel $\theta_z(n,f)$ domine dans le signal $e^{\text{echo}}(n,f)$. Quant à l'écho estimé $\widehat{y}$, il contient une information sur l'état de convergence du filtre $\mathcal{H}(f)$, c'est-à-dire sur la quantité d'écho réduite par le filtre $\mathcal{H}(f)$ (voir la partie 3.2). Nous ajoutons que seule la combinaison du signal d'entrée $\widehat{y}$ et du critère FSP permet d'améliorer significativement le SI-SDR par rapport à l'approche de Lee et al. [2015].

Après convergence du filtre $\mathcal{H}(f)$ La figure 3.9b présente les performances après la convergence du filtre $\mathcal{H}(f)$. Toutes les approches obtiennent un SI-SDR positif. L'approche proposée dépasse significativement l'approche de Valin [2007] en SI-SDR et SI-SAR. Toutefois, cette approche dépasse largement toutes les autres approches en SER et ERLE. L'approche de Valin [2007] réduit donc très fortement l'écho résiduel $z(n,f)$ avec le post-filtre $w_s^{\text{WF}}(n,f)$, au point de dégrader fortement la parole locale $s(n,f)$ (le SI-SAR le plus faible de toutes les approches). Dans le scénario de *parole simultanée*, il y a donc une surestimation du coefficient utilisé pour l'estimation la covariance de l'écho résiduel $|z(n,f)|^2$ dans l'approche de Valin [2007].

L'approche proposée obtient des SI-SDR et SI-SAR légèrement plus faibles que pour l'approche de Schwarz et al. [2013] obtient le meilleur SI-SDR. Toutefois, ces différences ne sont pas significatives. L'approche proposée dépasse largement l'approche de Schwarz et al. [2013] en SER (7 dB) et en ERLE (environ 10 dB). L'approche de Schwarz et al. [2013] obtient de plus le SER et l'ERLE le plus faible de toutes les approches. Autrement dit, après convergence du filtre $\mathcal{H}(f)$, l'approche de Schwarz et al. [2013] sous-estime la covariance de l'écho résiduel $|z(n,f)|^2$, ce qui permet de moins dégrader la parole locale $s(n,f)$ avec le post-filtre $w_s^{\text{WF}}(n,f)$. Ceci vient du fait que la parole distante x utilisée en entrée de l'approche de Schwarz et al. [2013] ne contient ni l'information de volume de l'écho résiduel $z(n,f)$, ni l'information du chemin d'écho $a_y(k,f',f)$. L'approche proposée parvient à obtenir des performances en distorsion globale quasi similaire à l'approche de Schwarz et al. [2013], tout en réduisant beaucoup plus l'écho résiduel $z(n,f)$.

Enfin, l'approche proposée dépasse significativement l'approche de Lee et al. [2015] selon toutes les métriques. Pour comprendre cette différence de performance, nous avons ajouté les deux variantes de l'approche de Lee et al. [2015] à la figure 3.9a. Nous remarquons que la différence est significativement plus élevée en SI-SDR et SI-SAR en

utilisant le critère FSP. Après la convergence du filtre $\mathcal{H}(f)$, où le niveau de SER est plus faible, l'utilisation de l'information de phase dans le critère d'apprentissage du post-filtre permet donc de diminuer les dégradations de la parole locale $s(n,f)$ par le post-filtre $w_s^{\text{FSP}}(n,f)$. De plus, nous remarquons que seule la combinaison du signal d'entrée $\widehat{y}$ et du critère FSP permet d'améliorer significativement les SER par rapport à l'approche de Lee et al. [2015], tout en améliorant les SI-SDR, SI-SAR et ERLE. L'information de la phase dans le critère d'optimisation est donc complémentaire à l'information contenue dans l'écho estimé $\widehat{y}$ pour l'estimation du post-filtre $w_s(n,f)$ après convergence du filtre $\mathcal{H}(f)$.

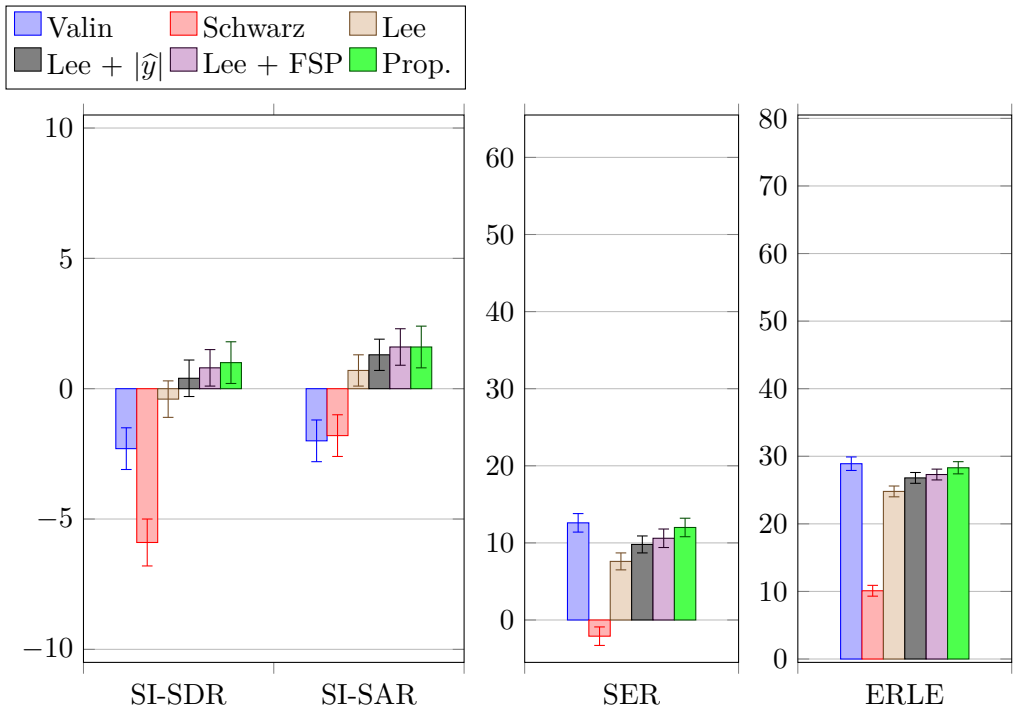
Comparaison entre les deux périodes Avant la convergence du filtre $\mathcal{H}(f)$, l'approche proposée est la meilleures approche en SI-SDR avec l'approche de Lee et al. [2015], tout en faisant partie des 2 meilleures approches réduisant l'écho résiduel $z(n,f)$ avec l'approche de Valin [2007]. Après la convergence du filtre $\mathcal{H}(f)$, l'approche proposée fait partie des 2 meilleures approches en SI-SDR avec l'approche de Schwarz et al. [2013], tout en faisant partie des 3 meilleures approches réduisant l'écho résiduel avec les approches de Valin [2007] et Lee et al. [2015]. C'est la raison pour laquelle l'approche proposée obtient le meilleur SI-SDR en moyenne sur les deux périodes, tout en réduisant fortement l'écho résiduel $z(n,f)$ (voir le tableau 4.1). Nous concluons que notre approche est la plus polyvalente pour la suppression d'écho résiduel sur les deux périodes.

3.4.3. Complexité

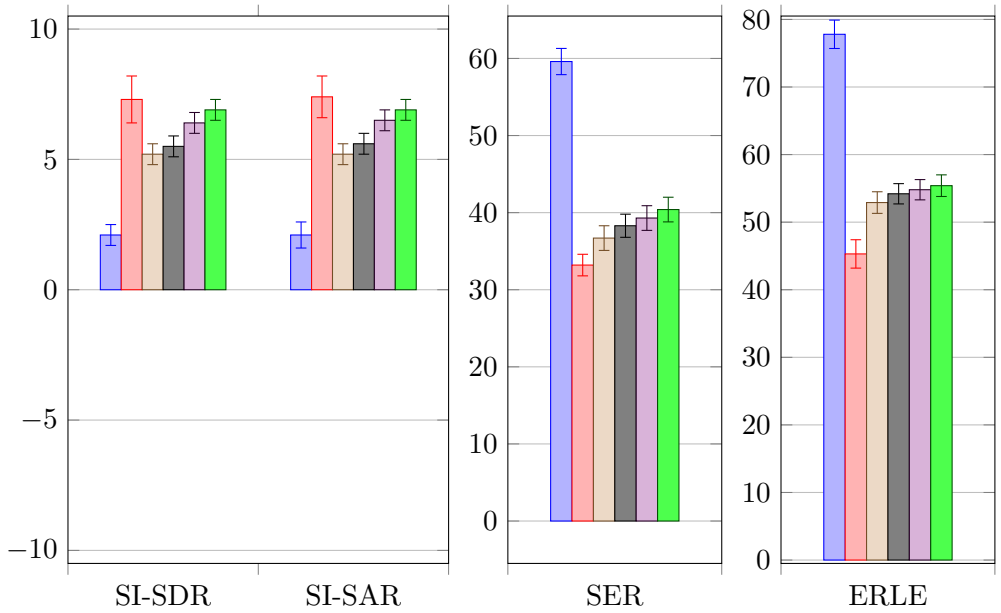
Le nombre de poids est de 321 pour l'approche de Valin [2007], 1 705 984 pour l'approche de Schwarz et al. [2013], 2 034 688 pour l'approche de Lee et al. [2015] et 2 363 392 pour l'approche proposée. Avec un processeur Intel Core i5 à 2, 70 GHz, l'annulation d'écho sur un enregistrement de 10 s prend 0, 335 s $\pm$0, 003 s . La suppression d'écho résiduel sur un enregistrement de 10 prend 0, 387 s $\pm$0, 002 s pour l'approche de Valin [2007], 0, 528 s $\pm$0, 005 s pour l'approche de Schwarz et al. [2013] , 0, 495 s $\pm$0, 004 s pour l'approche de Lee et al. [2015], et 0, 546 s $\pm$0, 004 s pour l'approche proposée. Nous concluons que toutes les approches peuvent être utilisées en temps réel.

3.5. Résumé

Dans ce chapitre, nous avons proposé une méthode de suppression d'écho résiduel qui estime directement le post-filtre $w_s(n,f)$ à l'aide d'un réseau de neurones prenant en entrée le signal après annulation d'écho $e^{\text{echo}}(n,f)$, le signal source de l'écho $x(n,f)$, et l'écho estimé $\widehat{y}(n,f)$ par le filtre d'annulation d'écho $\mathcal{H}(f)$ en amont. Cette méthode inclut l'information de phase dans le critère de détermination du post-filtre. Nous avons évalué cette méthode sur des enregistrements réels d'écho et de parole locale acquis avec un Triby dans plusieurs situations. La méthode proposée dépasse, en moyenne sur l'ensemble de ces situations, toutes les méthodes de référence. En analysant les périodes sur lesquelles la parole cible $s(n,f)$ est présente, la méthode proposée maintient un bon



(a) Avant la convergence du filtre $\mathcal{H}(f)$.



(b) Après la convergence du filtre $\mathcal{H}(f)$.

FIGURE 3.9. – Analyse des performances (en dB) durant les périodes de parole simultanée.

niveau de performances en matière de réduction de distorsion globale et de réduction d'écho, notamment lorsque le filtre $\mathcal{H}(f)$ n'a pas convergé. Cependant, l'estimation du post-filtre $w_s(n, f)$ reste séparée du filtre d'annulation d'écho $\mathcal{H}(f)$. Par conséquent, quand l'état de convergence du filtre $\mathcal{H}(f)$ est modifié par une perturbation du chemin d'écho (changement de position du Triby, locuteur local en mouvement, etc.), l'écho résiduel $z(n, f)$ change, et le post-filtre $\mathcal{H}(f)$ doit être à nouveau adapté. Afin d'améliorer la robustesse du système de réduction, le filtre $\mathcal{H}(f)$ et le post-filtre $w_s(n, f)$ doivent donc être optimisés conjointement.

4. Réduction conjointe de bruit, d'écho et de réverbération basée sur l'apprentissage profond

Ce chapitre présente notre méthode de réduction conjointe de bruit, d'écho et de réverbération en multicanal, qui repose sur l'application successive d'un filtre d'annulation d'écho, d'un filtre de déréverbération linéaire et d'un post-filtre de suppression de distorsions résiduelles. Comme les filtres interagissent entre eux, ils doivent être optimisés conjointement. Nous proposons de modéliser la parole cible et les signaux résiduels après annulation d'écho et déréverbération linéaire par des variables gaussiennes de moyenne nulle, et d'estimer conjointement leur spectre à court terme à l'aide d'un réseau de neurones. Nous développons un algorithme de montée par blocs de coordonnées pour mettre à jour les trois filtres. Nous évaluons notre méthode sur des enregistrements réels de bruit, d'écho et de réverbération acquis avec un Triby dans différentes situations. Nous comparons la méthode proposée à une combinaison en cascade des approches de réduction individuelle, et une autre méthode de réduction conjointe qui ne modélise pas les spectres à court terme de la parole cible et des signaux résiduels.

4.1. Formulation du problème

Nous rappelons ici brièvement le problème de la réduction conjointe de bruit, d'écho et de réverbération décrit dans la partie 2.3.3. Dans des scénarios réels, les trois types de distorsion peuvent être présents simultanément comme l'illustre la figure 4.1. Dans le domaine temps-fréquence, le mélange $\mathbf{d}(n,f)$ est la somme de la parole locale $\mathbf{s}(n,f)$, du signal de bruit $\mathbf{b}(n,f)$ et de l'écho $\mathbf{y}(n,f)$:

$$\mathbf{d}(n,f) = \mathbf{s}(n,f) + \mathbf{b}(n,f) + \mathbf{y}(n,f) \tag{4.1}$$

$$= \mathbf{s}_{\mathrm{e}}(n,f) + \mathbf{s}_{\mathrm{l}}(n,f) + \mathbf{b}(n,f) + \mathbf{y}(n,f). \tag{4.2}$$

Le but est d'extraire la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$ du mélange $\mathbf{d}(n,f)$. Pour cela, les méthodes de réduction individuelle de bruit, d'écho et de réverbération, présentées dans la partie 2.2, peuvent être combinées. La sélection des méthodes en tant que références pour la thèse est évoquée dans la partie 2.2.4. Pour la réduction d'écho, nous avons retenu le système combinant un filtre d'annulation d'écho avec un post-filtre de suppression d'écho résiduel. Pour la déréverbération, nous avons retenu le système combinant un filtre de déréverbération linéaire avec un post-filtre de suppression de réverbération résiduelle. Pour la réduction de bruit en multicanal, nous avons retenu les méthodes de

séparation de sources basées sur l'apprentissage profond, qui estiment les DSPs de la parole locale et du bruit avec un DNN et utilisent un modèle de rang plein pour les MCSs. Cependant, comme les filtres qui composent le système interagissent entre eux, ils doivent être optimisés conjointement, afin d'améliorer la robustesse du système de réduction.

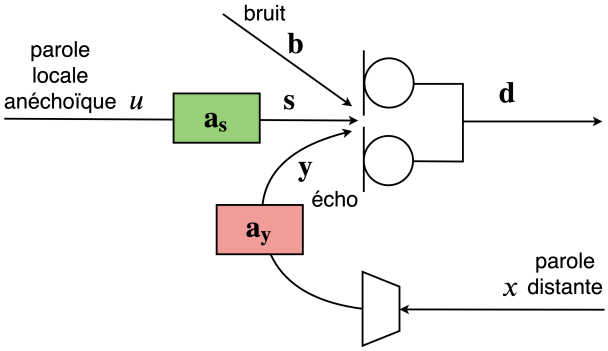


FIGURE 4.1. – Problème de l'écho acoustique, du bruit et de la réverbération.

Pour résoudre ce problème, Togami et Kawaguchi [2014] ont proposé une approche conjointe, illustrée sur la figure 4.2, qui combine un filtre d'annulation d'écho $\mathcal{H}(f)$, un filtre de déréverbération $\mathcal{G}(f)$ et un post-filtre $\mathbf{W}_{s_e}(n,f)$ (voir la partie 2.3.3.2). Ils appliquent une annulation d'écho à l'aide du filtre $\mathcal{H}(f)$. En parallèle, ils appliquent une déréverbération linéaire sur le mélange $\mathbf{d}(n,f)$ à l'aide du filtre $\mathcal{G}(f)$. Le signal $\mathbf{r}(n,f)$ qui résulte de ces deux filtragess s'exprime de la manière suivante :

$$\mathbf{r}(n,f) = \mathbf{d}(n,f) - \underbrace{\sum_{k=0}^{K-1} \mathbf{h}(k,f)x(n-k,f)}_{=\widehat{\mathbf{y}}(n,f)} - \underbrace{\sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\mathbf{d}(n-l,f)}_{=\widehat{\mathbf{d}}_{\mathrm{l}}(n,f)}. \tag{4.3}$$

Les signaux $\mathbf{z}_{\mathrm{e}}(n,f)$, $\widetilde{\mathbf{b}}_{\mathrm{r}}(n,f)$ et $\mathbf{b}_{\mathrm{r}}(n,f)$ sont définis de la manière suivante :

$$\mathbf{z}_{\mathrm{e}}(n,f) = \mathbf{y}_{\mathrm{e}}(n,f) - \widehat{\mathbf{y}}(n,f), \tag{4.4}$$
$$\widetilde{\mathbf{b}}_{\mathrm{r}}(n,f) = \mathbf{s}_{\mathrm{l}}(n,f) - \widehat{\mathbf{d}}_{\mathrm{l},s}(n,f) + \mathbf{y}_{\mathrm{l}}(n,f) - \widehat{\mathbf{d}}_{\mathrm{l},y}(n,f), \tag{4.5}$$
$$\mathbf{b}_{\mathrm{r}}(n,f) = \mathbf{b}(n,f) - \widehat{\mathbf{d}}_{\mathrm{l},b}(n,f), \tag{4.6}$$

où le signal $\mathbf{y}_{\mathrm{e}}(n,f)$ désigne la composante précoce de l'écho $\mathbf{y}(n,f)$, et les signaux

$\widehat{\mathbf{d}}_{\mathrm{l},s}(n,f)$, $\widehat{\mathbf{d}}_{\mathrm{l},y}(n,f)$ et $\widehat{\mathbf{d}}_{\mathrm{l},b}(n,f)$ sont les composantes latentes de $\widehat{\mathbf{d}}_{\mathrm{l}}(n,f)$ définies comme

$$\widehat{\mathbf{d}}_{\mathrm{l}}(n,f) = \sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\Big(\mathbf{s}(n-l,f) + \mathbf{y}(n-l,f) + \mathbf{b}(n-l,f)\Big) \tag{4.7}$$

$$= \underbrace{\sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\mathbf{s}(n-l,f)}_{=\widehat{\mathbf{d}}_{\mathrm{l},s}(n,f)} + \underbrace{\sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\mathbf{y}(n-l,f)}_{=\widehat{\mathbf{d}}_{\mathrm{l},y}(n,f)} + \underbrace{\sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\mathbf{b}(n-l,f)}_{=\widehat{\mathbf{d}}_{\mathrm{l},b}(n,f)}. \tag{4.8}$$

À titre de rappel, nous désignons $\mathbf{b}_{\mathrm{r}}(n,f)$ comme le signal de bruit *déréverbéré*, où le terme *déréverbéré* signifie « après application du filtre $\mathcal{G}(f)$ ».

À cause des raisons évoquées dans le chapitre 2, des signaux résiduels non désirés subsistent dans le signal $\mathbf{r}(n,f)$ et s'expriment de la manière suivante :

$$\mathbf{r}(n,f) - \mathbf{s}_{\mathrm{e}}(n,f) = \mathbf{z}_{\mathrm{e}}(n,f) + \widetilde{\mathbf{b}}_{\mathrm{r}}(n,f) + \mathbf{b}_{\mathrm{r}}(n,f), \tag{4.9}$$

où les signaux $\mathbf{z}_{\mathrm{e}}(n,f)$ et $\widetilde{\mathbf{b}}_{\mathrm{r}}(n,f)$ sont définis dans (2.84)–(2.85), et $\mathbf{b}_{\mathrm{r}}(n,f)$ est défini comme dans (2.72). Pour extraire la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$, Togami et Kawaguchi [2014] appliquent ensuite le post-filtre $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$ sur le signal $\mathbf{r}(n,f)$:

$$\widehat{\mathbf{s}}_{\mathrm{e}}(n,f) = \mathbf{W}_{s_{\mathrm{e}}}(n,f)\mathbf{r}(n,f). \tag{4.10}$$

Pour estimer les filtres $\mathcal{H}(f)$, $\mathcal{G}(f)$ et $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$, Togami et Kawaguchi [2014] modélisent les signaux $\mathbf{s}_{\mathrm{e}}(n,f)$ et $\mathbf{b}_{\mathrm{r}}(n,f)$ par des variables gaussiennes de moyenne nulle comme dans (2.27), et les signaux $\mathbf{z}_{\mathrm{e}}(n,f)$ et $\widetilde{\mathbf{b}}_{\mathrm{r}}(n,f)$ par des variables gaussiennes de moyenne non nulle. Les paramètres de ce modèle sont alors optimisés conjointement selon le critère du MV en utilisant l'algorithme EM.

Toutefois, cette approche souffre de quatre limites. Premièrement, aucune contrainte n'est imposée sur les DSPs $v_{s_{\mathrm{e}}}(n,f)$ et $v_{s_{\mathrm{r}}}(n,f)$. Deuxièmement, le signal $\widetilde{\mathbf{b}}_{\mathrm{r}}(n,f)$ défini dans (4.5) regroupe la composante $\mathbf{s}_{\mathrm{l}}(n,f) - \widehat{\mathbf{d}}_{\mathrm{l},s}(n,f)$ liée à la parole locale et la composante $\mathbf{y}_{\mathrm{l}}(n,f) - \widehat{\mathbf{d}}_{\mathrm{l},y}(n,f)$ liée à l'écho. Par conséquent, ces deux composantes vont partager les mêmes paramètres spectraux et spatiaux dans ce modèle, alors qu'en réalité ce n'est pas le cas puisqu'elles correspondent à deux sources différentes. Ces deux limites conduisent à une mauvaise estimation des filtres $\mathcal{H}(f)$, $\mathcal{G}(f)$ et $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$. Troisièmement, les filtres $\mathcal{H}(f)$ et $\mathcal{G}(f)$ opèrent en parallèle sur le mélange $\mathbf{d}(n,f)$ (voir la figure 4.2). Par conséquent, leurs composantes respectives $\widehat{\mathbf{y}}(n,f)$ et $\widehat{\mathbf{d}}_{\mathrm{l},y}$ sont susceptibles d'interférer entre elles, car elles sont soustraites à des composantes de la même source (l'écho $\mathbf{y}(n,f)$) dans (4.4) et (4.5), respectivement. Enfin, puisque l'écho $\mathbf{y}(n,f)$ est souvent bien plus fort que la parole locale $\mathbf{s}(n,f)$ et le bruit $\mathbf{b}(n,f)$, le filtre $\mathcal{G}(f)$ réduit surtout l'écho résiduel tardif $\mathbf{y}_{\mathrm{l}}(n,f)$ plutôt que la réverbération tardive de la parole locale $\mathbf{s}_{\mathrm{l}}(n,f)$.

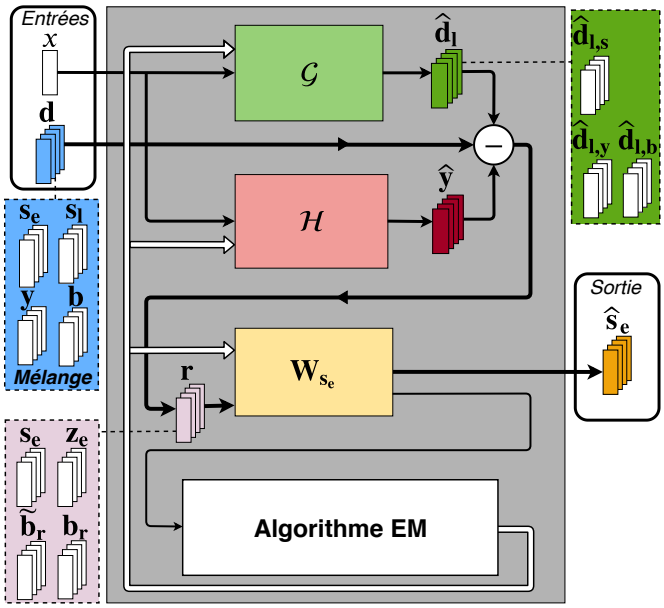


FIGURE 4.2. – Approche proposée par Togami et Kawaguchi [2014]. Les flèches en gras désignent les étapes de filtrage. Les lignes en pointillés désignent les composantes latentes des signaux. Les flèches minces désignent les signaux utilisés par l'algorithme EM. Les flèches blanches désignent les mises à jour des filtres.

4.2. Solution proposée

4.2.1. Modèle

La figure 5.2 illustre l'approche proposée. Tout d'abord, nous appliquons une annulation d'écho sur le mélange $\mathbf{d}(n,f)$ à l'aide du filtre $\mathcal{H}(f)$:

$$\mathbf{e}(n,f) = \mathbf{d}(n,f) - \underbrace{\sum_{k=0}^{K} \mathbf{h}(k,f)x(n-k,f)}_{=\widehat{\mathbf{y}}(n,f)}. \tag{4.11}$$

Le signal $\mathbf{e}(n,f)$ contient la parole locale $\mathbf{s}(n,f)$, le signal de bruit $\mathbf{b}(n,f)$ et l'écho résiduel $\mathbf{z}(n,f)$. Contrairement à Togami et Kawaguchi [2014], nous appliquons ensuite une déréverbération linéaire à l'aide du filtre $\mathcal{G}(f)$ non pas sur le mélange $\mathbf{d}(n,f)$ mais sur le signal $\mathbf{e}(n,f)$. À notre connaissance, c'est la première fois qu'un filtre de déréverbération $\mathcal{G}(f)$ est appliqué après un filtre d'annulation d'écho $\mathcal{H}(f)$ dans le contexte de la réduction conjointe de bruit, d'écho et de déréverbération. Le signal $\mathbf{r}(n,f)$ qui en résulte s'exprime de la manière suivante :

$$\mathbf{r}(n,f) = \mathbf{e}(n,f) - \underbrace{\sum_{l=\Delta}^{\Delta+L+1} \mathbf{G}(l,f)\mathbf{e}(n-l,f)}_{=\mathbf{e}_{\mathrm{l}}(n,f)}. \tag{4.12}$$

Les filtres $\mathcal{H}(f)$ et $\mathcal{G}(f)$ sont causaux. Pour $n < 0$, nous supposons que les signaux observés $\mathbf{d}(n, f)$ et $x(n, f)$ sont nuls. Il convient de remarquer que le filtre de déréverbération $\mathcal{G}(f)$ réalise une plus grande réduction de la réverbération tardive $\mathbf{s}_{\mathrm{l}}(n, f)$ que dans l'approche de Togami et Kawaguchi [2014]. En effet, le filtre $\mathcal{G}(f)$ opère ici sur le signal $\mathbf{e}(n, f)$, où l'écho $\mathbf{y}(n, f)$ a été réduit par le filtre $\mathcal{H}(f)$ dans (4.11), alors que dans l'approche de Togami et Kawaguchi [2014], le filtre $\mathcal{G}(f)$ opère sur le mélange $\mathbf{d}(n, f)$, où l'écho $\mathbf{y}(n, f)$ n'a pas été réduit.

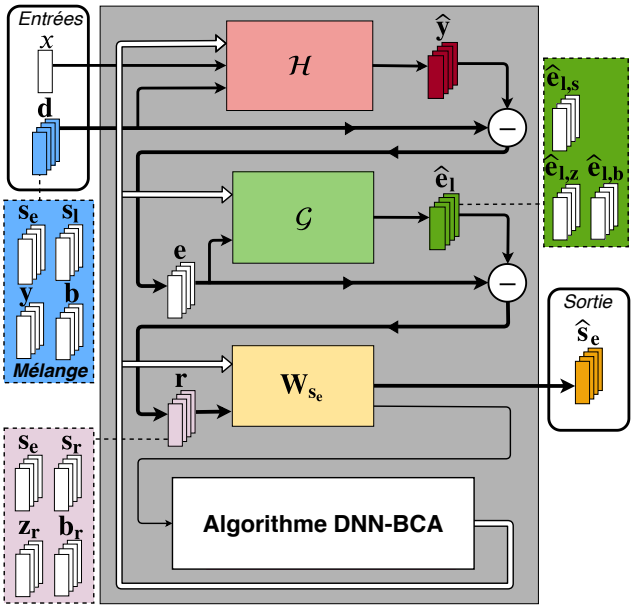


FIGURE 4.3. – Approche proposée. Les flèches et lignes ont la même signification que dans la figure 4.2.

À cause des raisons évoquées dans les parties 2.2 et 2.3, des signaux résiduels non désirés subsistent dans le signal $\mathbf{r}(n, f)$ et s'expriment de la manière suivante :

$$\mathbf{r}(n, f) - \mathbf{s}_{\mathrm{e}}(n, f) = \mathbf{s}_{\mathrm{r}}(n, f) + \mathbf{z}_{\mathrm{r}}(n, f) + \mathbf{b}_{\mathrm{r}}(n, f), \tag{4.13}$$

où $\mathbf{s}_{\mathrm{r}}(n, f)$ est la réverbération résiduelle de la parole locale, $\mathbf{z}_{\mathrm{r}}(n, f)$ l'écho résiduel *déréverbéré*, et $\mathbf{b}_{\mathrm{r}}(n, f)$ le bruit *déréverbéré*. À titre de rappel, le terme *déréverbéré* signifie « après application du filtre $\mathcal{G}(f)$ ». Les signaux $\mathbf{s}_{\mathrm{r}}(n, f)$, $\mathbf{z}_{\mathrm{r}}(n, f)$ et $\mathbf{b}_{\mathrm{r}}(n, f)$ sont définis de la manière suivante :

$$\mathbf{s}_{\mathrm{r}}(n, f) = \mathbf{s}_{\mathrm{l}}(n, f) - \widehat{\mathbf{e}}_{\mathrm{l},s}(n, f), \tag{4.14}$$
$$\mathbf{z}_{\mathrm{r}}(n, f) = \mathbf{z}(n, f) - \widehat{\mathbf{e}}_{\mathrm{l},z}(n, f), \tag{4.15}$$
$$\mathbf{b}_{\mathrm{r}}(n, f) = \mathbf{b}(n, f) - \widehat{\mathbf{e}}_{\mathrm{l},b}(n, f), \tag{4.16}$$

où les signaux $\widehat{\mathbf{e}}_{\mathrm{l},s}(n, f)$, $\widehat{\mathbf{e}}_{\mathrm{l},z}(n, f)$ et $\widehat{\mathbf{e}}_{\mathrm{l},b}(n, f)$ sont les composantes latentes de la réver-

bération tardive estimée $\widehat{\mathbf{e}}_{\mathrm{l}}(n,f)$ qui résultent de (4.12) :

$$\widehat{\mathbf{e}}_{\mathrm{l}}(n,f) = \sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\Big(\mathbf{s}(n-l,f)+\mathbf{z}(n-l,f)+\mathbf{b}(n-l,f)\Big) \tag{4.17}$$

$$= \underbrace{\sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\mathbf{s}(n-l,f)}_{=\widehat{\mathbf{e}}_{\mathrm{l},s}(n,f)} + \underbrace{\sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\mathbf{z}(n-l,f)}_{=\widehat{\mathbf{e}}_{\mathrm{l},z}(n,f)} + \underbrace{\sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\mathbf{b}(n-l,f)}_{=\widehat{\mathbf{e}}_{\mathrm{l},b}(n,f)}. \tag{4.18}$$

Pour extraire la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$ du signal $\mathbf{r}(n,f)$, nous appliquons un post-filtre court $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$ sur le signal $\mathbf{r}(n,f)$:

$$\widehat{\mathbf{s}}_{\mathrm{e}}(n,f) = \mathbf{W}_{s_{\mathrm{e}}}(n,f)\mathbf{r}(n,f). \tag{4.19}$$

Inspirés par les méthodes de filtrage inverse en déréverbération (voir la partie 2.2.3.2), nous estimons les filtres $\mathcal{H}(f)$, $\mathcal{G}(f)$ et $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$ en optimisant le critère ML. Pour cela, nous modélisons la parole cible $\mathbf{s}_{\mathrm{e}}(n,f)$ et les trois signaux résiduels $\mathbf{s}_{\mathrm{r}}(n,f)$, $\mathbf{z}_{\mathrm{r}}(n,f)$ et $\mathbf{b}_{\mathrm{r}}(n,f)$ par des variables gaussiennes de moyenne nulle avec le modèle gaussien local (voir la partie 2.2.1.4). Nous utilisons la notation générale $\mathbf{c}(n,f)$ pour désigner l'un des quatre signaux $\mathbf{s}_{\mathrm{e}}(n,f)$, $\mathbf{s}_{\mathrm{r}}(n,f)$, $\mathbf{z}_{\mathrm{r}}(n,f)$ et $\mathbf{b}_{\mathrm{r}}(n,f)$. Chacun de ces signaux est modélisé comme

$$\mathbf{c}(n,f) \sim \mathcal{N}(\mathbf{0}, v_c(n,f)\mathbf{R}_c(f)). \tag{4.20}$$

Nous estimons le post-filtre $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$ à partir des covariances $\mathbf{\Sigma}_c(n,f) = v_c(n,f)\mathbf{R}_c(f)$ en utilisant le filtre de Wiener. Celui-ci est formulé de la manière suivante pour le signal $\mathbf{c}(n,f)$:

$$\mathbf{W}_c(n,f) = v_c(n,f)\mathbf{R}_c(f)\left(\sum_{c'\in\,\mathcal{C}} v_{c'}(n,f)\mathbf{R}_{c'}(f)\right)^{-1}, \tag{4.21}$$

où $\mathcal{C} = \{\mathbf{s}_{\mathrm{e}}, \mathbf{s}_{\mathrm{r}}, \mathbf{z}_{\mathrm{r}}, \mathbf{b}_{\mathrm{r}}\}$ désigne l'ensemble des quatre composantes du signal $\mathbf{r}(n,f)$ dans (4.13). Le post-filtre $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$ est un cas spécifique de (4.21) où $\mathbf{c}(n,f) = \mathbf{s}_{\mathrm{e}}(n,f)$.

4.2.2. Vraisemblance

Pour estimer les paramètres de ce modèle, nous devons d'abord exprimer la vraisemblance de la suite des signaux observés $\mathcal{O} = \{\mathbf{d}(n,f), x(n,f)\}_{n,f}$, où la notation $\{\cdot\}_{n,f}$ désigne l'ensemble des paramètres sur toutes les trames $n \in \{0,\ldots,N-1\}$ et sur toutes les bandes de fréquence $f \in \{0,\ldots,F-1\}$. D'après (4.11), (4.12), (4.13) et (4.20), la

log-vraisemblance de la suite $\mathcal{O}$ correspond à

$$\mathcal{L}(\mathcal{O};\Theta_H,\Theta_G,\Theta_c)=\sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\log p\Big(\mathbf{d}(n,f)\Big|\mathbf{d}(n-1,f),\ldots,\mathbf{d}(0,f),x(n,f),\ldots,x(0,f)\Big), \tag{4.22}$$

$$=\sum_{f=0}^{F-1}\sum_{n=0}^{N-1}\log\mathcal{N}_{\mathbb{C}}\Big(\mathbf{d}(n,f);\boldsymbol{\mu}_{\mathbf{d}}(n,f),\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f)\Big), \tag{4.23}$$

où

$$\boldsymbol{\mu}_{\mathbf{d}}(n,f)=\sum_{k=0}^{K-1}\mathbf{h}(k,f)x(n-k,f)+\sum_{l=\Delta}^{\Delta+L-1}\mathbf{G}(l,f)\mathbf{e}(n-l,f), \tag{4.24}$$

$$\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f)=\sum_{c'\in\,\mathcal{C}}v_{c'}(n,f)\mathbf{R}_{c'}(f), \tag{4.25}$$

et $\Theta_H=\{\mathcal{H}(f)\}_f$, $\Theta_G=\{\mathcal{G}(f)\}_f$ et $\Theta_c=\{v_{c'}(n,f),\mathbf{R}_{c'}(f)\}_{c',n,f}$ sont les paramètres à estimer. Cette formulation est valide car $\mathbf{e}(n-l,f)$ est une combinaison linéaire du mélange $\mathbf{d}(n-l,f)$ et de la parole distante $x(n-k,f)$ (voir (4.11)). Comme ce problème d'optimisation n'a pas de solution analytique, nous avons besoin d'estimer les paramètres Θ_H, Θ_G et Θ_c avec une procédure itérative.

4.2.3. Algorithme itératif d'optimisation

Nous proposons un algorithme de montée par blocs de coordonnées (BCA, *block-coordinate ascent*) [Bertsekas, 1999, Chapitre 2], qui se base sur l'apprentissage profond pour l'optimisation de la vraisemblance. Nous désignons cet algorithme par le terme de DNN-BCA. La figure 4.4 illustre le processus d'optimisation. Chaque itération i est constituée de trois étapes de maximisation :

$$\widehat{\Theta}_H^{(i)}\leftarrow\underset{\Theta_H}{\operatorname{argmax}}\,\mathcal{L}\left(\mathcal{O};\Theta_H,\widehat{\Theta}_G^{(i-1)},\widehat{\Theta}_c^{(i-1)}\right), \tag{4.26}$$

$$\widehat{\Theta}_G^{(i)}\leftarrow\underset{\Theta_G}{\operatorname{argmax}}\,\mathcal{L}\left(\mathcal{O};\widehat{\Theta}_H^{(i)},\Theta_G,\widehat{\Theta}_c^{(i-1)}\right), \tag{4.27}$$

$$\widehat{\Theta}_c^{(i)}\leftarrow\underset{\Theta_c}{\operatorname{argmax}}\,\mathcal{L}\left(\mathcal{O};\widehat{\Theta}_H^{(i)},\widehat{\Theta}_G^{(i)},\Theta_c\right), \tag{4.28}$$

où l'exposant $(\cdot)^{(i)}$ indique la valeur des paramètres à l'itération i. Les solutions de (4.26) et (4.27) sont analytiques. Comme il n'y a pas de solution analytique pour (4.28), nous proposons d'utiliser une version modifiée de l'algorithme DNN-EM de Nugraha et al. [2016a].Il convient de noter qu'il est aussi possible d'optimiser les paramètres Θ_H, Θ_G et Θ_c avec l'algorithme EM en ajoutant un terme de nuisance à (4.13) [Ozerov et Fevotte, 2010]. Toutefois, cette approche serait moins efficace pour l'estimation des paramètres des filtres linéaires Θ_H et Θ_G. Dans les sous-parties suivantes, nous décrivons l'initialisation et les règles de mise à jour pour les étapes (4.26)–(4.28) à l'itération i de l'algorithme proposé. Le calcul de ces règles de mise à jour est détaillée dans l'annexe A.

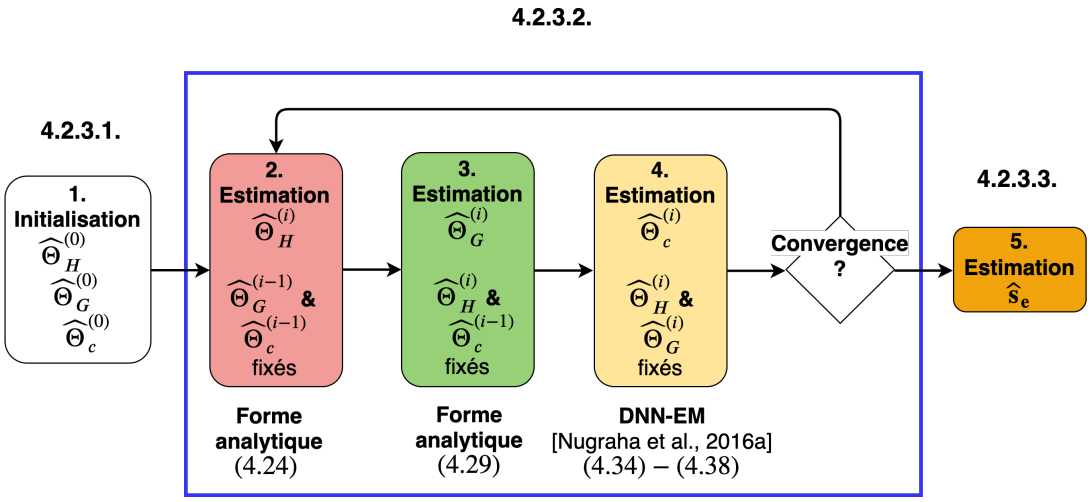


FIGURE 4.4. – Algorithme DNN-BCA proposé.

4.2.3.1. Initialisation

D'après (4.26), l'optimisation des trois paramètres Θ_H, Θ_G et Θ_c nécessite l'initialisation de Θ_G et Θ_c. Pour $\widehat{\Theta}_G^{(0)}$, nous initialisons le filtre $\mathcal{G}(f)$ à la valeur $\mathcal{G}^{(0)}(f)$, qui peut être définie en utilisant, par exemple, la méthode WPE [Yoshioka et Nakatani, 2012] (voir la partie 2.2.3.2). Pour $\widehat{\Theta}_c^{(0)}$, nous initialisons les MCS $\mathbf{R}_c(f)$ de la parole cible et des signaux résiduels $\mathbf{s}_{\mathrm{e}}(n,f)$, $\mathbf{s}_{\mathrm{r}}(n,f)$, $\mathbf{z}_{\mathrm{r}}(n,f)$ et $\mathbf{b}_{\mathrm{r}}(n,f)$ à la valeur $\mathbf{R}_c^{(0)}(f)$, qui peut être définie, par exemple, à la matrice identité $\mathbf{I}_{\mathrm{M}}$ ou en utilisant des méthodes de localisation de sources [Vincent et al., 2018, Chapitre 4]. Nous initialisons les DSPs $v_c(n,f)$ sont initialisées conjointement aux valeurs $v_c^{(0)}(n,f)$ à l'aide d'un DNN désigné par DNN_0, que nous avons pré-entraîné. Nous décrivons les entrées, les cibles et l'architecture de DNN_0 dans la partie 4.2.4 ci-après. En particulier, les entrées de DNN_0 sont calculées à partir des valeurs de $\widehat{\Theta}_H^{(0)}$ et $\widehat{\Theta}_G^{(0)}$ à l'initialisation. Pour $\widehat{\Theta}_H^{(0)}$, nous initialisons le filtre $\mathcal{H}(f)$ à la valeur $\mathcal{H}^{(0)}(f)$, qui peut être définie en utilisant, par exemple, la méthode de Valin [2007], comme à la partie 3.3.4.

4.2.3.2. Mise à jour des paramètres

Nous décrivons la mise à jour des paramètres Θ_H, Θ_G et Θ_c à l'itération i dans les paragraphes suivants. Il convient de rappeler qu'à chaque nouvelle itération i, nous utilisons les paramètres estimés $\widehat{\Theta}_G^{(i-1)}$ et $\widehat{\Theta}_c^{(i-1)}$ de l'itération précédente $i-1$.

Paramètres du filtre d'annulation d'écho Θ_H Le filtre d'annulation d'écho $\mathcal{H}^{(i)}(f)$ est mis à jour de la manière suivante :

$$\underline{\mathbf{h}}^{(i)}(f) = \mathbf{P}^{(i)}(f)^{-1}\mathbf{p}^{(i)}(f), \tag{4.29}$$

où le terme $\underline{\mathbf{h}}^{(i)}(f) = \left[\mathbf{h}^{(i)}(0,f)^T \dots \mathbf{h}^{(i)}(K-1,f)^T\right]^T \in \mathbb{C}^{MK\times 1}$ est une version vectorisée de $\mathcal{H}^{(i)}(f)$. La notation soulignée dans $\underline{\mathbf{h}}^{(i)}(f)$ désigne la concaténation des K trames $\mathbf{h}^{(i)}(k,f)$ du filtre $\mathcal{H}^{(i)}(f)$. Les termes $\mathbf{P}^{(i)}(f) \in \mathbb{C}^{MK\times MK}$ et $\mathbf{P}^{(i)}(f) \in \mathbb{C}^{MK\times 1}$ sont calculés de la manière suivante :

$$\mathbf{P}^{(i)}(f) = \sum_{n=0}^{N-1} \underline{\mathbf{X}}_{\mathrm{r}}^{(i-1)}(n,f)^H \mathbf{\Sigma}_{\mathbf{dd}}^{(i-1)}(n,f)^{-1} \underline{\mathbf{X}}_{\mathrm{r}}^{(i-1)}(n,f) \tag{4.30}$$

$$\mathbf{p}^{(i)}(f) = \sum_{n=0}^{N-1} \underline{\mathbf{X}}_{\mathrm{r}}^{(i-1)}(n,f)^H \mathbf{\Sigma}_{\mathbf{dd}}^{(i-1)}(n,f)^{-1} \mathbf{r}_d^{(i-1)}(n,f). \tag{4.31}$$

Le terme $\underline{\mathbf{X}}_{\mathrm{r}}^{(i-1)}(n,f) = \left[\mathbf{X}_{\mathrm{r}}^{(i-1)}(n,f) \dots \mathbf{X}_{\mathrm{r}}^{(i-1)}(n-K+1,f)\right] \in \mathbb{C}^{M\times MK}$contient les K trames $\mathbf{X}_{\mathrm{r}}^{(i-1)}(n-k,f) \in \mathbb{C}^{M\times M}$. La notation soulignée dans $\underline{\mathbf{X}}_{\mathrm{r}}^{(i-1)}(n,f)$ désigne la concaténation des K trames $\mathbf{X}_{\mathrm{r}}^{(i-1)}(n-k,f)$. Les K trames $\mathbf{X}_{\mathrm{r}}^{(i-1)}(n-k,f)$ sont des versions *déréverbérées* de la parole distante $x(n-k,f)$ obtenues de la manière suivante :

$$\mathbf{X}_{\mathrm{r}}^{(i-1)}(n-k,f) = x(n-k,f)\mathbf{I}_M - \sum_{l=\Delta}^{\Delta+L-1} x(n-k-l,f)\mathbf{G}^{(i-1)}(l,f). \tag{4.32}$$

Le terme $\mathbf{r}_d^{(i-1)}(n,f)$ dans (4.31) est une version *déréverbérée* du mélange $\mathbf{d}(n,f)$ obtenue comme suit :

$$\mathbf{r}_d^{(i-1)}(n,f) = \mathbf{d}(n,f) - \sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}^{(i-1)}(l,f)\mathbf{d}(n-l,f). \tag{4.33}$$

La mise à jour du filtre d'annulation d'écho $\mathcal{H}^{(i)}(f)$ est donc influencée par le filtre de déréverbération $\mathcal{G}^{(i-1)}(f)$ de l'itération précédente $i-1$ à travers les termes $\underline{\mathbf{X}}_{\mathrm{r}}^{(i-1)}(n,f)$ et $\mathbf{r}_d^{(i-1)}(n,f)$. Cette mise à jour permet d'empêcher le filtre d'annulation d'écho $\mathcal{H}^{(i)}(f)$ de réduire la composante de l'écho $\mathbf{y}(n,f)$ qui a déjà été réduite par le filtre de déréverbération $\mathcal{G}^{(i-1)}(f)$.

La mise à jour du filtre d'annulation d'écho $\mathcal{H}^{(i)}(f)$ dépend aussi des DSPs $v_c^{(i-1)}(n,f)$ et des MCSs $\mathbf{R}_c^{(i-1)}(f)$ à travers le terme $\mathbf{\Sigma}_{\mathbf{dd}}^{(i-1)}(n,f)$ défini dans (4.25). Puisque le post-filtre $\mathbf{W}_c^{(i-1)}(n,f)$ est utilisé à la fois dans la mise à jour des DSPs $v_c^{(i-1)}(n,f)$ (voir la partie 4.2.4.2) et des MCSs $\mathbf{R}_c^{(i-1)}(f)$ (voir le paragraphe sur les paramètres Θ_c ci-dessous), la mise à jour du filtre d'annulation d'écho $\mathcal{H}^{(i)}(f)$ est aussi influencée par le post-filtre $\mathbf{W}_c^{(i-1)}(n,f)$. L'obtention de cette mise à jour est détaillée dans l'annexe A. Le signal $\mathbf{e}^{(i)}(n,f)$ est ensuite calculé comme dans (4.11).

Paramètres du filtre de déréverbération Θ_G De la même manière qu'en déréverbération par filtrage inverse (voir la partie 2.2.3.2), le filtre $\mathcal{G}^{(i)}(f)$ est mis à jour de la manière suivante :

$$\underline{\mathbf{g}}^{(i)}(f) = \mathbf{Q}^{(i)}(f)^{-1}\mathbf{q}^{(i)}(f), \tag{4.34}$$

où le terme $\underline{\mathbf{g}}^{(i)}(f) = \left[\mathbf{g}_1^{(i)}(\Delta, f)^T \dots \mathbf{g}_M^{(i)}(\Delta, f)^T \dots\dots \mathbf{g}_1^{(i)}(\Delta + L - 1, f)^T \dots \mathbf{g}_M^{(i)}(\Delta + L - 1, f)^T\right]^T \in \mathbb{C}^{M^2L \times 1}$ est une version vectorisée de $\mathcal{G}^{(i)}(f)$. La notation soulignée dans $\underline{\mathbf{g}}^{(i)}(f)$ désigne la concaténation des $M \times L$ composantes $\mathbf{g}_m^{(i)}(l, f)$. Les termes $\mathbf{Q}^{(i)}(f) \in \mathbb{C}^{M^2L \times M^2L}$ et $\mathbf{q}^{(i)}(f) \in \mathbb{C}^{M^2L \times 1}$ sont calculés de la manière suivante :

$$\mathbf{Q}^{(i)}(f) = \sum_{n=0}^{N-1} \underline{\mathbf{E}}^{(i)}(n, f)^H \mathbf{\Sigma}_{\mathbf{dd}}^{(i-1)}(n, f)^{-1} \underline{\mathbf{E}}^{(i)}(n, f), \tag{4.35}$$

$$\mathbf{q}^{(i)}(f) = \sum_{n=0}^{N-1} \underline{\mathbf{E}}^{(i)}(n, f)^H \mathbf{\Sigma}_{\mathbf{dd}}^{(i-1)}(n, f)^{-1} \mathbf{e}^{(i)}(n, f), \tag{4.36}$$

Le terme $\underline{\mathbf{E}}^{(i)}(n, f) = \left[\mathbf{E}^{(i)}(n - \Delta, f) \dots \mathbf{E}^{(i-1)}(n - \Delta - L + 1, f)\right] \in \mathbb{C}^{M \times M^2L}$ contient les L trames $\mathbf{E}^{(i)}(n - l, f) \in \mathbb{C}^{M \times M^2}$. La notation soulignée dans $\underline{\mathbf{E}}^{(i)}(n, f)$ désigne la concaténation des L trames $\mathbf{E}^{(i)}(n - l, f)$. Les L trames $\mathbf{E}^{(i)}(n - l, f)$ sont des versions matricielles du signal $\mathbf{e}^{(i)}(n - l, f)$ obtenues de la manière suivante :

$$\mathbf{E}^{(i)}(n - l, f) = \mathbf{I}_M \otimes \mathbf{e}^{(i)}(n - l, f)^T, \tag{4.37}$$

où $\otimes$ est le produit de Kronecker et $\mathbf{e}^{(i)}(n - l, f)$ est le signal obtenu par application du filtre d'annulation d'écho $\mathcal{H}^{(i)}(f)$ comme dans (4.11). La mise à jour du filtre de déréverbération $\mathcal{G}^{(i)}(f)$ est donc influencée par le filtre d'annulation d'écho $\mathcal{H}^{(i)}(f)$ à travers les termes $\mathbf{e}^{(i)}(n, f)$ et $\underline{\mathbf{E}}^{(i)}(n, f)$. De même que pour le filtre d'annulation d'écho $\mathcal{H}^{(i)}(f)$, la mise à jour du filtre de déréverbération $\mathcal{G}^{(i)}(f)$ est aussi influencée par le post-filtre $\mathbf{W}_c^{(i-1)}(n, f)$ à travers les DSPs $v_c^{(i-1)}(n, f)$ et les MCSs $\mathbf{R}_c^{(i-1)}(f)$ utilisées dans le terme $\mathbf{\Sigma}_{\mathbf{dd}}^{(i-1)}(n, f)$ défini dans (4.25). L'obtention de cette mise à jour est détaillée dans l'annexe A. Le signal $\mathbf{r}^{(i)}(n, f)$ est ensuite calculé comme dans (4.12).

Paramètres des covariances Θ_c Comme il n'y a pas de solution analytique pour l'optimisation de la log-vraisemblance $\mathcal{L}\left(\mathcal{O}; \widehat{\Theta}_H^{(i)}, \widehat{\Theta}_G^{(i)}, \Theta_c\right)$ en fonction de Θ_c, nous estimons les paramètres des covariances à l'aide d'un algorithme EM. Sachant les trames précédentes du mélange $\mathbf{d}(n, f)$, le signal de la parole distante $x(n, f)$ et ses trames précédentes, et les filtres linéaires $\mathcal{H}(f)$ et $\mathcal{G}(f)$, le signal $\mathbf{r}(n, f)$ est conditionnellement distribué comme suit :

$$\mathbf{r}(n, f) \Big| \mathbf{d}(n - 1, f), \dots, \mathbf{d}(0, f), x(n, f), \dots, x(0, f), \mathcal{H}(f), \mathcal{G}(f) \sim \mathcal{N}_{\mathbb{C}}\Big(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{dd}}(n, f)\Big). \tag{4.38}$$

Le modèle de signal est conditionnellement identique au modèle gaussien local pour la séparation de sources (voir la partie 2.2.1.4). Toutefois, l'algorithme EM ne contraint ni les DSPs $v_c(n, f)$, ni les MCSs $\mathbf{R}_c(f)$, ce qui conduit à une ambiguïté de permutation. À la place, après chaque mise à jour des filtres linéaires $\mathcal{H}^{(i)}(f)$ et $\mathcal{G}^{(i)}(f)$ à l'itération i, nous proposons d'utiliser une itération de l'algorithme DNN-EM de Nugraha et al. [2016a] pour estimer les DSPs et les MCSs de la parole cible et des signaux résiduels $\mathbf{s}_{\mathrm{e}}^{(i)}(n, f)$,

$\mathbf{s}_{\mathrm{r}}^{(i)}(n,f)$, $\mathbf{z}_{\mathrm{r}}^{(i)}(n,f)$ et $\mathbf{b}_{\mathrm{r}}^{(i)}(n,f)$. À l'étape E, chacun des quatre signaux $\mathbf{c}^{(i)}(n,f)$ est estimé par

$$\widehat{\mathbf{c}}^{(i)}(n,f) = \mathbf{W}_c^{(i)}(n,f)\mathbf{r}^{(i)}(n,f), \tag{4.39}$$

où $\mathbf{W}_c^{(i)}(n,f)$ est le filtre de Wiener obtenu comme dans (4.21) à partir des paramètres $v_c^{(i-1)}(n,f)$ et $\mathbf{R}_c^{(i-1)}(f)$, et le signal $\mathbf{r}^{(i)}(n,f)$ est obtenu par application successive des filtres linéaires $\mathcal{H}^{(i)}(f)$ et $\mathcal{G}^{(i)}(f)$ sur le mélange $\mathbf{d}(n,f)$ comme dans (4.11)–(4.12). Le moment non centré d'ordre 2 $\widehat{\mathbf{\Sigma}}_c^{(i)}(n,f)$ est estimé par

$$\widehat{\mathbf{\Sigma}}_c^{(i)}(n,f) = \widehat{\mathbf{c}}^{(i)}(n,f)\widehat{\mathbf{c}}^{(i)}(n,f)^H + \Big(\mathbf{I} - \mathbf{W}_c^{(i)}(n,f)\Big)v_c^{(i-1)}(n,f)\mathbf{R}_c^{(i-1)}(f). \tag{4.40}$$

À l'étape M, nous considérons une forme pondérée de la mise à jour des MCSs [Nugraha et al., 2016b] :

$$\mathbf{R}_c^{(i)}(f) = \left(\sum_{n=0}^{N-1}\omega_c^{(i)}(n,f)\right)^{-1}\sum_{n=0}^{N-1}\frac{\omega_c^{(i)}(n,f)}{v_c^{(i-1)}(n,f)}\widehat{\mathbf{\Sigma}}_c^{(i)}(n,f), \tag{4.41}$$

où $\omega_c^{(i)}(n,f)$ désigne un poids associé au signal $\mathbf{c}(n,f)$. Lorsque $\omega_c^{(i)}(n,f) = 1$, (4.41) revient à l'algorithme EM exact. Ici, nous utilisons les poids suivants [Liutkus et al., 2015; Nugraha et al., 2016b] :

$$\omega_c^{(i)}(n,f)(n,f) = v_c^{(i-1)}(n,f). \tag{4.42}$$

Les expériences montrent que cette pondération réduit les mauvaises estimations dans certaines bandes de fréquences en augmentant l'importance des points temps-fréquence pour lesquels $v_c^{(i-1)}(n,f)$ est grand. Comme les DSPs sont contraintes, nous devons aussi contraindre les MCSs $\mathbf{R}_c^{(i)}(f)$ de manière à n'encoder que l'information spatiale des signaux $\mathbf{c}(n,f)$. Nous modifions (4.41) en normalisant $\mathbf{R}_c^{(i)}(f)$ après chaque mise à jour [Nugraha et al., 2016b] :

$$\mathbf{R}_c^{(i)}(f) \leftarrow \frac{M}{\mathrm{tr}\left(\mathbf{R}_c^{(i)}(f)\right)}\mathbf{R}_c^{(i)}(f). \tag{4.43}$$

Les opérations (4.41)–(4.43) correspondent à la mise à jour spatiale.

Les DSPs $v_c^{(i)}(n,f)$ des quatre signaux sont mises à jour conjointement à l'aide d'un DNN désigné par DNN_i, avec $i \geq 1$, que nous avons pré-entraîné. Cette opération correspond à la mise à jour spectrale. Nous décrivons les entrées, les cibles et l'architecture de DNN_i dans la partie 4.2.4 ci-après. En particulier, les entrées de DNN_i sont calculées à partir des valeurs de $\Theta_H^{(i)}$, $\Theta_G^{(i)}$ et $\mathbf{R}_c^{(i)}$ et $\widehat{\mathbf{c}}^{(i)}(n,f)$ à l'itération i.

4.2.3.3. Estimation de la composante précoce finale $\mathbf{s}_{\mathrm{e}}(n,f)$

Une fois que l'algorithme itératif d'optimisation a convergé après I itérations, nous avons l'estimation finale de parole cible $\widehat{\mathbf{s}}_{\mathrm{e}}^{(I)}(n,f)$ en utilisant (4.11), (4.12) et (4.19). Le pseudo-code de l'algorithme est détaillé dans l'annexe A.

4.2.4. Modèle spectral par réseau de neurones

Dans cette partie, nous définissons les entrées, les cibles et l'architecture des DNNs utilisés pour initialiser et mettre à jour les DSPs de la parole cible et des signaux résiduels.

4.2.4.1. Cibles

Il a été prouvé que l'estimation de la densité spectrale d'amplitude $\sqrt{v_c(n,f)}$ de la DSP $v_c(n,f)$ donne de meilleurs résultats [Nugraha et al., 2016a]. Par conséquent, nous choisissons $\left[\sqrt{v_{s_\mathrm{e}}(n,f)}\ \sqrt{v_{s_\mathrm{r}}(n,f)}\ \sqrt{v_{z_\mathrm{r}}(n,f)}\ \sqrt{v_{b_\mathrm{r}}(n,f)}\right]$ pour les cibles du DNN. Nugraha et al. [2016a] a défini la vérité terrain des DSPs de la manière suivante :

$$v_c(n,f) = \frac{1}{M}\|\mathbf{c}(n,f)\|^2. \tag{4.44}$$

Pour cela, nous avons besoin de connaître la vérité terrain des signaux $\mathbf{c}(n,f)$. Toutefois, la vérité terrain des signaux latents $\mathbf{s}_\mathrm{r}(n,f)$, $\mathbf{z}_\mathrm{r}(n,f)$ et $\mathbf{b}_\mathrm{r}(n,f)$ est inconnue.

Dans les ensembles d'apprentissage et de validation, il est possible de connaître la vérité terrain de la composante précoce $\mathbf{s}_\mathrm{e}(n,f)$ et des signaux $\mathbf{s}_\mathrm{l}(n,f)$, $\mathbf{y}(n,f)$ et $\mathbf{b}(n,f)$ (voir la partie 4.3.2). Ces trois derniers signaux correspondent aux signaux originaux de distorsion $\mathbf{s}_\mathrm{r}(n,f)$, $\mathbf{z}_\mathrm{r}(n,f)$ et $\mathbf{b}_\mathrm{r}(n,f)$, respectivement, lorsque les filtres linéaires $\mathcal{H}(f)$ et $\mathcal{G}(f)$ sont égaux à zéro. Pour déterminer la vérité terrain des signaux latents $\mathbf{s}_\mathrm{r}(n,f)$, $\mathbf{z}_\mathrm{r}(n,f)$ et $\mathbf{b}_\mathrm{r}(n,f)$, nous proposons donc d'utiliser un algorithme itératif similaire à l'algorithme DNN-BCA (voir la figure 4.4), où les filtres linéaires $\mathcal{H}(f)$ et $\mathcal{G}(f)$ sont initialisés à zéro.

À l'initialisation, nous assignons donc les signaux résiduels à la valeur des signaux originaux de distorsion :

$$\mathbf{s}_r(n,f) \leftarrow \mathbf{s}_\mathrm{l}(n,f), \tag{4.45}$$
$$\mathbf{z}_r(n,f) \leftarrow \mathbf{y}(n,f), \tag{4.46}$$
$$\mathbf{b}_r(n,f) \leftarrow \mathbf{b}(n,f). \tag{4.47}$$

Nous initialisons les DSPs $v_c(n,f)$ comme dans (4.44) et les MCSs $\mathbf{R}_c(f)$ à la matrice identité $\mathbf{I}_M$.

À chaque itération, nous mettons à jour les filtres linéaires $\mathcal{H}(f)$ et $\mathcal{G}(f)$ comme dans (4.29) et (4.34), respectivement. Pour mettre à jour les signaux résiduels $\mathbf{s}_\mathrm{r}(n,f)$, $\mathbf{z}_\mathrm{r}(n,f)$ et $\mathbf{b}_\mathrm{r}(n,f)$, nous appliquons les filtres linéaires $\mathcal{H}(f)$ et $\mathcal{G}(f)$ séparément à chacun des signaux originaux de distorsion $\mathbf{s}_\mathrm{l}(n,f)$, $\mathbf{y}(n,f)$ et $\mathbf{b}(n,f)$ comme dans (4.14), (4.15) et (4.16).

Le but de cette procédure itérative est d'obtenir la vérité terrain des DSPs $v_c(n,f)$. Toutefois, nous n'appliquons pas (4.44) sur la composante précoce $\mathbf{s}_\mathrm{e}(n,f)$ et les signaux latents $\mathbf{s}_\mathrm{r}(n,f)$, $\mathbf{z}_\mathrm{r}(n,f)$ et $\mathbf{b}_\mathrm{r}(n,f)$, car cette formule ne tient pas compte des MCSs $\mathbf{R}_c(f)$. À la place, nous proposons d'utiliser, dans la procédure itérative, une mise à jour spatiale et spectrale, inspirée de l'algorithme EM en séparation de sources [Duong et al., 2010]. À l'étape E, l'estimation du signal $\widehat{\mathbf{c}}(n,f)$ est remplacé par sa vérité terrain

$\mathbf{c}(n,f)$, et les moments non centrés d'ordre 2 $\widehat{\mathbf{\Sigma}}_c(n,f)$ sont remplacés par $\mathbf{c}(n,f)\mathbf{c}(n,f)^H$. À l'étape M, les DSPs $v_c(n,f)$ sont alors déterminées comme

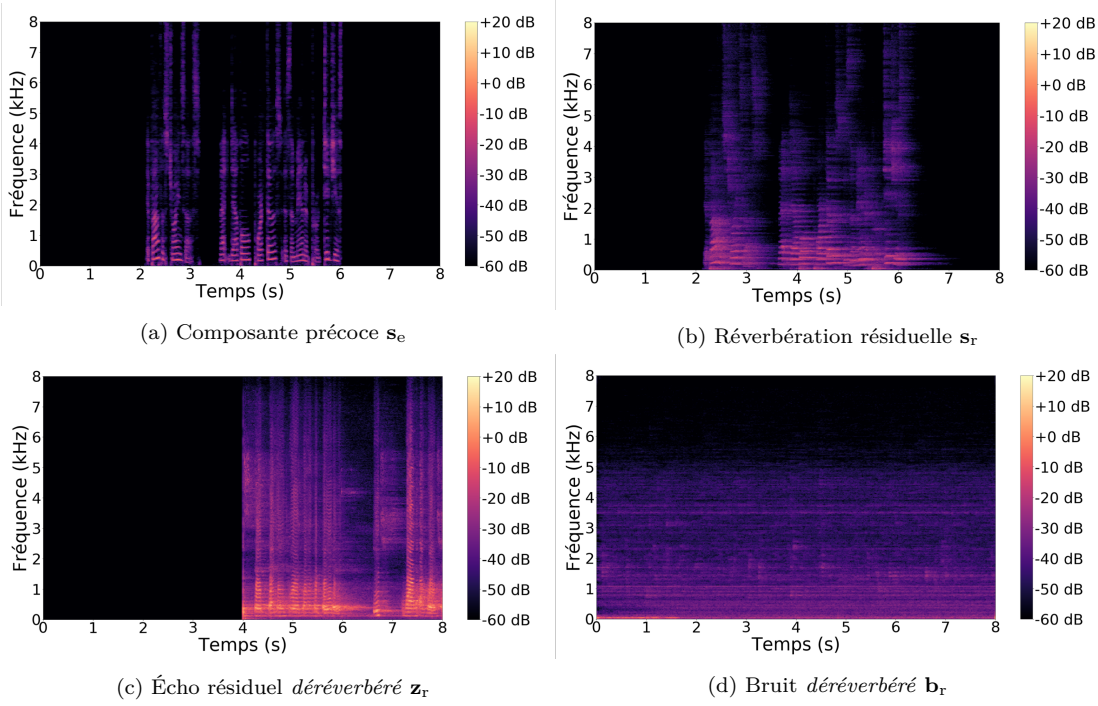
$$v_c(n,f) = \frac{1}{M}\mathrm{tr}\Big(\mathbf{R}_c(f)^{-1}\mathbf{c}(n,f)\mathbf{c}(n,f)^H\Big), \tag{4.48}$$

et les MCSs $\mathbf{R}_c(f)$ sont déterminées de la manière suivante :

$$\mathbf{R}_c(f) = \frac{1}{N}\sum_{n=0}^{N-1}\frac{1}{v_c(n,f)}\mathbf{c}(n,f)\mathbf{c}(n,f)^H. \tag{4.49}$$

Les MCSs sont ensuite normalisées comme dans (4.43) de manière à n'encoder que l'information spatiale des signaux. Une procédure similaire a été utilisée par Nugraha et al. [2016b].

Le pseudo-code de la procédure de détermination des vérités terrain des DSPs est détaillée dans l'annexe A. Après quelques itérations, nous observons la convergence des variables latentes $\mathbf{s}_\mathrm{r}(n,f)$, $\mathbf{z}_\mathrm{r}(n,f)$ et $\mathbf{b}_\mathrm{r}(n,f)$. La figure 4.5 illustre un exemple de vérité terrain des DSPs obtenues après convergence de la procédure itérative.



(a) Composante précoce $\mathbf{s}_\mathrm{e}$

(b) Réverbération résiduelle $\mathbf{s}_\mathrm{r}$

(c) Écho résiduel *déréverbéré* $\mathbf{z}_\mathrm{r}$

(d) Bruit *déréverbéré* $\mathbf{b}_\mathrm{r}$

FIGURE 4.5. – Exemple de vérité terrain des DSPs de la parole cible et signaux résiduels dans l'ensemble d'apprentissage.

4.2.4.2. Entrées

Nous utilisons les spectres en amplitude comme entrées de DNN_0 et DNN_i plutôt que les spectres en puissance, puisque il a été prouvé qu'ils donnent de meilleurs résultats

lorsque les cibles sont les densités spectrales d'amplitude $\sqrt{v_c(n,f)}$ [Nugraha et al., 2016a]. Les différentes entrées sont résumées sur la figure 4.6. Nous considérons tout d'abord le spectre en amplitude de la parole distante $|x(n,f)|$ et le spectre en amplitude monocanal $|\widetilde{d}(n,f)|$ obtenu à partir du mélange multicanal correspondant $\mathbf{d}(n,f)$ de la manière suivante [Nugraha et al., 2016a] :

$$|\widetilde{d}(n,f)| = \sqrt{\frac{1}{M}\|\mathbf{d}(n,f)\|^2}. \tag{4.50}$$

De plus, nous utilisons les spectres en amplitude $|\widetilde{y}^{(0)}(n,f)|$, $|\widetilde{e}^{(0)}(n,f)|$, $|\widetilde{e}_1^{(0)}(n,f)|$ et $|\widetilde{r}^{(0)}(n,f)|$ obtenus à partir des signaux multicanaux correspondants $\widehat{\mathbf{y}}^{(0)}(n,f)$, $\mathbf{e}^{(0)}(n,f)$, $\widehat{\mathbf{e}}_1^{(0)}(n,f)$ et $\mathbf{r}^{(0)}(n,f)$ en utilisant les filtres linéaires $\mathcal{H}^{(0)}(f)$ et $\mathcal{G}^{(0)}(f)$ à l'initialisation. En effet, dans le chapitre 3, nous avons montré que l'utilisation du spectre en amplitude de l'écho estimé $|\widehat{y}(n,f)|$ comme entrée supplémentaire du DNN permettait d'améliorer l'estimation du filtre $w_s(n,f)$ par le DNN. Nous désignons les entrées $|\widetilde{d}(n,f)|$, $|x(n,f)|$, $|\widetilde{y}^{(0)}(n,f)|$, $|\widetilde{e}^{(0)}(n,f)|$, $|\widetilde{e}_1^{(0)}(n,f)|$ et $|\widetilde{r}^{(0)}(n,f)|$ par le terme d'entrées de type I. Les entrées de DNN_0 correspondent à la concaténation suivante (voir la figure 4.6) :

$$\begin{aligned}\text{entrées}^{(0)} &= \text{entrées}_{\text{I}}^{(0)} && (4.51)\\ &= \left[|\widetilde{d}(n,f)|, |x(n,f)|, |\widetilde{y}^{(0)}(n,f)|, |\widetilde{e}^{(0)}(n,f)|, |\widetilde{e}_1^{(0)}(n,f)|, |\widetilde{r}^{(0)}(n,f)|\right]. && (4.52)\end{aligned}$$

Pour DNN_i, avec $i \geq 1$, nous considérons des entrées supplémentaires pour améliorer l'estimation du DNN. Nous utilisons en particulier les racines carrées $\sqrt{v_c^{\text{unc}}(n,f)}^{(i)}$ des DSPs non contraintes :

$$\sqrt{v_c^{\text{unc}}(n,f)}^{(i)} = \sqrt{\frac{1}{M}\text{tr}\left(\mathbf{R}_c^{(i)}(f)^{-1}\widehat{\mathbf{\Sigma}}_c^{(i)}(n,f)\right)}. \tag{4.53}$$

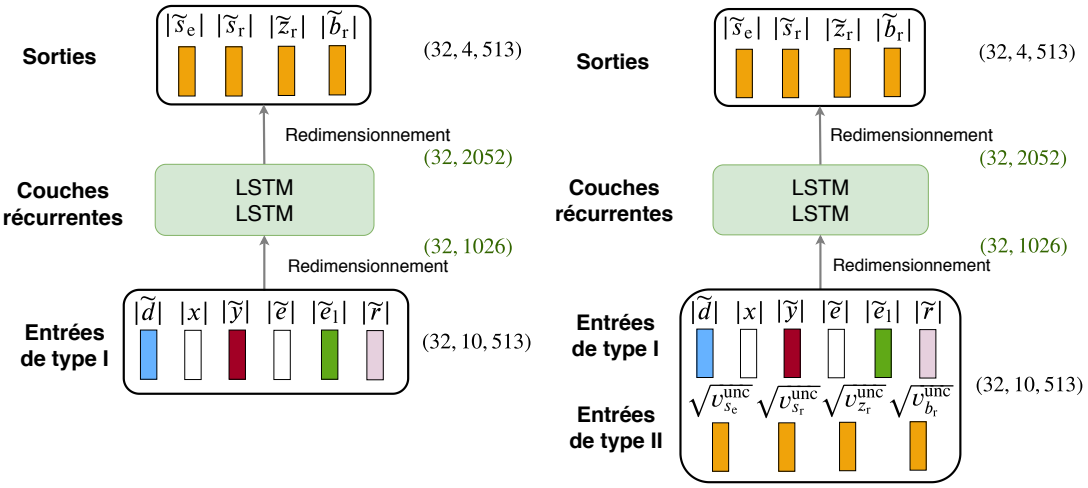
En effet, ces entrées exploitent l'information spatiale des signaux. Par ailleurs, il a été prouvé qu'elles donnent de meilleurs résultats en séparation de sources [Nugraha et al., 2016a]. Nous désignons les entrées $\sqrt{v_c^{\text{unc}}(n,f)}^{(i)}$ obtenues à partir de (4.53) par le terme d'entrées de type II. Les entrées de DNN_i avec $i \geq 1$ correspondent donc à la concaténation suivante (voir la figure 4.6) :

$$\text{entrées}^{(i)} = \left[\text{entrées}_{\text{I}}^{(i)} \text{entrées}_{\text{II}}^{(i)}\right]. \tag{4.54}$$

où $\text{entrées}_{\text{I}}^{(i)}$ correspond aux entrées de type I estimées à l'itération i de la même manière que (4.52), et

$$\text{entrées}_{\text{II}}^{(i)} = \left[\sqrt{v_{s_{\text{e}}}^{\text{unc}}(n,f)}^{(i)} \; \sqrt{v_{s_{\text{r}}}^{\text{unc}}(n,f)}^{(i)} \; \sqrt{v_{z_{\text{r}}}^{\text{unc}}(n,f)}^{(i)} \; \sqrt{v_{b_{\text{r}}}^{\text{unc}}(n,f)}^{(i)}\right]. \tag{4.55}$$

Il convient de noter que pour DNN_0, nous n'utilisons que les entrées de type I, car les entrées de type II ne sont pas accessibles à l'initialisation.



(a) DNN_0 (b) DNN_i, $i \geq 1$

FIGURE 4.6. – Architecture des DNNs avec une longueur de séquence de 32 intervalles de temps et $F = 513$ bandes de fréquence.

4.2.4.3. Fonction de coût

Nous désignons par $|\widetilde{c}(n,f)|$ la sortie du DNN pour le signal $\mathbf{c}(n,f)$. Comme indiqué précédemment, nous utilisons DNN_0 et DNN_i pour estimer conjointement les quatre paramètres spectraux $\left[|\widetilde{s}_{\mathrm{e}}(n,f)||\widetilde{s}_{\mathrm{r}}(n,f)||\widetilde{z}_{\mathrm{r}}(n,f)||\widetilde{b}_{\mathrm{r}}(n,f)|\right]$ (voir la figure 4.6). Nous utilisons la divergence de Kullback-Leibler comme fonction de coût :

$$\mathcal{D}_{KL} = \frac{1}{4FN} \sum_{c,n,f} \left(\sqrt{v_c(n,f)} \log \frac{\sqrt{v_c(n,f)}}{|\widetilde{c}(n,f)|} - \sqrt{v_c(n,f)} + |\widetilde{c}(n,f)| \right). \tag{4.56}$$

En effet, il a été prouvé que celle-ci donne les meilleurs résultats parmi plusieurs autres fonctions de coût pour l'apprentissage des DNNs dans le contexte de la séparation de sources [Nugraha et al., 2016a].

4.2.4.4. Architecture

Nous considérons une architecture LSTM à 1 couche cachée (voir la figure 4.6). Le nombre d'unités en entrée est de $6F$ pour DNN_0 et $10F$ pour DNN_i, où F désigne le nombre de bandes de fréquence. Les fonctions d'activation des couches sont des ReLUs. Nous ne considérons pas d'autre architecture de réseau ici, car la comparaison entre différentes architectures dépasse le cadre de ce travail.

4.3. Protocole expérimental

Dans cette partie, nous décrivons les données, les métriques, les méthodes de référence et le réglage des hyperparamètres utilisés pour évaluer l'algorithme proposé.

4.3.1. Scénario

Nous considérons la situation où un locuteur local interagit avec un correspondant distant à l'aide de Triby à une distance de 1,5 m dans un environnement bruyant. Chaque enregistrement a une durée de 8 s qui contient 4 s de parole locale et 4 s de parole distante qui se chevauchent pendant 2 s. Nous étudions 2 scénarios : dans le premier, un bruit ambiant est présent pendant tout l'enregistrement, et dans le second, le bruit ambiant y est absent. Chaque enregistrement du premier scénario, en présence de bruit, est composé de 4 périodes de 2 s, comme le montre la figure 4.7 : 1) *bruit seul*, 2) *bruit et parole locale*, 3) *bruit, parole locale et parole distante*, 4) *bruit et parole distante*. Le second scénario, en absence de bruit, possède la même organisation temporelle que le scénario précédent, sans le signal de bruit.

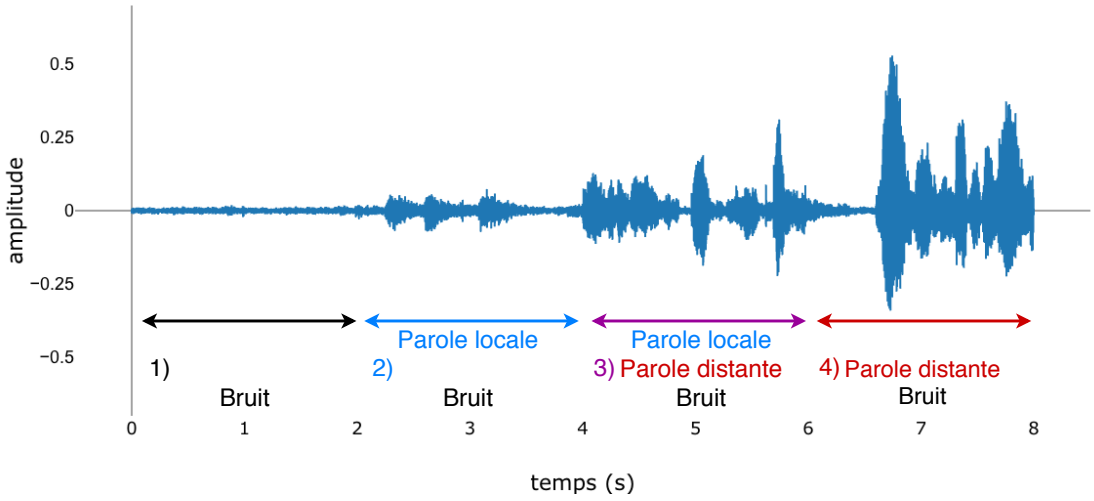


FIGURE 4.7. – Exemple d'enregistrement avec bruit (seul un canal est illustré).

4.3.2. Données

4.3.2.1. Description générale

Nous créons trois ensembles de données disjoints pour l'apprentissage, la validation et le test. Leurs caractéristiques sont résumées dans le tableau 4.1. Pour chaque ensemble de données, nous avons créé séparément l'écho acoustique $\mathbf{y}(t)$, la parole locale $\mathbf{s}(t)$ et le signal de bruit $\mathbf{b}(t)$, à partir de signaux originaux de parole et de bruit. Nous avons enregistré l'écho $y(t)$ avec un Triby, l'enceinte intelligente développée par Invoxia, qui

est dotée de 4 microphones. Toutefois, comme l'un des microphones était défectueux, nous avons considéré seulement $M = 3$ microphones. Pour la parole locale $\mathbf{s}(t)$ et le signal de bruit $\mathbf{b}(t)$, nous avons soit simulé, soit enregistré avec le Triby les signaux selon l'ensemble de données. Le signal du microphone $\mathbf{d}(t)$ est ensuite calculé comme dans (4.1). Ce protocole est nécessaire afin d'obtenir les vérités terrain des signaux pour l'apprentissage et la validation, ce qui n'est pas possible avec des enregistrements réels où les signaux $\mathbf{s}(t)$, $\mathbf{y}(t)$ et $\mathbf{b}(t)$ ne sont pas observés séparément. Les ensembles d'apprentissage et de validation correspondent à des conditions acoustiques invariantes au cours du temps, tandis que l'ensemble de test comprend deux sous-ensembles : l'un dont les conditions acoustiques sont invariantes au cours du temps, et l'autre dont ces conditions varient au cours du temps. Les paramètres d'enregistrement et de simulation sont détaillés dans l'annexe A.

Ensemble de données	**Apprentissage**	**Validation**	**Test**
Signaux **y**	enregistrés		enregistrés
Signaux **s**	RIRs $\mathbf{a}_s$ simulées		
Signaux **b**	simulés		
Salles	1-2-3	1-2	4
# paires de locuteurs	79	27	25
# phrases	13 572	4 536	4 500
# échantillons de bruit	36	36	6
Plage de SER (dB)	$[-45, +6]$		$[-45, -7]$
Plage de SNR (dB)	$[-21, +24]$		$[-20, +13]$

TABLEAU 4.1. – Caractéristiques des trois ensembles de données.

Signaux orignaux de parole et de bruit Les signaux originaux de parole proviennent du sous-ensemble « train-clean-360 » du corpus Librispeech [Panayotov et al., 2015], qui comprend 921 locuteurs lisant des extraits de livres, dont l'ensemble des segments par locuteur fait 25 min. Nous avons sélectionné 262 locuteurs que nous avons regroupés en 131 paires, réparties en sous-ensembles disjoints entre l'apprentissage, la validation et le test. Pour chaque paire, nous avons considéré chacun des deux locuteurs comme locuteur local et comme correspondant distant de manière alternée, et nous avons choisi plusieurs échantillons disjoints de 4 s de parole. Chaque échantillon de 4 s n'a été utilisé qu'une fois.

En ce qui concerne les signaux de bruit, nous avons considéré 6 types de bruit domestique : bribes de conversations, lave-vaisselle, réfrigérateur, micro-ondes, aspirateur et machine à laver. Nous avons sélectionné aléatoirement 78 échantillons disjoints de 8 s de bruit à partir de vidéos YouTube, et nous les avons regroupés en sous-ensembles disjoints pour l'apprentissage, la validation et le test.

Enregistrements d'écho réels Pour créer l'écho acoustique $\mathbf{y}(t)$, Togami et Kawaguchi [2014] ont convolué les signaux de parole distante $x(t)$ avec des chemins d'écho $\mathbf{a}_s(\tau)$

simulés, qui ne contiennent aucune non-linéarité. Dans les systèmes mains-libres, l'écho acoustique $y(t)$ contient des non-linéarités dues à la réponse non-linéaire du haut-parleur et des microphones, aux vibrations de l'enceinte et aux effets d'écrêtage de l'amplification (voir la partie 2.2.2). Pour réaliser des tests dans des conditions réalistes, nous créons l'écho $y(t)$ en enregistrant la rétroaction acoustique du haut-parleur du Triby vers les microphones du Triby. La parole distante $x(t)$ a été jouée à une fréquence d'échantillonnage de 16 kHz par le Triby, et l'écho correspondant a été enregistré par l'antenne linéaire de 4 microphones de ce même Triby. Toutefois, nous rappelons que comme l'un des microphones était défectueux, nous avons considéré seulement $M = 3$ microphones (voir la partie 4.3.2.1). La figure 4.8 montre une configuration de l'installation pour l'enregistrement de l'écho. Les enregistrements ont été réalisés avec le même Triby dans quatre salles de taille et de temps de réverbération T_{60} différents (voir le tableau 4.2).

FIGURE 4.8. – Installation pour l'enregistrement de l'écho pour l'ensemble d'apprentissage.

Salle	**Dimensions** (m)	$\mathbf{T_{60}}$ (s)
1	$4,4 \times 4,2 \times 4$	$1,0$
2	$3,8 \times 2,5 \times 3,5$	$0,5$
3	$3,4 \times 2,1 \times 3,3$	$0,8$
4	$5,9 \times 4,6 \times 4$	$1,3$

TABLEAU 4.2. – Caractéristiques des salles.

Parole locale et bruit Les signaux de parole locale $\mathbf{s}(t)$ et de bruit $\mathbf{b}(t)$ ont été soit simulés, soit enregistrés par le Triby, en fonction de l'ensemble de données considéré. Nous les décrivons les procédures de simulation et d'enregistrement dans les parties suivantes.

4.3.2.2. Ensemble d'apprentissage

Pour l'ensemble d'apprentissage, nous ne considérons que le scénario avec présence de bruit. Les enregistrements d'écho $\mathbf{y}(t)$ ont été réalisés dans les salles 1, 2 et 3 (voir le tableau 4.2). Pour créer la parole locale $\mathbf{s}(t)$, nous avons convolué des signaux originaux de parole $u(t)$ avec des RIRs de parole locale $\mathbf{a}_s(\tau)$. Les RIRs $\mathbf{a}_s(\tau)$ ont été simulées à l'aide du logiciel Roomsimove, de manière à ce qu'elles correspondent aux propriétés d'enregistrement de l'écho [Vincent et Campbell, 2008]. Les paramètres de simulation sont détaillés dans l'annexe A.

Parmi les 79 paires de locuteurs utilisées pour l'apprentissage, 54 ont été utilisées dans les salles 1 et 2 (voir le tableau 4.2). Nous avons joué et enregistré 4 536 signaux de parole distante $x(t)$ pour obtenir l'écho $\mathbf{y}(t)$. Nous avons simulé 4 536 RIRs $\mathbf{a}_s(\tau)$ appliquées à la parole locale anéchoïque $u(t)$ dans chacune des salles 1 et 2. Les 25 paires restantes ont été utilisées dans la salle 3 (voir le tableau 4.2). Nous avons joué et enregistré 4 500 signaux de parole distante $x(t)$ pour obtenir l'écho $\mathbf{y}(t)$. Nous avons simulé 4 500 RIRs $\mathbf{a}_s(\tau)$ appliquées à la parole locale anéchoïque $u(t)$ dans cette salle 3.

Pour créer le signal de bruit $\mathbf{b}(t)$, nous avons tout d'abord mesuré 14 RIRs dans chacune des salles 1, 2 et 3, soit 42 RIRs au total. Nous avons convolué un échantillon de bruit choisi au hasard parmi 36 échantillons de bruit (6 par type de bruit) utilisés pour l'apprentissage (voir le tableau 4.1) avec la moyenne de deux RIRs sélectionnées au hasard parmi les 42 RIRs mesurées. Cette procédure permet de créer un signal de bruit spatialement diffus, ce qui ne serait pas le cas avec des RIRs simulées.

Les niveaux d'écho $\mathbf{y}(t)$, de parole locale $\mathbf{s}(t)$ et de signal de bruit $\mathbf{b}(t)$ ont été choisis aléatoirement de manière à ce que le SER varie de -45 dB à $+6$ dB, et à ce que le SNR varie de -21 dB à $+24$ dB. Ces conditions sont très difficiles, d'autant plus que la réverbération domine dans la parole locale $\mathbf{s}(t)$. Au total, nous avons enregistré 13 572 signaux d'écho $\mathbf{y}(t)$, nous avons simulé 13 572 signaux de parole locale $\mathbf{s}(t)$ avec des RIRs $\mathbf{a}_s(\tau)$ simulées, et nous avons simulé 13 572 signaux de bruit $\mathbf{b}(t)$ avec des RIRs mesurées, soit environ 32 h d'audio (voir le tableau 4.1).

4.3.2.3. Ensemble de validation

Pour l'ensemble de validation, nous ne considérons que le scénario en présence de bruit. L'ensemble de validation a été généré de la même manière que l'ensemble d'apprentissage, en utilisant 27 paires de locuteurs et 36 échantillons de bruit qui ne sont pas dans l'ensemble d'apprentissage. Les enregistrements d'écho ont été réalisés dans les salles 1 et 2 (voir le tableau 4.2). Les RIRs de parole locale ont été simulées de manière similaire à la procédure de l'ensemble d'apprentissage.

Nous avons joué et enregistré 4 536 signaux de parole distante $x(t)$ pour obtenir l'écho $\mathbf{y}(t)$. Nous avons simulé 4 536 RIRs $\mathbf{a}_s(\tau)$ appliquées à la parole locale anéchoïque $u(t)$ dans chacune des salles 1 et 2. Nous n'avons pas considéré la salle 3 afin d'avoir une plus grande variété de salles dans l'apprentissage. Pour créer le bruit diffus, nous avons utilisé les mêmes 42 RIRs mesurées que dans l'ensemble d'apprentissage.

Les niveaux d'écho $\mathbf{y}(t)$, de parole locale $\mathbf{s}(t)$ et de signal de bruit $\mathbf{b}(t)$ ont été choi-

sis sur la même plage que pour l'ensemble d'apprentissage, ce qui aboutit aux mêmes conditions difficiles de SER et SNR. Au total, nous avons enregistré 4 536 signaux d'écho $\mathbf{y}(t)$, nous avons simulé 4 536 signaux de parole locale $\mathbf{s}(t)$ avec des RIRs $\mathbf{a}_s(\tau)$ simulées, et nous avons simulé 4 536 signaux de bruit $\mathbf{b}(t)$ avec des RIRs mesurées, soit environ 10 h d'audio (voir le tableau 4.1).

4.3.2.4. Ensemble de test invariant au cours du temps

Pour l'ensemble de test, nous considérons les deux scénarios, en présence et en absence de bruit. L'ensemble de test a été créé uniquement à partir d'enregistrements réels, en utilisant 25 paires de locuteurs et 6 échantillons de bruit qui ne sont ni dans l'ensemble d'apprentissage, ni dans l'ensemble de validation.

Pour le scénario avec présence de bruit, l'écho $\mathbf{y}(t)$, la parole locale $\mathbf{s}(t)$ et le signal de bruit $\mathbf{b}(t)$ sont tous enregistrés dans la salle 4 (voir le tableau 4.2). La figure 4.9 montre une configuration de l'installation pour l'enregistrement des trois signaux. Les paramètres d'enregistrement sont détaillés dans l'annexe A. Pour le scénario en absence de bruit, les mêmes enregistrements ont été utilisés, sans le signal de bruit $\mathbf{b}(t)$.

La parole locale $\mathbf{s}(t)$ a été obtenue en jouant le signal original de parole locale avec un haut-parleur Yamaha MSP5 Studio à un unique niveau de volume. Le signal de bruit $\mathbf{b}(t)$ a été obtenu en sélectionnant aléatoirement un signal original de bruit parmi les 6 échantillons de bruit (1 par type de bruit) et en le jouant simultanément sur les haut-parleurs de 4 Tribys. Les signaux de bruit $\mathbf{b}(t)$ résultant de cette procédure sont cependant moins diffus que dans les ensembles d'apprentissage et de validation.

Les niveaux enregistrés ont été choisis de manière à ce que le SER varie de -45 dB à -7 dB, et le SNR varie de -20 dB à $+13$ dB. Ces conditions sont difficiles mais néanmoins comprises dans les intervalles de SER et SNR des ensembles d'apprentissage et de validation. Pour le scénario en présence de bruit, nous avons joué et enregistré 4 500 signaux d'écho $\mathbf{y}(t)$, de parole locale $\mathbf{s}(t)$ et de bruit $\mathbf{b}(t)$ (voir le tableau 4.1). Pour le scénario en absence de bruit, nous avons réutilisé ces enregistrements, sans les signaux de bruit. Au total, nous avons enregistré 9 000 signaux d'écho, de parole locale et de bruit, soit environ 20 h d'audio.

4.3.2.5. Ensemble de test variant au cours du temps

Afin d'évaluer notre approche dans des conditions acoustiques qui varient au cours du temps, nous avons aussi considéré le scénario où le locuteur local parle pendant 4 s, se déplace à un autre endroit de la salle, et parle à nouveau pendant 4 s. Pour cela, nous avons concaténé des paires d'enregistrements de 8 s de parole locale $\mathbf{s}(t)$ provenant de l'ensemble de test invariant au cours du temps qui correspondent aux mêmes locuteur local et correspondant distant, à deux positions différentes du haut-parleur jouant la parole locale. Les sources de bruit, c'est-à-dire les haut-parleurs des 4 Tribys évoqués dans la sous-partie précédente, restent immobiles durant tout l'enregistrement. Contrairement à l'ensemble de test invariant au cours du temps, nous n'avons considéré que le scénario en présence de bruit, car le but ici n'est pas d'analyser les performances en présence ou

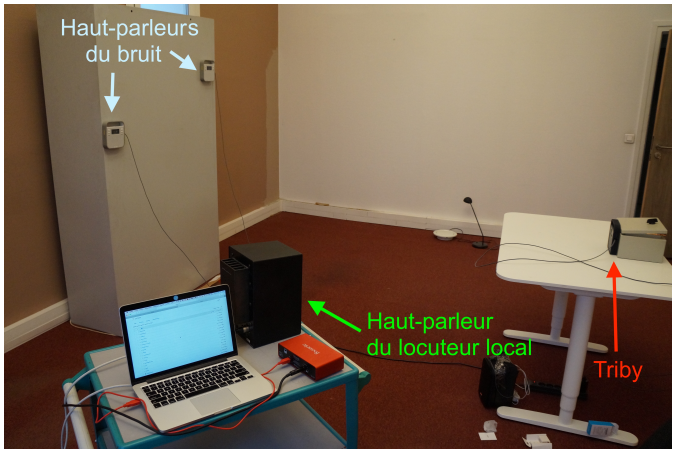


FIGURE 4.9. – Installation pour l'enregistrement de l'écho pour l'ensemble de test.

absence de bruit. Par conséquent, il y a 2 250 enregistrements de 16 s, soit environ 10 h d'audio.

4.3.3. Métriques

La composante précoce estimée $\widehat{\mathbf{s}}_{\mathrm{e}}$, obtenue avec (4.19), contient cinq composantes :

$$\widehat{\mathbf{s}}_{\mathrm{e}} = \mathbf{s}_{\mathrm{e}}^{\mathrm{post}} + \mathbf{s}_{\mathrm{r}}^{\mathrm{post}} + \mathbf{z}_{\mathrm{r}}^{\mathrm{post}} + \mathbf{b}_{\mathrm{r}}^{\mathrm{post}} + \mathbf{s}_{\mathrm{e}}^{\mathrm{art}}, \tag{4.57}$$

où $\mathbf{s}_{\mathrm{e}}^{\mathrm{post}}$ est la composante précoce potentiellement atténuée, $\mathbf{s}_{\mathrm{r}}^{\mathrm{post}}$, $\mathbf{z}_{\mathrm{r}}^{\mathrm{post}}$ et $\mathbf{b}_{\mathrm{r}}^{\mathrm{post}}$ sont les trois signaux post-résiduels de la réverbération tardive, de l'écho et bruit, respectivement, et $\mathbf{s}_{\mathrm{e}}^{\mathrm{art}}$ représente les artefacts introduits dans la composante précoce $\mathbf{s}_{\mathrm{e}}$. Ces composantes sont calculées de la même que dans (2.89)–(2.91). À partir de ces composantes, nous utilisons les métriques définies dans la partie 2.4.1 pour évaluer la réduction de chaque type de distorsion et la dégradation de la composante précoce. Ces métriques sont résumées dans le tableau 4.4. Les métriques sont calculées séparément sur chaque canal m, puis moyennées sur les M canaux.

Les ELR, SI-SAR et SI-SDR sont évalués uniquement dans les situations de *parole locale* et de *parole simultanée* (en présence et absence de bruit). Le SNR est évalué uniquement dans les deux situations de *parole locale* et *parole simultanée* en présence de bruit. Le SER est évalué uniquement sur dans la situation de *parole simultanée*. L'ERLE est évalué dans les trois situations de *parole locale*, *parole distante* et de *parole simultanée* (en présence et absence de bruit). Le tableau 4.3 résume la correspondance entre les types de distorsion présents et les situations considérées.

Puisque les performances peuvent varier en fonction de la présence de l'écho acoustique qui est le signal le plus fort, et aussi en fonction de la présence de bruit, nous calculons les métriques séparément dans chaque situation, en présence et en absence de bruit : *parole locale seule*, *parole simultanée* (paroles locale et distante actives simultanément) et *parole distante seule*. En particulier, les métriques dépendent de l'estimation

Numérotation	**Situation**	**Type de distorsion présent**
1	*parole locale*	réverbération
2	*parole simultanée*	réverbération + écho
3	*parole locale + bruit*	réverbération + bruit
4	*parole simultanée + bruit*	réverbération + écho + bruit
5	*parole distante seule*	écho

TABLEAU 4.3. – Correspondance entre les types de distorsion présents et les situations considérées.

d'un facteur d'échelle γ_c associé au signal c et défini comme dans (2.90) :

$$\gamma_c = \frac{\langle \widehat{s}_\mathrm{e}, c \rangle}{\|c\|^2}. \tag{4.58}$$

Dans chaque situation de l'ensemble de test (voir le tableau 4.3), nous supposons que γ_c est fixe. Toutefois, nous supposons que γ_c peut varier d'une situation à l'autre. Nous faisons ensuite la moyenne pondérée pour chaque métrique en fonction de la durée de chaque période sur laquelle sont calculées ces métriques, de la même manière que le calcul du SNR segmenté [Vincent et al., 2006].

Écho	ERLE	$10 \log_{10} \frac{\|y\|^2}{\|z_\mathrm{r}^\mathrm{post}\|^2}$	situations 2, 4, 5
	SER	$10 \log_{10} \frac{\|s_\mathrm{e}^\mathrm{post}\|^2}{\|z_\mathrm{r}^\mathrm{post}\|^2}$	situations 2, 4
Réverbération	ELR	$10 \log_{10} \frac{\|s_\mathrm{e}^\mathrm{post}\|^2}{\|s_\mathrm{r}^\mathrm{post}\|^2}$	situations 1, 2, 3, 4
Bruit	SNR	$10 \log_{10} \frac{\|s_\mathrm{e}^\mathrm{post}\|^2}{\|b_\mathrm{r}^\mathrm{post}\|^2}$	situations 3, 4
Artefacts	SI-SAR	$10 \log_{10} \frac{\|s_\mathrm{e}^\mathrm{post}\|^2}{\|s_\mathrm{e}^\mathrm{art}\|^2}$	situations 1, 2, 3, 4
Distorsion globale	SI-SDR	$10 \log_{10} \frac{\|s_\mathrm{e}^\mathrm{post}\|^2}{\|s_\mathrm{r}^\mathrm{post} + z_\mathrm{r}^\mathrm{post} + b_\mathrm{r}^\mathrm{post} + s_\mathrm{e}^\mathrm{art}\|^2}$	situations 1, 2, 3, 4

TABLEAU 4.4. – Métriques d'évaluation. Les formules sont données dans le cas monophonique ($M = 1$) et l'indice m du microphone est omis par souci de clarté.

D'après la formule $c^\mathrm{post} = \gamma_c\, c$, les composantes $\mathbf{s}_\mathrm{e}^\mathrm{post}$, $\mathbf{s}_\mathrm{r}^\mathrm{post}$, $\mathbf{z}_\mathrm{r}^\mathrm{post}$, $\mathbf{b}_\mathrm{r}^\mathrm{post}$ et $\mathbf{s}_\mathrm{e}^\mathrm{art}$ sont calculées à partir des vérités terrain des signaux $\mathbf{s}_\mathrm{e}$, $\mathbf{s}_\mathrm{l}$, $\mathbf{y}$ et $\mathbf{b}$ (voir la partie 2.4). La procédure de génération des données fournit immédiatement les vérités terrain de

l'écho $\mathbf{y}$ et du signal de bruit $\mathbf{b}$. Pour définir la parole cible $\mathbf{s}_\mathrm{e}$ et la vérité terrain de la réverbération tardive $\mathbf{s}_\mathrm{l}$, nous fixons le *temps de mélange* $t_\mathrm{e} = 64$ ms (voir la partie 2.1.1.2). Nous calculons ces deux composantes à l'aide de (2.5). Dans l'ensemble de test, comme la vérité terrain de la RIR de la parole locale $\mathbf{a}_s(\tau)$ est inconnue, nous utilisons la méthode de Yoshioka et al. [2011], qui estime la RIR de la parole locale $\mathbf{a}_s(\tau)$ en optimisant le critère MMSE entre la parole locale $\mathbf{s}$ et la parole locale anéchoïque u, comme pour déterminer le filtre d'annulation d'écho $\mathcal{H}(f)$ en réduction d'écho (voir la partie 2.2.2.1).

Nous évaluons aussi la qualité perçue la parole cible estimée $\widehat{\mathbf{s}}_\mathrm{e}$ à l'aide d'un test ABX, qui est un test de préférence entre deux signaux [Munson et Gardner, 1950]. Pour cela, nous sélectionnons 21 enregistrements dont la parole cible estimée $\widehat{\mathbf{s}}_\mathrm{e}$ par chacune des méthodes de référence possède un SI-SDR positif et est intelligible d'après notre écoute. Pour chaque enregistrement, nous faisons évaluer les estimations de la parole cible $\widehat{\mathbf{s}}_\mathrm{e}$ par deux méthodes de référence par 40 auditeurs, qui doivent choisir s'ils préfèrent la première ou la seconde estimation, ou bien s'ils n'ont pas de préférence entre les deux. Les auditeurs sont des personnes de notre lieu de travail ou de notre entourage à qui nous avons transmis un lien web afin qu'ils aient accès au test. Nous fournissons aux auditeurs la vérité terrain de la parole cible $\mathbf{s}_\mathrm{e}$, ainsi que le mélange $\mathbf{d}$, qui servent de signaux de référence.

4.3.4. Méthodes de référence

Nous comparons notre approche hors-ligne avec deux méthodes de l'état de l'art : 1) notre implémentation de l'approche de Togami et Kawaguchi [2014] et 2) une approche en cascade où le filtre d'annulation d'écho $\mathcal{H}(f)$, le filtre de déréverbération $\mathcal{G}(f)$ et le post-filtre court $\mathbf{W}_{s_\mathrm{e}}(n, f)$ sont estimés indépendamment et appliqués l'un après l'autre. L'annulation d'écho est réalisée avec SpeexDSP [1], qui est une implémentation de l'approche adaptative de Valin [2007] (voir la partie 2.2.2). La déréverbération linéaire est réalisée avec notre implémentation de la méthode WPE de Yoshioka et Nakatani [2012] (voir la partie 2.2.3.2). Le post-filtre court $\mathbf{W}_{s_\mathrm{e}}(n, f)$ est estimé en utilisant notre implémentation de l'algorithme DNN-EM de Nugraha et al. [2016a] (voir la partie 2.2.1.5).

4.3.5. Réglage des hyperparamètres

Le réglage des hyperparamètres des trois approches est décrit dans les sous-parties suivantes.

4.3.5.1. Initialisation des filtres linéaires

Pour l'annulation d'écho, nous calculons $\mathcal{H}^{(0)}(f)$ en appliquant SpeexDSP sur chaque canal m du mélange $\mathbf{d}(n, f)$, car SpeexDSP est conçu pour le traitement monocanal. Puisque SpeexDSP est basé sur des fenêtres de TFCT rectangulaires avec un chevauchement à 50%, nous utilisons une taille de trame $T_\mathrm{TFCT} = 512$ échantillons et un pas

1. https://github.com/xiph/speexdsp

d'avancement $P = 256$ échantillons. Nous fixons la longueur du filtre d'annulation d'écho à $0,208$ s dans le domaine temporel, soit $K = 13$ trames. Cette configuration permet un compromis entre réduction d'écho, et complexité de l'algorithme. SpeexDSP donne le signal de sortie $\mathbf{e}(t)$ dans le domaine temporel. Pour obtenir un filtre initial d'annulation d'écho $\mathcal{H}^{(0)}(f)$ invariant au cours du temps avec l'algorithme en ligne SpeexDSP, nous appliquons celui-ci 2 fois à chaque enregistrement afin d'obtenir la convergence de l'algorithme et ainsi approcher le fonctionnement d'un algorithme hors ligne.

Pour la déréverbération linéaire, nous calculons $\mathcal{G}^{(0)}(f)$ en réalisant 3 itérations de WPE sur le signal après annulation d'écho $\mathbf{e}(n, f)$. Pour cette opération, nous utilisons la TFCT avec une fenêtre de Hanning de taille $T_{\text{TFCT}} = 1\,024$ échantillons et un pas d'avancement $P = 256$ échantillons. Nous fixons la longueur du filtre de déréverbération à $0,208$ s dans le domaine temporel, soit $L = 10$ trames, avec un délai $\Delta = 3$ trames. Cette configuration permet un compromis entre déréverbération et complexité de l'algorithme.

4.3.5.2. Hyperparamètres des DNNs

Nous choisissons 1 026 neurones pour la couche cachée de l'architecture LSTM. L'apprentissage des DNNs est réalisé par rétropropagation avec une taille de mini-lots de 16 séquences, une taille de séquence de 32 trames et l'algorithme Adam avec les réglages par défaut pour l'optimisation [Kingma et Ba, 2015]. Pour éviter l'explosion du gradient avec les longues séquences, nous utilisons l'écrêtage du gradient avec un seuil à $1,0$. L'apprentissage est arrêté lorsque la fonction de coût de l'ensemble de validation ne diminue plus après 5 cycles.

4.3.5.3. Hyperparamètres de l'algorithme DNN-BCA

Pour l'algorithme DNN-BCA, la TFCT est calculée avec une fenêtre de Hanning de longueur $T_{\text{TFCT}} = 1\,024$ échantillons et un pas d'avancement $P = 256$ échantillons, pour obtenir $F = 513$ bandes de fréquence. La longueur du filtre d'annulation d'écho $\mathcal{H}^{(i)}(f)$ ($0,208$ s dans le domaine temporel) correspond ici à $K = 10$ trames. Les hyperparamètres du filtre de déréverbération $\mathcal{G}^{(i)}(f)$ sont identiques à ceux de $\mathcal{G}^{(0)}(f)$. Pour déterminer les cibles des DNNs durant l'apprentissage, nous réalisons 3 itérations de la procédure de détermination des vérités terrain des DSPs (voir la partie 4.2.4.1). Nous réalisons l'apprentissage de DNN_0, DNN_1 et DNN_2. Durant la phase de test, cela signifie que nous réalisons $I = 3$ itérations de l'algorithme DNN-BCA avec 1 mise à jour spectrale et 1 mise à jour spatiale à chaque itération i (voir la partie 4.2.3.2).

4.3.5.4. Hyperparamètres de l'approche conjointe de Togami et Kawaguchi [2014]

L'approche de Togami et Kawaguchi [2014] nécessite les valeurs initiales des filtres linéaires $\mathcal{H}(f)$ et $\mathcal{G}(f)$, et celles des DSPs de la parole locale $v_s(n, f) = \frac{1}{M}\|\mathbf{s}(n, f)\|^2$ et du signal de bruit $v_b(n, f) = \frac{1}{M}\|\mathbf{b}(n, f)\|^2$. Nous initialisons $\mathcal{H}(f)$ et $\mathcal{G}(f)$ en appliquant SpeexDSP (annulation d'écho) et WPE (déréverbération linéaire) sur le mélange $\mathbf{d}(n, f)$, respectivement, avec les mêmes hyperparamètres que pour l'algorithme DNN-BCA (voir

la partie 4.3.5.1). Puisque les auteurs ne spécifient pas comment initialiser les DSPs $v_c(n,f)$, nous les estimons à l'aide d'un DNN similaire à DNN_0 où l'entrée de type I $|\widetilde{e}_1(n,f)|$ est remplacée par $|\widetilde{d}_1(n,f)|$, qui est obtenue de la même manière que (4.50) à partir du signal multicanal $\widehat{\mathbf{d}}_1(n,f) = \sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\mathbf{d}(n-l,f)$ (voir la figure 4.2). Toutes les MCSs $\mathbf{R}_c(f)$ sont initialisées à la matrice identité $\mathbf{I}_M$. Nous utilisons les mêmes hyperparamètres et valeurs de K, L et Δ que dans notre approche. Nous réalisons 3 itérations de l'algorithme EM de Togami et Kawaguchi [2014].

4.3.5.5. Hyperparamètres de l'approche en cascade

Concernant l'approche en cascade, nous calculons et fixons les filtres linéaires $\mathcal{H}(f) = \mathcal{H}^{(0)}(f)$ et $\mathcal{G}(f) = \mathcal{G}^{(0)}(f)$ avec les mêmes hyperparamètres que pour l'approche proposée. L'architecture et les entrées des DNNs sont identiques à celles de notre approche, et la vérité terrain des DSPs est calculée avec la même procédure, où les filtres linéaires sont fixés à $\mathcal{H}(f) = \mathcal{H}^{(0)}(f)$ et $\mathcal{G}(f) = \mathcal{G}^{(0)}(f)$ (voir la partie 4.2.4.1). Il convient de remarquer que les valeurs des entrées de type I $|\widetilde{y}(n,f)|$, $|\widetilde{e}(n,f)|$, $|\widetilde{e}_1(n,f)|$ et $|\widetilde{r}(n,f)|$ restent fixes durant les itérations de l'algorithme DNN-EM à cause des filtres linéaires fixés.

4.3.5.6. Régularisation

Afin d'éviter des instabilités numériques et des matrices mal conditionnées, nous ajoutons un scalaire de régularisation ϵ au dénominateur dans (4.41) et une matrice de régularisation $\epsilon\mathbf{I}$ à la matrice à inverser dans (4.21), (4.29) et (4.34). Nous régularisons aussi la fonction de coût d'apprentissage $\mathcal{D}_{KL}$ dans (4.56) [Nugraha et al., 2016a] :

$$\mathcal{D}_{KL} = \frac{1}{4FN} \sum_{c,n,f} \left(\left(\sqrt{v_c(n,f)} + \epsilon \right) \log \frac{\sqrt{v_c(n,f)} + \epsilon}{|\widetilde{c}(n,f)| + \epsilon} - \sqrt{v_c(n,f)} + |\widetilde{c}(n,f)| \right). \quad (4.59)$$

Nous régularisons l'approche conjointe de Togami et Kawaguchi [2014] et l'approche en cascade de la même manière. Le paramètre de régularisation est fixé à $\epsilon = 10^{-5}$.

4.4. Résultats et discussion

Dans cette partie, nous comparons l'approche proposée pour la réduction conjointe de bruit, d'écho et de réverbération à l'approche conjointe de Togami et Kawaguchi [2014] et l'approche en cascade. Nous analysons tout d'abord les résultats des trois approches dans des conditions acoustiques invariantes dans le temps. Enfin, nous discutons de leurs performances lorsque les conditions acoustiques varient au cours du temps et nous comparons leur temps de calcul. Des exemples audio sont disponibles en ligne [2].

2. https://team.inria.fr/multispeech/demos

4.4.1. Conditions invariantes dans le temps

4.4.1.1. Performances moyennes

La figure 4.10 présente la moyenne des performances de l'approche proposée et des approches de référence dans des conditions acoustiques invariantes dans le temps. Toutes les approches ont un SI-SDR proche de 0 dB ou négatif, en raison des conditions difficiles de l'ensemble de test. L'approche proposée obtient le meilleur SI-SDR qui est de plus le seul positif.L'approche proposée dépasse l'appproche en cascade d'environ 2 dB en SI-SDR . La performance en SI-SDR par rapport à l'approche en cascade s'explique par le fait que l'approche proposée est meilleure que l'approche en cascade selon toutes les autres métriques.

L'approche proposée dépasse aussi l'approche de Togami et Kawaguchi [2014] d'environ 4 dB en SI-SDR. Bien que l'approche proposée réduise beaucoup moins l'écho (ERLE plus faible), la réverbération (ELR plus faible) et le bruit (SNR plus faible) que l'approche de Togami et Kawaguchi [2014], les performances de l'approche proposée en SI-SDR s'explique par le fait qu'elle introduit beaucoup moins de dégradations dans la parole cible $\mathbf{s}_{\mathrm{e}}(n, f)$ que l'approche de Togami et Kawaguchi [2014] (différence en SI-SAR d'environ $5,5$ dB). Par ailleurs, malgré une plus faible réduction d'écho, l'approche proposée obtient un meilleur SER car la parole cible $\mathbf{s}_{\mathrm{e}}(n, f)$ moins dégradée. Dans le cas de l'approche de Togami et Kawaguchi [2014], la dégradation de la parole cible $\mathbf{s}_{\mathrm{e}}(n, f)$ est due à l'application des filtres linéaires $\mathcal{H}(f)$ et $\mathcal{G}(f)$ en parallèle (voir la figure 4.2), ainsi qu'aux problèmes de modélisation des signaux dans le signal $\mathbf{r}(n, f)$ (voir la partie 4.1). En effet, nous avons remarqué que l'écho résiduel *déréverbéré* $\mathbf{z}_{\mathrm{r}}$ dans le signal $\mathbf{r}$ était systématiquement plus élevé avec leur approche qu'avec l'approche proposée et l'approche en cascade.

La figure 4.11 illustre un exemple de spectrogramme de la parole cible estimée $\widehat{\mathbf{s}}_{\mathrm{e}}(n, f)$ par l'approche proposée et les approches de référence. Nous pouvons remarquer que l'énergie de la parole cible $\mathbf{s}_{\mathrm{e}}(n, f)$ est faible par rapport à celle de l'écho $\mathbf{y}(n, f)$. Nous pouvons constater que l'approche de Togami et Kawaguchi [2014] dégrade de manière importante la parole cible $\mathbf{s}_{\mathrm{e}}(n, f)$. L'approche proposée parvient à mieux estimer la parole cible $\mathbf{s}_{\mathrm{e}}(n, f)$ en période de *parole simultanée + bruit*.

4.4.1.2. Interactions des composantes du système

Tandis que les résultats ci-dessus présentent la moyenne des performances sur toutes les situations (*parole locale seule*, *parole distante seule*, *parole simultanée*, en présence ou en absence de bruit), nous souhaitons analyser les interactions entre les différents filtres des systèmes, ainsi que les performances sur différents cas d'usage. Pour cela, nous avons besoin d'une analyse plus approfondie des performances lorsque seule la réverbération est présente, lorsque seul le bruit et la réverbération sont présents, lorsque seul l'écho et la réverbération sont présents, et enfin lorsque le bruit, l'écho et la réverbération sont présents simultanément. À titre de rappel, le tableau 4.3 résume la correspondance entre les types de distorsion présents et les situations considérées. Comme nous cherchons à transmettre la parole cible $\mathbf{s}_{\mathrm{e}}(n, f)$, nous ne considérons ici que l'analyse durant les

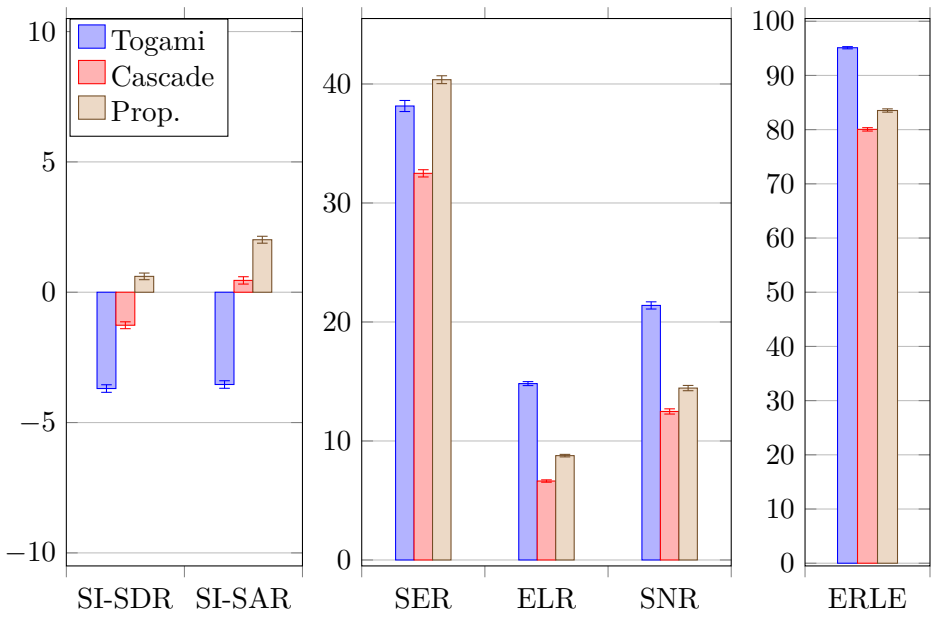
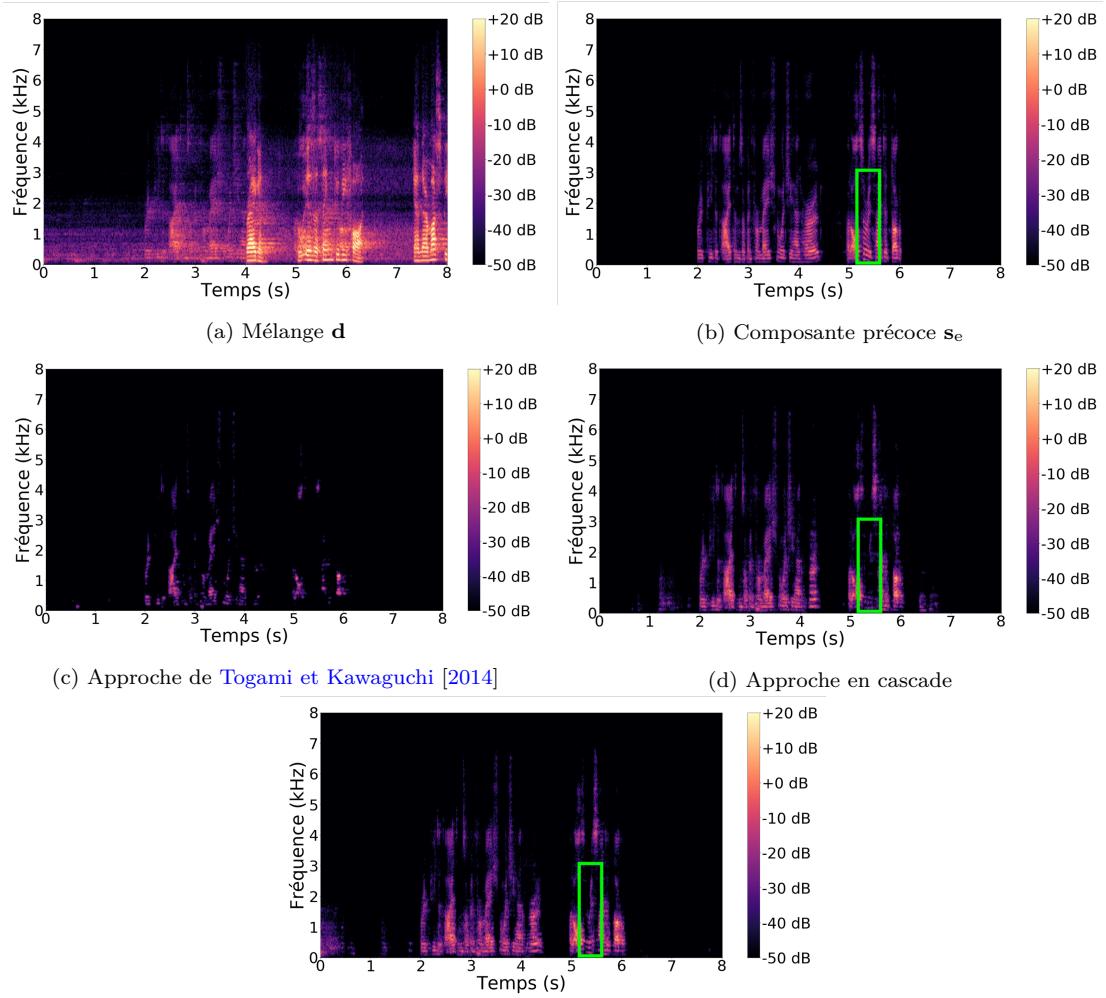


FIGURE 4.10. – Performances moyennes (en dB) des trois approches en conditions acoustiques invariantes dans le temps.

périodes de *parole locale seule* et *parole simultanée*, en présence et en absence de bruit. Cela exclut les situations de *parole distante*, *bruit seul*, et *parole distante + bruit*.

Interactions en présence de bruit La figure 4.12a présente les résultats dans la situation de *parole locale + bruit*. Les SER et ERLE ne sont pas évalués car l'écho $\mathbf{y}(n, f)$ est absent. Toutes les approches ont un SI-SDR positif ou proche de 0 dB. Bien que l'approche proposée obtienne un meilleur SNR que l'approche en cascade, l'approche proposée est seulement légèrement meilleure en SI-SDR et ELR que l'approche en cascade, et obtient un SI-SAR équivalent. Par conséquent, l'optimisation conjointe n'améliore que faiblement les performances en période de *parole locale + bruit*. Il convient d'ajouter qu'elle ne dégrade pas les performances. L'approche de Togami et Kawaguchi [2014] dépasse l'approche proposée en ELR et SNR, mais est dépassée de manière importante par l'approche proposée en SI-SDR. Ces performances en matière de distorsion globale s'expliquent par le fait que l'approche proposée dégrade beaucoup moins la parole cible $\mathbf{s}_{\mathrm{e}}(n, f)$ que l'approche de Togami et Kawaguchi [2014] (SI-SAR plus élevé). Les causes de cette dégradation importante de la parole locale ont été évoquées dans la partie 4.4.1.1.

La figure 4.12b présente les résultats dans la situation de *parole simultanée + bruit*. Les tendances sont similaires à celles observées sur la moyenne des performances sur toutes les situations (voir la figure 4.10). Toutefois, toutes les approches obtiennent des SI-SDR et SI-SAR négatifs, ce qui traduit des dégradations significatives de la parole cible $\mathbf{s}_{\mathrm{e}}(n, f)$.



(a) Mélange $\mathbf{d}$

(b) Composante précoce $\mathbf{s}_{\mathrm{e}}$

(c) Approche de Togami et Kawaguchi [2014]

(d) Approche en cascade

(e) Approche proposée

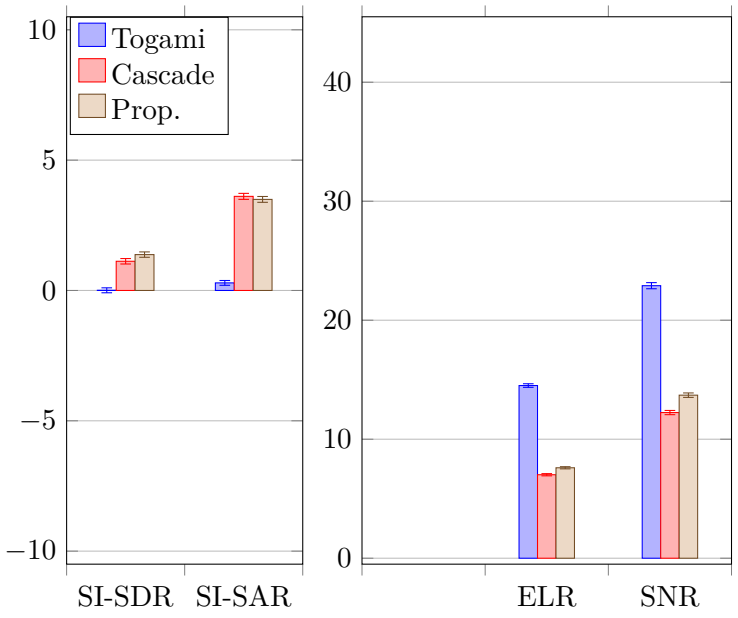
FIGURE 4.11. – Exemple de spectrogrammes de la composante précoce estimée $\widehat{\mathbf{s}}_{\mathrm{e}}$ avec les approches de référence et l'approche proposée, pour le scénario où les conditions acoustiques sont invariantes dans le temps (seul un canal est illustré). Le rectangle vert montre une zone du spectrogramme dont l'estimation est améliorée entre l'approche en cascade et l'approche proposée.

Entre les situations de *parole locale + bruit* et *parole simultanée + bruit*, les SI-SDR et SI-SAR décroissent de $5,1$ dB et $5,8$ dB, respectivement, pour l'approche proposée, de $6,1$ dB et $7,7$ dB, respectivement, pour l'approche en cascade, et de $8,7$ dB et $8,8$ dB, respectivement, pour l'approche de Togami et Kawaguchi [2014]. De plus, l'approche proposée est la meilleure en SI-SDR dans les deux situations de *parole locale + bruit* et *parole simultanée + bruit*, qui est la situation la plus difficile. Nous concluons que l'approche proposée d'optimisation conjointe améliore la robustesse en SI-SDR et SI-SAR lorsque l'écho, la réverbération et le bruit sont présents simultanément, tout en limitant la baisse de performances par rapport aux autres approches lorsque seulement la réverbération et le bruit sont présents.

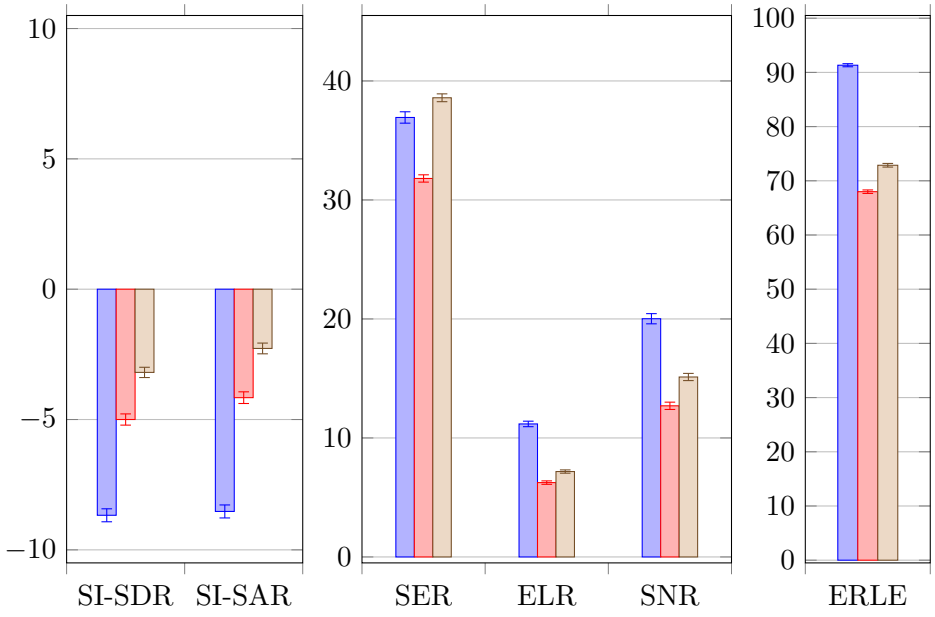
Interactions en absence de bruit La figure 4.13a présente les résultats dans la situation de *parole locale*. Les SER, SNR et ERLE ne sont pas évalués car le bruit $\mathbf{b}(n,f)$ et l'écho $\mathbf{y}(n,f)$ sont absents. Toutes les approches ont un SI-SDR positif. L'approche proposée obtient de bien meilleures performances que l'approche en cascade dans toutes les métriques (SI-SDR, SI-SAR et ELR). L'optimisation conjointe améliore nettement les performances en période de *parole locale seule* + absence de bruit par rapport à la période de *parole locale seule* + présence de bruit. L'approche de Togami et Kawaguchi [2014] dépasse l'approche proposée en ELR, mais est nettement dépassée par l'approche proposée en SI-SDR. Ces performances en matière de distorsion globale s'explique par le fait que l'approche proposée dégrade beaucoup moins la parole cible $\mathbf{s}_{\mathrm{e}}(n,f)$ que l'approche de Togami et Kawaguchi [2014] (SI-SAR plus élevé). Les causes de cette dégradation importante de la parole locale ont été évoquées dans la partie 4.4.1.1.

La figure 4.13b présente les résultats dans la situation de *parole simultanée*. Les tendances sont similaires à celles observées sur la moyenne des performances sur toutes les situations (voir la figure 4.10). Toutefois, l'écart de performances entre l'approche proposée et l'approche en cascade est plus élevé que pour la situation de *parole simultanée* + présence de bruit. L'écart de performances en SI-SDR, SI-SAR et SER est aussi plus élevé entre l'approche proposée et l'approche de Togami et Kawaguchi [2014] que pour la situation de *parole simultanée* + présence de bruit. L'approche proposée est la seule à obtenir un SI-SAR positif.

En absence de bruit, l'écart des performances entre l'approche proposée et l'approche en cascade est plus élevée qu'en présence de bruit. Une possible explication serait que les DNNs de chaque approche ont été entraînées seulement dans le scénario en présence de bruit. Or, en absence de bruit, les filtres linéaires $\mathcal{H}(f)$ et $\mathcal{G}(f)$ se comportent différemment qu'en présence de bruit. Puisque les entrées des DNNs sont calculées à partir des filtres $\mathcal{H}(f)$ et $\mathcal{G}(f)$, cela perturberait l'estimation des DNNs de l'approche en cascade. À l'inverse, dans l'approche proposée, l'estimation des DNNs ne serait pas perturbée car les DNNs optimisent conjointement les filtres $\mathcal{H}(f)$ et $\mathcal{G}(f)$. Nous concluons que l'approche proposée d'optimisation conjointe améliore la robustesse en SI-SDR et SI-SAR lorsque l'écho, la réverbération et le bruit sont présents simultanément, en améliorant les performances lorsque seulement la réverbération et l'écho sont présents.

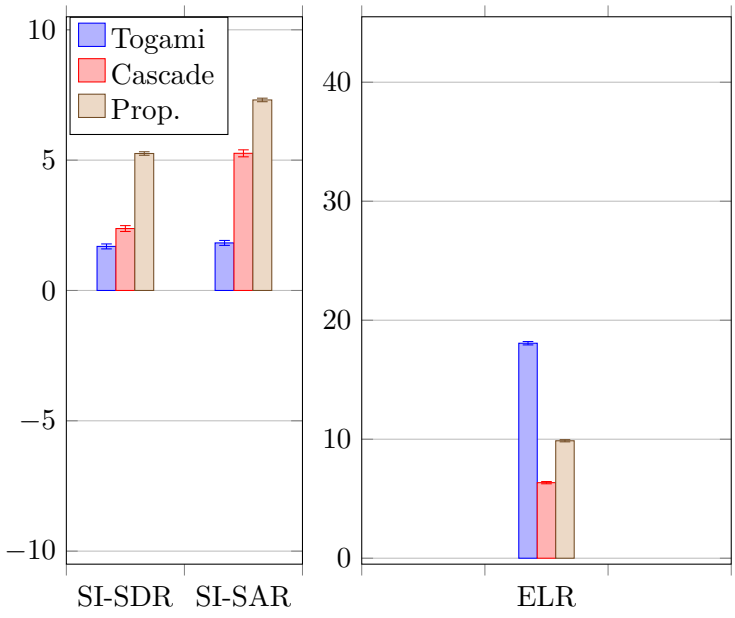


(a) *Parole locale + bruit.*

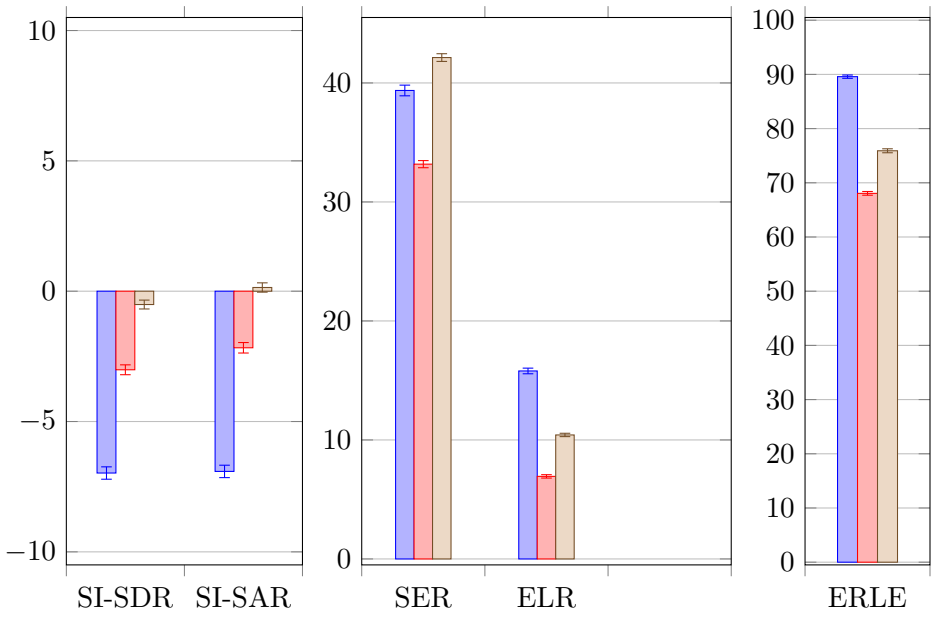


(b) *Parole simultanée + bruit.*

FIGURE 4.12. – Analyse des performances (en dB) en conditions acoustiques invariantes dans le temps, en présence de bruit ambiant.



(a) *Parole locale.*



(b) *Parole simultanée.*

FIGURE 4.13. – Analyse des performances (en dB) en conditions acoustiques invariantes dans le temps, en absence de bruit ambiant.

4.4.1.3. Test d'écoute

Tandis que les résultats ci-dessus présentent les métriques basées sur les rapports d'énergie, nous souhaitons maintenant analyser la qualité perceptuelle des différentes méthodes de référence avec un test ABX. D'après les résultats précédents, l'approche de Togami et Kawaguchi [2014] dégrade la parole cible $\mathbf{s}_{\mathrm{e}}(n, f)$ de manière plus importante que l'approche proposée et l'approche en cascade. Par conséquent, nous ne considérons ici que la comparaison entre l'approche proposée et l'approche en cascade.

Le tableau 4.5 présente le nombre de votes obtenus pour les choix suivants : *approche proposée*, *approche en cascade* et *pas de préférence*. Nous constatons qu'environ 40% des votes ont été soit pour l'approche en cascade, soit sans préférence. Environ 20% des votes ont été pour l'approche proposée. Afin de savoir si la différence entre l'approche proposée et l'approche en cascade est significative, il nous faut savoir comment traiter le choix *pas de préférence* par rapport aux deux autres choix. Pour cela, nous utilisons l'hypothèse suivante [Eskénazi et al., 2013, Chapitre 7] : chaque moitié des votes pour *pas de préférence* est assignée aux deux choix *approche proposée* et *approche en cascade* (voir le tableau 4.6). Cette hypothèse est justifiée par le fait que si nous n'avions pas mis le choix *pas de préférence*, les votes auraient été attribués au hasard entre *approche proposée* et *approche en cascade*. D'après le calcul des statistiques de test de préférence, nous obtenons une p-valeur $\approx 8,1 \cdot 10^{-7}$. La qualité perceptuelle de l'approche en cascade est donc significativement meilleure que l'approche proposée pour les enregistrements proposés. À titre de rappel, nous avions sélectionné les échantillons dont les estimations par chacune des deux approches possédaient un SI-SDR positif et étaient intelligibles d'après notre écoute (voir la partie 4.3.3). Le SI-SDR moyen de la parole cible estimée $\widehat{\mathbf{s}}_{\mathrm{e}}$ sur les 21 échantillons était de 2,8 dB pour l'approche en cascade et de 4,1 dB. Nous concluons que pour ce type d'estimation, le SI-SDR n'est pas corrélé à la qualité perçue.

Préférence	**Nombre de votes**	**% Total**
Approche proposée	181	21,5
Approche en cascade	324	38,6
Pas de préférence	335	39,9
Total	**840**	**100,0**

TABLEAU 4.5. – Résultats du test ABX.

4.4.2. Conditions variant au cours du temps

La figure 4.14 présente la moyenne des résultats lorsque les conditions acoustiques varient au cours du temps. Toutes les approches ont un SI-SDR négatif. Le SI-SDR diminue pour toutes les approches, car les propriétés spatiales de la parole cible $\mathbf{s}_{\mathrm{e}}(n, f)$ et de la réverbération résiduelle $\mathbf{s}_{\mathrm{r}}(n, f)$ varient au cours du temps, alors que leurs MCSs $\mathbf{R}_c(f)$ restent fixes.

Préférence	**Nombre de votes**	**% Total**	**Différence significative**
Approche proposée	$348,5$	$41,5$	**Oui**
Approche en cascade	$491,5$	$58,5$	(p-valeur $\approx 8,1 \cdot 10^{-7}$)
Total	**840**	**100,0**	-

TABLEAU 4.6. – Résultats du test ABX après l'hypothèse de répartition du choix *pas de préférence*.

Les tendances sont les mêmes que pour la moyenne des performances dans des conditions invariantes dans le temps (voir la figure 4.10), sauf en SER et ERLE. L'approche proposée est nettement dépassée par l'approche en cascade en SER et ERLE, ce qui traduit une réduction d'écho moins importante. Ceci s'explique par le fait que SpeexDSP est appliqué 2 fois à chaque enregistrement (voir la partie 4.3.5.1), dont la durée est ici de 16 s, au lieu de 8 s pour l'ensemble de données où les conditions acoustiques sont invariantes. Nous constatons que SpeexDSP réduit moins l'écho $\mathbf{y}(n,f)$ sur les 8 premières secondes d'un enregistrement de 16 s que pour un enregistrement de 8 s où les conditions acoustiques sont invariantes. L'écho résiduel *déréverbéré* $\mathbf{z}_{\mathrm{r}}(n,f)$ est alors plus fort. Le post-filtre $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$ de l'approche en cascade va alors être plus agressif en réduction d'écho que le post-filtre $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$ de l'approche proposée. Toutefois, il va potentiellement introduire plus de dégradations de la parole cible $\mathbf{s}_{\mathrm{e}}(n,f)$ que pour l'approche proposée. Cette hypothèse est confirmée en observant que l'approche proposée obtient de meilleurs SI-SDR et SI-SAR que l'approche en cascade. Cela suggère donc que l'approche en cascade est très agressive en réduction d'écho, au point de dégrader significativement la parole cible $\mathbf{s}_{\mathrm{e}}(n,f)$. Il convient d'ajouter que l'approche en cascade dépasse aussi l'approche de Togami et Kawaguchi [2014] en SER.

La figure 4.15 illustre un exemple de spectrogramme de la parole cible estimée $\widehat{\mathbf{s}}_{\mathrm{e}}(n,f)$ par l'approche proposée et les approches de référence. Nous pouvons constater que l'approche de Togami et Kawaguchi [2014] dégrade de manière importante la parole cible $\mathbf{s}_{\mathrm{e}}(n,f)$. L'approche proposée et l'approche en cascade obtiennent une estimation équivalente de la parole cible $\mathbf{s}_{\mathrm{e}}(n,f)$.

4.4.3. Temps de calcul

Nous évaluons le temps de calcul des trois approches. Nous ne considérons pas l'initialisation des filtres linéaires $\mathcal{H}(f)$ et $\mathcal{G}(f)$ car elle représente la même opération pour les trois approches. Le tableau 4.7 détaille le temps d'estimation de la parole cible $\widehat{\mathbf{s}}_{\mathrm{e}}$ sur un enregistrement de 8 s, avec un processeur Intel Core i5 à $2,7$ GHz. Comme l'estimation des DSPs $v_c(n,f)$ avec les DNNs dépend de l'architecture, nous décrivons aussi le temps de calcul propre aux algorithmes de mises à jour des filtres en excluant le temps d'estimation des DSPs $v_c(n,f)$ avec les DNNs, et le temps de calcul propre à l'estimation des DSPs $v_c(n,f)$ avec les DNNs. En excluant l'estimation des DSPs $v_c(n,f)$ avec les DNNs,

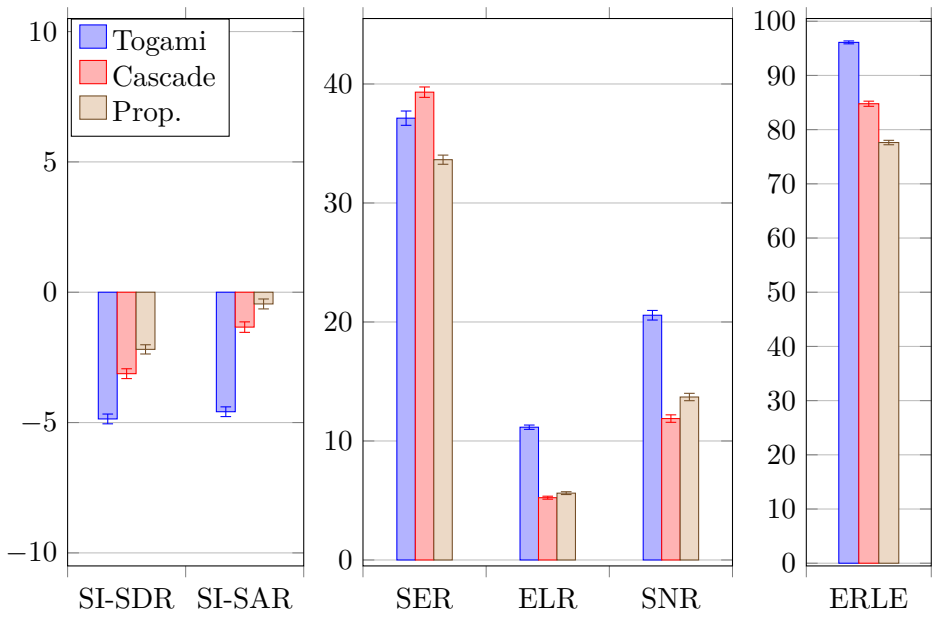
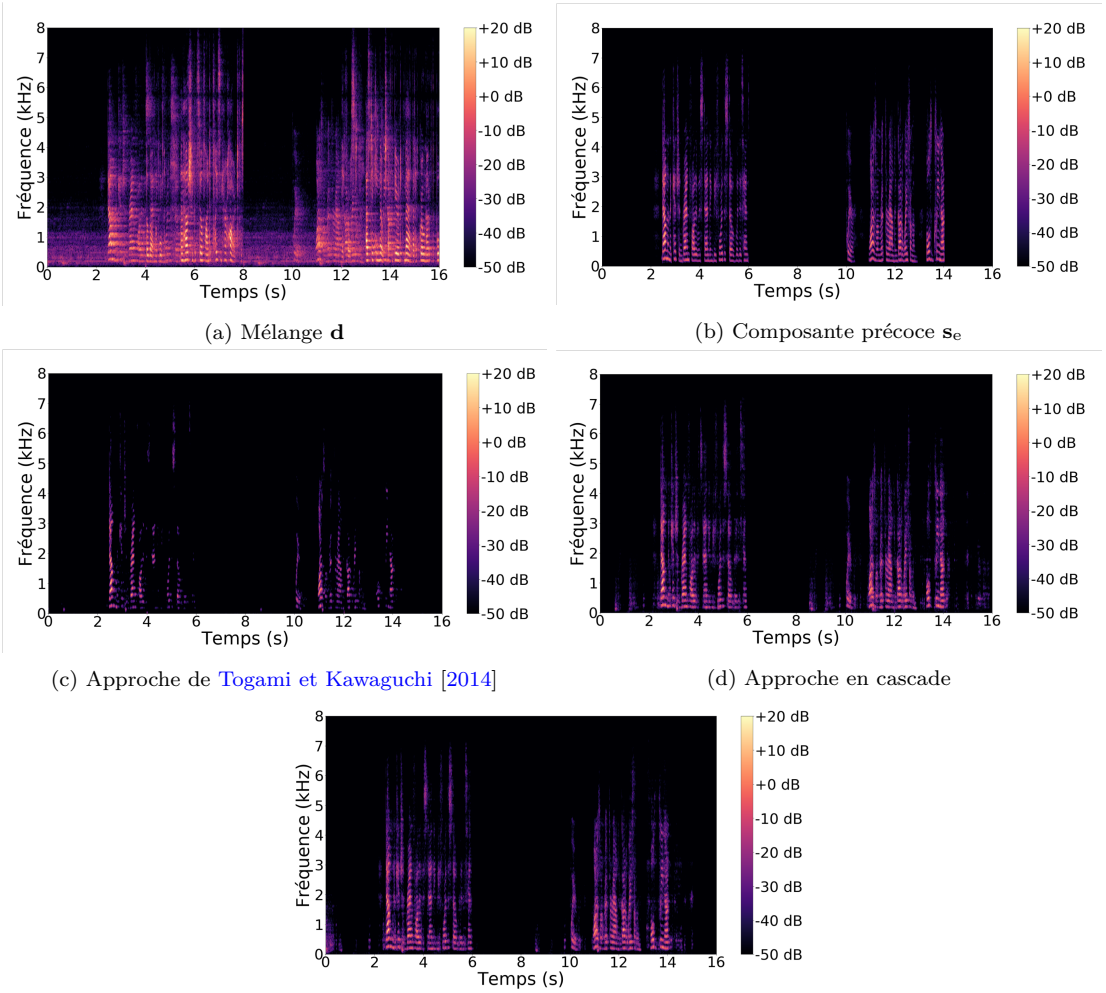


FIGURE 4.14. – Analyse des performances (en dB) en conditions acoustiques qui varient au cours du temps.

l'approche proposée prend environ 3 fois plus de temps que l'approche en cascade, car les filtres linéaires $\mathcal{H}(f)$ et $\mathcal{G}(f)$ doivent être mis à jour à chaque itération de l'algorithme. L'approche de Togami et Kawaguchi [2014] prend environ 2 fois plus de temps que l'approche proposée, car elle estime les filtres linéaires $\mathcal{H}(f)$ et $\mathcal{G}(f)$ comme un filtre unique. L'opération d'inversion matricielle associée, similaire à (4.29), augmente alors la complexité de l'algorithme d'optimisation par rapport à celui de l'approche proposée.

En ce qui concerne l'estimation des DSPs $v_c(n, f)$ avec les DNNs, l'approche proposée a le même temps de calcul que l'approche en cascade, car les DNNs ont le même nombre de paramètres, et les approches utilisent le même nombre de DNNs. L'approche de Togami est plus de 5 fois plus rapide, car elle n'utilise qu'un seul DNN (à l'initialisation) qui a de plus moins de paramètres (seulement 2 DSPs à estimer au lieu de 4). Le tableau 4.8 détaille le nombre de paramètres des DNNs dans chaque approche.

En prenant en compte l'estimation des DSPs $v_c(n, f)$ avec les DNNs, et par rapport à l'approche en cascade, l'estimation de la parole cible $\mathbf{s}_e$ prend +12% de temps pour l'approche proposée, et −45% de temps pour l'approche de Togami et Kawaguchi [2014]. Puisque l'approche en cascade fait partie de l'une des approches implémentées dans les appareils industriels actuels (sans tenir compte de cette architecture de DNN en particulier), nous concluons que l'approche proposée pourrait être implémentée en temps réel.



(a) Mélange $\mathbf{d}$

(b) Composante précoce $\mathbf{s}_{\mathrm{e}}$

(c) Approche de Togami et Kawaguchi [2014]

(d) Approche en cascade

(e) Approche proposée

FIGURE 4.15. – Exemple de spectrogrammes de la composante précoce estimée $\widehat{\mathbf{s}}_{\mathrm{e}}$ avec les approches de référence et l'approche proposée, pour le scénario où les conditions acoustiques varient au cours du temps (seul un canal est illustré).

Élement de l'algorithme	**Togami**	**Cascade**	**Prop.**
Mises à jour des filtres (DNNs exclus)	$46,8 \pm 0,5$	$8,0 \pm 0,1$	$22,0 \pm 1,0$
DNNs	$19,5 \pm 0,2$	$110,3 \pm 1,7$	$110,4 \pm 1,5$
Total	$66,2 \pm 0,7$	$120,2 \pm 2,1$	$131,2 \pm 1,8$

TABLEAU 4.7. – Temps de calcul (en s) des approches de références et de l'approche proposée sur un signal de 8 s. Les deux chiffres représentent la moyenne et l'intervalle de confiance.

Nombre de paramètres	**Togami**	**Cascade**	**Prop.**
DNN_0	25 272 432	42 119 352	42 119 352
DNN_i	-	50 540 760	50 540 760
Total	25 272 432	143 200 872	143 200 872

TABLEAU 4.8. – Nombre de paramètres des DNNs dans les approches de références et l'approche proposée. Le total prend en compte tous les DNN_i.

4.5. Résumé

Dans ce chapitre, nous avons proposé un algorithme de montée par blocs de coordonnées basé sur l'apprentissage profond pour la réduction conjointe multicanale de l'écho acoustique, la réverbération et le bruit. L'approche modélise conjointement à l'aide d'un DNN les spectres de la parole cible et des signaux résiduels après l'annulation d'écho et la déréverbération linéaire. Nous avons évalué notre système sur des enregistrements réels d'écho acoustique, de réverbération et de bruit acquis avec un Triby dans plusieurs situations différentes. Nous avons évalué les approches lorsque les conditions acoustiques sont invariantes dans le temps. Lorsque l'écho, la réverbération et le bruit sont présents simultanément, l'approche proposée dépasse les autres approches en réduction de la distorsion globale, sans dégrader les performances lorsque seulement un ou deux type de distorsion sont présents. Cependant, les performances en réduction la distorsion globale ne sont pas nécessairement corrélés à la qualité perceptuelle de la parole cible estimée. Lorsque les conditions acoustiques varient au cours du temps, l'approche proposée dépasse aussi les deux approches de référence en réduction de la distorsion globale. Toutefois, les performances sont limitées dans le cas où les conditions acoustiques varient au cours du temps, car notre approche suppose que les filtres linéaires sont invariants dans le temps, et que les sources ne se déplacent pas. Afin d'améliorer les performances en réduction globale de la distorsion, une version en ligne de l'approche proposée doit être formulée.

5. Variante en ligne de la réduction conjointe de bruit, d'écho et de réverbération basée sur l'apprentissage profond

Ce chapitre présente une version causale de notre méthode de réduction conjointe de bruit, d'écho et de réverbération en multicanal, présentée au chapitre 4. En effet, cette dernière souffre de deux limites. D'une part, elle suppose la connaissance des données futures et passées pour estimer les paramètres. D'autre part, elle suppose que les conditions acoustiques sont invariantes au cours du temps, ce qui n'est pas toujours le cas dans des situations réelles. Pour cela, nous proposons de modifier le modèle présenté au chapitre 4 en modélisant les filtres d'annulation d'écho et de déréverbération, ainsi que les caractéristiques spatiales de la parole cible et des signaux résiduels après annulation d'écho et déréverbération, comme des paramètres qui dépendent du temps. Nous développons un algorithme récursif pour mettre à jour tous les filtres de manière causale. Nous évaluons notre méthode sur les mêmes enregistrements réels qu'au chapitre 4. Nous la comparons à une version en ligne de la combinaison en cascade des approches de réduction individuelle de bruit, d'écho et de réverbération présentée au chapitre 4.

5.1. Formulation du problème

Nous rappelons ici brièvement le problème de la réduction conjointe de bruit, d'écho et de réverbération lorsque les conditions acoustiques varient au cours du temps. Dans des scénarios réels, les trois types de distorsion peuvent être présents simultanément comme l'illustre la figure 5.1. Dans le domaine temporel, la parole locale $\mathbf{s}(t)$ s'exprime comme dans (2.4) :

$$\mathbf{s}(t) = \sum_{\tau \geq 0} \mathbf{a}_s(t-\tau, \tau) u(t-\tau), \tag{5.1}$$

où la RIR de la parole locale $\mathbf{a}_s(t-\tau, \tau)$ dépend du temps t en raison des conditions acoustiques qui varient potentiellement dans le temps (par exemple, le locuteur local est en mouvement). De même, l'expression de l'écho $\mathbf{y}(t)$ dans (2.6) est formulée de la manière suivante :

$$\mathbf{y}(t) \approx \sum_{\tau \geq 0} \mathbf{a}_y(t-\tau, \tau) x(t-\tau), \tag{5.2}$$

où le chemin d'écho $\mathbf{a}_y(t-\tau, \tau)$ dépend aussi du temps t. Même si le système mains-libres est fixe, le chemin d'écho $\mathbf{a}_y(t-\tau, \tau)$ peut varier au cours du temps si le locuteur local

se déplace par exemple, car cela affecte les réflexions de l'écho acoustique dans la salle.

Dans le domaine temps-fréquence, la convolution temporelle dans (5.1) s'exprime de manière similaire à (2.12) :

$$\mathbf{s}(n,f) = \sum_{f'=0}^{F-1} \sum_{k \geq 0} \mathbf{a}_s(n,k,f',f) u(n-k,f'), \tag{5.3}$$

où la RIR $\mathbf{a}_s(n,k,f',f)$ dépend ici à la fois de la trame n, associée au temps t, et de la trame k, associée au délai τ. L'expression de l'écho dans (5.2) est formulée de manière similaire à (5.3) dans le domaine temps-fréquence. Le mélange $\mathbf{d}(n,f)$ est la somme de la parole locale $\mathbf{s}(n,f)$, du signal de bruit $\mathbf{b}(n,f)$ et de l'écho $\mathbf{y}(n,f)$:

$$\mathbf{d}(n,f) = \mathbf{s}(n,f) + \mathbf{b}(n,f) + \mathbf{y}(n,f) \tag{5.4}$$

$$= \mathbf{s}_{\mathrm{e}}(n,f) + \mathbf{s}_{\mathrm{l}}(n,f) + \mathbf{b}(n,f) + \mathbf{y}(n,f). \tag{5.5}$$

Le but est d'extraire la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$ du mélange $\mathbf{d}(n,f)$.

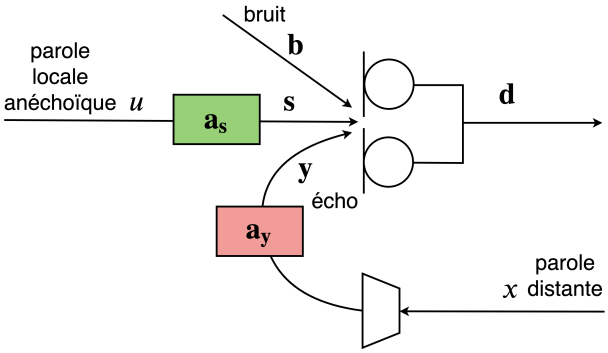


FIGURE 5.1. – Problème de l'écho acoustique, du bruit et de la réverbération.

En ce qui concerne l'état de l'art, nous avons présenté au chapitre 4 une combinaison de méthodes de réduction individuelle d'écho, de réverbération et de bruit. La sélection de ces méthodes avait été évoquée dans la partie 2.2.4. Pour la réduction d'écho, nous avions retenu le système combinant un filtre d'annulation d'écho avec un post-filtre de suppression d'écho résiduel. En particulier, les méthodes d'estimation du filtre d'annulation d'écho et du post-filtre sont très souvent adaptatives (voir la partie 2.2.2.1). Elles sont donc appropriées pour lorsque les conditions acoustiques varient au cours du temps. Toutefois, en réduction de bruit et en déréverbération, les méthodes sélectionnées sont toutes des méthodes hors-lignes qui utilisent les données futures et passées pour estimer les paramètres des filtres. De plus, ces méthodes supposent que les conditions acoustiques sont invariantes au cours du temps.

Pour la déréverbération, nous avions retenu le système combinant un filtre de déréverbération linéaire avec un post-filtre de suppression de réverbération résiduelle (voir la partie 2.2.3.2). En particulier, ce type de méthodes estime le filtre de déréverbération linéaire à l'aide de la méthode WPE. Toutefois, cette méthode suppose que le filtre de

déréverbération ne dépend pas du temps, ce qui n'est pas adapté aux scénarios où la RIR de la parole locale $\mathbf{a}_s(n,k,f',f)$ varie au cours du temps. Une version en ligne de la méthode WPE a été proposée pour estimer un filtre de déréverbération linéaire $\mathcal{G}(n,f)$ qui varie au cours du temps [Yoshioka et al., 2009b; Yoshioka et Nakatani, 2013]. Toutefois, aucune contrainte n'est imposée sur les paramètres spectraux de la parole cible $\mathbf{s}_e(n,f)$, ce qui conduit à une déréverbération limitée. Heymann et al. [2018] ont proposé une version en ligne de la méthode WPE où les paramètres spectraux de la parole cible sont contraints par un DNN.

Pour la réduction de bruit en multicanal, nous avions retenu les méthodes de séparation de sources basées sur l'apprentissage profond, qui estiment les DSPs à l'aide d'un DNN et les MCSs à l'aide d'un algorithme EM (voir la partie 2.2.1.5). Cependant, ce type de méthodes suppose que les MCSs sont invariantes au cours du temps, ce qui n'est pas adapté aux scénarios où le locuteur local et la source de bruit se déplacent. Togami [2011] a proposé une version en ligne de l'algorithme EM qui relâche l'hypothèse d'invariance des MCSs. Toutefois, cette méthode n'impose pas de contraintes sur les DSPs. Simon et Vincent [2012] ont proposé une version en ligne où les DSPs sont contraintes par une NMF. Des méthodes plus récentes contraignent les DSPs avec un DNN [Drude et al., 2018; Togami, 2019].

Comme les filtres qui composent le système interagissent entre eux, ils doivent être optimisés conjointement, afin d'améliorer la robustesse du système de réduction conjointe d'écho, de réverbération et de bruit. Toutefois, aucune méthode en ligne estimant conjointement tous les filtres n'a été proposée.

5.2. Solution proposée

5.2.1. Modèle

La figure 5.2 illustre l'approche proposée. Tout d'abord, nous appliquons une annulation d'écho à l'aide du filtre $\mathcal{H}(f) = [\mathbf{h}(n,0,f)\dots\mathbf{h}(n,K-1,f)] \in \mathbb{C}^{M\times K}$, qui dépend ici de la trame n, sur le mélange $\mathbf{d}(n,f)$:

$$\mathbf{e}(n,f) = \mathbf{d}(n,f) - \underbrace{\sum_{k=0}^{K} \mathbf{h}(n,k,f)x(n-k,f)}_{=\widehat{\mathbf{y}}(n,f)}, \tag{5.6}$$

où $\mathbf{h}(n,k,f) \in \mathbb{C}^{M\times 1}$ correspond au k-ième délai du filtre $\mathcal{H}(n,f)$. Le signal $\mathbf{e}(n,f)$ contient la parole locale $\mathbf{s}(n,f)$, le signal de bruit $\mathbf{b}(n,f)$ et l'écho résiduel $\mathbf{z}(n,f)$. Nous appliquons ensuite une déréverbération linéaire à l'aide du filtre $\mathcal{G}(n,f) = [\mathbf{G}(n,\Delta,f)\dots\mathbf{G}(n,\Delta+L-1,f)] \in \mathbb{C}^{M\times ML}$, qui dépend ici de la trame n, sur le signal $\mathbf{e}(n,f)$. Le signal $\mathbf{r}(n,f)$ qui en résulte s'exprime de la manière suivante :

$$\mathbf{r}(n,f) = \mathbf{e}(n,f) - \underbrace{\sum_{l=\Delta}^{\Delta+L+1} \mathbf{G}(n,l,f)\mathbf{e}(n-l,f)}_{=\mathbf{e}_l(n,f)}, \tag{5.7}$$

où $\mathbf{G}(n,l,f) \in \mathbb{C}^{M\times M}$ correspond au l-ième délai du filtre long $\mathcal{G}(n,f)$. Les filtres $\mathcal{H}(n,f)$ et $\mathcal{G}(n,f)$ sont causaux. Pour $n<0$, nous supposons que les signaux observés $\mathbf{d}(n,f)$ et $x(n,f)$ sont nuls. À cause des raisons évoquées dans les parties 2.2 et 2.3, ainsi que des conditions acoustiques qui varient au cours du temps, des signaux résiduels non désirés subsistent dans le signal $\mathbf{r}(n,f)$ et s'expriment de la manière suivante :

$$\mathbf{r}(n,f) - \mathbf{s}_{\mathrm{e}}(n,f) = \mathbf{s}_{\mathrm{r}}(n,f) + \mathbf{z}_{\mathrm{r}}(n,f) + \mathbf{b}_{\mathrm{r}}(n,f). \tag{5.8}$$

À titre de rappel, les termes $\mathbf{s}_{\mathrm{r}}(n,f)$, $\mathbf{z}_{\mathrm{r}}(n,f)$ et $\mathbf{b}_{\mathrm{r}}(n,f)$ désignent la réverbération résiduelle de la parole locale, l'écho résiduel *déréverbéré*, et le bruit *déréverbéré*, définis comme dans (4.14), (4.15) et (4.16), respectivement (voir la partie 4.2.1). Le terme *déréverbéré* signifie « après application du filtre $\mathcal{G}(n,f)$ ». Pour extraire la composante précoce $\mathbf{s}_{\mathrm{e}}(n,f)$ du signal $\mathbf{r}(n,f)$, nous appliquons un post-filtre court $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$ sur le signal $\mathbf{r}(n,f)$:

$$\widehat{\mathbf{s}}_{\mathrm{e}}(n,f) = \mathbf{W}_{s_{\mathrm{e}}}(n,f)\mathbf{r}(n,f). \tag{5.9}$$

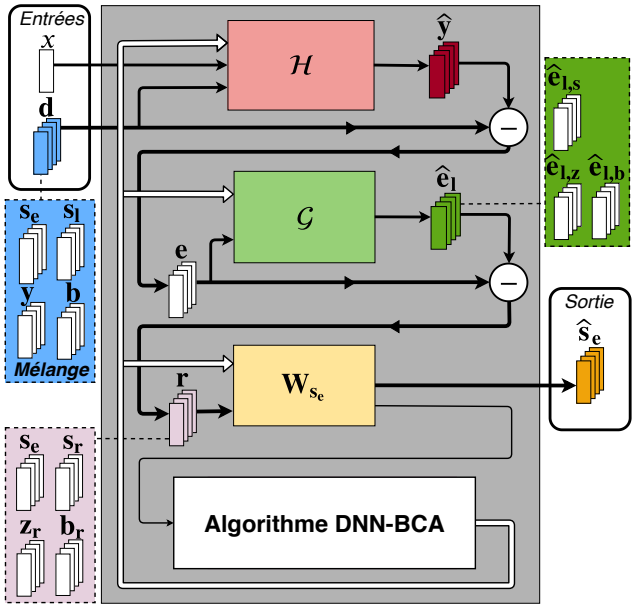


FIGURE 5.2. – Approche proposée. Les flèches et lignes ont la même signification que dans la figure 4.2.

Inspirés par les méthodes en ligne de Togami [2019] en séparation de sources, et de Heymann et al. [2018] en déréverbération, nous estimons les filtres $\mathcal{H}(n,f)$, $\mathcal{G}(n,f)$ et $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$ en utilisant une mise à jour en ligne de leurs paramètres respectifs. Pour cela, nous modélisons la parole cible $\mathbf{s}_{\mathrm{e}}(n,f)$ et les trois signaux résiduels $\mathbf{s}_{\mathrm{r}}(n,f)$, $\mathbf{z}_{\mathrm{r}}(n,f)$ et $\mathbf{b}_{\mathrm{r}}(n,f)$ par des variables gaussiennes de moyenne nulle avec le modèle gaussien local (voir la partie 2.2.1.4). Nous utilisons la notation générale $\mathbf{c}(n,f)$ pour désigner l'un des

quatre signaux $\mathbf{s}_{\mathrm{e}}(n,f)$, $\mathbf{s}_{\mathrm{r}}(n,f)$, $\mathbf{z}_{\mathrm{r}}(n,f)$ et $\mathbf{b}_{\mathrm{r}}(n,f)$. Chacun de ces signaux est modélisé comme suit

$$\mathbf{c}(n,f) \sim \mathcal{N}(\mathbf{0}, v_c(n,f)\mathbf{R}_c(n,f)), \tag{5.10}$$

où la MCS $\mathbf{R}_c(n,f)$ dépend ici de la trame n. Nous estimons le post-filtre $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$ à partir des covariances $\mathbf{\Sigma}_c(n,f) = v_c(n,f)\mathbf{R}_c(n,f)$ en utilisant le filtre de Wiener. Celui-ci est formulé de la manière suivante pour le signal $\mathbf{c}(n,f)$:

$$\mathbf{W}_c(n,f) = v_c(n,f)\mathbf{R}_c(n,f)\left(\sum_{c'\in\,\mathcal{C}} v_{c'}(n,f)\mathbf{R}_{c'}(n,f)\right)^{-1}, \tag{5.11}$$

où $\mathcal{C} = \{\mathbf{s}_{\mathrm{e}}, \mathbf{s}_{\mathrm{r}}, \mathbf{z}_{\mathrm{r}}, \mathbf{b}_{\mathrm{r}}\}$ désigne l'ensemble des quatre composantes du signal $\mathbf{r}(n,f)$ dans (4.13). Le post-filtre $\mathbf{W}_{s_{\mathrm{e}}}(n,f)$ est un cas spécifique de (5.11) où $\mathbf{c}(n,f) = \mathbf{s}_{\mathrm{e}}(n,f)$.

5.2.2. Vraisemblance

Pour estimer les paramètres de ce modèle de manière en ligne, nous choisissons d'optimiser la vraisemblance de la suite des signaux $\mathcal{O}(n) = \{\mathbf{d}(n,f), \ldots, \mathbf{d}(0,f), x(n,f), \ldots x(0,f)\}_f$ observés à la trame n en fonction des paramètres estimés aux trames précédentes $0,\ldots,n-1$. D'après (5.6), (5.7), (5.8) et (5.10), la log-vraisemblance de la suite $\mathcal{O}(n)$ correspond à

$$\begin{aligned}\mathcal{L}(\mathcal{O}(n);\Theta_H(n),&\Theta_G(n),\Theta_c(n)) \\ &= \sum_{f=0}^{F-1}\sum_{n'=0}^{n} \log p\Big(\mathbf{d}(n',f)\Big|\mathbf{d}(n'-1,f),\ldots,\mathbf{d}(0,f),x(n',f),\ldots,x(0,f)\Big), \qquad (5.12)\\ &= \sum_{f=0}^{F-1}\sum_{n=0}^{n} \log \mathcal{N}_{\mathbb{C}}\Big(\mathbf{d}(n',f);\boldsymbol{\mu}_{\mathbf{d}}(n',f),\mathbf{\Sigma}_{\mathbf{dd}}(n',f)\Big), \qquad (5.13)\end{aligned}$$

où

$$\boldsymbol{\mu}_{\mathbf{d}}(n',f) = \sum_{k=0}^{K-1}\mathbf{h}(n',k,f)x(n'-k,f) + \sum_{l=\Delta}^{\Delta+L-1}\mathbf{G}(n',l,f)\mathbf{e}(n'-l,f), \tag{5.14}$$

$$\mathbf{\Sigma}_{\mathbf{dd}}(n',f) = \sum_{c'\in\mathcal{C}} v_{c'}(n',f)\mathbf{R}_{c'}(n',f), \tag{5.15}$$

et $\Theta_H(n) = \{\mathcal{H}(n,f)\}_f$, $\Theta_G(n) = \{\mathcal{G}(n,f)\}_f$ et $\Theta_c(n) = \{v_{c'}(n,f), \mathbf{R}_{c'}(n,f)\}_{c',f}$ sont les paramètres à estimer à la trame n. Il convient de noter qu'à la trame n, nous n'estimons pas les paramètres des trames précédentes $0,\ldots,n-1$, car ces paramètres ont déjà été estimés. Comme ce problème d'optimisation n'a pas de solution analytique, nous avons besoin d'estimer les paramètres $\Theta_H(n)$, $\Theta_G(n)$ et $\Theta_c(n)$ avec une procédure itérative.

5.2.3. Algorithme itératif d'optimisation en ligne

Nous proposons une version en ligne de l'algorithme DNN-BCA présenté dans la partie 4.2.3 pour estimer $\Theta_H(n)$, $\Theta_G(n)$ et $\Theta_c(n)$. La figure 5.3 illustre le processus d'optimisation. À chaque trame n, nous effectuons I itérations de cet algorithme. Chaque itération i est constituée de trois étapes de maximisation :

$$\widehat{\Theta}_H^{(i)}(n) \leftarrow \underset{\Theta_H(n)}{\operatorname{argmax}}\, \mathcal{L}\left(\mathcal{O}(n); \Theta_H(n), \widehat{\Theta}_G^{(i-1)}(n), \widehat{\Theta}_c^{(i-1)}(n)\right), \tag{5.16}$$

$$\widehat{\Theta}_G^{(i)}(n) \leftarrow \underset{\Theta_G(n)}{\operatorname{argmax}}\, \mathcal{L}\left(\mathcal{O}(n); \widehat{\Theta}_H^{(i)}(n), \Theta_G(n), \widehat{\Theta}_c^{(i-1)}(n)\right), \tag{5.17}$$

$$\widehat{\Theta}_c^{(i)}(n) \leftarrow \underset{\Theta_c(n)}{\operatorname{argmax}}\, \mathcal{L}\left(\mathcal{O}(n); \widehat{\Theta}_H^{(i)}(n), \widehat{\Theta}_G^{(i)}(n), \Theta_c(n)\right), \tag{5.18}$$

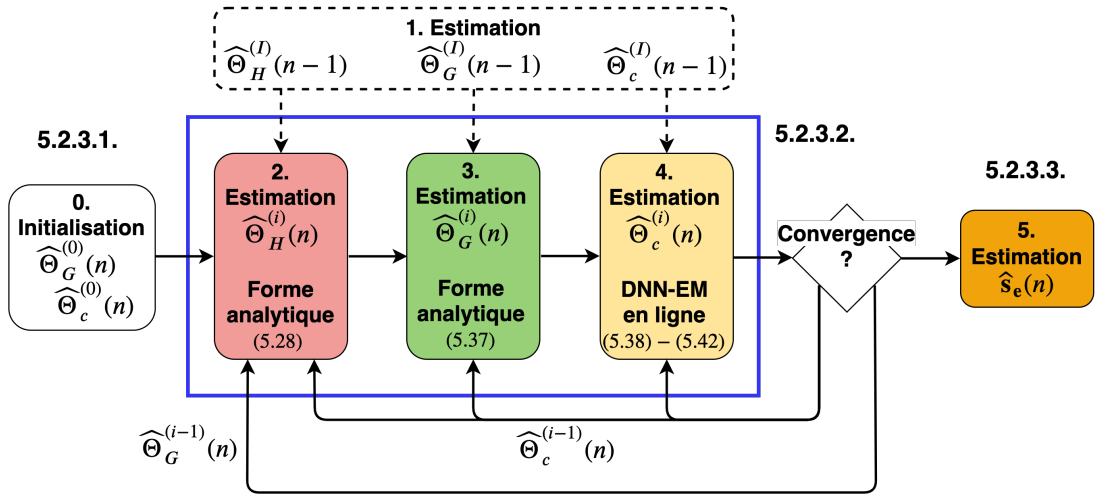
où l'exposant $(\cdot)^{(i)}$ indique la valeur des paramètres à l'itération i.

Les solutions de (5.16) et (5.17) sont analytiques. En ce qui concerne (5.18), comme il n'y a pas de solution analytique, nous proposons d'utiliser une version en ligne de l'algorithme DNN-EM utilisé au chapitre 4. Il convient de noter qu'ici, les mises à jour dépendent à la fois des paramètres $\widehat{\Theta}_H^{(I)}(n-1)$ $\widehat{\Theta}_G^{(I)}(n-1)$ et $\widehat{\Theta}_c^{(I)}(n-1)$ estimés à l'itération finale I de la trame précédente $n-1$, et des paramètres $\widehat{\Theta}_G^{(i-1)}(n)$ et $\widehat{\Theta}_c^{(i-1)}(n)$ estimés à l'itération $i-1$ de la trame courante n . Dans les sous-parties suivantes, nous décrivons l'initialisation et les règles de mise à jour pour les étapes (5.16)–(5.18) à l'itération i de l'algorithme proposé. Le calcul des règles de mise à jour de (5.16)–(5.17) est détaillée dans l'annexe B.

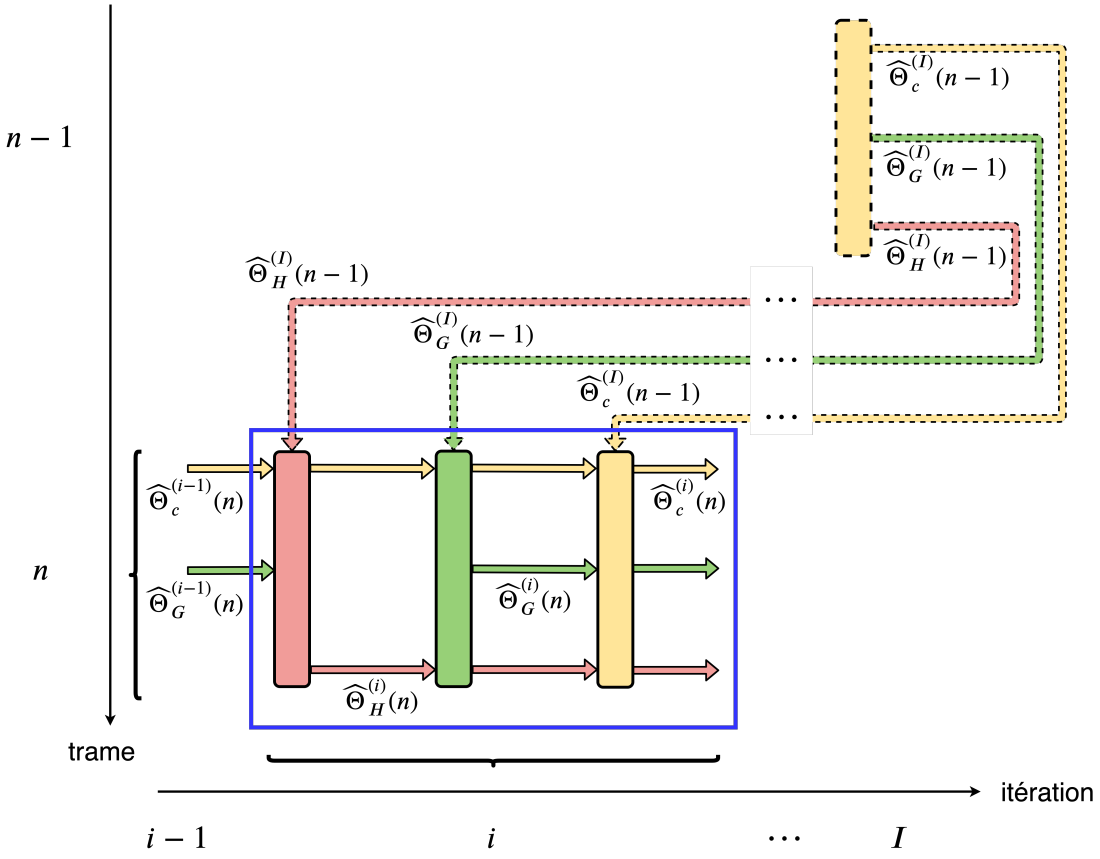
5.2.3.1. Initialisation

D'après (5.16), l'optimisation des trois paramètres $\Theta_H(n)$, $\Theta_G(n)$ et $\Theta_c(n)$ nécessite l'initialisation de $\Theta_G(n)$ et $\Theta_c(n)$. Pour $\widehat{\Theta}_G^{(0)}(n)$, nous initialisons le filtre $\mathcal{G}(n,f)$ à la valeur $\mathcal{G}^{(0)}(n,f) = \mathcal{G}^{(I)}(n-1,f)$, c'est-à-dire le filtre estimé à l'itération finale I de la trame précédente $n-1$. Pour $\widehat{\Theta}_c^{(0)}$, nous initialisons les MCS $\mathbf{R}_c(n,f)$ à la valeur $\mathbf{R}_c^{(0)}(n,f) = \mathbf{R}_c^{(I)}(n-1,f)$, c'est-à-dire les MCSs estimées à l'itération finale I de la trame précédente $n-1$. Nous initialisons les DSPs $v_c(n,f)$ sont initialisées conjointement à la valeur $v_c^{(0)}(n,f)$ à l'aide d'un DNN désigné par DNN_0, que nous avons pré-entraîné. Nous décrivons les entrées, les cibles et l'architecture de DNN_0 dans la partie 5.2.4 ci-après.

Pour les trames $n < 0$, nous considérons que les filtres linéaires $\mathcal{H}^{(I)}(n,f)$ et $\mathcal{G}^{(I)}(n,f)$ sont égaux à zéro. Les MCSs $\mathbf{R}_c^{(I)}(n,f)$ sont considérées comme étant égales à la matrice identité $\mathbf{I}_M$. L'algorithme proposé nécessite aussi l'initialisation de matrices inverses $\mathbf{P}^{(I)}(n,f)^{-1} \in \mathbb{C}^{MK\times MK}$ et $\mathbf{Q}^{(I)}(n,f)^{-1} \in \mathbb{C}^{M^2L\times M^2L}$, que nous définissons dans la sous-partie 5.2.3.2 ci-après. Nous considérons que les matrices inverses $\mathbf{P}^{(I)}(n,f)^{-1}$ et $\mathbf{Q}^{(I)}(n,f)^{-1}$ sont égales aux matrices identité $\mathbf{I}_{MK}$ et $\mathbf{I}_{M^2L}$, respectivement.



(a) Schéma du processus global d'optimisation.



(b) Schéma des mises à jour des paramètres de la trame n à l'itération i.

FIGURE 5.3. – Algorithme DNN-BCA en ligne proposé. Les mises à jour dépendent des paramètres estimés à l'itération finale I de la trame $n-1$ et des paramètres estimés à l'itération $i-1$ de la trame n.

5.2.3.2. Mise à jour des paramètres

Paramètres du filtre d'annulation d'écho $\Theta_H(n)$ À l'itération i, le filtre d'annulation d'écho $\mathcal{H}^{(i)}(n,f)$ est mis à jour de manière similaire à (4.29) :

$$\underline{\mathbf{h}}^{(i)}(n,f) = \mathbf{P}^{(i)}(n,f)^{-1}\mathbf{p}^{(i)}(n,f). \tag{5.19}$$

Le terme $\underline{\mathbf{h}}^{(i)}(n,f) = \left[\mathbf{h}^{(i)}(n,0,f)^T \dots \mathbf{h}^{(i)}(n,K-1,f)^T\right]^T \in \mathbb{C}^{MK\times 1}$ est une version vectorisée de $\mathcal{H}^{(i)}(n,f)$ comme dans la partie 4.2.3.2. Pour les termes $\mathbf{P}^{(i)}(n,f) \in \mathbb{C}^{MK\times MK}$ et $\mathbf{p}^{(i)}(n,f) \in \mathbb{C}^{MK\times 1}$, nous considérons des mises à jour différentes des mises à jour hors ligne dans (4.30)–(4.31). Les termes $\mathbf{P}^{(i)}(n,f) \in \mathbb{C}^{MK\times MK}$ et $\mathbf{p}^{(i)}(n,f) \in \mathbb{C}^{MK\times 1}$ sont ici calculés de manière récursive à partir des paramètres estimés à l'itération I de la trame précédente $n-1$, et des paramètres estimés à l'itération $i-1$ de la trame courante n :

$$\mathbf{P}^{(i)}(n,f) = \mathbf{P}^{(I)}(n-1,f) + \underline{\mathbf{X}}_{\mathrm{r}}^{(i-1)}(n,f)^H \mathbf{\Sigma}_{\mathbf{dd}}^{(i-1)}(n,f)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}^{(i-1)}(n,f) \tag{5.20}$$

$$\mathbf{p}^{(i)}(n,f) = \mathbf{p}^{(I)}(n-1,f) + \underline{\mathbf{X}}_{\mathrm{r}}^{(i-1)}(n,f)^H \mathbf{\Sigma}_{\mathbf{dd}}^{(i-1)}(n,f)^{-1}\mathbf{r}_d^{(i-1)}(n,f). \tag{5.21}$$

Le terme $\underline{\mathbf{X}}_{\mathrm{r}}^{(i-1)}(n,f) = \left[\mathbf{X}_{\mathrm{r}}^{(i-1)}(n,f) \dots \mathbf{X}_{\mathrm{r}}^{(i-1)}(n-K+1,f)\right] \in \mathbb{C}^{M\times MK}$ contient les K trames $\mathbf{X}_{\mathrm{r}}^{(i-1)}(n-k,f) \in \mathbb{C}^{M\times M}$ comme dans la partie 4.2.3.2. Les K trames $\mathbf{X}_{\mathrm{r}}^{(i-1)}(n-k,f)$ sont des versions *déréverbérées* de la parole distante $x(n-k,f)$ obtenues comme dans (4.32) :

$$\mathbf{X}_{\mathrm{r}}^{(i-1)}(n-k,f) = x(n-k,f)\mathbf{I}_M - \sum_{l=\Delta}^{\Delta+L-1} x(n-k-l,f)\mathbf{G}^{(i-1)}(n,l,f). \tag{5.22}$$

Le terme $\mathbf{r}_d^{(i-1)}(n,f)$ dans (5.21) est une version *déréverbérée* du mélange $\mathbf{d}(n,f)$ obtenue comme dans (4.33) :

$$\mathbf{r}_d^{(i-1)}(n,f) = \mathbf{d}(n,f) - \sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}^{(i-1)}(n,l,f)\mathbf{d}(n-l,f). \tag{5.23}$$

Il convient de remarquer que le filtre de déréverbération $\mathcal{G}^{(i-1)}(n,f)$ dans (5.22) et (5.23) correspond à l'estimation de l'itération $i-1$ de la trame courante n.

Pour s'adapter à des conditions acoustiques qui varient potentiellement au cours du temps, il est nécessaire d'« oublier » les états passés des termes $\mathbf{P}^{(i)}(n,f)$ et $\mathbf{p}^{(i)}(n,f)$. Plutôt que d'utiliser (5.20) et (5.21), nous considérons des mises à jour de $\mathbf{P}^{(i)}(n,f)$ et $\mathbf{p}^{(i)}(n,f)$ avec une moyenne glissante :

$$\mathbf{P}^{(i)}(n,f) = \alpha_h\mathbf{P}^{(I)}(n-1,f) + (1-\alpha_h)\underline{\mathbf{X}}_{\mathrm{r}}^{(i-1)}(n,f)^H \mathbf{\Sigma}_{\mathbf{dd}}^{(i-1)}(n,f)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}^{(i-1)}(n,f), \tag{5.24}$$

$$\mathbf{p}^{(i)}(n,f) = \alpha_h\mathbf{p}^{(I)}(n-1,f) + (1-\alpha_h)\underline{\mathbf{X}}_{\mathrm{r}}^{(i-1)}(n,f)^H \mathbf{\Sigma}_{\mathbf{dd}}^{(i-1)}(n,f)^{-1}\mathbf{r}_d^{(i-1)}(n,f). \tag{5.25}$$

où α_h est un coefficient d'oubli tel que $0 < \alpha_h < 1$. Il convient d'ajouter que le coefficient α_h permet un compromis entre vitesse de convergence et stabilité de l'algorithme en ligne. Généralement, le coefficient α_h est choisi tel que $0,99 < \alpha_h < 1$.

En ce qui concerne la complexité de la mise à jour du filtre $\mathcal{H}^{(i)}(n,f)$, il ne paraît pas raisonnable de calculer la matrice inverse $\mathbf{P}^{(i)}(n,f)^{-1} \in \mathbb{C}^{MK\times MK}$ dans (5.19) à chaque trame n et chaque itération i. En appliquant l'identité de Woodbury sur (5.24), nous réduisons la complexité de (5.19) en calculant la matrice inverse $\mathbf{P}^{(i)}(n,f)^{-1}$ de la manière suivante :

$$\mathbf{P}^{(i)}(n,f)^{-1} = \frac{1}{\alpha_h}\mathbf{P}^{(I)}(n-1,f)^{-1} - \frac{1}{\alpha_h}\mathbf{K}_{\mathrm{X}}^{(i)}(n,f)\underline{\mathbf{X}}_{\mathrm{r}}^{(i-1)}(n,f)\mathbf{P}^{(I)}(n-1,f)^{-1}. \quad (5.26)$$

où

$$\mathbf{K}_{\mathrm{X}}^{(i)}(n,f) = (1-\alpha_h)\mathbf{P}^{(I)}(n-1,f)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}^{(i-1)}(n,f)^H \left(\alpha_h\mathbf{\Sigma}_{\mathbf{dd}}^{(i-1)}(n,f) + (1-\alpha_h)\underline{\mathbf{X}}_{\mathrm{r}}^{(i-1)}(n,f)\mathbf{P}^{(I)}(n-1,f)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}^{(i-1)}(n,f)^H\right)^{-1}. \quad (5.27)$$

Ainsi, la complexité de la mise à jour du filtre d'annulation d'écho $\mathcal{H}^{(i)}(n,f)$ est diminuée en remplaçant l'inversion de la matrice $\mathbf{P}^{(i)}(n,f)$ de taille MK dans (5.19) par l'inversion d'une matrice carrée de taille M dans (5.27). Le filtre d'annulation d'écho $\mathcal{H}^{(i)}(n,f)$ est alors mis à jour de la manière suivante :

$$\underline{\mathbf{h}}^{(i)}(n,f) = \underline{\mathbf{h}}^{(I)}(n-1,f) + \mathbf{K}_{\mathrm{X}}^{(i)}(n,f)\left(\mathbf{r}_d^{(i-1)}(n,f) - \underline{\mathbf{X}}_{\mathrm{r}}^{(i-1)}(n,f)\underline{\mathbf{h}}^{(I)}(n-1,f)\right). \quad (5.28)$$

La mise à jour du filtre d'annulation d'écho $\mathcal{H}^{(i)}(n,f)$ dépend donc à la fois du filtre d'annulation d'écho $\mathcal{H}^{(I)}(n,f)$ estimé à l'itération I de la trame précédente $n-1$, et des paramètres estimés à l'itération $i-1$ de la trame courante n (voir la figure 5.3). L'obtention des mises à jour (5.26)–(5.28) est détaillée dans l'annexe B. Le signal $\mathbf{e}^{(i)}(n,f)$ est ensuite calculé comme dans (5.6).

Paramètres du filtre de déréverbération $\Theta_G(n)$ À l'itération i, le filtre de déréverbération linéaire $\mathcal{G}^{(i)}(n,f)$ est mis à jour de manière similaire à (4.34) :

$$\underline{\mathbf{g}}^{(i)}(n,f) = \mathbf{Q}^{(i)}(n,f)^{-1}\mathbf{q}^{(i)}(n,f), \quad (5.29)$$

Le terme $\underline{\mathbf{g}}^{(i)}(n,f) = \left[\mathbf{g}_1^{(i)}(n,\Delta,f)^T \dots \mathbf{g}_M^{(i)}(n,\Delta,f)^T \dots\dots \mathbf{g}_1^{(i)}(n,\Delta+L-1,f)^T \dots \mathbf{g}_M^{(i)}(n,\Delta+L-1,f)^T\right]^T \in \mathbb{C}^{M^2L\times 1}$ est une version vectorisée de $\mathcal{G}^{(i)}(n,f)$ comme dans la partie 4.2.3.2. Pour les termes $\mathbf{Q}^{(i)}(n,f) \in \mathbb{C}^{M^2L\times M^2L}$ et $\mathbf{q}^{(i)}(n,f) \in \mathbb{C}^{M^2L\times 1}$, nous considérons des mises à jour différentes des mises à jour hors ligne définies dans (4.35)–(4.36). Les termes $\mathbf{Q}^{(i)}(n,f)$ et $\mathbf{q}^{(i)}(n,f)$ sont ici calculés de manière récursive à partir des paramètres estimés à l'itération I de la trame précédente $n-1$, et des paramètres

estimés à l'itération $i-1$ de la trame courante n :

$$\mathbf{Q}^{(i)}(n,f) = \mathbf{Q}^{(I)}(n-1,f) + \underline{\mathbf{E}}^{(i)}(n,f)^H \mathbf{\Sigma}_{\mathbf{dd}}^{(i-1)}(n,f)^{-1} \underline{\mathbf{E}}^{(i)}(n,f) \tag{5.30}$$

$$\mathbf{q}^{(i)}(n,f) = \mathbf{q}^{(I)}(n-1,f) + \underline{\mathbf{E}}^{(i)}(n,f)^H \mathbf{\Sigma}_{\mathbf{dd}}^{(i-1)}(n,f)^{-1} \mathbf{e}^{(i)}(n,f). \tag{5.31}$$

Le terme $\underline{\mathbf{E}}^{(i)}(n,f) = \left[\mathbf{E}^{(i)}(n-\Delta,f) \dots \mathbf{E}^{(i-1)}(n-\Delta-L+1,f)\right] \in \mathbb{C}^{M \times M^2 L}$ contient les L trames $\mathbf{E}^{(i)}(n-l,f) \in \mathbb{C}^{M \times M^2}$. Les L trames $\mathbf{E}^{(i)}(n-l,f)$ sont des versions matricielles du signal $\mathbf{e}^{(i)}(n-l,f)$ obtenues de la manière suivante :

$$\mathbf{E}^{(i)}(n-l,f) = \mathbf{I}_M \otimes \mathbf{e}^{(i)}(n-l,f)^T, \tag{5.32}$$

où $\mathbf{e}^{(i)}(n-l,f)$ est le signal obtenu par application du filtre d'annulation d'écho $\mathcal{H}^{(i)}(n,f)$ comme dans (4.11).

De même que pour les termes $\mathbf{P}^{(i)}(n,f)$ et $\mathbf{p}^{(i)}(n,f)$ dans (5.24)–(5.25), nous considérons des mises à jour de $\mathbf{Q}^{(i)}(n,f)$ et $\mathbf{q}^{(i)}(n,f)$ dans (5.30)–(5.31) avec une moyenne glissante :

$$\mathbf{Q}^{(i)}(n,f) = \alpha_g \mathbf{Q}^{(I)}(n-1,f) + (1-\alpha_g) \underline{\mathbf{E}}^{(i)}(n,f)^H \mathbf{\Sigma}_{\mathbf{dd}}^{(i-1)}(n,f)^{-1} \underline{\mathbf{E}}^{(i)}(n,f), \tag{5.33}$$

$$\mathbf{q}^{(i)}(n,f) = \alpha_g \mathbf{q}^{(I)}(n-1,f) + (1-\alpha_g) \underline{\mathbf{E}}^{(i)}(n,f)^H \mathbf{\Sigma}_{\mathbf{dd}}^{(i-1)}(n,f)^{-1} \mathbf{e}^{(i)}(n,f). \tag{5.34}$$

où α_g est un coefficient d'oubli tel que $0 < \alpha_g < 1$.

De même que pour la mise à jour du filtre d'annulation d'écho $\mathcal{H}^{(i)}(n,f)$ dans (5.28), nous réduisons la complexité de (5.29) en calculant la matrice inverse $\mathbf{Q}^{(i)}(n,f)^{-1}$ à l'aide de l'identité de Woodbury appliquée sur (5.33) :

$$\mathbf{Q}^{(i)}(n,f)^{-1} = \frac{1}{\alpha_g} \mathbf{Q}^{(I)}(n-1,f)^{-1} - \frac{1}{\alpha_g} \mathbf{K}_{\mathrm{E}}^{(i)}(n,f) \underline{\mathbf{E}}^{(i)}(n,f) \mathbf{Q}^{(I)}(n-1,f)^{-1}, \tag{5.35}$$

où

$$\begin{aligned} \mathbf{K}_{\mathrm{E}}^{(i)}(n,f) = (1-\alpha_g) &\mathbf{Q}^{(I)}(n-1,f)^{-1} \underline{\mathbf{E}}^{(i)}(n,f)^H \\ &\left(\alpha_g \mathbf{\Sigma}_{\mathbf{dd}}^{(i-1)}(n,f) + \underline{\mathbf{E}}^{(i)}(n,f) \mathbf{Q}^{(I)}(n-1,f)^{-1} \underline{\mathbf{E}}^{(i)}(n,f)^H\right)^{-1}. \end{aligned} \tag{5.36}$$

Ainsi, la complexité de la mise à jour du filtre de déréverbération $\mathcal{G}^{(i)}(n,f)$ est diminuée en remplaçant l'inversion de la matrice $\mathbf{Q}^{(i)}(n,f)$ de taille $M^2 L$ dans (5.29) par l'inversion d'une matrice carrée de taille M dans (5.36). Le filtre de déréverbération $\mathcal{G}^{(i)}(n,f)$ est alors mis à jour de la manière suivante :

$$\underline{\mathbf{g}}^{(i)}(n,f) = \underline{\mathbf{g}}^{(I)}(n-1,f) + \mathbf{K}_{\mathrm{E}}^{(i)}(n,f)\left(\mathbf{e}^{(i)}(n,f) - \underline{\mathbf{E}}^{(i)}(n,f) \underline{\mathbf{g}}^{(I)}(n-1,f)\right). \tag{5.37}$$

La mise à jour du filtre de déréverbération $\mathcal{G}^{(i)}(n,f)$ dépend donc à la fois du filtre de déréverbération $\mathcal{G}^{(I)}(n,f)$ estimé à l'itération I de la trame précédente $n-1$, et des paramètres estimés à l'itération $i-1$ de la trame courante n (voir la figure 5.3). L'obtention des mises à jour (5.35)–(5.37) est détaillée dans l'annexe B. Le signal $\mathbf{r}^{(i)}(n,f)$ est ensuite calculé comme dans (5.7).

Paramètres des covariances Θ_c Comme il n'y a pas de solution analytique pour l'optimisation de la log-vraisemblance $\mathcal{L}\Big(\mathcal{O}(n); \widehat{\Theta}_H^{(i)}(n), \widehat{\Theta}_G^{(i)}(n), \Theta_c(n)\Big)$ en fonction de $\Theta_c(n)$, nous estimons les paramètres des covariances à l'aide d'un algorithme EM en ligne. Plus spécifiquement , après chaque mise à jour des filtres linéaires $\mathcal{H}^{(i)}(n,f)$ et $\mathcal{G}^{(i)}(n,f)$, nous proposons d'utiliser une itération d'une version en ligne de l'algorithme DNN-EM de Nugraha et al. [2016a] pour estimer les DSPs et les MCSs de la parole cible et des signaux résiduels $\mathbf{s}_{\mathrm{e}}(n,f)$, $\mathbf{s}_{\mathrm{r}}(n,f)$, $\mathbf{z}_{\mathrm{r}}(n,f)$ et $\mathbf{b}_{\mathrm{r}}(n,f)$. À l'étape E, chacun des quatre signaux $\mathbf{c}^{(i)}(n,f)$ est estimé par

$$\widehat{\mathbf{c}}^{(i)}(n,f) = \mathbf{W}_c^{(i)}(n,f)\mathbf{r}^{(i)}(n,f), \tag{5.38}$$

où $\mathbf{W}_c^{(i)}(n,f)$ est le filtre de Wiener obtenu comme dans (5.11) à partir des paramètres $v_c^{(i-1)}(n,f)$ et $\mathbf{R}_c^{(i-1)}(f)$, et le signal $\mathbf{r}^{(i)}(n,f)$ est obtenu par application successive des filtres linéaires $\mathcal{H}^{(i)}(n,f)$ et $\mathcal{G}^{(i)}(n,f)$ sur le mélange $\mathbf{d}(n,f)$ comme dans (5.6)–(5.7). Le moment non centré d'ordre 2 $\widehat{\mathbf{\Sigma}}_c^{(i)}(n,f)$ est estimé par

$$\widehat{\mathbf{\Sigma}}_c^{(i)}(n,f) = \widehat{\mathbf{c}}^{(i)}(n,f)\widehat{\mathbf{c}}^{(i)}(n,f)^H + \Big(\mathbf{I} - \mathbf{W}_c^{(i)}(n,f)\Big) v_c^{(i-1)}(n,f)\mathbf{R}_c^{(i-1)}(n,f). \tag{5.39}$$

À l'étape M, pour les MCSs $\mathbf{R}_c^{(i)}(f)$, nous considérons une forme pondérée de la mise à jour, inspirée des méthodes de réduction de bruit basée sur la SPP (voir la partie 2.2.1.3) :

$$\mathbf{R}_c^{(i)}(n,f) = \alpha_c^{(i)}(n,f)\mathbf{R}_c^{(I)}(n-1,f) + \Big(1-\alpha_c^{(i)}(n,f)\Big)\frac{1}{v_c^{(i-1)}(n,f)}\widehat{\mathbf{\Sigma}}_c^{(i-1)}(n,f), \tag{5.40}$$

où

$$\alpha_c^{(i)}(n,f) = \Big(1 - p_c^{(i)}(n,f)\Big) + \alpha_c p_c^{(i)}(n,f), \tag{5.41}$$

$p_c^{(i)}(n,f)$ est une probabilité de présence du signal $\mathbf{c}(n,f)$ à l'itération i que nous définissons comme

$$p_c^{(i)}(n,f) = \sqrt{v_c^{(i-1)}(n,f)}\left(\sqrt{\sum_{c'} v_{c'}^{(i-1)}(n,f)}\right)^{-1}, \tag{5.42}$$

et α_c est un coefficient d'oubli tel que $0 < \alpha_c < 1$. Les expériences montrent que cette pondération réduit les mauvaises estimations dans certaines bandes de fréquences en augmentant l'importance des points temps-fréquence pour lesquels $v_c^{(i-1)}(n,f)$ est grand. Ainsi, si la DSP $v_c^{(i-1)}(n,f)$ est nulle (le signal $\mathbf{c}(n,f)$ est absent), on a $p_c^{(i)}(n,f) = 0$. Par conséquent, $\alpha_c^{(i)}(n,f) = 1$, et $\mathbf{R}_c^{(i-1)}(n,f) = \mathbf{R}_c^{(I)}(n-1,f)$. Si le signal $\mathbf{c}(n,f)$ est le seul signal présent dans le mélange $\mathbf{d}(n,f)$, on a $p_c^{(i)}(n,f) = 1$. Par conséquent, $\alpha_c^{(i)}(n,f) = \alpha_c$, et $\mathbf{R}_c^{(i)}(n,f) = \alpha_c\mathbf{R}_c^{(I)}(n-1,f) + (1-\alpha_c)\frac{1}{v_c^{(i-1)}(n,f)}\widehat{\mathbf{\Sigma}}_c^{(i-1)}(n,f)$. Nous ne normalisons pas les MCSs $\mathbf{R}_c^{(i)}(n,f)$ après (5.40) car les expériences montrent que cela dégrade les performances. L'opération (5.40)correspond à la mise à jour spatiale.

Les DSPs $v_c^{(i)}(n,f)$ des quatre signaux sont mises à jour conjointement à l'aide d'un DNN désigné par DNN_i, avec $i \geq 1$, que nous avons pré-entraîné. Cette opération correspond à la mise à jour spectrale. Nous décrivons les entrées, les cibles et l'architecture de DNN_i dans la partie 5.2.4 ci-après.

5.2.3.3. Estimation de la composante précoce finale $\mathbf{s_e}(n,f)$

Une fois que l'algorithme itératif d'optimisation en ligne a convergé après I itérations, nous avons l'estimation finale de parole cible $\widehat{\mathbf{s}}_{\text{e}}^{(I)}(n,f)$ en utilisant (5.6), (5.7) et (5.9). Le pseudo-code de l'algorithme est détaillé dans l'annexe A.

5.2.4. Modèle spectral par réseau de neurones

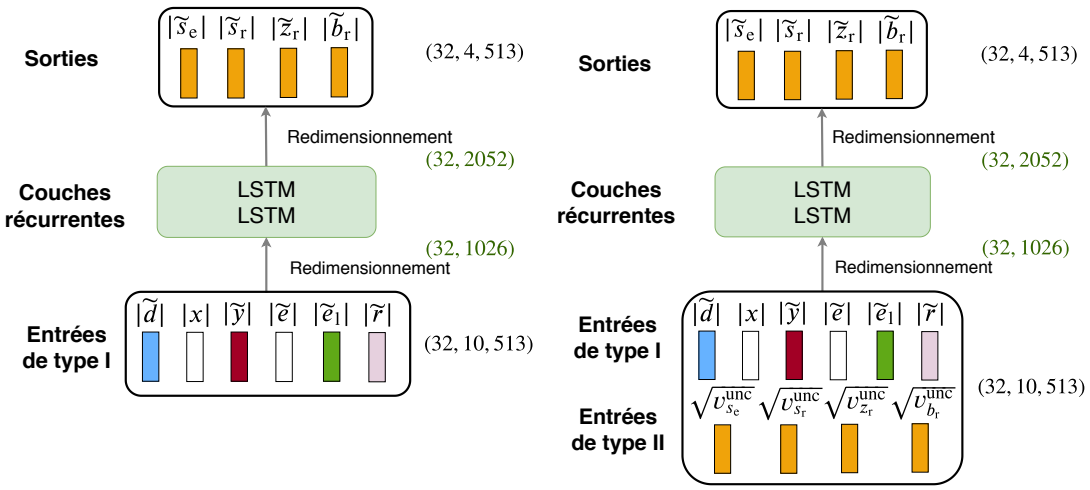
Nous considérons le même modèle spectral qu'au chapitre 4 (voir la partie 4.2.4). Les cibles et l'architecture de DNN_0 et DNN_i, avec $i \geq 1$, sont similaires et sont résumées sur la figure 5.4. Toutefois, il convient de préciser les entrées qui sont ici légèrement différentes.

Pour DNN_0, nous utilisons les entrées de type I obtenues comme dans (4.50). Les spectres en amplitude $|\widetilde{y}^{(0)}(n,f)|$, $|\widetilde{e}^{(0)}(n,f)|$, $|\widetilde{e}_1^{(0)}(n,f)|$ et $|\widetilde{r}^{(0)}(n,f)|$ obtenus à partir des signaux multicanaux $\widehat{\mathbf{y}}^{(0)}(n,f)$, $\mathbf{e}^{(0)}(n,f)$, $\widehat{\mathbf{e}}_1^{(0)}(n,f)$ et $\mathbf{r}^{(0)}(n,f)$ sont calculées à partir des valeurs des filtres linéaires $\mathcal{H}_{\text{DNN}}^{(0)}(n,f)$ et $\mathcal{G}_{\text{DNN}}^{(0)}(n,f)$. Les filtres linéaires $\mathcal{H}_{\text{DNN}}^{(0)}(n,f)$ et $\mathcal{G}_{\text{DNN}}^{(0)}(n,f)$ sont différents des filtres $\mathcal{H}^{(0)}(n,f) = \mathcal{H}^{(I)}(n-1,f)$ et $\mathcal{G}^{(0)}(n,f) = \mathcal{G}^{(I)}(n-1,f)$ (voir la partie 5.2.3.1). En effet, si nous considérions $\mathcal{H}_{\text{DNN}}^{(0)}(n,f) = \mathcal{H}^{(I)}(n-1,f)$ et $\mathcal{G}_{\text{DNN}}^{(0)}(n,f) = \mathcal{G}^{(I)}(n-1,f)$, c'est-à-dire les filtres linéaires estimés à l'itération I de la trame précédente $n-1$, alors il faudrait entrainer DNN_0 de bout-en-bout. Comme l'apprentissage bout-en-bout dépasse le cadre de ce chapitre, nous définissons la valeur des filtres linéaires $\mathcal{H}_{\text{DNN}}^{(0)}(n,f)$ et $\mathcal{G}_{\text{DNN}}^{(0)}(n,f)$ à la trame courante n. Ces valeurs sont définies avec des méthodes « extérieures » à l'algorithme itératif d'optimisation en ligne. La valeur du filtre d'annulation d'écho $\mathcal{H}_{\text{DNN}}^{(0)}(n,f)$ peut être définie en utilisant, par exemple, la méthode adaptative de Valin [2007], comme à la partie 4.3.4. La valeur du filtre de déréverbération $\mathcal{G}_{\text{DNN}}^{(0)}(n,f)$ peut être définie en utilisant, par exemple, la méthode en ligne DNN-WPE [Heymann et al., 2018].

Pour DNN_i, avec $i \geq 1$, nous utilisons les entrées de type I et II comme dans (4.54). Les entrées de type I sont calculées ici à partir des des valeurs des filtres linéaires $\mathcal{H}_{\text{DNN}}^{(i-1)}(n,f) = \mathcal{H}^{(i-1)}(n,f)$ et $\mathcal{G}_{\text{DNN}}^{(i-1)}(n,f) = \mathcal{G}^{(i-1)}(n,f)$ de l'itération $i-1$ de la trame courante n. En effet, les entrées de type I peuvent ici être calculées à partir des filtres linéaires de la trame courante n, et non de la trame précédente $n-1$ comme pour DNN_0. Les entrées de type II sont les racines carrées $\sqrt{v_c^{\text{unc}}(n,f))}$ des DSPs non contraintes définies dans (4.53) (voir la partie 4.2.4.2), calculées de la manière suivante :

$$v_c^{\text{unc}}(n,f)^{(i)} = \frac{1}{M}\text{tr}\left(\mathbf{R}_c^{(i)}(n,f)^{-1}\widehat{\mathbf{\Sigma}}_c^{(i-1)}(n,f)\right). \tag{5.43}$$

Nous utilisons DNN_0 et DNN_i pour estimer conjointement les quatre paramètres spec-



(a) DNN_0 (b) DNN_i, $i \geq 1$

FIGURE 5.4. – Architecture des DNNs avec une longueur de séquence de 32 intervalles de temps et $F = 513$ bandes de fréquence.

traux $\left[|\widetilde{s}_{\mathrm{e}}(n,f)||\widetilde{s}_{\mathrm{r}}(n,f)||\widetilde{z}_{\mathrm{r}}(n,f)||\widetilde{b}_{\mathrm{r}}(n,f)|\right]$ (voir la figure 5.4). Nous utilisons la divergence de Kullback-Leibler comme fonction de coût, définie comme dans (4.56).

5.3. Protocole expérimental

Dans cette partie, nous décrivons les données, les métriques, les méthodes de référence et le réglage des hyperparamètres utilisés pour évaluer l'algorithme proposé.

5.3.1. Scénario

Nous considérons la même situation qu'au chapitre 4, où un locuteur local interagit avec un correspondant distant à l'aide de Triby à une distance de 1, 5 m de ce système dans un environnement bruyant. Chaque enregistrement a une durée de 8 s qui contient 4 s de parole locale et 4 s de parole distante qui se chevauchent pendant 2 s. Nous étudions seulement le scénario où un bruit ambiant est présent pendant tout l'enregistrement. Chaque enregistrement est composé de 4 périodes de 2 s, comme le montre la figure 5.5 : 1) *bruit seul*, 2) *bruit et parole locale*, 3) *bruit, parole locale et parole distante*, 4) *bruit et parole distante*.

5.3.2. Données

Comme l'un des microphones était défectueux (voir la partie 4.3.2.1, nous considérons $M = 3$ microphones. Nous utilisons les ensembles de données d'apprentissage, de validation et de test utilisés au chapitre 4 correspondant au scénario avec présence de

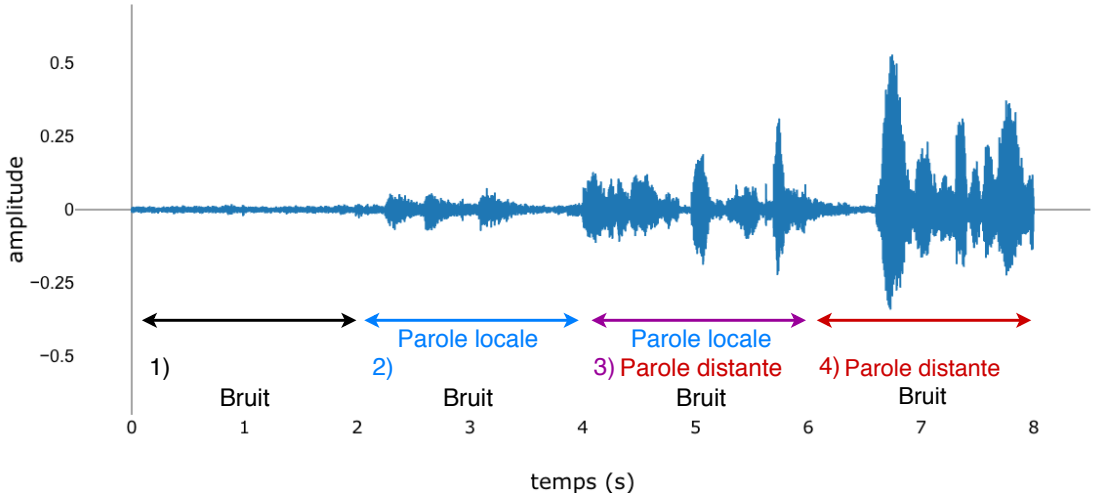


FIGURE 5.5. – Exemple d'enregistrement (seul un canal est illustré).

bruit (voir la partie 4.3.2). Les caractéristiques des ensembles de données sont résumées dans le tableau 5.1. À titre de rappel, les ensembles d'apprentissage et de validation correspondent à des conditions acoustiques invariantes au cours du temps, tandis que l'ensemble de test comprend deux sous-ensembles : l'un dont les conditions acoustiques sont invariantes au cours du temps, et l'autre dont ces conditions varient au cours du temps. L'ensemble de test invariant au cours du temps contient 4 500 enregistrements de 8 s, soit environ 10 h d'audio. Dans l'ensemble de test dont les conditions acoustiques varient au cours du temps, nous avons considéré le scénario où le locuteur local parle pendant 4 s, se déplace à un autre endroit de la salle, et parle à nouveau pendant 4 s. Cet ensemble contient 2 250 enregistrements de 16 s, soit environ 10 h. Dans les ensembles de test, nous considérons des conditions de SER et SNR moins difficiles que dans la partie 4.3.2.1, en appliquant des gains sur les signaux enregistrés d'écho $\mathbf{y}(t)$ et de bruit $\mathbf{s}(t)$. Il convient de préciser que les gains utilisés dans chacun des deux sous-ensembles de test sont différents. Les niveaux d'écho $\mathbf{y}(t)$ et de signal de bruit $\mathbf{b}(t)$ ont été choisis de manière à ce que le SER varie de -17 dB à -12 dB, et à ce que le SNR varie de $+5$ dB à $+10$ dB. Les caractéristiques des ensembles de données sont résumées dans le tableau 5.1.

5.3.3. Métriques

La composante précoce estimée $\widehat{\mathbf{s}}_{\mathrm{e}}$, obtenue avec (5.9), contient cinq composantes :

$$\widehat{\mathbf{s}}_{\mathrm{e}} = \mathbf{s}_{\mathrm{e}}^{\mathrm{post}} + \mathbf{s}_{\mathrm{r}}^{\mathrm{post}} + \mathbf{z}_{\mathrm{r}}^{\mathrm{post}} + \mathbf{b}_{\mathrm{r}}^{\mathrm{post}} + \mathbf{s}_{\mathrm{e}}^{\mathrm{art}}, \tag{5.44}$$

où $\mathbf{s}_{\mathrm{e}}^{\mathrm{post}}$ est la composante précoce potentiellement atténuée, $\mathbf{s}_{\mathrm{r}}^{\mathrm{post}}$, $\mathbf{z}_{\mathrm{r}}^{\mathrm{post}}$ et $\mathbf{b}_{\mathrm{r}}^{\mathrm{post}}$ sont les trois signaux post-résiduels de la réverbération tardive, de l'écho et bruit, respectivement, et $\mathbf{s}_{\mathrm{e}}^{\mathrm{art}}$ représente les artefacts introduits dans la composante précoce $\mathbf{s}_{\mathrm{e}}$. Ces

Ensemble de données	**Apprentissage**	**Validation**	**Test**
Signaux $\mathbf{y}$	enregistrés		enregistrés
Signaux $\mathbf{s}$	RIRs $\mathbf{a}_s$ simulées		
Signaux $\mathbf{b}$	simulés		
Salles	1-2-3	1-2	4
# paires de locuteurs	79	27	25
# enregistrements	13 572	4 536	4 500
# échantillons de bruit	36	36	6
Plage de SER (dB)	$[-45, +6]$		$[-17, -12]$
Plage de SNR (dB)	$[-21, +24]$		$[+5, +10]$

TABLEAU 5.1. – Caractéristiques des trois ensembles de données.

composantes sont calculées de la même que dans (2.89)–(2.91). À partir de ces composantes, nous utilisons les métriques définies dans la partie 2.4.1 pour évaluer la réduction de chaque type de distorsion et la dégradation de la composante précoce. Ces métriques sont résumées dans le tableau 5.3. Les métriques sont calculées séparément sur chaque canal m, puis moyennées sur les M canaux.

Contrairement à la partie 4.3.3, nous ne précisons pas ici les situations en présence de bruit avec le qualificatif *+ bruit*, car nous ne considérons justement que les situations en présence de bruit (voir la partie 5.3.1). Les ELR, SI-SAR et SI-SDR sont évalués uniquement dans les situations de *parole locale* et de *parole simultanée*. Le SNR est évalué uniquement dans les deux situations de *parole locale* et *parole simultanée*. Le SER est évalué uniquement sur dans la situation de *parole simultanée*. L'ERLE est évalué dans les trois situations de *parole locale*, *parole distante* et de *parole simultanée*. Le tableau 5.2 résume la correspondance entre les types de distorsion présents et les situations considérées.

Numérotation	**Situation**	**Type de distorsion présent**
1	*parole locale*	réverbération + bruit
2	*parole simultanée*	réverbération + écho + bruit
3	*parole distante seule*	écho

TABLEAU 5.2. – Correspondance entre les types de distorsion présents et les situations considérées. Le qualificatif *+ bruit* est omis.

De même qu'au chapitre 3, les performances peuvent aussi varier selon l'état de convergence des filtres linéaires $\mathcal{H}(n, f)$ et $\mathcal{G}(n, f)$. Toutefois, nous supposons que l'influence de la convergence des filtres linéaires $\mathcal{H}(n, f)$ et $\mathcal{G}(n, f)$ est négligeable par rapport à l'influence du type de situation considéré (voir le tableau 5.2) sur les performances. Par conséquent, nous ne considérons pas l'état de convergence des filtres dans le calcul des performances.

Puisque les performances peuvent varier en fonction de la présence de l'écho acous-

tique qui est le signal le plus fort, et aussi en fonction de la présence de bruit, nous calculons les métriques séparément dans chaque situation, en présence et en absence de bruit : *parole locale seule*, *parole simultanée* (paroles locale et distante actives simultanément) et *parole distante seule*. En particulier, les métriques dépendent de l'estimation d'un facteur d'échelle γ_c associé au signal c et défini comme dans (2.90) :

$$\gamma_c = \frac{\langle \widehat{s}_{\mathrm{e}}, c \rangle}{\|c\|^2}. \tag{5.45}$$

Dans chaque situation de l'ensemble de test (voir le tableau 4.3), nous supposons que γ_c est fixe. Toutefois, nous supposons que γ_c peut varier d'une situation à l'autre. Nous faisons ensuite la moyenne pondérée pour chaque métrique en fonction de la durée de chaque période sur laquelle sont calculées ces métriques, de la même manière que le calcul du SNR segmenté [Vincent et al., 2006].

Écho	ERLE	$10 \log_{10} \frac{\|y\|^2}{\|z_{\mathrm{r}}^{\mathrm{post}}\|^2}$	situations 2, 3
	SER	$10 \log_{10} \frac{\|s_{\mathrm{e}}^{\mathrm{post}}\|^2}{\|z_{\mathrm{r}}^{\mathrm{post}}\|^2}$	situations 2
Réverbération	ELR	$10 \log_{10} \frac{\|s_{\mathrm{e}}^{\mathrm{post}}\|^2}{\|s_{\mathrm{r}}^{\mathrm{post}}\|^2}$	situations 1, 2
Bruit	SNR	$10 \log_{10} \frac{\|s_{\mathrm{e}}^{\mathrm{post}}\|^2}{\|b_{\mathrm{r}}^{\mathrm{post}}\|^2}$	situations 1, 2
Artefacts	SI-SAR	$10 \log_{10} \frac{\|s_{\mathrm{e}}^{\mathrm{post}}\|^2}{\|s_{\mathrm{e}}^{\mathrm{art}}\|^2}$	situations 1, 2
Distorsion globale	SI-SDR	$10 \log_{10} \frac{\|s_{\mathrm{e}}^{\mathrm{post}}\|^2}{\|s_{\mathrm{r}}^{\mathrm{post}} + z_{\mathrm{r}}^{\mathrm{post}} + b_{\mathrm{r}}^{\mathrm{post}} + s_{\mathrm{e}}^{\mathrm{art}}\|^2}$	situations 1, 2

TABLEAU 5.3. – Métriques d'évaluation. Les formules sont données dans le cas monophonique ($M = 1$) et l'indice m du microphone est omis par souci de clarté.

D'après la formule $c^{\mathrm{post}} = \gamma_c \, c$, les composantes $\mathbf{s}_{\mathrm{e}}^{\mathrm{post}}$, $\mathbf{s}_{\mathrm{r}}^{\mathrm{post}}$, $\mathbf{z}_{\mathrm{r}}^{\mathrm{post}}$, $\mathbf{b}_{\mathrm{r}}^{\mathrm{post}}$ et $\mathbf{s}_{\mathrm{e}}^{\mathrm{art}}$ sont calculées à partir des vérités terrain des signaux $\mathbf{s}_{\mathrm{e}}$, $\mathbf{s}_{\mathrm{l}}$, $\mathbf{y}$ et $\mathbf{b}$ (voir la partie 2.4). La procédure de génération des données fournit immédiatement les vérités terrain de l'écho $\mathbf{y}$ et du signal de bruit $\mathbf{b}$. Pour définir la parole cible $\mathbf{s}_{\mathrm{e}}$ et la vérité terrain de la réverbération tardive $\mathbf{s}_{\mathrm{l}}$, nous fixons le *temps de mélange* $t_{\mathrm{e}} = 64$ ms (voir la partie 2.1.1.2). Nous calculons ces deux composantes à l'aide de (2.5). Dans l'ensemble de test, comme la vérité terrain de la RIR de la parole locale $\mathbf{a}_s(\tau)$ est inconnue, nous utilisons la méthode de Yoshioka et al. [2011], qui estime la RIR de la parole locale $\mathbf{a}_s(\tau)$ en optimisant le critère MMSE entre la parole locale $\mathbf{s}$ et la parole locale anéchoïque u,

comme pour déterminer le filtre d'annulation d'écho $\mathcal{H}(f)$ en réduction d'écho (voir la partie 2.2.2.1).

5.3.4. Méthodes de référence

Nous comparons notre approche en ligne avec deux méthodes de référence : 1) l'approche hors-ligne proposée au chapitre 4, et 2) une approche en cascade où le filtre d'annulation d'écho $\mathcal{H}(n, f)$, le filtre de déréverbération $\mathcal{G}(n, f)$ et le post-filtre court $\mathbf{W}_{s_e}(n, f)$ sont estimés en ligne et appliqués l'un après l'autre. L'annulation d'écho en ligne est réalisée avec SpeexDSP [1], qui est une implémentation de l'approche adaptative de Valin [2007] (voir la partie 2.2.2). La déréverbération linéaire en ligne est réalisée avec notre extension de la méthode en ligne DNN-WPE de Heymann et al. [2018], que nous détaillons dans le paragraphe suivant. Le post-filtre court $\mathbf{W}_{s_e}(n, f)$ est estimé en utilisant notre implémentation d'une version en ligne de l'algorithme DNN-EM de Nugraha et al. [2016a].

La méthode DNN-WPE de Heymann et al. [2018] n'a été définie que pour le problème à un locuteur. Cette méthode en ligne estime le filtre $\mathcal{G}(n, f)$ à partir d'une estimation de la DSP cible $v_{s_e}(n, f) = \frac{1}{M}\|\mathbf{s}_e(n, f)\|^2$ par un DNN. L'entrée de ce DNN est l'entrée de type I $|\widetilde{s}(n, f)| = \sqrt{\frac{1}{M}\|\mathbf{s}(n, f)\|^2}$. Dans notre cas, nous avons deux locuteurs et une source de bruit. Pour pouvoir appliquer cette méthode à notre cas, nous définissons la DSP cible $v_r(n, f) = \frac{1}{M}\|\mathbf{r}(n, f)\|^2$. Le signal $\mathbf{r}(n, f)$ est ici obtenu par application successive sur le mélange $\mathbf{d}(n, f)$ des filtres linéaires $\mathcal{H}(f)$ et $\mathcal{G}(f)$ « idéaux ». Ces filtres linéaires « idéaux » correspondent à des filtres ayant déjà convergé, comme pour l'approche en cascade hors-ligne considérée dans la partie 4.3.4. Pour obtenir ces filtres linéaires « idéaux », nous considérons une annulation d'écho hors-ligne et une déréverbération linéaire hors-ligne. Pour l'annulation d'écho hors-ligne, nous utilisons SpeexDSP « hors ligne » (voir la partie 4.3.5.1). Pour la déréverbération hors-ligne, nous utilisons la méthode WPE (hors ligne). L'entrée du DNN est l'entrée de type I $|\widetilde{e}(n, f)| = \sqrt{\frac{1}{M}\|\mathbf{e}(n, f)\|^2}$.

5.3.5. Réglage des hyperparamètres

Le réglage des hyperparamètres des trois approches est décrit dans les sous-parties suivantes.

5.3.5.1. Hyperparamètres de la version en ligne de l'algorithme DNN-BCA

Pour l'algorithme DNN-BCA en ligne, la TFCT est calculée avec une fenêtre de Hanning de longueur $T_{\text{TFCT}} = 1\,024$ échantillons et un pas d'avancement $P = 256$ échantillons, pour obtenir $F = 513$ bandes de fréquence. La longueur du filtre d'annulation d'écho $\mathcal{H}(n, f)$ est fixé à $K = 10$ trames ($0,208$ s dans le domaine temporel). Nous fixons la longueur du filtre de déréverbération $\mathcal{G}(n, f)$ à $0,208$ s dans le domaine temporel, soit $L = 10$ trames, avec un délai $\Delta = 3$ trames. Durant la phase de test, par souci

1. https ://github.com/xiph/speexdsp

de complexité algorithmique, nous réalisons seulement $I = 1$ itération de l'algorithme DNN-BCA en ligne. Ainsi, nous n'utilisons que DNN_0, dont nous détaillons l'apprentissage dans la sous-partie suivante. Les facteurs d'oubli des filtres linéaires $\mathcal{H}(n, f)$ et $\mathcal{G}(n, f)$ sont fixés à $\alpha_h = \alpha_g = 0,995$, et celui des MCSs $\mathbf{R}_c(n, f)$ à $\alpha_c = 0,90$. Ces valeurs permettent un compromis entre la vitesse de convergence et la stabilité des filtres linéaires et des MCS. Nous considérons 1 mise à jour spectrale et 1 mise à jour spatiale de l'algorithme DNN-BCA en ligne (voir la partie 5.2.3.2).

5.3.5.2. Hyperparamètres du modèle spectral

Hyperparamètres de DNN_0 Nous choisissons 1 026 neurones pour la couche cachée de l'architecture LSTM. L'apprentissage de DNN_0 est réalisé par rétropropagation avec une taille de mini-lots de 16 séquences, une taille de séquence de 32 trames et l'algorithme Adam avec les réglages par défaut pour l'optimisation [Kingma et Ba, 2015]. Pour éviter l'explosion du gradient avec les longues séquences, nous utilisons l'écrêtage du gradient avec un seuil à $1,0$. L'apprentissage est arrêté lorsque la fonction de coût de l'ensemble de validation ne diminue plus après 5 cycles. Il convient de souligner que la structure de DNN_0 est identique pour l'approche en ligne proposée et l'approche en ligne en cascade. Toutefois, les cibles $v_c(n, f)$ sont différentes car elles ne correspondent pas aux mêmes filtres linéaires $\mathcal{H}(f)$ et $\mathcal{G}(f)$ « idéaux », c'est-à-dire s'ils avaient déjà convergé.

Initialisation des filtres linéaires du modèle spectral Comme expliqué à la partie 5.2.4, nous avons besoin des valeurs des filtres linéaires $\mathcal{H}^{(0)}_{\text{DNN}}(n, f)$ et $\mathcal{G}^{(0)}_{\text{DNN}}(n, f)$ à la trame courante n pour obtenir les entrées de DNN_0. Pour l'annulation d'écho, nous calculons $\mathcal{H}^{(0)}_{\text{DNN}}(n, f)$ en appliquant SpeexDSP sur chaque canal m du mélange $\mathbf{d}(n, f)$, car SpeexDSP est conçu pour le traitement monocanal. Puisque SpeexDSP est basé sur des fenêtres de TFCT rectangulaires avec un chevauchement à 50%, nous utilisons une taille de trame $T_{\text{TFCT}} = 512$ échantillons et un pas d'avancement $P = 256$ échantillons. Nous fixons la longueur du filtre d'annulation d'écho à $0,208$ s dans le domaine temporel, soit $K = 13$ trames. Cette configuration permet un compromis entre réduction d'écho, et complexité de l'algorithme. SpeexDSP donne le signal de sortie $\mathbf{e}(t)$ dans le domaine temporel. Comme SpeexDSP est un algorithme en ligne, nous l'appliquons ici une seule fois à chaque enregistrement. À titre de rappel, au chapitre 4, nous appliquions SpeexDSP 2 fois par enregistrement pour approcher un fonctionnement hors ligne (voir la partie 4.3.5.1).

Pour la déréverbération linéaire, nous calculons $\mathcal{G}^{(0)}_{\text{DNN}}(n, f)$ en appliquant notre extension de la méthode en ligne DNN-WPE de Heymann et al. [2018] sur le signal après annulation d'écho $\mathbf{e}(n, f)$ (voir la partie 5.3.4). Pour cette opération, nous utilisons la TFCT avec une fenêtre de Hanning de taille $T_{\text{TFCT}} = 1\,024$ échantillons et un pas d'avancement $P = 256$ échantillons. Nous fixons la longueur du filtre de déréverbération $\mathcal{G}^{(0)}_{\text{DNN}}(n, f)$ à $0,208$ s dans le domaine temporel, soit $L = 10$ trames, avec un délai $\Delta = 3$ trames. Notre extension de la méthode DNN-WPE utilise un DNN pour estimer la DSP cible $v_r(n, f)$ (voir la partie 5.3.4). Nous détaillons l'apprentissage de ce DNN dans la

sous-partie suivante.

Phase d'apprentissage de notre extension de DNN-WPE Pour l'apprentissage du DNN de notre extension de la méthode DNN-WPE, nous utilisons le même ensemble d'apprentissage que pour DNN_0. Nous utilisons les mêmes hyperparamètres de DNN que pour DNN_0. Nous choisissons une procédure hors-ligne pour l'apprentissage du DNN. Pour cela, nous déterminons la DSP cible $v_r(n,f)$ du DNN en appliquant des filtres linéaires $\mathcal{H}(f)$ et $\mathcal{G}(f)$ « idéaux », c'est-à-dire ayant déjà convergé. Pour obtenir un filtre d'annulation d'écho $\mathcal{H}(f)$ « idéal » avec l'algorithme en ligne SpeexDSP, nous appliquons SpeexDSP 2 fois à chaque enregistrement afin d'obtenir la convergence de l'algorithme et ainsi approcher le fonctionnement d'un algorithme hors ligne. Pour obtenir le filtre de déréverbération $\mathcal{G}(f)$ « idéal », nous réalisons 3 itérations de l'algorithme hors-ligne WPE sur le signal $\mathbf{e}(n,f)$ après annulation d'écho avec le filtre « idéal » $\mathcal{H}(f)$. Durant l'apprentissage, l'entrée du DNN $|\widetilde{e}(n,f)| = \sqrt{\frac{1}{M}\|\mathbf{e}(n,f)\|^2}$ est calculée après annulation d'écho avec le filtre « idéal » $\mathcal{H}(f)$.

Phase d'apprentissage de DNN_0 Nous choisissons la même procédure d'apprentissage de DNN_0 que celle utilisée dans la version hors ligne de l'algorithme DNN-BCA (voir la partie 4.3.5.3). À titre de rappel, nous utilisions des filtres linéaires et les MCSs « idéaux », c'est-à-dire ayant déjà convergé. Nous avions donc $\mathcal{H}^{(0)}(n,f) = \mathcal{H}^{(0)}(f)$, $\mathcal{G}^{(0)}(n,f) = \mathcal{G}^{(0)}(f)$ et $\mathbf{R}_c^{(0)}(n,f) = \mathbf{R}_c^{(0)}(f)$. Les filtres linéaires « idéaux »$\mathcal{H}^{(0)}(f)$ et $\mathcal{G}^{(0)}(f)$ sont calculés avec SpeexDSP « hors-ligne » et WPE, respectivement. Pour obtenir un filtre d'annulation d'écho $\mathcal{H}^{(0)}(f)$ « idéal » avec l'algorithme en ligne SpeexDSP, nous appliquons celui-ci 2 fois à chaque enregistrement afin d'obtenir la convergence de l'algorithme et ainsi approcher le fonctionnement d'un algorithme hors ligne. Pour obtenir le filtre de déréverbération « idéal » $\mathcal{G}^{(0)}(f)$, nous réalisons 3 itérations de l'algorithme hors-ligne WPE sur le signal $\mathbf{e}(n,f)$ après annulation d'écho avec $\mathcal{H}_0(f)$. Pour déterminer les cibles de DNN_0, nous réalisons 3 itérations de la procédure de détermination des vérités terrain des DSPs, utilisée dans la version hors-ligne de l'algorithme DNN-BCA (voir la partie 4.2.4.1).

5.3.5.3. Hyperparamètres de la version en ligne de l'approche en cascade

Concernant l'approche en cascade en ligne, nous calculons et fixons les filtres linéaires à la valeur $\mathcal{H}(n,f) = \mathcal{H}^{(0)}_{\text{DNN}}(n,f)$ et $\mathcal{G}(n,f) = \mathcal{G}^{(0)}_{\text{DNN}}(n,f)$ avec les mêmes hyperparamètres que l'approche proposée. Durant la phase de test, par souci de complexité algorithmique, nous réalisons seulement 1 itération à chaque trame n de la version en ligne de l'algorithme DNN-EM. Ainsi, nous n'utilisons qu'un seul DNN, dont l'architecture, les entrées et la phase d'apprentissage sont identiques à celles de DNN_0. Nous choisissons la même procédure d'apprentissage du DNN que pour DNN_0, ce qui correspond à la phase d'apprentissage de la version hors ligne de l'approche en cascade (voir la partie 4.3.5.5). La vérité terrain des DSPs est calculée avec la même procédure que pour DNN_0, où les filtres linéaires « idéaux » sont fixés à $\mathcal{H}(f) = \mathcal{H}^{(0)}(f)$ et $\mathcal{G}(f) = \mathcal{G}^{(0)}(f)$. Les

filtres linéaires « idéaux »$\mathcal{H}^{(0)}(f)$ et $\mathcal{G}^{(0)}(f)$ sont calculés avec SpeexDSP « hors-ligne » et WPE, respectivement.

5.3.5.4. Hyperparamètres de la version hors-ligne de l'algorithme DNN-BCA

Concernant la version hors-ligne de l'algorithme DNN-BCA, nous réalisons $I = 1$ itération afin d'obtenir une équivalence d'estimation avec la version en ligne de l'algorithme DNN-BCA. Sur l'ensemble de test invariant dans le temps du chapitre 4, l'algorithme hors-ligne donnait un SI-SDR $= -1,0$ dB $\pm 0,1$ dB avec $I = 3$ itérations, et un SI-SDR $= -2,3$ dB $\pm 0,1$ dB avec $I = 1$ itération, pour le scénario en présence de bruit. Nous utilisons les mêmes hyperparamètres de filtres linéaires que pour la version en ligne de l'algorithme DNN-BCA. Nous utilisons DNN_0 pour le modèle spectral. Dans les ensembles de test, les entrées de DNN_0 sont calculées à partir des filtres linéaires $\mathcal{H}^{(0)}(f)$ et $\mathcal{G}^{(0)}(f)$ « idéaux », c'est-à-dire ayant déjà convergé. Pour obtenir un filtre d'annulation d'écho $\mathcal{H}(f)$ « idéal » avec l'algorithme en ligne SpeexDSP, nous appliquons SpeexDSP 2 fois à chaque enregistrement afin d'obtenir la convergence de l'algorithme et ainsi approcher le fonctionnement d'un algorithme hors ligne. Pour obtenir le filtre de déréverbération $\mathcal{G}(f)$ « idéal », nous réalisons 3 itérations de l'algorithme hors-ligne WPE sur le signal $\mathbf{e}(n, f)$ après annulation d'écho avec le filtre « idéal » $\mathcal{H}(f)$. Cette configuration des filtres linéaires correspond à celles de la phase d'apprentissage de DNN_0. Par conséquent, les estimations de DNN_0 dans cette configuration sont donc censées être meilleures que pour la version en ligne de l'algorithme DNN-BCA. En effet, dans le cas de la version en ligne, les filtres $\mathcal{H}^{(0)}_{\text{DNN}}(n, f)$ et $\mathcal{G}^{(0)}_{\text{DNN}}(n, f)$ n'ont pas nécessairement convergé dans les ensembles de test.

5.3.5.5. Régularisation

Afin d'éviter des instabilités numériques et des matrices mal conditionnées, nous ajoutons un scalaire de régularisation ϵ au dénominateur dans (5.40) et (5.42) et une matrice de régularisation $\epsilon\mathbf{I}$ à la matrice à inverser dans (5.11), (5.27) et (5.36). Nous régularisons aussi la fonction de coût d'apprentissage $\mathcal{D}_{KL}$ comme dans (4.59). Nous régularisons l'approche en cascade de la même manière. Le paramètre de régularisation est fixé à $\epsilon = 10^{-5}$.

5.4. Résultats et discussion

Dans cette partie, nous comparons l'approche en ligne proposée pour la réduction conjointe de bruit, d'écho et de réverbération à l'approche hors-ligne proposée au chapitre 4 et l'approche en cascade. Nous analysons tout d'abord les résultats des trois approches dans des conditions acoustiques invariantes dans le temps. Enfin, nous discutons de leurs performances lorsque les conditions acoustiques varient au cours du temps et nous comparons leur temps de calcul.

5.4.1. Conditions invariantes au cours du temps

La figure 5.6 compare la moyenne des performances de l'approche en ligne proposée à celles de l'approche hors-ligne proposée et l'approche en cascade en ligne dans des conditions acoustiques invariantes dans le temps. L'approche hors-ligne proposée dépasse les deux approches en ligne pour toutes les métriques. De plus, c'est la seule méthode qui obtient un SI-SDR positif. Ceci s'explique par le fait que l'approche hors-ligne est adaptée aux conditions acoustiques invariantes au cours du temps, qu'elle exploite l'enregistrement complet, et qu'elle se base sur des filtres ayant déjà convergé pour l'initialisation. De plus, dans l'ensemble de test, les entrées du modèle spectral de l'approche hors-ligne ont été calculées à partir d'algorithmes hors-ligne (SpeexDSP « hors-ligne » et WPE), ce qui correspond à la configuration d'apprentissage de DNN_0. À l'inverse, les entrées du modèle spectral des deux approches en ligne ont été calculées à partir d'algorithmes en ligne (SpeexDSP et extension de DNN-WPE), ce qui est différent des conditions d'apprentissage de DNN_0. Cela affecte la réduction d'écho de manière flagrante : les SER et ERLE sont nettement plus élevés pour l'approche hors-ligne proposée que pour les deux approches en ligne.

L'approche en ligne proposée dépasse l'approche en cascade en ligne pour toutes les métriques. En particulier, elle dépasse l'approche en cascade en ligne d'environ $0,5$ dB en SI-SDR. La performance en SI-SDR par rapport à l'approche en cascade s'explique par le fait que l'approche proposée est meilleure que l'approche en cascade selon toutes les autres métriques. Cela confirme que l'optimisation conjointe permet d'améliorer les performances pour la réduction en ligne de bruit, d'écho et de réverbération lorsque les conditions acoustiques sont invariantes dans le temps. Par ailleurs, nous avons remarqué que le filtre d'annulation d'écho $\mathcal{H}^{(0)}(n, f)$ convergeait plus vite avec l'approche en ligne proposée qu'avec l'approche en cascade. La figure 5.7 illustre un exemple de signal $\mathbf{r}(n, f)$ estimé avec l'approche en cascade en ligne et l'approche en ligne proposée. Nous pouvons remarquer que l'écho résiduel *déréverbéré* $\mathbf{z}_{\text{r}}(n, f)$ est nettement plus faible avec l'approche en ligne proposée. Ceci traduit une convergence du filtre d'annulation d'écho $\mathcal{H}^{(0)}(n, f)$ plus rapide.

La figure 5.8 illustre un exemple de spectrogramme de la parole cible estimée $\widehat{\mathbf{s}}_{\text{e}}(n, f)$ par l'approche en ligne proposée et les approches de référence. Nous pouvons constater que le spectrogramme de l'approche hors-ligne proposée est plus proche de celui de la parole cible $\widehat{\mathbf{s}}_{\text{e}}(n, f)$ que pour les approches en ligne. Les spectrogrammes des deux approches en ligne sont nettement moins bien estimés. L'approche en ligne proposée semble cependant réduire plus les distorsions résiduelles que l'approche en cascade en ligne, ce qui confirme les tendances observées sur les métriques.

5.4.2. Conditions variant au cours du temps

La figure 5.9 présente la moyenne des résultats lorsque les conditions acoustiques varient au cours du temps. L'approche hors-ligne proposée dépasse les deux approches selon toutes les métriques. Comme dans la sous-partie précédente, ceci s'explique par le fait la configuration de test des filtres linéaires utilisés dans le modèle spectral de l'approche

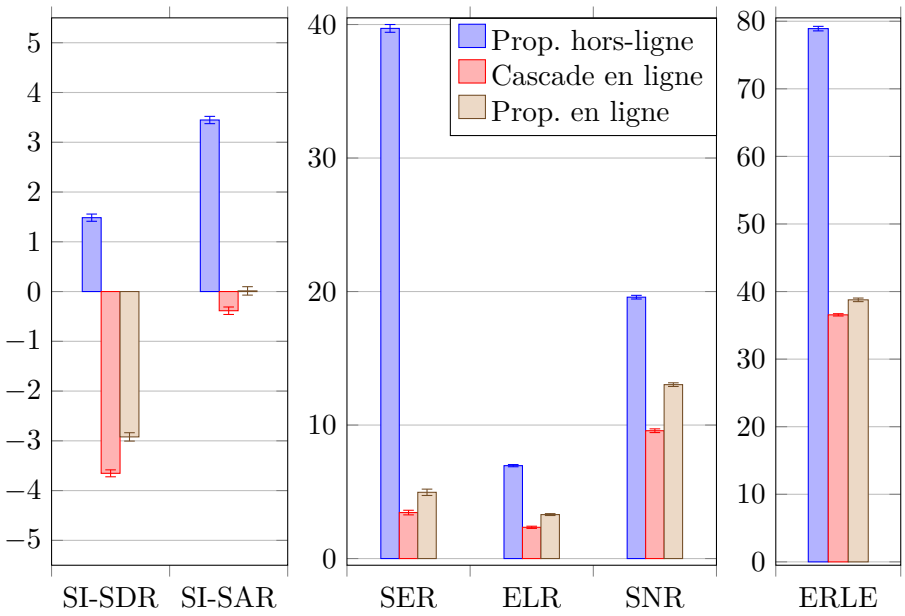
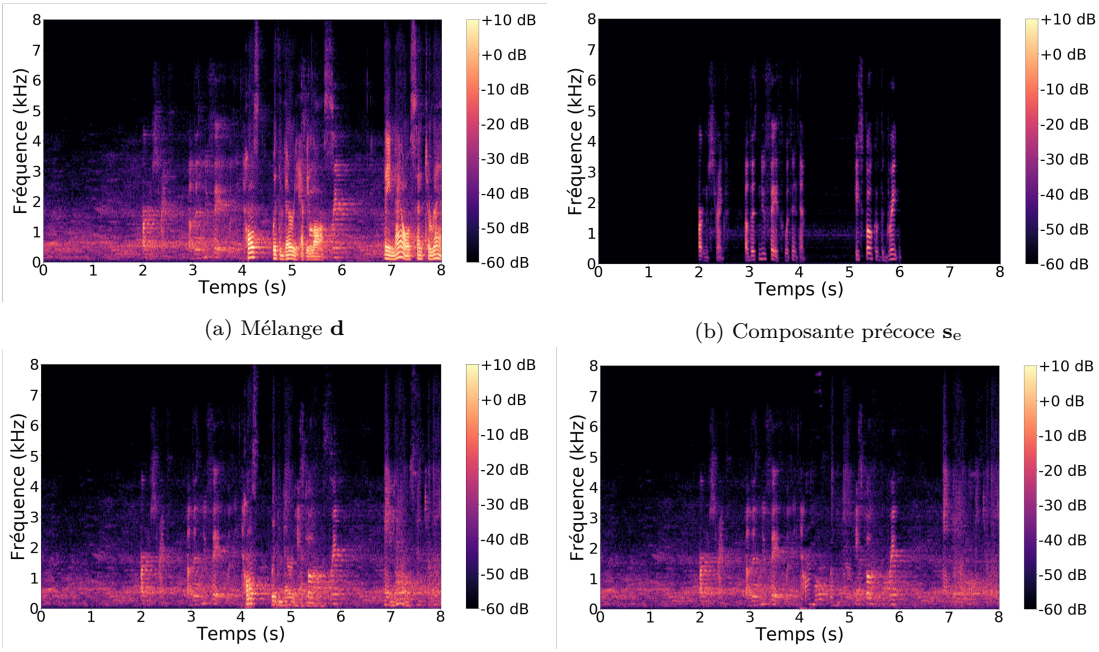


FIGURE 5.6. – Performances moyennes (en dB) des trois approches en conditions acoustiques invariantes dans le temps.

hors-ligne (SpeexDSP « hors-ligne » et WPE) proposée est proche de la configuration d'apprentissage du modèle spectral. Dans cette configuration, les filtres linéaires ont déjà convergé pour l'initialisation. De plus, l'approche hors-ligne proposée exploite l'enregistrement complet. Toutefois, ses performances sont plus faibles que lorsque les conditions acoustiques sont invariantes dans le temps. Ceci s'explique par le fait que les propriétés spatiales de la parole cible $\mathbf{s}_{\mathrm{e}}(n,f)$ et de la réverbération résiduelle $\mathbf{s}_{\mathrm{r}}(n,f)$ varient au cours du temps, alors que leurs MCSs $\mathbf{R}_c(f)$ dans l'approche hors-ligne restent fixes.

Les deux approches en ligne obtiennent de meilleures performances en SI-SDR, SI-SAR, SER et ERLE que dans l'exprérience précédente. Ceci s'explique par le fait que la durée de chaque enregistrement est ici de 16 s, et non de 8 s comme dans l'ensemble de test invariant au cours du temps. Sur la deuxième période de 8 s, les filtres linéaires $\mathcal{H}_{\mathrm{DNN}}(n,f)$ et $\mathcal{G}_{\mathrm{DNN}}(n,f)$, utilisés pour calculer les entrées du modèles spectral des deux approches en ligne, ont mieux convergé que dans la première période de 8 s. Les entrées du modèle spectral des deux approches en ligne correspondent donc plus à la configuration d'apprentissage de DNN_0.

L'approche en ligne proposée dépasse l'approche en cascade en ligne en SI-SDR, ELR et SNR. En particulier, elle dépasse l'approche en cascade en ligne d'environ $0,5$ dB en SI-SDR. Cela confirme que l'optimisation conjointe permet d'améliorer les performances en réduction globale de la distorsion lorsque les conditions acoustiques varient au cours du temps. Toutefois, l'approche en ligne proposée est nettement dépassée par l'approche en cascade en SER et ERLE, ce qui traduit une réduction d'écho moins importante pour l'approche en ligne proposée. Les performances en SER sont cohérentes avec un SI-SAR



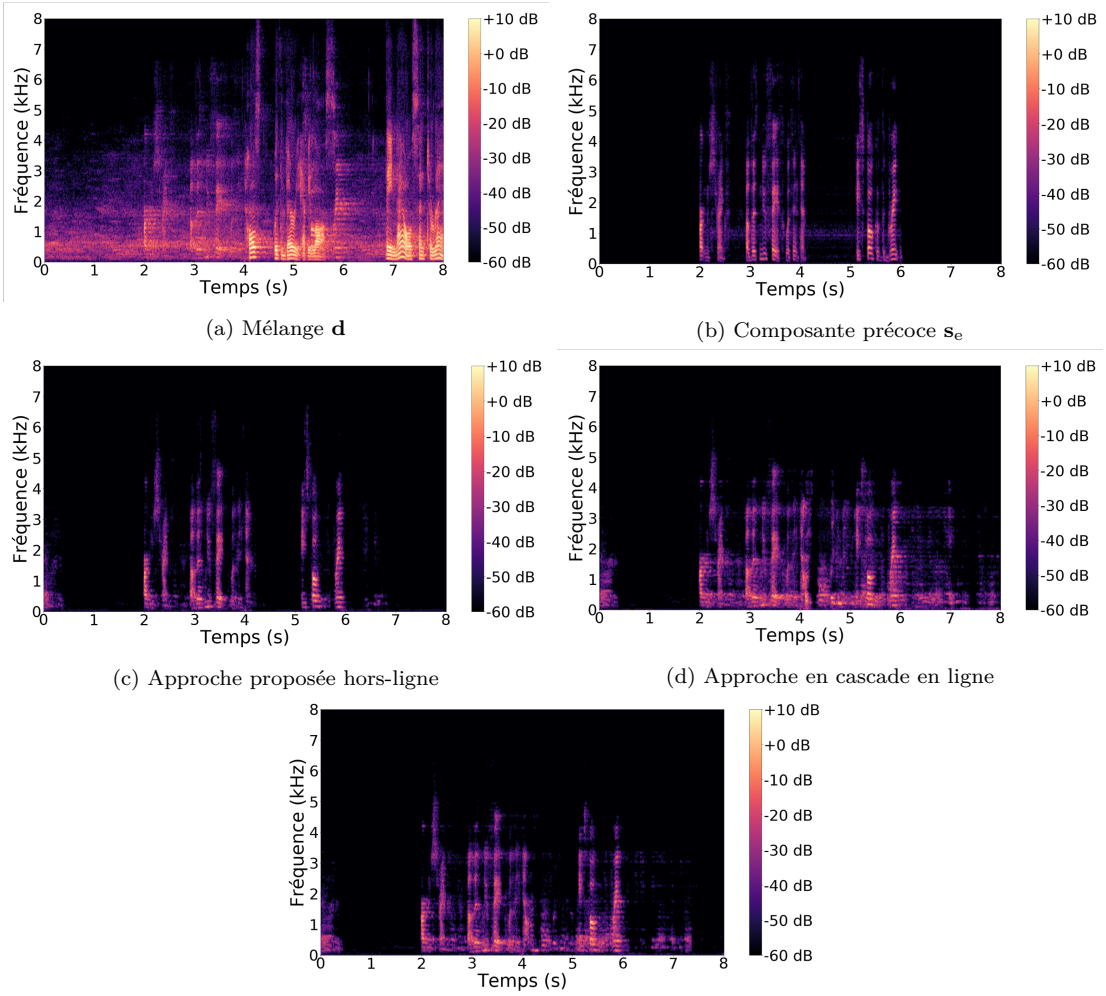
(a) Mélange $\mathbf{d}$

(b) Composante précoce $\mathbf{s}_{\mathrm{e}}$

(c) Signal $\mathbf{r}$ avec l'approche en cascade

(d) Signal $\mathbf{r}$ avec l'approche en ligne proposée

FIGURE 5.7. – Exemple de spectrogrammes du signal $\mathbf{r}$ estimé avec l'approche en cascade en ligne et l'approche en ligne proposée, pour le scénario où les conditions acoustiques sont invariantes dans le temps (seul un canal est illustré).



(a) Mélange $\mathbf{d}$

(b) Composante précoce $\mathbf{s}_{\mathrm{e}}$

(c) Approche proposée hors-ligne

(d) Approche en cascade en ligne

(e) Approche en ligne proposée

FIGURE 5.8. – Exemple de spectrogrammes de la composante précoce estimée $\widehat{\mathbf{s}}_{\mathrm{e}}$ avec les approches de référence et l'approche proposée, pour le scénario où les conditions acoustiques sont invariantes dans le temps (seul un canal est illustré).

équivalent pour les deux approches en ligne et un ERLE plus élevé pour l'approche en cascade en ligne. Cela suggère donc que l'approche en cascade en ligne est plus agressive que l'approche en ligne proposée en réduction d'écho, sans introduire plus d'artefacts dans la parole cible $\mathbf{s}_{\mathrm{e}}(n,f)$. Toutefois, l'approche en ligne proposée parvient à obtenir de meilleures performances en réduction de la distorsion globale, grâce à une meilleure réduction de bruit et de réverbération.

La figure 5.10 illustre un exemple de spectrogramme de la parole cible estimée $\widehat{\mathbf{s}}_{\mathrm{e}}(n,f)$ par l'approche proposée et les approches de référence. Nous pouvons constater que le spectrogramme de l'approche hors-ligne proposée est plus proche de celui de la parole cible $\widehat{\mathbf{s}}_{\mathrm{e}}(n,f)$ que pour les approches en ligne. Les spectrogrammes des deux approches en ligne sont nettement moins bien estimés. L'approche en ligne proposée semble atténuer plus les distorsions résiduelles que l'approche en cascade en ligne.

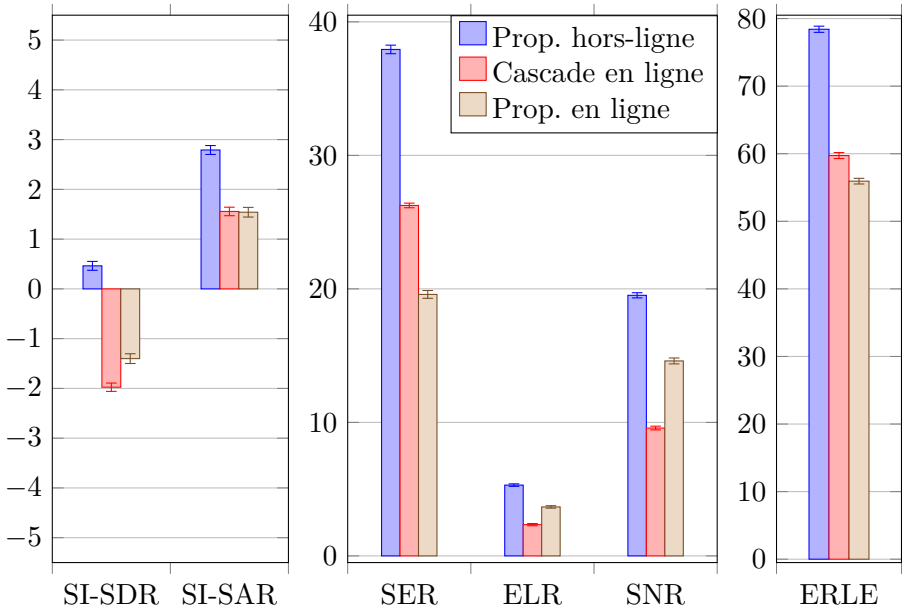
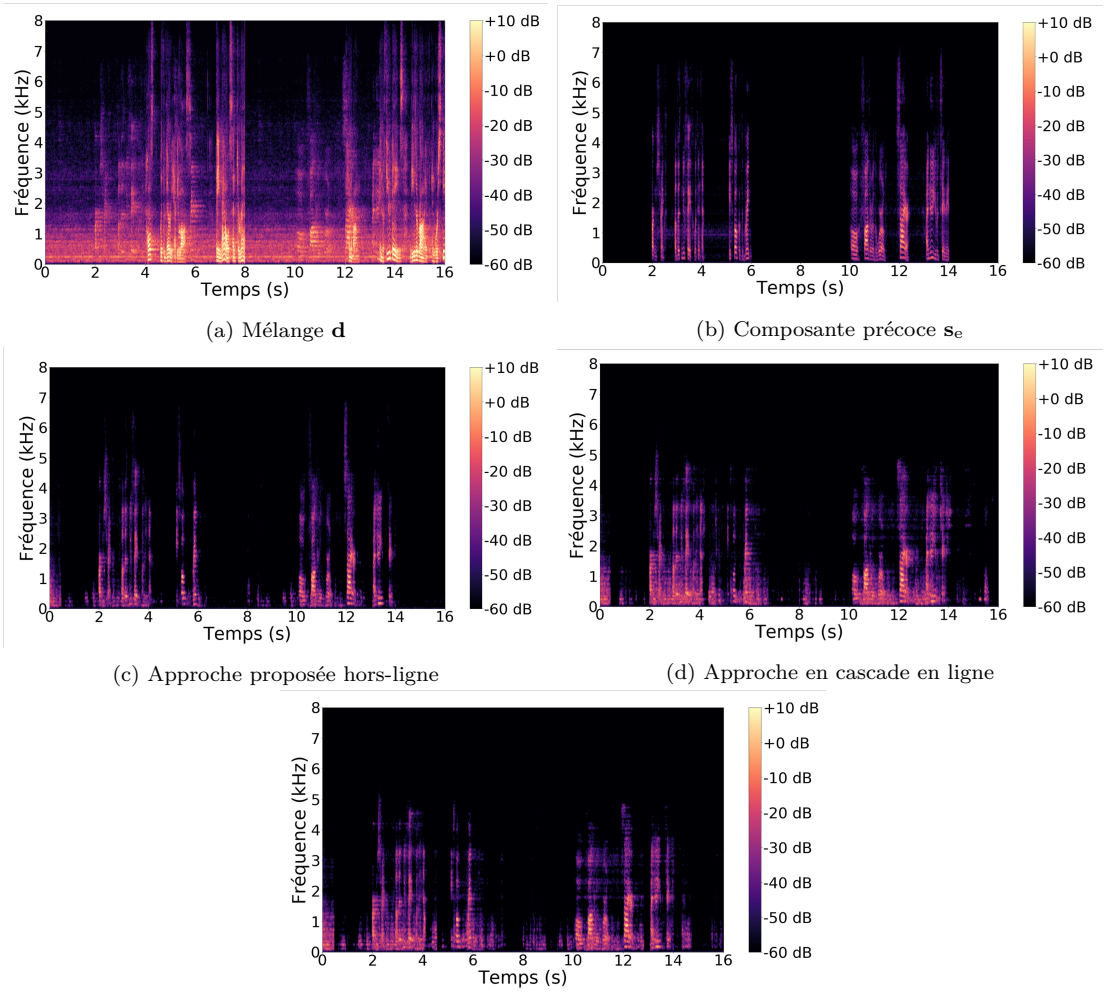


FIGURE 5.9. – Performances moyennes (en dB) des trois approches en conditions acoustiques variant au cours du temps.

5.4.3. Temps de calcul

Nous évaluons le temps de calcul des deux approches en ligne. Nous ne considérons pas le temps de calcul de l'approche hors-ligne proposée car elle ne correspond pas à une méthode fonctionnant en temps réel. Nous ne considérons pas l'initialisation des filtres linéaires $\mathcal{H}^{(0)}_{\mathrm{DNN}}(n,f)$ et $\mathcal{G}^{(0)}_{\mathrm{DNN}}(n,f)$ car elle représente la même opération pour les deux approches en ligne. Le tableau 5.4 détaille le temps d'estimation de la parole cible $\widehat{\mathbf{s}}_{\mathrm{e}}$ sur un enregistrement de 8 s, avec un processeur Intel Core i5 à 2,7 GHz. Comme l'estimation des DNNs dépend de l'architecture, nous décrivons aussi le temps de calcul



(a) Mélange $\mathbf{d}$

(b) Composante précoce $\mathbf{s}_{\mathrm{e}}$

(c) Approche proposée hors-ligne

(d) Approche en cascade en ligne

(e) Approche en ligne proposée

FIGURE 5.10. – Exemple de spectrogrammes de la composante précoce estimée $\widehat{\mathbf{s}}_{\mathrm{e}}$ avec les approches de référence et l'approche proposée, pour le scénario où les conditions acoustiques varient au cours du temps (seul un canal est illustré).

propre aux algorithmes de mises à jour des filtres en excluant le temps d'estimation des DSPs $v_c(n,f)$ avec les DNNs, et le temps de calcul propre à l'estimation des DSPs $v_c(n,f)$ avec les DNNs. En excluant l'estimation des DSPs $v_c(n,f)$ avec les DNNs, l'approche en ligne proposée prend environ 20 fois plus de temps que l'approche en cascade en ligne. En effet, les mises à jour des filtres linéaires $\mathcal{H}^{(0)}(n,f)$ et $\mathcal{G}^{(0)}(n,f)$ dans l'approche en ligne proposée impliquent notamment des inversions matricielles dans (5.27) et (5.36).

En ce qui concerne l'estimation des DSPs $v_c(n,f)$ avec les DNNs, l'approche en ligne proposée a le même temps de calcul que l'approche en cascade en ligne, car les DNNs ont le même nombre de paramètres, et les approches utilisent le même nombre de DNNs.

En prenant en compte l'estimation des DSPs $v_c(n,f)$ avec les DNNs, et par rapport à l'approche en cascade, l'estimation de la parole cible $\mathbf{s}_e$ prend environ 3 fois plus de temps par l'approche proposée. La complexité de l'approche en ligne proposée pourrait être diminuée en supposant que certaines MCSs $\mathbf{R}_c(n,f)$ sont diagonales ou égales à la matrice identité $\mathbf{I}_M$ pour l'estimation des filtres linéaires $\mathcal{H}(n,f)$ et $\mathcal{G}(n,f)$. Ainsi, les inversions matricielles dans (5.27) et (5.36) se ramèneraient à de simples divisions scalaires. Cette simplification est notamment utilisée dans la méthode DNN-WPE en déréverbération linéaire [Heymann et al., 2018]. Toutefois, il faudrait évaluer l'impact de cette simplification sur les performances de réduction.

Élément de l'algorithme	**Cascade en ligne**	**Prop. en ligne**
Mises à jour des filtres (DNNs exclus)	$2,4 \pm 0,1$	$43,4 \pm 0,2$
DNNs	$20,0 \pm 0,4$	$20,4 \pm 0,6$
Total	$22,4 \pm 0,3$	$63,8 \pm 0,3$

TABLEAU 5.4. – Temps de calcul (en s) de l'approche en cascade en ligne et de l'approche en ligne proposée sur un signal de 8 s. Les deux chiffres représentent la moyenne et l'écart-type.

5.5. Résumé

Dans ce chapitre, nous avons proposé une version en ligne de l'algorithme de montée par blocs de coordonnées du chapitre 4 pour la réduction conjointe multicanale de l'écho acoustique, la réverbération et le bruit. Cette variante en ligne modélise les filtres d'annulation d'écho et de déréverbération, ainsi que les caractéristiques spatiales de la parole cible et des signaux résiduels après annulation d'écho et déréverbération, comme des paramètres qui dépendent du temps. La méthode en ligne proposée estime conjointement à l'aide d'un DNN les spectres de la parole cible et des signaux résiduels après l'annulation d'écho et la déréverbération linéaire. Les filtres linéaires et le post-filtre sont estimés en mettant à jour leurs paramètres de manière récursive. Pour réduire la complexité d'estimation des filtres d'annulation d'écho et de déréverbération linéaire à chaque trame, nous avons proposé des mises à jour de leurs paramètres utilisant l'identité de Woodbury. Nous avons évalué notre système sur des enregistrements réels d'écho acoustique,

de réverbération et de bruit acquis avec un Triby dans plusieurs situations différentes. L'approche proposée dépasse l'approche en cascade en ligne en réduction de distorsion globale. Ceci confirme que l'optimisation conjointe permet d'améliorer les performances en réduction de la distorsion globale pour la réduction en ligne de bruit, d'écho et de réverbération. Toutefois, la complexité de l'approche en ligne proposée est plus élevée que pour l'approche en cascade en ligne. La complexité pourrait être diminuée à l'aide de simplifications sur la mise à jour des filtres linéaires, ce qui est susceptible de diminuer les performances de réduction. Par ailleurs, les performances de la version en ligne de l'approche proposée sont nettement inférieures à la version hors-ligne. Ceci vient du fait que les performances de l'approche en ligne proposée sont dépendantes de la configuration d'apprentissage du DNN. Dans ce chapitre, le DNN a été entraîné avec des filtres linéaires ayant convergé, alors que ce n'est pas toujours le cas en conditions réelles, et en particulier dans la phase de test considérée ici. Dans le chapitre 3, le DNN de la méthode proposée avait été entraîné avec un filtre d'annulation d'écho qui variait au cours du temps. Ceci suggère que la méthode en ligne proposée ici pourrait être améliorée en considérant une configuration d'apprentissage du DNN où les filtres linéaires varient au cours du temps.

6. Conclusion et perspectives

6.1. Conclusion

Compte tenu du contexte industriel de cette thèse, liée à l'amélioration de l'enceinte intelligente Triby développée par la société Invoxia, il s'agissait de mettre au point un algorithme en ligne de rehaussement de la parole permettant de réduire simultanément le bruit, l'écho acoustique et la réverbération. Les résultats de la thèse sont potentiellement applicables à tout système de communication mains-libres et non seulement au Triby.

Dans le chapitre 3, nous avons proposé une méthode monocanale de suppression d'écho résiduel. Ce type de méthode consiste à appliquer un post-filtre court visant à supprimer l'écho résiduel subsistant après l'annulation d'écho avec un filtre long. Pour estimer le post-filtre, les méthodes existantes supposent généralement que le filtre d'annulation a convergé. Par conséquent, elles modélisent l'écho résiduel comme un signal faible par rapport à la parole cible. Dans des scénarios réels, le filtre d'annulation d'écho n'a pas nécessairement convergé, ce qui amène l'écho résiduel à être potentiellement important durant ces périodes. Le post-filtre de suppression d'écho résiduel doit donc être adapté à la fois aux périodes antérieures et postérieures à la convergence du filtre d'annulation d'écho.

Pour cela, la méthode proposée estime les coefficients du post-filtre à l'aide d'un réseau de neurones combinant deux avantages. D'une part, il prend en entrée le signal après annulation d'écho, le signal source de l'écho, et l'écho estimé par le filtre d'annulation d'écho. L'écho estimé contient de l'information liée à l'écho résiduel qui est complémentaire aux deux autres signaux d'entrée. D'autre part, le critère d'apprentissage du réseau de neurones inclut l'information de phase. Ceci améliore les performances du post-filtre, notamment avant la convergence du filtre d'annulation d'écho, où le SER est faible.

Pour évaluer la méthode proposée, nous avons mené des tests en conditions acoustiques réalistes. Comme il n'existe pas de base de données audio en accès public pour le problème de réduction d'écho, nous avons collecté des données dans plusieurs situations à partir de signaux enregistrés avec le Triby. En particulier, nous avons réalisé des enregistrements de l'écho acoustique, car c'est un type de distorsion spécifique au système mains-libres difficile à modéliser. Nous avons comparé la méthode proposée à trois méthodes d'estimation du post-filtre. Deux de ces méthodes sont des méthodes de soustraction spectrale n'utilisant qu'un seul signal d'entrée. La troisième méthode est une méthode d'estimation directe n'utilisant que deux signaux d'entrée. Aucune de ces trois méthodes n'inclut l'information de la phase dans le critère de détermination du post-filtre.

Sur la moyenne des périodes avant et après convergence du filtre, la méthode proposée dépasse les méthodes de référence d'au moins 1, 5 dB en matière de réduction de la distorsion globale et de dégradation de la parole cible. Cependant, quand l'état de convergence du filtre d'annulation d'écho est modifié par une perturbation du chemin d'écho (changement de position du Triby, locuteur local en mouvement, etc.), le post-filtre de notre méthode doit être à nouveau adapté.

Dans le chapitre 4, nous avons proposé une méthode multicanale de réduction conjointe de bruit, d'écho et de réverbération. Cette méthode repose sur l'application successive d'un filtre d'annulation d'écho, d'un filtre de déréverbération linéaire et d'un post-filtre de suppression de distorsions résiduelles. Ces trois filtres interagissent entre eux. Lorsque l'état de convergence du filtre d'annulation d'écho ou du filtre de déréverbération est modifié par une perturbation des conditions acoustiques (changement de position du Triby, déplacement du locuteur local, etc.), les filtres qui suivent doivent être à nouveau adaptés. Ainsi, les filtres doivent être estimés conjointement, ce qui nécessite de modéliser les caractéristiques des signaux transformés par ces filtres. La méthode de Togami et Kawaguchi [2014] a été la seule méthode jusqu'à maintenant à estimer conjointement tous les filtres. Toutefois, cette méthode ne modélise pas séparément chaque signal transformé par ces filtres.

Pour optimiser conjointement les filtres, nous avons modélisé la parole cible et les signaux résiduels après annulation d'écho et déréverbération linéaire par des variables gaussiennes de moyenne nulle. Les paramètres spectraux de la parole cible et les signaux résiduels après annulation d'écho et déréverbération linéaire sont estimés avec un DNN. Les paramètres spatiaux de ces quatre signaux sont estimés avec un algorithme EM. Nous avons intégré ces estimations dans un algorithme de montée par blocs de coordonnées pour mettre à jour les trois filtres.

Pour évaluer la méthode proposée, nous avons mené des tests en conditions acoustiques réalistes. Comme il n'existe pas de base de données audio en accès public pour le problème de la réduction conjointe de bruit, d'écho et de réverbération, nous avons collecté des données dans plusieurs situations à partir de signaux de parole, d'écho et de bruit enregistrés avec le Triby. Nous avons comparé la méthode proposée à une combinaison en cascade des approches de réduction individuelle et à la méthode d'estimation conjointe de Togami et Kawaguchi [2014].

Nous avons évalué les approches lorsque les conditions acoustiques sont invariantes dans le temps. Lorsque l'écho, la réverbération et le bruit sont présents simultanément, l'approche proposée dépasse les autres approches d'au moins 2 dB en matière de réduction de la distorsion globale, sans dégrader les performances lorsque seulement un ou deux types de distorsion sont présents. Cependant, les performances en réduction la distorsion globale ne sont pas nécessairement corrélés à la qualité perceptuelle de la parole cible estimée. Lorsque les conditions acoustiques varient au cours du temps, l'approche proposée dépasse aussi les deux approches de référence en réduction de la distorsion globale. Toutefois, les performances sont limitées dans le cas où les conditions acoustiques varient au cours du temps, car notre approche suppose que les filtres linéaires

sont invariants dans le temps, et que les sources ne se déplacent pas.

Dans le chapitre 5, nous avons proposé une version en ligne de la méthode multicanale de réduction conjointe de bruit, d'écho et de réverbération proposée au chapitre 4. D'une part, les méthodes en ligne ne supposent pas la connaissance des données futures pour estimer les filtres. D'autre part, elles sont adaptées aux situations où les conditions acoustiques varient au cours du temps. Toutefois, lorsque plusieurs filtres sont appliqués successivement, une combinaison en cascade d'approches de réduction individuelle, même en ligne, peut ne pas être robuste à la modification de l'état de convergence du filtre d'annulation d'écho ou du filtre de déréverbération. C'est ce que nous avons constaté au chapitre 3 avec la combinaison du filtre d'annulation d'écho et du post-filtre de suppression d'écho résiduel. Les filtres doivent donc être estimés conjointement. Ceci nécessite de modéliser les caractéristiques des signaux transformés par ces filtres, en tenant compte ici des conditions acoustiques qui varient potentiellement au cours du temps. Cependant, aucune méthode en ligne estimant conjointement les filtres d'annulation d'écho, de déréverbération linéaire et le post-filtre de suppression de distorsions résiduelles n'a été proposée.

Pour optimiser conjointement les filtres, nous utilisons le même modèle gaussien complexe multicanal pour la parole cible et les signaux résiduels après annulation d'écho et déréverbération linéaire. Les paramètres spectraux de ces quatre signaux sont estimés avec un DNN. Pour mettre à jour les paramètres spatiaux, qui dépendent ici du temps, nous considérons ici une version en ligne d'un algorithme EM. Nous avons intégré ces estimations dans une version en ligne de l'algorithme de montée par blocs de coordonnées du chapitre 4 pour mettre à jour les trois filtres. Dans cette version, les filtres d'annulation d'écho et de déréverbération linéaire dépendent du temps. Ces deux filtres sont estimés à partir de mises à jour récursives de leurs paramètres. Afin de réduire la complexité d'estimation des filtres d'annulation d'écho et de déréverbération linéaire à chaque trame, nous avons proposé des mises à jour de leurs paramètres utilisant l'identité de Woodbury.

Nous avons évalué notre méthode sur les mêmes enregistrements réels de bruit, d'écho et de réverbération que dans le chapitre 4, avec des conditions de SER en SNR moins difficiles que dans le chapitre 4. Pour l'évaluation, nous avons considéré des conditions acoustiques invariantes et variant au cours du temps. En ce qui concerne les conditions invariantes dans le temps, nous voulions notamment comparer les performances à la version hors-ligne de l'approche proposée. Nous avons comparé la méthode en ligne proposée à une combinaison en cascade des versions en ligne de méthode de réduction individuelle, ainsi que la version hors-ligne de la méthode proposée.

L'approche en ligne proposée dépasse l'approche en cascade en ligne en réduction de distorsion globale. Ceci confirme que l'optimisation conjointe permet d'améliorer les performances en réduction de la distorsion globale pour la réduction en ligne de bruit, d'écho et de réverbération. Toutefois, la complexité de l'approche en ligne proposée est plus élevée que pour l'approche en cascade en ligne. La complexité pourrait être diminuée à l'aide de simplifications sur la mise à jour des filtres linéaires, ce qui pourrait

diminuer les performances de réduction. En outre, les performances de la version en ligne de l'approche proposée sont nettement inférieures à la version hors-ligne. Ceci vient du fait que les performances de l'approche en ligne proposée sont dépendantes de la configuration d'apprentissage du DNN. Or, le DNN a été entraîné avec des filtres linéaires ayant convergé, ce qui n'est pas nécessairement le cas en conditions réelles. Dans le chapitre 3, le DNN de la méthode proposée avait été entraîné avec un filtre d'annulation d'écho qui variait au cours du temps. Ceci suggère que la méthode en ligne proposée ici pourrait être améliorée en considérant une configuration d'apprentissage du DNN où les filtres linéaires varient au cours du temps.

6.2. Perspectives

Évaluation dans des conditions plus générales Une perspective à court terme serait d'évaluer la méthode actuelle dans des conditions plus générales. Premièrement, nous pourrions examiner les performances dans le cas monocanal, puis observer l'évolution des performances en considérant plus de microphones. En dérévérbération linéaire, le nombre de microphones a un impact important sur les performances. Dans notre cas, ceci permettrait d'étudier l'influence du nombre de microphones sur les filtres linéaires et le post-filtre.

Deuxièmement, nous pourrions considérer d'autres types de signaux d'écho acoustique, correspondant à d'autres cas d'usage. En particulier, lorsqu'un utilisateur interagit avec l'assistant personnel intégré dans l'enceinte intelligente, il est courant que l'enceinte soit en train de jouer de la musique. Il convient de noter que ce cas d'usage est plus facile dans le scénario considéré dans cette thèse. En effet, les caractéristiques spectrales de la musique et de la parole locale sont différentes, tandis que dans notre cas, les caractéristiques spectrales de la parole locale et de l'écho acoustique sont similaires.

Troisièmement, notre méthode pourrait s'appliquer dans des scénarios avec plusieurs locuteurs locaux, comme avec les systèmes de téléconférence. Ainsi, nous pourrions envisager une extension de notre approche avec plusieurs locuteurs locaux, où l'on ajouterait une variable gaussienne pour chaque locuteur local [Nugraha et al., 2016a]. Bien qu'ici le but final ne soit pas nécessairement de séparer la parole de chaque locuteur local pour le correspondant distant, il peut être nécessaire de modéliser chaque locuteur. En effet, les locuteurs ont des positions différentes dans la salle et peuvent potentiellement se déplacer, ce qui diminuerait les performances de rehaussement.

Enfin, nous pourrions étendre notre méthode à des systèmes avec plusieurs enceintes intelligentes ou des systèmes de visioconférence. Dans ce type de systèmes, la parole distante peut provenir de plusieurs canaux, comme dans un contexte stéréophonique (voir la figure 2.3a). Dans ce cas, la parole distante est un signal multicanal, et chaque haut-parleur joue un canal différent de la parole distante. L'écho $\mathbf{y}(t)$ n'est plus représenté par la convolution de la parole distante $x(t)$ par une RIR comme dans (2.6), mais par la somme de convolutions du signal $\mathbf{x}(t)$ par plusieurs RIRs. En réduction d'écho, cela modifie l'estimation du filtre l'annulation d'écho [Sondhi et al., 1995], mais aussi du post-filtre de suppression d'écho résiduel [Lee et al., 2014]. Dans notre cas, cela modifierait

l'estimation des trois filtres.

Application à la reconnaissance automatique de la parole Une évaluation non considérée dans cette thèse est celle de la reconnaissance automatique de la parole, ainsi que la reconnaissance du locuteur [Sarkar et al., 2016], pour les assistants personnels qui sont intégrés dans les enceintes intelligentes. En effet, le scénario considéré dans cette thèse avec le bruit, l'écho et la réverbération correspond à un cas d'utilisation usuel d'une enceinte intelligente, où l'utilisateur interagit avec celle-ci dans un environnement bruyant et réverbérant, et qu'elle joue de la musique ou bien diffuse de la radio.

Nakatani et Kinoshita [2019] ont récemment développé une approche de réduction conjointe de bruit et de réverbération, utilisant un filtre de déréverbération $\mathcal{G}(f)$ et un post-filtre $\mathbf{W}_{s_\mathrm{e}}(n, f)$. L'amélioration en performance de reconnaissance de la parole est significative par rapport à une version en cascade de cette approche. Cela laisse donc supposer que notre méthode permettrait une amélioration de la reconnaissance de la parole dans notre scénario d'étude. Pour la reconnaissance du locuteur en présence de bruit et de réverbération, Zhao et al. [2014] ont proposé un prétraitement de réduction de bruit à l'aide d'un filtre estimé par un DNN. Ce prétraitement est intégré à un système de reconnaissance du locuteur. Il serait donc possible de réaliser une intégration similaire de notre méthode.

Approche bout-en-bout Une des limitations de notre approche de réduction de bruit, d'écho et de réverbération est que la procédure d'apprentissage des DNNs est séparée de l'estimation des filtres $\mathcal{H}(n, f)$, $\mathcal{G}(n, f)$ et $\mathbf{W}_{s_\mathrm{e}}(n, f)$. En effet, nous avons défini une procédure dans la partie 4.2.4.1 pour déterminer la vérité terrain des DSPs $v_c(n, f)$, qui servent ensuite de cibles pour l'apprentissage des DNNs. Puisque nous avons exprimé les équations de mise à jour des paramètres $\Theta_H(n)$, $\Theta_G(n)$ et $\Theta_c(n)$ de manière en ligne dans le chapitre 5, il serait possible de réaliser l'apprentissage des DNNs de manière bout-en-bout en appliquant ces équations aux signaux $\mathbf{s}_\mathrm{e}(n, f)$, $\mathbf{s}_\mathrm{l}(n, f)$, $\mathbf{y}(n, f)$ et $\mathbf{b}(n, f)$, qui sont connus durant l'apprentissage.

Réduction de la complexité Plutôt que d'estimer les filtres $\mathcal{H}(f)$, $\mathcal{G}(f)$ et $\mathbf{W}_{s_\mathrm{e}}(n, f)$ à l'aide de l'algorithme DNN-BCA qui requiert plusieurs étapes à chaque itération, une perspective serait de regrouper ces trois filtres en un seul filtre, et de l'estimer au moyen d'un DNN. Dans le cas monocanal, cette méthode pourrait s'inspirer de la méthode monocanale de Luo et Mesgarani [2019], qui estime directement à l'aide d'un DNN la parole locale $\mathbf{s}(t)$ à partir du mélange $\mathbf{d}^{\mathrm{bruit}}(t)$ dans le domaine temporel. Pour appliquer la méthode de Luo et Mesgarani [2019] à notre problème, il faudrait la modifier pour qu'elle puisse donner la même représentation temps-fréquence du mélange $\mathbf{d}(t)$ et de la parole distante $x(t)$. Pour cela, les poids d'entrée du DNN pourraient être appliqués séparément sur chacun des deux signaux $\mathbf{d}(t)$ et $x(t)$, en contraignant les poids d'entrée du DNN à être égaux. Enfin, pour pouvoir traiter le cas multicanal, une architecture de type CNN pourrait traiter les différents canaux comme les canaux rouge-vert-bleu en traitement d'image.

Optimisation d'un critère perceptuel Comme nous l'avons présenté au chapitre 4, le critère d'optimisation considéré n'est pas nécessairement corrélé à la qualité perceptuelle. Avec un apprentissage bout-en-bout depuis le domaine temporel, il serait possible d'entraîner un DNN pour optimiser un critère perceptuel. En particulier, Kim et al. [2019] a récemment proposé un apprentissage bout-en-bout optimisant la métrique PESQ, qui donne un score perceptuel [Int. Telecomm. Union (ITU-T) Rec., 2001] (voir la partie 2.4.2).

Microphones distribués Enfin, une autre perspective serait de ne plus considérer une antenne de microphones compacte, mais plutôt des microphones distribués dans la salle, situés de plusieurs systèmes différents (enceinte intelligente, ordinateur, téléphone portable, etc.). Ce type de microphones est désigné par le terme d'antenne de microphones *ad hoc* [Bertrand et al., 2015]. Dans ce type de problèmes, le décalage temporel entre chaque microphone devient important par rapport à une antenne compacte, ce qui nécessite de modéliser différemment les signaux sur chaque dispositif.

A. Réduction conjointe de bruit, d'écho et de réverbération

Ce chapitre est l'annexe du chapitre 4 sur la réduction conjointe de bruit, d'écho et de réverbération. Dans les parties A.1 et A.2, nous expliquons en détail l'obtention des règles de mise à jour de l'algorithme itératif de montée par blocs de coordonnées qui optimise conjointement tous les filtres de notre approche. Nous fournissons le pseudo-code détaillé de cet algorithme. Dans cette annexe, nous présentons le pseudo-code détaillé de l'algorithme itératif qui détermine les cibles pour le réseau de neurones utilisé dans l'approche. Enfin, nous donnons des précisions sur les paramètres d'enregistrement et de simulation des données.

A.1. Calcul vectorisé de l'annulation d'écho et de la déréverbération linéaire

Puisque le contexte de filtrage est multicanal et multitrame, le problème d'optimisation de la vraisemblance défini dans (4.23) n'est pas séparable entre les canaux et les trames. Pour le résoudre, nous reformulons le terme $\widehat{\mathbf{y}}(n,f) = \sum_{k=0}^{K-1} \mathbf{h}(k,f)x(n-k,f)$ dans (4.24) de la manière suivante :

$$\widehat{\mathbf{y}}(n,f) = \underline{\mathbf{X}}(n,f)\underline{\mathbf{h}}(f), \tag{A.1}$$

où $\underline{\mathbf{h}}(f)$ est la version vectorisée du filtre $\mathcal{H}(f)$

$$\underline{\mathbf{h}}(f) = \begin{bmatrix} \mathbf{h}(0,f) \\ \vdots \\ \mathbf{h}(K-1,f) \end{bmatrix}, \tag{A.2}$$

$\underline{\mathbf{X}}(n,f) \in \mathbb{C}^{M\times MK}$ est la concaténation des K trames $\mathbf{X}(n-k,f) \in \mathbb{C}^{M\times M}$

$$\underline{\mathbf{X}}(n,f) = \Big[\mathbf{X}(n,f)\dots\mathbf{X}(n-K+1,f)\Big], \tag{A.3}$$

et $\mathbf{X}(n-k,f)$ est la version multicanale de $x(n-k,f)$ obtenue comme suit :

$$\mathbf{X}(n-k,f) = x(n-k,f)\mathbf{I}_M. \tag{A.4}$$

De la même manière, nous reformulons le terme $\widehat{\mathbf{e}}_{\mathrm{l}}(n,f) = \sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\mathbf{e}(n-l,f)$ dans (4.24) de la manière suivante :

$$\widehat{\mathbf{e}}_{\mathrm{l}}(n,f) = \underline{\mathbf{E}}(n,f)\underline{\mathbf{g}}(f), \tag{A.5}$$

où $\underline{\mathbf{g}}(f)$ est la version vectorisée du filtre $\mathcal{G}(f)$ [Yoshioka et al., 2011]

$$\underline{\mathbf{g}}(f) = \begin{bmatrix} \mathbf{g}_1(\Delta, f) \\ \vdots \\ \mathbf{g}_M(\Delta, f) \\ \vdots \\ \vdots \\ \mathbf{g}_1(\Delta + L - 1, f) \\ \vdots \\ \mathbf{g}_M(\Delta + L - 1, f) \end{bmatrix}, \tag{A.6}$$

$\underline{\mathbf{E}}(n, f) \in \mathbb{C}^{M \times M^2 L}$ est la concaténation des L trames $\mathbf{E}(n - l, f) \in \mathbb{C}^{M \times M^2}$ [Yoshioka et al., 2011]

$$\underline{\mathbf{E}}(n, f) = [\mathbf{E}(n - \Delta, f) \dots \mathbf{E}(n - \Delta - L + 1, f)], \tag{A.7}$$

et $\mathbf{E}(n - l, f)$ est la version multicanale de $\mathbf{e}(n - l, f)$ obtenue comme suit :

$$\mathbf{E}(n - l, f) = \mathbf{I}_M \otimes \mathbf{e}(n - l, f)^T. \tag{A.8}$$

Comme le problème d'optimisation défini dans (4.23) n'a pas de solution analytique, nous devons estimer les paramètres à l'aide d'une procédure itérative.

A.2. Algorithme itératif d'optimisation

Dans cette partie, nous détaillons l'obtention des règles de mise à jour des filtres linéaires $\mathcal{H}(f)$ et $\mathcal{G}(f)$, et nous décrivons le pseudo-code détaillé de l'algorithme DNN-ACB.

A.2.1. Paramètres du filtre d'annulation d'écho Θ_H

La dérivée partielle de la log-vraisemblance $\mathcal{L}\left(\mathcal{O};\Theta_H,\Theta_G,\Theta_c\right)$ définie dans (4.23) en fonction de $\underline{\mathbf{h}}(f)$ peut être calculée de la manière suivante :

$$\begin{aligned}
&\frac{\partial\mathcal{L}(\mathcal{O};\Theta_H,\Theta_G,\Theta_c)}{\partial\underline{\mathbf{h}}(f)} \\
&\quad= -\frac{\partial}{\partial\underline{\mathbf{h}}(f)}\sum_{n=0}^{N-1}\Big(\mathbf{d}(n,f)-\widehat{\mathbf{y}}(n,f)-\widehat{\mathbf{e}}_{\mathrm{l}}(n,f)\Big)^{H} \qquad \text{(A.9)}\\
&\qquad\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\Big(\mathbf{d}(n,f)-\widehat{\mathbf{y}}(n,f)-\widehat{\mathbf{e}}_{\mathrm{l}}(n,f)\Big)\\
&\quad= -\frac{\partial}{\partial\underline{\mathbf{h}}(f)}\sum_{n=0}^{N-1}\Big(\mathbf{d}(n,f)-\widehat{\mathbf{y}}(n,f)-\Big(\widehat{\mathbf{e}}_{\mathrm{l},s}(n,f)+\widehat{\mathbf{e}}_{\mathrm{l},z}(n,f)+\widehat{\mathbf{e}}_{\mathrm{l},b}(n,f)\Big)\Big)^{H}\\
&\qquad\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\Big(\mathbf{d}(n,f)-\widehat{\mathbf{y}}(n,f)-\Big(\widehat{\mathbf{e}}_{\mathrm{l},s}(n,f)+\widehat{\mathbf{e}}_{\mathrm{l},z}(n,f)+\widehat{\mathbf{e}}_{\mathrm{l},b}(n,f)\Big)\Big). \qquad \text{(A.10)}
\end{aligned}$$

De la même manière que (4.17)–(4.18), nous pouvons séparer le terme $\widehat{\mathbf{e}}_{\mathrm{l},z}(n,f)$ comme suit

$$\begin{aligned}
\widehat{\mathbf{e}}_{\mathrm{l},z}(n,f) &= \sum_{l=\Delta}^{\Delta+L-1}\mathbf{G}(l,f)\mathbf{z}(n-l,f) && \text{(A.11)}\\
&= \sum_{l=\Delta}^{\Delta+L-1}\mathbf{G}(l,f)\Big(\mathbf{y}(n-l,f)-\widehat{\mathbf{y}}(n-l,f)\Big) && \text{(A.12)}\\
&= \widehat{\mathbf{e}}_{\mathrm{l},y}(n,f)-\widehat{\mathbf{e}}_{\mathrm{l},\widehat{y}}(n,f), && \text{(A.13)}
\end{aligned}$$

et nous le remplaçons dans (A.10) :

$$\begin{aligned}
&\frac{\partial\mathcal{L}\left(\mathcal{O};\Theta_H,\Theta_G,\Theta_c\right)}{\partial\underline{\mathbf{h}}(f)}\\
&= -\frac{\partial}{\partial\underline{\mathbf{h}}(f)}\sum_{n=0}^{N-1}\Big(\mathbf{d}(n,f)-\widehat{\mathbf{y}}(n,f)-\Big(\widehat{\mathbf{e}}_{\mathrm{l},s}(n,f)+\widehat{\mathbf{e}}_{\mathrm{l},y}(n,f)-\widehat{\mathbf{e}}_{\mathrm{l},\widehat{y}}(n,f)+\widehat{\mathbf{e}}_{\mathrm{l},b}(n,f)\Big)\Big)^{H}\\
&\quad\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\Big(\mathbf{d}(n,f)-\widehat{\mathbf{y}}(n,f)-\Big(\widehat{\mathbf{e}}_{\mathrm{l},s}(n,f)+\widehat{\mathbf{e}}_{\mathrm{l},y}(n,f)-\widehat{\mathbf{e}}_{\mathrm{l},\widehat{y}}(n,f)+\widehat{\mathbf{e}}_{\mathrm{l},b}(n,f)\Big)\Big). \qquad \text{(A.14)}
\end{aligned}$$

Ainsi, dans (A.14), nous pouvons regrouper les termes liés aux signaux $\mathbf{s}(n,f)$, $\mathbf{y}(n,f)$ et $\mathbf{b}(n,f)$, puisqu'ils ne dépendent pas de $\underline{\mathbf{h}}(f)$:

$$\begin{aligned}\mathbf{d}(n,f) - \widehat{\mathbf{e}}_{\mathrm{l},s}(n,f) - \widehat{\mathbf{e}}_{\mathrm{l},b}(n,f) - \widehat{\mathbf{e}}_{\mathrm{l},y}(n,f)& \\ = \mathbf{d}(n,f) - \sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\Big(\mathbf{s}(n-l,f) + \mathbf{b}(n-l,f) + \mathbf{y}(n-l,f)\Big)& \qquad \text{(A.15)}\\ = \mathbf{d}(n,f) - \sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\mathbf{d}(n-l,f)& \qquad \text{(A.16)}\\ = \mathbf{r}_d(n,f),& \qquad \text{(A.17)}\end{aligned}$$

où $\mathbf{r}_d(n,f)$ la composante latente *déréverbérée* du signal $\mathbf{r}(n,f)$ obtenue après application du filtre de déréverbération $\mathcal{G}(f)$ sur le mélange $\mathbf{d}(n,f)$ sans annulation d'écho en amont (A.14) devient alors

$$\begin{aligned}\frac{\partial \mathcal{L}\left(\mathcal{O};\Theta_H,\Theta_G,\Theta_c\right)}{\partial \underline{\mathbf{h}}(f)} &= -\frac{\partial}{\partial \underline{\mathbf{h}}(f)} \sum_{n=0}^{N-1} \left(\mathbf{r}_d(n,f) - \left(\widehat{\mathbf{y}}(n,f) - \widehat{\mathbf{e}}_{\mathrm{l},\widehat{y}}(n,f)\right)\right)^H \\ &\quad \boldsymbol{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\left(\mathbf{r}_d(n,f) - \left(\widehat{\mathbf{y}}(n,f) - \widehat{\mathbf{e}}_{\mathrm{l},\widehat{y}}(n,f)\right)\right) \qquad \text{(A.18)}\\ &= -\frac{\partial}{\partial \underline{\mathbf{h}}(f)} \sum_{n=0}^{N-1} \left(\mathbf{r}_d(n,f) - \left(\widehat{\mathbf{y}}(n,f) - \sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\widehat{\mathbf{y}}(n-l,f)\right)\right)^H \\ &\quad \boldsymbol{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\left(\mathbf{r}_d(n,f) - \left(\widehat{\mathbf{y}}(n,f) - \sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\widehat{\mathbf{y}}(n-l,f)\right)\right). \qquad \text{(A.19)}\end{aligned}$$

En remplaçant (A.1) dans (A.19) :

$$\begin{aligned}&\frac{\partial \mathcal{L}\left(\mathcal{O};\Theta_H,\Theta_G,\Theta_c\right)}{\partial \underline{\mathbf{h}}(f)} \\ &\quad = -\frac{\partial}{\partial \underline{\mathbf{h}}(f)} \sum_{n=0}^{N-1} \left(\mathbf{r}_d(n,f) - \left(\underline{\mathbf{X}}(n,f)\underline{\mathbf{h}}(f) - \sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\underline{\mathbf{X}}(n-l,f)\underline{\mathbf{h}}(f)\right)\right)^H \\ &\qquad \boldsymbol{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\left(\mathbf{r}_d(n,f) - \left(\underline{\mathbf{X}}(n,f)\underline{\mathbf{h}}(f) - \sum_{l=\Delta}^{\Delta+L-1} \mathbf{G}(l,f)\underline{\mathbf{X}}(n-l,f)\underline{\mathbf{h}}(f)\right)\right). \qquad \text{(A.20)}\\ &\quad = -\frac{\partial}{\partial \underline{\mathbf{h}}(f)} \sum_{n=0}^{N-1} \left(\mathbf{r}_d(n,f) - \underline{\mathbf{X}}_{\mathrm{r}}(n,f)\underline{\mathbf{h}}(f)\right)^H \\ &\qquad \boldsymbol{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\left(\mathbf{r}_d(n,f) - \underline{\mathbf{X}}_{\mathrm{r}}(n,f)\underline{\mathbf{h}}(f)\right), \qquad \text{(A.21)}\end{aligned}$$

où $\underline{\mathbf{X}}_{\mathrm{r}}(n,f) = [\mathbf{X}_{\mathrm{r}}(n,f)\dots\mathbf{X}_{\mathrm{r}}(n-K+1,f)] \in \mathbb{C}^{M\times MK}$ est la concaténation des K trames $\mathbf{X}_{\mathrm{r}}(n,f) \in \mathbb{C}^{M\times M}$, qui sont des versions de $\mathbf{X}(n,f)$ obtenues par application du filtre de déréverbération $\mathcal{G}(f)$ sur les L trames précédentes du signal de parole distante $x(n-k-l,f)$, et par soustraction du signal résultant au signal $\mathbf{X}(n,f)$ de la manière suivante :

$$\mathbf{X}_{\mathrm{r}}(n,f) = \mathbf{X}(n,f) - \sum_{l=\Delta}^{\Delta+L-1} x(n-k-l,f)\mathbf{G}(l,f). \tag{A.22}$$

Ainsi, (A.21) devient :

$$\begin{aligned}\frac{\partial\mathcal{L}\left(\mathcal{O};\Theta_H,\Theta_G,\Theta_c\right)}{\partial\underline{\mathbf{h}}(f)} = &\sum_{n=0}^{N-1} 2\underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}(n,f)\underline{\mathbf{h}}(f)\\ &- 2\underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\mathbf{r}_d(n,f)\end{aligned} \tag{A.23}$$

La log-vraisemblance est optimale par rapport à $\underline{\mathbf{h}}(f)$ pour $\frac{\partial\mathcal{L}(\mathcal{O};\Theta_H,\Theta_G,\Theta_c)}{\partial\underline{\mathbf{h}}(f)} = 0$. Le filtre d'annulation d'écho $\mathcal{H}(f)$ est alors mis à jour de la manière suivante :

$$\underline{\mathbf{h}}(f) = \mathbf{P}(f)^{-1}\mathbf{p}(f), \tag{A.24}$$

où

$$\mathbf{P}(f) = \sum_{n=0}^{N-1} \underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}(n,f) \tag{A.25}$$

$$\mathbf{p}(f) = \sum_{n=0}^{N-1} \underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\mathbf{r}_d(n,f). \tag{A.26}$$

Il convient de remarquer que la matrice $\mathbf{P}(f)$ est une somme de termes de rang M, et nécessite donc au moins K termes pour être inversible.

A.2.2. Paramètres du filtre de déréverbération Θ_G

La dérivée partielle de la log-vraisemblance $\mathcal{L}\left(\mathcal{O};\Theta_H,\Theta_G,\Theta_c\right)$ définie dans (4.23) en fonction de $\underline{\mathbf{g}}(f)$ peut être calculée de la manière suivante :

$$\begin{aligned}\frac{\partial\mathcal{L}\left(\mathcal{O};\Theta_H,\Theta_G,\Theta_c\right)}{\partial\underline{\mathbf{g}}(f)} = -\frac{\partial}{\partial\underline{\mathbf{g}}(f)}&\sum_{n=0}^{N-1}\Big(\mathbf{d}(n,f) - \widehat{\mathbf{y}}(n,f) - \widehat{\mathbf{e}}_{\mathrm{l}}(n,f)\Big)^H\\ &\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\Big(\mathbf{d}(n,f) - \widehat{\mathbf{y}}(n,f) - \widehat{\mathbf{e}}_{\mathrm{l}}(n,f)\Big).\end{aligned} \tag{A.27}$$

Le terme $\mathbf{e}(n,f) = \mathbf{d}(n,f) - \widehat{\mathbf{y}}(n,f)$ ne dépend pas du filtre de déréverbération $\underline{\mathbf{g}}(f)$. Nous obtenons alors

$$\begin{aligned}\frac{\partial \mathcal{L}\left(\mathcal{O};\Theta_H,\Theta_G,\Theta_c\right)}{\partial \underline{\mathbf{g}}(f)} = -\frac{\partial}{\partial \underline{\mathbf{g}}(f)} \sum_{n=0}^{N-1} \Bigg(\mathbf{e}(n,f) - \widehat{\mathbf{e}}_{\mathrm{l}}(n,f)\Bigg)^H \\ \mathbf{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\Bigg(\mathbf{e}(n,f) - \widehat{\mathbf{e}}_{\mathrm{l}}(n,f)\Bigg).\end{aligned} \tag{A.28}$$

En remplaçant (A.5) dans (A.28) :

$$\begin{aligned}\frac{\partial \mathcal{L}\left(\mathcal{O};\Theta_H,\Theta_G,\Theta_c\right)}{\partial \underline{\mathbf{g}}(f)} = -\frac{\partial}{\partial \underline{\mathbf{g}}(f)} \sum_{n=0}^{N-1} \Bigg(\mathbf{e}(n,f) - \underline{\mathbf{E}}(n,f)\underline{\mathbf{g}}(f)\Bigg)^H \\ \mathbf{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\Bigg(\mathbf{e}(n,f) - \underline{\mathbf{E}}(n,f)\underline{\mathbf{g}}(f)\Bigg)\end{aligned} \tag{A.29}$$

$$\begin{aligned}&= \sum_{n=0}^{N-1} 2\underline{\mathbf{E}}(n,f)^H \mathbf{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\underline{\mathbf{E}}(n,f)\underline{\mathbf{g}}(f) \\ &- 2\underline{\mathbf{E}}(nf)^H \mathbf{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\mathbf{e}(n,f).\end{aligned} \tag{A.30}$$

La log-vraisemblance est optimale par rapport à $\underline{\mathbf{g}}(f)$ pour $\frac{\partial \mathcal{L}(\mathcal{O};\Theta_H,\Theta_G,\Theta_c)}{\partial \underline{\mathbf{g}}(f)} = 0$. Ainsi, le filtre de déréverbération $\mathcal{G}(f)$ est mis à jour de la manière suivante :

$$\underline{\mathbf{g}}(f) = \mathbf{Q}(f)^{-1}\mathbf{q}(f), \tag{A.31}$$

où

$$\mathbf{Q}(f) = \sum_{n=0}^{N-1} \underline{\mathbf{E}}(n,f)^H \mathbf{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\underline{\mathbf{E}}(n,f) \tag{A.32}$$

$$\mathbf{q}(f) = \sum_{n=0}^{N-1} \underline{\mathbf{E}}(n,f)^H \mathbf{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\mathbf{e}(n,f). \tag{A.33}$$

Il convient de remarquer que la matrice $\mathbf{Q}(f)$ est une somme de termes de rang M, qui nécessite donc au moins ML termes pour être inversible.

A.2.3. Estimation de la composante précoce finale $\mathbf{s_e}(n,f)$

Le pseudo-code de l'algorithme DNN-BCA est décrit dans l'Algorithme 1.

A.3. Détermination des cibles pour le réseau de neurones

Le pseudo-code de la procédure de détermination de la vérité terrain des DSPs $v_c(n,f)$ est décrit dans l'Algorithme 2.

Algorithme 1 : Algorithme DNN-BCA proposé pour la réduction conjointe de bruit, d'écho et de réverbération.

Entrées :
- $\mathbf{d}(n,f), x(n,f)$
- $\mathrm{DNN}_0, \mathrm{DNN}_1, \ldots, \mathrm{DNN}_I$ pré-entraînés

Initialisation :
- Initialisation des filtres linéaires
  - $\underline{\mathbf{h}}(f) \leftarrow \underline{\mathbf{h}}_0(f)$ (choisi par l'utilisateur, par exemple la méthode de Valin [2007])
  - $\underline{\mathbf{g}}(f) \leftarrow \underline{\mathbf{g}}_0(f)$ (choisi par l'utilisateur, par exemple la méthode WPE [Yoshioka et Nakatani, 2012])
- Initialisation des MCSs
  - $[\mathbf{R}_{s_\mathrm{e}}(f)\ \mathbf{R}_{s_\mathrm{r}}(f)\ \mathbf{R}_{z_\mathrm{r}}(f)\ \mathbf{R}_{b_\mathrm{r}}(f)] \leftarrow [\mathbf{I}_M\ \mathbf{I}_M\ \mathbf{I}_M\ \mathbf{I}_M]$
- Initialisation des entrées du DNN
  - entrées $\leftarrow$ (4.52)
- Initialisation des DSPs
  - $[v_{s_\mathrm{e}}(n,f)\ v_{s_\mathrm{r}}(n,f)\ v_{z_\mathrm{r}}(n,f)\ v_{b_\mathrm{r}}(n,f)] \leftarrow [\mathrm{DNN}_0(\text{entrées})]^2$

for *itération i de I* **do**
- Mise à jour du filtre d'annulation d'écho
  - $\underline{\mathbf{h}}(f) \leftarrow$ (4.29)
- Mise à jour du signal $\mathbf{e}(n,f)$
  - $\mathbf{e}(n,f) \leftarrow$ (4.11)
- Mise à jour du filtre de déréverbération
  - $\underline{\mathbf{g}}(f) \leftarrow$ (4.34)
- Mise à jour du signal $\mathbf{r}(n,f)$
  - $\mathbf{r}(n,f) \leftarrow$ (4.12)
- Mise à jour des MCSs
- **for** *mise à jour spatiale j of J* **do**
  - **for** *source* $\mathbf{c}$ *de* $[\mathbf{s}_\mathrm{e}, \mathbf{s}_\mathrm{r}, \mathbf{z}_\mathrm{r}, \mathbf{b}_\mathrm{r}]$ **do**
    - Mise à jour du post-filtre court
      - $\mathbf{W}_c(n,f) \leftarrow$ (4.21)
    - Mise à jour de l'estimation du signal
      - $\widehat{\mathbf{c}}(n,f) \leftarrow$ (4.39)
    - Mise à jour des statistiques *a posteriori*
      - $\widehat{\mathbf{\Sigma}}_c(n,f) \leftarrow$ (4.40)
    - Mise à jour de la MCS
      - $\mathbf{R}_c(f) \leftarrow$ (4.41)–(4.43)
  - **end**
- **end**
- Mise à jour des entrées du DNN
  - entrées $\leftarrow$ (4.54)
- Mise à jour des DSPs
  - $[v_{s_\mathrm{e}}(n,f)\ v_{s_\mathrm{r}}(n,f)\ v_{z_\mathrm{r}}(n,f)\ v_{b_\mathrm{r}}(n,f)] \leftarrow [\mathrm{DNN}_i(\text{entrées})]^2$

end

Calcul de la composante précoce finale
- $\widehat{\mathbf{s}}_\mathrm{e}(n,f) \leftarrow$ (4.11)+(4.12)+(4.19)

Sortie :
- $\widehat{\mathbf{s}}_\mathrm{e}(n,f)$

Algorithme 2 : Procédure itérative de détermination de la vérité terrain des DSPs.

Entrées :
$\mathbf{s}(n,f)$, $\mathbf{s}_{\mathrm{e}}(n,f)$, $\mathbf{s}_{\mathrm{l}}(n,f)$, $\mathbf{y}(n,f)$, $\mathbf{b}(n,f)$
Initialisation :
Initialisation des variables latentes
$\mathbf{s}_r(n,f) \leftarrow \mathbf{s}_{\mathrm{l}}(n,f)$
$\mathbf{z}_r(f) \leftarrow \mathbf{y}(n,f)$
$\mathbf{b}_r(f) \leftarrow \mathbf{b}(n,f)$
for *source* $\mathbf{c}$ *de* $[\mathbf{s}_{\mathrm{e}}, \mathbf{s}_{\mathrm{r}}, \mathbf{z}_{\mathrm{r}}, \mathbf{b}_{\mathrm{r}}]$ **do**
Initialisation de la DSP
$v_c(n,f) \leftarrow (4.44)$
Initialisation de la MCS
$\mathbf{R}_c(f) \leftarrow \mathbf{I}_M$
end
for *itération* i *de* I **do**
Mise à jour du filtre d'annulation d'écho
$\underline{\mathbf{h}}(f) \leftarrow (4.29)$
Mise à jour du filtre de déréverbération
$\underline{\mathbf{g}}(f) \leftarrow (4.34)$
Mise à jour des variables latentes
$\mathbf{s}_{\mathrm{r}}(n,f) \leftarrow (4.14)$
$\mathbf{z}_{\mathrm{r}}(n,f) \leftarrow (4.15)$
$\mathbf{b}_{\mathrm{r}}(n,f) \leftarrow (4.16)$
for *source* $\mathbf{c}$ *de* $[\mathbf{s}_{\mathrm{e}}, \mathbf{s}_{\mathrm{r}}, \mathbf{z}_{\mathrm{r}}, \mathbf{b}_{\mathrm{r}}]$ **do**
Mise à jour de la DSP
$v_c(n,f) \leftarrow (4.48)$
Mise à jour de la MCS
$\mathbf{R}_c(f) \leftarrow (4.49)+(4.43)$
end
end
Sortie :
$[v_{s_{\mathrm{e}}}(n,f)\ v_{s_{\mathrm{r}}}(n,f)\ v_{z_{\mathrm{r}}}(n,f)\ v_{b_{\mathrm{r}}}(n,f)]$

A.4. Paramètres d'enregistrement et de simulation

Dans cette partie, nous donnons des précisions sur les paramètres d'enregistrement et de simuation pour la création des données.

A.4.1. Enregistrements réels d'écho

La distance entre le haut-parleur et l'antenne de microphones du Triby est de 11 cm, et la distance entre les microphones est de 3 cm. L'enceinte intelligente Triby a été placée au centre de chaque salle (voir la figure 4.8) dans 2 positions différentes pour augmenter la diversité des chemins d'écho $\mathbf{a}_y(\tau)$. La parole distante $x(t)$ a été jouée à 3 réglages de volume différents du Triby, pour augmenter la diversité des non-linéarités : plus la parole distante $x(t)$ est forte, plus il y a de non-linéarités dans l'écho $y(t)$.

A.4.2. Simulations de RIR de la parole locale $\mathbf{a}_s(\tau)$

Nous avons réalisé les simulations de RIR de la parole locale $\mathbf{a}_s(\tau)$ à l'aide du logiciel Roomsimove [Vincent et Campbell, 2008]. Les caractéristiques des salles simulées correspondent aux mêmes caractéristiques de réverbération (T_{60} par bande d'octave) que les salles réelles où les enregistrements d'écho ont été réalisés (voir la figure A.1). Toutefois, les dimensions des salles simulées ont été choisies aléatoirement sur une plage de $\pm 20\%$ des dimensions des salles réelles. Les dimensions de l'antenne de microphones simulée sont identiques à celle du vrai Triby, et ses position et orientation sont les mêmes. La position du locuteur local a été choisi aléatoirement sur un demi-cercle de rayon $1,5$ m centré sur le milieu de l'antenne de microphones, et à une distante d'au moins 10 cm des murs (voir la figure A.2). Pour chaque RIR, une nouvelle salle et une position aléatoire ont été générées pour augmenter la diversité des RIRs.

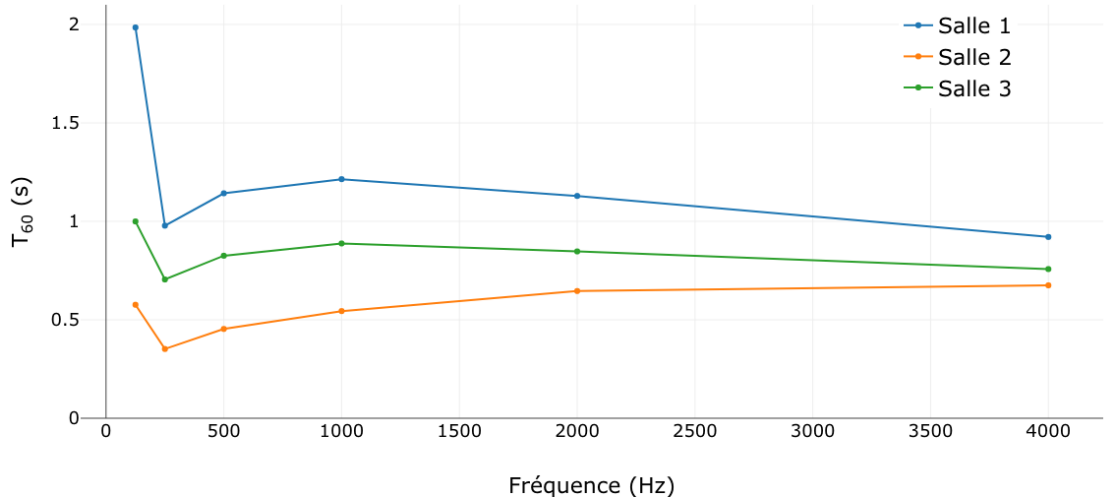


FIGURE A.1. – Temps de réverbération T_{60} par bande d'octave pour les salles 1, 2 et 3.

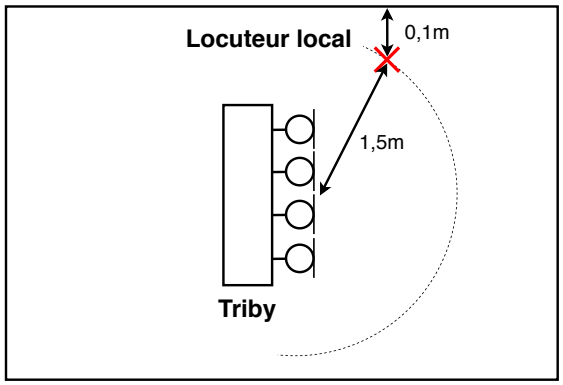
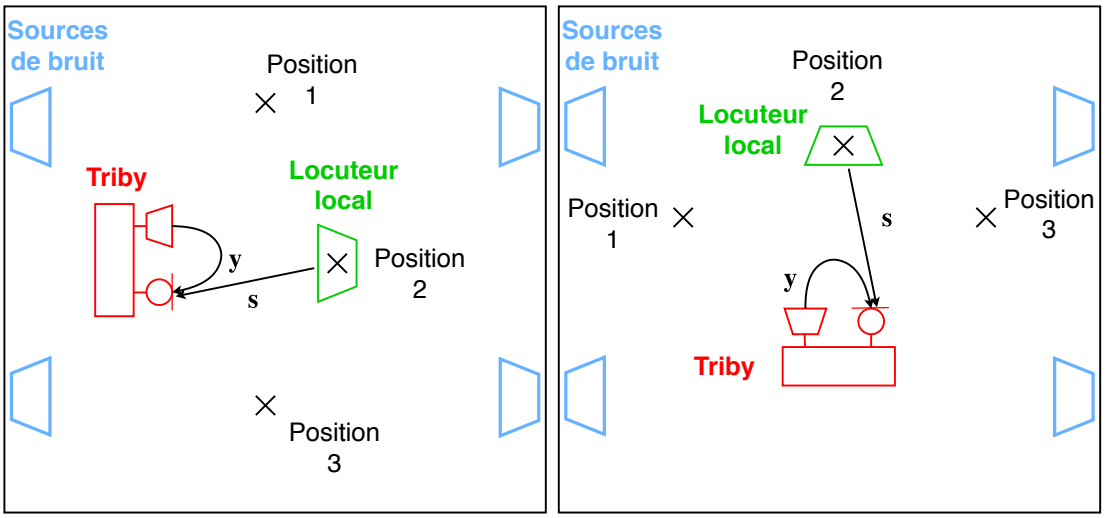


FIGURE A.2. – Paramètres de position du locuteur local dans les simulations.

A.4.3. Enregistrements réels de l'ensemble de test invariant dans le temps

Pour l'ensemble de test où les conditions acoustiques sont invariantes au cours du temps, le haut-parleur jouant la parole locale $\mathbf{s}(t)$ a été placé sur un demi-cercle de $1,5$ m de rayon, centré sur l'enceinte intelligente Triby, dans 3 positions : $0°$, $60°$, $-60°$(voir la figure A.3). Les quatre Tribys jouant le signal de bruit $\mathbf{b}(t)$ ont été placés à $2,5$ m de l'antenne de microphones, et leur position est restée la même sur tous les enregistrements (voir la figure 4.9). De la même manière que pour l'écho, les signaux de bruit ont été joués à 3 niveaux de volume différents. Nous avons considéré un scénario réaliste où l'utilisateur augmente le niveau de volume du Triby servant à la télécommunication lorsque le bruit devient plus fort. Ainsi, plus les signaux de bruit sont forts, plus l'écho est fort.



(a) Position 1 du Triby.

(b) Position 2 du Triby.

FIGURE A.3. – Schéma des configurations expérimentales d'enregistrement des signaux dans l'ensemble de test.

B. Variante en ligne de la réduction conjointe de bruit, d'écho et de réverbération basée sur l'apprentissage profond

Ce chapitre est l'annexe du chapitre 5 sur la variante en ligne de la réduction conjointe de bruit, d'écho et de réverbération. Dans les parties B.1 et B.2, nous expliquons en détail l'obtention des règles de mise à jour de la version en ligne de l'algorithme itératif de montée par blocs de coordonnées qui optimise conjointement et de manière causale tous les filtres de notre approche. Nous omettons l'exposant $(\cdot)^{(i)}$ lié à l'itération i par souci de clarté. Enfin, dans la partie B.3, nous fournissons le pseudo-code détaillé de l'algorithme itératif de montée par blocs de coordonnées.

B.1. Paramètres du filtre d'annulation d'écho Θ_H

Dans le cas en ligne, le filtre d'annulation d'écho $\mathcal{H}(n,f)$ est mis à jour comme dans (5.19) :

$$\underline{\mathbf{h}}(n,f) = \mathbf{P}(n,f)^{-1}\mathbf{p}(n,f), \tag{B.1}$$

où les termes $\mathbf{P}(n,f)$ et $\mathbf{p}(n,f)$ sont estimés de manière récursive. Pour cela, nous utilisons une moyenne glissante :

$$\mathbf{P}(n,f) = \alpha_h \mathbf{P}(n-1,f) + (1-\alpha_h)\underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H \mathbf{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}(n,f), \tag{B.2}$$

$$\mathbf{p}(n,f) = \alpha_h \mathbf{p}(n-1,f) + (1-\alpha_h)\underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H \mathbf{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\mathbf{r}_d(n,f). \tag{B.3}$$

où α_h est un coefficient d'oubli tel que $0 < \alpha_h < 1$. Nous cherchons à simplifier (B.1) pour éviter d'inverser le terme $\mathbf{P}(n,f) \in \mathbb{C}^{MK\times MK}$ à chaque trame n. En utilisant l'identité de Woodbury avec (B.2), le terme $\mathbf{P}(n,f)^{-1}$ devient alors

$$\begin{aligned}\mathbf{P}(n,f)^{-1} = \frac{1}{\alpha_h}\mathbf{P}(n-1,f)^{-1} - \frac{1-\alpha_h}{\alpha_h}\mathbf{P}(n-1,f)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H \\ \Big(\alpha_h \mathbf{\Sigma}_{\mathbf{dd}}(n,f) + (1-\alpha_h)\underline{\mathbf{X}}_{\mathrm{r}}(n,f)\mathbf{P}(n-1,f)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H\Big)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}(n,f)\mathbf{P}(n-1,f)^{-1}\end{aligned} \tag{B.4}$$

$$= \frac{1}{\alpha_h}\mathbf{P}(n-1,f)^{-1} - \frac{1}{\alpha_h}\mathbf{K}_{\mathrm{X}}(n,f)\underline{\mathbf{X}}_{\mathrm{r}}(n,f)\mathbf{P}(n-1,f)^{-1}, \tag{B.5}$$

où

$$\begin{aligned}\mathbf{K}_{\mathrm{X}}(n,f) = (1-\alpha_h)\mathbf{P}(n-1,f)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H \\ \Big(\alpha_h\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f) + (1-\alpha_h)\underline{\mathbf{X}}_{\mathrm{r}}(n,f)\mathbf{P}(n-1,f)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H\Big)^{-1}.\end{aligned} \tag{B.6}$$

Pour poursuivre la simplification de (B.1), nous devons formuler le terme $\mathbf{K}_{\mathrm{X}}(n,f) \in \mathbb{C}^{MK\times M}$ sous une autre forme. Ainsi, en déplaçant le membre inversé de droite dans (B.6) au membre de gauche, (B.6) devient

$$\begin{aligned}\mathbf{K}_{\mathrm{X}}(n,f)\Big(\alpha_h\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f) + (1-\alpha_h)\underline{\mathbf{X}}_{\mathrm{r}}(n,f)\mathbf{P}(n-1,f)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H\Big) \\ = (1-\alpha_h)\mathbf{P}(n-1,f)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H,\end{aligned} \tag{B.7}$$

$$\begin{aligned}\alpha_h\mathbf{K}_{\mathrm{X}}(n,f)\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f) + (1-\alpha_h)\mathbf{K}_{\mathrm{X}}(n,f)\underline{\mathbf{X}}_{\mathrm{r}}(n,f)\mathbf{P}(n-1,f)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H \\ = (1-\alpha_h)\mathbf{P}(n-1,f)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H.\end{aligned} \tag{B.8}$$

En déplaçant le second membre de gauche au membre de droite, (B.8) devient

$$\begin{aligned}\alpha_h\mathbf{K}_{\mathrm{X}}(n,f)\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f) \\ = (1-\alpha_h)\mathbf{P}(n-1,f)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H - (1-\alpha_h)\mathbf{K}_{\mathrm{X}}(n,f)\underline{\mathbf{X}}_{\mathrm{r}}(n,f)\mathbf{P}(n-1,f)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H,\end{aligned} \tag{B.9}$$

$$\begin{aligned}\mathbf{K}_{\mathrm{X}}(n,f)\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f) \\ = (1-\alpha_h)\frac{1}{\alpha_h}\Big(\mathbf{P}(n-1,f)^{-1} - \mathbf{K}_{\mathrm{X}}(n,f)\underline{\mathbf{X}}_{\mathrm{r}}(n,f)\mathbf{P}(n-1,f)^{-1}\Big)\underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H.\end{aligned} \tag{B.10}$$

D'après (B.5), (B.10) s'exprime alors comme

$$\begin{aligned}\mathbf{K}_{\mathrm{X}}(n,f)\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f) \\ = (1-\alpha_h)\underbrace{\frac{1}{\alpha_h}\Big(\mathbf{P}(n-1,f)^{-1} - \mathbf{K}_{\mathrm{X}}(n,f)\underline{\mathbf{X}}_{\mathrm{r}}(n,f)\mathbf{P}(n-1,f)^{-1}\Big)}_{=\mathbf{P}(n,f)^{-1}}\underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H.\end{aligned} \tag{B.11}$$

En déplaçant le terme $\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f)$ au membre de droite, (B.11) devient alors

$$\mathbf{K}_{\mathrm{X}}(n,f) = (1-\alpha_h)\mathbf{P}(n,f)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f)^{-1}. \tag{B.12}$$

En remplaçant (B.3) dans (B.1), on obtient

$$\underline{\mathbf{h}}(n,f) = \mathbf{P}(n,f)^{-1}\mathbf{p}(n,f) \tag{B.13}$$

$$= \mathbf{P}(n,f)^{-1}\Big(\alpha_h\mathbf{p}(n-1,f) + (1-\alpha_h)\underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\mathbf{r}_d(n,f)\Big) \tag{B.14}$$

$$= \alpha_h\mathbf{P}(n,f)^{-1}\mathbf{p}(n-1,f) + \underbrace{(1-\alpha_h)\mathbf{P}(n,f)^{-1}\underline{\mathbf{X}}_{\mathrm{r}}(n,f)^H\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f)^{-1}}_{=\mathbf{K}_{\mathrm{X}}(n,f)}\mathbf{r}_d(n,f). \tag{B.15}$$

En remplaçant le terme $\mathbf{P}(n,f)^{-1}$ par (B.5), (B.15) devient

$$\begin{aligned}\underline{\mathbf{h}}(n,f) &= \alpha_h\left(\frac{1}{\alpha_h}\mathbf{P}(n-1,f)^{-1} - \frac{1}{\alpha_h}\mathbf{K}_{\mathrm{X}}(n,f)\underline{\mathbf{X}}_{\mathrm{r}}(n,f)\mathbf{P}(n-1,f)^{-1}\right)\mathbf{p}(n-1,f) \\ &\quad + \mathbf{K}_{\mathrm{X}}(n,f)\mathbf{r}_d(n,f) \qquad \text{(B.16)} \\ &= \underbrace{\mathbf{P}(n-1,f)^{-1}\mathbf{p}(n-1,f)}_{=\underline{\mathbf{h}}(n-1,f)} - \mathbf{K}_{\mathrm{X}}(n,f)\underline{\mathbf{X}}_{\mathrm{r}}(n,f)\underbrace{\mathbf{P}(n-1,f)^{-1}\mathbf{p}(n-1,f)}_{=\underline{\mathbf{h}}(n-1,f)} \\ &\quad + \mathbf{K}_{\mathrm{X}}(n,f)\mathbf{r}_d(n,f) \qquad \text{(B.17)} \\ &= \underline{\mathbf{h}}(n-1,f) + \mathbf{K}_{\mathrm{X}}(n,f)\Big(-\underline{\mathbf{X}}_{\mathrm{r}}(n,f)\underline{\mathbf{h}}(n-1,f) + \mathbf{r}_d(n,f)\Big). \qquad \text{(B.18)}\end{aligned}$$

Ainsi, le filtre d'annulation d'écho $\mathcal{H}(n,f)$ est mis à jour de la manière suivante :

$$\underline{\mathbf{h}}(n,f) = \underline{\mathbf{h}}(n-1,f) + \mathbf{K}_{\mathrm{X}}(n,f)\Big(\mathbf{r}_d(n,f) - \underline{\mathbf{X}}_{\mathrm{r}}(n,f)\underline{\mathbf{h}}(n-1,f)\Big). \qquad \text{(B.19)}$$

Il convient de noter qu'en retirant le terme de pondération $(1-\alpha_h)$ de la mise à jour du filtre d'annulation d'écho $\mathcal{H}(n,f)$, cela reviendrait à minimiser un critère des moindres carrés récursifs [Haykin, 2002, Chapitre 13]. D'après (A.21), celui s'exprimerait de la manière suivante :

$$\begin{aligned}\min_{\underline{\mathbf{h}}(n,f)} \sum_{n'=0}^{n} \alpha_h^{n-n'} &\left(\mathbf{r}_d(n',f) - \underline{\mathbf{X}}_{\mathrm{r}}(n',f)\underline{\mathbf{h}}(n',f)\right)^H \\ &\mathbf{\Sigma}_{\mathbf{dd}}(n',f)^{-1}\left(\mathbf{r}_d(n',f) - \underline{\mathbf{X}}_{\mathrm{r}}(n',f)\underline{\mathbf{h}}(n',f)\right).\end{aligned} \qquad \text{(B.20)}$$

B.2. Paramètres du filtre de déréverbération Θ_G

Dans le cas en ligne, le filtre de déréverbération $\mathcal{G}(n,f)$ est mis à jour comme dans (5.29) :

$$\underline{\mathbf{g}}(n,f) = \mathbf{Q}(n,f)^{-1}\mathbf{q}(n,f), \qquad \text{(B.21)}$$

où les termes $\mathbf{Q}(n,f)$ et $\mathbf{q}(n,f)$ sont estimés de manière récursive. De même que pour l'annulation d'écho dans la sous-partie précédente, nous utilisons une moyenne glissante :

$$\mathbf{Q}(n,f) = \alpha_g\mathbf{Q}(n-1,f) + (1-\alpha_g)\underline{\mathbf{E}}(n,f)^H\mathbf{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\underline{\mathbf{E}}(n,f), \qquad \text{(B.22)}$$

$$\mathbf{q}(n,f) = \alpha_g\mathbf{q}(n-1,f) + (1-\alpha_g)\underline{\mathbf{E}}(n,f)^H\mathbf{\Sigma}_{\mathbf{dd}}(n,f)^{-1}\mathbf{e}(n,f). \qquad \text{(B.23)}$$

Nous cherchons à simplifier (B.21) pour éviter d'inverser le terme $\mathbf{Q}(n,f) \in \mathbb{C}^{M^2L\times M^2L}$ à chaque trame n. En reprenant la même logique qu'à la sous-partie précédente, où l'on remplace $\underline{\mathbf{X}}_{\mathrm{r}}(n,f)$ par $\underline{\mathbf{E}}(n,f)$, $\mathbf{r}_d(n,f)$ par $\mathbf{e}(n,f)$, $\mathbf{P}(n,f)^{-1}$ par

$$\mathbf{Q}(n,f)^{-1} = \frac{1}{\alpha_g}\mathbf{Q}(n-1,f)^{-1} - \frac{1}{\alpha_g}\mathbf{K}_{\mathrm{E}}(n,f)\underline{\mathbf{E}}(n,f)\mathbf{Q}(n-1,f)^{-1}, \qquad \text{(B.24)}$$

et $\mathbf{K}_{\mathrm{X}}(n,f)$ par

$$\begin{aligned}\mathbf{K}_{\mathrm{E}}(n,f) = (1-\alpha_g)\mathbf{Q}(n-1,f)^{-1}\underline{\mathbf{E}}(n,f)^H \\ \Big(\alpha_g\boldsymbol{\Sigma}_{\mathbf{dd}}(n,f) + (1-\alpha_g)\underline{\mathbf{E}}(n,f)\mathbf{Q}(n-1,f)^{-1}\underline{\mathbf{E}}(n,f)^H\Big)^{-1},\end{aligned} \tag{B.25}$$

on obtient la mise à jour suivante pour le filtre de déréverbération $\mathcal{G}(n,f)$:

$$\underline{\mathbf{g}}(n,f) = \underline{\mathbf{g}}(n-1,f) + \mathbf{K}_{\mathrm{E}}(n,f)\Big(\mathbf{e}(n,f) - \underline{\mathbf{E}}(n,f)\underline{\mathbf{g}}(n-1,f)\Big). \tag{B.26}$$

Il convient de noter qu'en retirant le terme de pondération $(1-\alpha_g)$ de la mise à jour du filtre de déréverbération $\mathcal{G}(n,f)$, cela reviendrait à minimiser un critère des moindres carrés récursifs [Haykin, 2002, Chapitre 13]. D'après (A.29), celui s'exprimerait de la manière suivante :

$$\begin{aligned}\min_{\underline{\mathbf{g}}(n,f)} \sum_{n'=0}^{n} \alpha_g^{n-n'} \Bigg(\mathbf{e}(n',f) - \underline{\mathbf{E}}(n',f)\underline{\mathbf{g}}(n',f)\Bigg)^H \\ \boldsymbol{\Sigma}_{\mathbf{dd}}(n',f)^{-1}\Bigg(\mathbf{e}(n',f) - \underline{\mathbf{E}}(n',f)\underline{\mathbf{g}}(n',f)\Bigg).\end{aligned} \tag{B.27}$$

B.3. Estimation de la composante précoce finale $\mathbf{s_e}(n,f)$

Le pseudo-code de la version causale de l'algorithme DNN-BCA est décrit dans l'Algorithme 3.

Algorithme 3 : Version causale de l'algorithme DNN-BCA proposé pour la réduction conjointe de bruit, d'écho et de réverbération.

Entrées :
$\mathbf{d}(n,f), x(n,f)$
$\mathrm{DNN}_0, \mathrm{DNN}_1, \ldots, \mathrm{DNN}_I$ pré-entraînés
$\alpha_h, \alpha_g, \alpha_c$
Initialisation : Initialisation des matrices inverses $\mathbf{P}^{(I)}(n,f)^{-1}$ et $\mathbf{Q}^{(I)}(n,f)^{-1}$
$\mathbf{P}^{(I)}(n,f)^{-1} \leftarrow \mathbf{I}_{MK}$
$\mathbf{Q}^{(I)}(n,f)^{-1} \leftarrow \mathbf{I}_{M^2L}$
Initialisation des MCSs
$\left[\mathbf{R}_c^{(I)}(n,f)\right]_c \leftarrow \left[\mathbf{I}_M\right]_c$
for *trame* n *de* N **do**
  Initialisation des filtres linéaires
    $\underline{\mathbf{h}}^{(0)}(n,f) \leftarrow$ méthode choisie par l'utilisateur (ex : SpeexDSP ou $\underline{\mathbf{h}}^{(I)}(n-1,f)$)
    $\underline{\mathbf{g}}^{(0)}(n,f) \leftarrow$ méthode choisie par l'utilisateur (ex : DNN-WPE causal ou $\underline{\mathbf{g}}^{(I)}(n-1,f)$)
  Initialisation des MCSs
    $\left[\mathbf{R}_c^{(0)}(n,f)\right]_c \leftarrow \left[\mathbf{R}_c^{(I)}(n-1,f)\right]_c$
  Initialisation des DSPs
    $\left[v_c^{(0)}(n,f)\right]_c \leftarrow \left[\mathrm{DNN}_0\right]^2$
  for *itération* i *de* I **do**
    Mise à jour du filtre d'annulation d'écho
      $\underline{\mathbf{h}}^{(i)}(f) \leftarrow$ (5.19)
    Mise à jour du signal $\mathbf{e}^{(i)}(n,f)$
      $\mathbf{e}^{(i)}(n,f) \leftarrow$ (5.6)
    Mise à jour du filtre de déréverbération
      $\underline{\mathbf{g}}^{(i)}(f) \leftarrow$ (5.29)
    Mise à jour du signal $\mathbf{r}^{(i)}(n,f)$
      $\mathbf{r}^{(i)}(n,f) \leftarrow$ (5.7)
    Mise à jour des MCSs
    for *source* $\mathbf{c}$ *de* $[\mathbf{s}_\mathrm{e}, \mathbf{s}_\mathrm{r}, \mathbf{z}_\mathrm{r}, \mathbf{b}_\mathrm{r}]$ **do**
      Mise à jour du post-filtre court
        $\mathbf{W}_c^{(i-1)}(n,f) \leftarrow$ (5.11)
      Mise à jour de l'estimation du signal
        $\widehat{\mathbf{c}}^{(i-1)}(n,f) \leftarrow$ (5.38)
      Mise à jour des statistiques *a posteriori*
        $\widehat{\mathbf{\Sigma}}_c^{(i-1)}(n,f) \leftarrow$ (5.39)
      Mise à jour de la MCS
        $\mathbf{R}_c^{(i)}(f) \leftarrow$ (5.40)
    end
    Mise à jour des DSPs
      $\left[v_c^{(i)}(n,f)\right] \leftarrow \left[\mathrm{DNN}_i\right]^2$
  end
end
Sortie :
$\left[\widehat{\mathbf{s}}_\mathrm{e}^{(I-1)}(n,f)\right]_n$

Bibliographie

Affes, S. et 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.

Allen, J. B. (1977). Short-term spectral analysis, and modification by discrete Fourier transform. *IEEE Transactions on Acoustics Speech and Signal Processing*, 25(3) :235–238.

Araki, S. et Nakatani, T. (2011). Hybrid approach for multichannel source separation combining time-frequency mask with multi-channel Wiener filter. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 225–228.

Attias, H. (2003). New EM algorithms for source separation and deconvolution with a microphone array. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 297–300.

Audio Software Engineering and Siri Speech Team (2018). Optimizing Siri on HomePod in far-field settings. *Apple Machine Learning Journal*, 1(12).

Avendano, C. (2001). Acoustic echo suppression in the STFT domain. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 175–178.

Ben Kheder, W., Matrouf, D., Ajili, M., et Bonastre, J.-F. (2018). A unified joint model to deal with nuisance variabilities in the i-vector space. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(3) :633–645.

Ben Kheder, W., Matrouf, D., Bonastre, J.-F., Ajili, M., et Bousquet, P.-M. (2015). Additive noise compensation in the i-vector space for speaker recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4190–4194.

Bendersky, D. A., Stokes, J. W., et Malvar, H. S. (2008). Nonlinear residual acoustic echo suppression for high levels of harmonic distortion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 261–264.

Benesty, J., Makino, S., et Chen, J. (2005). Springer Berlin Heidelberg.

Bertrand, A., Doclo, S., Gannot, S., Ono, N., et Waterschoot, T. v. (2015). Special issue on wireless acoustic sensor networks and ad hoc microphone arrays. *Signal Processing*, 107 :1 – 3.

Bertsekas, D. P. (1999). *Nonlinear Programming, 2nd edition.* Athena Scientific.

Birkett, A. N. et Goubran, R. A. (1995). Acoustic echo cancellation using NLMS-neural network structures. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages 3035–3038.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2) :113–120.

Brandstein, M. et Ward, D. (2001). *Microphone Arrays : Signal Processing, Techniques and Applications.* Springer.

Braun, S. et Habets, E. A. P. (2013). Dereverberation in noisy environments using reference signals and a maximum likelihood estimator. In *European Signal Processing Conference (EUSIPCO)*, pages 1–5.

Braun, S. et Habets, E. A. P. (2015). A multichannel diffuse power estimator for dereverberation in the presence of multiple sources. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1).

Braun, S. et Habets, E. A. P. (2016). Online dereverberation for dynamic scenarios using a Kalman filter with an autoregressive model. *IEEE Signal Processing Letters*, 23(12) :1741–1745.

Braun, S., Kuklasinski, A., Schwartz, O., Thiergart, O., Habets, E. A. P., Gannot, S., Doclo, S., et Jensen, J. (2018). Evaluation and comparison of late reverberation power spectral density estimators. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(6) :1056–1071.

Braun, S., Schwartz, B., Gannot, S., et Habets, E. A. P. (2016). Late reverberation PSD estimation for single-channel dereverberation using relative convolutive transfer functions. In *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–5.

Buchner, H., Aichner, R., et Kellermann, W. (2005). A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. *IEEE Transactions on Speech and Audio Processing*, 13(1) :120–134.

Chen, Z., Luo, Y., et Mesgarani, N. (2017). Deep attractor network for single-microphone speaker separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 246–250.

Chhetri, A. S., Surendran, A. C., Stokes, J. W., et Platt, J. C. (2005). Regression-based residual acoustic echo suppression. In *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 201–204.

Cohen, A., Stemmer, G., Ingalsuo, S., et Markovich-Golan, S. (2017). Combined weighted prediction error and minimum variance distortionless response for dereverberation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 446–450.

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.

Costa, J. P., Lagrange, A., et Arliaud, A. (2003). Acoustic echo cancellation using nonlinear cascade filters. In *IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP)*, volume 5, pages 389–392.

Crochiere, R. (1980). A weighted overlap-add method of short-time Fourier analysis/synthesis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(1) :99–102.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., et Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4) :788–798.

Dietzen, T., Doclo, S., Moonen, M., et Waterschoot, T. V. (2018). Joint multi-microphone speech dereverberation and noise reduction using integrated sidelobe cancellation and linear prediction. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 221–225.

Doclo, S., Moonen, M., et de Clippel, E. (2000). Combined acoustic echo and noise reduction using GSVD-based optimal filtering. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1061–1064.

Drude, L., Higuchi, T., Kinoshita, K., Nakatani, T., et Haeb-Umbach, R. (2018). Dual frequency- and block-permutation alignment for deep learning based block-online blind source separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 691–695.

Duong, N. Q. K., Vincent, E., et Gribonval, R. (2010). Under-determined reverberant audio source separation using a full-Rank spatial covariance model. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7) :1830–1840.

Duong, N. Q. K., Vincent, E., et Gribonval, R. (2011). An acoustically-motivated spatial prior for under-determined reverberant source separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9–12.

Emiya, V., Vincent, E., Harlander, N., et Hohmann, V. (2011). Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7) :2046–2057.

Enzner, G. et Vary, P. (2006). Frequency-domain adaptive Kalman filter for acoustic echo control in hands-free telephones. *Signal Processing*, 86(6) :1140–1156.

Ephraim, Y. et 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.

Erdogan, H., Hershey, J. R., Watanabe, S., et Le Roux, J. (2015). Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 708–712.

Erdogan, H., Hershey, J. R., Watanabe, S., Mandel, M. I., et Le Roux, J. (2016). Improved MVDR beamforming using single-channel mask prediction networks. In *Interspeech*, pages 1981–1985.

Erkelens, J. S. et Heusdens, R. (2010). Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7) :1746–1765.

Ernst, O., Chazan, S. E., Gannot, S., et Goldberger, J. (2018). Speech dereverberation using fully convolutional networks. In *European Signal Processing Conference (EUSIPCO)*, pages 390–394.

Eskénazi, M., Levow, G.-A., Meng, H., Parent, G., et Suendermann, D. (2013). *Crowdsourcing for Speech Processing : Applications to Data Collection, Transcription and Assessment*. Wiley.

Faller, C. et Chen, J. (2005). Suppressing acoustic echo in a spectral envelope space. *IEEE Transactions on Speech and Audio Processing*, 13(5) :1048–1062.

Favrot, A., Faller, C., et Kuech, F. (2012). Modeling late reverberation in acoustic echo suppression. In *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–4.

Frost, O. L. (1972). An algorithm for linearly constrained adaptive array processing. *Proceedings of the IEEE*, 60(8) :926–935.

Furuya, K. et Kataoka, A. (2007). Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction. *IEEE Transactions on Audio, Speech and Language Processing*, 15(5) :1579–1591.

Févotte, C., Bertin, N., et Durrieu, J.-L. (2009). Nonnegative matrix factorization with the Itakura-Saito divergence : with application to music analysis. *Neural Computation*, 21(3) :793–830.

Gannot, S., Burshtein, D., et Weinstein, E. (2001). Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing*, 49(8) :1614–1626.

Grais, E. M., Ward, D., et Plumbley, M. D. (2018). Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders. In *European Signal Processing Conference (EUSIPCO)*, pages 1577–1581.

Griffiths, L. et Jim, C. (1982). An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on Antennas and Propagation*, 30(1) :27–34.

Guérin, A., Faucon, G., et Le Bouquin-Jeannès, R. (2003). Nonlinear acoustic echo cancellation based on Volterra filters. *IEEE Transactions on Speech and Audio Processing*, 11(6) :672–683.

Gustafsson, S., Martin, R., Jax, P., et Vary, P. (2002). A psychoacoustic approach to combined acoustic echo cancellation and noise reduction. *IEEE Transactions on Speech and Audio Processing*, 10(5) :245–256.

Habets, E. A. P. (2005). Multi-channel speech dereverberation based on a statistical model of late reverberation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 173–176.

Habets, E. A. P. et Gannot, S. (2007). Dual-microphone speech dereverberation using a reference signal. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 901–904.

Habets, E. A. P., Gannot, S., et Cohen, I. (2008a). Speech dereverberation using backward estimation of the late reverberant spectral variance. In *IEEE Convention of Electrical and Electronics Engineers in Israel (IEEEI)*, pages 384–388.

Habets, E. A. P., Gannot, S., et Cohen, I. (2009). Late reverberant spectral variance estimation based on a statistical model. *IEEE Signal Processing Letters*, 16(9) :770–773.

Habets, E. A. P., Gannot, S., Cohen, I., et Sommen, P. C. (2008b). Joint dereverberation and residual echo suppression of speech signals in noisy environments. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8) :1433–1451.

Haykin, S. (2002). *Adaptive filter theory, 4th edition*. Prentice-Hall.

Herbordtt, W., Nakamura, S., et Kellermann, W. (2005). Joint optimization of LCMV beamforming and acoustic echo cancellation for automatic speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 77–80.

Hershey, J. R., Chen, Z., Le Roux, J., et Watanabe, S. (2016). Deep clustering : Discriminative embeddings for segmentation and separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35.

Heymann, J., Drude, L., et Haeb-Umbach, R. (2016). Neural network based spectral mask estimation for acoustic beamforming. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 196–200.

Heymann, J., Drude, L., Haeb-Umbach, R., Kinoshita, K., et Nakatani, T. (2018). Frame-online DNN-WPE dereverberation. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 466–470.

Higuchi, T. et Kameoka, H. (2015). Unified approach for audio source separation with multichannel factorial HMM and DOA mixture model. In *European Signal Processing Conference (EUSIPCO)*, pages 2043–2047.

Hoshuyama, O. et Sugiyama, A. (2006). An acoustic echo suppressor based on a frequency-domain model of highly nonlinear residual echo. In *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings (ICASSP)*, pages 269–272.

Hoshuyama, O., Sugiyama, A., et Hirano, A. (1999). A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters. *IEEE Transactions on Signal Processing*, 47(10) :2677–2684.

Huang, H., Hofmann, C., Kellermann, W., Chen, J., et Benesty, J. (2016). A multi-frame parametric Wiener filter for acoustic echo suppression. In *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–5.

Huang, P., Kim, M., Hasegawa-Johnson, M., et Smaragdis, P. (2015). Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12) :2136–2147.

Huemmer, C., Hofmann, C., Maas, R., et Kellermann, W. (2014a). The significance-aware EPFES to estimate a memoryless preprocessor for nonlinear acoustic echo cancellation. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 557–561.

Huemmer, C., Hofmann, C., Maas, R., Schwarz, A., et Kellermann, W. (2014b). The elitist particle filter based on evolutionary strategies as novel approach for nonlinear acoustic echo cancellation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1315–1319.

Hänsler, E. et Schmidt, G. (2004). *Acoustic Echo and Noise Control : a Practical Approach*. Wiley-Interscience.

Int. Telecomm. Union (ITU-T) Rec. (2001). Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs P. 862.

Int. Telecomm. Union (ITU-T) Rec. (2015). Recommendation itu-r bs.1534-3 : method for the subjective assessment of intermediate quality level of audio systems.

Ito, N., Araki, S., Yoshioka, T., et Nakatani, T. (2014). Relaxed disjointness based clustering for joint blind source separation and dereverberation. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 268–272.

Jukić, A., Waterschoot, T. v., Gerkmann, T., et Doclo, S. (2014). Speech dereverberation with convolutive transfer function approximation using map and variational deconvolution approaches. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 50–54.

Jukić, A., Waterschoot, T. v., Gerkmann, T., et Doclo, S. (2015). Multi-channel linear prediction-based speech dereverberation with sparse priors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9) :1509–1520.

Kagami, H., Kameoka, H., et Yukawa, M. (2018). Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35.

Kallinger, M. et Mertins, A. (2006). Multi-channel room impulse response shaping - a study. In *IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages 101–104.

Kang, T. G., Kwon, K., Shin, J. W., et Kim, N. S. (2015). NMF-based target source separation using deep neural network. *IEEE Signal Processing Letters*, 22(2) :229–233.

Kellermann, W. L. (2001). Acoustic echo cancellation for beamforming microphone arrays. In *Microphone Arrays : Signal Processing Techniques and Applications*, pages 281–306. Springer Berlin Heidelberg.

Kenny, P., Boulianne, G., Ouellet, P., et Dumouchel, P. (2007). Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 15(4) :1435–1447.

Kim, J., El-Kharmy, M., et Lee, J. (2019). End-to-End Multi-Task Denoising for joint SDR and PESQ Optimization. *http://arxiv.org/abs/1901.09146*, pages 1–10.

Kingma, D. P. et Ba, J. (2015). Adam : a method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Kinoshita, K., Delcroix, M., Kwon, H., Mori, T., et Nakatani, T. (2017). Neural network-based spectrum estimation for online WPE dereverberation. In *Interspeech*, pages 384–388.

Kinoshita, K., Delcroix, M., Nakatani, T., et 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.

Kodrasi, I. et Bourlard, H. (2018). Single-channel late reverberation power spectral density estimation using denoising autoencoders. In *Interspeech*, pages 1319–1323.

Kodrasi, I. et Doclo, S. (2017). Late reverberant power spectral density estimation based on an eigenvalue decomposition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 611–615.

Kodrasi, I. et Doclo, S. (2018a). Analysis of eigenvalue decomposition-based late reverberation power spectral density estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(6) :1106–1118.

Kodrasi, I. et Doclo, S. (2018b). Joint late reverberation and noise power spectral density estimation in a spatially homogeneous noise field. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 441–445.

Kodrasi, I., Gerkmann, T., et Doclo, S. (2014). Frequency-domain single-channel inverse filtering for speech dereverberation : theory and practice. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5177–5181.

Kodrasi, I., Goetze, S., et Doclo, S. (2013). Regularization for partial multichannel equalization for speech dereverberation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(9) :1879–1890.

Kuech, F., Mitnacht, A., et Kellermann, W. (2005). Nonlinear acoustic echo cancellation using adaptive orthogonalized power filters. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 105–108.

Kuklasiński, A., Doclo, S., Jensen, S. H., et Jensen, J. (2014). Maximum likelihood based multi-channel isotropic reverberation reduction for hearing aids. In *European Signal Processing Conference (EUSIPCO)*, pages 61–65.

Kuklasiński, A., Doclo, S., Jensen, S. H., et Jensen, J. (2016). Maximum likelihood PSD estimation for speech enhancement in reverberation and noise. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9) :1599–1612.

Le Bouquin-Jeannès, R., Scalart, P., Faucon, G., et Beaugeant, C. (2001). Combined noise and echo reduction in hands-free systems : a survey. *IEEE Transactions on Speech and Audio Processing*, 9(8) :808–820.

Le Roux, J., Hershey, J. R., et Weninger, F. (2015). Deep NMF for speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 66–70.

Le Roux, J., Wisdom, S., Erdogan, H., et Hershey, J. R. (2019). SDR – half-baked or well done ? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 5.

Lebart, K., Boucher, J. M., et Denbigh, P. N. (2001). A new method based on spectral subtraction for speech dereverberation. *Acta Acustica united with Acustica*, 87(3) :359–366.

Lee, C. M., Shin, J. W., et Kim, N. S. (2014). Stereophonic acoustic echo suppression incorporating spectro-temporal correlations. *IEEE Signal Processing Letters*, 21(3) :316–320.

Lee, C. M., Shin, J. W., et Kim, N. S. (2015). DNN-based residual echo suppression. In *IEEE International Conference on Digital Signal Processing (DSP)*, pages 1775–1779.

Lee, D. D. et Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755) :788–791.

Lee, S. Y. et Kim, N. S. (2007). A statistical model-based residual echo suppression. *IEEE Signal Processing Letters*, 14(10) :758–761.

Leglaive, S., Badeau, R., et Richard, G. (2016). Multichannel audio source separation with probabilistic reverberation priors. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 24(12) :2453–2465.

Leglaive, S., Girin, L., et Horaud, R. (2019). Semi-supervised multichannel speech enhancement with variational autoencoders and non-negative matrix factorization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 101–105.

Lim, F., Zhang, W., Habets, E. A. P., et Naylor, P. A. (2014). Robust multichannel dereverberation using relaxed multichannel least squares. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(9) :1379–1390.

Liutkus, A., Fitzgerald, D., et Rafii, Z. (2015). Scalable audio separation with light kernel additive modelling. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 76–80.

Liutkus, A., Fitzgerald, D., Rafii, Z., Pardo, B., et Daudet, L. (2014). Kernel additive models for source separation. *IEEE Transactions on Signal Processing*, 62(16) :4298–4310.

Loizou, P. C. (2007). *Speech Enhancement : Theory and Practice*. CRC Press.

Löllmann, H. W. et Vary, P. (2009). A blind speech enhancement algorithm for the suppression of late reverberation and noise. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3989–3992.

Lu, X., Tsao, Y., Matsuda, S., et Hori, C. (2013). Speech enhancement based on deep denoising autoencoder. In *Interspeech*, pages 436–440.

Luis Valero, M. et Habets, E. A. P. (2017). Multi-microphone acoustic echo cancellation using relative echo transfer functions. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 229–233.

Luis Valero, M. et Habets, E. A. P. (2019). Low-complexity multi-microphone acoustic echo control in the short-time Fourier transform domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(3) :595–609.

Luo, Y., Chen, Z., et Mesgarani, N. (2018). Speaker-independent Speech Separation with Deep Attractor Network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(4) :787–796.

Luo, Y. et 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.

Mack, W., Chakrabarty, S., Stöter, F.-R., Braun, S., Edler, B., et Habets, E. A. P. (2018). Single-channel dereverberation using direct MMSE optimization and bidirectional LSTM networks. In *Interspeech*, pages 1314–1318.

Mader, A., Puder, H., et Schmidt, G. U. (2000). Step-size control for acoustic echo cancellation filters – an overview. *Signal Processing*, 80(9) :1697–1719.

Madrid Portillo, J. (2017). Deep learning applied to acoustic echo cancellation. Thèse de master, Aalborg University.

Malek, J. et Koldovský, Z. (2016). Hammerstein model-based nonlinear echo cancelation using a cascade of neural network and adaptive linear filter. In *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–5.

Malik, S. et Enzner, G. (2012). State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(7) :2065–2079.

Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, 9(5) :504–512.

Martin, R. et Vary, P. (1996). Combined acoustic echo control and noise reduction for hands-free telephony - state of the art and perspectives. In *European Signal Processing Conference (EUSIPCO)*, pages 1–4.

Matsui, T., Kanno, T., et Furui, S. (1996). Speaker recognition using HMM composition in noisy environments. *Computer Speech & Language*, 10(2) :107–116.

McAulay, R. et Malpass, M. (1980). Speech enhancement using a soft-decision noise suppression filter. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(2) :137–145.

McCulloch, W. S. et Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4) :115–133.

Miyoshi, M. et Kaneda, Y. (1988). Inverse filtering of room acoustics. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(2) :145–152.

Mohammadiha, N., Smaragdis, P., et Doclo, S. (2015). Joint acoustic and spectral modeling for speech dereverberation using non-negative representations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4410–4414.

Munson, W. A. et Gardner, M. B. (1950). Standardizing auditory tests. *The Journal of the Acoustical Society of America*, 22(5) :675–675.

Myllylä, V. (2006). Residual echo filter for enhanced acoustic echo control. *Signal Processing*, 86(6) :1193–1205.

Nakatani, T. et Kinoshita, K. (2019). A unified convolutional beamformer for simultaneous denoising and dereverberation. *IEEE Signal Processing Letters*, 26(6) :903–907.

Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., et Juang, B. H. (2008). Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 85–88.

Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., et 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.

Naylor, P. A. et Gaubitch, N. D. (2010). *Speech Dereverberation*. Springer.

Ngia, K. S. H. et Sjobert, J. (1998). Nonlinear acoustic echo cancellation using a Hammerstein model. In *IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP)*, volume 2, pages 1229–1232.

Nugraha, A. A., Liutkus, A., et Vincent, E. (2016a). Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9) :1652–1664.

Nugraha, A. A., Liutkus, A., et Vincent, E. (2016b). Multichannel music separation with deep neural networks. In *European Signal Processing Conference (EUSIPCO)*, pages 1748–1752.

Ochiai, T., Watanabe, S., Hori, T., et Hershey, J. R. (2017). Multichannel end-to-end speech recognition. In *International Conference on Machine Learning (ICML)*, pages 2632–2641.

Oppenheim, A. V. et Lim, J. S. (1981). The importance of phase in signals. *Proceedings of the IEEE*, 69(5) :529–541.

Ozerov, A. et Fevotte, C. (2010). Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3) :550–563.

Ozerov, A., Vincent, E., et Bimbot, F. (2012). A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4) :1118–1133.

Paleologu, C., Benesty, J., et Ciochina, S. (2008). A variable step-size affine projection algorithm designed for acoustic echo cancellation. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8) :1466–1478.

Paleologu, C., Benesty, J., et Ciochină, S. (2013). Study of the general Kalman filter for echo cancellation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(8) :1539–1549.

Paleologu, C., Ciochină, S., Benesty, J., et Grant, S. L. (2015). An overview on optimized NLMS algorithms for acoustic echo cancellation. *EURASIP Journal on Advances in Signal Processing*, 2015(97) :1–19.

Panayotov, V., Chen, G., Povey, D., et Khudanpur, S. (2015). Librispeech : an ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Park, S. J., Cho, C. G., Lee, C., et Youn, D. H. (2002). Integrated echo and noise canceler for hands-free applications. *IEEE Transactions on Circuits and Systems II : Analog and Digital Signal Processing*, 49(3) :188–195.

Park, S. R. et Lee, J. W. (2017). A fully convolutional neural network for speech enhancement. In *Interspeech*, pages 1993–1997.

Park, Y.-S. et Chang, J.-H. (2009). Frequency domain acoustic echo suppression based on soft decision. *IEEE Signal Processing Letters*, 16(1) :53–56.

Park, Y.-S. et Chang, J.-H. (2012). Integrated acoustic echo and background noise suppression technique based on soft decision. *EURASIP Journal on Advances in Signal Processing*, 2012(11) :1–9.

Polack, J.-D. (1988). *La Transmission de l'Énergie Sonore dans les Salles.* Thèse de doctorat, Université du Maine.

Rethage, D., Pons, J., et Serra, X. (2018). A Wavenet for speech denoising. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5069–5073.

Reuven, G., Gannot, S., et Cohen, I. (2007). Joint noise reduction and acoustic echo cancellation using the transfer-function generalized sidelobe canceller. *Speech Communication*, 49(7-8) :623–635.

Rombouts, G. et Moonen, M. (2005). An integrated approach to acoustic noise and echo cancellation. *Signal Processing*, 85(4) :849–871.

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., et 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.

Sarkar, A. K., Bonastre, J. F., et Matrouf, D. (2016). A study on the roles of total variability space and session variability modeling in speaker recognition. *International Journal of Speech Technology*, 19(1) :111–120.

Sawada, H., Kameoka, H., Araki, S., et Ueda, N. (2013). Multichannel extensions of non-negative matrix factorization with complex-valued data. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5) :971–982.

Sawada, H., Mukai, R., Araki, S., et Makino, S. (2004). A robust and precise method for solving the permutation problem of frequency-domain blind source separation. *IEEE Transactions on Speech and Audio Processing*, 12(5) :530–538.

Scarpiniti, M., Comminiello, D., Parisi, R., et Uncini, A. (2011). Comparison of Hammerstein and Wiener systems for nonlinear acoustic echo cancelers in reverberant environments. In *International Conference on Digital Signal Processing (DSP)*, pages 1–6.

Schmid, D., Malik, S., et Enzner, G. (2012). An expectation-maximization algorithm for multichannel adaptive speech dereverberation in the frequency-domain. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 17–20.

Schwartz, B., Gannot, S., et Habets, E. A. P. (2015a). Online speech dereverberation using Kalman filter and EM algorithm. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(2) :394–406.

Schwartz, O., Braun, S., Gannot, S., et Habets, E. A. P. (2015b). Maximum likelihood estimation of the late reverberant power spectral density in noisy environments. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5.

Schwartz, O., Gannot, S., et Habets, E. A. P. (2015c). Multi-Microphone Speech Dereverberation and Noise Reduction Using Relative Early Transfer Functions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(2) :240–251.

Schwartz, O., Gannot, S., et Habets, E. A. P. (2016a). An Expectation-Maximization Algorithm for Multimicrophone Speech Dereverberation and Noise Reduction With Coherence Matrix Estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9) :1495–1510.

Schwartz, O., Gannot, S., et Habets, E. A. P. (2016b). Joint estimation of late reverberant and speech power spectral densities in noisy environments using frobenius norm. In *European Signal Processing Conference (EUSIPCO)*, pages 1123–1127.

Schwartz, O., Gannot, S., et Habets, E. A. P. (2016c). Joint maximum likelihood estimation of late reverberant and speech power spectral density in noisy environments. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155.

Schwarz, A., Hofmann, C., et Kellermann, W. (2013). Spectral feature-based nonlinear residual echo suppression. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–4.

Seki, S., Kameoka, H., Li, L., Toda, T., et Takeda, K. (2018). Generalized multichannel variational autoencoder for underdetermined source separation. In *European Signal Processing Conference (EUSIPCO)*, pages 1–5.

Seo, H., Lee, M., et Chang, J.-H. (2018). Integrated acoustic echo and background noise suppression based on stacked deep neural networks. *Applied Acoustics*, 133 :194–201.

Shi, Z., Lin, H., Liu, L., Liu, R., Hayakawa, S., Harada, S., et Han, J. (2019). End-to-end monaural speech separation with multi-scale dynamic weighted gated dilated convolutional pyramid network. In *Interspeech*, pages 4614–4618.

Simon, L. S. R. et Vincent, E. (2012). A general framework for online audio source separation. In *Latent Variable Analysis and Signal Separation*, pages 397–404.

Smaragdis, P. (2007). Convolutive speech bases and their application to supervised speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1) :1–12.

Sondhi, M., Morgan, D., et Hall, J. (1995). Stereophonic acoustic echo cancellation-an overview of the fundamental problem. *IEEE Signal Processing Letters*, 2(8) :148–151.

Sondhi, M. M. (1967). An adaptive echo canceller. *Bell System Technical Journal*, 46(3) :497–511.

Souden, M., Chen, J., Benesty, J., et Affes, S. (2010). Gaussian model-based multichannel speech presence probability. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5) :1072–1077.

Stenger, A. et Kellermann, W. (2000). Adaptation of a memoryless preprocessor for nonlinear acoustic echo cancelling. *Signal Processing*, 80(9) :1747–1760.

Stevens, S. S., Volkmann, J., et Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of Acoustical Society of America*, 8(3) :185–190.

Taal, C. H., Hendriks, R. C., Heusdens, R., et Jensen, J. (2010). A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4214–4217.

Takeda, R., Nakadai, K., Takahashi, T., Komatani, K., Ogata, T., et Okuno, H. G. (2009). ICA-based efficient blind dereverberation and echo cancellation method for barge-in-able robot audition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3677–3680.

Tan, K. et Wang, D. (2018). A convolutional recurrent neural network for real-time speech enhancement. In *Interspeech*, pages 3229–3233.

Togami, M. (2011). Online speech source separation based on maximum likelihood of local Gaussian modeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 213–216.

Togami, M. (2015). Variational Bayes state space model for acoustic echo reduction and dereverberation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 101–105.

Togami, M. (2019). Simultaneous optimization of forgetting factor and time-frequency mask for block online multi-channel speech enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2702–2706.

Togami, M. et Hori, K. (2011). Multichannel semi-blind source separation via local Gaussian modeling for acoustic echo reduction. In *European Signal Processing Conference (EUSIPCO)*, pages 496–500.

Togami, M. et Kawaguchi, Y. (2012). Speech enhancement combined with dereverberation and acoustic echo reduction for time varying systems. In *IEEE Statistical Signal Processing Workshop (SSP)*, pages 357–360.

Togami, M. et Kawaguchi, Y. (2014). Simultaneous optimization of acoustic echo reduction, speech dereverberation, and noise reduction against mutual interference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(11) :1612–1623.

Togami, M., Kawaguchi, Y., et Takashima, R. (2014). Frequency domain acoustic echo reduction based on Kalman smoother with time-varying noise covariance matrix. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5909–5913.

Togami, M., Kawaguchi, Y., Takeda, R., Obuchi, Y., et Nukaga, N. (2013). Optimized speech dereverberation from probabilistic perspective for time varying acoustic transfer function. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7) :1369–1380.

Tong, Y. et Gu, Y. (2016). Acoustic echo suppression based on speech presence probability. In *IEEE International Conference on Digital Signal Processing (DSP)*, pages 35–38.

Turbin, V., Gilloire, A., et Scalart, P. (1997). Comparison of three post-filtering algorithms for residual acoustic echo reduction. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 307–310.

Valero, M. L., Mabande, E., et Habets, E. A. P. (2014). Signal-based late residual echo spectral variance estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5914–5918.

Valin, J. M. (2007). On adjusting the learning rate in frequency domain echo cancellation with double-talk. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3) :1030–1034.

Vincent, E. (2012). Improved perceptual metrics for the evaluation of audio source separation. In *Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 430–437.

Vincent, E. et Campbell, D. R. (2008). Roomsimove. http://homepages.loria.fr/evincent/software/Roomsimove_1.4.zip.

Vincent, E., Gribonval, R., et 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., et Gannot, S. (2018). *Audio source separation and speech enhancement*. Wiley.

Wada, T. S. et Juang, B. H. (2012). Enhancement of residual echo for robust acoustic echo cancellation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1) :175–189.

Wang, Y., Narayanan, A., et Wang, D. (2014). On training targets for supervised speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12) :1849–1858.

Wang, Z., Vincent, E., Serizel, R., et Yan, Y. (2018a). Rank-1 constrained multichannel Wiener filter for speech recognition in noisy environments. *Computer Speech & Language*, 49 :37–51.

Wang, Z.-Q., Le Roux, J., et Hershey, J. R. (2018b). Multi-channel deep clustering : discriminative spectral and spatial embeddings for speaker-independent speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Wang, Z.-Q., Le Roux, J., Wang, D., et Hershey, J. R. (2018c). End-to-end speech separation with unfolded iterative phase reconstruction. In *Interspeech*, pages 2708–2712.

Weninger, F., Hershey, J. R., Le Roux, J., et Schuller, B. (2014a). Discriminatively trained recurrent neural networks for single-channel speech separation. In *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 577–581.

Weninger, F., Le Roux, J., Hershey, J., et Watanabe, S. (2014b). Discriminative NMF and its application to single-channel source separation. In *Interspeech*, pages 865–869.

Widrow, B., Glover, J. R., McCool, J. M., Kaunitz, J., Williams, C. S., Hearn, R. H., Zeidler, J. R., Dong, J. E., et Goodlin, R. C. (1975). Adaptive noise cancelling : Principles and applications. *Proceedings of the IEEE*, 63(12) :1692–1716.

Williamson, D. S. et Wang, D. (2017). Speech dereverberation and denoising using complex ratio masks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5590–5594.

Williamson, D. S., Wang, Y., et Wang, D. (2016). Complex ratio masking for monaural speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3) :483–492.

Winter, S., Kellermann, W., Sawada, H., et Makino, S. (2007). MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and -norm minimization. *EURASIP Journal on Advances in Signal Processing*, 2007(1) :1–12.

Wong, L. P. et Russell, M. (2001). Text-dependent speaker verification under noisy conditions using parallel model combination. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 457–460.

Wu, B., Yang, M., Li, K., Huang, Z., Siniscalchi, S. M., Wang, T., et Lee, C.-H. (2017). A reverberation-time-aware DNN approach leveraging spatial information for microphone array dereverberation. *EURASIP Journal on Advances in Signal Processing*, 2017(81) :1–13.

Wung, J., Wada, T. S., Juang, B.-H., Lee, B., Kalker, T., et Schafer, R. W. (2011). A system approach to residual echo suppression in robust hands-free teleconferencing. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 445–448.

Xia, B. et Bao, C. (2013). Speech enhancement with weighted denoising auto-encoder. In *Interspeech*, pages 3444–3448.

Xiao, X., Watanabe, S., Erdogan, H., Lu, L., Hershey, J., Seltzer, M. L., Chen, G., Zhang, Y., Mandel, M., et Yu, D. (2016). Deep beamforming networks for multi-channel speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5745–5749.

Yılmaz, O. et Rickard, S. (2004). Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7) :1830–1847.

Yoshioka, T. et Nakatani, T. (2012). Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(10) :2707–2720.

Yoshioka, T. et Nakatani, T. (2013). Dereverberation for reverberation-robust microphone arrays. In *European Signal Processing Conference (EUSIPCO)*, pages 1–5.

Yoshioka, T., Nakatani, T., et Miyoshi, M. (2009a). Integrated speech enhancement method using noise suppression and dereverberation. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(2) :231–246.

Yoshioka, T., Nakatani, T., Miyoshi, M., et Okuno, H. G. (2011). Blind separation and dereverberation of speech mixtures by joint optimization. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1) :69–84.

Yoshioka, T., Tachibana, H., Nakatani, T., et Miyoshi, M. (2009b). Adaptive dereverberation of speech signals with speaker-position change detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3733–3736.

Yu, D., Kolbæk, M., Tan, Z.-H., et Jensen, J. (2017). Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245.

Zhang, H. et Wang, D. (2018). Deep learning for acoustic echo cancellation in noisy and double-talk scenarios. In *Interspeech*, pages 3239–3243.

Zhang, W., Habets, E. A. P., et Naylor, P. A. (2010). On the use of channel shortening in multichannel acoustic system equalization. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*.

Zhao, X., Wang, Y., et Wang, D. (2014). Robust speaker identification in noisy and reverberant conditions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3997–4001.

Zhao, Y., Wang, Z.-Q., et Wang, D. (2017). A two-stage algorithm for noisy and reverberant speech enhancement. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5580–5584.