PARCOURS MASTER 2
ATiAM
Parcours multi-mentions du Master Sciences et Technologies
Université Pierre et Marie Curie - Paris 6
en collaboration avec TELECOM ParisTech et l'Ircam

UPMC
PARIS UNIVERSITAS

Méthodes Parcimonieuses pour la Déreverbération des Signaux Audio

Mémoire de Recherche

Nicolás López

5 août 2011

Stage de recherche du Master 2 ATIAM

Arkamys
Télécom ParisTech, Equipe AAO

Encadrants : Ivan Bourmeyster (Arkamys)
Gaël Richard (Telecom ParisTech)
Yves Grenier (Telecom ParisTech)

Arkamys

TELECOM ParisTech

Résumé

Les signaux audio, que ce soit de la parole ou de la musique, sont soumis à différentes dégradations lors de la prise de son. Dans le contexte particulier de la téléphonie sur des dispositifs mains libres, le signal de parole est soumis à la réverbération de la salle où il est produit. Des études récentes ont montré que la réverbération dégrade l'intélligibilité de la parole. Contrairement aux problèmes de réduction de bruit de fond et d'annulation des échos, le problème de le réverbération n'a été abordé que récemment. Le problème de la déréverbération est particulièrement compliqué parce que le signal parasite à éliminer est en forte corrélation avec le signal que l'on souhaite estimer.

Dans le cadre de ce stage de fin d'études, on se propose d'étudier le phénomène de la réverbération dans le but de développer des traitements performants permettant de la réduire. Comme on souhaite intégrer les algorithmes dans des dispositifs téléphoniques, on privilégie les méthodes pouvant fonctionner en temps réel et en utilisant un ou deux microphones pour capter le signal de parole.

Nous présentons d'abord une étude sur les caractéristiques de la réverbération ainsi qu'un état de l'art complet des techniques de déréverbération . De cette étude, deux approches seront choisies et présentées dans le détail avant d'être évaluées.

On présentera d'abord une technique de réduction de la réverbération basée sur un modèle de prédiction linéaire du signal capté par le microphone, proposée dans [20]. Elle vise à modéliser séparément le signal anéchoïque à estimer et la partie reverbérante. L'effet de ce dernier est réduit dans le domaine fréquentiel par soustraction spectrale. Ensuite, on étudiera une technique faisant intervenir des techniques de déconvolution par factorisation en matrices non-négatives, présentée dans [19]. Il s'agit d'un traitement dans le domaine fréquentiel pour estimer le spectrogramme du signal anéchoïque grâce à un algorithme itératif efficace.

Lors de l'évaluation, les deux traitements s'avèrent adaptés à la réduction de la réverbération. Tout de même, des améliorations doivent être introduites afin de réduire les dégradations du signal de parole introduites par les traitements de déréverbération.

Les résultats de cette étude serviront de référence pour développer des nouvelles techniques qui soient adaptées aux exigences de qualité de la parole et de latence des systèmes de télécommunication.

Remerciements

Je tiens tout d'abord à remercier mes encadrants de stage, Ivan Bourmeyster, Gaël Richard et Yves Grenier, qui m'ont proposé de travailler sur ce sujet de recherche innovant. Leurs compétences scientifiques ont été indispensables pour me permettre de prendre rapidement en main les principales problématiques rencontrées lors du stage. Leur suivi régulier et leur intêret m'ont permis d'avancer dans ma recherche malgré le début tardif de mon stage.

Je remercie aussi Philippe Tour, PDG d'Arkamys, de m'avoir accueilli au sein de son organisation et de me permettre de réaliser ce stage et le doctorat qui s'en suit.

Jean Michel Raczinski et Yves Grenier ont consacré un temps important à la mise en place de ce stage et de mon doctorat. C'est grâce à leurs démarches, que le projet de doctorat, qui s'ébauchait déjà l'an dernier, est devenu une réalité.

Je remercie également Romain Hennequin, Juan Restrepo et Marion Benassaya pour leur relecture et leurs conseils, qui ont été indispensables à la finalisation de ce document.

Enfin, je remercie l'ensemble des enseignants-chercheurs et des doctorants du département TSI pour leur accueil chaleureux et pour m'avoir permis de travailler dans une ambiance conviviale et encourageante pour la recherche.

Table des matières

Avant-propos

Avant de nous concentrer sur le déroulement du stage, il convient de rappeler le contexte particulier de sa réalisation. En effet, ce stage est une partie intégrante du doctorat sous convention CIFRE que j'ai décidé de réaliser avec la société Arkamys et Telecom Paristech. J'ai commencé mon doctorat (et mon stage) le 1er Mai 2011 sous la direction industrielle d'Ivan Bourmeyster et la direction académique de Gaël Richard et Yves Grenier. Ainsi, au moment de la soutenance du stage, uniquement 4 mois de travail se seront déroulés au bout desquels on présentera les premiers résultats de la recherche qui nous permettront de développer par la suite des solutions avancées pour le traitement de la réverbération des signaux audio.

Abréviations

AR	Auto-Régressif
BSD	Bark Spectral Distortion
CMS	Cepstral Mean Subtraction
DRR	Direct to Reverberant Ratio
FFT	Fast Fourier Transform
LP	Linear Prediction
LPC	Linear Prediction Coding
LSD	Log Spectral Distorsion
MOS	Mean Opinion Score
NLMS	Normalized Least Mean Squares
NMF	Non-Negative Matrix Factorization
NMFD	Non-Negative Matrix Factor Deconvolution
PESQ	Perceptual Evaluation of Speech Quality
PLLT	Prédiction Linéaire à Long Terme
SNR	Signal to Noise Ratio
SRR	Signal to Reverberant Ratio
TFCT	Transformée de Fourier à Court Terme
WER	Word Error Rate
segSRR	Segmental Signal to Reverberant Ratio

Chapitre 1

Introduction

Les signaux audio, que ce soit de la parole ou de la musique, sont soumis à différentes dégradations lors de la prise de son. Dans le contexte de la téléphonie en particulier, il est indispensable de transmettre un signal de parole dépourvu de signaux parasites afin de préserver la clarté de la conversation. Les signaux indésirables peuvent être classés en trois catégories : le bruit de fond, aléatoire et décorrélé de la source, les échos acoustiques, souvent présents lors de l'utilisation de systèmes de communication mains libres, et la réverbération due à l'effet de salle. Ces trois signaux parasites se superposent au signal d'interêt et introduisent des distorsions qui nuisent à l'intelligibilité de la parole. Les algorithmes visant à réduire l'influence des ces signaux sur le signal d'intêret sont dits de réhaussement de la parole (*speech enhancement*, en anglais). Ainsi, les trois principaux problèmes de recherche de l'amélioration de la parole sont : la réduction du bruit, l'annulation d'échos et la déréverbération [28].

Actuellement il existe de nombreuses solutions pour la réduction du bruit de fond en se basant sur un modèle statistique de celui-ci. Ainsi, dans [13] par exemple, à partir d'une observation du signal de bruit et en supposant que celui-ci est blanc, gaussien et stationnaire, on peut estimer sa variance et donc sa puissance afin de le soustraire de l'enregistrement en appliquant un masque de Wiener dans le domaine spectral. Le masque de Wiener [37] donne la meilleure approximation du signal débruité dans le sens de l'erreur quadratique moyenne.

De même, le problème de l'annulation des échos a été largement abordé. Les échos acoustiques apparaissent lorsqu'un microphone capte le son émis par un haut parleur situé dans la même salle. Cela est le cas notamment pour tous les systèmes de communication *mains libres* (dans les voitures, pour les téléconférences, ...), ce qui explique la généralisation de traitements d'annulation d'échos dans les dispositifs téléphoniques. Une des approches les plus communes consiste à approcher le filtre du canal qui génére l'écho afin de modéliser de façon indépendante le signal indésirable, en vue de sa soustraction du signal observé. L'estimation de la réponse impulsionnelle se fait, en général, en minimisant l'erreur quadratique moyenne de l'écho résiduel grâce à l'algorithme adaptatif Normalized Least Mean Squares (NLMS)[17].

Ces techniques sont suffisamment matures aujourd'hui pour considérer ces deux problématiques de recherche comme closes. Elles fonctionnent aussi bien en temps réel que hors ligne et de nombreuses améliorations pour réduire l'influence des artefacts introduits par le traitement (bruit musical, bruit résiduel) ont été présentées [7, 8, 9]. D'autre part, le problème de la réduction de la réverbération n'a été abordé que récemment. Ce problème est plus compliqué que les deux précédents car le signal parasite à réduire est très corrélé à court terme avec le signal anéchoïque souhaité. Il s'agit d'un bruit non stationnaire, contrairement au bruit de fond, ce qui exige l'utilisation de modèles plus complexes et fléxibles pour sa répresentation. Mon stage de fin d'études consiste donc à mettre en place des techniques de déréverbération performantes dans le but d'améliorer la qualité des enregistrements de parole.

Ce stage s'est déroulé dans le cadre du partenariat entre la société Arkamys, spécialisée dans le traitement du son, et le groupe de recherche AAO (Audio, Acoustique et Ondes) du département TSI (Traitement du Signal et des Images) de Telecom ParisTech. Il est également une partie intégrante du doctorat sous convention CIFRE que je réalise depuis le mois de Mai 2011 avec ces deux organismes sous la direction conjointe d'Ivan Bourmeyster, directeur technique d'Arkamys, de Gaël Richard, enseignant chercheur du département TSI et d'Yves Grenier, directeur du département TSI. On présentera brièvement comment mon stage s'intègre dans les besoins et le savoir faire d'Arkamys et Telecom ParisTech avant d'aborder une description détaillée des objectifs à atteindre au cours de ces premiers mois de doctorat qui constituent mon stage de fin d'études pour le Master ATIAM.

1.1 Présentation des organismes d'accueil

1.1.1 Arkamys

La société Arkamys se spécialise depuis 10 ans dans le traitement numérique du son. Cette PME qui a débuté en proposant à ses clients des algorithmes de spatialisation pour la production cinématographique, se consacre actuellement à developper des traitements temps réel de spatialisation, d'élargissement de la stéréo, de localisation de sources sonores et d'amélioration de la parole pour des domaines très variés qui vont de l'industrie du cinéma à l'électronique embarquée grand public. C'est dans ce dernier secteur qu'Arkamys propose sa plus large gamme de solutions autour de trois axes principaux d'application : la téléphonie mobile, l'audio dans les véhicules et les *home-cinémas*.

Dans le domaine de la téléphonie mobile, Arkamys fournit à ses clients des algorithmes de réduction du bruit de fond et d'annulation d'échos qui sont intégrés dans les dispositifs mobiles mais aussi dans les systèmes de communication mains libres intégrés dans l'autoradio des voitures. Dans ces derniers, le microphone qui capte la parole est éloigné du locuteur, le signal acquis est donc dégradé par le bruit ambiant (bruit de roulage, signaux parasites, ...), par les échos mais aussi par la réverbération intrinsèque à l'habitacle du véhicule.

Des études sur les effets de la réverbération sur l'intelligibilité de la parole ([2, 24, 26]) ont démontré qu'une composante de la réverbération, la réverbération tardive (voir 2.1), rend le signal moins intelligible tandis qu'une autre composante, les réflexions précoces, améliorent la perception de la parole. En partant de ce constat, Arkamys souhaite développer des algorithmes visant à réduire les effets nuisibles de la réverbération. La recherche menée au cours de mon stage et de mon doctorat a donc comme objectif final la mise en place de traitements temps réel de déréverbération qui seront ensuite intégrés dans la chaine de traitement de la parole de l'offre d'Arkamys.

1.1.2 Département TSI de Telecom ParisTech

Le département TSI de Telecom ParisTech concentre principalement sa recherche dans le traitement du signal audio visuel. Le département se décompose en quatre équipes : Statistiques et Applications (STA), Traitement et Interprétation des Images (TII), Audio Acoustique et Ondes (AAO) et Multimédia (MM). L'équipe AAO, à laquelle je suis rattaché, s'intéresse au traitement de toute la chaine du signal audio, que ce soit de la musique ou de la parole.

Dans le domaine du traitement de la parole, l'équipe AAO possède de fortes compétences en matière du codage de la parole, de sa modélisation mais aussi en analyse et synthèse vocale. Des études ont également été menées dans le débruitage de l'audio et dans la réduction des échos. Cependant aucune recherche précise n'a encore été réalisée dans le domaine de la déréverbération. Or, des travaux récents dans la reconnaissance de la parole pour des applications en robotique (Projet ROMEO [1]) ont mis en évidence la nécessité de développer des techniques pour réduire la réverbération afin d'améliorer la performance du système de reconnaissance vocale.

D'autre part, l'équipe AAO développe des méthodes de séparation des sources pour les signaux audio. Ces méthodes introduisent des modèles *a priori* sur la structure des signaux à séparer afin de pouvoir les identifier. Dans le cas de la parole on exploite le fait qu'à chaque trame du spectrogramme à court terme, il n'y a qu'un nombre restreint de fréquences qui sont activées. On dit que le signal de parole est parcimonieux. La propriété de parcimonie donne une représentation compacte des spectrogrammes, comme l'illustre la figure 1.1, et peut être également exploitée dans les techniques de déréverbération afin de ne garder que les quelques composantes spectrales du spectrogramme qui correspondent au signal anéchoique qu'on veut estimer.

1.2 Sujet du stage

Mon stage de fin d'études a donc pour but d'enrichir les compétences d'Arkamys et de Telecom ParisTech en matière du traitement de la réverbération. S'agissant du début de mon doctorat sur le sujet de la déréverbération, il doit me donner les bases qui permettront de faire avancer ma recherche. On souhaite disposer à la fin du stage

1. http ://www.projetromeo.com

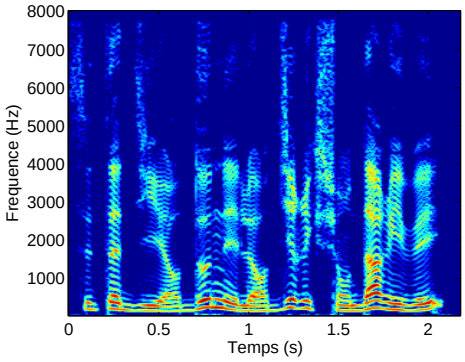


FIGURE 1.1 – Spectrogramme d'un signal anéchoique de parole échantillonné à 16kHz

d'une culture suffisante en matière de déréverbération ainsi que d'algorithmes fonctionnant sous Matlab. Ces algorithmes serviront de référence pour comparer les nouvelles méthodes de réduction de la réverbération des signaux audio, qui seront développées dans la suite du doctorat. Même si à la fin de ce travail de recherche on souhaite fournir des solutions pouvant fonctionner en temps réel, on laisse cette contrainte de côté pour l'instant. On mettra uniquement l'accent sur la recherche d'algorithmes permettant d'obtenir de bons résultats en matière de déréverbération qui pourront ultérieurement être adaptés à un fonctionnement en temps réel.

Dans un premier temps, on étudiera le phénomène de la réverbération afin de comprendre ses caractéristiques et les principaux modèles de répresentation proposés dans la littérature. On analysera également les effets de la réverbération sur les signaux de parole, la nature des dégradations introduites et leur influence sur l'intelligibilité des énonciations.

Ensuite, un parcours bibliographique approfondi des techniques de déréverbération sera réalisé. Plusieurs familles de traitements seront mises en évidence, étudiées et analysées afin de définir les types d'approche qui sont potentiellement adaptables à un fonctionnement en temps réel tout en garantissant une faible distorsion du signal de parole déréverbéré. Dans cette étude, nous ne nous limiterons pas aux approches parcimonieuses, l'idée étant d'avoir un aperçu global des techniques de déréverbération. Les contraintes de parcimonie ne seront introduites que de façon succincte pour comprendre comment elles peuvent s'intégrer aux traitements.

Enfin, deux approches différentes seront choisies pour être implémentées sous Matlab. Les performances de ces méthodes seront d'abord validées grâce à des écoutes informelles avant d'être quantifiées par des mesures de déréverbération proposées dans la littérature. Quelques modifications des techniques implémentées seront également proposées dans le but d'améliorer les résultats obtenus.

Nous établirons donc un état de l'art des techniques de déréverbération dans le Chapitre 2. Ensuite, dans les Chapitres 3 et 4 nous présenterons les deux techniques de déréverbération qui ont été implantées et testées lors du stage. Enfin dans le Chapitre 5 nous présenterons l'évaluation des différents algorithmes avant de conclure dans le Chapitre 6.

Chapitre 2

État de l'art

2.1 La réverbération

2.1.1 Définition de la réverbération

Lorsque le signal de parole se propage dans un espace fermé, le signal observé au niveau des microphones est formé par la superposition de plusieurs versions retardées et atténuées de la source. En effet, si on se place dans le cas de la figure 2.1 où on considère une source omnidirectionnelle, le microphone capte tout d'abord le signal source, encore appelé signal direct, mais également toutes les réflexions sur les parois. Les différentes versions du signal ont parcouru des chemins acoustiques de différentes longueurs et ont été atténuées par l'absorption des parois, la phase et l'amplitude des diverses versions de la source captées par le microphone sont donc différentes. Le signal observé peut également être dépourvu du signal direct si un obstacle est placé entre la source et le capteur, dans ce cas on observe uniquement les réflexions de la source.

On distingue deux types de réflexions : les réflexions précoces et la réverbération tardive. Les réflexions précoces arrivent avec un faible retard par rapport au son direct (de 0 à 30ms). Elles sont séparées temporellement et spatialement du son direct mais elles ne sont pas perçues séparément grâce à l'effet de précédence [35]. Ces réflexions sont identifiées par leur ordre, qui traduit le nombre de réflexions subies avant d'atteindre le capteur. Lorsqu'un chemin acoustique rencontre uniquement une paroi, on parle d'une réflexion de premier ordre, s'il en rencontre deux, il s'agit d'une réflexion de second ordre et ainsi de suite. Il convient de remarquer que l'ordre d'une réflexion n'est pas forcément lié à l'ordre d'arrivée des fronts d'onde, on peut très bien observer des réflexions du second ordre qui précèdent celles du premier ordre. Du fait de l'intégration temporelle de ces réflexions par l'oreille, certaines caractéristiques de la parole sont mises en relief, ce qui favorise son intelligibilité. Cela explique le fait qu'il soit plus aisé de tenir une conversation dans une salle qu'en plein air, par exemple.

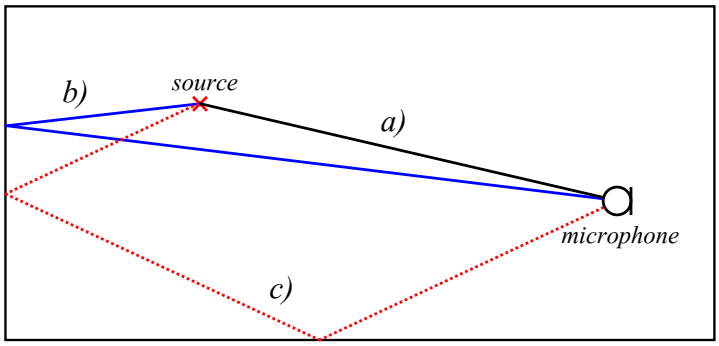


FIGURE 2.1 – Chemins Acoustiques multiples :
a) Chemin direct,
b) Réflexion de premier ordre,
c) Réflexion de second ordre

La réverbération tardive apparait environ 30ms après l'arrivée du son direct. Quand le delai augmente, les réflexions se succèdent et deviennent infiniment nombreuses et rapprochées dans le temps, il n'est plus possible de les séparer et parler de l'ordre des réflexions n'a plus de sens. Il est alors plus commode de considérer l'ensemble des réflexions comme une distribution aléatoire dont la densité augmente avec le temps [16]. Si elles arrivent avec un retard supérieur à 100ms, elles sont perçues comme de l'écho, sinon il s'agit bien de la réverbération. Cette composante du signal dégrade la qualité de la parole et son intelligibilité [26].

La composition du canal, et donc la distribution des différentes réflexions, peut être décrite par la Réponse Impulsionnelle du canal, notée $h(n)$. Il s'agit du signal observé au niveau du microphone lorsque la source est une impulsion de Dirac. La réponse impulsionnelle peut être interprétée comme le filtre qui est appliqué au signal anéchoïque $s(n)$ pour produire le signal réverbérant $x_0(n)$ tel que :

$$\begin{aligned} x_0(n) &= (s * h)(n) = \sum_{i=0}^{\infty} h(i)s(n-i) \\ &= \underbrace{h(0)s(n)}_{\text{signal direct}} + \underbrace{\sum_{i=1}^{\tau-1} h(i)s(n-i)}_{\text{réflexions précoces}} + \underbrace{\sum_{i=\tau}^{\infty} h(i)s(n-i)}_{\text{réverbération tardive}} \end{aligned}$$

où $\tau = 30ms$ est la limite communément fixée entre les réflexions précoces et la réverbération tardive.

La figure 2.2 donne un exemple de réponse impulsionnelle de salle simulée par la méthode de l'image [1]. On distingue bien les deux types de réflexion décrits dans cette partie. Le retard initial correspond au temps de parcours du signal direct entre la source et le capteur. En bleu, on observe les premières réflexions qui sont bien séparées les unes des autres et en vert, la réverberbération tardive, distribuée de façon continue. On constate également que la réponse impulsionnelle décroit de façon exponentielle avec le temps, on détaillera cette propriété par la suite.

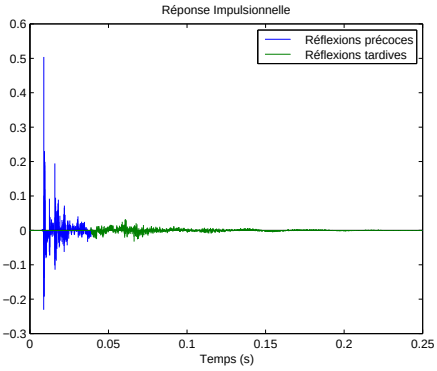


FIGURE 2.2 – Exemple de réponse impulsionnelle

2.1.2 Caractéristiques de la réverbération

La réponse impulsionnelle de la salle décrit l'évolution de la réverbération dans celle-ci. Dans [31], Polack modélise la réverbération comme une réalisation d'un processus stochastique non stationnaire. Selon ce modèle, la réponse impulsionnelle $h(t)$ est donnée par :

$$h(t) = b(t)e^{-\delta t}, \quad \text{pour } t \geq 0 \tag{2.1}$$

où $b(t)$ est un bruit gaussien centré et stationnaire, que l'on supposera blanc dans un premier temps, et δ est la constante de décroissance de la réponse impulsionnelle. En pratique on préfère définir ce temps de décroissance par une autre grandeur, le temps de réverbération, noté RT_{60} et donné par :

$$RT_{60} = \frac{3\ln(10)}{\delta}$$

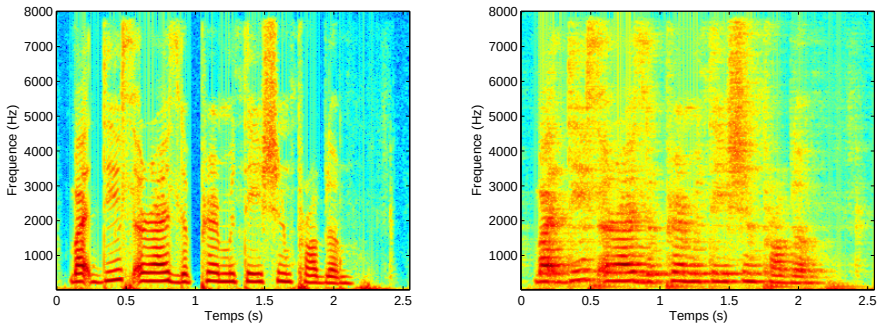
Le RT_{60} a d'abord été proposé par Sabine dans [33]. Il correspond au temps en secondes nécessaire pour observer une atténuation du niveau du signal de 60 dB à partir du moment où la source cesse d'émettre. De manière empirique, Sabine a établi la loi suivante :

$$RT_{60} = \frac{4\ln 10^6}{c}\frac{V}{Sa} \approx 0.1611\frac{V}{Sa} \tag{2.2}$$

qui définit le RT_{60} comme une fonction du volume de la salle V, de sa surface totale S, du coefficient d'absorption moyen des parois de la salle a et de la vitesse du son c. Cette loi peut également être utilisée pour calculer le coefficient d'absorption de la salle connaissant ses dimensions et son temps de réverbération.

2.1.3 Effets de la réverbération sur la parole

La réverbération modifie les caractéristiques temporelles et spectrales du signal de parole. La figure 2.3(a) montre le spectrogramme d'un signal anéchoïque. On observe que les transitoires des phonèmes sont bien localisés dans le temps et les formants de la parole apparaissent clairement. Également, les silences entre les phonèmes sont clairement identifiables. Ce même signal est filtré par une réponse impulsionnelle artificielle de 300ms de temps de réverbération ($RT_{60} = 300ms$), on observe son spectrogramme dans la figure 2.3(b). Le signal apparait étalé en temps et en fréquence. Les attaques sont donc moins nettes et les modulations des formants se perdent. Donc, les diffèrents phonèmes se superposent et occupent les zones de silence. Ces modifications affectent évidemment la perception de la parole comme on le verra par la suite.



(a) Signal anéchoïque (b) Signal réverbérant ($RT_{60} = 300ms$)

FIGURE 2.3 – Spectrogrammes d'une énonciation

Les réflexions précoces introduisent une sensation d'éloignement de la source ainsi qu'une coloration du signal. Ces modifications n'affectent pas pour autant l'intelligibilité de la parole, bien que leur existence dégrade les performances des systèmes de reconnaissance automatique de la parole [28]. Au contraire, il a été démontré que la présence de ces réflexions rendait plus intelligible le signal grâce à des interférences constructives entre le signal direct et ses versions retardées.

Dans [14], Haas étudie l'influence d'un simple écho sur la perception de la parole. Il apparait alors que lorsque l'écho a un retard compris entre 0 et 30ms, les auditeurs ne perçoivent pas l'écho mais remarquent une augmentation de l'intensité du signal ainsi qu'un élargissement de la source qui a un effet bénéfique pour la compréhension. Pour des retards supérieurs à 30ms, les auditeurs notent une dégradation de la parole.

Dans [26], Nábêlek étend l'étude au cas des signaux réverbérants et met en évidence deux effets de la réverbération sur la parole : le masquage propre (*self-masking*) et le masquage de recouvrement (*overlap-masking*). Le premier apparait au sein d'un même phonème et se traduit par un ralentissement de l'attaque des transitoires ainsi qu'une dégradation des transitions entre les formants des voyelles. Le masquage propre est principalement du aux réflexions précoces. Le masquage de recouvrement, d'autre part, apparait quand un phonème se superpose à celui qui le suit du fait de l'étalement temporel du signal. La queue de la réverbération occupe alors les régions de faible énergie du

signal, en réduisant ainsi son Index de Modulation [2], ce qui se traduit par une perte de précision dans la séparation des différents phonèmes. Ce masquage est une conséquence de la réverbération tardive et affecte l'intelligibilité du deuxième phonème.

On remarque donc que la réverbération tardive affecte l'intelligibilité du signal de parole de façon plus importante que les réflexions précoces. Les algorithmes de déréverbération que nous étudierons dans la section 2.2 abordent séparément les deux types de réflexions afin de s'adapter à leur structure particulière. La réduction de la réverbération tardive reste cependant le principal problème à résoudre lorqu'il s'agit d'améliorer la parole.

2.2 Techniques de déréverbération

Dans la suite nous appelerons *source* le signal anéchoïque que l'on souhaite estimer, *observation* le signal réverbérant acquis par le(s) microphone(s). La figure 2.4 illustre le cadre général des techniques de déréverbérationqu'on souhaite développer. La source $s(n)$ se propage dans l'espace réverbérant et subit le filtrage du canal acoustique $h(n)$. A ce signal s'ajoute le bruit additif $b(n)$ qui inclut le bruit du capteur, le bruit ambiant et tout signal parasite pouvant se superposer au signal. On reçoit ainsi, au niveau des microphones, le signal $x_0(n)$ qui sera traité par le module de déréverbération afin de produire le signal $\hat{s}(n)$, l'estimée de la source. En pratique, on suppose que le signal observé a été prétraité par un algorithme de réduction du bruit ce qui nous permet de négliger la perturbation $b(t)$ dans les modèles de déréverbération que nous allons analyser.

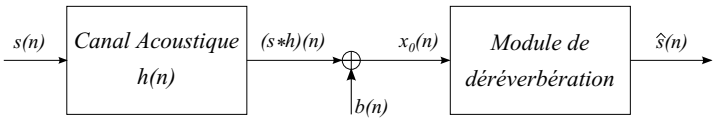


FIGURE 2.4 – Schéma général des algorithmes de déréverbération

On distingue deux approches pour la déréverbération : les algorithmes qui calculent directement une estimation du signal source à partir des observations et ceux qui cherchent plutôt à estimer le filtre du canal $h(n)$ pour ensuite appliquer un filtrage inverse à l'observation $x_0(n)$[27]. Dans le premier cas on parle de techniques de suppression de la réverbération et dans le deuxième de techniques d'annulation de la réverbération [15].

2.2.1 Techniques de suppression de la réverbération

Les méthodes d'estimation directe de la source requièrent le choix d'un modèle de la structure du signal de parole. Le modèle de prédiction linéaire (Linear Prediction, en anglais), très utilisé en codage de la parole, est souvent choisi. En travaillant sur le résiduel du modèle de prédiction linéaire du signal, Kinoshita [20] isole la réverbération

tardive pour ensuite établir un modèle de Prédiction Linéaire à Long Terme (PLLT). Ainsi il peut estimer le filtre $a(n)$ qui modélise au mieux la partie réverbérante du signal. Ensuite, en comparant les spectres de l'observation et de l'estimation de la réverbération, une soustraction spectrale est effectuée. Ensuite les réverbérations précoces sont réduites par soustraction de la moyenne cepstrale (Cepstral Mean Subtraction, en anglais), qui est une technique classique pour réduire l'effet du canal sur un signal de parole [4][3]. Des tests sont réalisés sur un système de reconnaissance vocale. La technique réduit le taux d'erreur de reconnaissance (Word Error Rate, en anglais) par rapport à un signal traité uniquement par CMS. La performance augmente avec le nombre de microphones. Des bons résultats sont obtenus avec 4 microphones mais même avec un seul capteur l'efficacité de la reconnaissance vocale est améliorée.

Dans [39], Yasuroaka modélise également le signal réverbéré par un processus LP, tandis que la source est modélisée par un modèle de mélanges gaussiens (Gaussian Mixture Model, en anglais) tel que chaque composante du mélange gaussien soit centrée sur une harmonique de la parole. Avec ces deux modèles, un masque de Wiener est calculé. Ce masque permet de réaliser le filtrage de la réverbération dans le domaine spectral.

Dans [21], la réponse impulsionnelle du canal est estimée à partir de l'enregistrement d'un claquement de main. Ensuite, le résiduel LP de l'observation est filtré par cette réponse pour approcher la partie réverbérante du signal. En passant dans le domaine spectral on peut récupérer le signal direct par simple soustraction spectrale de la forme $P_d(\kappa, \omega) = P_{obs}(\kappa, \omega) - P_{r\hat{e}v}(\kappa, \omega)$. Par ce biais, un masque de Wiener est également calculé afin de réduire le bruit de fond dans la même chaine de traitements.

Dans [10] et [11], la déreverbération est obtenue en appliquant des méthodes de Monte Carlo. Ces méthodes stochastiques permettent de construire un modèle plus fléxible et robuste du signal de parole que ce soit par des traitements causaux ou non-causaux. Dans [11], une étude comparative entre les approches Monte Carlo par chaines de Markov et les méthodes Monte Carlo séquentielles est presentée. Les méthodes séquentielles sont pertinentes pour un traitement en ligne. La structure à long terme du signal n'est pas prise en compte par l'algorithme, qui s'adapte aux évolutions échantillon par échantillon des observations. À l'opposé, les méthodes par chaines de Markov décomposent le signal dans une base de fonctions de référence et donnent un modèle qui permet de reconstruire exactement le signal, connaissant les paramètres. La performance du système dépend alors de la pertinence du choix des fonctions de base. Ces méthodes ne s'appliquent pas aux traitements en temps réel. En effet, pour établir le modèle on doit connaitre à l'avance l'intégralité du signal. Cependant, le fait de connaitre le signal à l'avance permet d'utiliser des méthodes plus simples, avec des hypothèses moins fortes. Dans le cas séquentiel [10], l'évolution des paramètres du modèle de parole est calculée par un *random walk* d'une chaine de Markov d'ordre 1. Ainsi, la densité de probabilité de la source est approchée à chaque itération par un nuage de particules (variables aléatoires). Lorsque la chaine de Markov atteint sa loi stationnaire, on récupère une estimation de la source de façon efficace.

Kameoka [19] présente un algorithme itératif qui permet d'estimer le spectrogramme de la source par une variante de la Factorisation en Matrices Non-négatives (Non-Negative Matrix Factorization, en anglais). En minimisant une fonction de coût régularisée par des contraintes de parcimonie de la source, l'algorithme exprime le spectrogramme de l'observation comme le produit de convolution de deux spectrogrammes :

celui de la source et celui du canal. Ainsi on converge vers une version dérverbérée du signal observé. Cette solution est très efficace et peut être aisément implémentée en temps réel.

2.2.2 Techniques d'annulation de la réverbération

Les techniques d'annulation de la réverbération consistent à estimer le filtre du canal acoustique pour ensuite appliquer le filtre inverse à l'observation. Le principal problème de cette approche est que souvent les filtres du canal ne sont pas à minimum de phase, leur inverse est donc instable [29]. On doit donc chercher des approximations stables du filtre inverse pour pouvoir récupérer le signal désiré. Miyoshi démontre le *Multiple Input/Output Inverse Theorem* (MINT) dans [25]. Ce théorème établit que dans le cas d'une observation multicanal il est possible de calculer l'inverse stable exact du filtre à condition que les différents chemins acoustiques ne possèdent aucun zéro en commun. Dans [12], Furuya utilise le principe MINT pour calculer le filtre inverse à partir de la matrice de corrélation du signal observé. Ainsi, l'effet des réflexions précoces est atténué par le filtrage, ensuite la réverbération tardive est supprimée par soustraction spectrale. En addition à cela, un lissage temporel est appliqué sur la matrice d'autocorrélation pour rendre la méthode robuste aux déplacements du locuteur par rapport au microphone et donc aux variations de la réponse impulsionnelle du canal.

Hazrati décompose le signal en sous-bandes grâce à un banc de filtres *gammatone* [18]. Un filtre inverse est calculé pour chaque sous-bande en appliquant le principe MINT puis le signal est filtré et les réverbérations tardives sont réduites par soustraction spectrale.

Il est également possible d'estimer le filtre de canal par minimisation de l'erreur quadratique moyenne de la prédiction. Dans [27], plusieurs variantes de cette technique sont introduites, cependant la plupart d'entre elles manquent de robustesse aux erreurs de modélisation de la source et du canal.

Le principal inconvénient de cette approche est que la réponse impulsionnelle de la salle dépend de la position de la source et du récepteur et dans le cas de la téléphonie, le locuteur déplace sa tête au long de la conversation, il modifie ainsi la réponse impulsionnelle. Ceci oblige à recalculer la réponse impulsionnelle en permanence, ce qui est couteux en temps de calcul.

2.3 Mesures de déréverbération

Une fois la déréverbération effectuée on doit pouvoir évaluer les différents algorithmes selon des critères objectifs et subjectifs. Bien qu'il existe de nombreuses métriques pour comparer deux signaux, leur évaluation peut ne pas être cohérente avec l'évaluation subjective. Ainsi, par exemple dans les algorithmes de débruitage on sait que l'application d'un filtrage de Wiener donne les meilleurs résultats perceptuels, cependant le rapport signal à bruit apparait dégradé. Une partie du stage consiste donc à étudier plusieurs mesures de réverbération afin de choisir celle ou celles qui sont le mieux corrélées avec la

perception de la réduction de la réverbération apportée par un algorithme donné. Dans la plupart des cas, on calcule des mesures intrusives, c'est à dire qu'on se place dans la situation idéale où l'on dispose du signal de référence qu'on souhaite estimer. Si, au contraire, on calcule les métriques à partir de la seule observation, sans utiliser un signal de référence, on parle de mesures non intrusives. Nous passons en revue certaines de ces mesures par la suite.

2.3.1 Signal to Noise Ratio (SNR)

Une première idée consiste à évaluer le rapport signal à bruit que l'on notera par la suite SNR. Le SNR est donné par :

$$\text{SNR} = 20\log_{10}\left(\frac{||s||_2}{||s-\hat{s}||_2}\right) \text{dB} \tag{2.3}$$

où s désigne la source anéchoïque à estimer, $\hat{s}$ l'estimée de la source après déréverbération et l'opérateur $||.||_2$ désigne la norme euclidienne. Lorsque les signaux sont proches, le SNR est élévé, au contraire lorsque les signaux diffèrent le SNR tend vers $-\infty$. Bien que le SNR donne une idée de la similarité des formes d'onde entre le signal original et le signal estimé, cette mesure générique ne donne que très peu d'information sur la présence de réverbération. En effet, il suffit d'imaginer le cas où on introduit des bruits parasites et des artefacts en effectuant la déréverbération. Dans ce cas, les bruits parasites introduits feront diminuer le SNR même si la réverbération en soit a été réduite.

2.3.2 Signal to Reverberant Ratio (SRR)

Le rapport signal à réverbérant, noté SRR, est une mesure intrusive de déréverbération donnée par :

$$\text{SRR} = 20\log_{10}\left(\frac{||s_d||_2}{||s_d-\hat{s}||_2}\right) \quad \text{(dB)} \tag{2.4}$$

Pour le calculer, on doit connaitre s_d, qui contient la composante directe du signal observé et les premières réflexions. En général, on compare les valeurs du SRR du signal réverbérant avec celles du signal traité, en espérant constater une augmentation de SRR. Cette mesure illustre bien l'effet de la réverbération à condition d'estimer correctement s_d. Ceci est le cas lorsqu'on travaille avec une réponse impulsionnelle connue ou simulée par la méthode de l'image [1], par exemple.

Il est parfois préférable de mesurer le SRR localement, sur des segments courts du signal : on parle alors de Segmental Signal to Reverberant Ratio (segSRR). Dans ce cas, on divise le signal à mesurer en trames courtes (de 32ms par exemple) avec ou sans recouvrement. Ensuite on calcule le SRR pour chaque trame et on prend la moyenne le long des trames pour obtenir le SRR_{seg}.

$$\text{SRR}_{\text{seg}} = \frac{1}{N_{seg}} \sum_{l=0}^{N_{seg}-1} \text{SRR}(l) \quad \text{(dB)} \tag{2.5}$$

où N_{seg} est le nombre de segments de parole à traiter.

2.3.3 Log Spectral Distorsion (LSD)

Le calcul de la distorsion spectrale, notée LSD, se fait dans le domaine fréquentiel. La LSD est une des premières mesures introduites pour évaluer les algorithmes de traitement de la parole. Il s'agit d'une mesure symétrique de la similarité entre deux spectres. Pour l'obtenir, on calcule d'abord la Transformée de Fourier à Court Terme (TFCT) du signal de référence $s(n)$ et du signal estimé $\hat{s}(n)$. On obtient ainsi les spectrogrammes $S(l,k)$ et $\hat{S}(l,k)$, où l est l'indice de la trame et k le canal fréquentiel. On définit alors la distorsion spectrale de chaque trame par :

$$\mathrm{LSD}(l) = \left(\frac{1}{K} \sum_{k=0}^{K-1} \left| L\{\hat{S}(l,k)\} - L\{S(l,k)\} \right|^2 \right)^{\frac{1}{2}} \quad \text{(dB)} \tag{2.6}$$

où $L\{X(l,k)\} \triangleq max\{20\log_{10}(|X(l,k)|), \delta\}$ est le spectre en échelle logarithmique du signal, restreint à une dynamique de 50 dB ($\delta = max_{l,k}\{20\log_{10}(|X(l,k)|)\} - 50$).

2.3.4 Direct to Reverberant Ratio (DRR)

Le rapport direct à réverbérant, noté DRR est une mesure de la déréverbération basée uniquement sur le canal. On n'a donc pas besoin de connaitre le signal anéchoïque de référence pour la calculer mais uniquement la réponse impulsionnelle du canal. Le DRR est donné par :

$$\mathrm{DRR} = 10\log_{10} \left(\frac{\sum_{n=0}^{n_d} h^2(n)}{\sum_{n=n_d+1}^{\infty} h^2(n)} \right) \quad \text{(dB)} \tag{2.7}$$

où $h(n)$ est la réponse impulsionnelle du canal et n_d est l'instant d'arrivée du signal direct au capteur. On peut facilement estimer la réponse impulsionnelle du signal traité par un algorithme de déréverbération grâce à l'algorithme NLMSqui à partir de la source anéchoïque estime le filtre permettant d'obtenir le signal observé. Lorsque le signal est stationnaire, le DRR et le SRR sont équivalents [15].

2.3.5 Bark Spectral Distortion (BSD)

La distorsion spectrale de Bark est calculée sur le spectre du signal, projeté sur l'échelle perceptuelle de Bark. L'échelle de Bark est un échelle psychoacoustique qui traduit la non linéarité de l'oreille dans la perception des sons. Dans [36], Wang démontre qu'il existe une forte corrélation entre la BSD et le Mean Opinion Score (MOS) (voir 2.3.6). La BSD entre la source $s(n)$ et son estimée $\hat{s}(n)$ est calculée sur les spectres de Bark $B_s(l,k)$ et $B_{\hat{s}}(l,k)$ des deux signaux. Selon Wang, le spectre de Bark se construit

à partir du spectre du signal réalisant un filtrage en bandes critiques, suivi d'une préaccentuation de l'intensité et d'une conversion de l'echelle des phons à celle des sones [38]. Une fois les spectres de Bark calculés, la BSD est donnée par :

$$\mathrm{BSD} = \frac{1}{L}\sum_{l=0}^{L-1}\frac{\sum_{k=1}^{K}(B_s(l,k) - B_{\hat{s}}(l,k))^2}{\sum_{k=1}^{K}(B_s(l,k))^2} \tag{2.8}$$

où L est le nombre de trames et K le nombre de canaux du spectre de Bark.

2.3.6 Mesures subjectives

Les mesures objectives sont un bon indicateur des performances d'un algorithme. Cependant, il faut s'assurer que ces mesures reflètent bien la perception des auditeurs, d'où la nécessité de mener également une évaluation subjective. Ceci est réalisé grâce à des tests d'écoute qu'un panel d'auditeurs plus ou moins avisés doit passer. Lors du test chaque sujet doit donner une note dans une échelle d'opinion selon la qualité du son perçu. Ensuite la moyenne sur tous les sujets est calculée afin d'établir la note de l'algorithme, le Mean Opinion Score (MOS).

Le MOS est une mesure communément utilisée pour valider la qualité de la chaine du traitement de la parole dans les télécommunications. Cependant, une évaluation fiable nécéssite un panel d'auditeurs assez grand ce qui est difficile et coûteux à réaliser en pratique. Une solution consiste alors en la construction de mesures objectives qui sont en forte corrélation avec le MOS, c'est le cas de la BSD (voir 2.3.5), qui a un coefficient de corrélation supérieur à 0.9, ou du Perceptual Evaluation of Speech Quality (PESQ) [32] qui est une méthode recommandée par l'ITU-T dans le but de valider les codeurs de parole dans le contexte des télécommunications. Cette estimation se fait aussi bien dans le domaine temporel que dans le domaine spectral.

Chapitre 3

Déréverbération par prédiction linéaire à long terme

Dans [20], Kinoshita *et al.* optent pour une approche par prédiction linéaire pour estimer la source anéchoïque et ainsi réduire la réverbération. À partir d'une ou plusieurs observations d'un signal, l'approche proposée cherche à estimer un modèle de Prédiction Linéaire à Long Terme (PLLT) de la partie réverbérante du signal afin de réduire son influence par soustraction spectrale [22]. L'algorithme est mis à l'épreuve en l'intégrant dans un système de reconnaissance vocale automatique, on observe une nette diminution du taux d'erreurs de reconnaissance. Moyennant une augmentation du temps de calcul, les performances de la méthode sont améliorées dès que l'on dispose de plusieurs versions, acquises par différents microphones, du signal à traiter. Dans le contexte de ce stage, on cherche à privilégier les méthodes utilisant un ou deux microphones et étant facilement adaptables à un fonctionnement en temps réel. On présentera donc dans la suite, la version de l'algorithme qui fonctionne en utilisant uniquement une observation du signal.

La figure 3.1 résume l'ensemble des étapes de l'algorithme. En entrée du système, le signal $x_0(n)$, acquis par le microphone, est passé par un filtre blanchisseur qui a pour but de supprimer la corrélation à court terme, intrinsèque au signal de parole. On suppose alors que le signal résultant, $x_{white}(n)$, contient la partie réverbérante du signal, qui est corrélée à long terme avec la source anéchoïque $s(n)$. En calculant le modèle de PLLT de $x_{white}(n)$ on calcule le filtre $a(n)$ qui modélise correctement la réverbération tardive. En filtrant $x_0(n)$ par $a(n)$ on obtient le signal $x_{late}(n)$, une estimation de la réverbération tardive. Ensuite on calcule $X_0(l,k)$ et $X_{late}(l,k)$, la TFCT de l'observation et de $x_{late}(n)$ respectivement, pour atténuer la réverbération tardive dans le domaine spectral par soustraction spectrale. On récupère ainsi le signal $x_1(n)$ par TFCT inverse. Ce dernier est dépourvu de la réverbération tardive. Enfin on effectue la soustraction de la moyenne cepstrale (CMS) [30] de ce signal afin d'atténuer les réflexions précoces et obtenir l'estimée du signal anéchoïque $\hat{s}(n)$. L'ensemble du traitement est appliqué énonciation par énonciation, on suppose donc dans un premier temps qu'on dispose d'enregistrements de parole préalablement segmentés par un détecteur d'activité vocale. Nous détaillerons chacune de ces étapes par la suite.

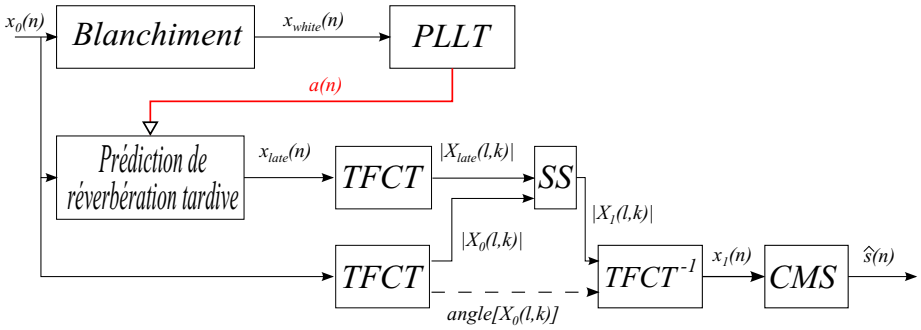


FIGURE 3.1 – Schéma de l'algorithme

3.1 Blanchiment du signal

L'opération de blanchiment du signal (*prewhitening*, en anglais) consiste à modifier le spectre du signal de sorte à le rapprocher du spectre d'un bruit blanc, c'est à dire d'un spectre constant quelque soit la fréquence (Figure 3.2). Pour le signal de parole, cela correspond à retirer l'effet des formants de la voix, i.e. des résonances du conduit vocal, sur le spectre. Le signal résultant est donc proche d'un bruit blanc pour les parties non voisées de la voix. Les parties voisées du signal persistent dans le signal blanchi.

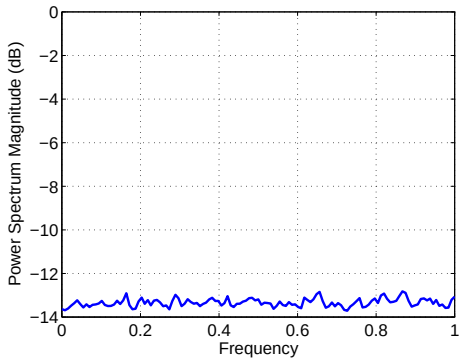


FIGURE 3.2 – Spectre d'un bruit blanc gaussien

Le blanchiment est le principe de base du codage de la parole par prédiction linéaire (Linear Prediction Coding, en l'anglais). En théorie du codage, on suppose que la corrélation du signal de parole peut être exploitée pour approcher l'échantillon $x(n)$ comme combinaison linéaire des P échantillons qui le précèdent [5]. On définit ainsi le filtre prédicteur $w(n)$ tel que :

$$x(n) = \sum_{i=1}^{P} w(i)x(n-i) + e(n) \tag{3.1}$$

où $e(n)$ est appelé l'erreur de prediction. L'erreur de prédiction, encore appelée résiduel de prédiction, est un signal moins redondant que celui d'origine, son entropie est

donc moindre ce qui est intéressant du point de vue de la compression. C'est justement cette propriété qui est utilisée pour le codage et transmission de la parole où au lieu de transmettre le signal d'origine on transmet le filtre $w(n)$ et l'erreur de prédiction $e(n)$, qui sont suffisants pour reconstruire le signal complet à la réception.

À partir de la fonction d'autocorrélation d'ordre P de $x(n)$ et en utilisant (3.1) on établit les équations de Yule-Walker. Une démonstration complète de ces équations est donnée dans [5] ou [4]. En inversant ce système et notant $r(n)$ la fonction d'autocorrélation de $x(n)$, on obtient la relation :

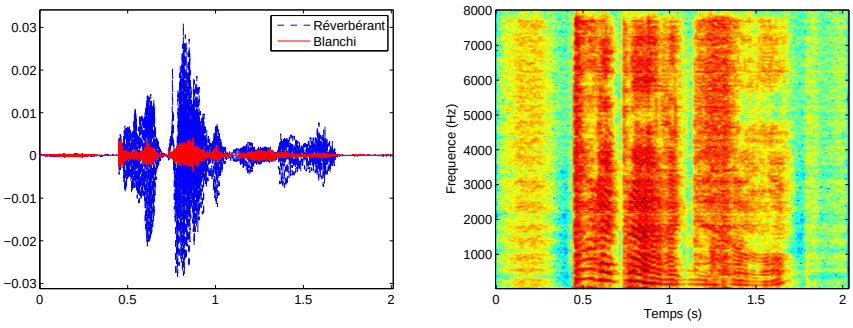
$$\mathbf{w} = R^{-1}\mathbf{r} \tag{3.2}$$

où $\mathbf{w} = [w(1), \cdots, w(P)]^{\mathrm{T}}$, $\mathbf{r} = [r(1), \cdots, r(P)]^{\mathrm{T}}$ et $R = toeplitz([r(0), \cdots, \mathrm{r(P\text{-}1)}])$. Dans la pratique, la matrice R est inversée avec l'algorithme récursif de Levinson Durbin [5] pour l'inversion rapide de matrices Toeplitz.

Le résiduel du modèle de prédiction linéaire du signal est donc décorrélé à court terme, il s'agit donc d'une version blanchie du signal d'origine. Or, en présence de réverbération, on observe encore des corrélations entre le signal direct et ses versions retardées. Le résiduel contient donc aussi l'effet de ces réflexions et son analyse permet d'estimer le signal de réverbération tardive.

Dans notre implémentation, le blanchiment de l'observation est réalisé sur une fenêtre glissante de durée 10ms, sans recouvrement. Pour chaque trame le filtre $w(n)$ d'ordre $P = 20$ est calculé et le signal résiduel est extrait. On construit ainsi le signal $x_{white}(n)$ de la forme :

$$x_{white}(n) = x_0(n) - \sum_{i=1}^{P} w(i)x_0(n-i) \tag{3.3}$$



(a) Signal réverbérant et signal blanchi (b) Spectrogramme du signal blanchi

FIGURE 3.3 – Forme d'onde et spectrogramme du signal blanchi

La figure 3.3 illustre un signal réverbérant, sa version blanchie et son spectrogramme. Le spectre de chaque trame du spectrogramme correspond bien à celui d'un bruit large bande, de puissance presque constante pour toutes les fréquences.

3.2 Prédiction linéaire à long terme

Pour extraire la réverbération tardive du signal blanchi, on introduit un modèle de Prédiction Linéaire à Long Terme (PLLT). On définit la PLLT par :

$$x_{white}(n) = \sum_{p=1}^{N} a(p)x_{white}(n-p-D) + e(n), \tag{3.4}$$

où D est le retard du modèle et N est l'ordre du modèle tel que $N \geq D$. Ce modèle est analogue à celui de prédiction linéaire présenté dans la section 3.1 à ceci près qu'on cherche à exploiter la corrélation à long terme du signal. Comme on souhaite prédire la réverbération tardive du signal, on choisit de travailler avec un retard $D = 30ms$ afin d'approcher les réflexions tardives que l'on veut atténuer. En partant de l'équation (3.4) on peut aisément établir les équations analogues à celles de Yule-Walker (voir Annexe A) pour le calcul du filtre $a(n)$ de la forme :

$$\mathbf{a} = R^{-1}\mathbf{r_D} \tag{3.5}$$

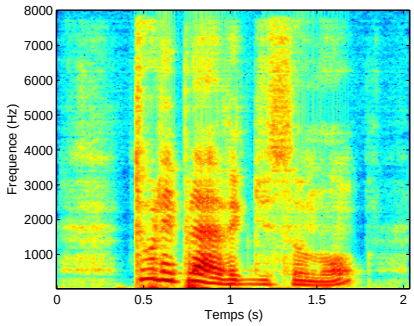
où $\mathbf{a} = [a(1), \cdots, a(N)]^{\mathrm{T}}$ et $R = toeplitz([r(0), \cdots, r(N-1)])$, sont définies de la même façon que pour le filtre blanchisseur, à partir de de la fonction d'autocorrélation $r(n)$ du signal. Le retard D de ce nouveau modèle intervient uniquement dans le terme $\mathbf{r_D} = [r(D+1), \cdots, r(D+N)]^{\mathrm{T}}$.

Le filtre $a(n)$ est un prédicteur de la réverbération tardive estimé à partir du signal blanchi $x_{white}(n)$. Dans l'implémentation de l'algorithme on choisit $N = 250ms$, $D = 30ms$, ce qui implique que le traitement pourra prédire correctement la réverbération tardive à condition que $RT_{60} \leq N$. La fonction d'autocorrélation est calculée sur $N + D$ échantillons et le système peut être inversé par l'algorithme de Levinson Durbin également. Notons tout de même que, pour des signaux échantillonnés à 16kHz, cela revient à inverser une matrice carrée de 4480 lignes, ce qui peut être très lourd en temps de calcul. Il existe cependant des algorithmes rapides d'inversion de matrices Toeplitz de taille importante qui seront ultérieurement étudiés.

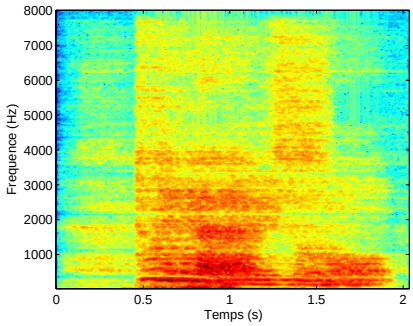
Une fois qu'on a calculé le prédicteur de la réverbération tardive, il suffit de convoluer le signal observé $x_0(n)$ avec le filtre $a(n)$ pour récupérer la partie réverbérante du signal dans $x_{late}(n) = (x_0 * a)(n)$. Dans la figure 3.4, on observe les spectrogrammes et les formes d'onde du signal original et de la réverbération tardive, respectivement.

3.3 Soustraction spectrale

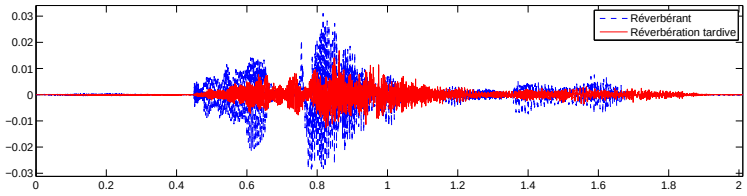
Une fois la réverbération tardive estimée, on peut procéder à son atténuation. Celle-ci se fait dans le domaine spectral, dans le but de limiter les distorsions de la parole



(a) Spectrogramme du signal réverbérant



(b) Spectrogramme de la réverbération tardive



(c) Formes d'onde des deux signaux

FIGURE 3.4 – Spectrogrammes et formes d'onde du signal original et de la réverbération tardive estimée

résultante. On calcule donc les spectres à court terme de $x_0(n)$ et de $x_{late}(n)$ par TFCT avec une fenêtre de Hamming de 512 points et un recouvrement de 75%. On les note $X_0(l,k) = |X_0(l,k)|e^{j\phi_0}$ et $X_{late}(l,k) = |X_{late}(l,k)|e^{j\phi_{late}}$, avec l l'indice des trames, k l'indice des canaux fréquentiels et $\phi_0(l,k)$ et $\phi_{late}(l,k)$ les phases de $X_0(l,k)$ et $X_{late}(l,k)$ respectivement. La fenêtre de Hamming est très utilisée en traitement de la parole car elle constitue un bon compromis entre la largeur de son lobe principal et l'attenuation de ses lobes secondaires.

La TFCT, notée $X(l,k)$, est donnée par :

$$\begin{aligned} TFCT[x(n)] = X(l,k) &= \sum_{n=-\infty}^{\infty} x(n-l)w(n)e^{-jkm} \\ &= \sum_{n=0}^{L-1} x(n-l)w(n)e^{-jkm} \end{aligned} \tag{3.6}$$

où $w(n)$ est une fenêtre de longueur finie L, et t et k désignent respectivement les indices temporels et fréquentiels de la transformée.

Ensuite, on réalise la réduction de la réverbération par soustraction spectrale pour construire $|X_1(l,k)|$ le spectrogramme de puissance du signal estimé. La soustraction spectrale est une technique classique pour la réduction du bruit de fond [6, 22]. Elle fournit l'amplitude spectrale du signal à débruiter à partir de l'estimateur au sens du

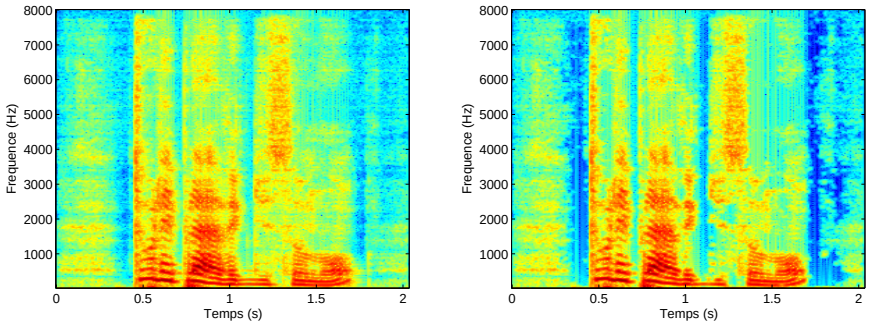
maximum de vraisemblance de la variance de chaque composante spectrale [8]. L'amplitude spectrale du signal déréverbéré est donnée par la relation suivante :

$$|X_1(l,k)| = \begin{cases} \sqrt{|X_0(l,k)|^2 - |X_{late}(l,k)|^2}, & \text{si } |X_0(l,k)|^2 - |X_{late}(l,k)|^2 \geq 0 \\ 0, & \text{sinon} \end{cases} \tag{3.7}$$

La phase du spectrogramme de $x_0(n)$, notée $\phi(X_0(l,k))$, n'est pas modifiée et est utilisée pour resynthétiser le spectrogramme du signal dépourvu de la réverbération tardive, tel que :

$$X_1(l,k) = |X_1(l,k)|e^{j\phi(X_0(l,k))} \tag{3.8}$$

Dans la figure 3.5 on observe les spectrogrammes du signal réverbérant et du signal après atténuation de la réverbération tardive.



(a) Spectrogramme du signal réverbérant ($RT_{60} = 200ms$)

(b) Spectrogramme après réduction de la réverbération tardive

FIGURE 3.5 – Spectrogrammes du signal original et après déréverbération tardive

3.4 Cepstral Mean Subtraction (CMS)

À ce stade, on a supprimé la réverbération tardive, il nous reste donc à traiter les réflexions précoces. Leur effet est atténué par analyse cepstrale. L'analyse cepstrale est un type de transformation homomorphique qui permet de transformer les mélanges convolutifs en simples mélanges additifs. Une transformation homomorphique est une projection d'un signal sur un domaine où son comportement est linéaire. Dans le cadre de notre étude, nous avons considéré le signal observé comme la convolution du signal anéchoïque par la réponse impulsionnelle de la salle : $x_0(n) = (s * h)(n)$. Les coefficients cepstraux de $x_0(n)$, encore appelés cepstres, sont calculés à partir du module de la transformée de Fourier du signal comme illustré dans la figure 3.6. La transformée de

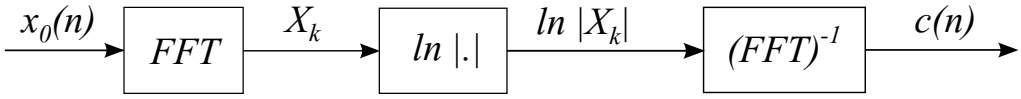


FIGURE 3.6 – Calcul du cepstre réel

Fourier est calculée par l'algorithme de Transformée de Fourier Rapide (Fast Fourier Transform, en anglais). Le diagramme (3.9) montre comment un produit de convolution, non linéaire par construction, est équivalent à une simple addition dans le domaine cepstral, ce qui explique l'interêt de passer dans ce domaine pour réduire l'influence du canal par simple soustraction cepstrale.

$$(x * h)(n) \xrightarrow{TF} X(k)H(k) \xrightarrow{TF^{-1}[\ln|.|]} c_x(n) + c_h(n) \tag{3.9}$$

Les cepstres donnent une information sur la vitesse de variation des composantes spectrales. La parole, localement stationnaire dans notre modèle, produit des cepstres non nuls proches de l'origine. Ces cepstres illustrent la réponse impulsionnelle du conduit vocal. Dans [30], Oppenheim montre que la présence d'echos importants introduit des fluctuations spectrales, qui se traduisent par des pics cepstraux éloignés de l'origine des temps. Si on annule ces cepstres là, on supprime les échos. Dans le cas de la réverbération tous les coefficients cepstraux sont modifiés. On ne peut donc pas annuler les cepstres affectés par la réverbération sans modifier les caractéristiques de la parole. Si on veut réduire l'effet de salle il vaut mieux soustraire la moyenne temporelle des cepstres. En effet, l'étude empirique de ces coefficients a permis de constater que la moyenne temporelle des cepstres d'un signal de parole est nulle. Le filtrage du signal par le canal, modifie cette moyenne. On suppose donc que la soustraction de la moyenne des cepstres permet de réduire l'influence du canal par déconvolution dans le domaine cepstral.

Dans la pratique, si on soustrait la moyenne cepstrale à l'intégralité des cepstres, on modifie aussi les premiers coefficients, ce qui provoque une modification du timbre de la voix lors de la resynthèse du signal. Cette modification est tolérable pour les systèmes de reconnaissance automatique de la parole mais n'est pas envisageable dans le cadre des télécommunications, où la qualité de la voix doit toujours être préservée. On doit donc s'assurer de ne pas modifier les premiers cepstres. Le problème consiste à déterminer l'instant à partir duquel la soustraction sera faite. En effet, il n'existe pas de preuve formelle permettant de fixer une frontière théorique entre les cepstres qui representent uniquement la voix et le reste. Cette frontière est donc choisie de façon heuristique à partir d'écoutes informelles, de façon à garder le seuil qui dégrade le moins le timbre de la parole tout en réduisant l'effet des réflexions précoces.

Ensuite on resynthétise le signal de parole en appliquant la procédure inverse que pour le calcul des cepstres, comme illustré à la figure 3.7. Afin de garder la cohérence temporelle du signal on utilise la phase de $x_{late}(n)$, qui n'a pas été modifiée.

La figure 3.8(a) montre le spectrogramme d'un signal anéchoïque de parole. On distingue bien les formants de la voix ainsi que les silences entre les phonèmes. Ce signal a été filtré par une réponse impulsionnelle artificielle avec un temps de réverbération de

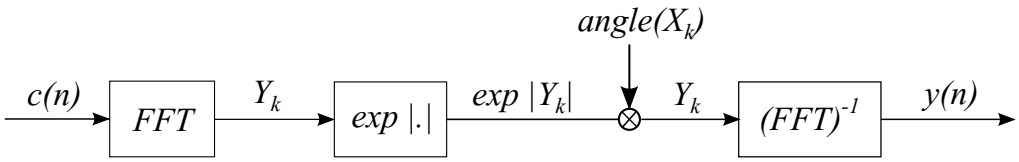
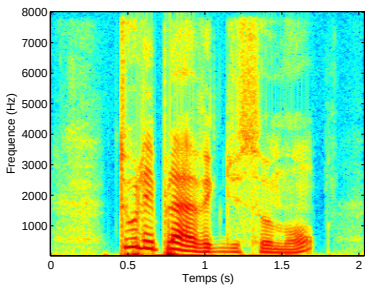
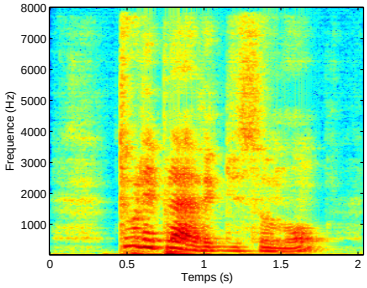


FIGURE 3.7 – Reconstruction du signal à partir des cepstres

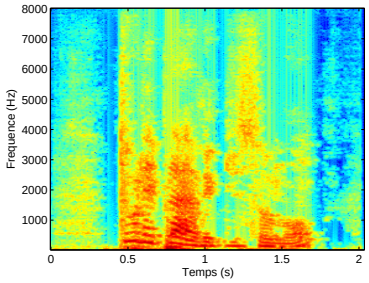
200ms. On observe son spectrogramme dans la figure 3.8(b). La structure des formants est dégradée par la réverbération et les silences sont masqués par la queue de la réverbération. À la fin du dernier phonème on constate bien l'étalement temporel induit par la réverbération. Enfin, la figure 3.8(c) montre le spectrogramme du signal estimé par la méthode de Kinoshita *et al.* L'énergie du signal a bien été atténuée dans les zones de silence et la structure des formants est moins dégradée que dans la version réverbarante.



(a) Spectrogramme du signal anéchoïque



(b) Spectrogramme du signal réverbérant ($RT_{60} = 200ms$)



(c) Spectrogramme du signal déréverbéré avec Soustraction Spectrale

FIGURE 3.8 – Spectrogrammes pour l'algorithme de Kinoshita

3.5 Propositions d'amélioration

L'algorithme décrit dans cette partie a été implémenté sous Matlab. Tous les paramètres ont été fixés aux valeurs décrites dans [20]. Ces valeurs ont été choisies pour optimiser les performances de l'algorithme dans le cadre de la reconnaissance automatique

de parole. L'apprentissage des valeurs est réalisé sur un corpus de 40000 énonciations issues de la base JNAS (Japanese Newspapers Article Sentences). Ces énonciations ont été traitées par la réponse impulsionnelle simulée d'une salle avec un temps de réverbération de 650ms. 200 énonciations sont mises à l'épreuve afin d'obtenir le taux d'erreurs de reconnaissance (Word Error Rate, en anglais) en fonction de la distance entre la source et le récepteur. Kinoshita observe une nette diminution du taux d'erreurs entre le signal non traité et sa version déréverbérée. On passe ainsi de 20% d'erreurs à seulement 8% pour une distance de 50cm, et de 67% à 25% pour une distance de 2 mètres.

N'ayant pas accès à la même base de données utilisée par les auteurs, nous ne pouvons pas valider les résultats de la déréverbération à partir des résultats présentés dans [20]. Dans le chapitre 5 nous décrirons les tests spécifiques mis en place pour valider et améliorer la méthode. Pour l'instant on se limite à une évaluation subjective de l'algorithme.

En comparant un signal anéchoïque à sa version déréverbérée on constate tout d'abord que la CMS introduit une modification du timbre de la parole. Cette coloration est réduite lorsqu'on ne modifie pas les premiers coefficients cepstraux, comme nous l'avons vu dans la partie 3.4. Nous avons donc évalué le résultat du traitement en faisant varier le nombre de coefficients cepstraux qui ne sont pas filtrés par la moyenne cepstrale. Ainsi, on a constaté que si la soustraction cepstrale est réalisé à partir du 21^{ieme} coefficient cepstral (pour des signaux échantillonnés à $16kHz$), le timbre de la parole n'est que légèrement affecté, tout en réduisant l'effet des réflexions précoces. Il convient, cependant, d'explorer d'autres techniques de réduction des réflexions précoces qui introduisent moins de dégradations du point de vue perceptuel. Cette problématique n'a pas été abordée pendant le stage.

Nous constatons également que la soustraction spectrale introduit des pics fréquentiels isolés dans le spectre à court terme du signal, ces pics se traduisent par du bruit musical. Cet effet indésirable dégrade la qualité perceptuelle de la parole. On souhaite donc utiliser d'autres règles de suppression de la réverbération tardive. Le filtrage de Wiener [37] est une méthode de filtrage dans le domaine spectral qui permet d'atténuer ces effets, à condition de disposer d'un modèle du signal parasite. Dans notre cas, on dispose bien d'une représentation de la réverbération tardive à supprimer. On calcule donc le masque de Wiener $M_{Wiener}(l,k)$ à partir des composantes spectrales $X_0(l,k)$ et $X_{late}(l,k)$ par :

$$M_{Wiener}(l,k) = \frac{|X_0(l,k)|^2}{|X_0(l,k)|^2 + |X_{late}(l,k)|^2} \tag{3.10}$$

Ce masque est appliqué au spectrogramme de l'observation pour construire le signal déréverbéré. Subjectivement, on entend moins d'artéfacts lorsque la règle de Wiener est utilisée pour l'atténuation.

Dans [8, 9], Ephraim et Malah construisent un estimateur de l'amplitude spectrale de signaux bruités qui limite l'apparition du bruit musical. En faisant intervenir un lissage temporel de l'estimateur d'amplitude, un masque $M_{EM}(k,l)$est construit tel que :

$$M_{EM}(l,k) = \frac{\sqrt{\pi \upsilon_k}}{\gamma_k} exp(-\frac{\upsilon_k}{2})[(1+\upsilon_k)I_0(\frac{\upsilon_k}{2}) + \upsilon_k I_1(\frac{\upsilon_k}{2})] \tag{3.11}$$

où $I_0(.)$ et $I_1(.)$ désignent les fonctions de Bessel modifiées d'ordre 0 et 1, respectivement. On définit υ_k par :

$$\upsilon_k \triangleq \frac{\xi_k}{1+\xi_k}\gamma_k \tag{3.12}$$

avec ξ_k et γ_k, les rapports signal à bruit *a priori* et *a posteriori* tel qu'ils ont été introduits dans [8].

Enfin, nous avons remarqué que l'implémentation actuelle ne fonctionne que si le temps de réverbération est de l'ordre de grandeur de la longueur du filtre prédicteur. Si le temps de réverbération est plus important, l'atténuation de la réverbération tardive est moins performante. On souhaite donc automatiser le choix de la taille du modèle de prédiction en ajoutant un bloc d'estimation du RT_{60} en amont du traitement. Cette adaptation sera réalisée ultérieurement dans nos travaux de recherche.

Chapitre 4

Déréverbération par déconvolution de spectrogramme

Dans [19], Kameoka *et al.* présentent une méthode de déconvolution aveugle pour effectuer la déréverbération. La déréverbération est réalisée dans le domaine spectral et on vise à estimer l'enveloppe de puissance de chaque sous-bande du spectrogramme du signal. En supposant la non-négativité des enveloppes de puissance et en partant d'une réponse impulsionnelle *a priori*, on fait un filtrage temporel de l'enveloppe du signal observé. Ce traitement exploite en plus la parcimonie du spectrogramme de la parole, il s'agit donc d'une façon de m'introduire aux notions de parcimonie qui seront probablement utilisées ultérieurement pour le doctorat.

La méthode se base sur une méthode performante de séparation de sources, la factorisation en matrices non-négatives (Non-Negative Matrix Factorization, en anglais). La Non-Negative Matrix Factorization (NMF)[23] consiste en la décomposition des spectrogrammes, vus comme des matrice de rang élevé, comme un produit de deux matrices de rang inférieur. À partir d'une matrice non-négative $\mathbf{V}$ de taille $M \times N$, on veut trouver une matrice $\mathbf{W}$, de taille $M \times R$, et une matrice $\mathbf{H}$, de taille $R \times N$, telles que $\mathbf{V} \approx \mathbf{W} \cdot \mathbf{H}$. En plus, $\mathbf{W}$ et $\mathbf{H}$ doivent également être non négatives. On dit qu'une matrice est non-négative si tous ses coefficients sont positifs ou nuls.

Dans le cas des spectrogrammes audio, chaque colonne de $\mathbf{W}$ est interprétée comme une fonction de base qui définit les fréquences du spectrogramme qui sont effectivement présentes dans le signal. D'autre part, chaque ligne de $\mathbf{H}$ correspond au poids affecté à chaque composante spectrale présente dans $\mathbf{W}$. On dit que $\mathbf{H}$ définit les instants d'activation des fonctions de base. Cette décomposition permet de faire une séparation de sources robuste pour les signaux musicaux. En effet, chaque note peut être décrite par les fréquences qui la forment et par l'instant auquel elle est jouée. Ces deux informations apparaissent dans $\mathbf{W}$ et $\mathbf{H}$, respectivement.

Cette technique est ensuite adaptée dans [34] pour tenir compte de l'évolution temporelle des fonctions de base décrites dans $\mathbf{W}$, on parle alors d'un problème de déconvolution par factorisation de matrices non négatives, encore appelé Non-Negative Matrix

Factor Deconvolution (NMFD), en anglais. La NMFD, généralise la NMF dans le cas des mélanges convolutifs. Un cas particulier de cette méthode est utilisé par Kameoka *et al* pour aborder le problème de la déréverbération. La technique présentée ici vise donc à décomposer le spectrogramme de puissance du signal réverbérant, comme la convolution du spectrogramme du signal anéchoïque et de celui de la réponse impulsionnelle de la salle.

4.1 Modèle en sous-bandes du signal réverbérant

Le signal $x(n)$, observé au niveau du microphone, est décomposé en sous-bandes par Transformée de Fourier à Court Terme. La TFCT donne une répresentation temps-fréquence de $x(n)$, notée $X(t,k)$. Elle a été définie dans 3.3. Si on fixe $t = t_0$, $X(t_0, k)$ est le spectre d'une trame de longueur L du signal. Si, au contraire, on fixe $k = k_0$, $X(t, k_0)$ correspond à un signal temporel. Il correspond à la sous-bande k_0 du signal observé. S'agissant d'un signal temporel, on note $x_k(t)$ le signal de la k^{ieme} sous-bande. La figure 4.1 montre le signal original et trois de ses sous-bandes. La méthode de choisie consiste donc à traiter séparément chaque sous-bande de l'observation pour estimer chaque sous-bande du signal anéchoïque.

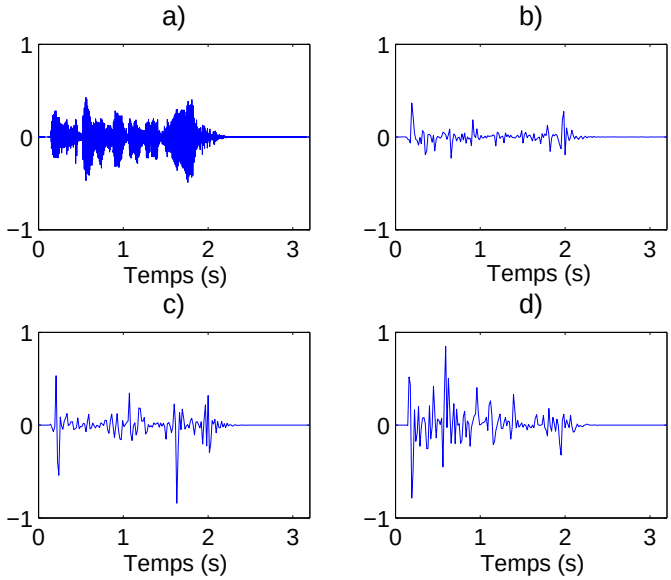


FIGURE 4.1 – Décomposition en sous bandes d'un signal.
a) Signal original
b), c) et d) Trois sous-bandes du signal

On suppose donc que $x_k(t)$ peut s'exprimer comme la convolution de la k^{ieme} sous-bande du signal anéchoïque, $s_k(t)$ et de celle de la réponse impulsionnelle de la salle,

$h_k(t)$, soit :

$$x_k[t] = \sum_{\tau=0}^{\infty} s_k[\tau] h_k[t-\tau] \tag{4.1}$$

Partant de (4.1) et en notant que $h_k[t] = |h_k[t]|e^{i\phi_k[t]}$, on exprime l'enveloppe de puissance de la k^{ieme} sous-bande comme :

$$|x_k[t]|^2 = \sum_{\tau,\tau'} s_k^*[t-\tau] s_k[t-\tau'] |h_k[\tau]||h_k[\tau']| e^{-i\phi_k[\tau]} e^{i\phi_k[\tau']} \tag{4.2}$$

La phase, $\phi_k[t]$, de la réponse impulsionnelle est affectée par les déplacements du locuteur dans l'espace réverbérant. Pour atténuer l'effet de la variabilité de la réponse impulsionnelle, on considère que sa phase est une variable aléatoire distribuée uniformement sur l'intervalle $[-\pi, \pi]$. Pour nous affranchir de ce terme, on travaille sur l'espérance de l'obervation. Ainsi, en prenant l'espérance de l'équation (4.2), on aboutit au modèle :

$$\mathbb{E}[|x_k[t]|^2] = \sum_{\tau} |s_k[t-\tau]|^2 |h_k[\tau]|^2 \tag{4.3}$$

Ce modèle suppose donc que, dans le sens de l'espérance mathématique, l'enveloppe de puissance du signal observé est obtenue comme la convolution en sous-bandes de l'enveloppe de signal anéchoïque et de celle de la réponse impulsionnelle de la salle. On aborde donc le problème comme une déconvolution aveugle des enveloppes de puissance des sous-bandes du signal.

4.2 Formulation du problème

Pour alléger la notation, nous posons $S_k[t] \triangleq |s_k[t]|^2$ et $H_k[t] \triangleq |h_k[t]|^2$. On définit alors le modèle de l'enveloppe de la k^{ieme} sous-bande de l'observation comme :

$$X_k[t] \triangleq \sum_{\tau} S_k[\tau] H_k[t-\tau] \tag{4.4}$$

En notant $Y_k[t]$ l'enveloppe de la k^{ieme} sous-bande du signal observé au niveau du microphone, le problème de convolution consiste à construire une enveloppe $X_k[t]$, telle que $Y_k[t] \simeq X_k[t]$. On introduit alors l'erreur de reconstruction, $\epsilon_k[t]$, telle que :

$$Y_k[t] = X_k[t] + \epsilon_k[t] \tag{4.5}$$

On suppose que $\epsilon_k[t]$ est un bruit blanc gaussien centré et de variance σ^2 que l'on souhaite minimiser dans notre estimation de l'enveloppe de la source, $S_k[t]$. On suppose également que le spectrogramme de puissance S, que l'on souhaite estimer, est parcimonieux puisqu'il représente un signal anéchoïque de parole. Ainsi, on considère que la distribution du spectrogramme S s'écrit :

$$P(S) = \prod_{k,t} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|S_k[t]|^p}{b^p}\right) \tag{4.6}$$

où b et p sont des paramètres qui permettent de jauger le degré de parcimonie du signal estimé.

En régularisant le problème de minimisation de $\epsilon_k[t]$ par la parcimonie du spectrogramme S, on peut le résumer à la minimisation de la fonction de coût suivante :

$$f(S,H) = \sum_{k,t} (Y_k[t] - X_k[t])^2 + 2\lambda \sum_{k,t} |S_k[t]|^p \tag{4.7}$$

où λ permet de déterminer l'importance de la parcimonie dans la résolution du problème. En effet, le deuxième terme de (4.7) fait augmenter la valeur de la fonction $f(S,H)$ puisque $\lambda > 0$ et $|S_k[t]| > 0$. Pour minimiser la fonction, on doit donc minimiser le deuxième terme et donc favoriser les solutions S avec un nombre important de coefficients nuls, les solutions parcimonieuses. Ce terme permet, en plus, de s'assurer que l'algorithme converge vers une solution non triviale. Ce problème est contraint par la non-négativité des enveloppes des sous-bandes. On suppose, en plus, que l'énergie de la réponse impulsionnelle est toujours égale à 1, ce qui nous assure de trouver une unique solution au problème.

On vise donc à résoudre le problème suivant :

$$\begin{array}{rc} \text{minimiser} & f(S,H) \text{ par rapport à } S \text{ et } H \\ \text{contraint par} & \sum_t H_k[t] = 1,\ H_k[t] \geq 0,\ S_k[t] \geq 0,\ \forall (k,t) \end{array} \tag{4.8}$$

4.3 Résolution

Pour résoudre le problème posé en (4.8) de façon efficace on souhaite mettre en place un algorithme itératif du même type que ceux utilisés dans le cadre de la NMF. L'algorithme de factorisation de matrices proposé dans [23] consiste à faire évoluer, à chaque itération, les matrices qu'on souhaite extraire. Pour ce faire, on extrait des règles multiplicatives de mise à jour qui permettent d'avoir une décroissance de la fonction de coût, que l'on minimise pour estimer les matrices $\mathbf{H}$ et $\mathbf{W}$. Ainsi, on peut démontrer

qu'après un nombre fini d'itérations l'algorithme converge vers la solution voulue. Le fait d'utiliser des règles multiplicatives de mise à jour des matrices, permet d'avoir des implémentations rapides de l'algorithme, ce qui peut être intéressant dans notre contexte d'application.

Dans [19], les auteurs proposent des règles de mise à jour pour les matrices S et H qui sont cohérentes avec le problème (4.8). La mise à jour de S se fait en partant de l'inégalité :

$$\begin{aligned} f(S,H') \leq \sum_{k,t,\tau} \frac{S'_k[\tau]H'_k[t-\tau]}{X'_k[t]} \left(Y_k[t] - \frac{S_k[\tau]}{S'_k[\tau]} X'_k[t]\right)^2 \\ +2\lambda \sum_{k,t} (pS'_k[t]^{p-2} s_k[t]^2 + 2|S'_k[t]|^p - p|S'_k[t]|^p) \end{aligned} \tag{4.9}$$

où $H'_k[t]$ et $S'_k[t]$ sont les valeurs respectives des paramètres à l'itération précédente. En dérivant le deuxième membre de l'équation (4.9) par rapport à $S'_k[t]$ et en annulant la dérivée on définit la règle de mise à jour de S par :

$$S_k[\tau] = S'_k[\tau] \frac{\sum_t H'_k[t-\tau]Y_k[t]}{\sum_t H'_k[t-\tau]X'_k[t] + \lambda p|S'_k[\tau]|^{p-1}} \tag{4.10}$$

avec $X'_k[t] = \sum_\tau S'_k[\tau]H'_k[t-\tau]$. Tous les termes de l'équation (4.10) sont positifs ce qui garantit la non-négativité de $S_k[t]$ lors de la mise à jour. De la même manière en partant de :

$$\begin{aligned} f(S',H) \leq \sum_{k,t,\tau} \frac{S'_k[t-\tau]H'_k[\tau]}{X'_k[t]} \left(Y_k[t] - \frac{H_k[\tau]}{H'_k[\tau]} X'_k[t]\right)^2 \\ +2\lambda \sum_{k,t} |S'_k[t]|^p, \end{aligned} \tag{4.11}$$

et en dérivant le deuxième membre par rapport à $H_k[t]$, on définit une règle de mise à jour pour H de la forme :

$$H_k[\tau] = H'_k[\tau] \frac{\sum_t S'_k[t-\tau]Y_k[t]}{\sum_t S'_k[t-\tau]X'_k[t]}. \tag{4.12}$$

Là aussi, tous les termes de la règle de mise à jour sont positifs, la non-négativité de $H_k[t]$ est donc assurée. On constate que λ, le poids de la parcimonie, intervient uniquement dans la mise à jour de $S_k[t]$. Ceci était attendu puisque c'est bien ce spectrogramme que l'on veut rendre parcimonieux. Enfin, on remarque que pour l'instant la contrainte $\sum_t H_k[t] = 1$ n'a pas été utilisée. On choisit donc de la faire intervenir après avoir effectué la mise à jour de S et H, en normalisant $H_k[t]$ par $\sum_t H_k[t]$.

4.4 Implémentation et optimisations

L'algorithme de déconvolution proposé dans [19] consiste donc à faire évoluer les matrices S et H à partir d'un état initial , jusqu'à converger vers une solution S^*, qui approche le spectrogramme du signal anéchoïque de la source.

Ainsi, on choisit d'initialiser la matrice S par le spectrogramme Y du signal observé. Ce spectrogramme est obtenu en calculant la TFCT du signal sur des trames de 64ms avec un recouvrement de 50%. Chaque trame est pondérée par une fenêtre de Hanning. La matrice H est initialisée avec une exponentielle décroissante pour chaque canal fréquentiel de la forme :

$$H_k[t] = e^{-\frac{t}{\delta}}, \quad \forall k \tag{4.13}$$

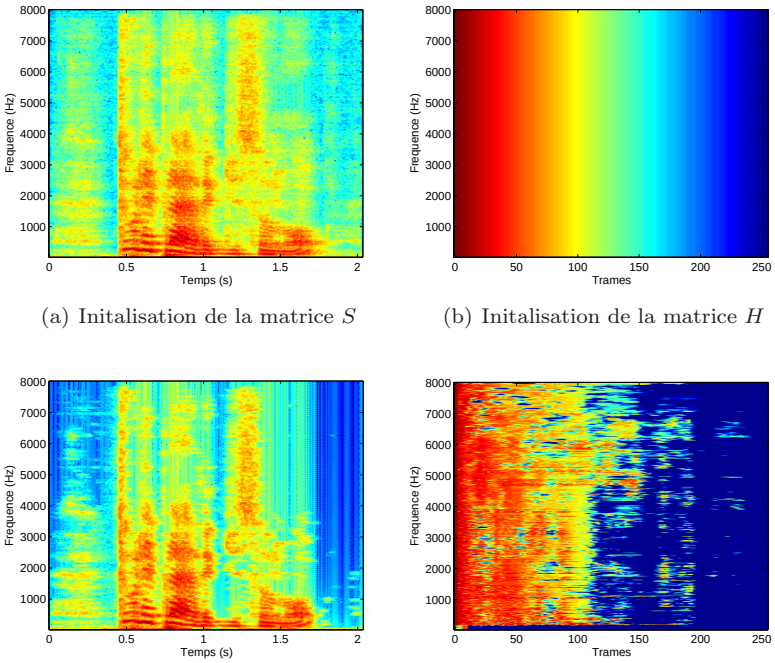
où δ permet de calibrer la vitesse de décroissance de la réponse impulsionnelle pour s'adapter au temps de réverbération de la salle. Ceci est cohérent avec le modèle de la réponse impulsionnelle d'une salle présenté par Polack [31] et introduit dans le chapitre 2.

Nous avons implémenté cette méthode sous Matlab, en utilisant les mêmes réglages que ceux proposés dans [19]. Ainsi, nous fixons $p = 1.2$ et $\lambda = E^{2-p}$ avec $E = \sum_{k,t} Y_k[t] \times 10^{-8}$. En rappelant que $Y_k[t] = |y_k[t]|^2$, on constate que λ dépend l'énergie totale du signal à déréverbérer.

À partir de l'état initial, les matrices S et H sont mises à jour à chaque itération. La condition d'arrêt de l'algorithme est donnée par le nombre d'itérations, fixé à $N = 20$.

Dans [19], l'algorithme est évalué en termes du SNR, que ce soit dans des conditions réelles de réverbération ou en utilisant des réponses impulsionnelles artificielles. Dans le cas d'une réponse impulsionnelle simulée, telle que $RT_{60} = 500ms$, le SNR passe de 2.30dB, pour le signal réverbérant, à 2.94dB. Cette augmentation n'est pas très importante mais, d'un point de vue subjectif la, déréverbération est bien réussie par cette méthode. Dans le chapitre 5, nous présenterons les résultats de l'évaluation de cette technique en utilisant d'autres mesures objectives parmi celles présentées dans le chapitre 2.

La figure 4.2 montre l'initialisation de l'algorithme et le résultat après convergence. À gauche, on observe le spectrogramme du signal réverbérant et celui du signal estimé. L'énergie du signal dans les zones de silence a bien été atténuée et les formants ont également été rehaussés. Ceci réduit l'effet de la réverbération tardive. À droite, on represente l'état initial et l'état final de la matrice de la réponse impulsionnelle en sous bandes du canal. À l'état initial, tous les canaux sont affectés à la même exponentielle décroissante. À la convergence, cette matrice est cohérente avec celle d'une réponse impulsionnelle. Les premiers coefficients sont prédominants et les derniers sont très faibles, ce qui donne une idée du temps de réverbération de la réponse. On observe également que la durée de la réponse diffère d'un canal à l'autre, ce qui n'a pas été pris en compte dans le modèle utilisé.



(a) Initalisation de la matrice S

(b) Initalisation de la matrice H

(c) Matrice S estimée

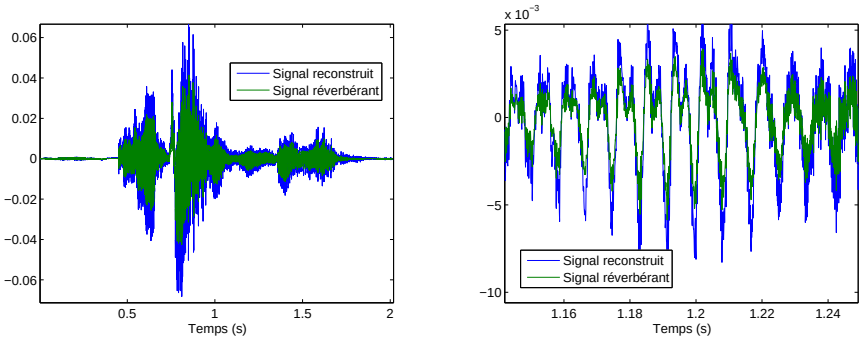
(d) Matrice H estimée

FIGURE 4.2 – État initial et état final des matrices S et H

On a vérifié la qualité de la déconvolution on reconstituant un signal par convolution rapide des matrices S et H. On illustre le résultat dans la figure 4.3. Les signaux obtenus sont très proches. On constate cependant que le signal reconstruit a une amplitude plus importante que le signal réverbérant. Une étude plus approffondie devre être menée pour trouver l'origine de ce gain.

Cette technique peut être aisément optimisée pour l'adapter à un fonctionnement en temps réel. On constate d'abord que la mise à jour des matrices S et H, fait intervenir des produits de convolution sur toute la durée du segment analysé. Dans [19], cette mise à jour est réalisée par convolution rapide dans le domaine spectral. La convolution rapide exploite le fait que les produits de convolution dans le domaine temporel sont équivalents à des simples produits, dans le domaine spectral. Si on s'assure de prendre des signaux $x(n)$ et $y(n)$, dont la longueur N est une puissance de 2, on peut calculer leur transformée de Fourier en utilisant la Fast Fourier Transform (FFT), comme l'illustre la figure 4.4. Ensuite on fait le produit terme à terme des signaux à convoluer avant de calculer la transformée de Fourier inverse. Cette procédure calcule la convolution circulaire des deux signaux. C'est à dire la convolution entre $x(n)$ et une version périodisée de $y(n)$. Ceci introduit des effets de bord, indésirables pour le calcul de la convolution classique. Pour éviter ce phénomène, la FFT doit être calculée sur au moins $2N$ points, ainsi que la transformée inverse. Pour récupérer le résultat de la convolution on tronque le signal résultant pour ne garder que les N premier termes, qui correspondent au résultat recherché.

Pour l'adapter à un fonctionnement en temps réel, l'algorithme doit, en plus, être



(a) Formes d'onde (b) Agrandissement des formes d'onde

FIGURE 4.3 – Signal réverbérant et signal reconstruit par convolution des matrices résultants

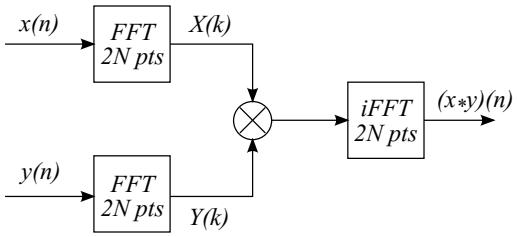


FIGURE 4.4 – Calcul de la convolution rapide

rendu causal. En effet, la méthode qu'on a présenté nécessite de connaitre l'intégralité du signal à déconvoluer, avant de commencer le traitement.

Une première analyse des matrices résultantes de cette méthode de déréverbération nous permet de constater que les réponses impulsionnelles estimées ne suivent pas toujours le modèle de décroissance exponentielle. On se propose donc d'améliorer les résultats en faisant intervenir une contrainte supplémentaire dans les équations de mise à jour de H, qui permet de privilégier les solutions présentant une décroissance exponentielle. Ceci, sera réalisé plus tard, dans la suite du doctorat.

D'autre part, l'initialisation de la matrice H proposée par les auteurs suppose que la réponse impulsionnelle a le même temps de décroissance dans chaque sous-bande. Or, on sait que les hautes fréquences sont atténuées plus rapidement que les basses fréquences. En introduisant cet a priori dans notre modèle, on peut raffiner l'estimation de la réponse impulsionnelle en sous bandes et on espère pouvoir améliorer la procédure de déconvolution.

Chapitre 5

Évaluation

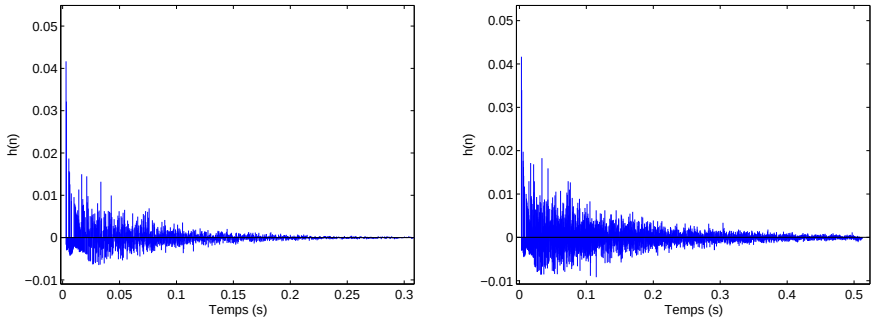
5.1 Protocole expérimental

On souhaite évaluer les deux algorithmes présentés dans le but de quantifier l'effet de la déréverbération. On utilisera donc un certain nombre de mesures objectives, parmi celles présentées dans le chapitre 2. Dans cette étude, on se limitera à mesurer l'efficacité des méthodes en fonction du temps de réverbération de la réponse impulsionnelle (RT_{60}). Nous expliquons dans la suite le protocole mis en place pour réaliser cette évaluation.

Comme on l'a déjà vu dans le chapitre 2, la plupart des métriques de déréverbération sont des mesures intrusives qui nécessitent la connaissance du signal anéchoïque que l'on souhaite estimer. On a donc construit une base de données composée de 15 énonciations, enregistrées dans une salle anéchoïque. Chaque énonciation a été segmentée manuellement de façon à avoir toujours du silence au début et à la fin du fichier (fig :forme donde). La durée des énonciations est de 3 secondes au plus et leur fréquence d'échantillonnage vaut 16kHz. Parmi les 15 échantillons, 13 correspondent à des voix masculines et 2 à des voix féminines. Deux énonciations sont en anglais et le reste en français.

Pour l'évaluation objective des algorithmes, il est pratique de connaitre le temps de réverbération de la salle dans laquelle on travaille. À moins de disposer d'une salle à acoustique variable, il est difficile de faire varier le RT_{60}. On décide donc de travailler sur des réponses impulsionnelles de salle simulées par la méthode de l'image d'Allen [1]. La méthode de l'image est un algorithme permettant de calculer la réponse impulsionnelle d'une salle parallélépipédique. En connaissant les dimensions du parallélépipède, la position de la source et du microphone et les coefficients d'absorption de chaque paroi, on simule tous les rayons acoustiques pouvant atteindre le microphone. Il s'agit d'une méthode d'acoustique géométrique. On a vu dans le chapitre 2 que la loi de Sabine donne le temps de réverbération en fonction des grandeurs caractéristiques de la salle. En inversant cette relation, on peut exprimer l'absorption moyenne de la salle en fonction du RT_{60}, soit :

$$a = \frac{4\ln 10^6}{c} \frac{V}{S.RT_{60}} \tag{5.1}$$



(a) $RT_{60} = 225ms$ (b) $RT_{60} = 500ms$

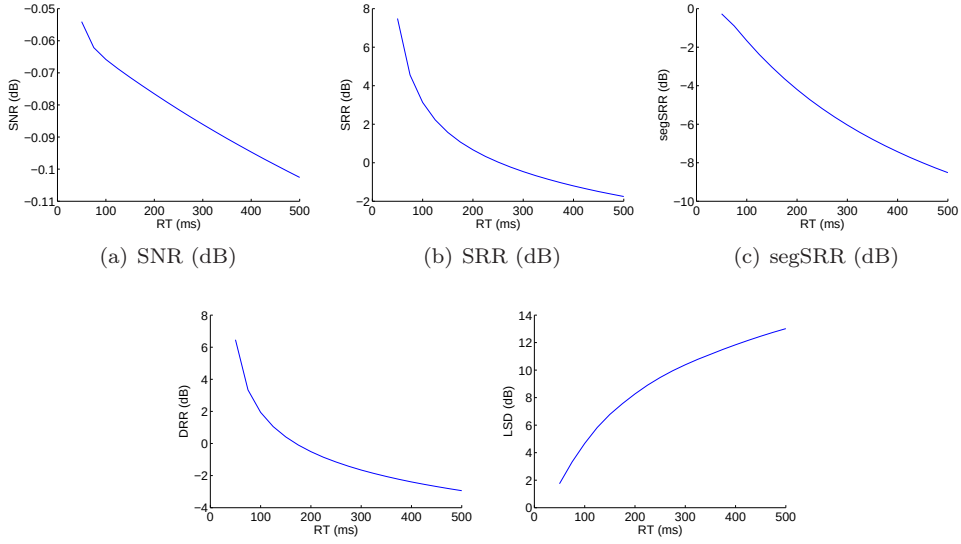
FIGURE 5.1 – Réponses impulsionnelles simulées par la méthode de l'image

De cette manière nous pouvons simuler des réponses impulsionnelles en calculant l'absorption moyenne de la salle à partir du RT_{60} souhaité. Nous avons donc généré des réponses impulsionnelles pour une salle de dimensions 2m x 1.5m x 1m, avec un locuteur situé aux coordonnées 0.7m x 0.3m x 0.7m et un microphone situé à 0.3m x 0.5m x 0.95m. Ces dimensions sont une approximation grossière de celles de l'habitacle d'une voiture, supposé parallélépipédique. Les positions de la source et du microphone sont cohérentes avec l'agencement classique à l'intérieur d'une voiture.

Nous avons ainsi généré 19 réponses impulsionnelles avec des temps de réverbération allant de 50ms à 500ms, par pas de 25ms. Chaque réponse impulsionnelle a une longueur de 4096 échantillons. La figure 5.1 montre des exemples des réponses impulsionnelles générées.

Chaque énonciation a été convoluée par chacune des réponses impulsionnelles et le signal réverbérant ainsi produit est traité par les deux algorithmes. Pour l'algorithme de prédiction linéaire présenté dans le chapitre 3, nous avons testé la version présentée dans [20] et deux versions modifiées dans lesquelles l'atténuation de la réverbération se fait par le filtrage de Wiener et par le filtrage d'Ephraim et Malah, au lieu d'effectuer la soustraction spectrale. Chaque fichier sera donc traité par 4 algorithmes différents.

Les résultats sont ensuite évalués avec 5 métriques différentes : le Signal to Noise Ratio (SNR), le Signal to Reverberant Ratio (SRR), le Segmental Signal to Reverberant Ratio (segSRR), le Direct to Reverberant Ratio (DRR) et la Log Spectral Distorsion (LSD). Toutes ces métriques ont été introduites dans la section 2.3. Aucune des métriques choisies ne fait intervenir des propriétés perceptuelles, on se concentre uniquement sur l'évaluation objective des algorithmes. La figure 5.2 montre l'évolution de chacune de ces métriques en fonction du temps de réverbération. La LSD est la seule métrique qui croit avec le temps de réverbération, ce qui est normal puisque plus la réverbération est importante, plus on introduit de la distorsion spectrale. Les autres métriques ont une évolution décroissante, cela traduit le fait que le signal anéchoïque est de plus en plus noyé dans la réverbération.



(a) SNR (dB) (b) SRR (dB) (c) segSRR (dB)

(d) DRR (dB) (e) LSD (dB)

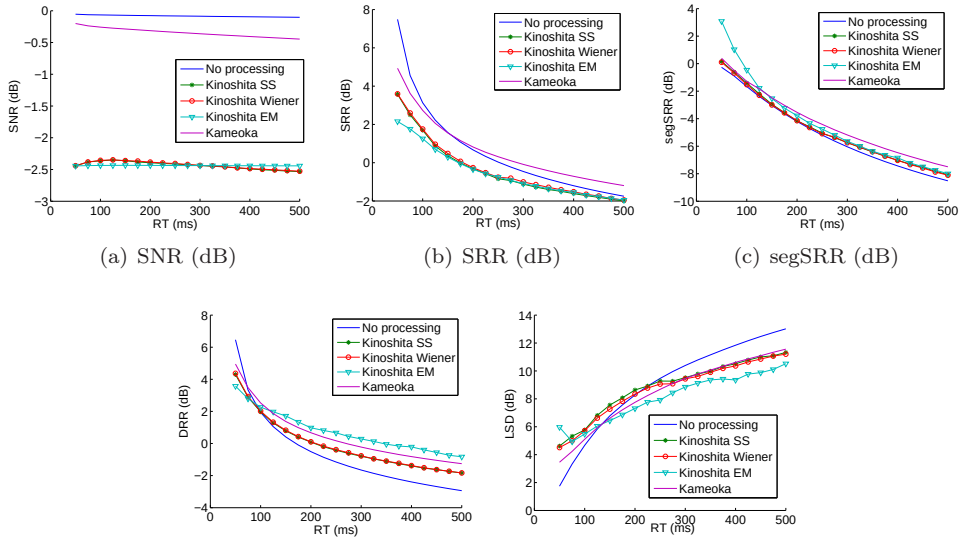
FIGURE 5.2 – Évolution des métriques en fonction du temps de réverbération

5.2 Résultats

5.2.1 Évaluation objective

Cette évaluation a un double objectif. D'une part, on veut exprimer une mesure de la déréverbération effectuée et, d'autre part, on souhaite étudier le comportement de chaque métrique pour justifier leur pertinence dans l'évaluation des algorithmes mis à l'épreuve. Pour chaque algorithme et pour chaque métrique, on prend la moyenne des mesures sur l'ensemble des énonciations pour observer leur performance en fonction du temps de réverbération. La figure 5.3 présente l'ensemble des résultats obtenus. On observe les métriques pour le signal réverbérant (bleu), pour le signal traité avec la méthode de Kinoshita avec soustraction spectrale (vert), avec filtrage de Wiener (rouge), et avec filtrage d'Ephraim et Malah (cyan), et pour le signal traité avec la méthode de Kameoka (violet).

- **Évolution du SNR**
  L'évolution du SNR avec le temps de réverbération est illustrée dans la figure 5.3(a). On observe que le SNR du signal réverbérant non traité ne varie presque pas, malgré une présence croissante de réverbération. Le même phénomène a lieu pour les signaux déréverbérés. Les trois variantes de l'algorithme de Kinoshita présentent des performances similaires et inférieures de 2.5dB par rapport au signal non traité. Ceci n'est pas cohérent avec le résultat perçu, où la déréverbération a effectivement lieu. L'algorithme de Kameoka donne des meilleurs résultats dans le sens du SNR. Celui-ci reste, cependant, inférieur à celui du signal non traité. En plus, quelque soit l'algorithme, le SNR dépend très peu du RT_{60}. Ceci montre qu'il ne s'agit pas d'une métrique adaptée à la mesure de la réverbération.



(a) SNR (dB) (b) SRR (dB) (c) segSRR (dB)

(d) DRR (dB) (e) LSD (dB)

FIGURE 5.3 – Résultats de la déréverbération en fonction du RT_{60}

- **Évolution du SRR**
  Les courbes correspondant au SRR sont présentées dans la figure 5.3(b). La décroissance de toutes les courbes est cohérente avec l'augmentation du temps de réverbération : plus la réverbération est longue, plus le signal direct est masqué par celle-ci. L'algorithme de Kinoshita présente des rapports signal à réverbérant inférieurs à ceux du signal non traité. Les trois variantes de cet algorithme donnent des résultats presque identiques ce qui montre l'incapacité de cette métrique pour évaluer la présence d'artéfacts dans le signal. L'algorithme de Kameoka, d'autre part, présente un meilleur SRR que le signal non traité à partir de $RT_{60} = 150ms$. Ceci correspond à la durée de décroissance de la réponse impulsionnelle choisie pour l'initialisation de l'algorithme. On pourrait penser qu'à partir d'un certain temps de réverbération l'algorithme réussit toujours à rehausser la parole. Une étude précise sur ce comportement devra être menée dans la suite des travaux de recherche.

- **Évolution du segSRR**
  La figure 5.3(c) montre les courbes du segSRR des signaux. Là tous les algotithmes améliorent la métrique de déréverbération. L'algorithme de Kameoka présente des résultats légérement meilleurs (de l'ordre de 0.5dB) que les autres. Cependant, la méthode de Kinoshita qui utilise le filtre d'Ephraim et Malah se démarque du reste pour les faibles RT_{60}.

- **Évolution du DRR**
  En ce qui concerne le DRR, les quatre algorithmes améliorent la note par rapport au signal non traité, sauf pour les faibles valeurs du RT_{60}. Pour cette métrique, illustrée dans la figure 5.3(d), c'est la méthode de Kinoshita avec filtrage d'Ephraim

et Malah qui présente les meilleurs résultats. L'algorithme de Kameoka produit également des bons résultats. Ces deux méthodes font augmenter le DRR d'environ 2dB alors que les deux autres techniques n'induisent qu'une augmentation de 1dB.

- **Évolution de la LSD**
  Enfin, la LSD est présentée dans la figure 5.3(e). Encore une fois, le filtrage d'Ephraim et Malah donne les meilleurs résultats, avec une dimution de l'ordre de 2dB par rapport au signal non traité. Les trois autres traitements tendent vers une même valeur du LSD quand RT_{60} augmente. On observe que tous les algorithmes présentent des résultats supérieurs au LSD de référence, sauf pour les RT_{60} faibles.

5.2.2 Évaluation Subjective

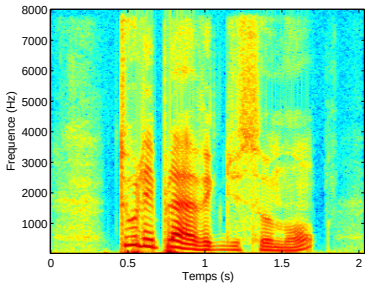
En plus de ces mesures, nous avons comparé les spectrogrammes des signaux obtenus par les quatre algorithmes. Un exemple de ces spectrogrammes est donné dans la figure 5.4. On observe que dans tous les cas, la queue de la réverbération a été atténuée par rapport à celle du signal réverbérant de la figure 5.4(c). Cependant, en aucun cas on atteint le spectrogramme du signal anéchoïque de la figure 5.4(a)). Les formants de la parole ne sont pas entièrement reconstruits et on observe encore de l'étalement temporel.

On a également réalisé des écoutes informelles pour réaliser une évaluation perceptuelle préliminaire. Dans tous les cas, on ressent un raccourcissement de la réverbération. Cependant, il est très difficile de percevoir des améliorations subtiles, par exemple dans le cas des RT_{60} faibles. On remarque aussi que, même si la réverbération est réduite, il existe toujours des différences perceptibles entre le signal rehaussé et le signal anéchoïque.

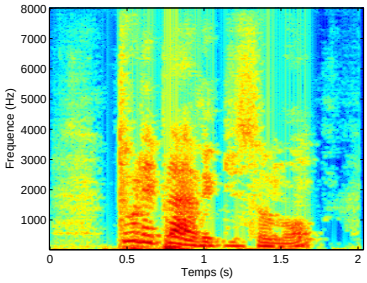
Comme on l'a remarqué dans le chapitre 3, le timbre de la parole est modifié par la soustraction cepstrale. À cause de cela, il est plus difficile d'apprécier de façon subjective si la réverbération a effectivement été réduite ou si c'est le changement de timbre qui donne cette sensation.

5.2.3 Commentaires

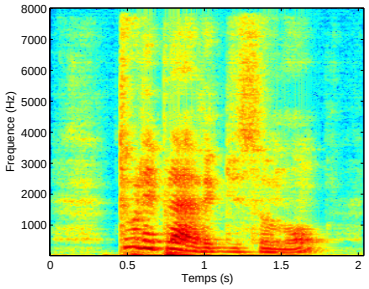
L'évaluation menée sur l'ensemble de la base de données nous a permis de confirmer que les deux méthodes présentées réussissent bien la tâche de déréverbération. L'effet de la déréverbération est d'autant plus perceptible que la réverbération est longue. Cependant, pour les temps de réverbération trop importants, on perçoit encore de la réverbération résiduelle dans le signal estimé. En plus, des améliorations restent à faire pour réduire la présence d'artefacts et pour préserver la qualité de la parole dans le signal rehaussé. Les modifications introduites dans l'algorithme de Kinoshita, permettent de réduire les artéfacts inhérents au traitement. Cependant, pour la plupart des métriques testées, aucune différence n'est mesurée entre le filtrage de Wiener et la soustraction spectrale. Le filtrage d'Ephraim et Malah, lui, permet d'améliorer légérement la note de l'algorithme.



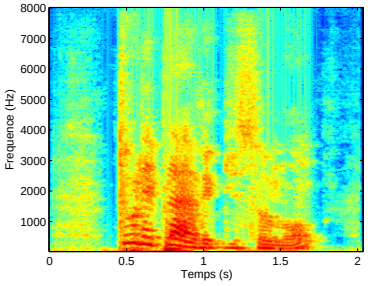
(a) Signal anéchoïque



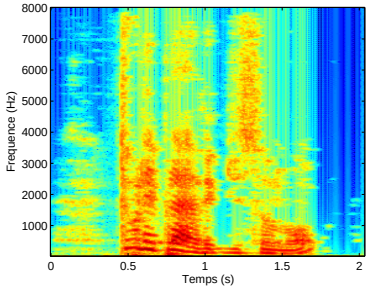
(b) Suppression par Soustraction Spectrale



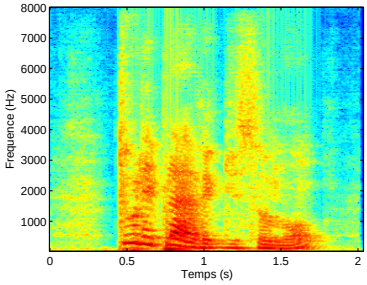
(c) Signal réverbérant ($RT_{60} = 200ms$)



(d) Suppression par filtrage de Wiener



(e) Signal déréverbéré par la méthode de Kameoka



(f) Suppression par filtrage d'Ephraim et Malah

FIGURE 5.4 – Spectrogrammes du signal original, du signal réverbérant et des résultats de la déréverbération. À droite, les trois variantes de l'algorithme de Kinoshita

Les deux algorithmes présentés sont assez rapides ce qui est encourageant pour une implémentation en temps réel. Dans leur implémentation actuelle, il s'agit d'algorithmes non causaux. On devra donc adapter le traitement à un fonctionnement en ligne. Le fait d'utiliser le filtrage d'Ephraim et Malah, à la place de la soustraction spectrale dans l'algorithme de Kinoshita, rend le traitement plus lent puisqu'il implique le calcul de deux fonctions de Bessel modifiées. Si les améliorations introduites par cette modification s'avèrent pertinentes, on peut envisager de précalculer des valeurs de ces fonctions. Ceci permet d'optimiser le temps de calcul.

Cette étude expérimentale nous a permis, en plus, de mettre en évidence la difficulté que pose l'évaluation des algorithmes de déréverbération. On a vu que les métriques objectives ne correspondent pas toujours à la perception de la réverbération . On a aussi constaté que selon le type d'algorithme utilisé, on trouve des métriques plus ou moins adaptées. Cela pose le problème de la comparaison d'algorithmes de déréverbération par des méthodes objectives. La mise en place de mesures de la déréverbération qui exploitent la perception humaine de ce phénomène est donc indispensable pour valider nos algorithmes et pour garantir la qualité de la parole dans le contexte de la téléphonie.

Dans la suite de nos travaux de recherche, nous approfondirons notre analyse des métriques perceptuelles telles que la BSD ou la PESQ, introduites dans le chapitre 2. Ainsi, on pourra mettre en place un protocole expérimental plus adapté à l'évaluation des algorithmes de déréverbération.

Chapitre 6

Conclusion

Lors de ce stage nous avons étudié le phénomène de la réverbération et ses effets sur les signaux de parole. En partant du consat qu'une composante de la réverbération, la réverbération tardive, dégrade l'intélligibilité de la voix, nous avons étudié différentes méthodes de déréverbération.

Nous avons présenté un état de l'art avec les différentes approches de déréverbération proposées à ce jour. On distingue deux types de traitement. D'une part, les traitements de suppression de la réverbération visent à estimer le signal anéchoïque à partir du signal observé. D'autre part, les approches d'annulation de la réverbération cherchent à estimer le filtre du canal acoustique pour ensuite appliquer un filtrage inverse à l'observation. C'est dernières sont plus difficiles à mettre en place parce que le filtre inverse est souvent instable, on doit donc chercher des solutions approchées au problème.

Nous avons également réalisé une étude des métriques pour quantifier l'effet de la réverbération. Des métriques objectives et subjectives ont été présentées et leur pertinence a été mise à l'épreuve.

Deux méthodes de suppression de la réverbération ont été choisies, étudiées dans le détail et implémentées sous Matlab. La première se base sur la prédiction linéaire à long terme pour estimer la partie réverbérante du signal de parole. La deuxième, réalise un déconvolution aveugle dans le domaine spectral par une technique de factorisation de matrices non-négatives.

Ces deux méthodes ont été évaluées avec les métriques présentées. Bien que d'un point de vue perceptuel la déréverbération a bien eu lieu, les mesures de la réverbération n'illustrent pas toujours ce fait. Des études supplémentaires devront être menées pour mettre en place des métriques qui soient fortement corrélées avec la perception de la réduction de la réverbération.

Le résultat de la déréverbération par les deux algorithmes n'est pas dépourvu de signaux parasites. Des modifications ont été proposées pour réduire l'effet de ces artéfacts.

L'ensemble des résultats présentés dans ce document constituent une référence pour développer des algorithmes performants pour la réduction de la réverbération au cours du doctorat que je réalise depuis de début du stage.

Bibliographie

[1] J.B. Allen and D.A. Berkley. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am*, 65(4) :943–950, 1979.

[2] Carlos Avendano. Study on the dereverberation of speech based on temporal envelope filtering. In *International Conference on Spoken Language Processing*, 1996.

[3] D. Bees, M. Blostein, and P. Kabal. Reverberant speech enhancement using cepstral processing. In *ICASSP '91 Proceedings of the Acoustics, Speech, and Signal Processing*, 1991.

[4] Jacob Benesty, M. Mohan Sondhi, and Yiteng Huang. *Springer Handbook of Speech Processing*. Springer, Berlin, 2008.

[5] R. Boite. *Traitement de la parole*. Presses polytechniques et universitaires romandes, 2000.

[6] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 27(2) :113–120, 1979.

[7] O. Cappé. Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. *IEEE Trans. Speech, Audio Process.*, 2(2) :345–349, 1994.

[8] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32 :1109–1121, 1984.

[9] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on*, 33(2), 1985.

[10] C. Evers, J.R. Hopgood, and J. Bell. Acoustic models for online blind source dereverberation using sequential monte carlo methods. In *ICASSP*, pages 4597–4600, 2008.

[11] C. Evers, J.R. Hopgood, and J. Bell. Blind speech dereverberation using batch and sequential monte carlo methods. In *ISCASSP*, pages 3226–3229, 2008.

[12] K. Furuya and A. Kataoka. Robust speech dereverberation using multichannel blind deconvolution with spectral subtraction. *IEEE Transactions on Audio, Speech & Language Processing*, 15(5) :1579–1591, 2007.

[13] S. Godsill, P. Rayner, and O. Cappé. Digital audio restoration. *Applications of digital signal processing to audio and acoustics*, pages 133–194, 2002.

[14] H. Haas. The influence of a single echo on the audibility of speech. *J. Audio Eng. Soc*, 20(2) :146–159, 1972.

[15] E.A.P. Habets. *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement*. PhD thesis, Technische Universiteit Eindhoven, 2007.

[16] E.A.P. Habets, N.D. Gaubitch, and P. A. Naylor. Temporal selective dereverberation of noisy speech using one microphone. In *ICASSP*, Las Vegas, USA, April 2008.

[17] S. Haykin. *Adaptive Filter Theory (4th Edition)*. Prentice Hall, September 2001.

[18] O. Hazrati, K. Kokkinakis, and P.C. Loizou. A blind subband-based dereverberation algorithm. In *ICASSP*, pages 4714–4717, 2010.

[19] H. Kameoka, T. Nakatani, and T. Yoshioka. Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '09, pages 45–48. IEEE Computer Society, 2009.

[20] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi. Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. *IEEE Transactions on Audio, Speech & Language Processing*, 17(4) :534–545, 2009.

[21] E.K. Kokkinis, A. Tsilfidis, E. Georganti, and J. Mourjopoulos. Joint noise and reverberation suppression for speech applications. In *Audio Engineering Society Convention 130*, 5 2011.

[22] K. Lebart, J.M. Boucher, and P.N. Denbigh. A new method based on spectral subtraction for speech dereverberation. *Acta Acustica United With Acustica*, 87(3) :359–366, 2001.

[23] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2001.

[24] B. Libbey and P.H. Rogers. The effect of overlap-masking on binaural reverberant word intelligibility. *The Journal of the Acoustical Society of America*, 116 :3141, 2004.

[25] M. Miyoshi and Y. Kaneda. Inverse filtering of room acoustics. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(2) :145–152, February 1988.

[26] A.K. Nábêlek, T.R. Letwoski, and F.M. Tucker. Reverberant overlap-and self-masking in consonant identification. *Journal of the Acoustical Society of America*, 1989.

[27] P. A. Naylor and N. D. Gaubitch. Speech dereverberation. In *Proc. Int. Workshop on Acoustic Echo and Noise Control*, 2005.

[28] P.A. Naylor and N.D. Gaubitch. *Speech Dereverberation*. Springer, 2010.

[29] S.T. Neely and J.B. Allen. Invertibility of a room impulse response. *Journal of the Acoustical Society of America*, 66(1) :165–169, 1979.

[30] A. V. Oppenheim, R. W. Schafer, and T. G. Stockham. Nonlinear filtering of multiplied and convolved signals. *Proceedings of the IEEE*, 56(8) :1264–1291, 1968.

[31] J.D. Polack. *La Transmission de l'énergie sonore dans les salles*. PhD thesis, Université du Maine, Le Mans, 1988.

[32] ITU Recommendation. P.862 perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. 2001.

[33] W.C. Sabine. *Collected papers on acoustics*. Harvard University Press, 1922.

[34] P. Smaragdis. Non-negative matrix factor deconvolution ; extraction of multiple sound sources from monophonic inputs. *Independent Component Analysis and Blind Signal Separation*, pages 494–499, 2004.

[35] H. Wallach, E.B. Newman, and M.R. Rosenzweig. The precedence effect in sound localization. *The American journal of psychology*, 62(3) :315–336, 1949.

[36] S. Wang, A Sekey, and A. Gersho. An objective measure for predicting subjective quality of speech coders. *IEEE Journal on Selected Areas in Communications*, 10 :819–829, 1992.

[37] N. Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. The MIT Press, 1964.

[38] W. Yang. *Enhanced modified bark spectral distortion (embsd) : an objective speech quality measure based on audible distortion and cognition model*. PhD thesis, Philadelphia, PA, USA, 1999. AAI9938714.

[39] N. Yasuraoka, T. Yoshioka, T. Nakatani, A. Nakamura, and H.G. Okuno. Music dereverberation using harmonic structure source model and wiener filter. In *ICASSP*, pages 53–56, 2010.

Annexe A

Calcul du prédicteur à long terme

On part du modèle de prédiction linéaire à long terme du signal $x(n)$:

$$x(n) = \sum_{p=1}^{N} a(p)x(n-p-D) + e(n), \tag{A.1}$$

où $e(n)$ est un bruit blanc, gaussien, indépendant et identiquement distribué. On procède par récurrence pour établir les équations permetttant de calculer le prédicteur $a(n)$. On veut faire intervenir la fonction d'autocorrélation du signal, on va donc multiplier l'équation (A.1) par $x(n-D-l)$, où $l \geq 1$. Pour des questions de lisibilité, on note $x_n = x(n)$, $a_n = a(n)$ et $e_n = e(n)$ par la suite.

Cas $\mathbf{l = 1}$

- On multiplie (A.1) par x_{n-D-1} et on prend l'espérance :

$$\mathbb{E}[x_n x_{n-D-1}] = \mathbb{E}\left[\sum_{p=1}^{N} a_p x_{n-p-D} x_{n-D-1}\right] + \mathbb{E}[e_n x_{n-D-1}] \tag{A.2}$$

- Comme $e(n)$ est un processus aléatoire, indépendant et identiquement distribué et que $D > 0$, on a $\mathbb{E}[e_n x_{n-D-1}] = 0$. En utilisant en plus la linéarité de l'espérance on a :

$$\mathbb{E}[x_n x_{n-D-1}] = \sum_{p=1}^{N} a_p \mathbb{E}[x_{n-p-D} x_{n-D-1}] \tag{A.3}$$

- On introduit alors la fonction d'autocorrélation r_k, définie par $r_k = \mathbb{E}[x_n x_{n-k}]$:

$$r_{D+1} = \sum_{p=1}^{N} a_p r_{p-1} \tag{A.4}$$

Cas $\mathbf{l = 2}$

- On multiplie (A.1) par x_{n-D-2} et on prend l'espérance :

$$\mathbb{E}[x_n x_{n-D-2}] = \mathbb{E}\left[\sum_{p=1}^{N} a_p x_{n-p-D} x_{n-D-2}\right] + \mathbb{E}[e_n x_{n-D-2}] \tag{A.5}$$

- On a toujours $\mathbb{E}[e_n x_{n-D-2}] = 0$. D'où :

$$\mathbb{E}[x_n x_{n-D-2}] = \sum_{p=1}^{N} a_p \mathbb{E}[x_{n-p-D} x_{n-D-2}] \tag{A.6}$$

- En utilisant la fonction d'autocorrélation r_k, on a :

$$r_{D+2} = \sum_{p=1}^{N} a_p r_{p-2} \tag{A.7}$$

$\mathbf{l = k}$

- On multiplie (A.1) par x_{n-D-k} et on prend l'espérance :

$$\mathbb{E}[x_n x_{n-D-k}] = \mathbb{E}\left[\sum_{p=1}^{N} a_p x_{n-p-D} x_{n-D-k}\right] + \mathbb{E}[e_n x_{n-D-k}] \tag{A.8}$$

- On a toujours $\mathbb{E}[e_n x_{n-D-2}] = 0$. D'où :

$$\mathbb{E}[x_n x_{n-D-k}] = \sum_{p=1}^{N} a_p \mathbb{E}[x_{n-p-D} xn - D - k] \tag{A.9}$$

- En introduisant la fonction d'autocorrélation, on obtient alors des équation analogues à celles de Yule-Walker, de la forme :

$$r_{D+k} = \sum_{p=1}^{N} a_p r_{p-k} \quad \forall k \geq 1 \tag{A.10}$$

En utilisant la symétrie de la fonction d'autocorrélation ($r_k = r_{-k}$), on réecrit alors le système (A.10) sous forme matricielle :

$$\begin{pmatrix} r_{D+1} \\ r_{D+2} \\ \vdots \\ r_{D+N} \end{pmatrix} = \begin{pmatrix} r_0 & r_1 & r_2 & \cdots & r_{N-1} \\ r_1 & r_0 & r_1 & \cdots & r_{N-2} \\ \vdots & & \vdots & & \vdots \\ r_{N-1} & r_{N-2} & r_{N-3} & \cdots & r_0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{pmatrix} \tag{A.11}$$

Ce qui s'écrit aussi, en notant $\mathbf{a} = [a_1, \cdots, a_N]^{\mathrm{T}}$ et $\mathbf{r_D} = [r_{D+1}, \cdots, r_{D+N}]^{\mathrm{T}}$

$$\mathbf{r_D} = R\mathbf{a} \tag{A.12}$$

Ainsi, le prédicteur à long terme est obtenu en inversant la matrice Toeplitz R définie dans (A.11) :

$$\mathbf{a} = R^{-1}\mathbf{r_D} \tag{A.13}$$