Mémoire de Recherche
Master ATIAM

Méthodes de Déréverbération Tardive de la Parole

Auteur :
Arthur BELHOMME

Responsable :
Éric HUMBERT

29 Juillet 2014

Résumé

Lorsque l'on réalise une prise de son, le signal émis par une source sonore est différent de celui acquis par le système de captation : différentes perturbations (principalement du bruit et de la réverbération) viennent s'ajouter au son direct. Contrairement au bruit, bien souvent considéré comme stationnaire et émis par une source différente [8], la réverbération est le résultat des multiples réflexions de l'onde sonore sur les différentes surfaces qu'elle rencontre. Alors qu'une faible réverbération peut être utile pour colorer un son, donner une sensation d'espace [26], une réverbération trop forte vient dégrader le signal et peut nuire à la compréhension d'un locuteur lorsqu'il s'agit de parole. Plus la source est éloignée du microphone plus l'effet de salle est présent et, donc, moins bonne est l'intelligibilité [43] ; cas très fréquent dans des conversations téléphoniques de type mains-libres.

C'est justement dans ce cadre que se pose la problématique de mon stage de fin d'études : *Méthodes de déréverbération tardive de la parole*. Ce dernier consiste à étudier le phénomène de réverbération et les différentes méthodes de déréverbération connues jusqu'à ce jour afin d'en implémenter une en temps réel, embarquée dans des téléphones de conférence type mains-libres. Ce problème étant relativement complexe (élimination d'un signal fortement corrélé à court terme, décorrélé à long terme), de nombreuses méthodes figurent dans l'état de l'art, abordant ce sujet sous différents angles.

Après une étude du phénomène de réverbération nous présenterons les différentes techniques issues de l'état de l'art, en mono- et multi-canal, pour n'en conserver que deux : estimation de la réverbération par prédiction linéaire à long terme [28] ou alors basée sur le modèle stochastique de réponse de salles [20]. Cette énergie de réverbération tardive sera ensuite supprimée en utilisant deux types de gains : soustraction spectrale [33] ou Optimally-Modified Log Spectral Amplitude [11].

Les performances de ces méthodes seront ensuite analysées de manière subjective et objective s'appuyant sur des critères développés par Loizou dans [25]. En fonction de ces résultats, la méthode retenue sera présentée et justifiée en fonction des produits dans lesquels elle sera implémentée.

Au cours de ce stage j'ai également pu me pencher sur l'étude de la phase des signaux, et mettre en avant son utilité dans le processus de déréverbération. Je présenterai dans la dernière partie les résultats obtenus jusqu'à présent, dont les travaux de recherches se poursuivront au cours de ma thèse orientée vers ce même sujet.

Mots-clés Déréverbération, acoustique des salles, réseau de microphones, traitement du signal.

Abstract

When a signal is recorded through a microphone, the output signal will be different from the original one : many disturbances are added to the direct sound (mainly noise and reverberation). Unlike noise, which is frequently assumed to be stationary and issued from a diffrent source [8], reverberation is the result of the many reflexions of the acoustic wave on the different surfaces it meets. While a soft reverberation would be usefull to color a sound, to give space feeling [26], a strong reverberation deteriorates the signal and may jeopardize speaker's intelligibility. The more the microphone is far from the source, the more the reverberation is noticeable, and so the worst is inteligibility [43] ; which is a very common context for hands-free communications.

This is the context of my research internship, entitled *Suppression of late reverberation effect on speech signals*. It consists in studying the reverberation phenomenon and the different known dereverberation processes in order to implement one, with real-time performance, loaded in hand-free devices. This relatively complex problem (suppression of a short term highly-correlated signal, but long term decorrelated) has been adressed with different approaches in the state of the art.

After a study of the reverberation phenomenon, I will present the different methods contained in the state of the art, for single- or multi-microphone, and choose two of them : estimation of late reverberation based on multiple-step linear prediction [28] or on the statistical model of room impulse [20]. This late reverberant energy is then suppressed with two diffrent gains : spectral substraction [33] or Optimally-Modified Log Spectral Amplitude [11].

Performance will be analysed from a subjective and objective viewpoint, using Loizou criteria developed in [25]. According to these results, I will expose the chosen method and explain my choice toward the products where it will be loaded.

During this internship I also studied the signal's phase, and hilighted its value in a dereverberation process. The first results I obtained will be discussed in the last section, whose future work will continue during my Ph.D.

Keywords Dereverberation, room acoustics, microphone array, signal processing.

Table des matières

Chapitre 1

Introduction

Lorsque l'on cherche à enregistrer un son issu d'une source localisée, on obtient rarement sa reproduction exacte. Dès que l'onde acoustique se propage dans l'espace, des perturbations viennent la modifier et bien souvent dégrader le signal. Ces perturbations sont propres à l'environnement dans lequel l'onde évolue : un bruit ambiant (moteur, bruit de fond, etc.), de l'écho, de la réverbération. Le bruit est souvent considéré comme étant stationnaire et décorrélé du son visé [8] tandis que la réverbération et l'écho ne le sont pas, compliquant alors leur élimination.

Ces phénomènes sont dûs aux réflexions de l'onde acoustique sur les différentes surfaces qu'elle rencontre. En fonction du type de lieu où l'on se trouve, on n'obtient plus uniquement le son direct mais une somme de versions plus ou moins atténuées et espacées dans le temps. Ce phénomène peut être intéréssant dans certains cas pour donner une coloration ou une sensation d'espace [26], par exemple en production musicale, mais peut devenir gênant lorsqu'il n'est pas maîtrisé ou trop marqué. Dans le cas de la parole, la réverbération nuit à l'intelligibilité d'un discours ou d'une communication [42] ; que ce soit pour l'Homme ou pour des systèmes de reconnaissance automatique de la parole (commandes vocales principalement).

Pour les conversations téléphoniques utilisant un combiné, le problème ne se pose pas vraiment : le locuteur est suffisament proche du microphone pour que l'onde acoustique ne subisse pas l'effet de salle, tant la distance parcourue est faible. En revanche pour des systèmes de type mains-libres ou de conférence, le locuteur peut se retrouver à une distance suffisante du microphone pour que l'énergie de l'onde directe soit du même ordre de grandeur que celle des diverses réflexions. On entend alors ce que dit la personne, mais de façon moins claire, moins intelligible ; au même titre qu'observer une image floue ne nous empêche pas de percevoir ce qu'elle représente, à condition de faire un effort supplémentaire.

Afin de répondre à cette problématique, de nombreux travaux de recherche ont été effectués en déréverbération afin d'être, bien souvent, embarqués dans des systèmes de téléphonie mains-libres (conférence, automobile, etc.). Deux grandes familles se démarquent : l'annulation et la supression [20]. Tandis que la première cherche à estimer la réponse impulsionnelle de la salle afin de l'inverser et re-convoluer le signal avec, la seconde cherche à estimer l'énergie propre à la réverbération afin de la retirer du signal. Les méthodes d'annulation, plus exactes mais moins robustes et plus lentes (problèmes de stabilité dans l'inversion de

matrice) ne seront pas étudiées en profondeur dans le cadre de ce stage, préférant les méthodes par suppression plus adaptées à un usage robuste et temps réel. Ce stage de recherche s'effectuant en entreprise, il est entendu d'adapter le sujet d'étude aux produits développés par *Invoxia*, dont je vais à présent faire une présentation.

1.1 Présentation de l'entreprise

Fondée en 2010 par Éric Carreel et Serge Renouard (anciennement Inventel), la société *Invoxia* se spécialise dans les appareils de télécommunication. De la conception à la commercialisation, en passant par le développement et la réalisation, l'entreprise propose un moyen de communication de meilleure qualité que la téléphonie "classique". Pour cela, elle se sert de la technique de *Voix sur IP* dont l'augmentation constante du débit permet d'accéder à une bande passante plus large et donc une qualité de communication supérieure.

La cible principale étant les entreprises, les produits *Invoxia* ont de réels atouts en ce qui concerne la conférence à plusieurs. Ils permettent des conversations type mains-libres de très bonne qualité, quel que soit le contexte de l'appel (disposition des participants, distance à la base, etc.).

Afin d'obtenir de telles performances, les téléphones disposent de particularités tant *hardware* que logicielle. Pour ce qui est du matériel, les produits *Invoxia* sont équipés d'un réseau de six à huit haut-parleurs, avec autant de microphones. Cela permet de capter bien plus d'informations qu'un combiné standard monocanal et donc de développer des traitements d'amélioration du signal, transmis ensuite sur un support adapté.

Parmi ces traitements, on peut trouver un *Beamformer* permettant de favoriser la prise de son dans une direction et ainsi réduire les perturbations venant de directions autres que celle du locuteur. Le réseau de haut-parleurs permet ensuite de projeter les différentes sources dans différentes zones de l'espace afin de recréer un champ sonore de haute fidélité.

Afin de supprimer le bruit qui vient parasiter les conversations (bruit de fond, source indésirable, etc.) ou encore un potentiel écho, les appareils sont équipés d'un algorithme de débruitage et d'annulateur d'écho. Reste alors à traiter le phénomène de réverbération afin de parfaire la qualité et l'intelligibilité des conversations des produits *Invoxia*, d'où le sujet de mon stage : *Méthodes de déréverbération tardive de la parole.*

1.2 Organisation du mémoire

Avant d'aborder la déréverbération il faut bien comprendre ce qu'est la réverbération. C'est pourquoi nous commencerons dans le **Chapitre 2** par rappeler les prérequis d'acoustique des salles afin d'expliquer le phénomène de réverbération et son impact sur l'intelligibilité d'une conversation.

Le **Chapitre 3** présentera une partie conséquente de l'état de l'art en matière de déréverbération. À la suite de cette étude, nous indiquerons les méthodes développées et implémentées en justifiant ces choix vis-à-vis du contexte du stage.

Le processus de déréverbération peut se diviser en deux grandes étapes : l'estimation de la réverbération puis sa suppression. Le **Chapitre 4** a pour but de présenter, pour chacune

de ces étapes, les différentes méthodes utilisées en mono- et multicanal.

C'est dans le **Chapitre 5** que seront présentés les différents algorithmes testés. Les résultats seront évalués par des critères subjectifs et objectifs. Enfin, ce chapitre exposera la méthode retenue, qui sera ensuite implémentée dans les appareils *Invoxia*.

Le **Chapitre 6**, dernier chapitre de ce mémoire de recherche, présentera mes premiers résultats sur l'utilité de la phase dans un processus de déréverbération. Ces travaux de recherche seront poursuivis tout au long de ma thèse réalisée dans les laboratoires de *TélécomParisTech*, en collaboration avec *Invoxia*.

Chapitre 2

Acoustique des salles

Ce chapitre a pour but de rappeler les bases de l'acoustique afin d'expliquer le phénomène de réverbération et d'en analyser les conséquences sur notre perception du son. Nous présenterons d'abord la structure des ondes acoustique afin de définir la fonction de transfert acoustique, et donc la réponse impulsionnelle d'une salle. À partir de cette réponse, nous pourrons caractériser la réverbération et étudier ses conséquence en fonction de ses paramètres.

2.1 Prérequis d'acoustique

2.1.1 L'onde acoustique

Lorque l'on parle d'onde acoustique, il est question de l'onde de pression évoluant dans le milieu considéré : bien souvent, l'air. Cette onde peut être stationnaire (oscillation malgré des ventres et noeuds de vibrations fixes) ou progressive (déplacement des ventres et noeuds de vibrations de l'onde). Elle est bien souvent considérée comme plane : ses fronts d'ondes sont des plans supposés infinis, ce qui peut se vérifier même pour une source sphérique si on s'en éloigne suffisament [31].

Sous ce modèle, une onde de pression $p(\mathbf{r}, t)$ évoluant dans un milieu fluide non visqueux et homogène satisfait l'équation d'Euler linéarisée :

$$\nabla^2 p(\mathbf{r}, t) - \frac{1}{c^2} \frac{\partial p(\mathbf{r}, t)}{\partial t} = 0 \tag{2.1}$$

Le vecteur $\mathbf{r}$ représente la position de l'observation : $\mathbf{r} = (x, y, z)$ en cartésien donnant lieu à $\nabla = (\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z})$; $\mathbf{r} = (u, \theta, z)$ en cylindrique donnant lieu à $\nabla = (\frac{\partial}{\partial u}, \frac{1}{u}\frac{\partial}{\partial \theta}, \frac{\partial}{\partial z})$. Ce modèle reste fidèle pour des ondes vérifiant $|p(\mathbf{r}, t)| \ll \rho_0 c$, où ρ_0 et c sont respectivement la densité et la vitesse du son dans le milieu considéré.

Lorsque l'on rajoute une source $s(\mathbf{r}, t)$ à un endroit $\mathbf{r}_s$ l'équation (2.1) se transforme en :

$$\nabla^2 p(\mathbf{r}, t) - \frac{1}{c^2} \frac{\partial p(\mathbf{r}, t)}{\partial t} = -s(\mathbf{r}, t) \tag{2.2}$$

Enfin, si on applique à (2.2) la transformée de Fourier $\mathcal{F}\{g(\mathbf{r}, t)\}(\omega) = \int_{-\infty}^{\infty} g(\mathbf{r}, t) e^{-i\omega t} dt$ on obtient :

$$\nabla^2 P(\mathbf{r}, \omega) - k^2 P(\mathbf{r}, \omega) = -S(\mathbf{r}, \omega) \tag{2.3}$$

Avec $k = \frac{\omega}{c}$ le nombre d'onde, ω la pulsation.

On définit alors la fonction de transfert acoustique $H(\mathbf{r}, \omega)$ comme étant la réponse de la pression lorsque cette dernière est soumise à un dirac (c'est donc une fonction de Green), c'est-à-dire lorsque la source en $\mathbf{r}_s$ se traduit par $S(\mathbf{r}, \omega) = \delta(\mathbf{r} - \mathbf{r}_s)$. L'équation (2.3) donne alors :

$$\nabla^2 H(\mathbf{r}, \omega) - k^2 H(\mathbf{r}, \omega) = -\delta(\mathbf{r} - \mathbf{r}_s) \tag{2.4}$$

La transformée de Fourier inverse $\mathcal{F}^{-1}\{G(\mathbf{r}, \omega)\}(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(\mathbf{r}, \omega) e^{i\omega t} d\omega$ de la fonction de transfert acoustique nous donne la réponse impulsionnelle acoustique $h(\mathbf{r}, t)$. C'est cette dernière que l'on défini comme étant la réponse de la salle lorsque l'on travail en acoustique des salles.

Alors, une source émettant une onde acoustique $s(\mathbf{r}, t)$ dans une salle de réponse impulsionnelle $h(\mathbf{r}, t)$, générera une onde $y(\mathbf{r}, t)$ telle que :

$$y(\mathbf{r}, t) = (s * h)(\mathbf{r}, t) + b(\mathbf{r}, t) \tag{2.5}$$

avec b un bruit parasite (bruit blanc, brouhaha, moteur, etc.). On voit bien que l'effet de la réverbération est contenu dans la réponse impulsionnelle de la salle puisque cette dernière permet de passer d'un son anéchoïque au son réellement perçu dans la salle. C'est ce que nous allons à présent étudier dans cette seconde sous partie.

2.1.2 Caractérisation de la réverbération

L'onde acoustique issue de la source s, évoluant dans la salle, va se réfléchir sur les différentes surfaces de son milieu. Considérons une salle vide, les surfaces rencontrées sont donc les quatre murs, le sol et le plafond. Chacune d'elle, en fonction son coefficient de réflexion (dépendant du matériau utilisé, son absorption) et sa distance par rapport à la source et au récepteur (augmentant ou diminuant le trajet total parcouru), va renvoyer une version plus ou moins atténuée et retardée de s.

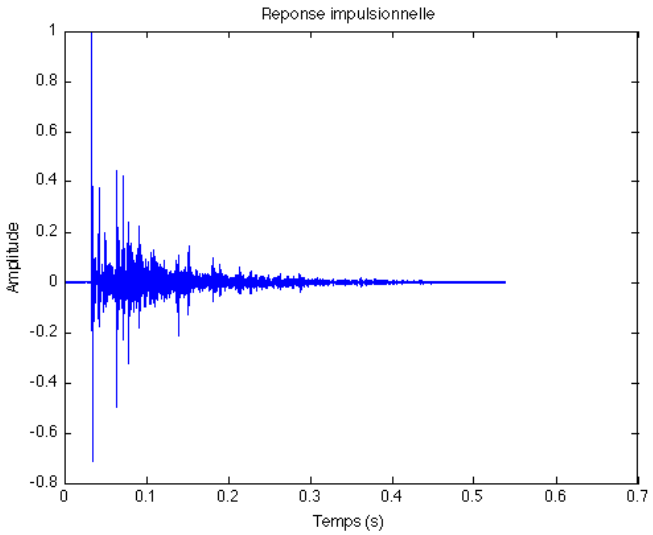

FIGURE 2.1 – Exemple de réponse impulsionnelle

On comprend donc à travers (2.5) que chaque pic de h va traduire le retard et l'atténuation de s, en fonction de son amplitude et emplacement par rapport au premier pic. Il est alors possible de décomposer le son perçu en trois parties via l'expression de h. Si on échantillonne le temps t en échantillons n à une fréquence f_e, on peut écrire (2.5) comme :

$$y(n) = s(n)h(0) + \sum_{k=1}^{T_l} s(n-k)h(k) + \sum_{k=T_l+1}^{\infty} s(n-k)h(k) + b(n) \tag{2.6}$$

La première partie de (2.6) correspond au son direct $y_{\text{dir}}(n) = \alpha s(n)$ tandis que les deux autres sont propres à la réverbération. On peut alors discerner les réflexions précoces, contenues dans les T_l secondes suivant l'impulsion, de la réverbération tardive, à la suite de ces premières réflexions et jusqu'à la disparition totale du son.

Il est alors possible de caractériser la réverbération en fonction de la réponse impulsionnelle de la salle. Le temps de réverbération ainsi que le rapport d'énergie directe sur l'énergie de réverbération sont deux paramètres représentatifs de la réverbération [20].

Temps de réverbération

Quand on parle de temps de réverbération, sans autres précisions, on fait référence au temps de réverbération à 60 décibels (dB), noté RT_{60}. Ce paramètre correspond au temps mis à l'énergie d'un son pour diminuer de 60 dB, après son apparition [31]. Pour le calculer on peut se servir de la courbe de diminution de l'énergie (*Energy Decay Curve*, EDC en anglais) et regarder le temps mis pour diminuer de 60 dB.

$$\text{EDC}(t) = \int_t^{\infty} h^2(\tau) d\tau$$

On remarque que ce temps RT_{60} va dépendre directement de la réponse impulsionnelle de la salle. Sabine a dédié une partie de ses travaux à l'étude du temps de réverbération et a pu montrer que ce dernier était proportionnel au volume de la salle utilisée, mais inversement proportionnel à son absorbtion [47].

Rapport d'énergie directe et énergie de réverbération

Généralement désigné sous le sigle DRR (*direct to reverberant ratio*, en anglais), ce paramètre permet d'exprimer le rapport entre l'énergie propre au chemin direct de l'onde sonore et celle due à ses multiples réflexions. Une fois de plus, cet indicateur de réverbération se calcule directement à partir de la réponse impulsionnelle de la salle (on considère que le son direct arrive à l'échantillon n_d) :

$$\text{DRR} = 10\log_{10}\left(\frac{E_d}{E_r}\right) = 10\log_{10}\left(\frac{\sum_{k=0}^{n_d} h^2(n)}{\sum_{k=n_d+1}^{\infty} h^2(n)}\right)$$

Cette valeur, en décibels, nous donne une information sur l'importance de la réverbération. Elle correspondrait au réglage "*Dry/Wet*" que l'on peut trouver dans les effets de réverbérations en production musical.

Modèles de réponse de salle

Ces caractéristiques ont permis de développer des modèles de réponse impulsionnelle de salle, afin de pouvoir travailler de façon aussi bien pratique que théorique avec le phénomène de réverbération. C'est Polack, suite à une première intuition de Moorer [39], qui développa le modèle stochastique de la réverbération [44] :

$$h(t) = \begin{cases} b(t)e^{-\delta t} & \text{si } t \geq 0 \\ 0 & \text{sinon} \end{cases} \tag{2.7}$$

Ici b est un bruit blanc Gaussien centré et le paramètre δ est lié au temps de réverbération par :

$$\delta = \frac{3\log_e(10)}{\mathrm{RT}_{60}}$$

Il faut préciser que $\delta(f)$ et $\mathrm{RT}_{60}(f)$ varient en fonction de la fréquence : $\delta(f)$ augmente avec la fréquence puisque le temps de réverbération est plus long en basses fréquences, cf. Fig. 2.2. Mais pour alléger les notations nous considérerons que nous nous plaçons toujours dans une bande de fréquences donnée, les raisonnements étant identiques entre chaque bande [20].

Ce modèle statistique de la réverbération est valable à partir du moment où le nombre d'échos simultanés arrive au nombre de 10 [45] et que la distance entre le microphone et la source est plus grande que la distance critique [31]. La distance critique étant la distance pour laquelle l'énergie propre au chemin direct est égale à celle des diverses réflexions ; on a donc un DRR de zéro décibels.

Ce modèle nous permet de calculer l'énergie de la réponse impulsionnelle en fonction des paramètres de la réverbération :

$$\mathcal{E}_h\{h^2(t)\} = \sigma^2 e^{-2\delta t},$$

où σ^2 correspond à la variance du bruit blanc et l'opérateur $\mathcal{E}\{.\}$ à la moyenne sur les différentes réalisations de h. Cette expression nous permettra d'estimer l'énergie propre à la réverbération tardive d'un signal, dans le **Chapitre 4**.

Afin de s'affranchir de la condition "Distance source/microphone $\geq$ Distance critique" sans pour autant surestimer l'énergie de réverbération tardive, Habets a généralisé le modèle de Polack [20]. La principale différence réside dans l'utilisation de deux bruits blancs, de variances différentes, afin de représenter la réponse de la salle :

$$h(t) = \begin{cases} b_d(t)e^{-\delta t} & \text{si } 0 \leq t < T_r \\ b_r(t)e^{-\delta t} & \text{si } t \geq T_r \\ 0 & \text{sinon} \end{cases} \tag{2.8}$$

Le temps T_r représente le temps incluant le chemin direct, les bruits blancs Gaussien centrés b_d et b_r sont décorrélés. On obtient alors deux énergies pour une même réponse, différenciant le chemin direct de la réverbération à travers deux variances différentes σ_d^2 et σ_r^2. Evidemment, on suppose $\sigma_d^2 \geq \sigma_r^2$ puisque la répartition des pics est bien plus dense en fin de réponse qu'au début. On retrouve le modèle stochastique de Polack lorsque ces dernières sont égales.

Avant d'estimer l'énergie propre à la réverbération tardive des signaux, grâce à ce modèle de réponse de salle, observons son effet sur des signaux de parole.

2.2 Effet de la réverbération sur la parole

2.2.1 Impact de la réverbération

Comme expliqué précédement, la réponse impulsionnelle d'une salle peut être réduite en trois parties : un premier pic de grande amplitude (le chemin direct de l'onde acoustique), une série de quelques pics bien distants (deux à cinq en général, correspondant aux premières réflexions de l'onde), une décroissance exponentielle d'une série de pics très proches tel un bruit blanc (correspondant au modèle de Polak pour la partie tardive de la réverbération).

La plupart des articles abordant la déréverbération commencent leur introduction en rappelant que les premières réflexions ne sont pas nuisibles, au contraire. Comme expliqué dans [20] cette réverbération précoce n'est pas perçue comme étant une source différente tant que son retard n'excède pas les 80-100ms suivant le son direct. En plus de donner une couleur au son, d'apporter une information d'espace et de distance, elle permet également de renforcer le chemin direct. C'est ce que l'on retrouve dans l'effet de précédence (de Haas) en psychoacoustique [18], montrant alors l'apport de ces réflexions précoces dans l'intelligibilité de la parole. Les réflexions sont suffisament proches du chemin direct qu'elles l'amplifient, ce qui explique pourquoi on arrive mieux à avoir une discution dans un espace fermé plutot qu'en plein air.

En revanche, la partie tardive de la réverbération va, elle, dégrader sévèrement l'intelligibilité du discours. Cette succession de réflexions débute généralement suffisamment longtemps après le chemin direct pour que notre cerveau traite cette information comme étant une source différente. La forte densité des pics en fin de réponse modélise bien une superposition d'un nombre important de versions retardées et atténuées du signal source. Cette superposition a pour conséquence un étalement horizontal et vertical du spectre du signal anéchoïque.

Étalement horizontal

C'est sûrement la conséquence que l'on se représente le plus facilement : un son réverbéré va rester audible plus longtemps, en disparaissant petit à petit. On s'attend donc à voir un spectrogramme comportant des trainées, là où nous aurions eu des évenement ponctuels dans un signal anéchoïque. C'est ce que l'on constate en observant le spectrogramme d'un "calp" de main soumis à une forte réverbération (Fig. 2.2).

On devine alors aisément que la forme d'onde va se retrouver également modifiée, avec des variations plus lentes de l'amplitude. Il est alors plus difficile de discerner les différentes parties du discours juste en observant la forme d'onde, ce que l'on peut constater Fig. 2.3.

Étalement vertical

En réalité, c'est la conséquence de l'étalement horizontal mais sur des signaux de fréquences variables. En effet, la réverbération ne se fait sentir que lorsque l'on quitte un état stationnaire : c'est pourquoi on se rendra compte qu'une sinusoïde pure est réverbérée qu'à son apparition et extinction (idem pour un dirac, puisqu'on excite l'ensemble des fréquences dans une durée infiniment petite).

En revanche, dès que les composantes fréquentielles d'un signal réverbéré varient au cours du temps, on entend bien le phénomène de réverbération. L'étalement vertical que l'on peut apercevoir sur le spectrogramme est dû à un étalement horizontal sur une très courte durée :

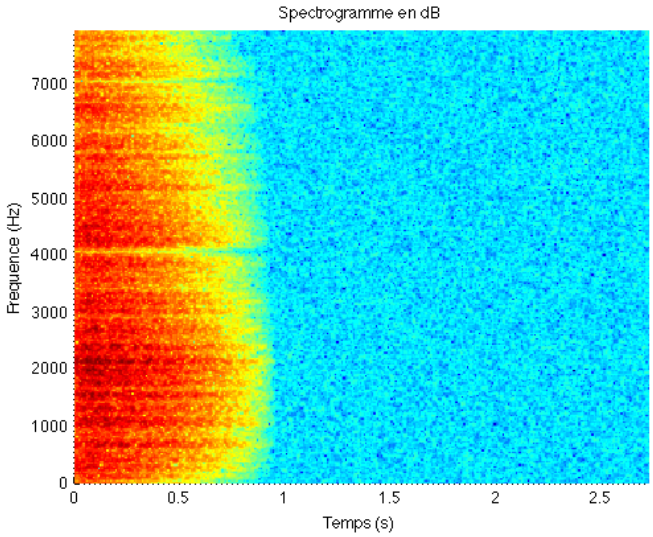

FIGURE 2.2 – Spectrogramme d'un "clap" de main réverbéré, sur 512 points, échantillonné à 16kHz

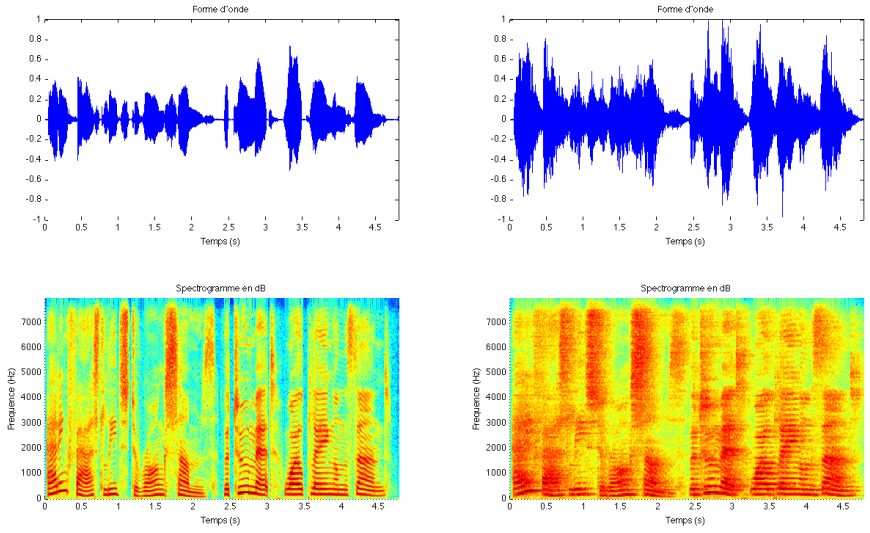

FIGURE 2.3 – Forme d'onde et spectrogramme d'un extrait de parole (échantillonné à 16kHz, sur 512 points) anéchoïque à gauche, réverbéré à droite

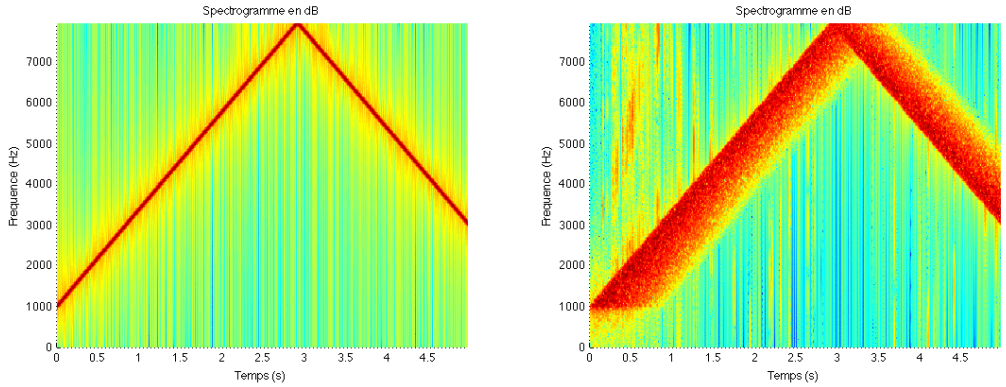

FIGURE 2.4 – Spectrogramme du signal anéchoïque à gauche, réverbéré à droite

si une sinusoïde passe d'une fréquence f_1 à f_2 entre les temps t_1 et t_2, on observe à l'instant t_2 la fréquence f_2 -liée au chemin direct- mais également le reste de l'énergie de f_1 qui avait eu lieu à l'instant t_1. On se rend bien compte de ces phénomènes en observant la différence entre le spectrogramme d'un *sweep* (onde monochromatique dont la fréquence varie de façon linéaire) anéchoïque et réverbéré (Fig. 2.4).

Or, la voix humaine peut être modélisée comme somme de sinusoïdes à fréquence variable, en rapport harmonique ([36], [32]), on comprend donc que la réverbération va avoir de sérieuses conséquences sur l'intelligibilité de la parole.

2.2.2 Conséquences sur la parole

On peut les résumer en deux phénomènes distincts : l'auto-masquage (*self-masking* en anglais) et le masquage de superposition (*overlap-masking* en anglais). Ces deux types de masquages ont été introduits dans [9] et [42] et sont des conséquences directes des effets de la réverbération décrits précédemment.

Le premier, auto-masquage, a un effet relativement moins gênant sur l'intelligibilité que le second, masquage de superposition [30]. Il correspond à la distorsion temporelle et fréquentielle de chaque phonème, suite aux deux types d'étalements énoncés précédemment. Les transitions des formants sont alors moins nettes, ainsi que l'enveloppe des consonnes qui se retrouve lissée.

Le second, masquage de superposition, se rapporte plutôt à l'étalement horizontal du spectre qu'induit la réverbération. Ce phénomène apparait lorsqu'il y a un recouvrement spectral entre deux phonèmes consécutifs. Pour une élocution très lente et bien détachée, l'énergie de réverbération vient "remplir" les silences entre les mots, ne perturbant donc pas vraiment l'intelligibilité du discours. En revanche pour une élocution normale (sans silences marqués entre chaque mots, voire phonèmes) l'énergie de réverbération de la fin d'un phonème vient se superposer à l'énergie de la suivante. On discrétise donc moins bien les mots de notre interlocuteur, dégradant sévèrement l'intelligibilité de son discours.

Ces effets se retrouvent non seulement à travers l'oreille humaine mais également pour

des systèmes électroniques commandés par la parole. Ces derniers sont équipés d'algorithmes de reconnaissance automatique de la parole (*Automatic Speech Recognition*, ASR en anglais) qui analysent le spectre et l'enveloppe temporelle des signaux de paroles afin d'en déduire les messages contenus. On peut donc, à partir des résultats obtenus, calculer un taux d'erreurs par rapport à ce qui est réellement dit par le locuteur. Ce dernier, appelé taux d'erreur de mots (*Word Error Rate*, WER en anglais), donne alors un bon indicateur de performance de ces systèmes, pouvant également mettre en valeur les conséquences de la réverbération sur l'intelligibilité de la parole.

Habets montra dans [20] que ce pourcentage d'erreur augmente rapidement à partir d'un temps de réverbération de $0.2s$ (pour une distance source-microphone de 3 mètres), et qu'il augmente de façon proportionelle avec la distance entre la source et le microphone. Il constata également que l'effet de la réverbération tardive dégradait davantage le WER sur des séquences entières de paroles que sur des mots isolés, ce qui est cohérent avec le phénomène de masquage de superposition énnoncé précédemment.

L'ensemble de ces résultats corroborent bien les premiers travaux, datant des années 1980, soulignant l'effet de la réverbération sur les différents marqueurs de l'intelligibilité de la parole ([43], [3], [6]).

2.3 Conclusion

Cette partie nous a permis de rappeler le modèle physique de l'onde de pression acoustique, vérifiant l'équation d'Euler. En transformant cette équation dans le domaine de Fourier, et en supposant un dirac en guise de source, nous avons pu définir la fonction de transfert d'une salle. À partir de cette dernière nous avons défini la réponse impulsionnelle d'une salle, qui n'est autre que sa transformée de Fourier inverse.

L'information contenue dans cette dernière permet de caractériser la réverbération présente dans une salle. En fonction de la répartition des différents pics de ce filtre, le signal anéchoïque sera acompagné de ses versions plus ou moins atténuées et retardées dans le temps. Nous avons également vu que des modèles statistiques de ces salles ont été établis afin de pouvoir modéliser le phénomène de réverbération, et donc de pouvoir le traiter.

On peut extraire de cette réponse impulsionnelle certains paramètres permettant de décrire l'intensité ou la durée de la réverbération. Ces derniers sont liés à la dégradation de l'intelligibilité des signaux de parole ; constat subjectif, mais également validé par les systèmes de reconnaissance automatique de la parole.

Contrairement à la partie précoce de la réverbération, la réverbération tardive nuit à la bonne communication entre interlocuteurs. C'est pourquoi nous allons à présent étudier les différentes méthodes issues de l'état de l'art permettant d'éliminer ce phénomène nuisible.

Chapitre 3

État de l'art

La réverbération est un phénomène qui peut devenir gênant dans certaines situations (communications téléphoniques mains-libres, systèmes à commandes vocales, etc.), il est donc compréhensible que de nombreuses méthodes de déréverbération figurent dans l'état de l'art afin de s'en débarrasser. Une des premières solutions pour minimiser la réverbération se trouve dans un amplificateur breveté par Ryall en 1939 [46], dans lequel les basses fréquences étaient moins amplifiées, ces dernières étant plus sujettes à la réverbération.

Depuis, différentes approches ont vu le jour pour déréverbérer proprement un signal sans pour autant trop le distordre. Malgré la forte distinction entre le traitement monocanal et multicanal, il est préférable de les classer dans deux autres grandes familles : méthodes par annulation ou par suppression [41]. C'est sous ces deux approches que nous allons différencier les méthodes présentes dans l'état de l'art, en rappelant les spécificités de ces deux familles. L'utilisation récente (années 2010) des dictionnaires en traitement du signal permet de considérer une troisième approche de la déréverbération, que nous détaillerons également.

3.1 Méthodes par annulation

Nous avons vu dans le chapitre précédent que lorsque une source émet un son ce dernier est modifié par son environnement, ce que l'on peut modéliser en l'absence de bruit par :

$$x(t) = (s * h)(t)$$

où s est le signal émis par la source, h la réponse impulsionnelle de la salle. Les méthodes par annulation visent à estimer la réponse h, et donc sa transformée de Fourier H, afin de l'inverser. Une fois que cette réponse inverse est estimée, on filtre le signal x avec, dans le but d'estimer s. Contrairement aux problèmes de déconvolution classique de systèmes linéaires-invariants, où l'on connait l'entrée et la sortie, l'entrée s est inconnue. Cette inconnue rend ce problème plus difficile, on parle alors de "déconvolution aveugle".

Seulement, la réponse d'une salle est très sensible aux positions et orientations du microphone et de la source. Une légère modification de ces derniers entraîne une nouvelle réponse à estimer, or les techniques par annulation mettent en jeu des calculs assez lourds puisqu'ils doivent généralement inverser des matrices de grande taille. Ces deux désavantages, en terme de robustesse et de rapidité, nous amènent à ne pas utiliser ces méthodes dans le

cadre du stage. C'est pourquoi la description de ces dernières sera plus succincte que celle des méthodes par suppression. Le lecteur pourra se reporter aux travaux de Haykin ([23], [24]) ou encore d'Habets ([20]) pour une présentation plus détaillée.

3.1.1 Approximation de filtrage inverse

Cette technique consiste à calculer un gain proche de l'inverse du filtre acoustique entre une source et un microphone. Ce procédé, nommé HERB pour *Harmonicity based dEReverBeration*, se base sur le modèle harmonique de la voix : le gain est calculé en fonction du rapport entre le signal réverbéré entrant et ce même signal filtré par un filtre harmonique, qui s'adapte à chaque trame de calcul. Kinoshita montra dans [29] que cette méthode améliore les performances des systèmes de reconnaissance automatique de la parole puisqu'elle diminue le taux d'erreur WER. En revanche, il faut plus d'une heure de données afin d'obtenir un filtre convenable [20], ce qui n'est donc pas adapté à un usage en temps réel.

3.1.2 Méthode LIME

Cette méthode proposée par Delcroix dans [12] permet de déréverbérer de manière presque parfaite des signaux soummis à de courtes réponses impulsionnelles. Comme son nom (*LInear-predictive Multi-input Equalization*) l'indique, cette méthode se situe dans les techniques multicanal, basée sur le principe de prédiction linéaire.

Elle consiste à estimer, à partir d'un signal réverbéré blanchi, les coefficients de prédiction linéaire afin de calculer l'erreur de prédiction. Cette dernière est ensuite filtrée par le filtre inverse du modèle auto-régressif de la parole [37] afin d'obtenir le signal déréverbéré. Nous reviendrons sur l'utilité du blanchiment dans le **Chapitre 4** puisqu'il figure parmi les techniques implémentées durant le stage.

3.1.3 Méthode MINT

Le principe de *Multiple-input/output INverse Theorem* a été introduit par Miyoshi et Kaneda dans [38]. Si on considère M acquisitions $x_m(n) = (h_m * s)(n)$ d'une source s que l'on vient filtrer une à une afin d'obtenir ŝ, l'estimation de s, on peut écrire :

$$\hat{\mathrm{s}}(\mathrm{n}) = \sum_{m=1}^{M} (g_m * x_m)(n) \tag{3.1}$$

avec g_m les M filtres d'égalisation. Si on suppose à présent que notre source à estimer est un dirac $\delta(n-\tau)$ on peut réécrire (3.1) comme :

$$\delta(n-\tau) = \sum_{m=1}^{M} (g_m * h_m)(n)$$

Ce qui nous donne, dans le domaine de Laplace :

$$\sum_{m=1}^{M} G_m(z) H_m(z) = 1 \tag{3.2}$$

où G_m et H_m sont respectivement l'expression de g_m et h_m dans le domaine de Laplace. Or l'existence d'une solution de (3.2) est assurée par l'identité de Bézout dans le cas où les

H_m n'ont pas de zéros multiples. Les solutions G_m sont alors calculées suivant l'algorithme figurant dans [38].

3.2 Méthodes par suppression

Comme expliqué précédemment, cette catégorie de techniques ne nécessite pas de connaître la réponse impulsionnelle de la salle afin de déréverbérer. Les problèmes de stabilité induits dans les méthodes par annulation ne se posent pas ici, rendant cette approche plus robuste. C'est pourquoi, dans le cadre d'une implémentation en temps et conditions réels, nous préférerons les méthodes par suppression plutôt que par annulation.

3.2.1 Méthodes basées sur l'enveloppe temporelle

Une des premières différences que l'on observe entre un signal anéchoïque et réverbéré se fait sur la forme de l'enveloppe temporelle : la réverbération lisse les attaques et fins des mots, transformant une enveloppe bien définie avec de fortes variations d'amplitude en une enveloppe qui varie peu au cours du temps (cf. Fig. 2.3).

C'est pourquoi une des premières intuitions fut de retrouver l'enveloppe du signal anéchoïque, à partir du réverbéré, pour minimiser l'effet de la réverbération. Afin de travailler de façon précise, ce traitement a lieu par bande de fréquences, comme ce que l'on peut lire dans un brevet de Berkley et Mitchell datant de 1979 [7].

Une approche par déconvolution d'enveloppe a été développée par Mourjopoulos, en travaillant toujours par bandes de fréquences, afin de retrouver l'enveloppe du signal anéchoïque [40]. Ces travaux furent poursuivis par Hirobayashi, qui définit un filtre inverse permettant de retrouver l'enveloppe anéchoïque :

$$H_{\text{inv}}(z) = \sigma^2 \left(1 - e^{-\frac{13.8}{RT_{60}f_e}} z^{-1}\right)$$

On voit bien que ce filtre dépend du temps de réverbération à 60dB et du bruit blanc considéré dans le modèle stochastique de salle établi par Polack [44], puisqu'il met en jeu sa variance σ. Malheureusement, ils ne testèrent leur méthode que sur des signaux et réponses de salles synthétiques.

3.2.2 Méthodes basées sur le modèle de la voix

Introduit par Hardwick en 1993 [22], le modèle par excitation duale (*dual excitation speech model* en anglais) consiste à éliminer le bruit entre les M composantes harmoniques d'un signal de parole, relativement à une fréquence fondamentale f_0. Pour ce faire, on resynthétise M composantes harmoniques $\tilde{A}_m$ issues de l'analyse des amplitudes harmoniques $|A_m|$, en s'aidant d'un seuil de bruit N_{oise} sous la forme :

$$\tilde{A}_m = \begin{cases} 0 & \text{si } |A_m| < N_{\text{oise}} \\ A_m & \text{sinon} \end{cases}$$

Ce modèle a été généralisé dans [53] afin de suivre les variations de la fréquence fondamentale. Contrairement aux méthodes par soustraction spectrale qui avaient vu le jour jusqu'alors, cette méthode a l'avantage de générer moins de bruit musical [53].

Une version multicanale a été élaborée par Brandstein afin de généraliser les résultats à des réseaux de microphones [10] et l'adapter à la déréverbération. Ce modèle a été utilisé conjointement avec une approche probabiliste par Attias dans [5] afin de débruiter et déréverbérer les signaux de paroles.

3.2.3 Méthodes basées sur le *beamforming*

Les techniques de formation de faisceaux (*beamforming* en anglais) consistent à priviligier une direction d'arrivée de l'onde acoustique. En utilisant un réseau de microphones -on travaille forcément en multicanal- on obtient M versions du signal, prises à différents points de l'espace. Des filtres, principalement retard et égalisation fréquentielle, sont appliqués à ces M acquisitions qui seront ensuite sommées afin de donner un signal monocanal. En fonction des filtres appliqués, on amplifiera une certaine direction d'arrivée du son, permettant de "viser" une zone de l'espace à enregistrer, grâce aux ajouts constructifs ou destructifs des différents canaux. Ces filtres peuvent être fixes (on privilégie toujours la même direction), on parle alors de *beamforming* fixe ; ou bien peuvent s'adapter dynamiquement à l'espace sonore, on parle de *beamforming* adaptatif.

L'avantage de ce procédé dans le cadre de la déréverbération est que l'on peut ajouter de façon constructive l'onde du chemin direct et celle des premières réflexions, et ajouter de façon destructive les réflexions tardives. Cette approche, visant initialement à maximiser le rapport signal sur bruit, a été introduite par Affes dans [1] et a pour conséquence d'augmenter également le *Direct to Reverberant Ratio*, donc de déréverbérer.

3.2.4 Méthodes par prédiction linéaire

Ces méthodes se basent sur le modèle source-filtre de la parole : un bruit blanc (la source) est filtré par un filtre dont les coefficients varient au cours du temps (le conduit vocal) [27]. Puisque le conduit vocal ne se déforme pas en présence de réverbération, on suppose que les coefficients paramètrant le filtre non plus. Ces méthodes consistent donc à estimer les coefficients de ce filtre par prédiction linéaire, à partir du signal réverbéré, afin d'en extraire le résiduel par filtrage inverse. En ne gardant que les premiers pics du résiduel -ou en lissant les pics secondaires- il est possible de retrouver les impulsions du signal anéchoïque, et donc de le re-synthétiser.

Différents travaux se sont basés sur ces méthodes ([17], [51], [15]), pour finalement utiliser le modèle auto-régressif de la parole, dont les coefficients sont équivalents en situation anéchoïque ou réverbérée [14]. Nous verrons qu'il est également possible d'utiliser les méthodes par prédiction linéaire afin d'estimer l'énergie propre à la partie tardive de la réverbération. Cette dernière est ensuite retirée par soustraction spectrale ; c'est pourquoi ce type de méthode est également rattaché à la catégorie "*Méthode par réhaussement de la parole*" que nous allons à présent étudier.

3.2.5 Méthodes par réhaussement de la parole

Ces méthodes consistent à modifier directement la transformée de Fourier à court terme (TFCT) du signal réverbéré afin d'estimer le signal anéchoïque. On applique donc un gain $G(f,t)$ dépendant de la fréquence, pour chaque instant t, au spectre du signal réverbéré. Ces méthodes englobent donc celles par soustraction spectrale, où l'on vient retirer au spectre du signal à réhausser l'énergie propre à la réverbération, estimée au préalable. On obtient

un gain de type $G(k,l) = \left(1 - \frac{|\text{TFCT}_{\text{rev}}(k,l)|}{|\text{TFCT}_{\text{entrée}}(k,l)|}\right)$, pour chaque bin de fréquence k, à la trame l.

On peut trouver dans [33] une méthode par soustraction spectrale qui consiste à estimer l'énergie de réverbération tardive grâce au modèle stochastique de Polack [44], pour la retirer du spectre du signal réverbéré. Afin d'éviter le phénomène de bruit musical qui peut avoir lieu lorsque certaines "cases" du spectrogramme sont brutalement mises à zéro, Lebart introduit un seuil spectral (*spectral floor* en anglais).

En effet, en soustraction spectrale classique, si l'on cherche à retirer le spectre $R(k,l)$ du spectre $X(k,l)$, l'amplitude du spectre de sortie $|S(k,l)|$ sera de la forme [28] :

$$|S(k,l)| = \begin{cases} \sqrt{|X(k,l)|^2 - |R(k,l)|^2} & \text{si } |X(k,l)| \geq |R(k,l)| \\ 0 & \text{sinon} \end{cases}$$

Or ces fameux zéros produisent des irrégularités dans le spectre, produisant ce que l'on appelle le "bruit musical". Afin de palier à ce phénomène Lebart propose de remplacer les zéros par le spectre du signal d'entrée atténué d'un facteur λ, aux coordonnées (temps, fréquence) correspondantes. Le signal de sortie est alors calculé par transformée de Fourier inverse sur le spectre $S(k,l)$ en utilisant la phase du signal réverbéré. Nous expliquerons en détail comment estimer l'énergie de réverbération dans le **Chapitre 4**, puisque cette méthode figure parmi celles retenues dans la suite des travaux.

Une autre approche consiste à d'utiliser des notions de psychoacoustique afin d'éliminer la réverbération de façon plus adéquate au fonctionnement de l'oreille. Dans [49], Tsoukalas part du phénomène psychoacoustique de masquage ([54]) afin de déterminer des seuils de masquage du bruit. Sous ces seuils, à une fréquence et un temps donné, il est inutile de supprimer le bruit puisque ce dernier est masqué par la simple présence du discours à réhausser.

Xia conserva ces seuils de masquage dans [52] afin de définir un gain basé sur les propriétés de l'oreille humaine, limitant alors le bruit musical :

$$G(k,l) = \begin{cases} \sqrt{1 - \alpha(k,l)\frac{|\tilde{D}(k,l)|^2}{|Y(k,l)|^2}} & \text{si } \frac{|Y(k,l)|^2}{|\tilde{D}(k,l)|^2} > \alpha(k,l) + \beta(k,l) \\ \sqrt{\beta(k,l)\frac{|\tilde{D}(k,l)|^2}{|Y(k,l)|^2}} & \text{sinon.} \end{cases}$$

Avec $(\alpha(k,l), \beta(k,l))$ dépendant des seuils de masquage calculés, $Y(k,l)$ la TFCT du signal reçu et $\tilde{D}(k,l)$ la TFCT de la réverbération tardive estimée au préalable.

Kinoshita propose dans [28] une méthode par prédiction linéaire à long terme afin d'estimer l'énergie propre à la réverbération tardive. Après blanchiment du signal réverbéré, il généralise la prédiction linéaire usuelle (à un échantillon passé) à D échantillons passés. Cela permet d'imposer un délai $D * f_{\text{ech}}$, correspondant au temps mis avant de se trouver dans la partie tardive de la réverbération. Une fois que l'énergie propre à la réverbération tardive est estimée, on la retire par soustraction spectrale au signal réverbéré.

Cette technique, ainsi que le type de gain utilisé, sera détaillée dans le **Chapitre 4** puisqu'elle fait partie des méthodes implémentées durant le stage.

3.3 Méthodes par dictionnaires

Les méthodes par dictionnaires, en pleine expansion ces dernières années, ont aussi trouvé une application en déréverbération. Après des premiers travaux en traitement d'images à haute-résolution [50] ou en débruitage de signaux [48], des algorithmes basés sur l'utilisation de dictionnaires offrent une nouvelle approche à la déréverbération.

Pour le *Reverb Challenge 2014*, Moshirynia proposa une méthode en deux étapes afin de déréverbérer les signaux de parole [34]. Tout d'abord, il faut concaténer un dictionnaire de signaux anéchoïques avec un dictionnaire de signaux réverbérés. Une fois l'apprentissage fait de façon hors ligne (on minimise l'erreur en fixant successivement le dictionnaire et le code associé), on peut procéder à la déréverbération.

Pour cela, il faut tout d'abord déconvoluer le spectre afin de supprimer l'influence au long terme de la pièce, via des méthodes de déconvolution en matrices non-négatives. Une fois la déconvolution effectuée, on vient extraire le code permettant de reconstituer ce spectre dans le dictionnaire concaténé. L'algorithme *LARC* ([48]) nous donne le code correspondant, que l'on applique uniquement au dictionnaire de signaux anéchoïques afin d'obtenir la version déréverbérée du signal. Le dictionnaire est ensuite mis à jour par la méthode *K-SVD* [2].

L'auteur ajoute une étape de post-traitement afin d'améliorer la phase du signal estimé, au lieu de prendre la phase du signal réverbéré. Cette partie sera discutée plus en détail dans le **Chapitre 6**, puisqu'elle est en lien avec la suite de mes travaux.

3.4 Conclusion

Les méthodes figurant dans l'état de l'art peuvent bien souvent être utilisées aussi bien en monocanal qu'en multicanal, c'est pourquoi la dimension du réseau de microphone utilisé ne peut pas être un critère de classification. Il est donc préférable de différencier deux grands types de méthodes : celles basées sur la réponse impulsionnelle de la salle, méthodes par annulation ; celles basées sur la modification de la transformée de Fourier à court terme, méthodes par suppression.

Les méthodes par annulation permettent théoriquement de retrouver exactement le signal anéchoïque puisqu'elles consistent à inverser la réponse impulsionnelle h (qui modifia le signal anéchoïque s) afin de re-filtrer le signal réverbéré. Seulement, ces méthodes font appel à des inversions de matrices de grandes tailles, ce qui augmente la durée du traitement, et ne sont pas très robustes à un environnement réel. C'est pourquoi nous n'utiliserons pas cette famille de méthode dans le cadre d'une implémentation en temps et conditions réels.

Nous préférerons travailler avec les méthodes par suppression, bien plus robustes puisqu'elles utilisent l'amplitude du spectre du signal réverbéré afin d'estimer l'énergie tardive de la réverbération. Cette dernière est ensuite filtrée à l'aide de différents gains plus ou moins fidèles à la qualité du signal d'entrée. Nous reviendrons sur deux de ces méthodes dans le **Chapitre 4**, en mono- et multicanal : estimation par prédiction linéaire [28] ou par modèle stochastique des salles [44] ; réhaussement par suppression spectrale [33] ou par gain *Optimally-Modified Log-Spectral Amplitude* [11].

Enfin, des méthodes basées sur l'utilisation de dictionnaires commencent à voir le jour dans le cadre de la déréverbération. Elles consistent à apprendre conjointement des dictionnaires de signaux anéchoïques et réverbérés afin de coder un signal réverbéré. Le code utilisé est alors appliqué au dictionnaire de signaux anéchoïques afin d'obtenir le signal déréverbéré.

Chapitre 4

Méthodes implémentées

Ce chapitre est consacré aux principales méthodes de déréverbération implémentées durant le stage : estimation de la réverbération tardive par prédiction linéaire ou basée sur le modèle stochastique des salles ; élimination par soustraction spectrale ou *Optimally-Modified Log-Spectral Amplitude*. D'autres techniques figurant dans le **Chapitre 3** ([49], [52], [33]) ont également été implémentées, à titre d'essai. Elles n'ont pas été approfondies à cause des résultats jugés moins satisfaisants que ceux issus des techniques proposées dans ce chapitre.

4.1 Estimation de la réverbération tardive

Comme expliqué Eq. 2.6, on peut modéliser l'acquisition y d'une source s dans une salle de réponse impulsionnelle h, soumise à un bruit parasite b comme :

$$y(n) = h(0)s(n) + \sum_{k=1}^{\infty} h(k)s(n-k) + b(n)$$

Afin de faciliter les calculs nous supposons que nous travaillons dans un environnement non bruyant, c'est à dire $b(n) = 0\ \forall n \in \mathbb{N}$. Cette supposition n'est pas dérangeante puisque l'ajout du bruit se traduit par un simple biais corrigeable facilement dans l'estimation de la réverbération [20]. De plus, les algorithmes se situent après un annulateur de bruit présent dans les produits *Invoxia*.

On peut donc écrire :

$$y(n) = y_{\text{direct}}(n) + y_{\text{rev}}(n) \tag{4.1}$$

avec $y_{\text{direct}}(n) = h(0)s(n)$ correspondant au chemin direct de l'onde sonore et $y_{\text{rev}}(n) = \sum_{k=1}^{\infty} h(k)s(n-k)$ sa partie réverbérée. Or nous avons conclu dans le **Chapitre 2** que la partie précoce de la réverbération n'était pas gênante et qu'il fallait donc supprimer uniquement la partie tardive. C'est pourquoi nous séparons y_{rev} en deux parties, précoce et tardive tel que :

$$y_{\text{rev}}(n) = y_{\text{précoce}}(n) + y_{\text{tardive}}(n) \tag{4.2}$$

Avec

$$y_{\text{précoce}}(n) = \sum_{k=1}^{T_l} h(k)s(n-k) \tag{4.3}$$

$$y_{\text{tardive}}(n) = \sum_{k=T_l+1}^{\infty} h(k)s(n-k) \tag{4.4}$$

On rappelle que T_l correspond à l'échantillon à partir duquel, à la suite d'une impulsion, on se retrouve dans la partie tardive de la réverbération. Le but des méthodes de dérévebération tardive de la parole est donc d'estimer uniquement l'énergie du signal y_{tardive} afin de la retirer à celle du signal réverbéré.

4.1.1 Estimation basée sur le modèle stochastique de réponse de salles

Cette méthode est développée dans [20] et se base sur la généralisation du modèle de réponse de salles, introduit par Polack dans [44].

On considère donc que la réponse de la salle peut s'écrire comme :

$$h(t) = \begin{cases} b_d(t)e^{-\delta t} & \text{si } 0 \leq t < T_r \\ b_r(t)e^{-\delta t} & \text{si } t \geq T_r \\ 0 & \text{sinon} \end{cases} \tag{4.5}$$

avec T_r le temps incluant uniquement le chemin direct de l'onde acoustique, σ_d et σ_r les variances des bruits b_d et b_r.

Si on pose $h_d(t) = b_d(t)e^{-\delta t}$ et $h_r(t) = b_r(t)e^{-\delta t}$, le signal reçu peut s'écrire sous la forme :

$$y(t) = \int_{t-T_r}^{t} s(\theta)h_d(t-\theta)d\theta + \int_{-\infty}^{t-T_r} s(\theta)h_r(t-\theta)d\theta \tag{4.6}$$

Calculons sa densité spectrale de puissance via la fonction d'autocovariance $r_z(t, t+\tau) = \mathcal{E}_z\{z(t)z(t+\tau)\}$ [20] :

$$r_y(t, t+\tau) = \int_{-\infty}^{t}\int_{-\infty}^{t+\tau} \mathcal{E}_s\{s(\theta)s(\theta')\}\mathcal{E}_h\{h(t-\theta)h(t-\theta'+\tau)d\theta d\theta' \tag{4.7}$$

Or :

$$\mathcal{E}_{h_{d,r}}\{h_{d,r}(t-\theta)h_{d,r}(t-\theta'+\tau) = \delta_{\{\theta-\theta'+\tau\}}\sigma_{d,r}^2 e^{-2\delta t}e^{\delta(\theta+\theta'-\tau)}$$

avec $\delta_{\{.\}}$ le symbole de Kronecker.

On peut donc réécrire l'équation (4.7) en se servant de (4.6) sous la forme :

$$r_y(t, t+\tau) = r_{y_d}(t, t+\tau) + r_{y_r}(t, t+\tau) \tag{4.8}$$

Avec :

$$r_{y_d}(t, t+\tau) = \sigma_d^2 e^{-2\delta t} \int_{t-T_r}^{t} r_s(\theta, \theta+\tau) e^{2\delta\theta} d\theta \tag{4.9}$$

et

$$r_{y_r}(t, t+\tau) = \sigma_r^2 e^{-2\delta t} \int_{-\infty}^{t-T_r} r_s(\theta, \theta+\tau) e^{2\delta\theta} d\theta \tag{4.10}$$

Si on scinde l'expression de r_{y_r} en faisant apparaitre à nouveau le temps T_r, on peut écrire r_{y_r} en fonction de r_{y_d} :

$$r_{y_r}(t, t+\tau) = \underbrace{\sigma_r^2 e^{-2\delta t} \int_{-\infty}^{t-2T_r} r_s(\theta, \theta+\tau) e^{2\delta\theta} d\theta}_{r_{y_r}(t-T_r, t-T_r+\tau) e^{-2\delta T_r}} + \underbrace{\sigma_r^2 e^{-2\delta t} \int_{t-2T_r}^{t-T_r} r_s(\theta, \theta+\tau) e^{2\delta\theta} d\theta}_{\frac{\sigma_r^2}{\sigma_d^2} r_{y_d}(t-T_r, t-T_r+\tau) e^{-2\delta T_r}} \tag{4.11}$$

En posant $\kappa = \frac{\sigma_r^2}{\sigma_d^2}$ et en utilisant (4.8), on obtient une estimation de la fonction d'autocovariance de la partie réverbérée de y :

$$r_{y_r}(t, t+\tau) = (1-\kappa) e^{-2\delta T_r} r_{y_r}(t-T_r, t-T_r+\tau) + \kappa e^{-2\delta T_r} r_y(t-T_r, t-T_r+\tau) \tag{4.12}$$

Maintenant que nous avons une estimation de la fonction d'autocovariance de y_r il nous faut extraire la partie tardive, correspondant à y_{tardive}. On rappelle que l'on nomme T_l l'instant à partir duquel nous nous trouvons dans la partie tardive de la réverbération. D'après (4.10) on a :

$$r_{y_r}(t, t+\tau) = \underbrace{\sigma_r^2 e^{-2\delta t} \int_{t-T_l}^{t-T_r} r_s(\theta, \theta+\tau) e^{2\delta\theta} d\theta}_{\text{Partie précoce}} + \underbrace{\sigma_r^2 e^{-2\delta t} \int_{-\infty}^{t-T_l} r_s(\theta, \theta+\tau) e^{2\delta\theta} d\theta}_{\text{Partie tardive}} \tag{4.13}$$

Mais aussi :

$$r_{y_r}(t-T_l+T_r, t-T_l+T_r+\tau) = e^{2\delta(T_l-T_r)} \sigma_r^2 e^{-2\delta t} \int_{-\infty}^{t-T_l+T_r-T_r} r_s(\theta, \theta+\tau) e^{2\delta\theta} d\theta \tag{4.14}$$

On voit bien que la partie tardive de (4.13) $r_{y_{\text{tardive}}}$ peut s'écrire sous la forme :

$$r_{y_{\text{tardive}}}(t, t+\tau) = e^{-2\delta(T_l-T_r)} r_{y_r}(t-T_l+T_r, t-T_l+T_r+\tau) \tag{4.15}$$

Si on travaille avec des transformées de Fourier à court terme, avec un avancement de trame de taille R, on peut approximer le temps T_r à $\frac{R}{f_e}$ puisque c'est la plus petite discrimination temporelle disponible. La transformée de Fourier de la fonction d'autocovariance

permet d'estimer la densité spectrale de puissance (notée λ_z pour un signal z), donc si on applique une transformée de Fourier à court terme, d'avancement R, aux équations (4.12) et (4.15) on obtient :

$$\lambda_{y_r}(k,l) = (1-\kappa)e^{-2\delta\frac{R}{f_e}}\lambda_{y_r}(k,l-1) + \kappa e^{-2\delta\frac{R}{f_e}}\lambda_y(k,l-1) \tag{4.16}$$

$$\lambda_{y_{\text{tardive}}}(k,l) = e^{-2\delta(T_l-\frac{R}{f_e})}\lambda_{y_r}(k,l-N_l+1) \tag{4.17}$$

avec k le bin de fréquence considéré, l la trame considérée et $N_l = T_l f_e$ le nombre d'échantillons correspondants à T_l. Il est important de rappeler que δ et κ dépendent également du bin de fréquence k considéré, nous ometons de signaler cette dépendance afin d'alléger la notation. Habets propose dans [20] une méthode afin d'estimer κ trame par trame, avec un pas d'adaptation μ_κ :

$$\hat{\kappa}(l+1) = \begin{cases} \hat{\kappa}(l) + \mu_\kappa \left(1 - \frac{\sum_{k=1}^{\frac{N_{\text{fft}}}{2}} \lambda_{y_r}(k,l)}{\sum_{k=1}^{\frac{N_{\text{fft}}}{2}} |Y(k,l)|^2(k,l)}\right) & \text{si il y a de la parole} \\ \hat{\kappa}(l) & \text{sinon} \end{cases} \tag{4.18}$$

Nous avons donc à disposition une première méthode pour estimer la densité spectrale de puissance de la réverbération tardive du signal y. Cette dernière peut s'adapter simplement à une configuration multicapteurs, améliorant alors les résultats.

Extension multicanal

Habets montra dans [19] qu'il était possible d'améliorer l'estimation en moyennant les spectres des M acquisitions disponibles, si ces dernières étaient alignées dans le temps.

C'est à dire que là où on avait $\lambda_y(k,l) = \beta\lambda_y(k,l-1) + (1-\beta)|Y(k,l)|^2$ en monocanal [20], on utilise désormais :

$$\lambda_y(k,l) = \beta\lambda_y(k,l-1) + (1-\beta)\frac{1}{M}\sum_{m=1}^{M}|Y_m(k,l)|^2 \tag{4.19}$$

où $Y_m(k,l)$ est la transformée de Fourier à court terme de $y_m(t)$, acquisition de $y(t)$ avec le m$^{\text{ième}}$ microphone, et β un paramètre de lissage.

Étudions à présent une autre approche afin d'estimer l'énergie propre à la réverbération tardive, basée cette fois-ci sur le principe de prédiction linéaire.

4.1.2 Estimation basée sur la prédiction linéaire à long terme

Cette méthode a été développée par Kinoshita dans [28], elle se décompose en deux étapes : blanchiment du signal, calcul des coefficients de prédiction. L'utilité du blanchiment se comprend dans le calcul des coefficients de prédiction, c'est pourquoi nous commencerons par détailler ce dernier avant d'étudier le blanchiment.

Prédiction linéaire à long terme

La production d'un signal de voix se base sur le modèle source-filtre [27] comme expliqué dans la section 3.2.4 : un bruit blanc b (la source) est filtré par un filtre à réponse finie d'ordre P a (le conduit vocal) afin de générer un signal de parole s. Ce signal est modifié par la réponse impulsionnelle h de la pièce dans laquelle il est émis, pour être capté par un microphone sous la forme y.

On a donc :

$$y(n) = \sum_i h(i)s(n-i) = \sum_i h(i) \sum_{k=0}^{P} a(k)b(n-i-k) \tag{4.20}$$

Si on effectue le changement de variable $l = i + k$ on peut réécrire (4.20) sous la forme :

$$y(n) = \sum_i \left(\sum_{k=0}^{P} h(l-k)a(k) \right) b(n-l)$$

Alors, en posant

$$g(l) = \sum_{k=0}^{P} h(l-k)a(k) \tag{4.21}$$

on obtient :

$$y(n) = \sum_l g(l)b(n-l) \tag{4.22}$$

On peut alors reformuler (4.22) sous forme de matrice, si on considère une succession de $N+1$ échantillons de y on obtient :

$$\mathbf{y}(n) = \mathbf{G}\mathbf{b}(n) \tag{4.23}$$

Avec :

$$\mathbf{y}(n) = [y(n), y(n-1), \dots, y(n-N)]^T \tag{4.24}$$

$$\mathbf{G} = \begin{bmatrix} g(0) & g(1) & \dots & g(T-1) & 0 & \dots & \dots & 0 \\ 0 & g(0) & g(1) & \dots & g(T-1) & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & g(0) & g(1) & \dots & g(T-1) \end{bmatrix} \tag{4.25}$$

$$\mathbf{b}(n) = [b(n), b(n-1), \dots, b(n-(T+N-1))]^T \tag{4.26}$$

La matrice $\mathbf{G}$ est de taille $(N+1) \times (N+T)$ et peut s'écrire plus simplement, en posant $\mathbf{g} = [g(0), g(1), \dots, g(T-1)]$ sous la forme :

$$\mathbf{G} = \begin{bmatrix} \mathbf{g} & 0 & \dots & \dots & 0 \\ 0 & \mathbf{g} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & \mathbf{g} \end{bmatrix} \tag{4.27}$$

Si on s'intéresse maintenant au modèle de prédiciton linéaire à long terme, tel qu'il a été proposé initialement dans [16], on peut écrire :

$$y(n) = \sum_{p=0}^{N} w(p)y(n-p-D) + e(n) \tag{4.28}$$

avec N l'ordre de la prédiction linéaire, e l'erreur de prédiction et D le délai introduit dans la prédiction linéaire. Ce délai correspond donc, dans notre cas, au nombre d'échantillons à partir duquel on se trouve dans la partie tardive de la réverbération. On définit à cette occasion $\mathbf{g}_{\text{late}}$, la version tronquée de $\mathbf{g}$ et complétée avec des zéros, tel que :

$$\mathbf{g}_{\text{late}} = [g(D), g(D+1), \dots, g(T-1), \underbrace{0, \dots, 0}_{(T-D+1)}]^T$$

Les coefficients de $\mathbf{w} = [w(0), \dots, w(N)]^T$ peuvent être obtenus en minimisant l'erreur de prédiction e. La minimisation de la moyenne de l'énergie de l'erreur mène à l'équation suivante [28] :

$$\mathcal{E}\{\mathbf{y}(n-D)\mathbf{y}^T(n-D)\}\mathbf{w} = \mathcal{E}\{\mathbf{y}(n-D)y(n)\}$$

On en déduit donc :

$$\mathbf{w} = \left(\mathcal{E}\{\mathbf{y}(n-D)\mathbf{y}^T(n-D)\}\right)^{-1} \mathcal{E}\{\mathbf{y}(n-D)y(n)\} \tag{4.29}$$

En utilisant (4.23) on peut exprimer la première partie de (4.29) à l'aide de $\mathbf{G}$:

$$\mathcal{E}\{\mathbf{y}(n-D)\mathbf{y}^T(n-D)\} = \mathcal{E}\{\mathbf{G}\mathbf{b}(n-D)\mathbf{b}^T(n-D)\mathbf{G}^T\} = \mathbf{G}\mathcal{E}\{\mathbf{b}(n-D)\mathbf{b}^T(n-D)\}\mathbf{G}^T$$

Si on pose σ la variance du bruit blanc b on obtient :

$$\mathcal{E}\{\mathbf{y}(n-D)\mathbf{y}^T(n-D)\} = \sigma^2\mathbf{G}\mathbf{G}^T \tag{4.30}$$

La seconde partie de (4.29) peut s'exprimer en fonction de $\mathbf{G}$, mais aussi de $\mathbf{g}$:

$$\mathcal{E}\{\mathbf{y}(n-D)y(n)\} = \mathcal{E}\{\mathbf{G}\mathbf{b}(n-D)\mathbf{b}^T(n)\mathbf{g}^T\} = \mathbf{G}\mathcal{E}\{\mathbf{b}(n-D)\mathbf{b}^T(n-D)\mathbf{g}_{\text{late}}^T\}$$

On obtient donc :

$$\mathcal{E}\{\mathbf{y}(n-D)y(n)\} = \sigma^2\mathbf{G}\mathbf{g}_{\text{late}}^T \tag{4.31}$$

On peut donc exprimer $\mathbf{w}$ en remplaçant (4.30) et (4.31) dans (4.29), tel que :

$$\mathbf{w} = \left(\mathbf{G}\mathbf{G}^T\right)^{-1} \mathbf{G}\mathbf{g}_{\text{late}}^T \tag{4.32}$$

Montrons à présent qu'il est possible d'estimer l'énergie de réverbération tardive en filtrant chaque portion de signal $\mathbf{y}$ par le filtre de prédiction $\mathbf{w}$, ce qui se traduit par l'opération matricielle "$\mathbf{y}^T(n)\mathbf{w}$". Calculons alors l'énergie de ce signal :

$$\begin{aligned}\mathcal{E}\{\left(\mathbf{y}^T(n)\mathbf{w}\right)^2\} &= \mathcal{E}\{\left(\mathbf{b}^T(n)\mathbf{G}^T\mathbf{w}\right)^2\} = \mathcal{E}\{\mathbf{w}^T\mathbf{G}\mathbf{b}(n)\mathbf{b}^T(n)\mathbf{G}^T\mathbf{w}\} \\ &= \sigma^2\mathbf{w}^T\mathbf{G}\mathbf{G}^T\mathbf{w} = \sigma^2\mathbf{g}_{\text{late}}\mathbf{G}^T\left(\mathbf{G}\mathbf{G}^T\right)^{-1}\mathbf{G}\mathbf{G}^T\left(\mathbf{G}\mathbf{G}^T\right)^{-1}\mathbf{G}\mathbf{g}_{\text{late}}^T \\ &= \sigma^2\mathbf{g}_{\text{late}}\mathbf{G}^T\left(\mathbf{G}\mathbf{G}^T\right)^{-1}\mathbf{G}\mathbf{g}_{\text{late}}^T \\ &= ||\sigma\mathbf{g}_{\text{late}}||^2\end{aligned}$$

On voit qu'une estimation de l'énergie de réverbération tardive est possible en appliquant les coefficients de prédiction linéaire à long terme au signal réverbéré. Seulement, on obtient une énergie correspondant à la réponse g et non pas h. Le biais entre g et h est causé par le filtre a du conduit vocal (cf. (4.21)). Afin de réduire ce biais dans l'estimation de la réverbération tardive, un pré-traitement est effectué : le blanchiment.

Blanchiment

Pre-whitening en anglais, cette technique consiste à "blanchir" un signal, le rendre large bande. Cela permet donc de réduire la corrélation à court terme du filtre de parole a, en retrouvant un signal proche du bruit blanc modélisé comme source du modèle source-filtre.

Une solution pour blanchir un signal est le filtrage adapté. Si on considère un signal modélisé par un processus auto-régressif d'ordre P, dont les coefficients constituent un filtre a (cf. (4.28) avec $D = 1$ et en replaçant w par a), le filtrage adapté consiste à filtrer de nouveau ce signal par un filtre de réponse impulsionnelle $a(-n)$, version "retournée" de a. On garde ensuite le résiduel afin d'éliminer toute la partie prédictive du signal d'entrée, et ainsi réduire son autocorrélation. Les coefficients de a peuvent être estimés par résolution des équations de Yule-Walker [35]. On peut voir la différence entre les spectres de la figure 4.1 : l'énergie est presque identiquement distribuée une fois le signal blanchi, que ce soit pour les consonnes ou les voyelles.

Extension multicanal

Cette méthode peut également s'adapter à un réseau de M microphones, mettant à notre disposition M acquisitions y_m du signal source, à différents endroits de l'espace.

Le raisonnement est identique, tout comme l'algorithme général. Seulement cette fois-ci au lieu d'avoir un unique signal y, dont le modèle de prédiction à long terme s'appliquait suivant (4.28), on a M signaux y_m dont le i$^{\text{ème}}$ peut s'exprimer en fonction des autres :

$$y_i(n) = \sum_{m=1}^{M}\sum_{p=0}^{N} w_{m,i}(p)y_m(n-p-D) + e_i(n) \tag{4.33}$$

avec D correspondant au délai avant de se trouver dans la partie tardive de la réverbération, e_i l'erreur de prédiction associée au modèle de y_i. Les coefficients $w_{m,i}$ correspondent

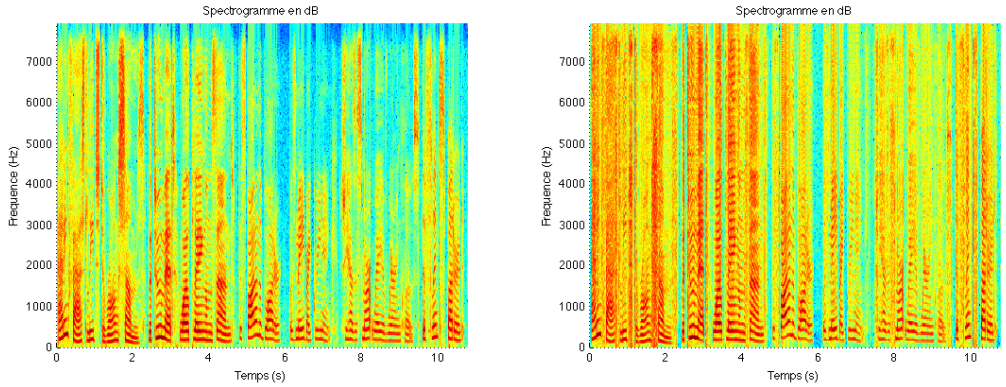

FIGURE 4.1 – Signal original à gauche, blanchi à droite avec un filtre de prédiction linéaire d'ordre 20

aux coefficients de prédiction au m$^{\text{ième}}$ microphone, lorsque l'on cherche à prédire y_i. Pour chaque acquisition i on associe un vecteur de prédiction $\mathbf{w}_i$ de taille $M \times (N+1)$ composé des coefficients de prédiction de y_i pour chaque microphone :

$$\mathbf{w}_i = [w_{1,i}(0), \dots, w_{1,i}(N), w_{2,i}(0), \dots, w_{2,i}(N), \dots, w_{M,i}(0), \dots, w_{M,i}(N)]^T$$

La résolution de $\mathbf{w}$ s'effectue comme dans (4.29), avec cette fois-ci un vecteur $\mathbf{y}$ de taille $M \times (N+1)$ contenant les $N+1$ coefficients antérieurs pour les M microphones disponibles :

$$\mathbf{y}(n) = [y_1(n), \dots, y_1(n-N), y_2(n), \dots, y_2(n-N), \dots, y_M(n), \dots, y_M(n-N)]^T$$

On peut alors estimer la réverbération tardive pour chaque canal, et ainsi les déréverbérer. On somme ensuite les M signaux obtenus, en les réalignant temporellement, afin d'avoir notre signal de sortie.

Nous avons étudié deux méthodes différentes afin d'estimer l'énergie propre à la réverbération tardive d'un signal de parole. Observons à présent les techniques utilisées afin de la retirer du signal réverbéré.

4.2 Déréverbération

La déréverbération s'effectue en retirant du spectre du signal réverbéré la partie propre à la réverbération tardive, estimée par une des méthodes figurant section 4.1. On peut considérer que les réflexions précoces et tardives de la réverbération sont statistiquement indépendantes [20], nous permettant de traiter le signal de réverbération tardive comme un bruit additif, décorrélé du signal à estimer.

Nous allons étudier deux types de gain : un premier proche de la soustraction spectrale utilisée par Lebart dans [33], un second basé sur l'amplitude logarithmique des signaux ainsi que sur la probabilité de présence de parole.

4.2.1 Soustraction spectrale

Supposons avoir estimé la densité spectrale de puissance de la réverbération tardive $\lambda_{y_{\text{tardive}}}$ par la méthode basée sur le modèle stochastique de salles, ou bien par prédiction linéaire. Supposons également que notre signal comporte un bruit additif, de densité spectrale de puissance λ_b, estimé suivant une technique que nous ne développerons pas. Le gain de soustraction spectrale appliqué à la transformée de Fourier à court terme du signal reçu $Y(k,l)$, pour chaque bin de fréquence k et trame l, peut s'écrire sous la forme :

$$G(k,l) = \left(1 - \left(\frac{1}{\gamma(k,l)}\right)^{\beta_1}\right)^{\beta_2}$$

avec $\gamma(k,l)$ le rapport signal sur interférence :

$$\gamma(k,l) = \frac{|Y(k,l)|^2}{\lambda_{y_{\text{tardive}}}(k,l) + \lambda_b(k,l)} \tag{4.34}$$

En fonction du jeu de paramètres (β_1, β_2) on peut parler de soustraction d'amplitude $(\beta_1 = \frac{1}{2}\,;\, \beta_2 = 1)$, de puissance $(\beta_1 = 1\,;\, \beta_2 = \frac{1}{2})$, ou encore filtre de Wiener $(\beta_1 = \beta_2 = 1)$. Habets montra dans [20] que le gain par soustraction d'amplitude donnait de meilleurs résultats. On travaillera avec un gain proche de celui utilisé par Boll ([8]), de la forme :

$$G(k,l) = 1 - \frac{1}{\sqrt{\gamma(k,l)}}$$

avec $\gamma(k,l)$ défini suivant (4.34). On obtient alors la transformée de Fourier à court terme du signal déréverbéré $Z(k,l)$ par filtrage fréquentiel de l'acquisition $Y(k,l)$:

$$Z(k,l) = G(k,l)Y(k,l) \tag{4.35}$$

Le rapport signal sur interférence peut être exprimé en fonction de la transformée à court terme du signal déréverbéré en écrivant :

$$G(k,l) = 1 - \frac{1}{\sqrt{\xi(k,l)+1}} \tag{4.36}$$

Avec :

$$\xi(k,l) = \gamma(k,l) - 1 = \frac{\mathcal{E}\{|Z(k,l)|^2\}}{\lambda_{y_{\text{tardive}}}(k,l) + \lambda_b(k,l)} \tag{4.37}$$

Or, $Z(k,l)$ n'est pas disponible à la trame l puisque c'est ce que l'on cherche à estimer à partir de $Y(k,l)$. En revanche on dispose de $Z(k,l-1) = G(k,l-1)Y(k,l-1)$, c'est pourquoi Ephraim et Malah proposèrent dans [13] une méthode afin d'estimer $\xi(k,l)$ à la trame l :

$$\hat{\xi}(k,l) = \eta\frac{|Z(k,l-1)|^2}{\lambda_{y_{\text{tardive}}}(k,l-1) + \lambda_b(k,l-1)} + (1-\eta)\max\left(\gamma(k,l) - 1, 0\right) \tag{4.38}$$

avec η un paramètre contrôlant le rapport déréverbération/distorsion, compromis systématiquement rencontré en déréverbération [20]. Plus cette valeur est élevée plus on déréverbera mais plus on distordra le signal.

Comme expliqué brièvement section 3.2.5, si l'estimation de $(\lambda_{y_{\text{tardive}}} + \lambda_b)$ dépasse $|Y|^2$ durant la trame l, à la fréquence k, on se retrouve avec une valeur négative de $|Z(k,l)|$, ce qui n'est évidemment pas possible. Pour éviter ce phénomène on pose une condition sur le signe de $|Y(k,l)|^2 - (\lambda_{y_{\text{tardive}}} + \lambda_b)$, afin de fixer à 0 l'amplitude de sortie $|Z(k,l)|$ lorsque ce scénario apparait.

Seulement, des variations brutales du spectre de z provoquent un sifflement peu naturel appelé "bruit musical". Afin d'éviter ce problème -extrêmement gênant dans le cadre de communications téléphoniques, moindre dans des systèmes de reconnaissance automatique de la parole- il est possible d'utiliser un seuil, développé par Lebart dans [33]. Lorsque la situation "$|Y(k,l)|^2 - (\lambda_{y_{\text{tardive}}}(k,l) + \lambda_b(k,l)) \leq 0$" apparaît, on impose $Z(k,l) = \alpha Y(k,l)$, avec α un paramètre d'atténuation (généralement inférieur à 10%).

Le gain de soustraction spectrale utilisé se résume donc à :

$$G(k,l) = \begin{cases} 1 - \dfrac{1}{\sqrt{\hat{\xi}(k,l)+1}} & \text{si } |Y(k,l)|^2 - (\lambda_{y_{\text{tardive}}}(k,l) + \lambda_b(k,l)) > 0 \\ \alpha & \text{sinon} \end{cases} \tag{4.39}$$

Le signal de sortie est alors obtenu en filtrant le signal d'entrée y suivant (4.35) afin d'obtenir Z. On récupère ensuite le signal déréverbéré temporel en appliquant une transformée de Fourier inverse sur Z, avec addition-recouvrement.

Extension multicanal

L'apport d'un réseau de microphones dans l'utilisation de ce gain se fait sur le calcul du rapport signal sur interférence. On peut remplacer $|Y(k,l)|^2$ dans (4.34), par la moyenne $|Y_{\text{moy}}(k,l)|^2$ des M spectrogrammes Y_m acquis par les différents microphones, après avoir été tous alignés temporellement [19] :

$$|Y_{\text{moy}}(k,l)|^2 = \frac{1}{M} \sum_{m=1}^{M} |Y_m(k,l)|^2$$

Ce moyennage permet d'améliorer les résultats obtenus, en déréverbérant mieux et en générant moins de bruit grâce au lissage des spectrogrammes. C'est ce que nous constaterons **Chapitre 5** dans l'analyse des résultats obtenus grâce à ces méthodes.

4.2.2 Gain OM-LSA

Dans [13] Ephraim et Malah proposent une méthode afin de restaurer un signal dégradé par du bruit, basé sur un modèle statistique Gaussien. Ils considèrent que chaque suite de coefficients de Fourier d'un signal est statistiquement indépendante, modélisée par une variable aléatoire Gaussienne. La moyenne des coefficients est supposée nulle puisque le processus impliqué est de moyenne nulle ; leur variance varie au cours du temps puisque les signaux de paroles sont non-stationnaires.

Le gain *Log Spectral Amplitude* (LSA) minimise alors l'erreur quadratique moyenne du logarithme du spectre à estimer, à savoir :

$$\mathcal{E}\left\{ \left(\log(A(k,l)) - \log(\hat{A}(k,l)) \right)^2 \right\}$$

avec $A(k,l) = |Z(k,l)|$ l'amplitude de la transformée de Fourier à court terme du signal déréverbéré, au bin de fréquence k et à la trame l, $\hat{A}(k,l)$ son estimation.

À partir du modèle statistique Gaussien des coefficients de Fourier, on définit le gain minimisant l'erreur quadratique moyenne par :

$$G_{\text{LSA}} = \frac{\xi(k,l)}{1+\xi(k,l)} \exp\left(\frac{1}{2}\int_{\zeta(k,l)}^{\infty} \frac{e^{-t}}{t}\mathrm{d}t\right) \tag{4.40}$$

avec $\xi(k,l)$ et $\gamma(k,l)$ définis comme dans (4.37) et (4.34). La variable $\zeta(k,l)$ s'exprime en fonction de ces deux derniers sous la forme :

$$\zeta(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)}\gamma(k,l) \tag{4.41}$$

Afin d'adapter ce gain de façon optimale en fonction de la présence ou non de parole, d'où le nom *Optimally-Modified Log Spectral Amplitude* (OM-LSA) en anglais, on lui ajoute une probabilité de présence de parole.

Là où il y aura de la parole dans le signal, on filtrera avec le gain $G_1 = G_{\text{LSA}}$, là où n'y en aura pas on filtrera avec $G_2 = \alpha$. Le gain utilisé durant les absences de parole correspond à celui utilisé dans (4.39) lorsque l'on surestime les densités spectrales de puissance de la réverbération et du bruit. Cela correspond donc à restituer le spectre atténué du signal d'entrée, permettant de limiter bruits et artéfacts.

Si on pose $p(k,l)$ la probabilité de présence de parole dans le bin de fréquence k à la trame l de la transformée de Fourier à court terme du signal d'entrée, le gain OM-LSA se définie alors comme une pondération des gains G_1 et G_2 :

$$G_{\text{OM-LSA}}(k,l) = G_1(k,l)^{p(k,l)} G_2(k,l)^{1-p(k,l)} \tag{4.42}$$

Il faut alors se référer aux travaux de Cohen dans [11] afin de calculer la probabilité de présence $p(k,l)$.

Probabilité de présence de parole

Afin de discerner les situations parlées des zones de silence, Cohen définit deux hypothèses $H_0(k,l)$ et $H_1(k,l)$ associées respectivement à l'absence ou présence de parole [11]. Le probabilité de présence de parole $p(k,l)$ au bin de fréquence k et à la trame l du spectrogramme $Y(k,l)$ d'entrée se traduit par :

$$p(k,l) = P\left(H_1(k,l)|Y(k,l)\right)$$

Or, les coefficients de la transformée de Fourier à court terme sont supposés suivre une distribution Gaussienne [13], on peut donc écrire :

$$P\left(Y(k,l)|H_0(k,l)\right) = \frac{1}{\pi\lambda_b(k,l)} \exp\left\{-\frac{|Y(k,l)|^2}{\lambda_b(k,l)}\right\} \tag{4.43}$$

$$P\left(Y(k,l)|H_1(k,l)\right) = \frac{1}{\pi(\lambda_p(k,l)+\lambda_b(k,l))} \exp\left\{-\frac{|Y(k,l)|^2}{\lambda_p(k,l)+\lambda_b(k,l)}\right\} \tag{4.44}$$

avec $\lambda_b(k,l)$ et $\lambda_p(k,l)$ respectivement la densité spectrale de puissance du bruit et du signal de parole. En appliquant le théorème de Bayes on peut écrire :

$$p(k,l) = \left\{1 + \frac{q(k,l)}{1-q(k,l)}(1+\xi(k,l))\exp(-\zeta(k,l))\right\}^{-1} \tag{4.45}$$

avec $q(k,l) = P\left(H_0(k,l)\right)$, $\xi(k,l)$ et $\zeta(k,l)$ défini respectivement dans (4.37) et (4.41).

Reste alors à calculer $q(k,l)$. Pour cela, Cohen définit dans [11] trois probabilités différentes $P_{\text{local}}, P_{\text{global}}$ et P_{trame} tels que :

$$q(k,l) = 1 - P_{\text{local}}(k,l)P_{\text{global}}(k,l)P_{\text{trame}}(l) \tag{4.46}$$

Ces paramètres se basent sur la forte corrélation de la présence de parole dans les bins de fréquence et trames voisines. On définit alors Θ comme étant une version du rapport signal sur bruit, lissée à l'aide d'un paramètre β :

$$\Theta(k,l) = \beta\Theta(k,l-1) + (1-\beta)\xi(k,l-1)$$

À partir de cette fonction lissée temporellement, on procède également à un lissage fréquentiel afin de bien représenter la corrélation temporelle et fréquentielle. Le moyennage fréquentiel peut se faire à l'aide de deux fenêtres de tailles différentes : une de taille $2\omega_{\text{local}}+1$ donnant lieu à P_{local}, une plus grande de taille $2\omega_{\text{global}}+1$ donnant lieu à P_{global}.

On a donc deux versions de $\Theta(k,l)$ en fonction de la taille de la fenêtre h_λ utilisée :

$$\Theta_\lambda(k,l) = \sum_{i=-\omega_\lambda}^{\omega_\lambda} h_\lambda(i)\Theta(k-i,l)$$

avec λ correspondant à "local" ou "global". On calcul ensuite les probabilités $P_\lambda(k,l)$ en fonction des seuils $\Theta_{\min}$ et $\Theta_{\max}$ que l'on se fixe :

$$P_\lambda(k,l) = \begin{cases} 0 & \text{si } \Theta_\lambda(k,l) \leq \Theta_{\min} \\ 1 & \text{si } \Theta_\lambda(k,l) \geq \Theta_{\max} \\ \dfrac{\log(\Theta_\lambda(k,l)/\Theta_{\min})}{\log(\Theta_{\max}/\Theta_{\min})} & \text{sinon} \end{cases} \tag{4.47}$$

Afin d'avoir un paramètre permettant d'atténuer le bruit dans les trames ne comportant que du bruit on définit Θ_{trame} ne dépendant que de la trame de la transformée de Fourier à court terme dans laquelle nous travaillons :

$$\Theta_{\text{trame}}(l) = \text{mean}_{1\leq k\leq N_{\text{fft}}/2+1}\{\Theta(k,l)\} \tag{4.48}$$

avec N_{fft} le nombre de points sur lesquels la transformée de Fourier est calculée. Cohen livre dans [11] un pseudo code afin de calculer $P_{\text{trame}}(k,l)$:

if $\Theta_{\text{trame}}(l) > \Theta_{\min}$ **then**
  if $\Theta_{\text{trame}}(l) > \Theta_{\text{trame}}(l-1)$ **then**
    $P_{\text{trame}}(l) = 1$
    $\Theta_{\text{pic}}(l) = \min\{\max\{\Theta_{\text{trame}}(l), \Theta_{\text{trame min}}\}, \Theta_{\text{trame max}}\}$

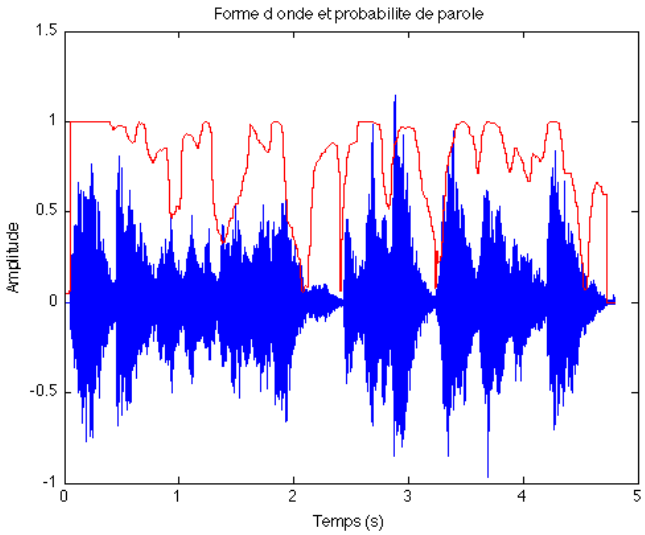

FIGURE 4.2 – Forme d'onde en bleu et probabilité de présence de parole en rouge

else
$P_{\text{trame}}(l) = \mu(l)$
end if
else
$P_{\text{trame}}(l) = 0$
end if
avec :

$$\mu(l) = \begin{cases} 0 & \text{si } \Theta_{\text{trame}}(l) \leq \Theta_{\text{pic}}(l)\Theta_{\min} \\ 1 & \text{si } \Theta_{\text{trame}}(l) \geq \Theta_{\text{pic}}(l)\Theta_{\max} \\ \dfrac{\log(\Theta_{\text{trame}}(l)/\Theta_{\text{pic}}(l)/\Theta_{\min})}{\log(\Theta_{\max}/\Theta_{\min})} & \text{sinon} \end{cases} \tag{4.49}$$

On constate bien Fig. 4.2 que la probabilité de présence de parole suit globalement l'enveloppe temporelle du signal de voix, en chutant lors des pauses.

Extension multicanal

L'utilisation de ce gain dans une configuration avec un réseau de M microphones ajoute deux modifications à la description précédente. La première consiste à moyenner les M spectrogrammes obtenus après avoir réaligné temporellement les signaux. Cela permet de lisser le signal à traiter et diminuer la distorsion [20].

Mais le réel avantage du multicanal dans l'utilisation de ce gain est l'ajout d'une probabilité supplémentaire dans l'expression de q, nommée P_{spat}. Comme son nom le laisse sous-entendre, cette probabilité se base sur l'information spatiale produite par le réseau de microphones, inaccessible en mono-capteur.

Au même titre qu'une forte corrélation dans les bins de fréquences et trames voisines pouvait indiquer la présence de parole, une forte corrélation entre deux signaux de micros

différents permet d'indiquer une présence de parole. Si les acquisitions sont réalignées dans le temps, les pics de corrélation entre deux canaux permettent même de détecter le chemin direct de l'onde acoustique puisque les réflexions arrivent de façon non synchrones, les microphones n'étant pas aux mêmes emplacements.

Habets développe dans [21] le calcul de cette nouvelle probabilité spatiale $P_{\text{spat}}(k,l)$, au bin de fréquence k et à la trame l de la transformée de Fourier à court terme du signal entrant. Pour cela il fait intervenir une fonction de cohérence quadratique moyenne (*Mean Square Coherence*, MSC en anglais) :

$$\Phi_{\text{MSC}}(k,l) = \frac{2!(M-2)!}{M!} \sum_{i=0}^{M-1} \sum_{j=i+1}^{M-1} \frac{\mathcal{S}\{Y_{j,i}(k,l)\}}{\mathcal{S}\{Y_i(k,l)\}\mathcal{S}\{Y_j(k,l)\}} \tag{4.50}$$

avec $Y_{j,i}(k,l) = Y_j(k,l)Y_i(k,l)^*$ et $\mathcal{S}\{.\}$ un opérateur de lissage temporel de paramètre β : $\mathcal{S}\{Z(k,l)\} = \beta\mathcal{S}\{Z(k,l-1)\} + (1-\beta)|Z(k,l)|^2$.

Une fois qu'on a calculé cette fonction de cohérence, on la lisse fréquentiellement à l'aide d'une fenêtre h_{spat} de taille $2\omega_{\text{spat}}+1$ afin d'obtenir $\tilde{\Phi}_{\text{MSC}}(k,l)$:

$$\tilde{\Phi}_{\text{MSC}}(k,l) = \sum_{i=-\omega_{\text{spat}}}^{\omega_{\text{spat}}} h_{\text{spat}}(i)\Phi_{\text{MSC}}(k-i,l)$$

Il ne reste qu'à fixer les valeurs extrêmes de la fonction de cohérence ([21]) afin de pouvoir calculer P_{spat} :

$$P_{\text{spat}}(k,l) = \begin{cases} 0 & \text{si } \tilde{\Phi}_{\text{MSC}}(k,l) \leq \Phi_{\min} \\ 1 & \text{si } \tilde{\Phi}_{\text{MSC}}(k,l) \geq \Phi_{\max} \\ \dfrac{\tilde{\Phi}_{\text{MSC}}(k,l) - \Phi_{\min}}{\Phi_{\max} - \Phi_{\min}} & \text{sinon} \end{cases} \tag{4.51}$$

On peut voir l'apport de cette nouvelle probabilité dans l'expression de $q(k,l) = 1 - P_{\text{local}}(k,l)P_{\text{global}}(k,l)P_{\text{trame}}(l)P_{\text{spat}}(k,l)$ en observant la différence entre le tracé en pointillé (sans P_{spat}) et le tracé plein (avec P_{spat}) sur la figure 4.3. La probabilité est plus précise avec l'information spatiale et permet d'améliorer les résultats. C'est ce que nous allons à présent étudier dans le **Chapitre 5**.

4.3 Conclusion

Nous avons étudié deux méthodes différentes afin d'estimer la réverbération tardive d'un signal réverbéré. La première se base sur le modèle stochastique de réponse des salles : on sépare le chemin direct des réflexions tardives pour calculer les différentes densités spectrales de puissance du signal. L'idée générale étant de dire que la réverbération tardive à un instant t correspond à ce que comportait le signal quelques instants plus tôt, atténué d'un facteur lié à la décroissance exponentielle du modèle de réponse de salle.

La seconde méthode implémentée se base sur la prédiction linéaire à long terme. On cherche à estimer le signal avec une combinaison linéaire de ses échantillons passés. Mais

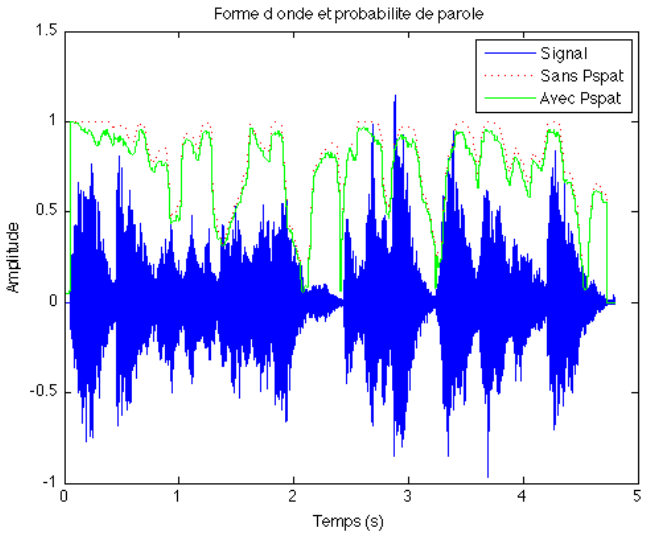

FIGURE 4.3 – Forme d'onde en bleu, probabilité de présence de parole sans P_{spat} en pointillés rouges, avec en trait plein vert

lorsque l'on applique ces coefficients de prédiction au signal réverbéré, l'énergie du signal filtré correspond à l'énergie de la réponse de la salle, une fois les réflexions précoces passées. Un biais est cependant présent, que l'on réduit en blanchissant le signal en amont.

Nous avons ensuite étudié deux façons de retirer du spectre du signal réverbéré la contribution de la réverbération tardive. Pour cela on peut utiliser une soustraction spectrale "classique" en utilisant un seuil afin de minimiser l'apparition du bruit musical. On peut également utiliser le gain OM-LSA, qui se base sur un modèle de distribution Gaussienne des coefficients de la transformée de Fourier du signal. Ce gain s'adapte à l'évolution du signal, grâce à un calcul de probabilité de présence de parole.

Ces techniques s'adaptent très bien en multicanal, dont le principal avantage est le moyennage des signaux obtenues grâce au réseau de microphones. Dans le cas du gain OM-LSA, on peut améliorer la précision de la probabilité de présence de parole en ajoutant une probabilité basée sur l'information spatiale, donnée par les différents microphones. Étudions à présent les résultats obtenus pour chacune de ces méthodes.

Chapitre 5

Résultats expérimentaux

Ce chapitre présente les résultats obtenus pour les différents algorithmes implémentés dans l'environnement *Matlab*, correspondant aux méthodes décrites **Chapitre 4**. Lorsque dans la première partie on parle de réponse de salle synthétique, cela correspond au code *Matlab* implémenté par Mc Govern suivant l'article de Allen [4], disponible en *open-source* sur Internet [1].

Les voix utilisées appartiennent au corpus *Harvard sentence list* [2], contenant des enregistrements anéchoïques de voix d'hommes et de femmes. En revanche, dans la partie "conditions réelles" de la section 5.2.3, j'utilise ma voix, enregistrée à travers un produit *Invoxia*.

Mais avant d'analyser les résultats obtenus, définissons les critères objectifs sur lesquels nous nous appuierons afin d'évaluer les performances de ces méthodes.

5.1 Critères objectifs

Lorsque nous écoutons un son réverbéré puis un son déréverbéré, il est possible de commenter de façon subjective l'efficacité de la déréverbération. On peut non seulement quantifier la réverbération qui a été retirée, mais également évaluer la distorsion introduite. Seulement, nous aurons tendance à dire "peu", "beaucoup", "imperceptible","gênant". Même si ces critères subjectifs sont dominants dans le cadre de communications téléphoniques (puisque le témoin de la qualité est un être humain), ils ne permettent pas de hiérarchiser ou discrétiser de façon précise les différentes méthodes.

C'est pourquoi un jeu de marqueurs est régulièrement utilisé afin d'évaluer les performances de la déréverbération, dont on peut trouver un inventaire très précis dans [20]. Cependant, je ne voulais pas implémenter moi-même ces marqueurs afin de ne pas altérer la crédibilité des résultats. C'est pourquoi j'ai utilisé des codes existants, basés sur les travaux de Loizou [25] et implémentés par l'auteur lui-même. Ces derniers se trouvent en accès libre sur sa page Internet [3]. Certains de ces critères sont également utilisés afin d'évaluer les performances des algorithmes testés dans le *Reverb Challenge 2014* [4], compétition

1. http ://www.mathworks.com/matlabcentral/fileexchange/5116-room-impulse-response-generator
2. IEEE Subcommittee on Subjective Measurements IEEE Recommended Practices for Speech Quality Measurements. IEEE Transactions on Audio and Electroacoustics. vol 17, 227-46, 1969
3. http ://ecs.utdallas.edu/loizou/speech/software.htm
4. http ://reverb2014.dereverberation.com/

Note	Qualité	Distorsion
5	Excellente	Imperceptible
4	Bonne	Perceptible mais pas dérangeante
3	Acceptable	Légèrement dérangeante
2	Pauvre	Dérangeante
1	Mauvaise	Très dérangeante

TABLE 5.1 – Tableau des notes MOS

de déréverbération à des fins de communications ou de reconnaissance automatique de la parole. Parmis les sept marqueurs disponibles dans [25] nous n'en utiliserons que quatre (PESQ, segSNR, WSS, LLR), les trois autres (C_{sig}, C_{bak}, C_{ovl}) étant peu utilisés en dehors de l'article de Loizou.

Le PESQ De l'anglais *Perceptual Evaluation of Speech Quality* (PESQ), cet indicateur permet d'évaluer la qualité d'un signal de voix transmise par un système de télécommunication (téléphone fixe, mobile, voix sur IP, etc.). Il est normalisé par l'Union Internationale des Télécommunications (ITU) sous le nom *P.862* [5]. Il permet de délivrer une note sur 5 équivalente à celle qui aurait été donnée par de réels auditeurs sur la qualité du signal perçu, suivant la note d'opinion moyenne (*Mean Opinion Score*, MOS en anglais) figurant dans le tableau 5.1.

Le segSNR Initialement utilisé pour évaluer le débruitage des signaux, le *segmental signal to noise ratio* est implémenté de la même manière que le *segmental signal to reverberant ratio* utilisé par Habets dans [20] pour évaluer la déréverbération. Il montra également que cet indice est proportionnel au DRR, donc plus cet indice sera élevé meilleure sera la déreverbération.

Pour le calculer il faut avoir le signal anéchoïque z ainsi que son estimation, issue de l'algorithme, $\tilde{z}$. Cet indice se calcule pour chaque trame l, de longueur N et d'avancement R suivant :

$$\text{segSNR}(l) = 10\log_{10}\left(\frac{\sum_{n=lR}^{lR+N-1} z^2(n)}{\sum_{n=lR}^{lR+N-1}(z(n)-\tilde{z}(n))^2}\right)$$

On peut calculer la moyenne sur l'ensemble des trames pour avoir une évaluation globale. Habets a montré dans [20] que cette valeur est mieux corrélée avec la qualité du signal (du point de vue coloration, temps de réverbération, qualité générale) que le DRR. L'indicateur utilisé par Loizou prend également en compte une pondération fréquentielle, prenant alors le nom de *frequency-weighted segmental signal to noise ratio, fwSNRseg* en anglais.

La WSS Cette distance en décibels calcule la différence pondérée des pentes spectrales $S(k,l)$ pour chaque bande de fréquence k à la trame l. On comprend alors son nom anglais *Weighted Spectral Slope* (WSS). Pour un nombre de trames N_t et K bandes de fréquences, on peut calculer :

5. https ://www.itu.int/rec/T-REC-P.862/fr

$$\mathrm{WSS} = \frac{1}{N_t} \sum_{l=0}^{N_t-1} \frac{\sum_{k=1}^{K} W(k,l)(S_a(k,l) - S_d(k,l))^2}{\sum_{k=1}^{K} W(k,l)}$$

Il faut dont également avoir la version anéchoïque du signal traité afin de calculer S_a et S_d les pentes spectrales respectives du signal anéchoïque et déréverbéré. La pente spectrale est, comme son nom le sous-entend, la différence entre deux amplitudes successives du spectre, en décibels. La pondération est faite à travers le filtre $W(k,l)$ au bin de fréquence k à la trame l. On cherche donc à diminuer cette distance, témoin de la distorsion introduite par le traitement.

Le LLR De l'anglais *Log-Likelihood Ratio* (LLR), cet indicateur correspond au logarithme de la fonction de vraisemblance entre le signal traité et le signal anéchoïque. Plus cette fonction est proche de un, donc le LLR proche de zéro, plus forte est la ressemblance.

Cette dernière est évaluée en fonction des vecteurs de prédiction linéaire du signal anéchoïque $\mathbf{pred}_a$ et du signal déréverbéré $\mathbf{pred}_d$, ainsi que de la matrice d'autocorrelation du signal anéchoïque $\mathbf{R}$. On définit pour chaque trame :

$$\mathrm{LLR} = \log\left(\frac{\mathbf{pred}_d \mathbf{R} \mathbf{pred}_d^T}{\mathbf{pred}_a \mathbf{R} \mathbf{pred}_a^T}\right)$$

On calcule ensuite la valeur moyenne sur l'ensemble des trames du signal afin d'obtenir la valeur finale.

5.2 Résultats obtenus

Les résultats des sections 5.2.1 et 5.2.2 ont été obtenus avec une réponse de salle synthétique, convoluée avec des signaux de voix issus du corpus *Harvard sentence list* échantillonnés à 16kHz. La réponse de la salle se base sur la méthode image ([4], [20]), considérant qu'un rayon qui se réfléchit sur une paroi peut être vu comme arrivant d'une salle voisine.

Le but de la méthode image est de calculer la distance parcourue entre une source acoustique et un récepteur, après chaque réflexion de l'onde sur une surface. Il suffit alors de prendre le symétrique de la source par rapport au mur (l'image) et de calculer la distance entre cette dernière et le microphone, en passant par le point où a lieu la réflexion sur le mur (cf. Fig. 5.1).

Puisque SRS' est un triangle isocèle en R, on voit que les distances RS' et RS sont identiques. Si l'on souhaite considérer une réflexion supplémentaire, il suffit de réitérer en partant cette fois-ci de S' pour placer son image S", et ainsi de suite (cf. Fig. 5.2).

On peut alors calculer la distance entre n'importe quelle source image et le microphone, donc le temps que prend chaque réflexion pour atteindre le microphone. Cela nous permet de calculer la fonction transfert entre la source et le récepteur, dont on prend la transformée de Fourier inverse afin d'obtenir la réponse impulsionnelle de notre salle.

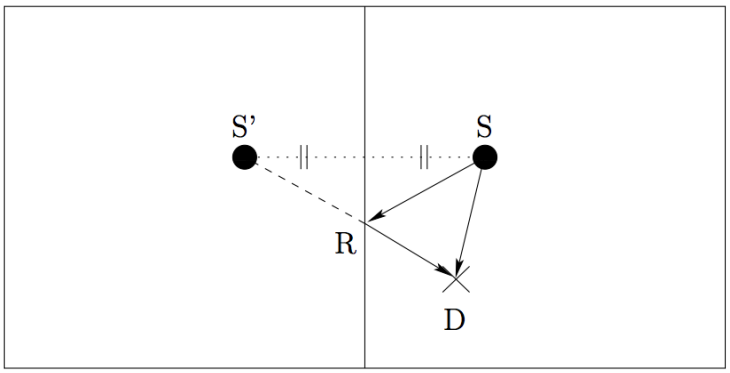

FIGURE 5.1 – S' est la source image de S, R le point d'incidence de la réflexion, D la source. Images provenant de [20]

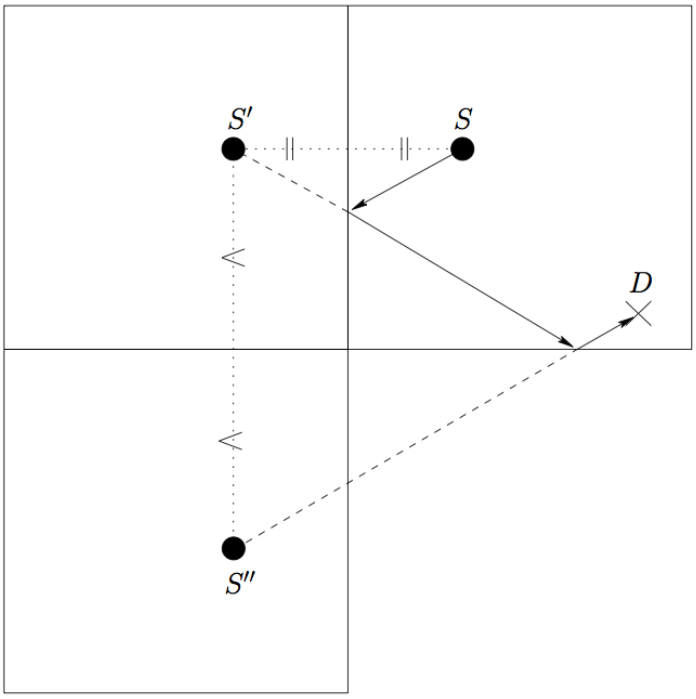

FIGURE 5.2 – Deux réflexions. Images provenant de [20]

h_{local}	Hanning	ω_{local}	1
h_{global}	Hanning	ω_{global}	15
Θ_{min}	-10 dB	Θ_{max}	-5 dB
$\Theta_{\text{trame min}}$	0 dB	$\Theta_{\text{trame max}}$	10 dB
β	0.8	α	0.1

TABLE 5.2 – Paramètres du gain OM-LSA

5.2.1 Monocanal

Les dimensions de la salle sont fixes : 9m de profondeur, 7m de largeur et 5m de hauteur. Le temps de réverbération est fixé à 750ms, via le coefficient de réflexion des surfaces de la salle. La source est également fixe, à 1m en partant du fond de la salle, décalée de 2m du mur latéral, à 1m de hauteur. Le microphone se situe dans le plan parallèle au sol, passant par la source, avec le même décalage par rapport au mur latéral. Sa distance va en revanche varier de 1m à 5m de la source.

On nomme "Méthode H-" lorsque l'on estime la réverbération via le modèle stochastique de réponse de salles, "Méthode K-" lorsque l'on utilise le modèle de prédiction linéaire à long terme. On nomme "Méthode -O" lorsque l'on restaure le signal à l'aide du gain OM-LSA, "Méthode -S" lorsque l'on utilise la soustraction spectrale. Cela permet donc d'évaluer quatre méthodes croisées : Méthode HO, Méthode HS, Méthode KO, Méthode KS. Pour chacune d'elles, on travaille avec des transformées de Fourier à court terme, utilisant des fenêtres de Hamming de 512 points, avec un recouvrement de 75%.

Les paramètres relatifs au gain OM-LSA sont récapitulés dans le tableau 5.2, ils sont conservés pour l'ensemble des essais. Pour ce qui est de la soustraction spectrale, on fixe $\alpha = 0.1$ (dans (4.39)) et $\eta = 0.8$ (dans (4.38)).

Pour l'estimation basée sur le modèle stochastique de réponse de salles (section 4.1.1), on fixe le paramètre T_l à 50ms. Le temps de réverbération RT_{60} est estimé par maximum de vraisemblance à partir de l'algorithme de Marco Jeub, disponibile sur Internet [6]. Le paramètre κ est initialisé à 0.01 ainsi que le pas d'avancement μ_κ.

Lorsque l'on estime la réverbération tardive par prédiction linéaire, l'ordre pour le blanchiment est fixé à $P = 20$, celui pour la prédiction linéaire à long terme est fixé à $N = 4000$, ce qui correspond à 250ms pour un échantillonnage à 16kHz.

Si l'on observe la figure 5.3, on peut voir les résultats des différents marqueurs présentés dans la section 5.1, pour le signal d'entrée (réverbéré) et de sortie (déréverbéré avec la méthode HO ou HS). On remarque que plus la distance entre la source et le microphone augmente, moins bons sont les marqueurs : le PESQ se dégrade (signal de moins en moins bonne qualité) ; le fwSNRseg diminue (signal de plus en plus réverbéré) ; le LLR ainsi que la WSS augmentent, témoignant de l'écart croissant avec le signal anéchoïque.

Cependant, on peut constater que le signal de sortie a un PESQ et un fwSNRseg plus élevé, que ce soit pour la méthode HO ou HS. Cela signifie que le signal de sortie est moins

6. http ://www.mathworks.com/matlabcentral/fileexchange/35740-blind-reverberation-time-estimation

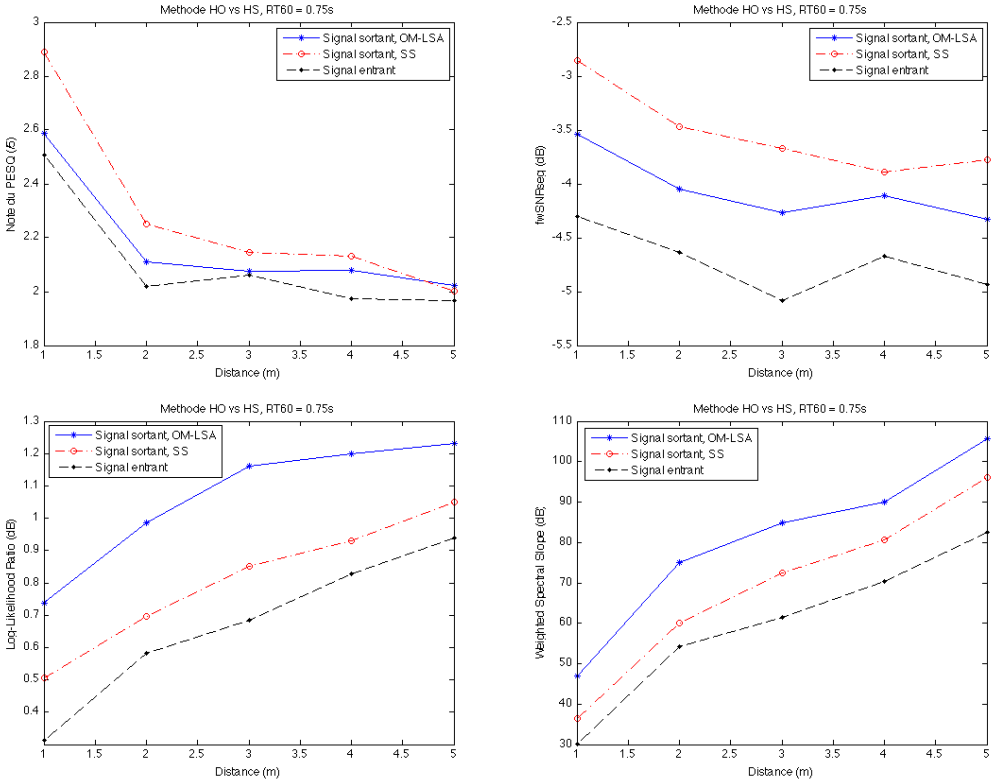

FIGURE 5.3 – Confrontation des méthodes HO/HS

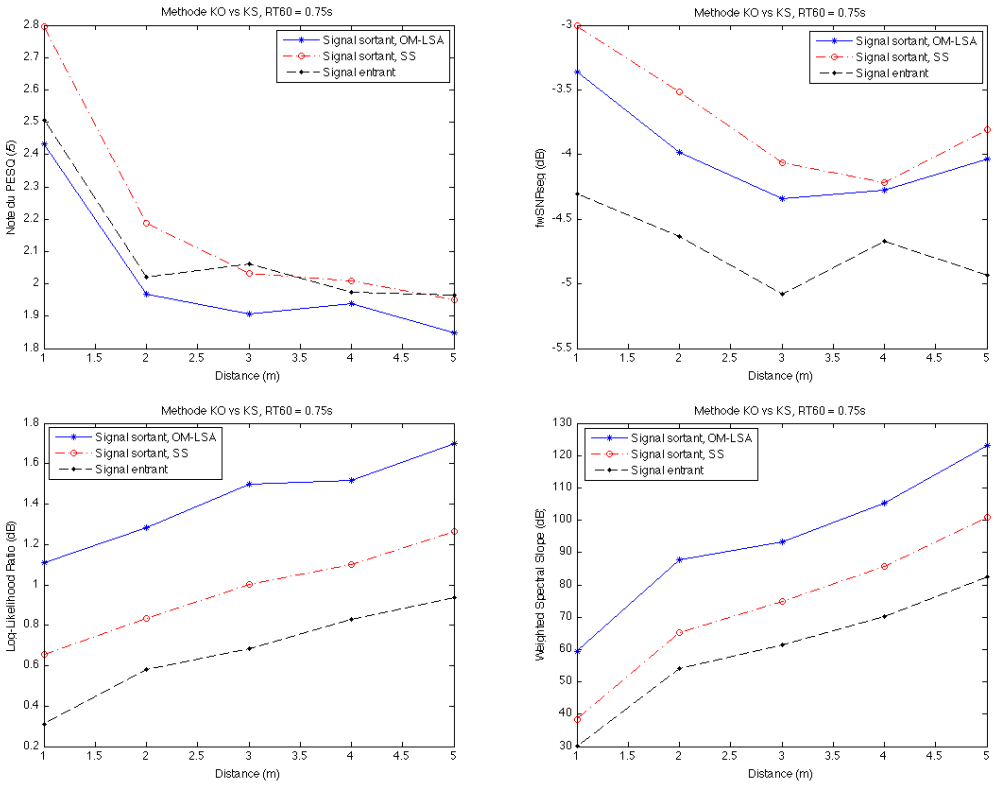

FIGURE 5.4 – Confrontation des méthodes KO/KS

réverbéré et jugé de meilleure qualité. En revanche l'augmentation générale du LLR et de la WSS montre que le signal traité comporte plus de distorsion que le signal réverbéré d'entrée. En revanche, à choisir entre la méthode HO ou HS, il vaut mieux choisir le gain de soustraction spectrale plutôt que OM-LSA puisque ce dernier donne un PESQ et un fwSNRseg plus élevé, un LLR et une WSS plus faible.

La figure 5.4 nous donne cette fois-ci les résultats obtenus par les méthodes KO et KS. L'analyse faite précédement sur les signaux d'entrée est toujours valable. Il en est de même pour les différences entre les méthodes KO et KS : on obtient de meilleurs résultats avec la méthode KS. Il semblerait donc que la déréverbération est meilleure lorsque l'on utilise le gain de soustraction spectrale défini dans la section 4.2.1 plutôt que le gain OM-LSA défini dans la section 4.2.2. Comparons alors les méthodes HS et KS.

On peut voir les différences entre les méthodes HS et KS dans la figure 5.5. On constate que pour une même méthode de déréverbération, on obtient de meilleurs résultats si on estime la réverbération tardive avec le modèle stochastique de réponse de salle décrit dans la section 4.1.1. En effet, la méthode HS donne un PESQ et un fwSNRseg plus élevés, ainsi qu'un LLR et une WSS plus faibles qu'avec la méthode KS.

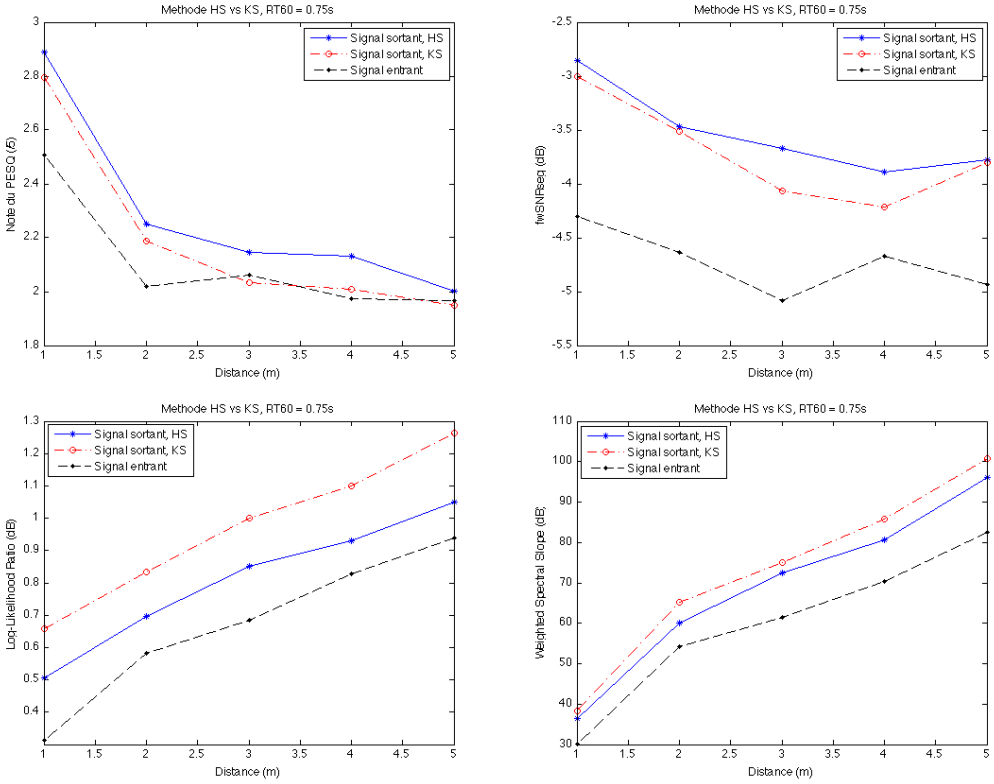

FIGURE 5.5 – Confrontation des méthodes HS/KS

À l'oreille la différence est très subtile, même si les signaux déréverbérés par la méthode HS semblent être effectivement plus fidèles au son réel. En revanche, le temps de calcul bien plus important pour la méthode KS (du à l'inversion de matrices de grandes tailles) nous encourage d'autant plus à utiliser la méthode HS. Observons à présent ce que l'on obtient en utilisant un réseau de microphones.

5.2.2 Multicanal

On teste désormais les méthodes en multicanal, avec un réseau de capteurs allant jusqu'à 10 microphones. Dans un premier temps observons l'apport de l'information spatiale en comparant les résultats de la méthode HO avec ou sans la probabilité P_{spat}. Pour cela nous avons testé deux montages du réseau de microphones : un premier dit "colinéaire" (les capteurs sont alignés avec la direction de propagation de l'onde acoustique) ; le second dit "orthogonal" (les capteurs sont alignés avec une direction orthogonale à celle de la propagation de l'onde). Les microphones sont espacés de 50cm dans la formation colinéaire, de 10cm dans la formation orthogonale.

On peut voir dans la figure 5.6 les résultats obtenus avec ou sans l'utilisation de la probabilité spatiale P_{spat}. On voit clairement que l'utilisation de cette dernière permet d'améliorer les performances de la déréverbération puisqu'on augmente le PESQ ainsi que le fwSNRseg, et diminue le LLR et la WSS. Les résultats suivants on été réalisés sur la formation orthogonale, pour deux temps de réverbération différents : 375ms (pièce meublée) et 750ms (pièce vide).

La figure 5.7 présente des résultats satisfaisants. Force est de constater que le fait de moyenner les différents signaux, alignés dans le temps, permet déjà de déréverbérer légèrement. C'est la conséquence du *beamforming* évoquée section 3.2.3. De plus, on voit bien que la méthode HO permet d'augmenter le PESQ ainsi que le fwSNRseg. En revanche, de la distorsion est introduite (augmentation du LLR et de la WSS) même si cette dernière diminue avec l'augmentation du nombre de microphones.

Si on compare ces résultats à la méthode HS, on s'aperçoit que cette dernière offre de meilleurs résultats. On peut voir sur la figure 5.8 qu'utiliser la soustraction spectrale au lieu du gain OM-LSA permet d'augmenter le PESQ ainsi que le fwSNRseq, et de diminuer le LLR et la WSS. De plus, le temps de calcul est bien plus court pour la méthode HS, c'est pourquoi nous choisirons d'utiliser cette dernière dans les produits *Invoxia*.

5.2.3 Conditions réelles

Réponse de salle réelle

Dans cette partie, nous avons testé les méthodes implémentées sur une vraie réponse de salle. Jusqu'à présent nous utilisions le script `rir.m` afin de réverbérer nos signaux, permettant de changer les paramètres de la salle (volume, temps de réverbération, etc.) et donc d'effectuer des tests sur différentes salles. La réponse utilisée ici, représentée figure 5.10, a été mesurée dans un bureau d'*Invoxia*, avec un produit de l'entreprise : le *NVX-620* (cf. Fig. 5.9). Le temps de réverbération mesuré est de $\text{RT}_{60} = 1.95s$.

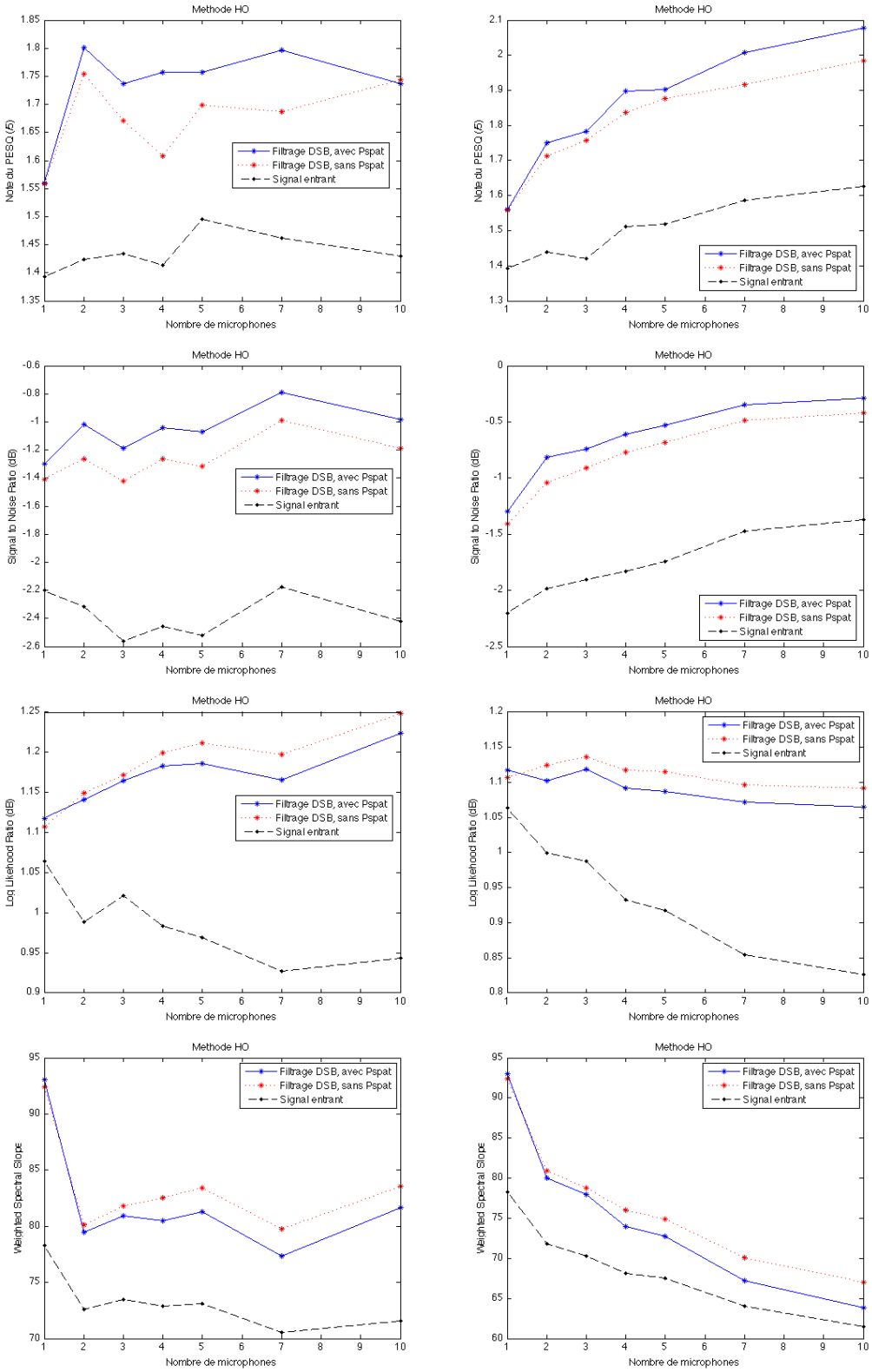

FIGURE 5.6 – Influence de l'information spatiale, montage colinéaire à gauche, orthogonal à droite

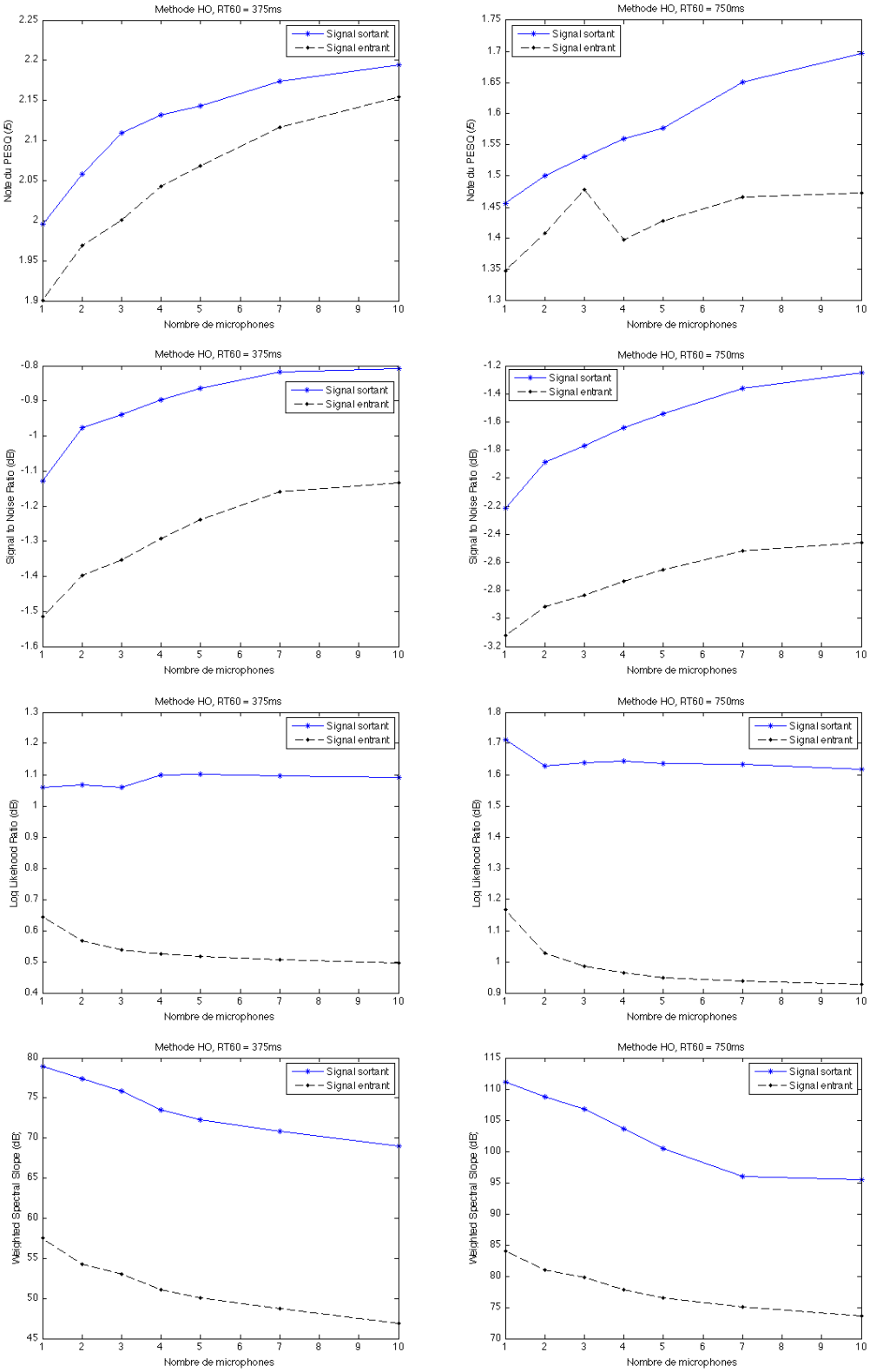

FIGURE 5.7 – Résultats de la méthode HO en multicanal pour $RT_{60} = 375$ms à gauche, $RT_{60} = 750$ms à droite

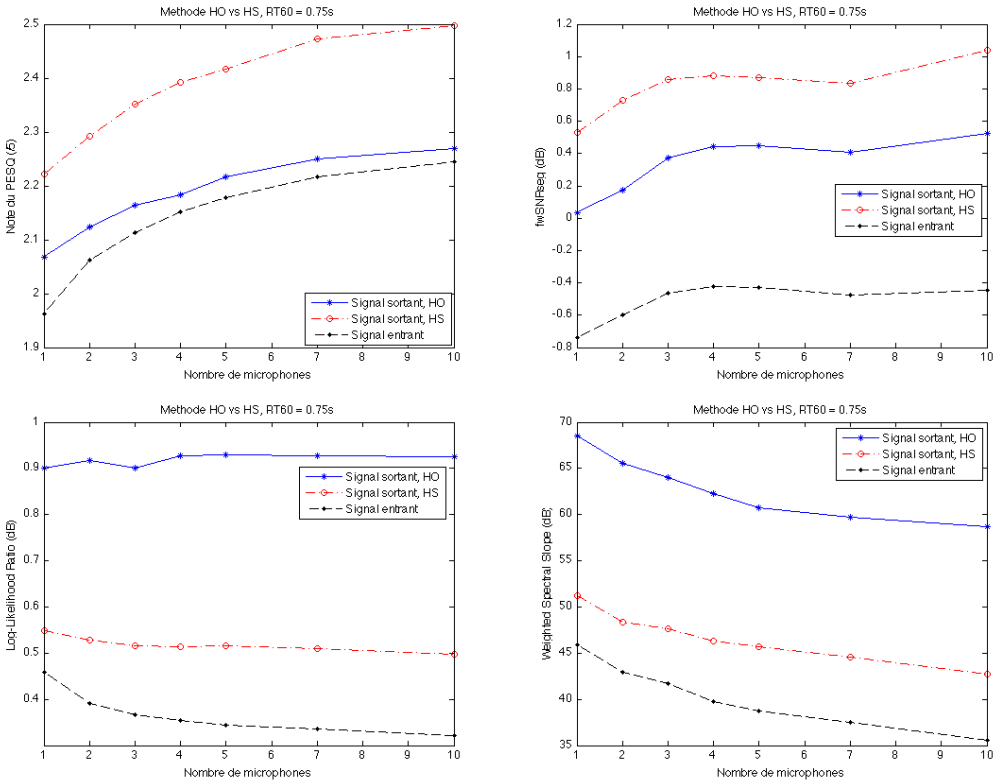

FIGURE 5.8 – Résultats des méthodes HO et HS en multicanal pour $RT_{60} = 750ms$

FIGURE 5.9 – Le *NVX-620*

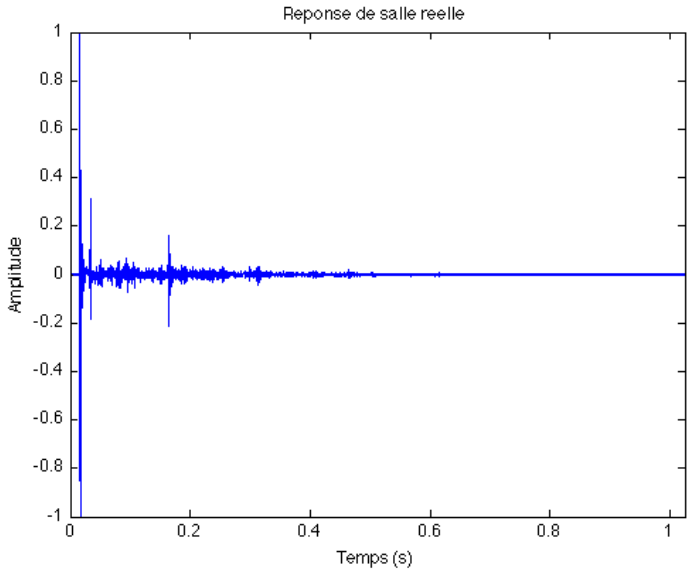

FIGURE 5.10 – Réponse de salle mesurée

	PESQ (/5)	fwSNRseg (dB)	LLR (dB)	WSS (dB)
Entrée	1.67	-4.58	7.67	88.46
Méthode HO	1.72	-3.03	5.02	118.74
Méthode HS	1.88	-2.63	5.28	98.55

TABLE 5.3 – Résultats obtenus pour une réponse de salle mesurée

On peut observer les résultats obtenus avec cette vraie réponse de salle dans le tableau 5.3. Le PESQ et le fwSNRseg augmentent et le LLR diminue avec les méthodes HO et HS, ce qui est satisfaisant. En revanche la WSS augmente, témoignant de la distorsion introduite dans le signal. On voit clairement que la méthode HS donne de meilleures performances en matière de déréverbération, en plus d'être moins couteuse en calculs.

Conditions réelles

Cette fois-ci on teste les méthodes en conditions réelles. J'ai enregistré ma voix, répétant le même texte à une distance de un à cinq mètres de la base (toujours un *NVX-620*). En me connectant localement au téléphone, je récupère un fichier monocanal correspondant au discours ayant subi l'effet de la salle ainsi que les divers traitements présents dans l'appareil (débruitage, *beamforming*). La salle de test est une pièce des locaux d'*Invoxia*, de dimensions $[3.36, 3.46, 2.26]\mathrm{m}^3$, avec un temps de réverbération mesuré à 430ms.

N'ayant pas le signal anéchoïque, nécessaire dans l'évaluation de Loizou, il n'est pas possible d'obtenir nos marqueurs habituels (PESQ, fwSNRseq, LLR, WSS). D'un point de vue subjectif, j'estime que les résultats sont très satisfaisants, aussi bien à une distance d'un mètre que de cinq. La différence entre les méthodes HO et HS est toujours aussi peu perceptible à mon oreille. Cependant, on pourra trouver dans les figures 5.11, 5.12 et 5.13 les spectrogrammes des signaux réverbérés et déréverbérés, pour différentes distances source/microphone. Cela permet d'observer les modifications effectuées sur les spectres des signaux, principalement le raccourcissement des queues des phonèmes et la disparition des trainées du spectre.

5.3 Conclusion

Dans cette partie, nous avons pu évaluer les performances des méthodes implémentées au cours du stage. Nous nous sommes servi des critères proposés par Loizou dans [25], implémentés en Matlab par ses soins. Les marqueurs retenus pour l'ensemble des tests sont ceux utilisés notamment pour le *Reverb Challenge 2014*, à savoir le PESQ, fwSNRseg, LLR et WSS. Nous avons évalué quatre méthodes nommées "HO" (estimation basée sur le modèle stochastique de réponses de salles, restauration par gain OM-LSA) ; "HS" (estimation basée sur le modèle stochastique de réponses de salles, restauration par soustraction spectrale) ; "KO" (estimation basée sur la prédiction linéaire à long terme, restauration par gain OM-LSA) ; "KS" (estimation basée sur la prédiction linéaire à long terme, restauration par soustraction spectrale).

Pour l'ensemble des tests, la restauration par soustraction spectrale propose de meilleurs résultats, en plus d'avoir un temps de calcul plus court. Pour ce qui est de l'estimation, celle développée par Habets (méthodes H-) dans [20] semble être plus précise que celle par prédiction linéaire (méthodes K-), comme on peut le voir en monocanal Fig. 5.5. N'ayant pas pu implémenter la méthode par prédiction linéaire en multicanal, nous n'avons pu comparer uniquement les méthodes HO et HS en multicanal. Les résultats sont satifsfaisants et permettent de distinguer les méthodes, même si d'un point de vue subjectif (à l'oreille) les différences entre ces dernières sont subtiles.

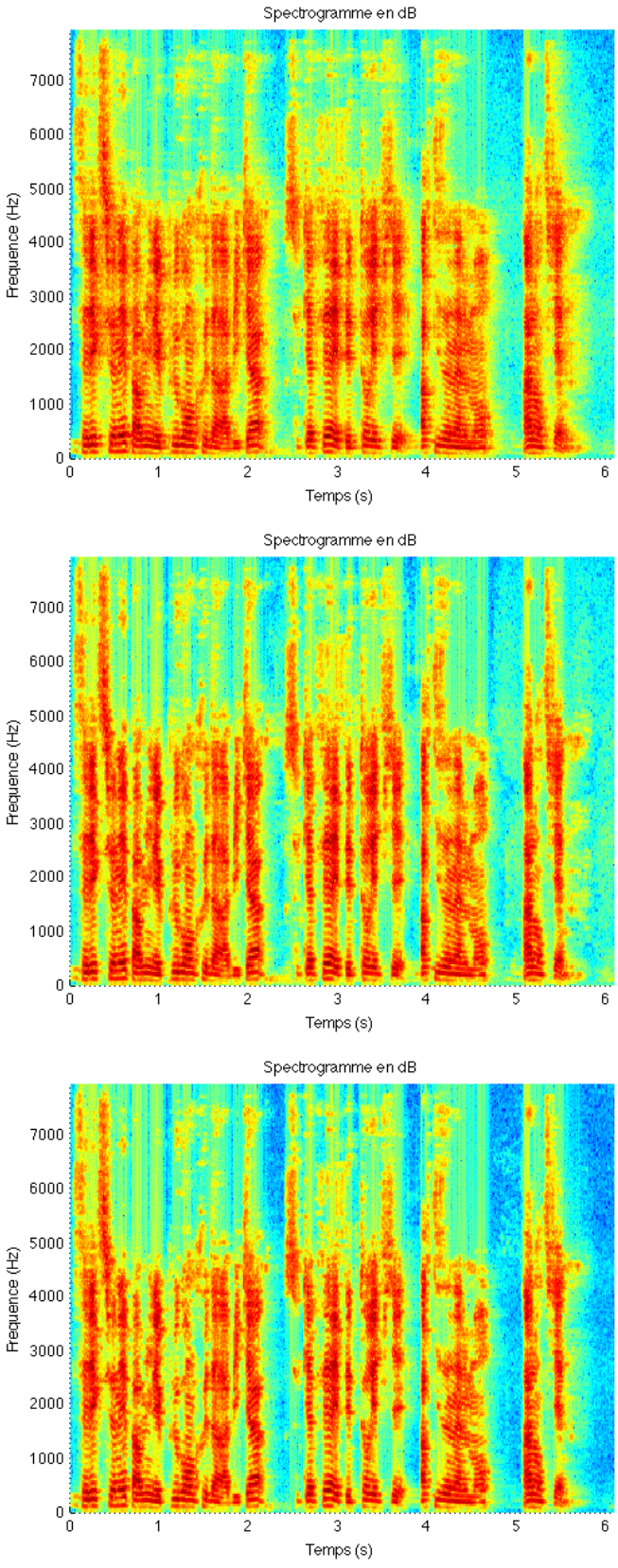

FIGURE 5.11 – Spectrogramme du signal réverbéré (en haut), déréverbéré par HS (au milieu), par KS (en bas) pour $RT_{60} = 430$ms, à 1m du microphone

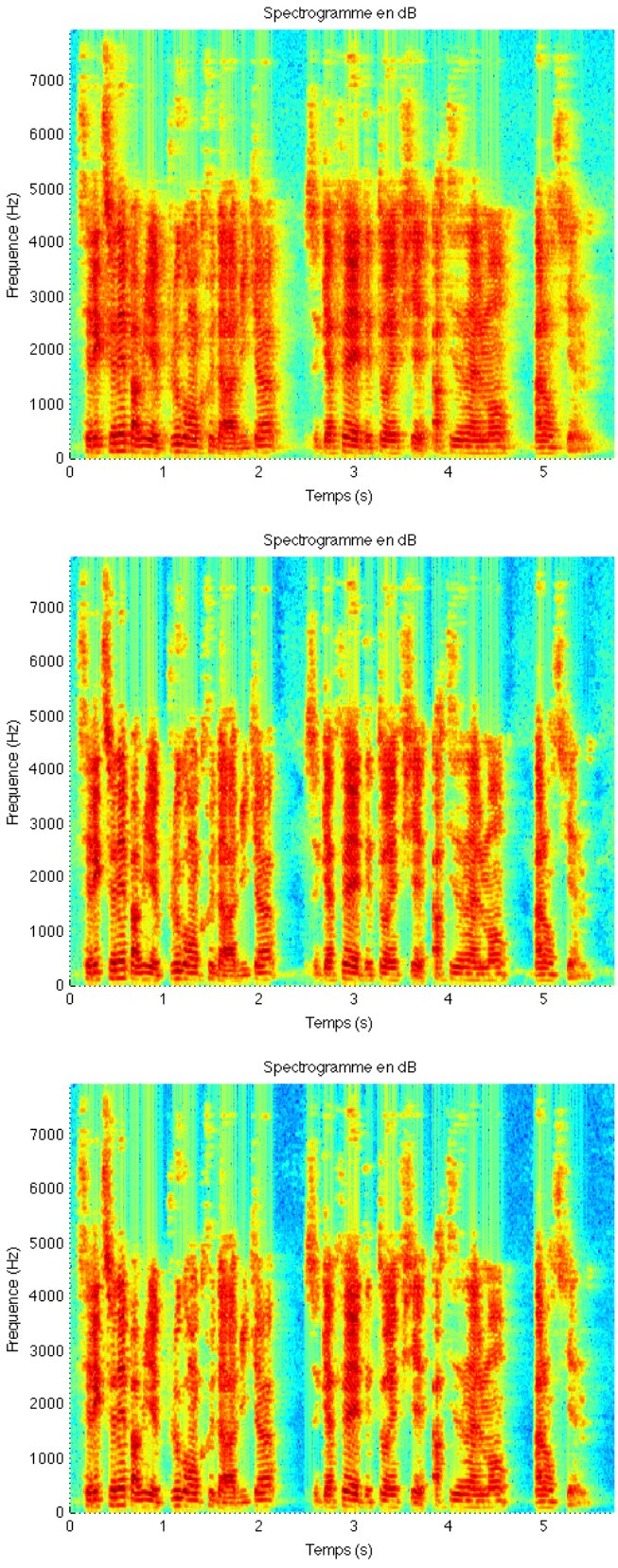

FIGURE 5.12 – Spectrogramme du signal réverbéré (en haut), déréverbéré par HS (au milieu), par KS (en bas) pour $RT_{60} = 430$ms, à 3m du microphone

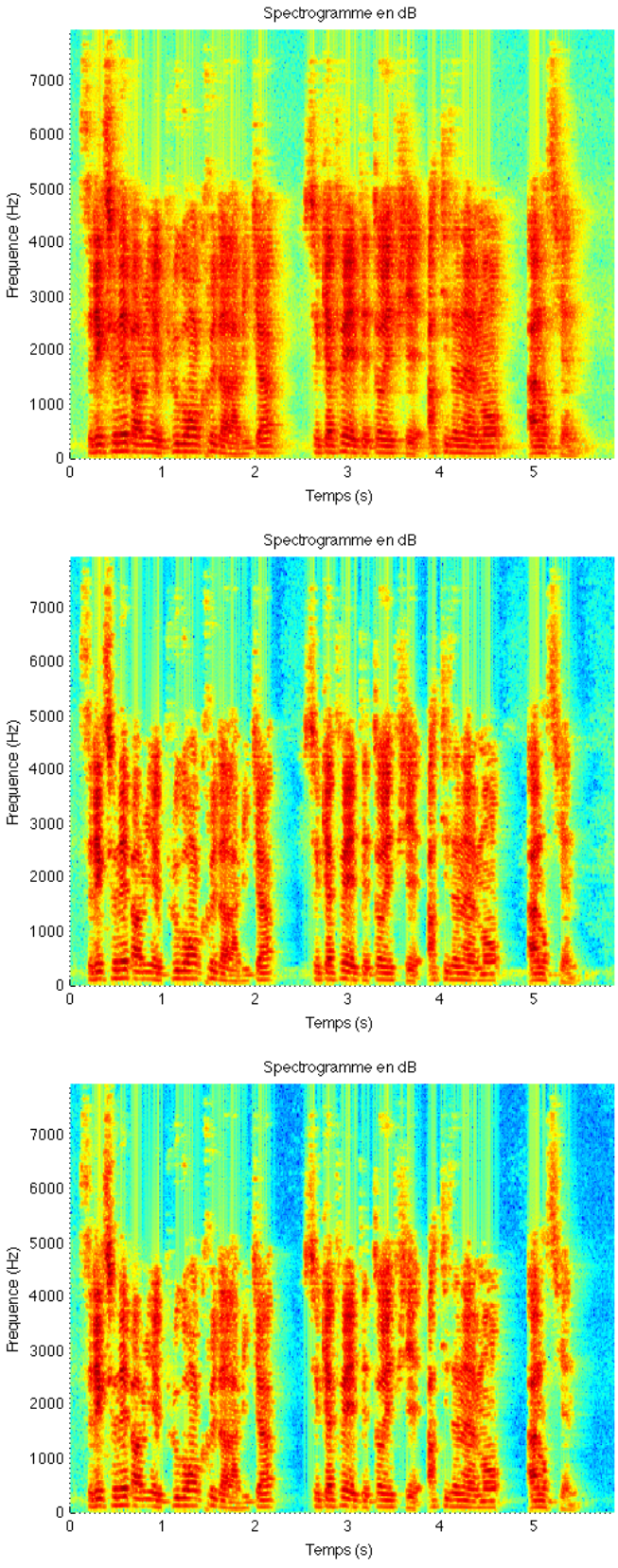

FIGURE 5.13 – Spectrogramme du signal réverbéré (en haut), déréverbéré par HS (au milieu), par KS (en bas) pour $RT_{60} = 430$ms, à 5m du microphone

Finalement ce sera la méthode HS qui sera implémentée dans les produits *Invoxia*, principalement due à sa vitesse de calcul. Les résultats obtenus en conditions réelles sont tout autant satisfaisants que ceux effectués en simulation de salle, avec des voix issues d'un corpus. En revanche un problème subsiste : même si on arrive à déréverbérer correctement d'un point de vue horizontal (disparition des traînées du spectre), on conserve une forte distorsion dans la trajectoire des formants. Nous allons montrer dans le **Chapitre 6** que ceci peut être corrigé en se penchant sur la phase des signaux, délaissée jusqu'à présent.

Chapitre 6

Importance de la phase en déréverbération

Dans l'ensemble des méthodes abordées dans ce document, on travaille sur l'amplitude de la transformée de Fourier à court terme des signaux à traiter. Une fois les modifications effectuées le signal de sortie est synthétisé par transformée de Fourier inverse, en se servant de la phase du signal réverbéré.

Certes, travailler sur l'amplitude des signaux plutôt que sur leur phase est plus confortable du point de vue robustesse et prédiction. Seulement, l'information contenue dans la phase est très importante et peut jouer un rôle clef pour parfaire les méthodes de déréverbération existantes. Car même si on arrive désormais à bien déréverbérer d'un point de vue temporel (on raccourcit les queues des phonèmes, supprime les trainées dans le spectre, etc.), la distorsion des formants est encore trop présente. Or l'utilisation de la phase du signal anéchoïque permet de régler ce problème.

6.1 Utilisation de la phase du signal anéchoïque

Comme expliqué section 2.2.1, la réverbération, en plus de rallonger et "d'étaler" le spectrogramme sur l'axe temporel, va également dégrader l'axe fréquentiel. Lorsqu'une sinusoïde change de fréquence de manière continue, la réverbération va grossir sa trajectoire dans le spectrogramme ; c'est ce que l'on constate figure 2.4. L'étalement est dû à la somme des versions atténuées et retardées du signal dont la fréquence fondamentale varie au cours du temps. C'est donc pour cela que la réverbération introduit de la distorsion dans la voix, puisque cette dernière perturbe la trajectoire des formants de la voix.

Or ces trajectoires sont guidées par la phase de la transformée de Fourier, tandis que la répartition de l'énergie est régie par son amplitude. C'est pourquoi la phase du signal anéchoïque est un bon candidat pour récupérer les trajectoires formantiques du signal anéchoïque, et donc de participer à la déréverbération.

Afin de mettre en valeur cette influence, j'ai appliqué directement la phase du signal anéchoïque au signal réverbéré. Si on nomme $Y_{\text{ané}}(k,l)$ le coefficient de la transformée de Fourier à court terme du signal anéchoïque (au bin de fréquence k et à la trame l) et

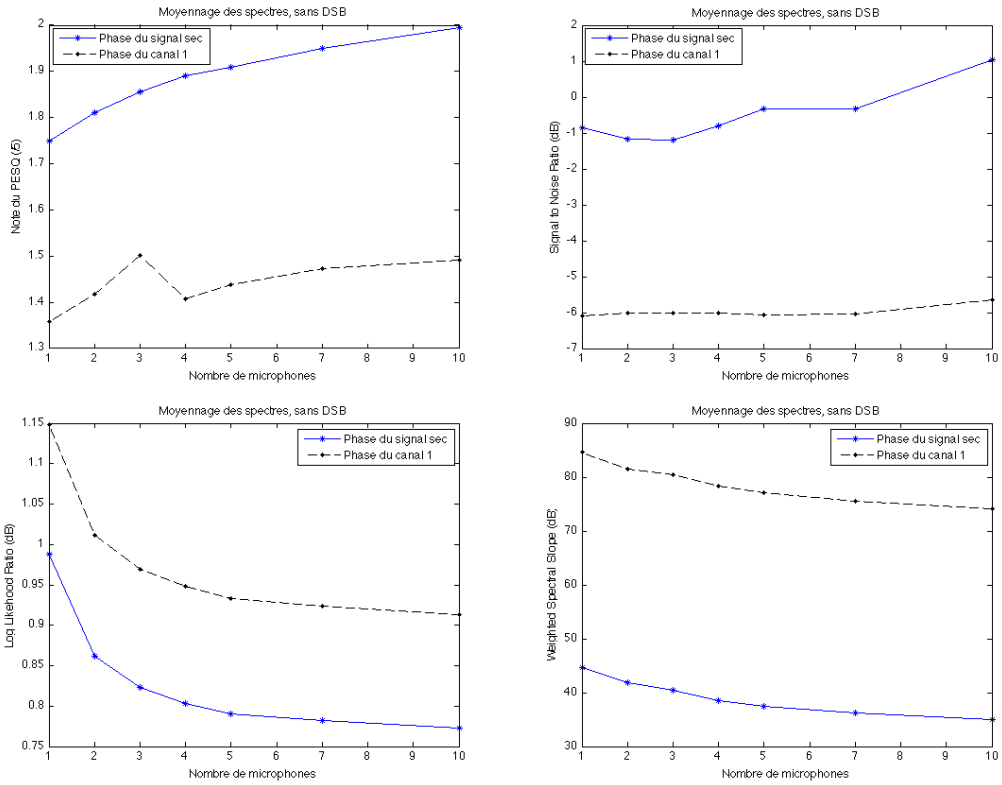

FIGURE 6.1 – Résultats obtenus en appliquant la phase du signal anéchoïque à une mixture de signaux réverbérés

$Y_{\text{rev}}(k,l)$ celui de sa version réverbérée, je créé une transformée de Fourier à court terme Z de la forme :

$$Z(k,l) = |Y_{\text{rev}}(k,l)|e^{i\arg\{Y_{\text{ané}}(k,l)\}} \tag{6.1}$$

Ce signal est ensuite synthétisé par transformée de Fourier inverse, avec addition-recouvrement, afin de donner le signal $z(t)$.

Lorsque l'on écoute le signal $z(t)$ ce dernier semble en effet moins réverbéré, même si la déréverbération est beaucoup moins évidente qu'avec les méthodes citées jusqu'à présent. Cette impression de déréverbération est validée par les critères objectifs utilisés dans le **Chapitre 5**. La figure 6.1 présente les résultats obtenus pour le PESQ, fwSNRseg, LLR et WSS pour des signaux obtenus via (6.1). La courbe "Phase du canal 1" correspond à la moyenne en amplitude des M acquisitions disponibles, dont la phase provient du signal réverbéré du premier microphone : $Y_{\text{rev}}(k,l) = \frac{1}{M}\sum_{m=1}^{M}|Y_m(k,l)|e^{i\arg\{Y_1(k,l)\}}$. La courbe "Phase du signal sec" provient du signal traité, obtenu suivant (6.1).

On voit clairement que le simple ajout de cette phase joue un rôle crucial dans la déréverbération : on augmente le PESQ ainsi que le fwSNRseg, mais on diminue également le LLR ainsi que la WSS.

On peut observer les modifications effectuées sur le spectre en regardant la figure 6.2. Force est de constater que les formants sont mieux définis dans la figure du bas, et qu'on retrouve même les pauses du discours, que ce soit pour les voyelles ou les consonnes. Le spectrogramme du logiciel *Adobe Audition* permet également de bien voir la restauration des formants, comme on peut voir sur la figure 6.3. Ces manipulations ont été réalisées en connaissant en amont la phase du signal anéchoïque, on comprend alors la motivation de pouvoir estimer cette phase anéchoïque à partir du signal réverbéré.

6.2 Estimation de la phase du signal anéchoïque

Après des premiers essais peu concluants utilisant la réponse de la salle, j'ai choisi d'utiliser une approche par synthèse de voix, employée de façon différente dans [34]. À partir du signal réverbéré, j'essaye de synthétiser la voix à l'aide de sinusoïdes. Une fois ce signal de synthèse obtenu, je récupère la phase de sa transformée de Fourier à court terme afin de l'appliquer au signal réverbéré comme dans (6.1). On espère alors retrouver les formants de la voix du signal anéchoïque, détériorés par la réverbération.

6.2.1 Synthèse de voix

Si on considère le modèle harmonique de la voix, on peut écrire un passage voisé s comme une somme de N_p partiels, chacun représenté par un cosinus :

$$s(n) = \sum_{p=1}^{N_p} a_p \cos(2\pi f_p n + \phi_p) \text{ avec } f_p = p f_0$$

Afin de synthétiser le signal de voix on réalise dans un premier temps une transformée de Fourier à court terme sur le signal observé, avec une fenêtre de taille N_w et un recouvrement de X%, donnant un avancement de trame de $R = (1 - X) * N_w$ échantillons.

On veut alors recréer un signal continu à l'aide d'une transformée de Fourier à court terme de N_t trames, donc d'informations obtenues tous les R échantillons : il faut interpoler sur R points, entre deux trames successives. Avant toute chose, il faut récupérer le pitch $f_0(n)$ du signal, c'est-à-dire la fréquence du partiel le plus bas. Pour cela je repère le premier pic significatif dans la transformée de Fourier à court terme, trame par trame, avec un prétraitement afin de faciliter la recherche (passe-bas à 500Hz et seuillage des pics). J'obtiens donc un vecteur de N_t points m'indiquant pour chaque trame dans quel bin de fréquence de la transformée de Fourier à court terme se trouve le fondamental de la voix.

Donc si à la trame i le fondamental est au bin k_i, on va générer sur R échantillons :

$$s_i(n) = \sum_{p=1}^{N_p} a_{p,i} \cos(2\pi p(\frac{k_i}{N_w} f_e)n + \phi_p)$$

Pour ce qui est du calcul des $a_{p,i}$ on interpole de façon linéaire :

$$a_{p,i}(n) = |Y(k_i, i)| + n * \frac{|Y(k_{i+1}, i+1)| - |Y(k_i, i)|}{R}$$

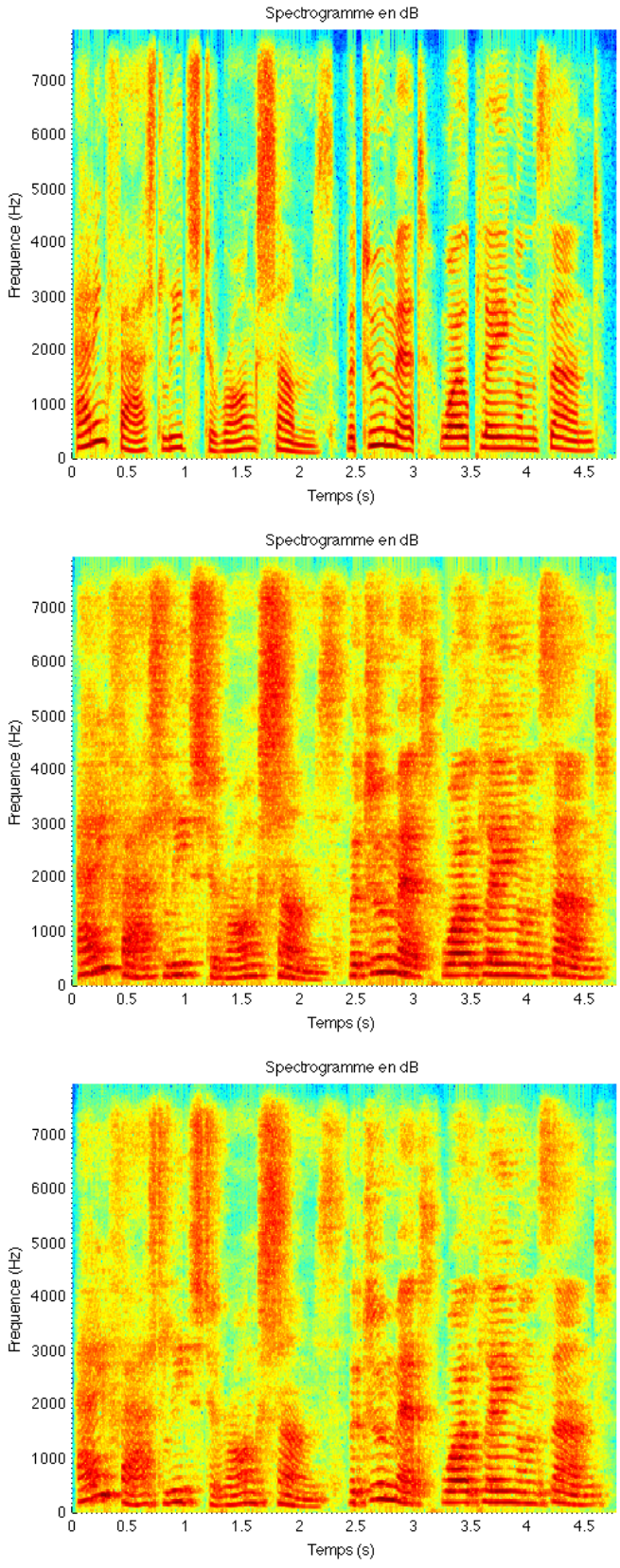

FIGURE 6.2 – Spectrogramme du signal anéchoïque (en haut), réverbéré (au milieu), réverbéré avec la phase anéchoïque (en bas)

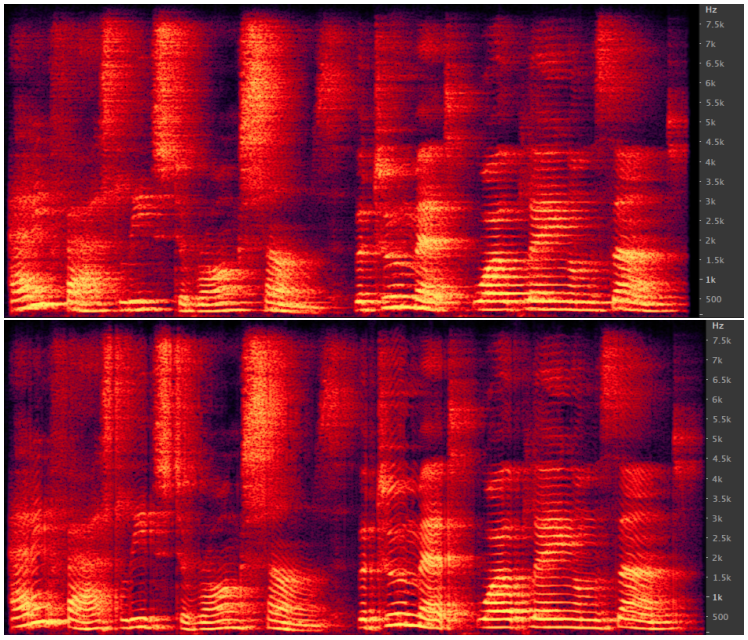

FIGURE 6.3 – Spectrogramme du signal réverbéré (en haut), réverbéré avec la phase anéchoïque (en bas)

Cette technique correspond au vocodeur de phase, où l'on considère la fréquence fixe sur la longueur de la fenêtre d'analyse. Afin que ce modèle soit valable il faut donc des fenêtres courtes, mais qui dit fenêtre courte dit faible résolution fréquentielle. Or les variations du fondamental sont faibles puisque ce dernier évolue dans une faible bande de fréquence. Il faut donc une fenêtre d'au moins 2048 points afin d'avoir une résolution de moins de 10Hz (pour des signaux échantillonnés à 16kHz).

C'est pourquoi on ne doit plus considérer la fréquence comme étant fixe sur les R points, mais ayant une évolution linéaire. On calcule alors la fréquence instantanée à injecter dans les cosinus :

$$\Phi_i(t) = 2\pi \int_0^t f_i(\tau)\mathrm{d}\tau + \Phi_i(0) = 2\pi \int_0^t \left(f_i + \tau \frac{\mathrm{d}f_i}{\mathrm{d}\tau} \right) \mathrm{d}\tau + \Phi_i(0)$$

Au début de chaque trame i on crée donc un signal de R points suivant :

$$s_i(n) = \sum_{p=1}^{N_p} a_{p,i} \cos(\Phi_{p,i}(n))$$

Avec :

$$\Phi_{p,i}(n) = 2\pi p \frac{k_i}{N_w} n + 2\pi p \left(\frac{k_{i+1} - k_i}{R N_w} \right) \frac{n^2}{2} + \Phi_{p,i}(0)$$

6.2.2 Résultats obtenus

Lorsque l'analyse se fait sur le signal anéchoïque, la synthèse est satisfaisante, on ne perçoit quasiment pas de différence entre le signal réel et synthétisé (cf. Fig. 6.4). En revanche, lorque l'on fait l'analyse sur le signal réverbéré, la synthèse n'est pas fidèle.

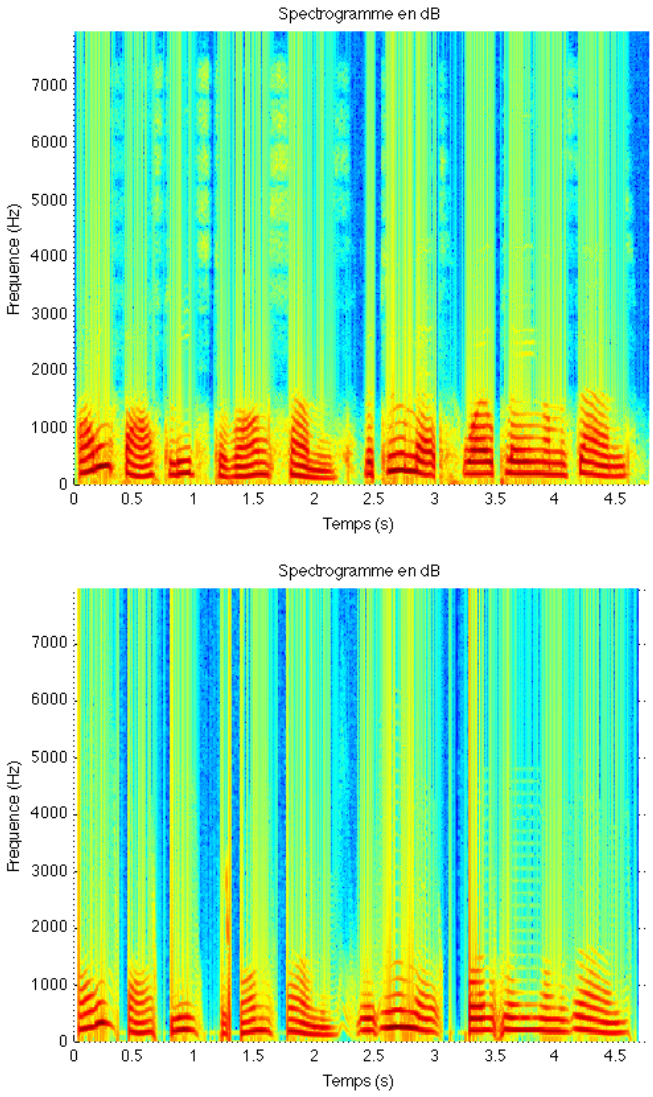

FIGURE 6.4 – Spectrogramme du signal d'origine en haut, synthétisé en bas

Puisque le signal est bruité et que les phonèmes sont rallongés, l'algorithme n'arrive pas à bien détecter les fins de phrases et donc la chute des formants (lorsque l'amplitude devient nulle). De plus, puisque le signal réverbéré est une somme de versions retardées du signal anéchoïque, on observe un retour de la même fréquence fondamentale une trame sur deux.

Ceci est dû à l'arrivée d'une version retardée du signal d'origine. Cela donne donc un vecteur de fréquence fondamentale ayant de nombreuses variations, et donc un effet *tremolo* sur la voix synthétisée (cf. Fig. 6.5).

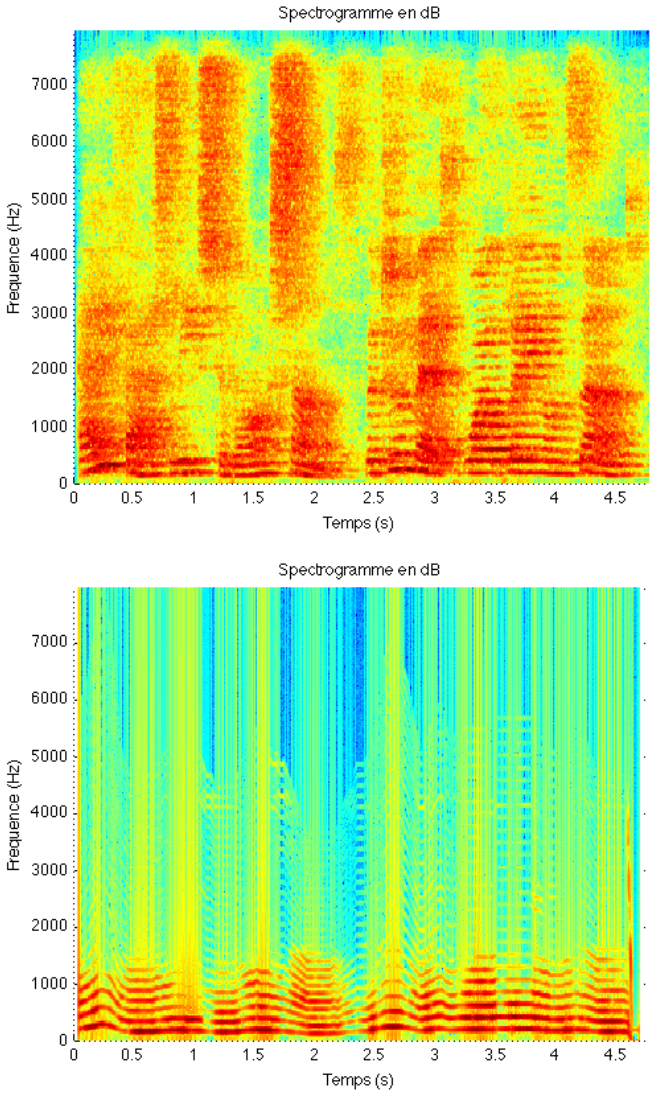

FIGURE 6.5 – Spectrogramme du signal d'origine en haut, synthétisé en bas

En revanche, le signal obtenu par transformée de Fourier inverse et addition-recouvrement (de la transformée de Fourier à court terme construite suivant (6.1)), basée sur la phase du signal de synthèse additive, possède les mêmes trajectoires formantiques. De fait, en améliorant notre algorithme de synthèse on pourrait obtenir les résultats espérés dans la section 6.1.

6.3 Travaux futurs

Mes travaux sur la phase se poursuivront durant ma thèse CIFRE encadrée par *Télécom ParisTech*, au sein de l'entreprise *Invoxia*. Cette dernière s'intitulera *Utilisation de la phase en déréverbération des signaux de parole*.

Les questions à résoudre seront initialement :
- Trouver comment la phase est impactée par la réverbération tardive, d'abord en mono-capteur puis en multi-capteurs,
- Modéliser cet impact afin de proposer un traitement adapté à ce modèle,
- Qualifier l'influence perceptive des traitements sur la phase.

Il faudra donc reprendre de façon assidue les différents modèles de phase établis jusqu'à présent, afin de bien comprendre son rôle sur des signaux simples. Les modèles compris sur ces signaux seront étendus à des signaux de plus en plus complexes, afin d'aboutir sur des signaux de parole. Le but étant de retrouver la phase d'un signal anéchoïque à partir de sa version réverbérée.

6.4 Conclusion

Nous avons montré que la phase joue un rôle important dans la déréverbération. Délaissée au profit de traitements sur l'amplitude du spectre, une modification de la phase dans le but d'estimer celle du signal anéchoïque, permet de déréverbérer les signaux de parole.

Une façon de l'estimer peut se faire en utilisant la synthèse additive de la voix, en récupérant la phase du signal de synthèse (une somme de cosinus en rapport harmonique). Seulement, si l'analyse-synthèse se déroule bien lorsque l'on travaille avec des signaux anéchoïques, il n'en est pas de même avec des signaux réverbérés. Les changements de pitch ainsi que la perte des zones de silence, due à la réverbération, dégradent le signal synthétisé, rendant son utilisation peu efficace.

La suite des travaux commencera par une étude théorique conséquente de la phase, en observant les modifications dues à une puis plusieurs réflexions, sur des signaux de plus en plus complexes. Le but étant de pouvoir reconstruire la phase du signal anéchoïque à partir de sa version réverbérée, afin de l'appliquer au signal déréverbéré par une méthode connue.

Chapitre 7

Conclusion

Ce stage de recherche s'est articulé autour de la réverbération. Ce phénomène naturel d'acoustique des salles peut être séparé en deux composantes : une partie précoce souvent utile et recherchée, une partie tardive. Nous avons constaté que cette dernière intervient de façon marquée dans des conversations téléphoniques de type mains-libres, et nuit à l'intelligibilité du discours. C'est pourquoi nous avons étudié différentes méthodes de déréverbération afin d'en implémenter une dans les produits *Invoxia*.

Après avoir rappelé les connaissances nécessaires pour expliquer et caractériser la réverbération, nous avons présenté un état de l'art des méthodes de déréverbération, afin de positionner celles utilisées durant le stage. On peut alors séparer les méthodes dites d'inversion, où l'on estime la réponse impulsionnelle de la salle afin de l'inverser, des méthodes de suppression, où l'on estime la réverbération au sein du signal afin de la retirer du spectre. Ces dernières seront préférées aux méthodes d'inversion, moins stables et plus coûteuses en calculs.

Nous avons donc implémenté deux méthodes d'estimation et deux méthodes de suppression. La première se base sur le modèle stochastique de réponses de salles afin d'estimer la densité spectrale de puissance propre à la réverbération tardive. La seconde utilise la prédiction linéaire à long terme pour obtenir un filtre permettant d'estimer la partie tardive de la réverbération, à partir du signal réverbéré.

Une fois la densité spectrale de puissance obtenue, on la retire du spectre du signal réverbéré. Pour cela on peut utiliser un gain de soustraction spectrale minimisant le bruit musical à l'aide de seuils. Ou alors, un gain pondéré par la probabilité de présence de parole dans le signal, minimisant l'erreur du logarithme de l'amplitude du spectre.

Les résultats obtenus ont ensuite été analysés par des critères objectifs et subjectifs. Nous nous sommes rapproché petit à petit des conditions réelles, avec une réponse de salle synthétique puis réelle, des voix issues de corpus à des voix enregistrées avec les appareils *Invoxia*, afin de tester les méthodes. Ces dernières sont très satisfaisantes et permettent une utilisation en temps réel.

Enfin, j'ai pu me pencher sur l'utilité de la phase des signaux en déréverbération. Nous avons vu que l'ajout de la phase du signal anéchoïque permet également de déréverbérer,

mais aussi de résoudre des problèmes de distorsion non réglés par les techniques actuelles. Les premiers essais, encourageants, se basent sur la synthèse de la voix et seront à poursuivre au cours de ma thèse intitulée *Utilisation de la phase en déréverbération des signaux de parole.*

Bibliographie

[1] S. Affes and Y. Grenier. A signal subspace tracking algorithm for microphone array processing of speech. *Speech and Audio Processing, IEEE Transactions on*, 5(5) :425–437, Sep 1997.

[2] M. Aharon, M. Elad, and A Bruckstein. K -svd : An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, 54(11) :4311–4322, Nov 2006.

[3] J. B. Allen. Effects of small room reverberation on subjective preference. *The Journal of the Acoustical Society of America*, 71(S1) :S5–S5, 1982.

[4] Jont B Allen and David A Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4) :943–950, 1979.

[5] H. Attias, J. Platt, Alex Acero, and Li Deng. Speech denoising and dereverberation using probabilistic models. In *NIPS*, November 2000.

[6] D.A. Berkley. *Acoustical factors affecting hearing aid performance*, chap Normal listeners in typical rooms :pp 3–24, 1980.

[7] D.A. Berkley and O.M. Mitchell. Removing reverberative echo components in speech signals,. *U.S. Patent No. 4166924*, 1979.

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

[9] R. H. Bolt and A. D. MacDonald. Theory of speech masking by reverberation. *The Journal of the Acoustical Society of America*, 21(6) :577–580, 1949.

[10] M.S. Brandstein. On the use of explicit speech modeling in microphone array applications. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 6, pages 3613–3616 vol.6, May 1998.

[11] I Cohen. Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. *Signal Processing Letters, IEEE*, 9(4) :113–116, April 2002.

[12] Marc Delcroix, Takafumi Hikichi, and Masato Miyoshi. Blind dereverberation algorithm for speech signals based on multi-channel linear prediction. *Acoustical Science and Technology*, 26(5) :432–439, 2005.

[13] Yariv Ephraim and David Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 32(6) :1109–1121, 1984.

[14] N. D. Gaubitch, D. B. Ward, and P. A. Naylor. Statistical analysis of the autoregressive modeling of reverberant speech. *Journal of the Acoustical Society of America*, 120(6) :4031–4039, December 2006.

[15] Nikolay D. Gaubitch, Patrick A. Naylor, and Darren B. Ward. On the use of linear prediction for dereverberation of speech. In *In Proceedings of the IEEE International Workshop on Acoustic Echo and Noise Control*, pages 99–102, 2003.

[16] D. Gesbert and P. Duhamel. Robust blind channel identification and equalization based on multi-step predictors. In *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, volume 5, pages 3621–3624 vol.5, Apr 1997.

[17] Scott Griebel and Michael Brandstein. Wavelet transform extrema clustering for multi-channel speech dereverberation. In *IN IEEE WORKSHOP ON ACOUSTIC ECHO AND NOISE CONTROL, POCONO*, pages 27–30, 1999.

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

[19] E. Habets. Multi-channel speech dereverberation based on a statistical model of late reverberation. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 4, pages iv/173–iv/176 Vol. 4, March 2005.

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

[21] E. A. P. Habets, S. Gannot, and I. Cohen. Dual-microphone speech dereverberation in a noisy environment. In *Proc. IEEE Intl. Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 651–655, Vancouver, Canada, August 2006.

[22] J. Hardwick, C.D. Yoo, and J.S. Lim. Speech enhancement using the dual excitation speech model. In *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, volume 2, pages 367–370 vol.2, April 1993.

[23] S. Haykin. *Blind Deconvolution*. Prentice Hall information and system sciences series. Prentice Hall, 4th ed. edition, 1994.

[24] S. Haykin. *Unsupervised Adaptive Filtering*. John–Wiley and Sons, 2nd ed. edition, 2000.

[25] Yi Hu and P.C. Loizou. Evaluation of objective quality measures for speech enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(1) :229–238, Jan 2008.

[26] Patrick A. Naylor Jimi Y.C Wen. An evaluation measure for reverberant speech using decay tail modelling. In *EUSIPCO*, Florence, Italy, September 2006.

[27] J.G. Proakis J.R. Deller and J.H.L. Hansen. *Discrete-Time Processing of Speech Signals,*. MacMillan ; New York, 1993.

[28] Keisuke Kinoshita, Marc Delcroix, Tomohiro Nakatani, and Masato Miyoshi. Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. *IEEE Transactions on Audio, Speech and Language Processing*, 17 :534–545, 2009.

[29] Keisuke Kinoshita, Tomohiro Nakatani, and Masato Miyoshi. Harmonicity based dereverberation for improving automatic speech recognition performance and speech intelligibility. *IEICE Transactions*, 88-A(7) :1724–1731, 2005.

[30] Kostas Kokkinakis and Philipos C Loizou. The impact of reverberant self-masking and overlap-masking effects on speech intelligibility by cochlear implant listeners. *J Acoust Soc Am*, 130(3) :1099–102, 2011.

[31] Heinrich. Kuttruff. *Room acoustics / Heinrich Kuttruff.* Elsevier Applied Science London ; New York, 5th ed. edition, 2009.

[32] J. Laroche, Y. Stylianou, and E. Moulines. Hns : Speech modification based on a harmonic+noise model. In *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, volume 2, pages 550–553 vol.2, April 1993.

[33] K Lebart and JM Boucher. A new method based on spectral subtraction for speech dereverberation. *ACUSTICA*, 87(3) :359–366, 2001.

[34] F. Razzazi M. Moshirynia and A. Haghbin. Speech dereverberation method using adaptive sparse dictionary learning. *Speech and Audio Processing, IEEE Transactions on*, 2014.

[35] John Makhoul. Linear prediction : A tutorial review. *Proceedings of the IEEE*, 63(4) :561–580, 1975.

[36] R. McAulay and T.F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 34(4) :744–754, Aug 1986.

[37] M. Miyoshi. Estimating ar parameter-sets for linear-recurrent signals in convolutive mixtures. pages 585–589, Nara, Japan, April 2003.

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

[39] James A. Moorer. About this reverberation business. *Computer Music Journal*, 3(2) :pp. 13–28, 1979.

[40] J. Mourjopoulos and J. Hammond. Modelling and enhancement of reverberant speech using an envelope convolution method. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '83.*, volume 8, pages 1144–1147, Apr 1983.

[41] P.A. Naylor and N.D. Gaubitch. Speech dereverberation. In *Proc. of the International Workshop on Acoustic Echo and Noise Control (IWAENC'05)*, 2005.

[42] Anna K. Nábělek, Tomasz R. Letowski, and Frances M. Tucker. Reverberant overlap and self masking in consonant identification. *The Journal of the Acoustical Society of America*, 86(4) :259–1265, 1989.

[43] V. M. A. Peutz. Articulation loss of consonants as a criterion for speech transmission in a room. *J. Audio Eng. Soc*, 19(11) :915–919, 1971.

[44] Jean-Dominique Polack. *La transmission de l'energie sonore dans les salles.* PhD thesis, 1988. Thèse de doctorat dirigée par Bruneau, Michel Physique Le Mans 1988.

[45] Jean-Dominique Polack. Playing billiards in the concert hall : The mathematical foundations of geometrical room acoustics. *Applied Acoustics*, 38(2–4) :235 – 244, 1993.

[46] L.E. Ryall. Improvements in electric signal amplifiers incorporating voice-operated devices,. *G.B. Patent No. 509613*, 1939.

[47] W.C. Sabine. *Collected papers on acoustics (originally 1921)*, 1993.

[48] C.D. Sigg, T. Dikk, and J.M. Buhmann. Speech enhancement using generative dictionary learning. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(6) :1698–1712, Aug 2012.

[49] D.E. Tsoukalas, J.N. Mourjopoulos, and G. Kokkinakis. Speech enhancement based on audible noise suppression. *Speech and Audio Processing, IEEE Transactions on*, 5(6) :497–514, Nov 1997.

[50] Jianchao Yang, J. Wright, T.S. Huang, and Yi Ma. Image super-resolution via sparse representation. *Image Processing, IEEE Transactions on*, 19(11) :2861–2873, Nov 2010.

[51] B. Yegnanarayana and P.S. Murthy. Enhancement of reverberant speech using lp residual signal. *Speech and Audio Processing, IEEE Transactions on*, 8(3) :267–281, May 2000.

[52] Bing yin Xia, Yan Liang, and Chang chun Bao. A modified spectral subtraction method for speech enhancement based on masking property of human auditory system. In *Wireless Communications Signal Processing, 2009. WCSP 2009. International Conference on*, pages 1–5, Nov 2009.

[53] C.D. Yoo and J.S. Lim. Speech enhancement based on the generalized dual excitation model with adaptive analysis window. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 832–835 vol.1, May 1995.

[54] Eberhard Zwicker and Hugo Fastl. *Psychoacoustics : Facts and Models (Springer Series in Information Sciences) (v. 22)*. 2nd updated ed. edition, apr 1999.