ENS Louis-Lumière
Mémoire de Fin d'Etudes
Département Son
Mai 2013

INTÉRÊTS ET DÉFAUTS DES PROCÉDÉS DE DÉ-RÉVERBÉRATION DANS LE TRAITEMENT DE LA VOIX AU CINÉMA

Nicolas **J**aillot

Directeur **I**nterne	**L**aurent **M**illot
Directeur **E**xterne	**C**yril **H**oltz
Rapporteur	**S**téphane **I**sidore

Remerciements

A mes deux directeurs Laurent Millot et Cyril Holtz pour leur soutien,

A Galaad Hemsi, Renaud Duguet et les Productions du Désert, pour *Le Premier Rôle,*

A Géraud Bec, Hassan Kamrani et Shaman Labs pour leur accueil chaleureux, leurs conseils et leur disponibilité,

A Michael Werner et Commune Image pour leur confiance,

A Claude Gazeau, Philippe Simonet, Matthieu Langlet et Lionel Thiriet pour leurs conseils et leur ouverture d'esprit,

A Zinedine Sadaoui, Nicholas Warwick et Radio France pour leur compréhension.

A Erwan Boulay, Florent Ollivier et Clément Grariel pour leur matériel,

A mes amis et ma famille pour tout le reste.

Résumé

La voix est un élément central dans bon nombre de bandes son au cinéma. Elle est un signal de nature complexe, propre à chaque individu et servant de support à la narration. Elle renseigne alors sur les lieux dans lesquels les personnages évoluent en adoptant leur signature acoustique. Si elle permet aussi la localisation d'une source, la réverbération excessive de certaines prises de sons peut parfois nuire l'intelligibilité du discours, ou ne pas satisfaire aux exigences de spatialisations. Par quels moyens le mixeur d'une bande son peut-il enlever la réverbération présente sur une voix lors d'une captation monophonique distante ?

Ce mémoire passe en revue l'éventail des possibilités offertes à l'ingénieur du son pour répondre à ce problème. En analysant les outils existants et en proposant l'utilisation de la dé-convolution sur un cas concret, il fait apparaître les contraintes propres à ces applications dans le domaine du son à l'image.

Abstract

For most feature films, the voice is the most important element of the soundtrack. It's a complex signal that customizes each character and supports the storytelling. The voice also gives much information on the film locations because it merges with their acoustical print. If this reverberation gives localization clues to the audience, an over-reverberated signal can lead to an intelligibility loss or an inappropriate spatialization. How the sound mixer can remove reverberation from a monophonic voice recording?

This study focuses on the possibilities available to the sound engineer to dereverberate a signal in these conditions. After analysing the existing tools, we tried to apply a deconvolution-based method on a practical case, to finally underline the limitations of these approaches in the movie postproduction.

“Que dites-vous ?... C'est inutile ?... Je le sais ! Mais on ne se bat pas dans l'espoir du succès ! Non ! non, c'est bien plus beau lorsque c'est inutile !”

Cyrano de Bergerac, Edmond Rostand, acte V, scène 6.

Table des Matières

Introduction

Au cours des dernières décennies, une multitude d'outils ont étés élaborés afin d'ajouter ou de simuler des réflexions acoustiques à un message sonore. Cela tient au fait que la réverbération est une caractéristique essentielle d'un enregistrement, reflétant la vie acoustique des sons captés. Elle joue un rôle majeur dans la localisation des sources ou dans l'identification des espaces. Cependant si l'ajout de réverbération a rapidement été possible dans l'histoire des techniques sonores (chambre d'écho, réverbérateurs à ressorts ou à plaques), le processus inverse visant à réduire, voire à supprimer ces réflexions, constitue toujours un réel défi.

Parvenir à soustraire la réverbération pourrait engendrer d'importants changements dans la manière de traiter une source. Cela élargirait les possibilités de spatialisation offertes au mixage tout en permettant de compenser d'éventuels défauts d'intelligibilité. Au cinéma ces questions sont actuellement résolues par un recours à la postsynchronisation. On tente alors d'intégrer les voix réenregistrées en studio par l'ajout de réverbération et de bruitage pour reconstituer la scène sonore.

Cependant les difficultés éprouvées par les acteurs et le surcoût engendré par l'abandon du son capté au moment du tournage, en font une étape redoutée par bon nombre de productions. De plus le réenregistrement des voix en studio n'est pas toujours possible et il est difficilement concevable d'avoir recours à la postsynchronisation lorsqu'il ne s'agit pas d'un film de fiction.

Si dès les années 60 les laboratoires Bell se sont penchés sur les techniques de réduction de bruit, l'étude de la dé-réverbération a réellement pris son essor au cours des 20 dernières années avec la téléphonie et la reconnaissance vocale. Cependant, c'est seulement en 2010 que le premier outil spécifiquement destiné à la post-production a fait son apparition. Quelles sont aujourd'hui les possibilités d'actions pour supprimer la réverbération ?

Sont-elles transposables au traitement de la parole au cinéma ? Souvent considérée comme le « squelette du mixage », la voix est dans ce cas soumise à des exigences qualitatives très élevées.

Après avoir étudié les principes théoriques mis en œuvre dans les outils actuellement disponibles en studio, nous tenterons de mettre en place un traitement par dé-convolution. Bon nombre de studios sont aujourd'hui pourvus de convolueurs, permettant d'utiliser l'empreinte acoustique du tournage pour replacer la source anéchoïque dans ce lieu. Serait-ce possible de détourner ces outils de manière à supprimer la réverbération ? Notre étude s'articulera autours du film *Premier Rôle* réalisé par Galaad Hemsi. Nous nous rendrons dans le lieu du tournage pour capturer la réponse impulsionnelle du lieu et ainsi tenter d'annuler l'effet de ces réflexions.

1. MISE EN SITUATION

Dans ce chapitre nous commencerons par décrire les mécanismes donnant naissance à la voix. Nous suivrons sa propagation, pour comprendre comment le lieu de diffusion influe sur ce message. Nous traiterons enfin de la réverbération et de son influence sur l'intelligibilité et la perception de l'espace sonore. Ces considérations nous amèneront à mieux cerner les enjeux de notre étude pour envisager la dé-réverbération de la voix dans le prochain chapitre.

1.1. La voix

La voix caractérise l'ensemble des sons susceptibles d'être produits par l'Homme, elle est l'organe de la parole et du chant. Si elle résulte d'un mode de production qui nous est commun, ses caractéristiques sont très variables suivant les individus ce qui rend son traitement délicat.

Aussi pour la caractériser plusieurs niveaux d'étude sont possibles. On parle d'une description acoustique lorsque l'on s'intéresse à la caractérisation objective du signal véhiculé par la voix. Si la voix est le support de la parole, nous pouvons faire intervenir la phonétique, qui se concentre sur la manière dont le son est mis en forme par le système articulatoire et si l'on considère une description de plus haut niveau on fait appel phonologie qui est l'interface entre la phonétique et les autres formes de caractérisations du langage (analyses morphologique, syntaxiques, sémantiques et pragmatiques). La phonologie définit le concept de phonème, comme la plus petite unité discrète du discours qui même s'il n'a pas d'existence propre, constitue pourtant le matériau de base à la création du langage. Chaque langue comporte en effet un nombre limité de phonèmes dont la mise en relation est porteuse de sens et constitue un mot.

1.1.1 Mécanismes de production de la parole

La parole est le résultat de la modulation du flux d'air généré par l'appareil respiratoire. Poussé par le diaphragme hors des poumons, l'air s'écoule à travers la trachée artère puis traverse le larynx sur lequel on rencontre les cordes vocales. La taille de l'ouverture formée par les cordes vocales (appelée glotte) est variable. Lorsque la voix est chuchotée l'air y circule librement on dit alors qu'il s'agit d'un signal « non-voisé ». Si la glotte se resserre, une différence de pression de part et d'autre des cordes vocales fera apparaître une vibration pseudopériodique de la colonne d'air. Ce phénomène sera alors amplifié par les cavités buccales, nasales et modulé par le système articulatoire (langue, dentition, lèvres).

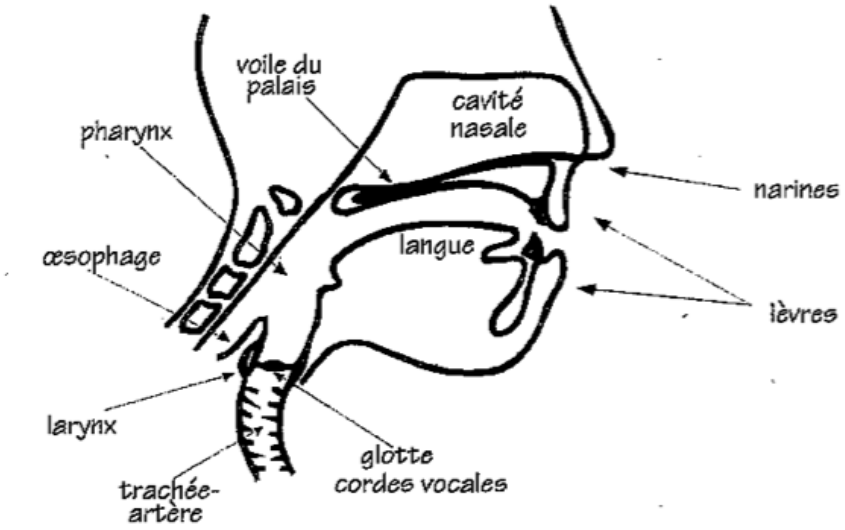


Fig 1-1. *Vue en coupe de l'appareil phonatoire, d'après [1] p.13.*

1.1.2 Outils de description

Certains ouvrages considèrent communément que le contenu spectral de la voix peut s'étendre de 50 Hz à 15 000 Hz. Le signal vocal, lorsqu'il est harmonique est régit par la morphologie de chaque individu et classifié en musique par la notion de tessiture. On considère alors qu'une voix de basse chantée comporte des fondamentales comprises entre 80 Hz (Mi 1) et 350 Hz (Ré 3) alors que celles d'une soprano s'étendent sur une plage allant de 260 Hz (Do 3) à environ 1700 Hz (Fa 5 – Sol5).

L'excitation des résonateurs pharyngien, buccal et nasal donne lieux à des harmoniques qui constituent le matériau de base à la création des voyelles. Ces résonnances sont appelées formants et chaque voyelle est alors caractérisée comme une combinaison linéaire des deux premiers formants. Le premier est proportionnellement lié à l'ouverture de la bouche, le second à la position de la langue tandis que les harmoniques supérieures participent au timbre de la voix.

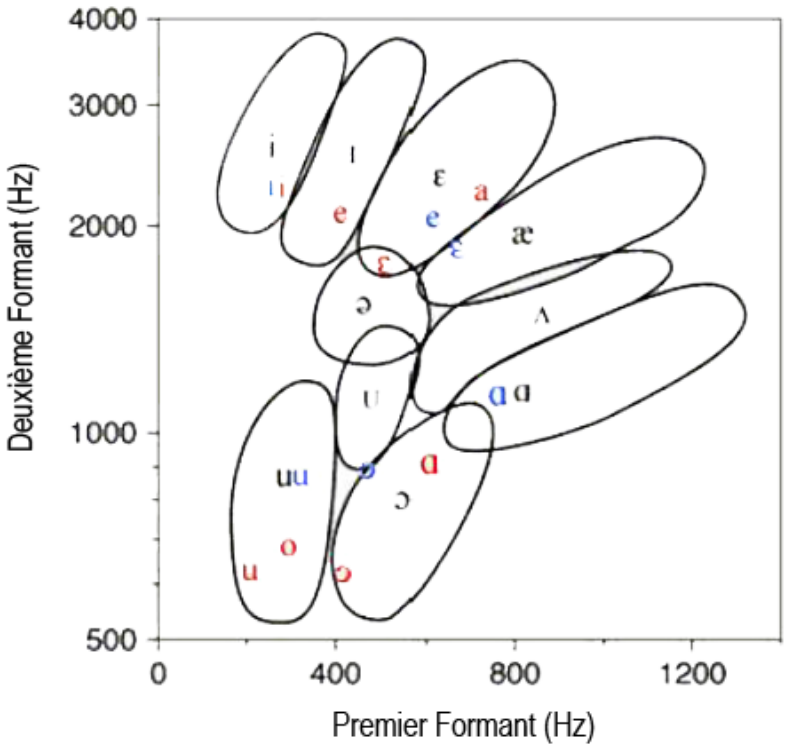


Fig 1-2. *Représentation de L'alphabet phonétique international dans le plan Formantique F1-F2, d'après [2].*

Lorsque l'écoulement de l'air est obstrué ou modifié par le système articulatoire, il y a production d'une consonne. Par exemple l'occlusion totale et temporaire du conduit génère une plosive (p, t, k, b, d, g). Ces modifications produisent des signaux transitoires sans réelles fréquences fondamentales et on parle alors plus volontiers de coloration, notamment dans le cas de sifflantes (ou fricatives) pour (f, s, ch, v, j, z) qui peuvent générer un contenu fréquentiel au delà de 15 000 Hz. Comme il s'agit de signaux de faible énergie souvent situés 30 à 40 dB en dessous du niveau des voyelles, les

consonnes sont donc des éléments de langages fragiles et difficiles à traiter. Si par exemple, il reste possible de comprendre un texte enregistré auquel nous avons retiré les voyelles. Il s'avère en revanche beaucoup plus délicat de comprendre un message sans les consonnes.

1.1.3 Rayonnement et propagation

La voix est une source directive. Le diagramme de la figure 1-3 montre les variations spectrales constatées dans les plans de l'azimut et de l'élévation. Ces mesures ont été réalisées en chambre semi-anéchoïque par A. H. Marshall et Meyer, J. dans le cadre d'une étude intitulée "*The Directivity and Auditory Impressions of Singers*". On constate alors que le spectre est le plus homogène pour un azimut de ±30° et pour une élévation de -30°.

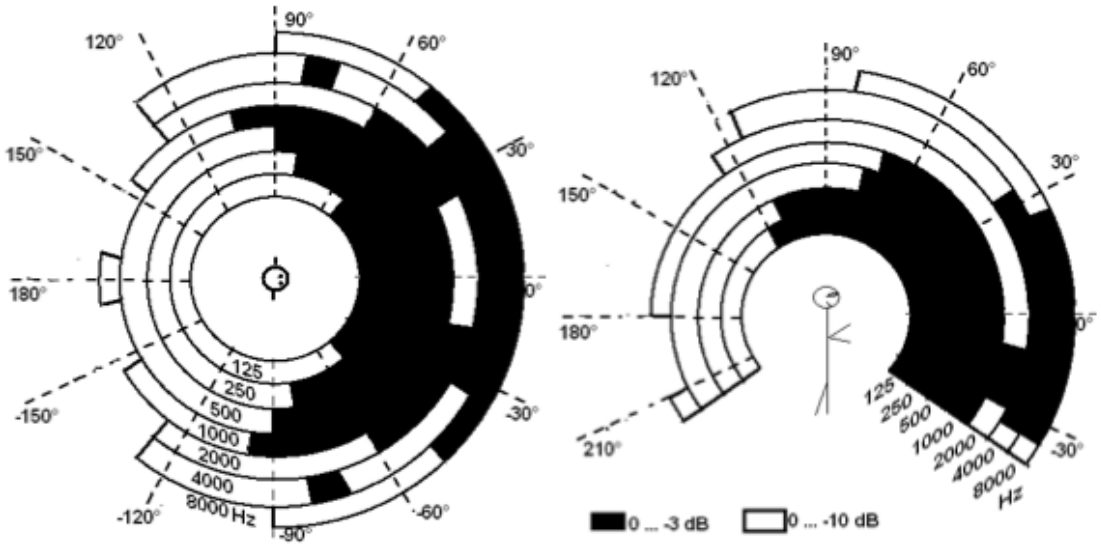


Fig 1-3. *Rayonnement de la voix, d'après [3].*

Or dans le cas d'une captation cinématographique une telle configuration de prise de s'avère difficilement envisageable.

Ce diagramme souligne l'importance du milieu environnant dans la diffusion la voix puisqu'ils sont réfléchissants, le sol, les murs latéraux et le plafond dans une moindre mesure joueront un rôle important sur la manière dont la source sera captée ou perçue.

1.2. Caractérisation de l'effet de salle

1.2.1 Principes Elémentaires

Les réflexions tridimensionnelles auxquelles nous venons de faire référence, engendreront d'abord une modification du timbre de la source puis des réflexions multiples qui viendront s'ajouter au signal d'origine, modifiant alors l'enveloppe du signal.

Nous pouvons illustrer ces problématiques par un le cas simple de la figure 1-4 où nous considérons une configuration de captation en plein air d'une voix dont le locuteur se trouverait sur un sol entièrement réfléchissant.

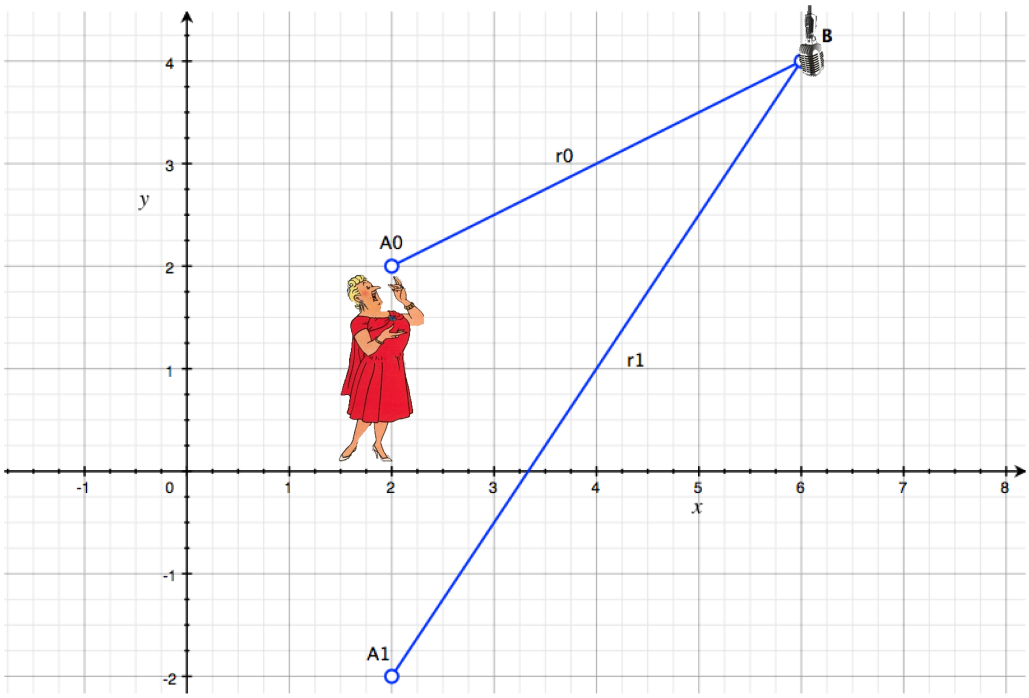


Fig. 1-4 *Configuration de prise de son en plein-air avec sol réfléchissant.*

Soient A_0 le point source, A_1 la source image formée par la réflexion sur le sol et B le point de captation.

$$A0 = \begin{cases} \text{xa0} \\ \text{ya0} \end{cases} ; A1 = \begin{cases} \text{xa0} \\ -\text{ya0} \end{cases} ; B \begin{cases} \text{xb0} \\ \text{yb0} \end{cases}$$

Si $S_A(t)$ est le son émis au point A, on capte au point B : $S_B(t) = S_A(t - \tau_0) + S_A(t - \tau_1)$

Où ∂_0 et ∂_1 correspondent au retard de propagation des distances parcourues (respectivement r_0 et r_1) à la vitesse c_0.

$$t0 = \frac{\sqrt{(\mathrm{xb0} - \mathrm{xa0})^2 + (\mathrm{yb0} - \mathrm{ya0})^2}}{c0}$$

$$t1 = \frac{\sqrt{(\mathrm{xb0} - \mathrm{xa0})^2 + (\mathrm{yb0} + \mathrm{ya0})^2}}{c0}$$

On obtient alors un effet de filtrage en peigne par la sommation de deux signaux de phase différente. Après numérisation du signal à la fréquence d'échantillonnage Fs, on peut aussi écrire :

$$S_B[n] = S_A[n-\Delta_0] + S_A[n-\Delta_1] \ ; \ (Avec\ \Delta_i = \tau i * Fs)$$

Puis après passage à la transformée en Z :

$$S_B = S_A * (Z^{-\Delta 0} + Z^{-\Delta 1})$$

On voit ainsi apparaitre la structure de filtrage doné en figure 1-5.

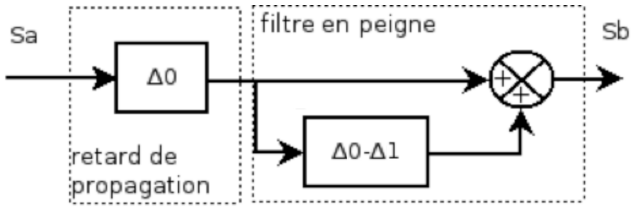


Fig 1-5. *Structure de filtre en Peigne*

En ayant recours à la transformée de fourrier à temps discret on peut alors déterminer le gain du filtre en peigne :

$$S_B = S_A * Z^{-\Delta 0} (1 + Z^{\Delta 0 - \Delta 1})$$

$$H[f] = e^{\frac{\pi f(\Delta 0 - \Delta 1)}{Fs}} (e^{\frac{-\pi f(\Delta 0 - \Delta 1)}{Fs}} + e^{\frac{\pi f(\Delta 0 - \Delta 1)}{Fs}})$$

$$H[f] = 2e^{\frac{\pi f(\Delta 0 - \Delta 1)}{Fs}} \cos\left(\frac{\pi f(\Delta 0 - \Delta 1)}{Fs}\right)$$

Ce qui permet d'écrire pour le gain

$$|H[f]| = 2\left|\cos\left(\frac{\pi f(\Delta 0 - \Delta 1)}{Fs}\right)\right| = 2|\cos(\pi f(\tau 0 - \tau 1))|$$

En utilisant les valeurs indiquées sur le schéma, cela signifie que le spectre de la voix du notre chanteuse sera modifié par la fonction tracée en figure 1-6.

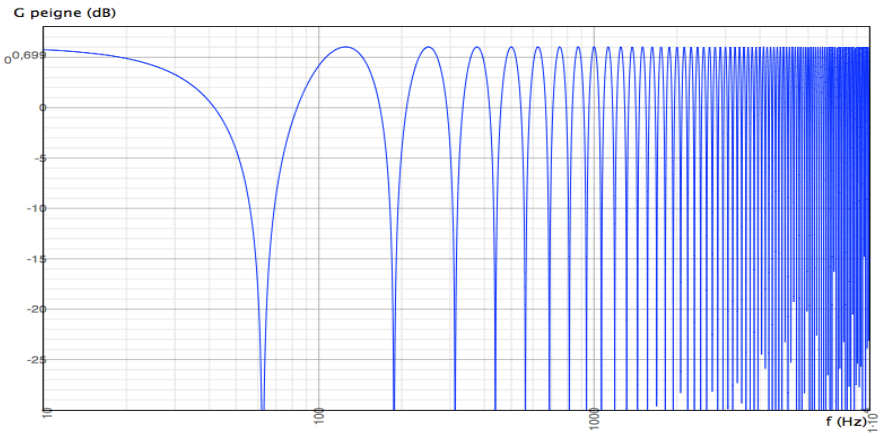


Fig 1-6. *Allure du gain du filtre en peigne.*

Ici nous avons considéré la source comme le point de captation fixes. Si ce n'est plus le cas, les fréquences de coupure de la figure 1-6 seront amenées à évoluer de manière continue, proportionnellement à la vitesse du déplacement. De plus nous avons traité uniquement le cas de la première réflexion sans prendre en compte la directivité de la source, ni l'atténuation due à son rayonnement, à l'absorbation du milieu de propagation, ou celle du sol...

On voit alors apparaître les notions de son direct et de champ réverbéré. Le son direct correspond à la portion du signal qui se propage librement, sans rencontrer d'obstacle engendrant une réflexion tandis que le champ réverbéré est constitué par opposition de l'ensemble des réflexions émises par les différentes surfaces excitées. On dit qu'il y a écho franc lorsque que la première réflexion arrive suffisamment tard (40 ms) pour ne plus fusionner avec le son direct. Notre cerveau interprète cette information comme une source totalement dé-corrélée du son direct.

Si l'on suppose que l'on place la cantatrice dans une boite close suffisamment grande tout en conservant le même système de captation, on observe alors des réflexions d'ordre supérieur, s'atténuant progressivement comme illustré à la figure 1-7.

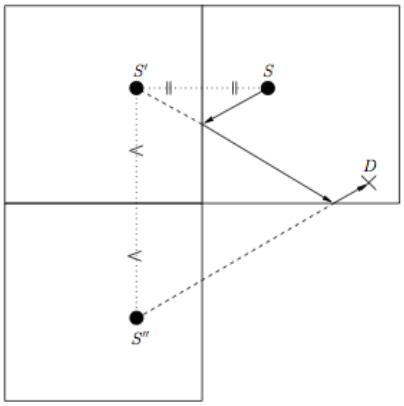


Fig 1-7. D*eux premières sources Images – Pattern 2D*

Cette situation, se reproduisant théoriquement infiniment en trois dimensions, est régie dans la pratique par la décroissance énergétique liée au milieu de propagation ainsi qu'à l'absorbation de chacune des parois. On voit alors apparaître une multitude de réflexions qui génèrent alors le champ diffus.

Une manière usuelle de représenter ces phénomènes dans le domaine temporel d'avoir recours à l'échogramme qui permet de distinguer en théorie (cf. figure 1-8) le champ direct, les premières réflexions et le champ diffus constituants du champ réverbéré.

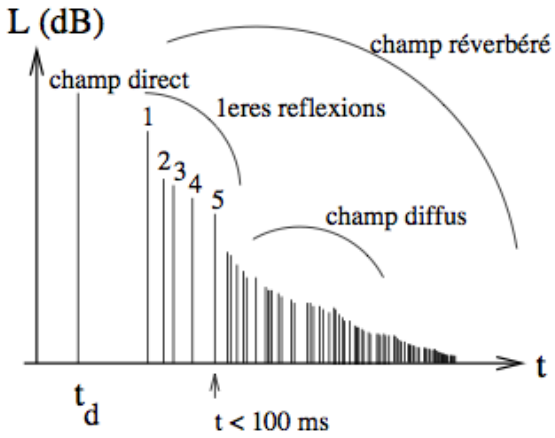


Fig 1-8. *Topologie du champ réverbéré.*

Pour une salle (ou une acoustique) donnée, on définit la distance critique comme le résultat de la situation ou les points de captation et de diffusion sont espacés d'une distance telle que l'énergie du champ réverbéré est identique à celle du son direct.

1.2.2 Evaluation subjective

L'effet de la réverbération est avant tout sensible et elle est une des caractéristiques du lieu dans lequel le son est produit. Nos habitudes d'écoute permettent alors de contextualiser un son pour lui imaginer une vie acoustique propre sans connaître a priori ses conditions de création.

A la fin des années 1980 l'acousticien Leo Beranek a réalisé en se basant sur ce principe sur ce principe, une campagne de caractérisation des salles de concerts les plus réputées, se utilisant des paramètres subjectifs. A partir d'enquêtes réalisées avec des musiciens et des chefs d'orchestres, il a proposé un jeu de critères pour qualifier l'ensemble des lieux évoqués. Il a aussi essayé de mettre en relation ces observations subjectives avec des critères objectifs, permettant de mesurer, ou d'anticiper la réponse de chaque salle...

- La **réverbérance** : qualifie la vitesse de décroissance de l'énergie acoustique. Elle est à relier avec le TR60, temps mis par le champ réverbéré pour décroitre de 60 dB.

- La **vivacité** : correspond généralement au temps de décroissance des composantes fréquentielles médium, comprises en 340 Hz et 1400 Hz.

- La **sensation d'espace** : impression de largeur créée par les premières réflexions. La source parait plus large qu'elle n'est en réalité. Les acousticiens peuvent la mesurer avec l'Index de Qualité Binaural (BQI).

- La **force sonore** : sensation (exprimée en décibels) liée au volume sonore disponible dans la salle. Elle peut être mesurée globalement à l'aide d'un sonomètre ou ne prendre en compte que le champ direct et ses premières réflexions.

- L'**agressivité** : reflète le caractère diffusant des surfaces responsables des premières réflexions. Des géométries planes créeront davantage d'agressivité que des surfaces irrégulières.

- **Timbre et couleur tonale** : décrit l'équilibre spectral de la salle. Ce critère nécessite le recours à une analyse fréquentielle.

- **Clarté** : Caractérise la bonne compréhension du message. Une clarté à 50 ms et une autre à 80 ms ont été définies comme le rapport de l'énergie présente au cours des 50 ou 80 premières millisecondes comparée à l'énergie associée à la durée globale.

- **L'enveloppement** : traduit la sensation d'immersion physique dans l'espace concerné. Plus l'enveloppement est grand, plus l'auditeur a la sensation que le champ réverbéré vient de toute part et pas seulement d'une direction. Il est possible d'avoir recours au coefficient d'inter-corrélation binaural pour qualifier l'enveloppement.

1.2.3 Avantages d'une perception binaurale

Nous venons de voir que la sensation d'enveloppement est très liée à la perception binaurale. Une corrélation marquée traduit un faible enveloppement ; les signaux captés par l'oreille gauche et par l'oreille droite sont quasiment identiques. Leur trajet est donc sensiblement le même et l'on peut alors parler de source monophonique. A l'inverse une corrélation nulle suppose la présence d'un champ très diffus et des signaux gauche/droite très différents.

Ce principe d'inter-corrélation est aussi celui mis en œuvre dans le bien connu effet « cocktail party ». En comparant les signaux de nos deux oreilles, nous sommes capables de focaliser notre attention sur un locuteur en particulier dans un milieu bruité. Certaines techniques mettent en œuvre ce principe en élaborant des antennes microphoniques visant à supprimer la réverbération. Grâce à ces techniques,

il est possible de déterminer la part significative de son direct présent entre l'ensemble des microphones constituant l'antenne. Nous n'étudierons pas ces procédés car même si les résultats théoriques de ces méthodes semblent encourageant, ces procédés sont difficilement transposables à l'enregistrement sonore sur un tournage (ils sont souvent encombrants, couteux et long à mettre en place)...

1.2.4 Evaluation de l'intelligibilité

La réverbération modifie aussi les propriétés dynamiques du signal et donc sa courbe Attack Decay Sustain Release (ADSR). Comme l'illustre la figure 1-9, plus la proportion de signal réverbérée sera importante, plus les transitoires seront lissés, les zones de maintien et de décroissances prolongées. L'énergie est alors répartie dans le temps et le facteur de crête (Xpeak/Xrms) diminue.

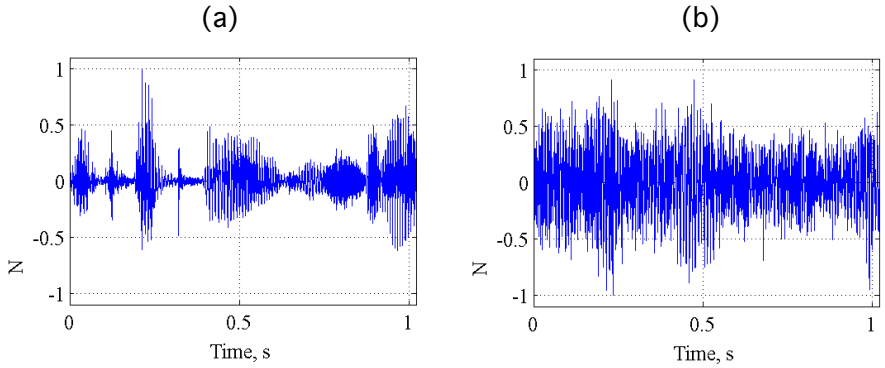


Fig. 1-9 *Effet dynamique de la réverbération - (a) signal d'origine (b) signal réverbéré.* Facteurs de crêtes associés : (a) 17.4dB, (b) 12.2dB.

Dans la section précédente (section 1.1) nous avons vu que les consonnes étaient des informations transitoires essentielles à la compréhension du discours, situés 30 à 40 dB en dessous du niveau des voyelles. Dans certains cas, la réverbération peut conduire à la disparition des consonnes par masquage temporel et comme illustré en figure 1-10, ce phénomène est accentué si le pré-délai (écart temporel séparant les premières réflexions du son direct) est court, que le temps de réverbération est long et que le point de captation est situé au delà du rayon critique.

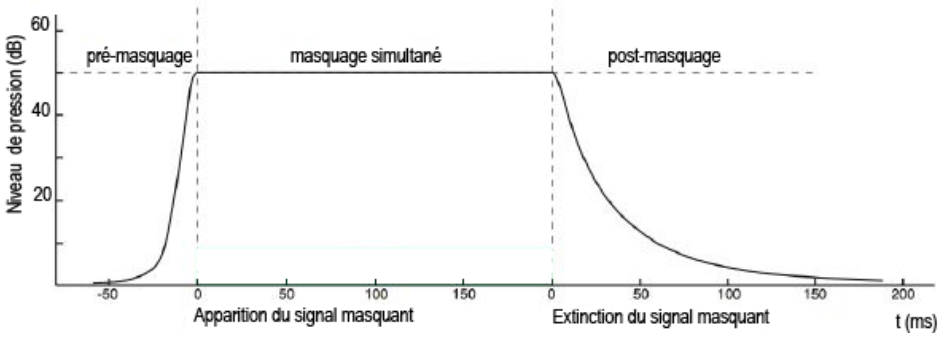


Fig 1-10. *Courbe MPEG de masquage temporel, D'après [4] p. 170.*

Les figures 1-10 et 1-12 ont été obtenues dans le cadre des recherches du Movie Picture Expert Group lors du développement des algorithmes de réduction de débit. En 1-10 on illustre le seuil d'audibilité dans le domaine temporel engendré par la présence d'un son masquant sur un période de 200 ms. Si le masquage est le plus important au moment où le son masquant et présent, les zones précédant et succédant cet événement sont elles aussi affectées. La perception des éléments transitoires est donc modifiée par la présence d'un signal masquant.

La figure 1-9 est l'illustration de ce phénomène. Elle représente temporellement un fragment de voix (a) convolué à la réponse impulsionnelle (captée par la société Audio Ease) de la cathédrale de chartes à 19 m sans pré-délai. La prolongation de la zone de maintient engendré par la réverbération peut être interprétée comme la prolongation du signal masquant et donc une modification du seuil de discrimination temporel.

Comme beaucoup de réverbérations dites « naturelles », le contenu spectral de cette réponse impulsionnelle s'atténue rapidement à mesure que la fréquence augmente. C'est ce qu'illustre la figure 1-11.

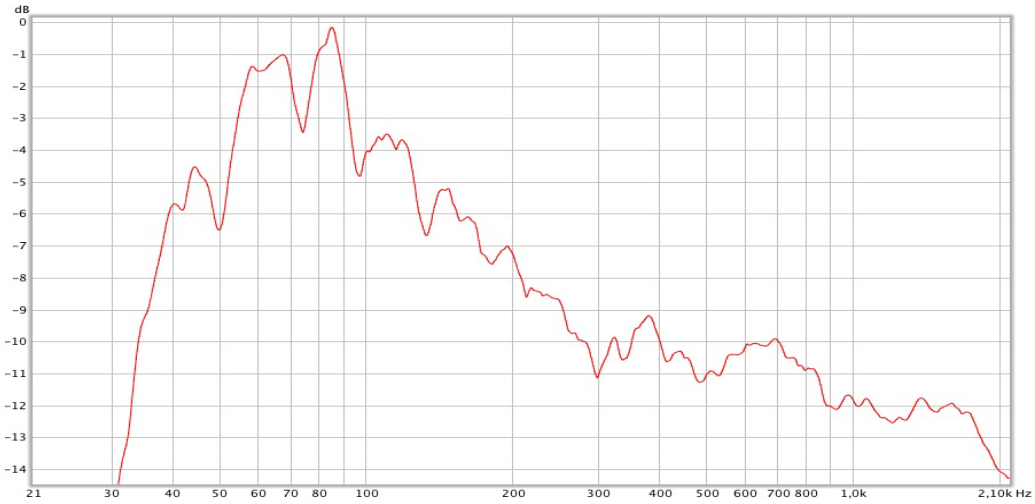


Fig. 1-11 Réponse Fréquentielle – Cathédrale de Chartres.

Ce phénomène est lié à l'absorbation de l'air et des surfaces réfléchissantes et cet effet s'accentue à mesure de l'accroissement des distances parcourues par chacun des rayons acoustiques. Ainsi, de part la nature de son contenu spectral, la réverbération est susceptible d'engendrer d'importants phénomènes masquage fréquentiel.

La figure 1-12 montre l'évolution de ce seuil d'audibilité pour des sons purs. Un signal sinusoïdal à 70 Hz engendre une modification du seuil de perception dans la bande fréquentielle représentée bleu. On remarque alors que les graves sont plus masquant que les aigus.

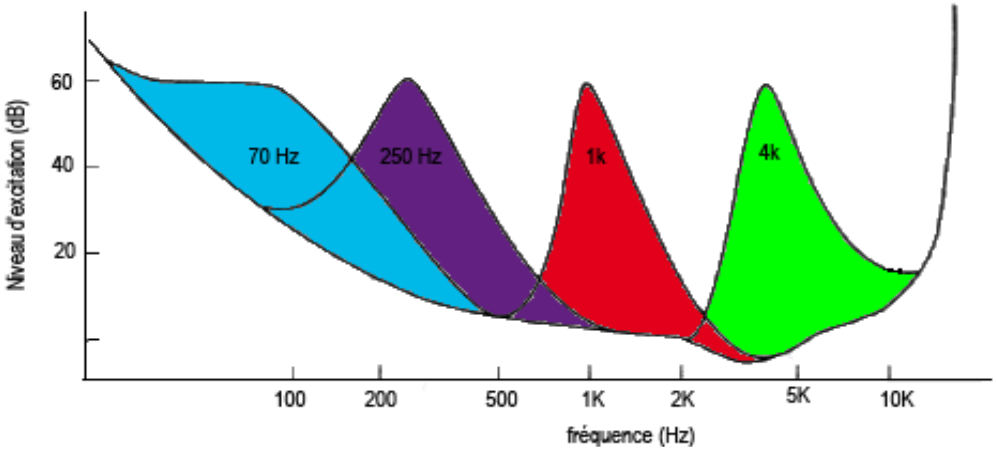


Fig 1-12. *Courbe MPEG de masquage fréquentiel, D'après [4] p. 168.*

De ce fait la réverbération peut nuire à la compréhension du message et donc dégrader l'intelligibilité. Aussi le gain d'intelligibilité est souvent considéré comme l'aboutissement des travaux de recherches visant à dé-réverbérer un signal. Puisque sauf volonté esthétique contraire, la compréhension du dialogue est une nécessité au cinéma, la question de la mesure de l'intelligibilité d'une voix dans un mixage est une question importante pour les applications que nous visons dans cette étude.

L'évaluation précise de l'intelligibilité s'avère un sujet relativement complexe, qui peut s'effectuer selon des orientations subjectives, objectives ou prédictives. Même dans ces deux derniers cas, ces mesures font appel à des facteurs cognitifs puisqu'il s'agit de noter la quantité d'informations comprises par un auditeur, ce qui explique l'existence d'une variabilité sensible. Ces résultats dépendent du degré de familiarité avec la langue, ou de l'acuité auditive du sujet. On a alors recours aux statistiques pour limiter ces écarts. Parmi ces critères objectifs, il existe une donnée acoustique dont le but est de lier le caractère réverbérant d'un lieu à la perte des consonnes. Il s'agit de l'ALCONS (Articulation Loss Consonants) souvent exprimé en pourcentage.

$$\%ALcons = \frac{200\, r^2\, T_{60}{}^2\, (1+n)}{V\, Q\, M}$$

Avec r la distance de la source la plus proche, n le nombre de sources, V le volume de la pièce, Q le facteur de directivité de la source la plus proche et M un coefficient image de la puissance du champ réverbéré. Plus ce pourcentage est faible, meilleur sera l'intelligibilité.

Ce critère nous permettra rarement de qualifier l'intelligibilité d'une voix dans le cadre de la postproduction car il ne prend en compte ni la présence de bruit de fond ni de l'effort d'articulation du locuteur. Dans les années 40, le développement de la téléphonie a conduit à la création de L'Articulation Index (AI) dans le but de qualifier la qualité de transmission. Il s'agit d'une mesure du rapport signal à bruit sur 20 bandes de fréquences pondérées pour prendre en compte les phénomènes de masquage. Cet index a connu diverses améliorations qui ont conduit au Speech Intelligibility Index (SII) puis au Speech Transmission Index (STI). Le SII applique une pondération différente et prend en

compte l'effort d'articulation du locuteur et comme le AI il s'adapte d'avantage à des transmissions uniformément bruitées (soit dans notre cas, des mixages à l'esthétique très affirmée) ! Le calcul du STI se base sur le calcul d'une fonction de modulation de transfert (FMT) pour sept octaves. Cette FMT est l'image de l'effet nivelant de la réverbération sur l'enveloppe du signal, elle s'exprime comme le rapport de la transformée de fourrier du carré de la réponse impulsionnelle du canal de transmission sur l'énergie totale, comme le montre l'équation suivante :

$$m(\Omega) = \frac{\int_0^\infty h^2(t)e^{-i\omega t}dt}{\int_0^\infty h^2(t)dt}$$

Le calcul objectif de l'intelligibilité nécessite alors la connaissance de la réponse impulsionnelle du canal acoustique dans son ensemble.

Les protocoles subjectifs permettent de s'affranchir de ces mesures intrusives. Ils consistent bien souvent en la reconnaissance de phrases, de mots ou de syllabes porteuses de sens ou totalement incohérentes. Mais comme ils nécessitent un panel d'auditeurs important, ces études subjectives sont souvent difficiles à réaliser. Puisque la difficulté de réside dans la lutte contre les phénomènes d'habitude, ces tests sont généralement peu reproductibles. Aussi pour certaines applications (notamment la téléphonie) la mesure d'intelligibilité est parfois confiée à un automate de reconnaissance vocal capable de nous renseigner sur le taux d'erreur de mot. La figure ci-dessous indique les relations entre les principaux tests subjectifs et les STI mesurés.

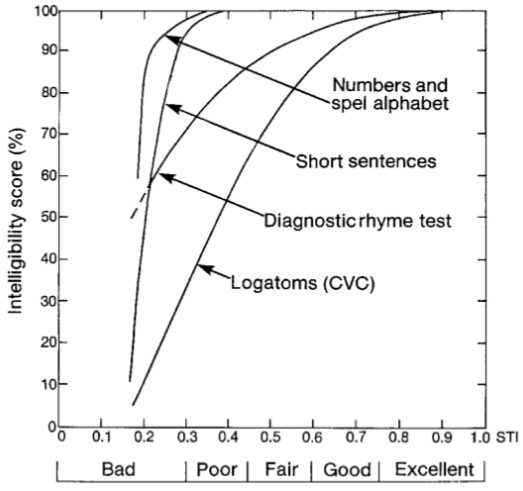


Fig. 1-13 *Evaluation perceptive et objective de l'intelligibilité, d'après [5] p.4.*

1.3. Caractérisation du canal acoustique

Nous venons de dresser un rapide portrait du canal acoustique en partant de l'émission de la voix d'un locuteur pour caractériser l'effet de salle et enfin tenter de définir des outils pour mesurer l'intelligibilité. L'accroissement de la puissance de calcul des ordinateurs a permis au cours de ces dix dernières années le recours à la convolution pour simuler la réponse impulsionnelle du lieu, en un point de captation. Ce sujet ayant déjà été traité (notamment dans plusieurs mémoires de fin d'étude), notre but n'est pas ici de décrire les apports de la convolution dans le monde de l'audio, mais simplement de résumer les bases théoriques nécessaires pour envisager la convolution comme moyen de dé-réverbération. Pour plus d'informations le lecteur pourra par exemple se rapporter aux mémoires de Mathieu Langlet (ENSLL 2002) ou de Guillaume Couturier (ENSLL 2010).

1.3.1 Principe

Un système linéaire et invariant dans le temps peut être décrit notamment grâce à sa réponse impulsionnelle, que l'on note h(t) lorsque l'on considère un système analogique et h[n] dans le cas numérique.

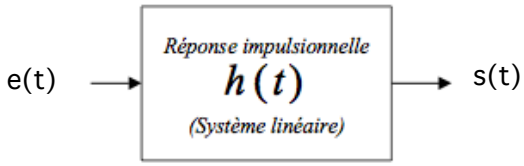


Fig. 1-15 : *Description d'un système linéaire.*

h(t) est définit théoriquement comme la réponse du système à une distribution de Dirac, impulsion infiniment courte et d'énergie infinie.

Dans le domaine analogique l'équation de convolution s'écrit :

$$s(t) = \int_{-\infty}^{+\infty} h(\tau).e(t-\tau)d\tau$$

après numérisation du signal, on a dans le domaine discret :

$$\begin{cases} s[n] = \sum_{m=0}^{n} h[m]\, e[n-m] & si\ n < Nh-1 \\ s[n] = \sum_{m=0}^{Nh-1} h[m]\, e[n-m] & sinon \end{cases}$$

1.3.2 Limitations pratique de la convolution

Le domaine temporel

Une implémentation directe de l'équation ci-dessus s'avère très coûteuse en terme de calcul. Comme le décrit par exemple Guillaume Couturier [6], la convolution directement dans le domaine temporel d'une réponse impulsionnelle de 2 secondes engendre $7.7*10^9$ opérations par seconde (à 44.1kHz). Si l'on raisonne en terme de cycles processeur (pour lesquels la multiplication s'effectue en plusieurs phases), ce nombre peut être bien plus important. L'utilisation de la convolution temporelle dans le monde de l'audio reste pour l'heure relativement limitée, bien qu'elle offre une précision beaucoup plus fine que celle obtenue par FFT (cf. annexe précision des algorithmes de convolution/déconvolution)

Le domaine fréquentiel et ses aberrations

Dans le domaine fréquentiel, la convolution se résume à la multiplication des spectres. Cette propriété est utilisée dans la majorité des convolueurs et rend possible l'implémentation de la convolution sur bon nombre stations de travail audio. Cependant la transformation dans le domaine fréquentiel n'est pas sans artefacts. Elle nécessite de considérer les signaux à convoluer par bloc d'échantillons dont la taille va directement influer sur le résultat de convolution. Un grand nombre d'échantillon permet une meilleure résolution fréquentielle, donc un meilleur rendu spectral. Cependant à cause du paradoxe temps fréquence inhérent à la transformée de fourier, on ne peut bénéficier de la même précision dans le domaine temporel. Si l'on veut privilégier les transitoires du

signal, il faudra considérer des blocs d'échantillons d'une durée moins importante, ce qui affectera aussi le timbre. Une large fenêtre d'analyse engendre une latence importante (de 64 échantillons - 16 ms à 48 kHz jusqu'à 2048 échantillons - 43 ms à 48 kHz). Pour diminuer le temps traitement il est possible d'avoir recours à la technique du « zero padding » qui consiste à modifier la durée du signal d'entrée (en lui ajoutant des échantillons nuls) afin que le nombre total d'échantillons soit une puissance de 2. L'organisation des données prend la forme d'un arbre binaire beaucoup plus structuré, ce qui accélère l'ensemble du processus.

1.3.3 La convolution sur les tournages, état des lieux dix ans plus tard

Depuis la commercialisation du premier convolueur spécialement dédié à l'audio au début des années 2000, la convolution s'est largement répandue dans les studios. Souvent préférée pour son caractère réaliste, elle facilite l'intégration d'éléments postsynchronisés ou de bruitages par exemple. Elle permet aussi de simuler des espaces difficilement modélisables par des traitements algorithmiques classiques tel que des habitacles de véhicules, ou des acoustiques extérieures.

Le choix du stimuli

Nous venons de définir la réponse impulsionnelle comme le résultat d'une excitation par impulsion de Dirac. Cette conception est aussi vraie théoriquement que non réalisable en pratique ; aussi créer une impulsion infiniment brève et infiniment forte n'est pas sans s'opposer à quelques limites pratiques. Les systèmes linéaires que l'on cherche à modéliser ne le sont certainement pas pour tous les niveaux d'excitation... L'effet du Dirac est donc à rapporter à un facteur d'échelle dépendant du système à modéliser. Un coup de feu, un ballon qui éclate, peuvent fournir des approximations satisfaisantes du signal de Dirac. La simplicité de mise en œuvre de ces stimuli constitue le principal avantage de cette méthode. Une récente étude *« The Hand Clap as an Impulse Source for Measuring Room Acoustics »* de Prem Seetharaman1, Stephen P.

Tarzia, a même montré la pertinence du claquement de mains dans l'estimation du TR60, de la décroissance fréquentielle et de la réponse spectrale pour des fréquences supérieures à 300 Hz. L'étendue fréquentielle et l'énergie acoustique d'un claquement de mains sont plus faibles que celle résultant de l'explosion d'un ballon de baudruche. Il s'agit d'une source colorée mais identifiable, dont il est possible de compenser sa réponse.

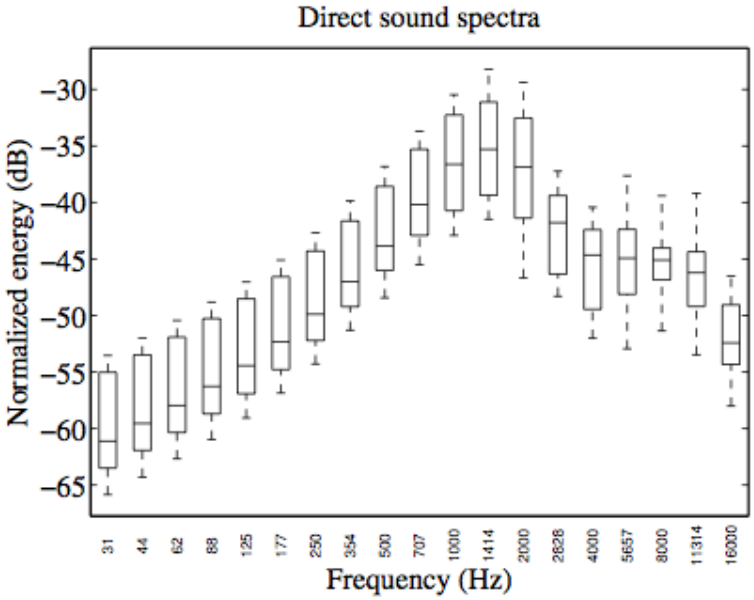


Fig 1-16. Répartition Fréquentielle d'un battement de main (son direct seul), d'après [7], p. 7.

Ces signaux très courts sont relativement sensibles aux bruits environnants qui seront alors directement intégrés dans la réponse impulsionnelle. D'autre part il est assez délicat de déterminer la longueur de la réponse impulsionnelle dans ces conditions, afin de ne garder que la décroissance du champ réverbéré, sans inclure le bruit de fond. Une méthode de détermination du temps de réverbération consiste à utiliser le point d'inflexion pour déterminer avec précision l'instant pour lequel le champ réverbéré se confond avec le bruit ambiant (cf. figure 1-17).

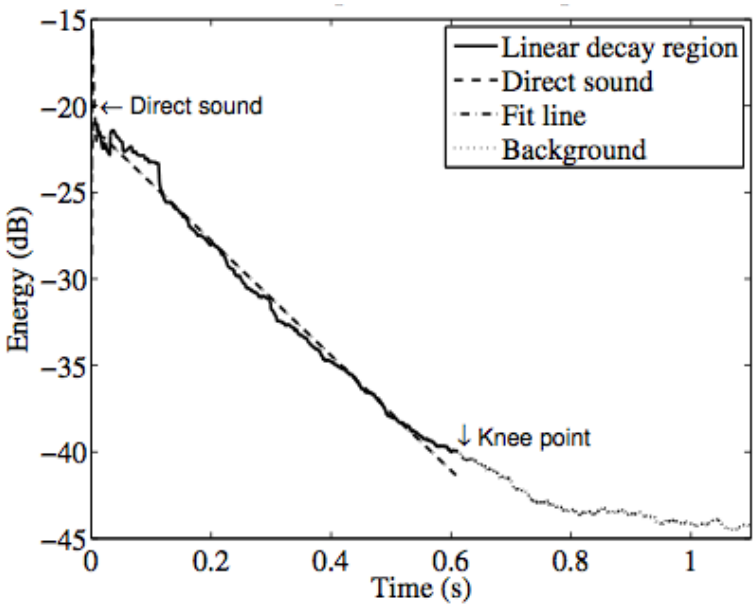


Fig 1-17: *Détermination de la longueur de la réponse impulsionnelle par recherche du point d'inflexion.*

C'est pourquoi, afin de s'affranchir de ces diverses imprécisions, il est possible d'avoir recours à des signaux d'excitation polychromatique. En effet, d'un point de vue mathématiques, la transformée de Fourrier d'une impulsion de Dirac est identique à celle d'un bruit blanc. Il est alors possible de transposer le phénomène temporel directement dans le domaine fréquentiel. Aujourd'hui la manière de capturer l'impulse semble aussi s'être standardisée. Parmi l'ensemble des signaux de références le recours au balayage monochromatique (ou sweep) semble s'être imposé.

Dans son article intitulé *"Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique",* Angelo Farina compare les résultats obtenus par diffusion d'une séquence MLS (Maximum Length Sequence) à ceux produit par sweep logarithmique (excitation monochromatique dont la fréquence évolue exponentiellement en fonction du temps) que l'on l'oppose généralement au sweep linéaire, pour qui l'évolution fréquentielle est directement proportionnelle au temps. La version logarithmique présente l'unique avantage d'un meilleur rapport signal à bruit en basses fréquences. En effet plus la mesure est longue, plus on minimise l'influence du bruit présent à la captation.

Le registre binaire générant la séquence MLS doit être suffisamment long pour ne pas engendrer d'artefacts lors de la dé-convolution. Quand le nombre d'échantillons est inférieur à la durée de la réponse impulsionnelle, on voit apparaître un phénomène de repliement temporel puisque la partie de la réponse impulsionnelle située après la deuxième période de la séquence, se retrouve en début de fenêtre temporelle. Cette particularité est liée à l'utilisation d'une dé-convolution cyclique, permettant de répéter plusieurs fois la séquence dans le but de maximiser le rapport signal à bruit. Un autre défaut de cette méthode réside dans la mise en forme de ce signal (signal créneau) qui présente de fait un facteur de crête important. La diffusion de cette séquence s'avère problématique sur bon nombre de systèmes qui ne sont pas capables de suivre cette évolution dynamique importante. Elle entraine le vieillissement accéléré des haut-parleurs ainsi que l'apparition de distorsions dans la réponse impulsionnelle.

Contrairement à la génération d'une séquence pseudo aléatoire (MLS), le sweep ne nécessite pas la synchronie parfaite du lecteur et de l'enregistreur. Les problèmes d'aliasing temporels sont simplement résolus par l'ajout de silence à la fin du stimulus. Elle paraît en ce sens plus simple à mettre en place. Elle est aussi plus robustes à de légères variations de fréquences d'échantillonnages et permet une visualisation directe des non linéarités du système (visibles sur la partie non causale de la réponse impulsionnelle).

Néanmoins, si de nombreuses études se sont penchées sur ces problématiques, la question choix du stimulus reste encore ouverte à l'heure actuelle. La manière de capter une réponse impulsionnelle doit-elle être unique ou nécessite-elle de s'adapter à la situation concrète d'utilisation? La convolution suppose la linéarité du système, or certains sonorisateurs rapportent l'existence potentielle de phénomènes de cuivrage de salle pour des niveaux d'excitation importants. Ces derniers peuvent caractériser le passage vers l'acoustique non linéaire. On peut donc supposer que la nature du stimulus semble aussi importante que son niveau de diffusion.

Afin de s'approcher au plus près du phénomène observé, nous pourrions aussi avoir recours à des stimuli réalistes tels qu'une une voix parlée dans notre cas, ou de la musique à fort niveau en sonorisation par exemple). Notre objectif serait alors de sortir d'une situation de laboratoire pour se rapprocher des caractéristiques réalistes, plus proches du signal à modéliser et de nos sensations.

Aujourd'hui dans le monde du son à l'image un outil a su se démarquer au point de devenir l'outil standard dès qu'une convolution est mise en œuvre, il s'agit d'Altiverb de la société Audio Ease. Au delà de l'accomplissement technologique, on peut attribuer son succès commercial à sa facilité d'utilisation, au renouvellement constant d'une vaste base de donnée mais aussi au développement des habitudes de travail. Altiverb a lui aussi contribué à la standardisation des modes de capture en fournissant des signaux de référence, rendant accessible la dé-convolution à tous les utilisateurs. On peut cependant constater que ces mesures prennent en compte l'ensemble de la chaîne de diffusion (conversion analogique-numérique, enceinte et lieu de restitution et microphone). La méthode proposée par Altiverb ne permet donc pas la dé-convolution relative à un microphone placé en champ proche de l'enceinte.

Si le modèle théorique est valable pour de nombreuses applications il est en revanche difficile de mesurer à des impulses associées ou nécessitant de forts niveaux d'excitation, telles que les acoustiques extérieures, qui comptent parmi les plus difficiles à capturer. Dans le but d'obtenir un rapport signal à bruit suffisant, il faudra générer de forts niveaux acoustiques et donc avoir recours à des systèmes à la bande passante moins bien contrôlée.

1.4. Son à l'image et dé-réverbération

Au delà des aspects précédemment évoqués, la réverbération permet avant tout la mise en espace par la création des plans sonores. A la manière du sfumato en peinture, elle génère une profondeur permettant d'organiser ces différents plans pour créer une sensation de relief. C'est d'ailleurs une des raisons pour lesquelles la perche a autant d'importance sur un tournage, puisqu'elle offre - entre autre - une perception beaucoup plus naturelle de l'espace. Cette manière de marquer les plans permet aussi la hiérarchisation de l'information.

Au cours de ses travaux sur de la distance subjective, Pavel Zahorik a montré que notre perception n'était pas linéaire. Nous avons tendance à surestimer les distances inférieures à 1m puis à sous-estimer les distances les plus lointaines. C'est ce qu'illustre la figure 1-18 pour des sujets soumis à un stimulus sonore seul.

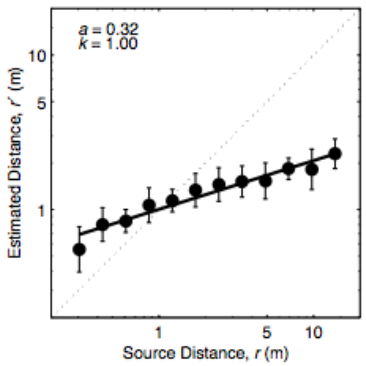


Fig. 1-18 : *Distance sonore perçue, d'après [8], p. 2.*

Cette courbe a été obtenue en soumettant les sujets à des stimulus pour lesquels l'intensité, le timbre, la différence inter-orale et le rapport direct à réverbéré ont été contrôlés de manière à recréer les conditions réelle de perception d'une source distante. La méthode utilisée est décrite dans [9].

Au cinéma notre perception est bimodale puisque l'image joue donc un rôle essentiel dans la localisation d'une source (effet ventriloque), elle peut donc accroitre la

précision de notre jugement, ou le contredire. Comme cette contradiction peut-être porteuse de sens ou au contraire nuire à l'immersion du spectateur, le travail de spatialisation est alors éminemment subjectif. Il peut nécessiter d'aller au delà d'une représentation naturelle, au profit d'un ressenti plus proche des attentes du spectateur. La dé-réverbération pourrait en ce sens contribuer à plus de réalisme, en agissant sur l'un des paramètres décrits par Pavel Zahorik. On ne peut pour l'heure marquer ces plans sonores qu'en les éloignant, par action sur le timbre et l'ajout de réverbération.

Cependant il serait illusoire de d'envisager qu'un tel outil puisse permettre de s'affranchir d'un preneur de son. Le travail du perchman en tournage consiste principalement à optimiser le placement microphonique dans le but de capter l'ensemble des dialogues en accord avec le jeu des acteurs. Il doit donc anticiper leur dynamique et réagir à leurs mouvements. Comme le montre la figure 1-3, de grandes variations du timbre existent du fait de la directivité de la voix.

S'il est parfois possible de corriger ces variations par l'utilisation ponctuelle d'un filtre, il est en revanche très difficile de corriger des défauts trop importants dûs à un mauvais placement. Le signal utile se retrouve dans ce cas très près du bruit de fond. Ce dernier va donc augmenter à la mesure du gain apportée par notre correction. Dans certains cas la réverbération fait partie de ces bruits de fond, la dé-réverbération pourrait alors accroitre la marge de correction.

Dans certains cas extrêmes, la perte d'intelligibilité due à la réverbération peut-être tolérable sur une courte période, mais s'avérer très fatigante en condition d'écoute prolongée. C'est pourquoi combattre la réverbération peut grandement améliorer le confort d'écoute ainsi que la compatibilité de l'œuvre. En effet le lieu de diffusion est un facteur important dans la transmission du message. On constate d'ailleurs que les tolérances sur le temps de réverbération des salles de cinéma sont relativement faibles. La figure 1-19 représente l'intervalle de tolérance Dolby pour les temps de réverbération des auditoriums de mixage. La valeur Tm est le temps de réverbération nominal, amené à évoluer en fonction du volume de l'auditorium. Si l'on prend l'exemple d'un studio de 1000 m^3 (20 m*10 m*5 m), elle vaut environ 540 ms, ce qui signifie que le mixage doit

s'effectuer, selon ces critères, dans un lieu plutôt mat. Le film lui pourra amené à être diffusé dans des conditions bien loin de ces situations normalisées. Si le cycle de vie d'un film en salle peut-être relativement court, sa commercialisation sur des supports vidéo peut elle être d'une durée beaucoup plus longue. L'utilisation dans le cadre domestique peut être l'une de ces conditions difficilement contrôlable qui aurait pour effet d'accentuer les problèmes précédemment décrits.

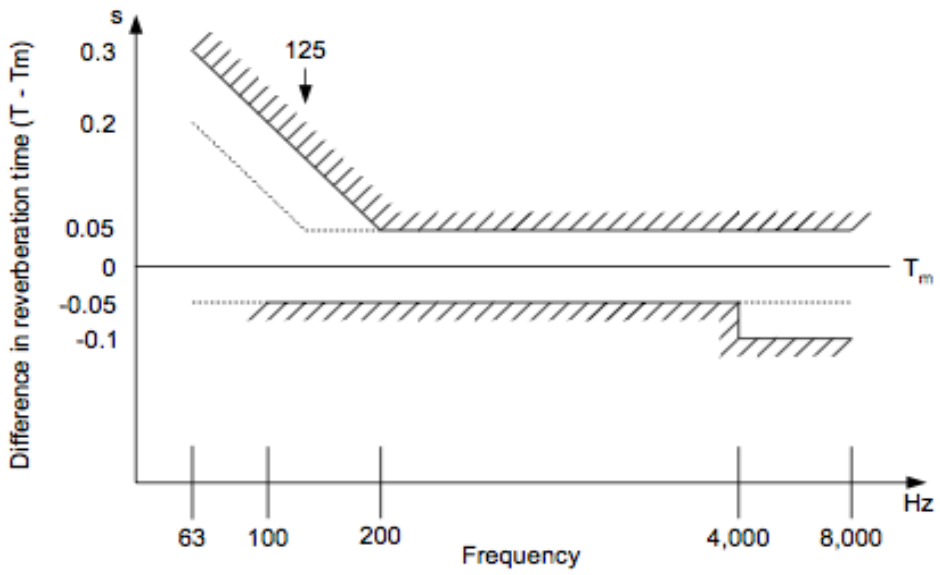


Fig 1-19 : *Intervalle de tolérance sur le temps de réverbération, d'après [10]. p. 3.*

Ainsi, en suivant les grandes étapes de la captation d'une voix depuis son émission jusqu'à sa diffusion, nous avons fait émerger l'influence du lieu de propagation. S'il participe à la caractérisation de la source en modifiant sa dynamique et ses propriétés spectrales, ce lieu peut engendrer des problèmes d'intelligibilités ou de localisation. La sensation d'éloignement augmente à mesure que l'énergie du champ direct s'atténue au profit de celle réverbérée (cf. fig. 1-18). Lors du mixage, nous pourrions donc être tenté de corriger ces deux problèmes par l'utilisation d'outils permettant de séparer le son direct de ces multiples réflexions. Le chapitre suivant sera consacré à l'étude des différentes approches envisagées ou utilisées pour supprimer la réverbération.

2. LES PROCEDES DE DE-REVERBERATION

Les procédés de dé-réverbération faisant appel à des technologies relativement récentes, l'ingénieur du son a parfois essayé des techniques détournées pour supprimer le champ réverbéré. Nous pouvons séparer ces techniques en deux catégories : les traitements dynamiques et les traitements de suppression de bruit (denoising). Après avoir étudié les potentialités de ces deux approches nous nous focaliserons sur des méthodes encore peu implantées dans les studios : prédiction linéaire, déconvolutions et estimation du champ réverbéré par apprentissage supervisé, en vue d'envisager leur utilisation sur un cas concret dans le prochain chapitre.

2.1. Les techniques pionnières

2.1.1. Les traitements dynamiques

Ces méthodes visent à atténuer voire à supprimer le sustain engendré par la réverbération. Dans des cas extrêmes, pour lesquels le TR60 et le pré-délai étaient exceptionnellement longs et le rapport direct à réverbéré suffisamment élevé ; certains mixeurs avaient recours aux ciseaux ou à la porte de bruit appliqués sur une bande défilant en sens inverse. Le retournement temporel offrait une meilleure conservation des transitoires d'attaque, alors non modifiés par le retard de déclenchement de la porte de bruit. Le seuil et le retard d'activation définissent la quantité de réverbération supprimée. Il en résulte cependant un son mutilé aux caractéristiques peu « naturelles ». Ce traitement « ancestral » affecte aussi les événements de faible niveau de la parole comme les respirations en préservant l'ensemble des premières réflexions. On tentait alors d'intégrer ce résultat au mixage avec diverses ambiances de comblage...

Aujourd'hui le travail sur l'enveloppe s'est affiné avec l'apparition des sculpteurs de transitoires (« transient designer » en anglais). Leurs effets, parfois comparés à ceux obtenus par compression-expansion, peuvent aussi offrir des résultats intéressants en matière de dé-réverbération. Leur fonctionnement est régi par un signal de contrôle d'amplification (VCA) obtenu par comparaison d'enveloppes préalablement générées. Trois enveloppes sont couramment utilisées : celle du signal d'origine, l'enveloppe du signal d'attaque et celle de maintien. L'objectif est de traiter séparément les zones d'Attack et de Sustain de la courbe ADSR. Comme il est possible de le lire dans la documentation du SPL *Transient Desginer* que la commande d'amplification pour l'attaque est issue d'une différence entre l'enveloppe réelle et l'enveloppe au temps de montée amorti (durée fixe). A l'inverse la gestion de la zone de maintien s'obtient par soustraction d'une enveloppe prolongée, comme nous l'indique la figure 2-1.

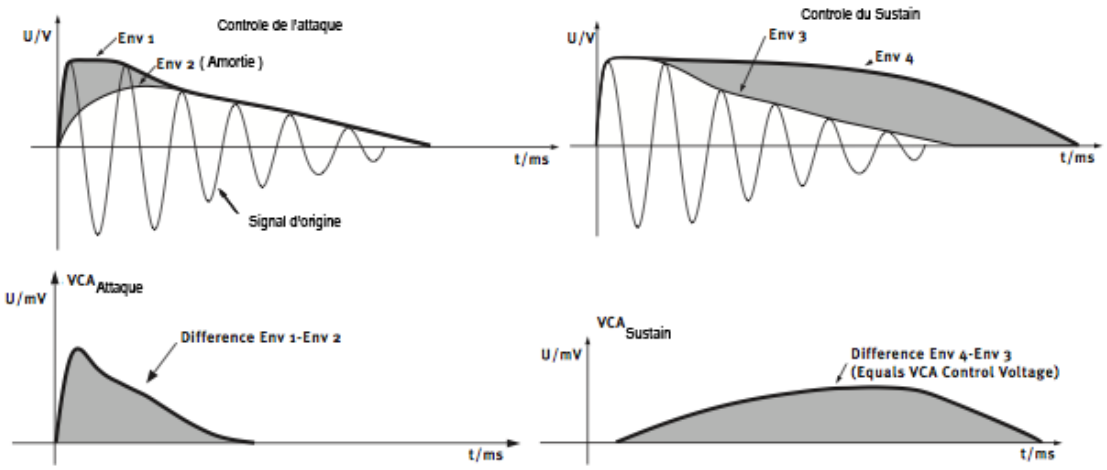


Fig. 2-1 : *Effets du Transient Designer d'après [11].*

L'action cumulée sur ces deux signaux permet l'accentuation des transitoires et la diminution de la zone de maintien. Elle provoque donc l'accroissement du facteur de crête, qui diminue à mesure que la réverbération augmente. L'utilisation de ces traitements en vue d'une dé-réverbération est parfois relativement efficace lorsqu'il s'agit de sons synthétiques ; pour lesquels l'enveloppe du son direct reste constante. Mais comme le profil dynamique de la voix est au contraire très variable, on voit

apparaître des artefacts comparables de modulations d'amplitude plus ou moins marquées.

L'automatisation de ce type de suivi d'enveloppe est une piste de recherches initiée par Berkley et Mitchell des laboratoires Bell depuis la fin des années 80, elle est appelée Filtrage d'Enveloppe Temporelle (*Temporal Envelope Filtering)*. Cela consiste en l'analyse d'un coefficient de modulation. Plus la réverbération sera longue plus la décroissance tardive va remplir les zones de faible énergie du signal. A partir de l'enveloppe du signal, ce procédé va extraire des informations sur un TR60 estimé pour modifier la courbe ADSR du son réverbéré dans le but d'accroitre le coefficient de modulation ci-dessous :

$$m(\Omega) = \frac{\int_0^\infty h^2(t)e^{-i\omega t}dt}{\int_0^\infty h^2(t)dt}$$

Ce procédé ne traite pas directement les modifications spectrales introduites par l'acoustique du lieu d'enregistrement ou ses moyens de captation. Cependant en agissant de la sorte dans le domaine temporel il est possible de faire émerger la zone d'intelligibilité de la voix (2kHz-4kHz) en amplifiant les transitoires d'attaque. L'atténuation de la zone de maintient peut aussi contribuer au démasquage fréquentiel, en contrôlant plus précisément les résonnances.

2.1.2 Le Denoising

Ces outils ont pour but de filtrer par bloc certaines régions du spectre du signal entrant. Cette fonction est mise en œuvre par différents algorithmes qui peuvent laisser la détermination du filtrage à l'utilisateur ou l'estimer par l'analyse du bruit de fond stationnaire. Or la réverbération ne peut être considérée comme stationnaire puisqu'elle est au contraire fortement corrélée avec le signal anéchoïque. Même si l'utilisation de ces outils dans le but de combattre la réverbération n'est pas valide théoriquement, des procédés de dé-bruitage tel que ceux développés par « Cedar Audio LTD » sont parfois utilisés en vue d'une dé-réverbération partielle. Les technologies évoquées font l'objet de brevet et il nous est difficile de connaître précisément les algorithmes implémentés, mais Il semble qu'ils mettent en œuvre plusieurs traitements dynamiques répartis selon

un découpage fréquentiel adapté et modulable. Nous aurons l'occasion d'évoquer ces effets dans le troisième chapitre de ce mémoire.

2.2. Etat des lieux de la recherche appliquée

Nous venons de voir que ces outils « conventionnels » sont d'une aide bien limitée lorsque l'on désire minimiser les effets de la réverbération. Au cours de ces dernières années de nouvelles méthodes développées dans le secteur des télécommunications ont fait leur apparition dans le domaine de l'audio.

Ces techniques de dé-réverbération à proprement parler se divisent en deux catégories : l'une visant à annuler le champ réverbéré, l'autre visant à le supprimer. Cette légère différence sémantique cache en fait deux approches du phénomène. L'annulation vise à déterminer les propriétés du canal de transmission pour trouver le filtre inverse de ce dernier, ce qui de manière plus imagée, revient à annuler l'effet de chacun des rayons acoustiques. La suppression a pour but de connaître les caractéristiques du signal d'origine et d'en déduire la part réverbéré du signal.

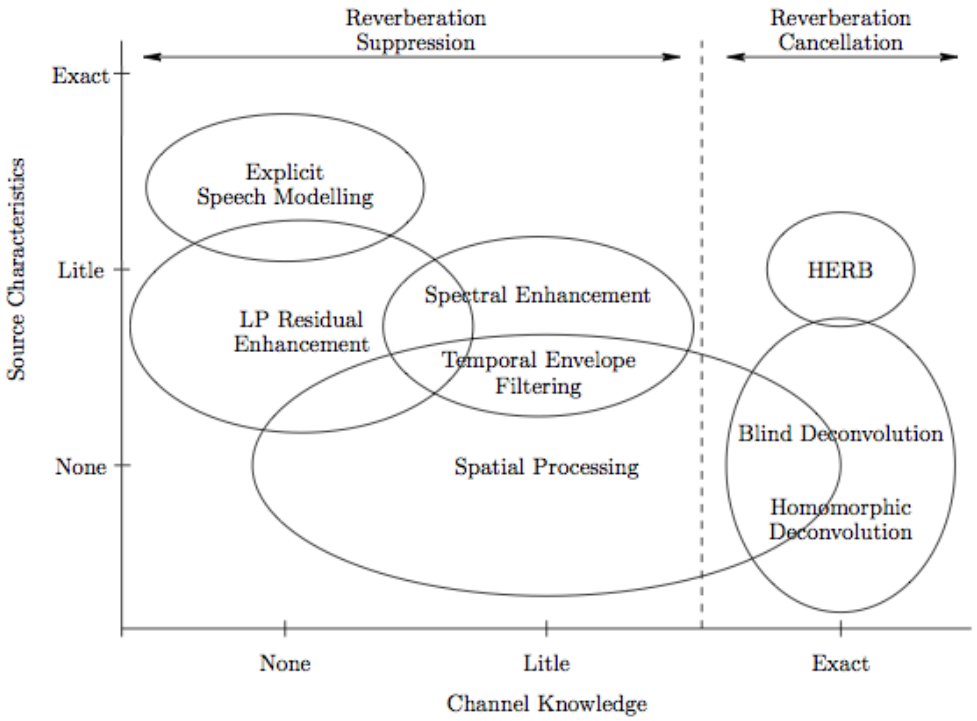


Fig. 2-2 : *Domaines d'étude des procédés de dé-réverbération, d'après [12] p.54.*

2.2.1 La prédiction linéaire

En traitement du signal, les mécanismes de la voix évoqués au cours du premier chapitre sont approchés par un modèle de filtre autorégressif, dont les coefficients dépendent des états précédents du signal d'entrée. Excité par un train d'impulsions périodiques (de même fréquence que celle de vibration des cordes vocale) ce filtre produit le signal dont les caractéristiques se rapprochent d'un son voisé. S'il est stimulé par un bruit blanc de moyenne nulle et de variance unitaire, on obtient une approximation de son non-voisé.

Le filtre autorégressif modélise les caractéristiques spectrales du conduit vocal (formants et antiformants), mais cette modélisation n'est valable que pour des portions où la voix est suffisamment stable. Pour rendre ce signal stationnaire, on l'analyse et on le segmente à intervalles réguliers. Le modèle est alors recalculé à chaque nouvelle itération. Certaines études à ce sujet *[1]* préconisent un fenêtrage temporel d'une durée de 10 ms dont les performances peuvent être améliorées par superposition des intervalles (overlapping). Cela suppose tout de même de connaître l'activité vocale et de pouvoir détecter les différents modes d'expression. Le taux de passage à 0 est alors souvent utilisé pour déterminer s'il s'agit d'un signal voisé ou non.

Lorsqu'elle est émise la voix traverse le canal acoustique du milieu de propagation, puis elle est captée par un microphone, deux entités qui sont elles aussi caractérisées par leur réponse impulsionnelle.

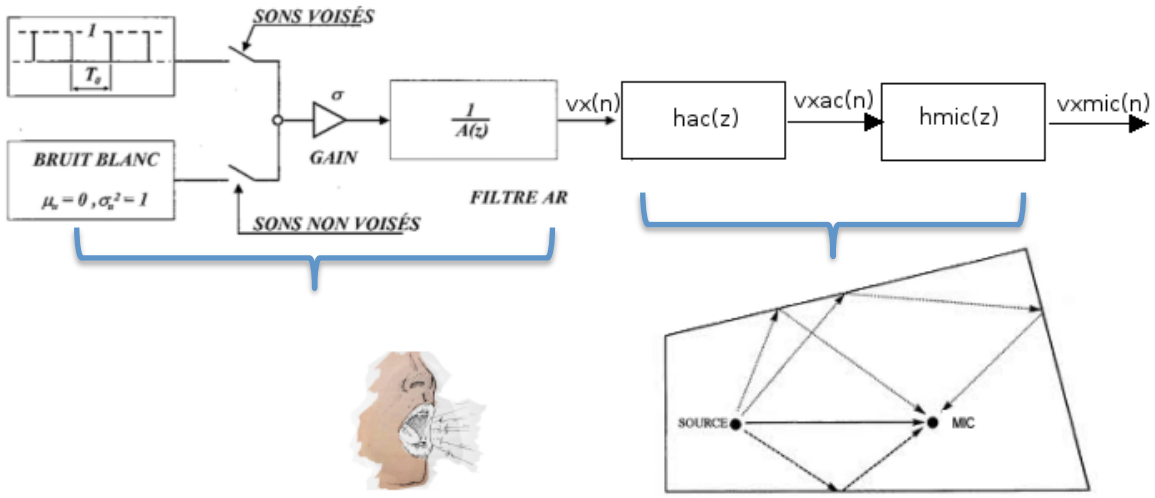


Fig 2-3. *Modélisation de la parole émise et captée en un lieu clos.*

Nous ne connaissons pour l'heure aucune de ces réponses impulsionnelles. Tout l'enjeu de cette étude réside dans la précision de l'estimation séparée de hac[z] et A[z]. La prédiction linéaire consiste en la modélisation d'un signal directement dans le domaine temporel sans passer par une transformée de Fourier. On considère par hypothèse que l'état courant du système peut être approché par une combinaison linéaire des états précédents. Pour le système décrit en figure 2-3 on aurait alors, à l'entrée du microphone :

$$vxmic[n] = \sum_{i=0}^{P} \alpha_i * vxmix[n-i] + e[n]$$

$e(n)$ représente un signal d'erreur d'ordre *n* et *P* le nombre d'échantillons qui précèdent vxmic(n) tandis que α_i est appelé prédicateur. C'est le coefficient qui va pondérer chaque contribution des précédents états. Comme ce modèle suppose le signal semblable à lui même au-cours du temps (auto-corrélé), il n'est pas possible de prédire un bruit blanc qui par définition résulte d'un phénomène aléatoire. Nous pouvons alors nous interroger sur la validité de la prédiction linéaire dans le cas d'un signal non voisé...

L'estimation des coefficients α_i offre la possibilité d'approcher le filtre équivalent aux associations du filtre AR, de la réponse impulsionnelle de la salle et celle du microphone. Ce calcul est réalisé par minimisation de la variance du signal d'erreur qui permet de faire apparaître une matrice d'autocorrélation de dimension P et de résoudre

un système d'équations linéaires. Cependant nous cherchons à supprimer uniquement la réverbération. Il faut donc éliminer les contributions des filtres AR et de hmic. C'est ici qu'émergent certaines différences dans la manière d'aborder le problème. Nous utiliserons plus tard un programme (NML_RevCon-RR) ayant été implémenté suivant le principe de la prédiction linéaire multi-pas (Multi Step Linear Prediction, MLSP), nous nous limiterons à la description de cet algorithme.

Il n'est pas prévu dans le programme évoqué de compenser l'effet du microphone utilisé lors de la captation. Ce détail a son importance puisqu'un dé-timbrage peut aussi jouer sur la perception de la distance. Ce paramètre est souvent utilisé en mixage pour éloigner une source par atténuation de l'octave de 2 à 4 kHz. S'il reste théoriquement possible d'appliquer un traitement global sur l'ensemble du signal pour compenser ces effets. Dans la pratique, si le blanchiment du micro est réalisé après dé-réverbération, il risque de générer de artefacts.

Si on se place dans le cas d'une configuration de captation idéale comprenant un microphone omnidirectionnel à la réponse plate pour toutes les fréquences, l'objectif est donc de séparer la réverbération du signal anéchoïque. Or la prédiction linéaire ne permet pas de déterminer complètement hac. Par cette seule méthode il n'est possible d'estimer que les réflexions tardives. L'algorithme Multi Step Linear Prediction (ou MLSP) définit une durée arbitraire D séparant le son direct et ses premières réflexions du champ diffus. Cela revient, d'un point de vue mathématiques, à considérer le produit de convolution en deux étapes :

$$vxac[n] = \underbrace{\sum_{i=0}^{D-1} vxac[n] * hac[n-i]}_{\text{Son direct + premières réflexions}} + \underbrace{\sum_{i=D}^{L} vxac[n] * hac[n-i]}_{\text{Champ diffus}}$$

la prédiction linéaire devient alors :

$$vxac[n] = \sum_{i=0}^{D-1} \alpha_i * vxac[n-D-i] + e[n]$$

et nous sommes ainsi capable de supprimer la réverbération tardive par inversion du filtre.

$$hac_D^{-1}[n] = vxac[n] - \sum_{i=0}^{L-1} \alpha_i * vxac[n - D - i] - e[n]$$

L'estimation du retard D est donc essentielle. La détermination de ce paramètre vient de l'observation des caractéristiques du signal vocal. Ce dernier est corrélé à court terme alors que les réflexions tardives le sont à long terme. Il faut donc définir la durée minimale pendant laquelle la parole va être suffisamment corrélée afin de préserver le son direct et ses premières réflexions. Diverses études [13] préconisent de choisir D de l'ordre de 30 ms, ce qui est par ailleurs proche du temps de fusion de l'oreille...
La longueur du filtre de prédiction définit lui aussi le temps de réverbération que l'on est capable de traiter, grossièrement approché par (L-D).Fs avec D<L.

La prédiction linéaire seule ne permet donc pas de supprimer l'intégralité du champ réverbéré, mais seulement une partie du champ diffus dans les conditions que nous venons d'énoncer. Cependant si l'on en croit l'étude de Kinoshita et Nakatami [13], l'algorithme MSLP réduit considérablement le taux d'erreur réalisé par un algorithme de reconnaissance vocale (de plus de 65% à 2%, dans une situation contrôlée, pour une distance source microphone de 2 m). Nous détaillerons les effets sensibles de cette méthode dans le cadre de notre partie pratique.

A ce stade nous pouvons cependant rester insatisfait face à ce résultat et chercher à supprimer les réflexions précoces. Du fait de sa forte corrélation avec le signal anechoïc, une des pistes consiste à se tourner vers les traitements homomorphiques.

2.2.3 La déconvolution cepstrale

Nous changeons donc de domaine d'étude et quittons les procédés de suppression pour nous tourner vers ceux d'annulation. L'objectif consiste maintenant à trouver un espace d'analyse dans lequel nous pourrons séparer l'effet du canal de transmission quel que soit le signal qu'il l'excite. Nous pourrions alors réaliser un filtrage dans ce domaine pour essayer de supprimer le champ diffus et les réflexions précoces.

Une transformée homomorphique a pour but de transposer un phénomène dans un domaine d'étude qui se veut linéaire, de manière à l'appréhender plus simplement. La transformée cepstrale est une de ces transformations pour laquelle l'image du produit de convolution est linéaire puisqu'il s'agit d'une simple addition. Elle s'obtient par l'algorithme illustré en figure 2-4, où x_0 est le signal d'origine et c(n) sa transformée cepstrale :

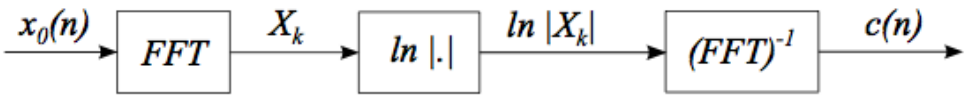


Fig 2-4 : *Transformée cepstrale.*

C'est une grandeur très utilisée dans le traitement de la voix pour des applications de filtrage, de détermination de hauteur tonale, ou de reconnaissance vocale. Elle permet dans une certaine mesure, de séparer l'influence du canal de transmission. Le cepstre de la voix donne l'image de ses variations, c'est d'ailleurs par ce biais que la dé-réverbération est envisagée. Le signal vocal varie plus lentement que ses multiples réflexions. Les représentations de la figure 2-5 Illustrent l'effet d'une réflexion unique à située 20 ms après l'émission du mot « merci »...

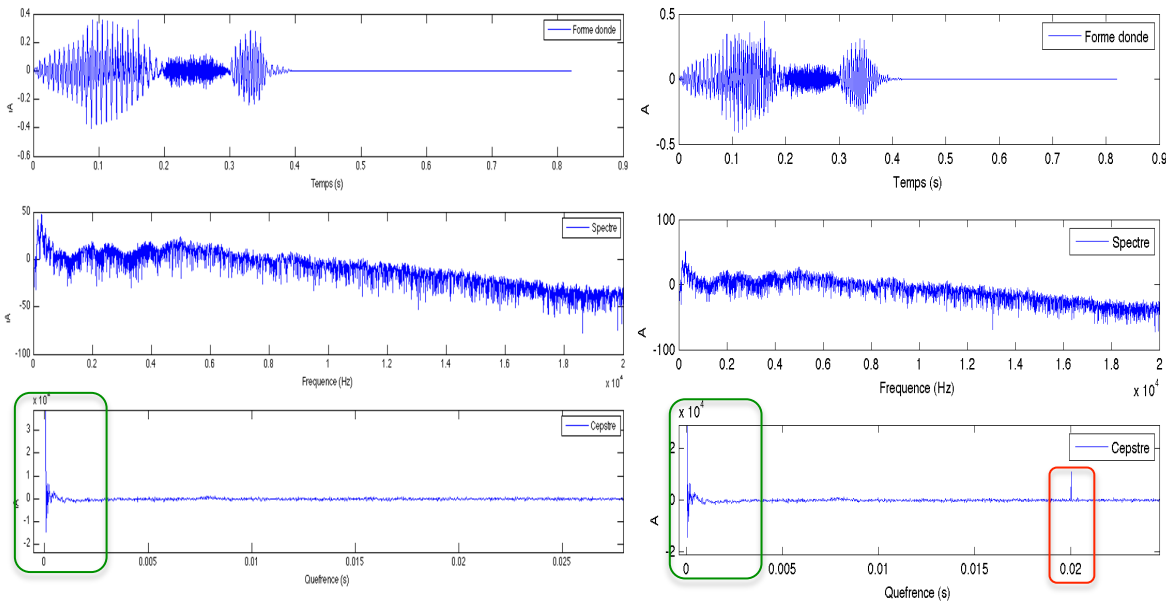


Fig 2-5 : *Effet d'un écho franc à 20 ms sur le cepstre.*

On parle alors de quéfrence et liftrage pour reprendre, de manière analogue au domaine spectral, les termes de fréquence et de filtrage. Les composantes du signal vocal sont regroupées près de l'origine quéfrentielle. Il est donc théoriquement possible de supprimer cet écho par liftrage (c'est–à-dire par soustraction dans le domaine cepstral) et de reconstruire le signal d'origine à partir de son cepstre complexe. En pratique ces méthodes sont bien plus difficiles à mettre en œuvre lorsque l'on considère le signal audio dans sa continuité. Il faut alors fragmenter le signal en vue d'une décomposition en série de fourrier, ce qui génère des artefacts (paradoxe temps-fréquence et effet de fenêtrage).

Il s'avère de plus difficile de séparer le signal utile de la réverbération. Le cas simple évoqué précédemment est bien éloigné d'une situation réverbérée qui modifie l'ensemble du cepstre à différente mesure. Du fait du fenêtrage les réflexions tardives peuvent empiéter dans la fenêtre d'analyse suivante et se mélanger à d'autres composantes. Il s'agit probablement des raisons pour lesquels je n'ai pas pu pour l'heure expérimenter de rendu suffisamment peu bruités pour envisager une utilisation dans le domaine de l'audio. Les résultats sont certes moins réverbérés mais toujours très saturés [14].

Pour contourner ces problèmes Nicolas Lopez dans « Méthodes parcimonieuses pour la dé-réverbération des signaux audio » propose de soustraire la moyenne du cepstre sur un intervalle de temps dont la durée serait définie par des évaluations subjectives. Il constate en effet qu'en présence de réverbération, la moyenne cepstrale du signal vocal devient non nulle. Cependant nous n'avons pas pu mesurer l'efficacité de cette approche faute d'accès à l'algorithme déjà implémenté.

2.2.4 A la découverte de Unveil

Au cours de l'année 2012 un programme développé par la société allemande Zynaptiq a fait une apparition remarquée dans le milieu de la post-production. Ce plug-in se nomme « Unveil ». Il est le deuxième outil à avoir été développé à destination de la post-production en vue de dé-réverbérer un signal monophonique. Unveil est protégé par un brevet, donc il est relativement difficile d'obtenir des informations relatives à son fonctionnement. Cependant pour répondre à de nombreuses questions posées par de potentiels utilisateurs Denis H. Goekdag, directeur de la société est allé un peu plus loin dans l'explication des paramètres de contrôle et donc des processus mis en œuvre. Unveil est le résultat d'une combinaison de deux algorithmes : le Mixed-signal Audio Processing (MAP) de Zynaptiq et l'Adaptive Phase Error Minimisation (APEM) développé par Prosoniq et acquis par Zynaptiq en 2012.

Si l'on ne connaît pas directement le mode de fonctionnement de cet outil, on sait en revanche qu'il n'utilise ni la prédiction linéaire, ni la dé-convolution. Le signal audio ne subit pas de FFT. Unveil est conçu sur la base d'un réseau neuronal visant à estimer la part significative du signal entrant (son anéchoïque) en vue de supprimer le bruit corrélé à ce dernier (son réverbéré).

La minimisation de l'erreur de phase présente dans l'APEM pourrait être comparée à l'algorithme évoqué par John Usher dans [15]. Cette technique consiste à mesurer en temps réel la réponse impulsionnelle du système de diffusion en vue de développer un filtre inverse pour annuler l'effet de la chaine de diffusion. La figure 2-6 illustre les principaux éléments mis en œuvre dans ce système bouclé.

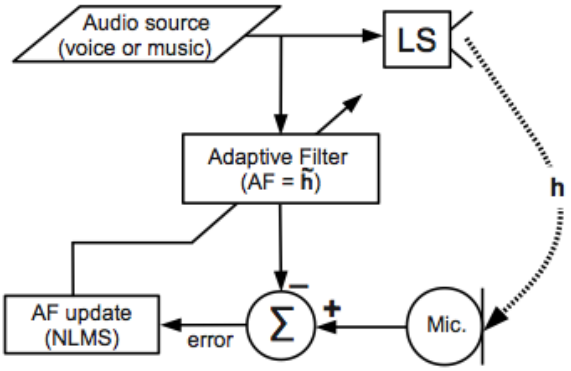


Figure 2-6 : *Estimation de la réponse Impulsionnelle en temps réel. Algorithme de J. Usher, d'après [15].*

Les caractéristiques du canal de transmission sont approchées par le filtre adaptatif. Comparé au signal provenant du microphone, on génère un signal d'erreur qui va entrainer la modification du filtre. Le but est de minimiser cette erreur pour rendre $\tilde{h}$ le plus proche de *h*. C'est le rôle de la fonction Normalized Least Mean Square (NLMS). Des signaux vocaux, ou musicaux peuvent être utilisés pour estimer la réponse impulsionnelle mais le filtre obtenu ne sera valide que sur la plage fréquentielle contenant suffisamment d'énergie pour être estimé avec précision. Le schéma présenté à la Fig 2-6 est en fait très proche de celui d'un suppresseur d'écho acoustique tel que ceux développés pour la téléphonie [16]. C'est donc par ce principe que l'on peut diminuer la diaphonie générée par la proximité du haut-parleur et du microphone d'un combiné. Cette diaphonie entraine la réinjection de la voix dans le haut parleur du locuteur placé en bout de ligne. Sans ce filtre adaptatif nous entendrions...drions notre voix retardée...tardée à mesure que l'on parle... parle !

Une fois ce filtre calculé on procède à la détermination de son inverse. Nous traiterons de ces problématiques dans un chapitre spécifique (c.f. Obtention du filtre inverse). L'approche schématique de J. Usher est détaillée à la figure 2-7 et on peut constater qu'avant de traverser le canal de diffusion, le signal d'origine est filtré par l'inverse approximé de la réponse impulsionnelle.

Nous pourrions supposer que Unveil adopte un fonctionnement similaire car cette technique n'utilise ni FFT, ni prédiction linéaire...

Dans le principe adopté J. Usher, le signal de mesure est obtenu par un microphone situé en un point de la zone d'audience. Si l'on en croit les déclarations de Denis H. Goekdag, le réseau neuronal d'Unveil serait susceptible de jouer ce rôle et à partir d'un modèle psycho-acoustique, il pourrait être capable d'estimer la part du signal considérée comme indésirable. La figure 2-8 illustre cette hypothèse.

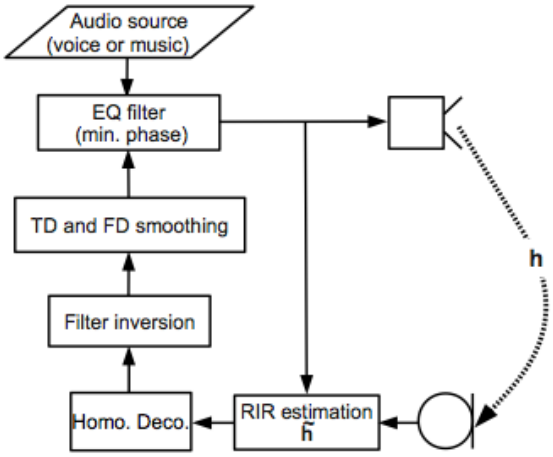


Fig. 2-7*: Inversion du filtre Adaptatif Algorithme de J. Usher, d'après [15].*

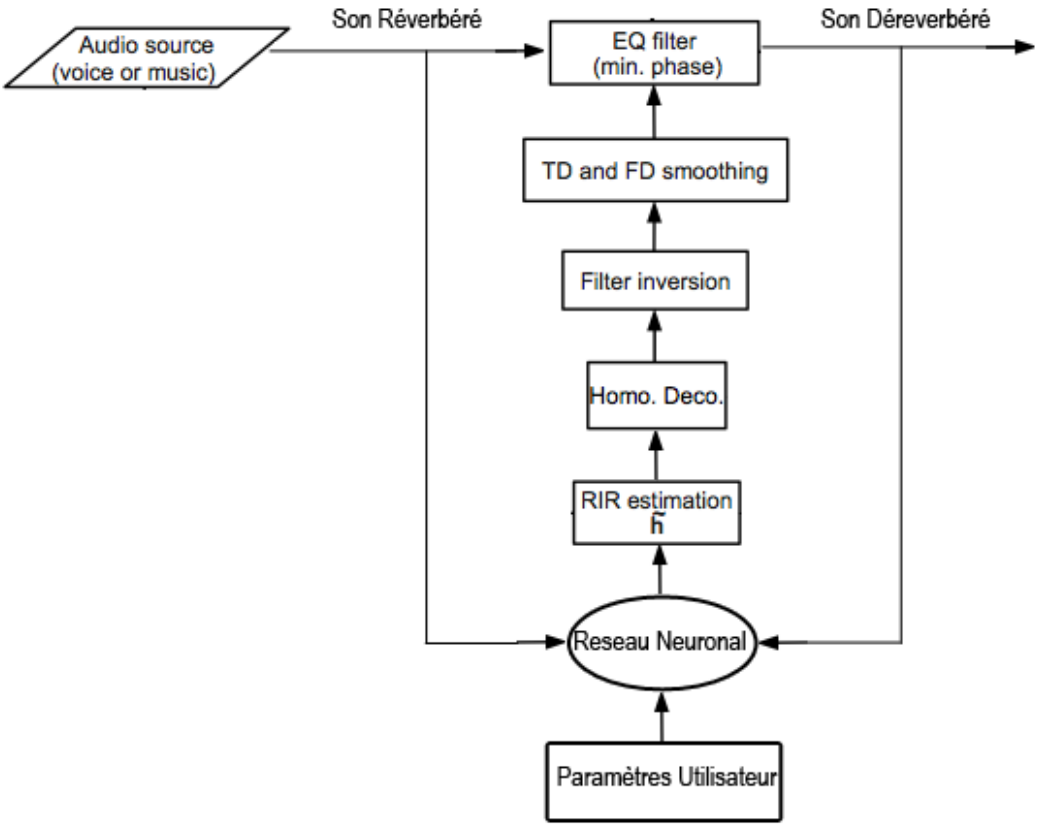


Fig. 2-8: *Hypothèse sur le fonctionnement de Unveil.*

Le réseau neuronal artificiel serait donc l'élément central de ce procédé. Il s'agit d'un type d'algorithme inspiré du fonctionnement des neurones physiologiques, capable de modifier sa structure en vue de traiter un phénomène. On dit alors qu'il met en œuvre des processus d'apprentissage par l'expérience similaires à ceux réalisés par le cerveau humain. Un réseau neuronal simple peut être représenté par la figure 2-9.

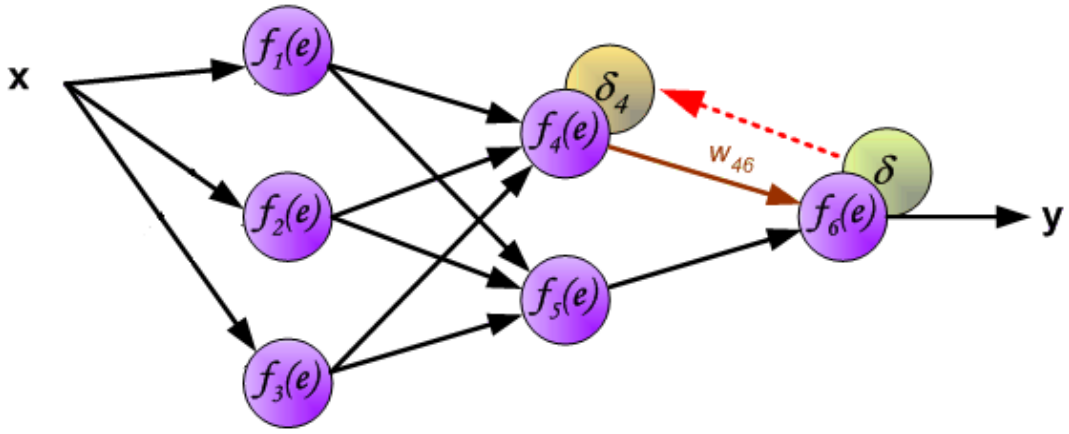


Fig 2-9 *: Réseau Neuronal avec rétro-propagation du signal d'erreur.*

Le signal d'entrée est traité par différentes fonctions toutes liées entre elles pour former des couches différentes et chacune de ces connections est pondérée. Ainsi le signal de sortie peut prendre différentes valeurs correspondant à chaque fois à une combinaison linéaire des différentes fonctions. Dans le cas d'un système à apprentissage non supervisé, la sortie obtenue pour l'échantillon x est comparée à une valeur théorique de référence. On élabore alors un signal d'erreur qui va servir à modifier les pondérations apportées à chacune de ces connections. Ainsi un tel système est relativement long puisqu'il doit élaborer et propager cette erreur à chaque modification du réseau, pour chaque valeur du signal d'entrée. Lorsque l'erreur est minimisée, le réseau est optimisé pour traiter le phénomène souhaité.

Il semblerait que Unveil soit un système mixte. Si ses concepteurs le décrivent comme pré-entraîné à discriminer le signal utile de la réverbération, plusieurs paramètres peuvent être modifiés par l'utilisateur. Il peut donc intervenir sur certains points du réseau neuronal. Par exemple « t [REFRACT] » détermine directement la longueur de l'analyse, puisqu'il agit sur le temps de réaction du réseau de neurones pour effectuer la séparation direct à réverbéré. Une courte durée d'analyse favorise le traitement des réflexions précoces mais est susceptible de générer des artefacts. Le paramètre « t [Adaptation] » renseigne le réseau sur la durée de la réverbération. « t/f [Localize] » organise les priorités de détection de manière analogue à une fenêtre d'analyse dans le cas d'un traitement par FFT.

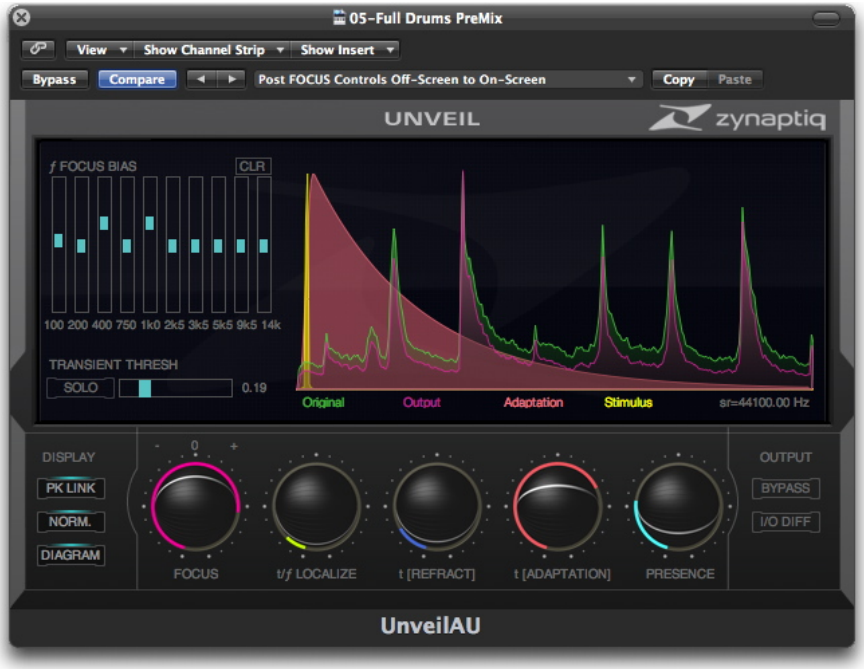


Fig 2-10 : *Interface du plug-in Unveil*

2.3. Vers une approche directe ?

Les traitements précités visent à tous supprimer ou à annuler la réverbération du signal sans connaître la réponse impulsionnelle du circuit acoustique tandis que la détermination de cette dernière constitue la clé de voute de toutes les problématiques liées à l'annulation de la réverbération. Au cours des dernières décennies le développement des techniques de convolutions audio a démocratisé la capture de réponse impulsionnelle. Si ces moyens restent une approximation du phénomène acoustique à bien des niveaux, ils ont néanmoins permis d'apporter un degré de réalisme satisfaisant pour bon nombre d'applications ce qui nous conduit à nous poser la question de la possibilité d'utiliser la réponse impulsionnelle du lieu de tournage en vue dé-réverbérer le signal enregistré. Dans l'hypothèse où nous parvenions à recréer précisément l'ensemble du phénomène capté, serait-il possible de lever une partie voire la totalité du problème d'indétermination du canal de transmission ? Si telle était le cas,

nous pourrions alors envisager la dé-convolution par cette réponse impulsionnelle. En quoi consiste la dé-convolution et quelles seraient ces limites théoriques ?

2.3.1 Problèmes inverses et moyens de résolution

La déconvolution est l'opérateur inverse donnant accès aux causes du système linéaire et invariant dans le temps, modélisé par l'équation de convolution donnée ci-dessous :

$$s(u) = \int_{-\infty}^{+\infty} h(t-u)e(u)du$$

La convolution peut donc être vue comme la moyenne pondérée de chaque instant du signal d'entrée par la réponse impulsionnelle. Cet opérateur a un rôle lissant dans la mesure où le résultat de la convolution hérite des propriétés de continuité de *h* ou de *e*. Les irrégularités du signal d'entrée sont donc gommées et tout l'enjeu de la dé-convolution est de parvenir à retrouver les informations perdues lors de cette opération.

En pratique, il n'est pas toujours possible de trouver une solution à ce problème inverse. Cette quête fait partie de problèmes décrits comme mathématiquement « *mal posés* » car il ne répondent pas aux critères énoncés par le mathématicien français J. S. Hadamard, à savoir :
« Un problème est bien posé si : la solution existe, est unique et dépend continument des données. »

Cependant dans le chapitre consacré à la convolution nous avons vu que l'obtention d'une réponse impulsionnelle était couramment réalisée à partir de la diffusion d'un sweep, d'une séquence MLS ou d'un bruit rose.... Comment faire pour ne conserver que la réponse à cette excitation ? Il faut recourir à la dé-convolution. La solution de cette opération passe nécessairement par la numérisation du signal. Comme la convolution, l'étape réciproque peut être vue sous des aspects temporels ou fréquentiels. Nous détaillerons par la suite les principes théoriques mis en jeux par ces différentes approches.

2.3.2 La déconvolution fréquentielle

Il s'agit de la méthode la plus couramment mise en œuvre pour l'obtention des réponses impulsionnelles. Pour la réaliser, le stimulus et la réponse enregistrée subissent une transformée de fourrier rapide (FFT), une division dans le domaine fréquentiel puis FFT inverse de ce résultat. Ces étapes impliquent donc un fenêtrage qui, comme pour la convolution, nécessite d'être adapté à la grandeur à traiter. Dans le cas de signaux stationnaires, ce dernier peut-être long pour accroitre la précision fréquentielle.
Cette méthode présente l'avantage d'une implémentation très efficace et peu couteuses en ressources de calcul, surtout si l'on utilise le « zero padding » pour optimiser la structure des données.

Cependant la division qui est mise en œuvre peut poser d'importants problèmes pratiques. Pour certains signaux, il est possible qu'une bande fréquentielle de faible niveau dans le signal d'origine fasse diverger ce quotient. On obtient alors une réponse impulsionnelle fortement bruitée, voire saturée. On peut néanmoins limiter ces problèmes en définissant un seuil de traitement afin de borner les valeurs du quotient spectral. On n'obtiendra pas dans ce cas le signal anéchoïque mais un rendu moins bruité, qualitativement plus acceptable. D'autre part, les signaux de référence, contenant l'ensemble du spectre à dé-convoluer ne posent pas ce genre de problèmes (bruit rose, MLS, sweep).

Nous pouvons illustrer ces problèmes de blanchiment par un simple filtrage coupe-haut à 1 kHz sur une voix. En envoyant un échantillon à la pleine échelle de quantification dans le filtre utilisé, nous obtenons sa réponse impulsionnelle. Afin de simuler les conditions de l'enregistrement nous ajoutons à la voix filtrée un bruit blanc dont la valeur RMS est située 30 dB en dessous de celle de la voix. Après déconvolution FFT on obtient alors un signal fortement saturé comme le montre la figure 2-11.

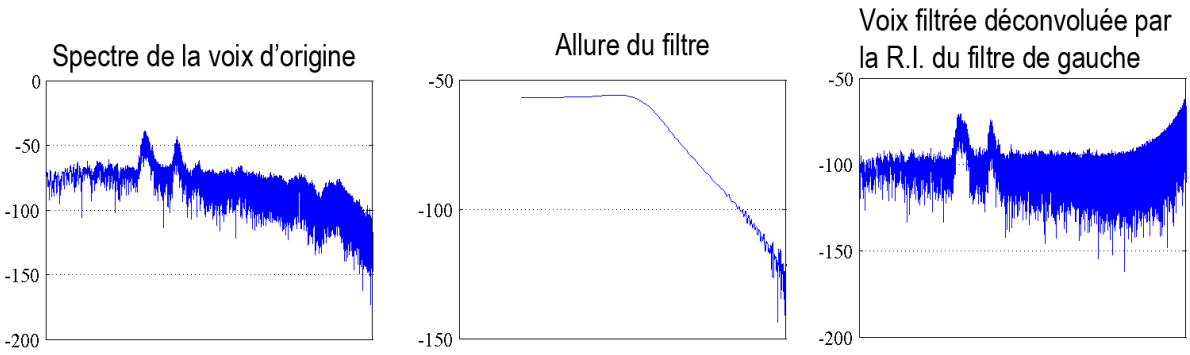


fig 2-11 : *Effet d'une dé-convolution FFT dans une opération de blanchiment.*

2.3.3 La déconvolution temporelle

L'approche progressive

En appliquant directement la formule de convolution précédemment citée, il est possible de déterminer *h*, la réponse impulsionnelle lorsque *e* (le stimulus) et *s* (la réponse de la salle) sont tous les deux connus. Par commutativité de la convolution il est aussi possible de trouver *e* (le son anéchoïque) si *h* et *s* sont donnés. La réponse impulsionnelle peut donc s'obtenir de la manière suivante :

$$s[n] = \sum_{m=0}^{n} h[m]\, e[n-m]$$

$$s[n] = h[n]e[0] + \sum_{m=0}^{n-1} h[m]\, e[n-m]$$

$$h[n] = \frac{1}{e[0]}\left(s[n] - \sum_{m=0}^{n-1} h[m]\, e[n-m]\right)$$

En regardant ce qui se passe pour chaque échantillon de *h* on constate l'importance de la valeur du premier échantillon de *e*. Par cette équation récursive, on fait apparaître des puissances croissantes du premier échantillon. C'est la raison pour laquelle, afin d'éviter que la suite ne diverge, on devra imposer la valeur -1 au premier échantillon de *e* si l'on cherche à déterminer *h*.

Dans notre cas, notre objectif est de supprimer la réverbération de *s* grâce aux informations contenues dans *h*. C'est pourquoi nous devons imposer le au premier échantillon de *h* la valeur -1 (h[0]=-1). Cette transformation engendre alors une aberration puisque l'on doit modifier la réponse impulsionnelle de la salle que l'on cherche à modéliser. Cependant il est possible de contourner ce problème par la suppression du pré-délai et la normalisation de la réponse impulsionnelle à 0 dBfs, l'éventuel changement de signe pourra alors être compensée par une inversion de phase dans le signal déconvolué.

Nous venons de présenter la méthode de dé-convolution progressive mais il est aussi possible d'envisager une dé-convolution rétrograde, partant de la fin de l'enregistrement, dont le but serait de déterminer *e* ou *h*. Il serait alors difficile de connaître de manière précise la fin du signal utile et le début du bruit de fond...

L'approche polynomiale

Une autre manière d'envisager la dé-convolution temporelle consiste à raisonner en terme de filtrage. Nous pouvons en effet considérer l'influence du lieu d'enregistrement comme un filtre numérique à réponse impulsionnelle finie, dont les coefficients seraient accessibles par la transformée en Z (c.f. Principes Elémentaires). La division polynomiale de *s* par les coefficients de *h* nous donne alors accès au signal d'origine. En appliquant la transformée inverse, on recrée ainsi le signal anéchoïque dans le domaine temporel.

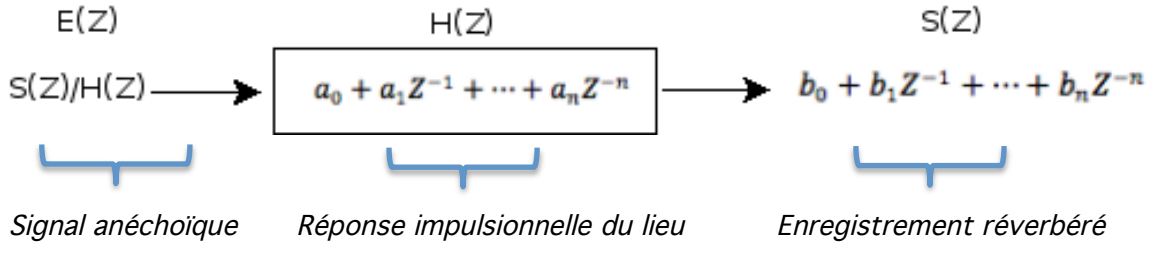


Fig 2-12 : *Principe de l'approche polynomiale.*

La figure 2-18 à la fin du chapitre permet de comparer les effets des différentes méthodes de dé-convolutions en situation contrôlée. Le signal réverbéré a été obtenu après convolution d'une voix de femme dans le domaine temporel par une réponse impulsionnelle monophonique. Le résultat de dé-convolution par cette même réponse impulsionnelle permet de comparer le signal dé-réverbéré au son d'origine par soustraction terme à terme. On peut voir que le résultat obtenu par dé-convolution polynomial est strictement identique à celui d'origine. C'est une des raisons pour lesquelles la dé-convolution progressive n'a pas été implémenté. La déconvolution polynomiale fait donc appel à un algorithme de déconvolution différent (recours à la transformée en Z) mais génère des coefficients identiques à ceux obtenus par deconvolution progressive. Ces deux méthodes sont donc fondamentalement identiques.

Cette dé-convolution polynomiale reste malgré tout très couteuse en ressources. Son utilisation s'avère très délicate lorsqu'il s'agit de traiter des fichiers de quelques dizaines de secondes, cela peut prendre plusieurs heures avec un processeur standard...

La convolution par stimulus inverse

En restant dans le domaine temporel, il est possible d'envisager la dé-convolution comme un filtrage inverse. Nous aurons plus tard l'occasion de développer les limitations théoriques et pratiques liées aux problèmes inverses. Le principe est simple puisqu'il s'agit trouver un filtre qui permettrait d'annuler l'effet du stimulus pour accéder à la réponse impulsionnelle.

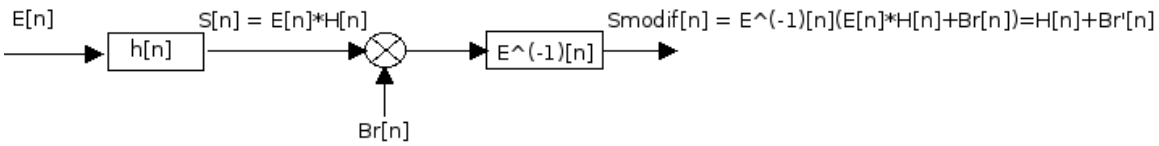


Fig 2-13 : *Approche dé-convolutive par détermination du filtre inverse.*

Cette opération n'est pas toujours réalisable car elle dépend du stimulus que l'on cherche à inverser. Toutefois s'il s'agit d'un sinus glissant (sweep logarithmique ou linéaire), cet inverse correspond au même sweep d'amplitude adapté renversé temporellement. La dé-convolution devient alors une convolution directe avec ce signal au facteur de normalisation près. Si l'évolution fréquentielle est suffisamment lente, il est possible de ne conserver que la fréquence sollicitée par le stimulus.

Par ce moyen, on accroit de manière considérable le rapport signal à bruit et l'on réduit l'influence des éléments non linéaires de la chaîne de mesure. L'enceinte utilisée pour la diffusion est souvent considérée comme la cause principale de ces distorsions car stabilité en fréquence et grande efficacité est un paradoxe difficile à résoudre. Ainsi le taux de distorsion croît à mesure que l'on augmente le niveau de diffusion. La méthode évoquée permet d'écarter l'influence de ces harmoniques de la réponse utile, mais aussi de les quantifier. L'amplitude des oscillations de la partie non causale de la réponse impulsionnelle donne alors la mesure des non linéarités du système.

2.3.4 Obtention du filtre inverse

Si peu d'auditoriums sont équipés de dé-convolueur, beaucoup d'entre eux sont en revanche pourvus de convolueurs (Altiverb, Waves-IR1...). Dans ces conditions serait-il possible d'utiliser ces outils afin d'annuler la réverbération ? Il faudrait pour ce faire déterminer précisément la réponse impulsionnelle inverse de la chaine de captation.

Or ce filtre inverse n'existe pas toujours, il est même très rare dans le cas de la réverbération. Ces réponses impulsionnelles sont dites à phase non linéaire, ce signifie que l'évolution de la phase n'est pas directement proportionnelle à la fréquence. En d'autres termes, l'ordre de départ et d'arrivée des différentes composantes n'est pas le même, on a alors affaire à une distorsion de phase relative (cf. figure 2-14).

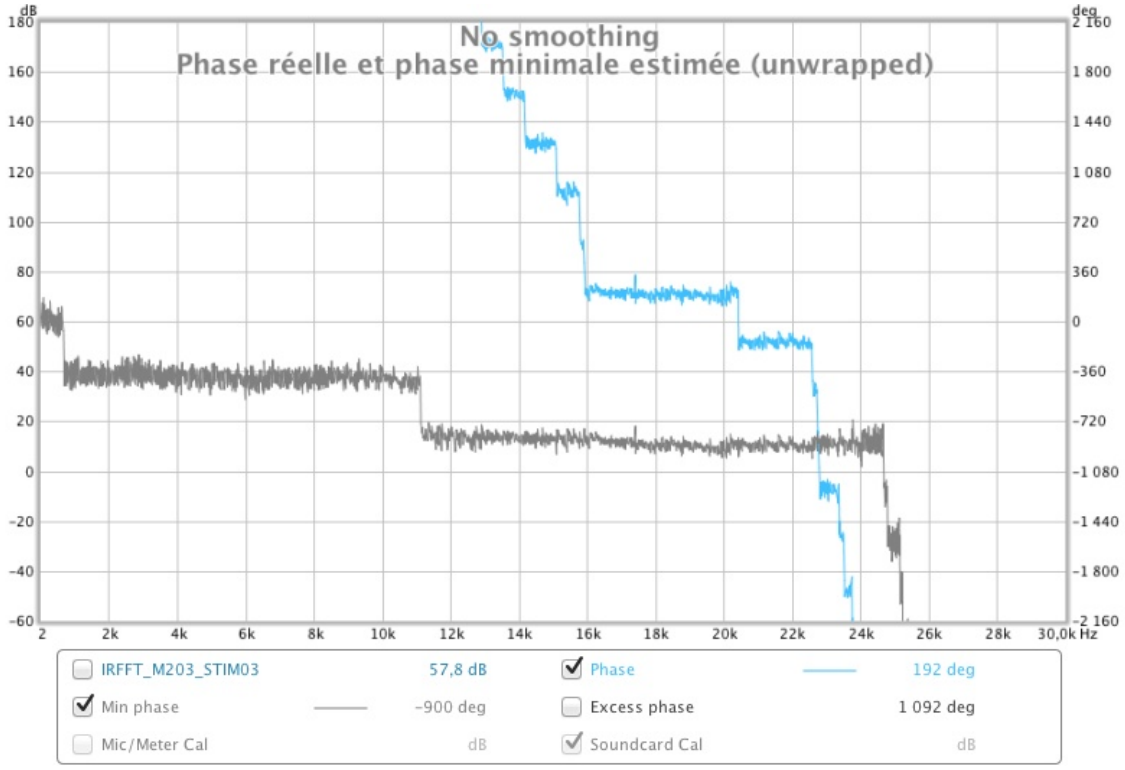


Fig 2-14: *Phase d'une réponse impulsionnelle acoustique « naturelle » non minimale (en bleu) et minimale par partie (en noir).*

La **figure 2-14** illustre ce phénomène et une estimation du filtre à phase minimale approchée par portions. Pour un confort de lecture la phase à été déroulée pour tenter de limiter les sauts de phase présents dans une représentation à 360°.

Selon S. Nelly et J. Allen dans [17], si la réponse impulsionnelle était à phase minimale, il serait possible déterminer complètement le filtre inverse et ainsi de dé-réverbérer parfaitement notre signal. Or dans le cas présent le filtre inverse n'est ni stable, ni causal. Cela signifie que cet inverse engendrera nécessairement des saturations, un retard de traitement et un effet de précédence. Les mêmes auteurs précisent cependant qu'il est possible de scinder cette réponse réelle en un filtre à phase minimale et un filtre déphaseur. Si ce passe-tout est un filtre à retard de groupe constant (déphasage identique à toutes les fréquences), il est possible de ne considérer que la partie à phase minimale. Dans ce cas on peut déterminer complètement la réponse impulsionnelle inverse et compenser ensuite le retard modélisé par le déphaseur.

Dans les autres cas nous pourrons chercher à ne conserver que la partie à phase minimale de la réponse réelle en vue d'obtenir un inverse stable mais approximé. Nous étudierons l'efficacité de cette méthode dans la phase expérimentale de cette étude.

Pour limiter les problèmes de stabilité, il est aussi possible d'approcher ce filtre inverse par une réponse impulsionnelle finie (FIR). Il faudra alors déterminer un nombre de coefficients très important pour tendre vers ce phénomène. Comme nous l'avons esquissé dans les *principes élémentaires* de la *caractérisation des salles*, le nombre de coefficients est directement lié au nombre de réflexions à modéliser. Ainsi plus la réponse impulsionnelle est longue, plus cette approximation sera couteuses en calcul et éloignée de la réalité. Il est donc délicat de traiter des réponses impulsionnelles complexes de cette manière.

En restant dans le domaine temporel il est possible d'envisager un algorithme d'inversion de la réponse impulsionnelle.

Les graphiques de la figure 2-15 illustrent la réponse impulsionnelle de la cathédrale de Chartres et son inverse, déterminé par inversion de la matrice de Toeplitz. Sur ces courbes, la représentation temporelle de droite illustre le caractère non-causal de l'inverse. La succession d'échantillons centrés à 0 dBfs fait apparaître une saturation relative au caractère instable du filtre. Même s'ils ne sont pas directement opposés, les diagrammes de gains sont d'allure relativement complémentaires.

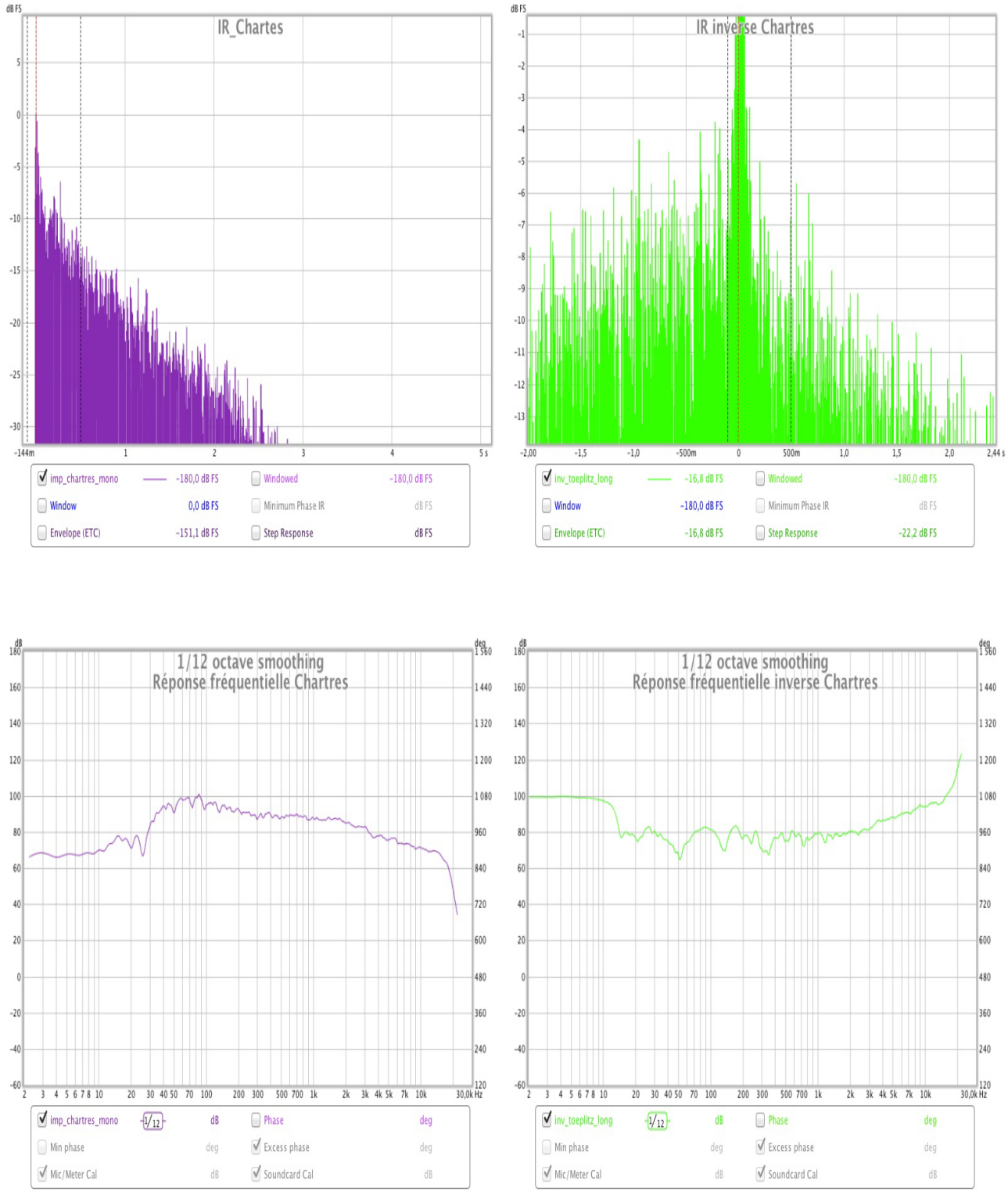


Fig 2-15 : *Allure de la réponse impulsionnelle de Chartres et de son inverse.*

En appliquant la formule de convolution discrète, il est aussi possible de formaliser le problème sous forme d'un produit matriciel, décrit par la figure 2-16.

$$\begin{pmatrix} y(0) \\ y(1) \\ y(2) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ y(L-1) \end{pmatrix} = \begin{pmatrix} h(0) & 0 & \cdots & \cdots & 0 \\ \vdots & h(0) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ \vdots & \vdots & & & h(0) \\ h(N-1) & \vdots & & & \vdots \\ 0 & h(N-1) & & & \vdots \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & h(N-1) \end{pmatrix} \begin{pmatrix} x(0) \\ x(1) \\ \vdots \\ \vdots \\ \vdots \\ x(M-1) \end{pmatrix}$$

Fig 2-16 : *Convolution et Matrice de Toeplitz.*

Les vecteurs *x* et *y* représentent l'entrée et le résultat de convolution alors que la matrice constituée des vecteurs *h* fait intervenir la réponse impulsionnelle. Le problème de dé-convolution semble ainsi se résoudre à l'inversion de cette matrice, quand c'est possible. Elle est de plus d'une forme identifiée, celle de Toeplitz, ce qui facilite la résolution du problème.

Or parvenir à l'obtention de l'inverse n'est pas parvenir à une dé-convolution acceptable. En effet pour des raisons mathématiques dites *de conditionnement*, l'inverse obtenu peut ne pas fournir les résultats attendus. Le conditionnement caractérise la complexité de l'inversion de cette matrice. Plus la fréquence d'échantillonnage sera importante, plus il devient difficile de trouver l'inverse car la taille de la matrice constituée des échantillons de *h* va augmenter.

Ainsi la dé-convolution de signaux quelconques rencontre des limitations théoriques importantes. Les difficultés énoncées font apparaître les problèmes de réversibilité de la réponse impulsionnelle. Ce n'est pas parce que l'on peut convoluer par

une réponse impulsionnelle que l'on peut nécessairement déconvoluer avec n'importe quelle approche.

La figure 2-18 permet de comparer l'effet d'une convolution et d'une dé-convolution dans les domaines temporel et fréquentiel. On voit clairement que le travail dans le domaine spectral génère plus d'artéfacts. Cependant il faut relativiser ces résultats car le bruit généré par ces traitements se trouve dans ce cas très en dessous du seuil perceptible et le gain en rapidité occasionné par le traitement FFT est considérable. Il est vraisemblable que ce signal d'erreur soit aussi dû pour partie, à des approximations engendrées par la résolution de calcul.

Ces courbes ont été obtenues en situation contrôlée. Les convolutions ont été réalisée à partir d'une voix anéchoïque et d'une réponse impulsionnelle enregistrée au studio *Lola sous la Lune,* rue Clavelle à Paris aussi la réponse impulsionnelle n'est pas à phase minimale. Ses caractéristiques sont données par la figure 2-17. Elle a été obtenue par génération d'un sweep avec dé-convolution FFT. Le choix de cette méthode a été fait de manière à limiter le temps de calcul. Le TR60, relativement court permet aussi une approximation plus fiable pour la génération des inverses.

La figure 2-19 Illustre les problématiques rencontrées pour la détermination des réponses impulsionnelles inverses. Deux méthodes sont comparées : l'inversion de la matrice de Toeplitz et l'approche par filtre à réponse impulsionnelle finie.

Les scripts Matlab utilisés pour ces calculs sont disponibles en annexe. Ces courbes illustrent le caractère non causal et instable de l'inverse. Pour approcher la non causalité, la réponse impulsionnelle obtenue par FIR (d'une durée D) est translatée (d'un facteur D) puis symétrisée par rapport à cet instant. La durée D est un paramètre laissé au choix de l'utilisateur, ce qui explique les décalages temporels différents dans les signaux dé-convolués.

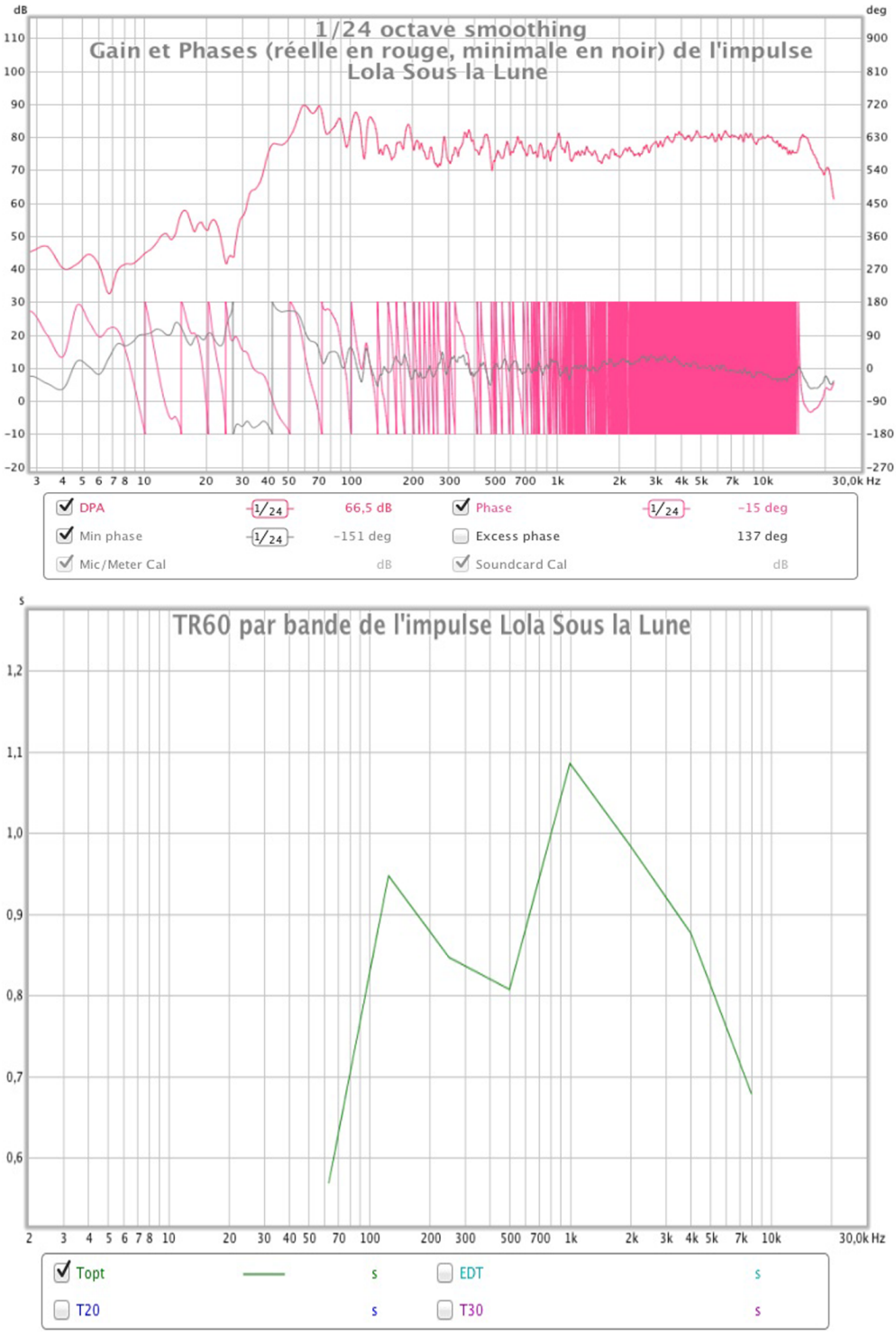
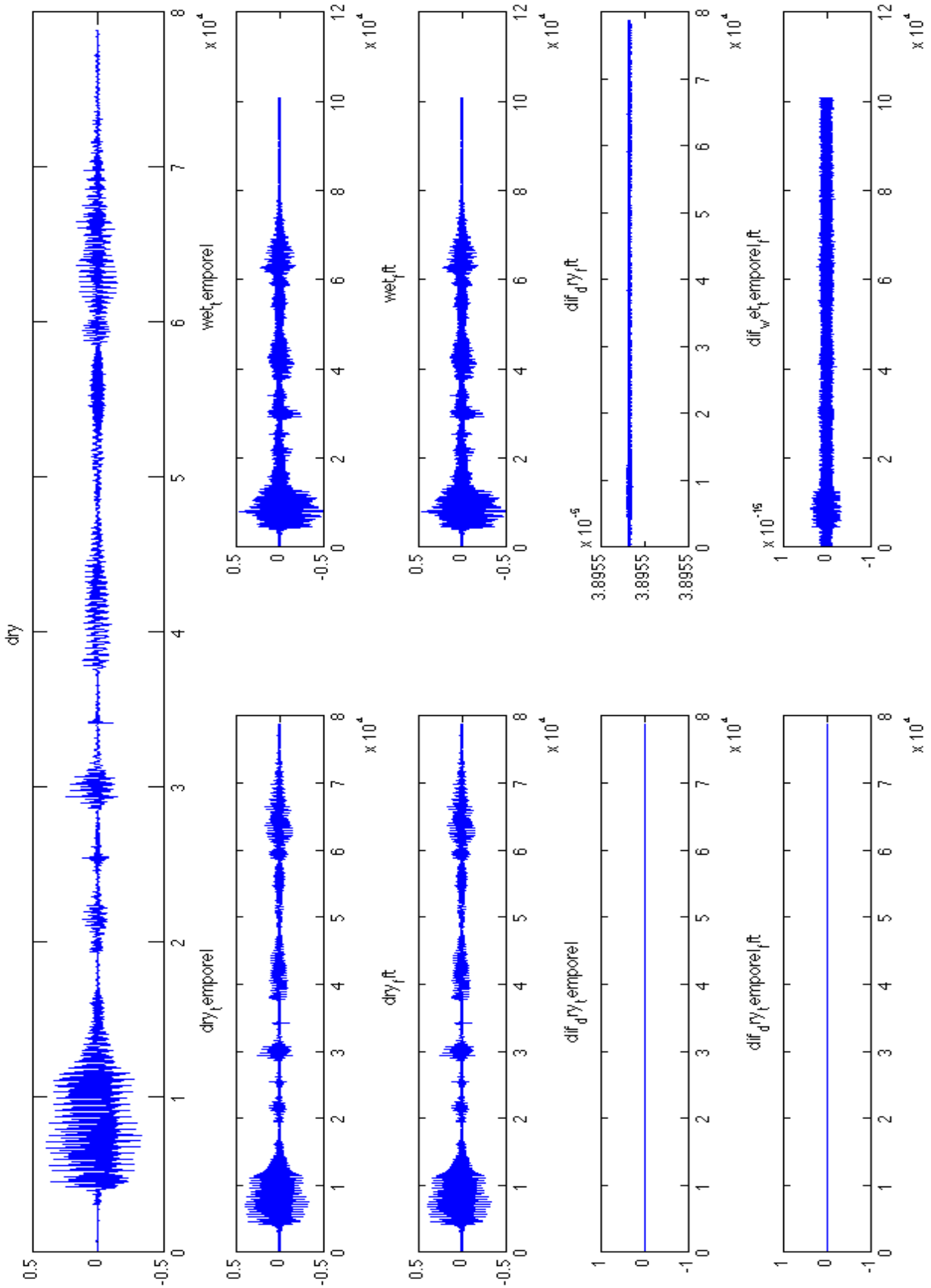


Fig 2-17 : *Caractéristiques de la réponse impulsionnelle du studio Lola Sous La Lune*

Fig 2-18 : *Comparaisons des dé-convolution polynomiales et FFT.*



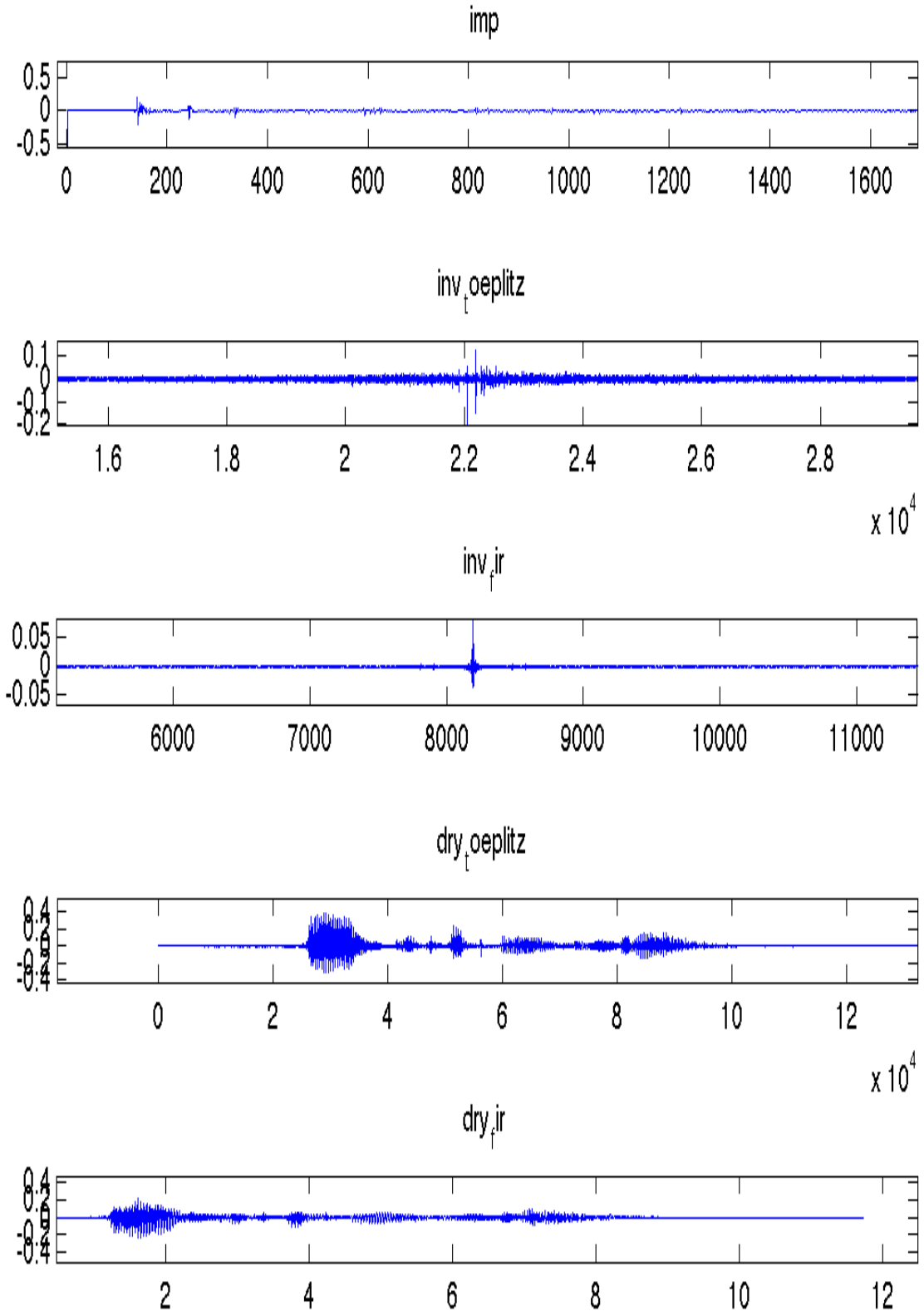


Fig 2-19 : *Convolution par les réponses inverses obtenues par matrice de Toeplitz et approximation FIR.*

2.4 Conclusions de l'étude théorique

Dans cette étude théorique nous avons décrit le fonctionnement des principaux outils ou méthodes pour contrer la réverbération, en envisageant sa suppression ou son annulation. Le travail sur l'enveloppe et la prédiction linéaire sont les deux seules vraies méthodes pouvant être catégorisées comme traitement de suppression. Nous avons donc vu apparaître une limite théorique de ces procédés dans le cadre de notre étude, restreinte à une situation de tournage à un seul microphone, puisqu'on ne peut modifier les premières réflexions par la prédiction linéaire ou le travail sur l'enveloppe sans altérer le son direct.

Ce constat nous a amenés à considérer des moyens de dé-convolution sans connaître la réponse impulsionnelle du canal de transmission. Nous avons alors envisagé la dé-convolution cepstrale comme une piste de développement. Si cette méthode s'avère valide sur le plan théorique, ses applications restent pour l'heure relativement limitées. Elle pourraient cependant constituer un bon complément aux techniques de prédictions linéaires si l'on parvenait à améliorer la différenciation du son direct et des premières réflexions.

Nous avons alors tenté de percer les secrets de l'application Unveil, dans le but de mieux cerner les principes sur lesquels ce logiciel repose. Selon le model envisagé nous pourrions l'apparenter aux traitements de dé-convolutions aveugles. La particularité de cet outil réside dans l'utilisation d'un réseau neuronal offrant une certaine flexibilité. Ce genre de démarches encore relativement marginales, semble offrir des perspectives intéressantes en vue de développement futur car elles permettent d'adapter le fonctionnement d'un algorithme aux conditions d'utilisation. Si l'on en croit, « *Extracting Room Reverberation Time from Speech Using Artificial Neural Networks* » [22], il serait

possible d'accroitre l'automatisation de ce dé-réverbérateur en utilisant le réseau neuronal pour déterminer le temps de réverbération.

En constatant la complexité des traitements mis en œuvres dans les démarches de détermination d'une réponse impulsionnelle inconnue, nous nous sommes alors demandé si un recours à la dé-convolution directe était envisageable. Nous avons esquissé les limitations théoriques de ces problèmes et constaté l'efficacité potentielle de ces traitements en situation contrôlée. Le chapitre suivant rend compte de l'expérimentation que nous avons mené afin de tester ces méthodes dans le cas d'une situation « réelle » correspondant à l'enregistrement d'un film avec un microphone.

3. APPLICATION PRATIQUE DES TECHNIQUES DE DE-REVEBERATION ENVISAGEES

Cette partie pratique s'articule autour d'un cas concret, la dé-réverbération de certaines séquences dialoguées du film « Premier Rôle » réalisé par Galaad Hemsi et produit par Désert Production. Ce film s'avère constituer un terrain d'application idéal pour notre projet car il concentre l'ensemble des problématiques soulevées par notre sujet. Il s'agit en effet d'un tournage ayant eu lieu dans des conditions très particulières nécessitant un recours à la dé-réverbération.

Nos étudions dans un premier temps la dé-convolution directe par une réponse impulsionnelle associée au lieu de tournage. Puis nous analysons les possibilités offertes et les limites intrinsèques des différentes méthodes de dé-réverbération que nous avons détaillées dans le chapitre précédent en nous focalisant sur les questions de savoir si ces outils engendrent des artefacts, s'il est possible d'intégrer les sons dé-réverbérés dans un mixage et si les effets de ces traitements sont tolérables pour le projet envisagé. Nous avons utilisé principalement quatre outils : NML_RevConvRR, Unveil, Cedar et Transient Master de Native Instrument. Les essais de dé-réverbérations relèvent de la phase de postproduction pendant le montage son et ont pour objectif de produire des éléments dé-réverbérés en vue du mixage. L'utilisation ou non de ces éléments traités sera alors laissée à l'appréciation du mixeur et du réalisateur lors du pré-mixage des paroles.

3.1. Présentation du projet

3.1.1 Le film

Inspiré par des expériences précédentes en tant qu'organisateur de casting pour des émissions de télé-réalité, Galaad Hemsi a voulu tourner un film sur ces moments de vie où tout peut basculer. Les acteurs se succèdent dans un décor identique, celui du casting. Des personnages en personnages, de situations en situations, une intrigue prend forme peu à peu. Ils pensent jouer pour décrocher le rôle mais les différents acteurs participent en fait à l'intrigue du même film. Rémi, le protagoniste est un jeune trentenaire qui part à la recherche de son identité à l'heure d'importants changements dans sa vie.

Cette histoire devient alors le moyen d'expression des enjeux du casting. Aux moments de jeux succèdent de vraies réactions, des doutes, de la colère, de la joie... Les acteurs subissent le rapport de force que leur inflige cette situation. Jusqu'où sommes nous capable d'aller pour parvenir à nos fin ?

3.1.2 Contexte de production

Le tournage s'est déroulé dans les locaux de « *Prép'Art* », une école d'art située au 23 rue de Ménilmontant, 75011 Paris. L'espace de tournage pourrait s'apparenter à un atelier d'artiste situé au rez-de-chaussée pour lequel le sol, le plafond et les murs, sont en béton nus ou recouvert de panneaux en bois. L'aspect dépouillé de ce lieu est une volonté de réalisation. Quelques fenêtres simple vitrage donnent sur la rue très passagère et la lumière du jour qu'elles laissent passer est utilisée pour compléter l'éclairage artificiel. Au plafond des néons donnent une lumière froide et génère un signal harmonique qui augmente le bruit de fond tandis que le sol est occupé par une dizaine de tabourets, quelques tables en bois ainsi que des travaux d'étudiants.

Pour tourner ce film l'ensemble de ce matériel a été déplacé pour être positionné près des fenêtres comme le montre la figure 3-1.

Fig 3-1 *: Le lieu de tournage vu de l'entrée.*

Durant le tournage six à huit personnes étaient présentes dans la salle. Les acteurs évoluent dans une zone délimitée, seuls ou par groupe (jusqu'à quatre personnes au maximum). Différents axes de caméra engendrent des placements microphoniques variés mais relativement localisés. La hauteur sous plafond est de l'ordre de 2,2 m. Le choix d'adopter plusieurs axes de caméra sensiblement différent engendre le recours à des placements microphoniques variés mais relativement localisés. La hauteur sous plafond est de l'ordre de 2,2 m. Plus de détails sont lisibles sur le schéma de l'annexe A qui présente une vue en coupe de ce lieu.

Comme le volume total de la pièce n'est pas très grand (80 m^3 environ), le temps de réverbération n'est donc pas très long bien que les surfaces en présence soient toutes très réfléchissantes : le coefficient d'absorption moyen pour le béton de l'ordre

de 0,01 et ceux du bois et du verre respectivement voisins de 0,03 et 0,02. Les comédiens évoluant dans une zone proche de ces surfaces, les premières réflexions sont donc marquées et le rayon critique est assez faible. Si l'on prend en compte les critères de Beranek, nous avons affaire à une salle claire et relativement agressive (car il y a peu de surfaces diffusantes hormis les objets présents au sol). Si le timbre de la réverbération est assez riche on note un coloration bas médium tout à fait logique.

3.1.3 Le dispositif de captation

Lors de ce tournage, la volonté initiale était de placer les acteurs dans des conditions de captation les plus proches de celles rencontrées lors d'un réel casting. Les comédiens ont été pré-sélectionnés, puis mis en situation, face à la caméra. Le tournage s'est déroulé en plusieurs jours et pour ne pas éveiller les soupçons, le réalisateur a d'abord choisi d'opter pour une équipe légère, ce qui explique pourquoi ce film a été tourné pour partie sans ingénieur du son. Comme on peu le voir sur certains plans, un pied de micro remplace le perchman tandis que deux microphones électrostatiques cardioïdes AKG C-451 ont été enregistrés directement sur la caméra Sony Z5 du tournage. Ces deux microphones ont été placés en bord cadres, leur capsules pointant parfois vers les acteurs mais pas toujours donc, du fait de la directivité de ces microphones ainsi que de leur placement particulier, d'importants effets de dé-timbrage ont été introduits dans les prises de son et ils sont audibles. Ces microphones placés sur la caméra sont de plus éloignés de quelques mètres de la source ce qui les situe parfois hors du rayon critique.

Si les deux microphones forment un couple ORTF il nous a semblé difficile d'exploiter les signaux enregistrés comme tel car les sources sont parfois en mouvement et/ou l'image stéréophonique obtenue n'est pas cohérente par rapport à l'image filmée... Aussi, lors du montage son, il a été décidé de n'utiliser qu'un seul de ces deux microphones. Il est évident qu'une telle configuration s'inscrit dans un contexte économique plutôt particulier.

Suite au visionnage des premiers rushs il a été décidé de faire appel à Renaud Duguet, ingénieur du son, afin d'assurer la prise de son. Ce dernier a donc équipé les acteurs de microphone cravate et d'émetteur haute fréquences tout en assurant une captation à la perche. Comme Renaud Duguet a aussi introduit de la moquette afin d'accroitre le coefficient d'absorption de certaines zones du sol et du plafond, l'influence acoustique du lieu est pour ces séquences très maîtrisée par rapport au reste des séquences tournées.

Les autres séquences peuvent ponctuellement poser des problèmes d'intelligibilité. Le manque de présence pour les séquences tournées le premier jour engendre une fatigue auditive sensible, qui rend le raccord entre les séquences tournées avec ou sans ingénieur du son plutôt délicat. De plus, les plans sonores du premier jour sont très dépendants de la position du système de prise de son et non de la valeur choisie pour le cadre puisque la caméra est munie d'un zoom et a parfois été tenue à l'épaule.

Pour résoudre ces problèmes le réalisateur et les acteurs ont souhaité éviter la postsynchronisation afin de conserver les réactions obtenues lors du casting puisqu'elles constituent la base même du film. Ceci nous conduit à chercher à disposer d'un outil qui permettrait de réduire de manière sensible l'influence de la réverbération afin de gagner en intelligibilité et de rendre plus facile le montage des différentes séquences. L'objectif n'est pas de supprimer complètement cette acoustique mais de la réduire pour améliorer le confort d'écoute, la compréhension du message et obtenir une spatialisation plus cohérente.

3.2. Protocole de mesure

Comme nous l'avons évoqué dans le chapitre précédent, il semble envisageable sous certaines conditions de réduire l'influence du champ réverbéré en ayant recours à la déconvolution. Le cadre de notre étude semble tout à fait se prêter à cette approche puisque toutes les scènes sont tournées en un même lieu et que les personnages comme le système de captation sont relativement statiques. Il a de plus été possible d'accéder au lieu du tournage afin de chercher à capturer sa réponse impulsionnelle. Aussi après avoir présenté les limitations théoriques de la déconvolution, nous étudierons dans ce chapitre si elle peut être envisagée pratiquement.

3.2.1. Capture de réponse impulsionnelle

Pour tester l'efficacité de la déconvolution nous devons dans un premier temps mesurer la réponse impulsionnelle du lieu de tournage. Et comme cette étape conditionne l'ensemble des traitements ultérieurs, il faut définir une chaîne de captation adaptée au problème que nous devons résoudre.

Conception de la séquence de stimuli

Les scènes que nous souhaitons traiter contiennent des voix françaises de nature et d'intonation différentes aussi pour se rapprocher des signaux à traiter, nous avons construit une séquence de stimuli contenant plusieurs registres vocaux. Nous avons donc choisi d'utiliser des voix de comédiens enregistrées lors de séances de postsynchronisation pour nous affranchir au maximum de l'acoustique du lieu de captation. Notre séquence débute avec une voix d'homme murmurant le texte suivant : *« C'est une attaque bionique à laquelle nous ne sommes pas préparés ».* Le stimulus suivant est une voix de femme portée puisqu'il s'agit d'une tirade de Roxane dans Cyrano de Bergerac. Le troisième stimulus est une voix d'homme elle aussi portée,

correspondant à la situation où un commandant de bord fait une annonce aux passagers d'un avion. Le quatrième stimulus est la voix de Roxane murmurée. Le cinquième stimulus est une voix brésilienne portée avec des sifflantes plus marquées.

Ce panel de stimuli a été constitué afin de nous permettre de couvrir les différents registres des voix présentes dans les séquences à traiter mais, quelles soient portées ou murmurées, ces voix présentent peu d'énergie en dessous de 200 Hz comme au dessus de 5 kHz. On peut donc supposer que l'estimation de la réponse impulsionnelle pour ces deux bandes de fréquences sera incertaine. Ainsi afin d'assurer une capture optimale nous diffusons aussi un bruit rose puisque ce signal comporte toutes les fréquences du domaine audible, bien que son caractère aléatoire rende délicate la déconvolution dans le domaine temporel. On a aussi choisi de diffuser un sweep exponentiel pour être en mesure de déconvoluer nos prises de son par convolution du signal inverse de ce sweep. De plus, comme nous l'avons précédemment indiqué cette étape devrait permettre la limitation des distorsions harmoniques introduites par le système de diffusion.

Enfin, pour rendre plus perceptibles les effets du lieu de diffusion nous avons aussi utilisé un extrait de l'enregistrement anéchoïque des Noces de Figaro extrait du disque « Anechoic Orchestral Music Recording » édité par Denon en 1989. La mise en relation de ces différents stimuli devrait permettre le choix de la réponse impulsionnelle la plus adaptée aux prises de son que nous devons traiter.

Nous avons construit une séquence test rassemblant chacun de ces stimuli. Si les prises de son du film ont été enregistrées sous 48 kHz, 24 bits avec un format linéaire, nous avons préféré réaliser nos mesures à la fréquence d'échantillonnage de 96 kHz pour accroitre la précision temporelle. Nous pouvons alors facilement revenir au format d'origine, sans trop altérer nos mesures grâce à une simple décimation de facteur 2. La dynamique de quantification adoptée est de 24 bits afin d'optimiser le rapport signal à bruit car, comme la convolution est une opération nécessitant un grand nombre de calculs, le choix de la dynamique de codage a une importance non négligeable compte tenu des multiples étapes de sommation mis en jeu.

Afin de pouvoir déterminer les retards temporels relatifs entre la séquence test et de les mesures sur site, chaque stimulus est mis en forme suivant le même schéma : 1,5 s de silence numérique ; un clic correspondant à un unique échantillon négatif à pleine échelle; 5 s de silence numérique ; le stimulus original ; à nouveau 5 s de silence numérique ; un unique échantillon positif à pleine échelle puis 1,5 s de silence. Ces durées ont été choisies de manière arbitraire en estimant à l'oreille le temps de réverbération du lieu de tournage en cherchant que chaque plage de silence soit suffisamment longue pour permettre à la décroissance de se développer complètement tout en restant suffisamment courte pour minimiser les temps de mesures.

Le niveau de diffusion est une donnée importante puisqu'en effet les stimuli doivent être assez forts pour laisser émerger la réverbération du bruit de fond tout en respectant la dynamique de la source utilisée. Or, comme notre séquence test rassemble des signaux avec des profils dynamiques assez différents il est absolument nécessaire de procéder à l'étalonnage de chacun des stimuli. Puisque le bruit rose et le sweep ont un niveau moyen important tandis que l'extrait musical ou les enregistrements vocaux présentent un niveau crête élevé associé à un niveau RMS assez faible, nous avons donc choisi de compresser la dynamique de ces enregistrements et d'atténuer les signaux de synthèses. L'objectif recherché étant la maîtrise de la forte dynamique des stimuli induite par une captation de proximité. Ces enregistrement on étés réalisés à l'aide d'un microphone TLM 49 placé à une vingtaine de centimètres de la bouche des comédiens.

Ce travail réalisé, on a calculé du profil dynamique de chaque stimulus à l'aide du plug-in « *Dolby Media Meter »* en utilisant comme critère la mesure de loudness intégrée (integrated EBU) puisqu'elle tient compte de l'ensemble du phénomène. La valeur cible de -20 dB FS a semblé constituer une valeur cohérente, puisque de nature à conserver une réserve dynamique suffisamment importante pour les crêtes des signaux de parole. Nous avons bien sûr veillé à ne pas dépasser le 0 dB True Peak.

3.2.2 Organisation Matérielle

Découpage de l'espace

Afin de s'adapter au mieux aux différentes possibilités de localisation de la source et du micro lors du tournage, on procède au quadrillage de l'espace. Un maillage de 1*1m nous permet de couvrir l'ensemble de la zone concernée. Chaque case se trouve alors à l'intersection d'une lettre et d'un chiffre (c. f. Annexe B) .

Système de diffusion et de captation

La lecture et l'enregistrement de la séquence ont été réalisés à partir du séquenceur Samplitude ProX installé sur un MacBook Pro 13" 2.8GHz core i7 sur une partition Boot Camp et Windows 7. L'interface audionumérique utilisée est une RME Fireface 800 en firewire 800 permettant la réalisation de mesures au format 96 kHz/24 bits.

La source sonore utilisée pour ces tests est une enceinte de monitoring active JBL LSR4326P, ce choix étant le fruit d'un compromis. En effet, dans le temps qui nous était imparti, seules deux enceintes étaient disponibles : une enceinte Tannoy *Systeme 600* et une enceinte JBL LSR4326P. Or, la mise en œuvre de l'enceinte Tannoy aurait été plus fastidieuse car il aurait fallut amener puis installer sur site un amplificateur d'une centaine de watt qui aurait constitué une source de bruit supplémentaire s'il avait été placé dans le lieu de capture.

Si les fiches caractéristiques (cf. figure 3-2) de ces deux modèles d'enceintes annoncent des performances fréquentielles similaires. La documentation des enceintes Tannoy ne nous renseigne pas sur le niveau maximal délivrable en continu alors que l'enceinte JBL est censée pouvoir fournir 106 dB RMS à un 1 m et laisser passer les crêtes du signal jusqu'à des niveaux de 112 dB.

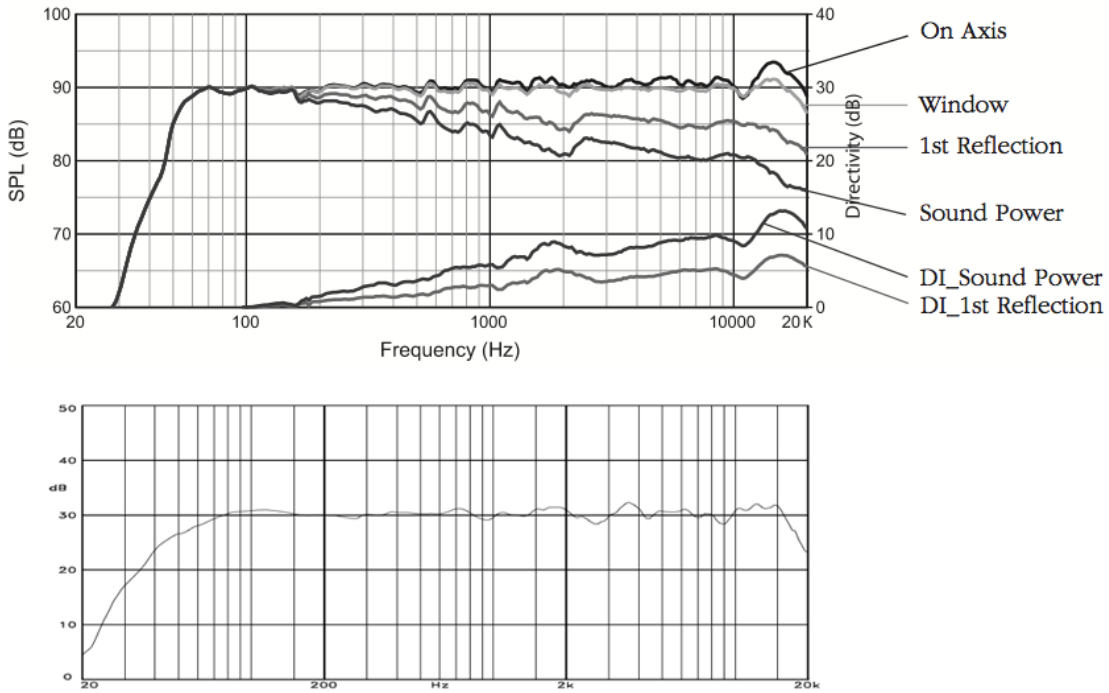


Fig 3-2 : *Courbes constructeur du niveau en fonction de la fréquence*. (en haut) *JBLLSR4326P au dessus (1 graduation = 5dB) [18] – (en bas) Tannoy system 600 en dessous (1 graduation =10dB) [19].*

Ceci constitue le point faible de cette enceinte pour cette application. En effet le plancher de bruit du local se trouve aux environs de 40 dB SPL donc, si on veut pouvoir mesurer l'intégralité du TR60, il faut disposer d'une source pouvant fournir au moins 100 dB à 1 m en continu. Comme certains stimulus présentent une dynamique importante (jusqu'à -2 dB True Peak), il faut que l'enceinte puisse développer cette marge dynamique ce qui signifie qu'il est nécessaire d'aligner le -2 dB True Peak sur la valeur de 112 dB SPL crête. Cela revient à faire correspondre notre -20 dB intégré EBU à un niveau mesuré au droit de l'enceinte de 112-20 = 92 dB SPL.

Comme le signal capté à quelques mètres est beaucoup plus faible cette valeur s'est vite avérée insuffisante. C'est la raison pour laquelle nous avons d'abord envisagé une diffusion à 95 dB SPL pour finalement adopter une diffusion à 100 dB SPL en régime permanent à 1 m de l'enceinte. Les mesures ont débuté dans la matinée puis se sont poursuivies dans l'après-midi du samedi 30 mars. Le bruit de la rue n'a cessé d'augmenter au fil de la journée aussi, afin de limiter les phénomènes de saturation

potentiels liés à l'enceinte utilisée, le niveau des stimuli trop dynamiques a été baissé de 5 dB en cours de journée.

Afin d'améliorer le rapport signal à bruit nous aurions éventuellement pu tester l'utilisation d'une enceinte plus puissante telle que celles des séries 108p/112p de L-Acoustics. Des mesures nocturnes auraient pu constituer une alternative intéressante mais, comme le lieu du tournage se situe en zone résidentielle, nous avons choisi de renoncer à cette option car le règlement d'une contravention pour tapage nocturne aurait représenté une part conséquente du budget alloué pour réaliser ce mémoire de fin d'études...

Calibrage de la source

L'étalonnage du niveau de diffusion s'est fait dans le lieu à l'aide d'un sonomètre placé au droit de l'enceinte à 1 m. A l'aide du séquenceur nous avons généré un bruit rose et procédé au réglage du gain numérique au sein du logiciel jusqu'à atteindre le niveau sonore choisi.
Nous avons utilisé pour la diffusion l'entrée SPDIF de la JBL_LSR26P car elle opère la conversion numérique analogique et l'amplification du signal. L'inteface Fireface était donc reliée à l'enceinte par un câble coaxial de 75 ohms d'impédance caractéristique. La hauteur de cette enceinte a été déterminée après analyse des scènes tournées. Comme pour les séquences les plus problématiques les acteurs sont assis sur un tabouret, nous avons donc pris cette hauteur comme référence pour le placement vertical de l'enceinte.

Système de captation

Puisque la réponse impulsionnelle caractérise l'ensemble du canal acoustique, nous avons souhaité mettre en œuvre un dispositif identique à celui utilisé lors du tournage. Nous avons alors utilisé un des deux microphones AKG C-451, branché directement sur la caméra Sony Z5. Nous avons aussi enregistré le signal provenant d'un autre microphone AKG C-451 dans le séquenceur servant à la diffusion. Mais ces microphones cardioïdes présentent une coloration prononcée.

Comme le montre la figure 3-3, le haut du spectre (5-20 kHz) est favorisé par ce microphone, tandis que le registre grave (20-100 Hz) est légèrement atténué. Et du fait de sa directivité, le détimbrage de la source sera d'autant plus marqué si cette dernière dépasse les 60 degrés d'incidence.

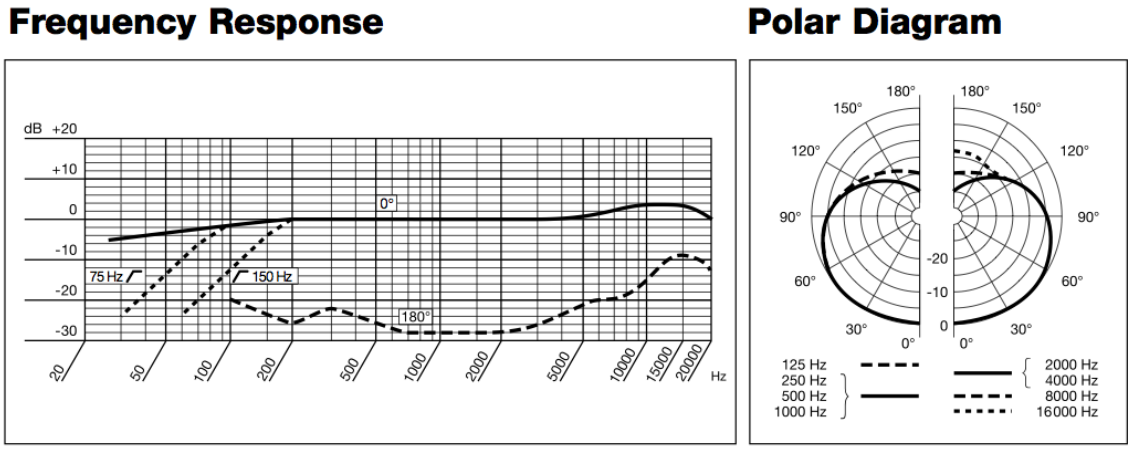


Fig 3-3 : *Réponse en fréquence et directivité du microphone AKG C-451, d'après [20].*

Afin de pouvoir s'affranchir de ces effets et de comparer ces microphones à un système plus neutre, nous avons aussi utilisé deux microphones DPA 4006 TL équipés d'ogives visant à les rendre omnidirectionnels pour une plus grande plage spectrale. Si l'on ne considère que la réponse en fréquence associé à cet appareillage des microphones, ce choix semble peu opportun car une remontée d'environ 5 dB peut être notée pour l'octave 10-20 kHz. Mais en considérant son diagramme de directivité, cet effet est compensé par une incidence normale à l'axe du microphone, raison pour laquelle nous avons utilisé ces microphones verticalement lors des mesures.

Un microphone DPA 4006 a donc été placé au même endroit que le microphone utilisé pendant le tournage pour capter la réponse impulsionnelle en ce point de la salle. Afin de limiter l'influence de l'enceinte, nous avons placé en champ proche de cette dernière un autre microphone DPA 4006 (apparié au précèdent), microphone qui a été utilisé comme microphone de référence.

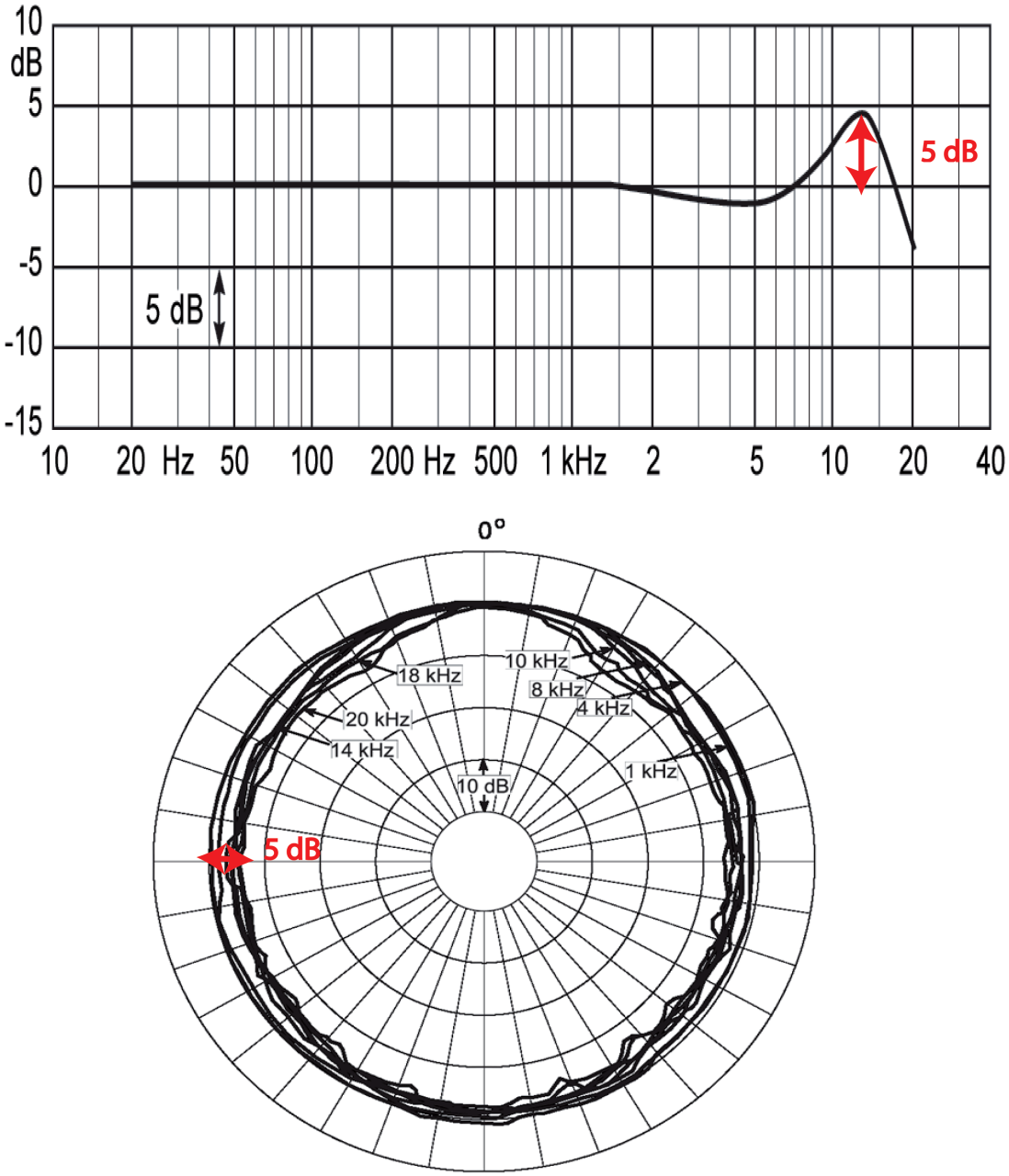


Fig 3-4 *: Réponse en fréquence et directivité DPA 4006 TL équipé du « nose cone », d'après [21].*

L'enceinte utilisée comporte deux voies raccordées au voisinage de 2,4 kHz associées à deux membranes non coïncidentes. Nous cherchons à exciter l'ensemble du spectre lors de la diffusion de notre séquence test. Aussi puisque l'objectif consiste à réaliser une déconvolution en prenant comme référence le champ acoustique direct et non le stimulus, il a fallu déterminer le placement optimal du microphone au droit de

l'enceinte en agissant suivant les trois axes. On a donc, dans un premier temps, choisi de placer le microphone de manière équidistante par rapport aux deux membranes, tout en le plaçant dans l'axe de l'enceinte. Mais, comme nous avons pu observer que la majeure partie de l'énergie des signaux vocaux était restituée par le boomer, nous avons choisi de privilégier légèrement celui-ci afin d'obtenir une meilleure précision pour cette zone spectrale.

Il a aussi fallu placer ce microphone suffisamment proche de l'enceinte pour ne pas être perturbé par le champ réverbéré tout en étant conscient qu'une distance trop faible aurait pour effet d'atténuer la contribution de l'une des deux membranes.

Enfin, d'après les travaux menés par Johann Lescure dans « Simulation Temps Réel de Prises de son Multicanales » [22], le champ acoustique en proximité d'une enceinte ferait apparaître des variations importantes de la vitesse de propagation acoustique comme le montre la figure 3-5.

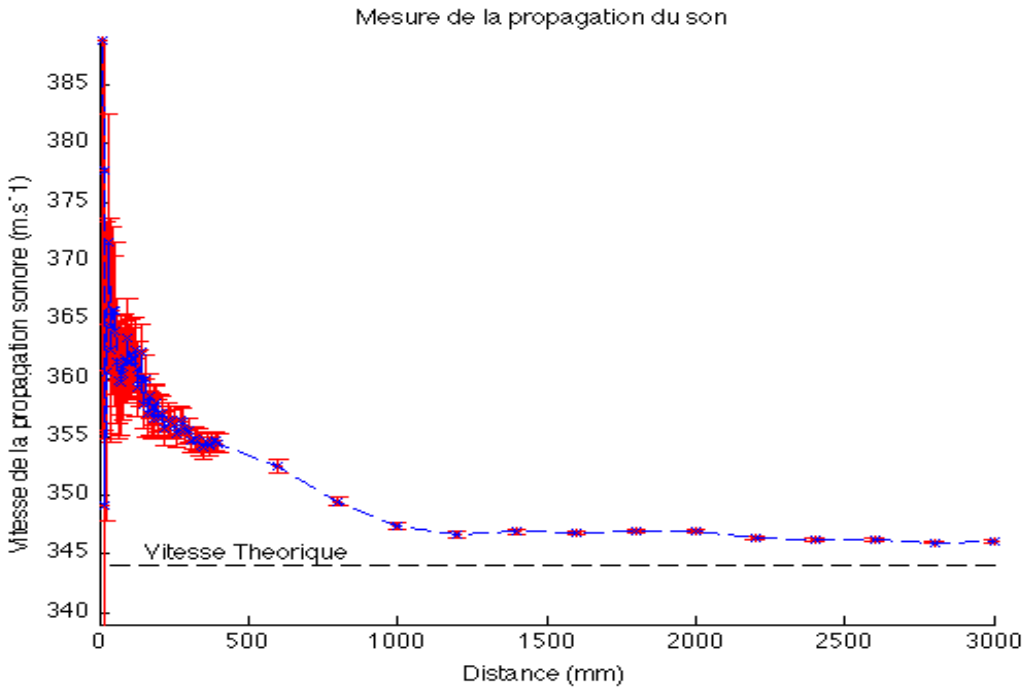


Fig 3-5 : *Vitesse de propagation acoustique en champ proche, d'après [22], p. 95.*

Afin de minimiser l'influence de ce phénomène, il apparaît nécessaire de se placer à une distance d'au moins 50 cm de la source, mais puisque cette distance s'est avérée trop grande pour obtenir un rapport direct à réverbéré suffisant dans le cadre de la configuration expérimentale testée, nous avons rapproché ce microphone d'une dizaine de centimètres en veillant, à l'écoute, à conserver un rendu aussi homogène que possible.

La pré-amplification et la conversion du signal s'effectuent à l'aide de RME Micstasy qui constituent l'interface d'acquisition via une liaison SMUX comme le montre la figure 3-6. Le réglage des gains d'entrée est réalisé en plaçant tous les microphones au point le plus proche de la source associé au placement du microphone de référence. Quand nous plaçons les microphones à l'endroit où se trouvait le comédien, nous conservons le réglage de gain adopté pour le placement en proximité de l'enceinte. Cela nous permet d'accéder à l'atténuation effective de niveau entraînée par la propagation en milieu réverbéré depuis la position utilisée comme référence pour la déconvolution (le microphone placé en ultra proximité) jusqu'au point où se situait le comédien.

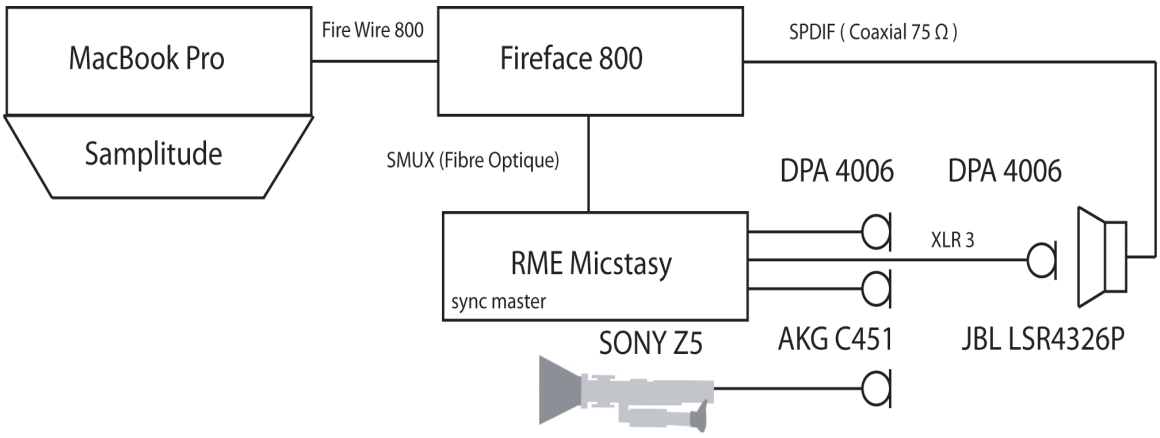
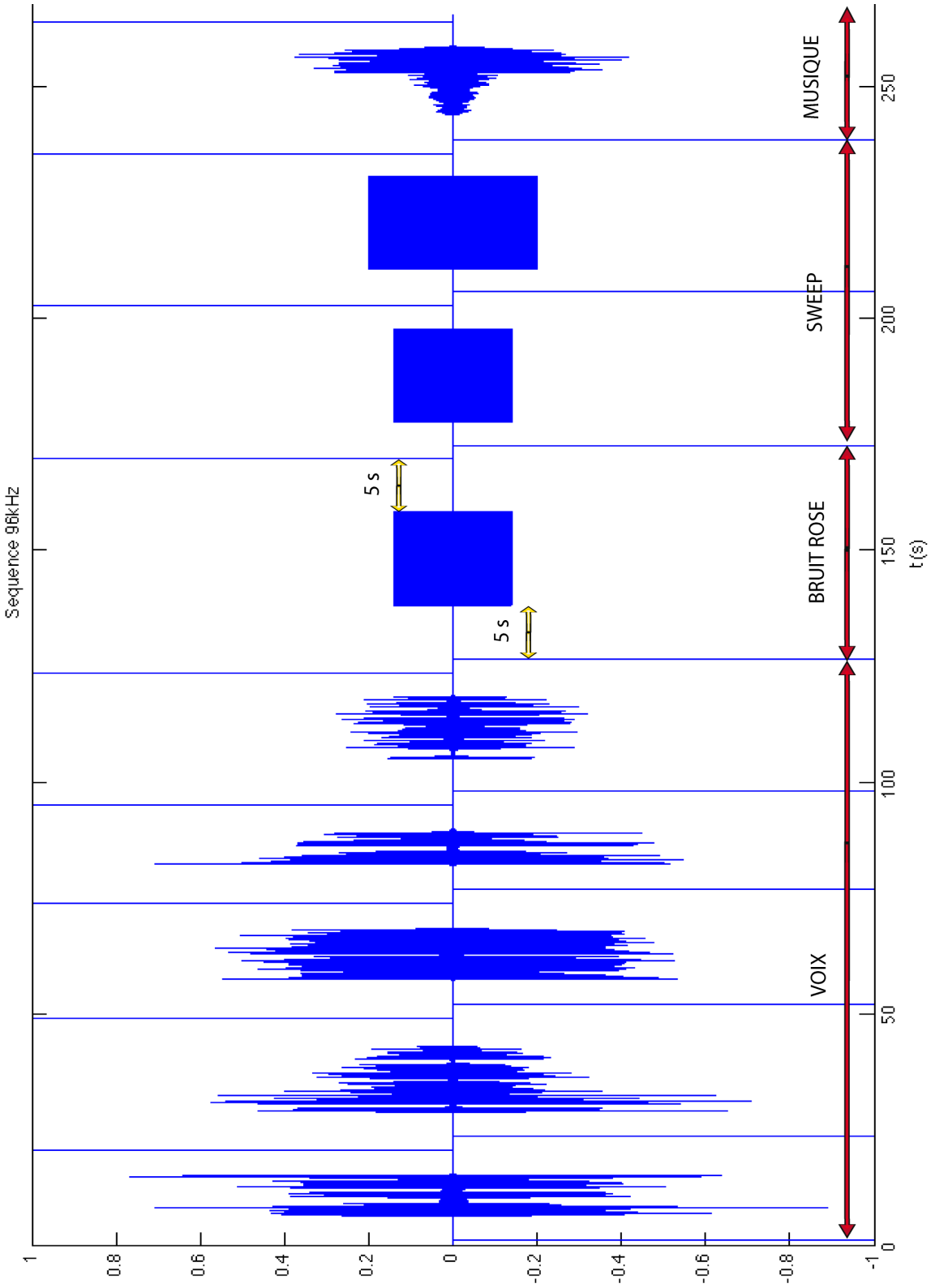
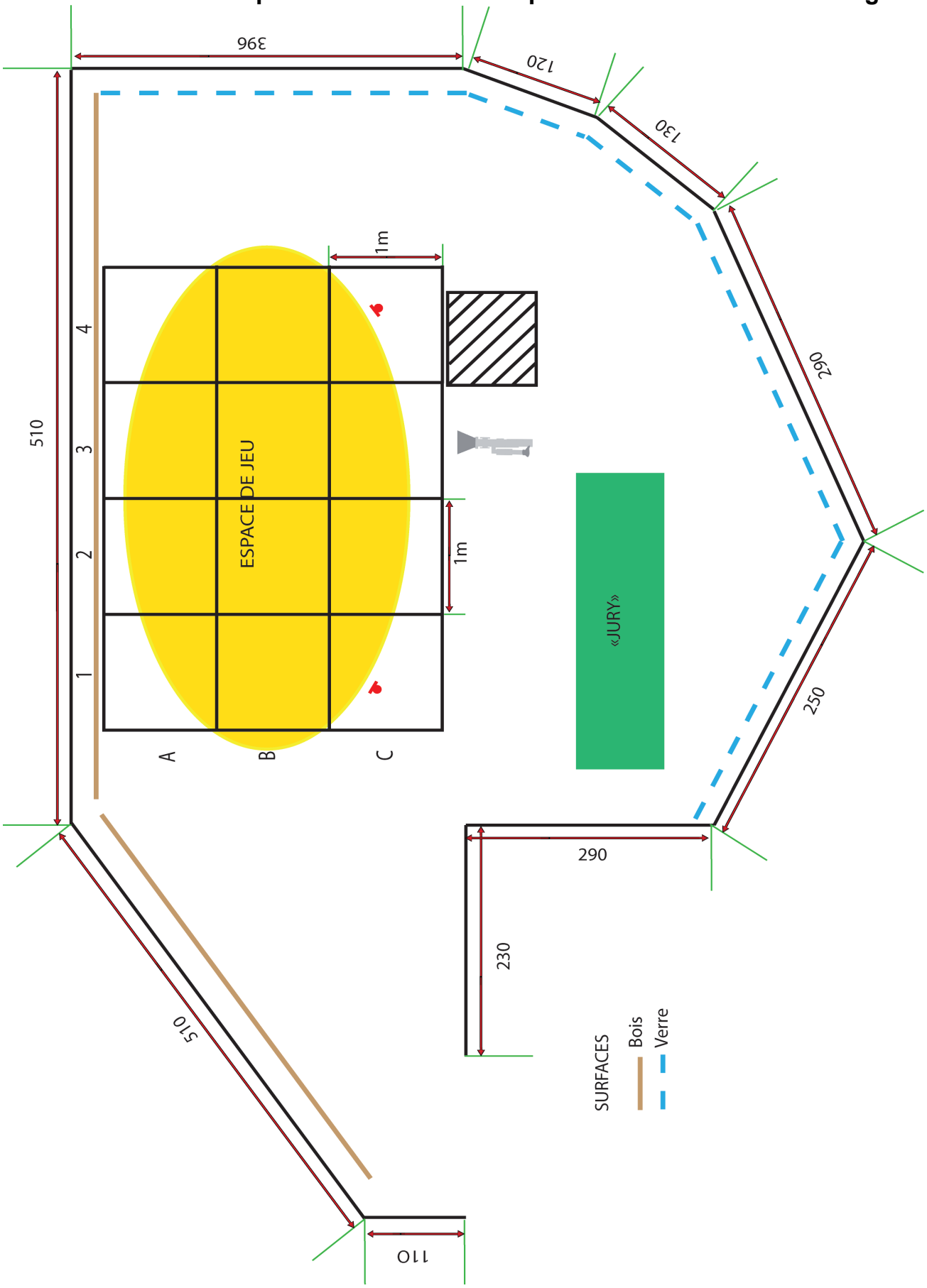


FIg 3-6 : *Synoptique de la chaine de captation des réponses impulsionnelles.*

ANNEXE A : Représentation Temporelle de la Séquence de Stimulus



ANNEXE B : Représentation Schématique de la Situation de Tournage



3.3 Déconvolutions appliquées à la séquence du notaire.

Parmi les séquences problématiques, nous avons choisi de nous concentrer sur la scène dans laquelle Rémi rencontre le notaire afin d'évaluer la pertinence du procédé mis en œuvre. Les deux personnages sont statiques (ils sont assis face à face) et il est possible de déterminer l'emplacement du microphone car ce dernier est visible sur certains plans (cf. figure 3-7). Grâce à l'analyse de ces images, on identifie la position du notaire, qui constitue la source active pour la scène étudiée, et l'on peut placer l'enceinte en A3. Les microphones M2 (DPA) et M3 (AKG) sont eux placés de manière coïncidente en C1. On oriente l'enceinte en direction des microphones et l'on diffuse la séquence test.

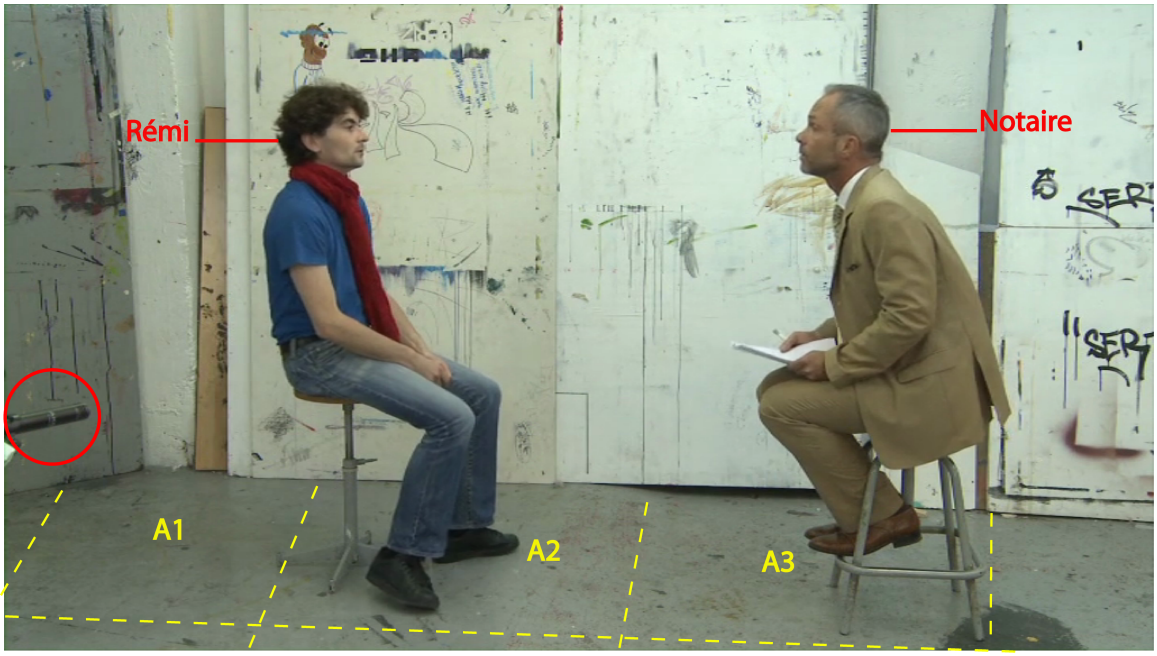


Fig 3-7 : *Schématisation de la configuration de tournage – Séquence du Notaire. Image extraite de « Le Premier Rôle » - © Désert Production.*

3.3.1 Application de l'approche fréquentielle

On procède alors au calcul des réponses impulsionnelles associées à chaque stimulus par déconvolution FFT grâce la fonction Matlab BruteDecon.m dont le listing est reporté en annexe. Avec les réponses obtenues on est à même de déconvoluer les signaux enregistrés pour s'approcher du stimulus d'origine. Cependant cette opération ne se fait pas sans artefacts puisque des effets de précédence plus ou moins marqués apparaissent. Ils ne concernent généralement pas tout le spectre, mais seulement certaines régions du spectrogramme.

L'annexe C montre les effets de la déconvolution pour le stimulus de sweep exponentiel. La présence de réverbération se traduit par des harmoniques évoluant parallèlement au signal d'origine. Après dé-réverbération, ces composantes sont atténuées mais l'on voit alors apparaître les artefacts de la déconvolution matérialisés par la présence de traits horizontaux sur le spectrogramme. On constate alors que les zones pour lesquelles ce phénomène est le plus marqué correspondent à des minimas du spectre de la réponse impulsionnelle.

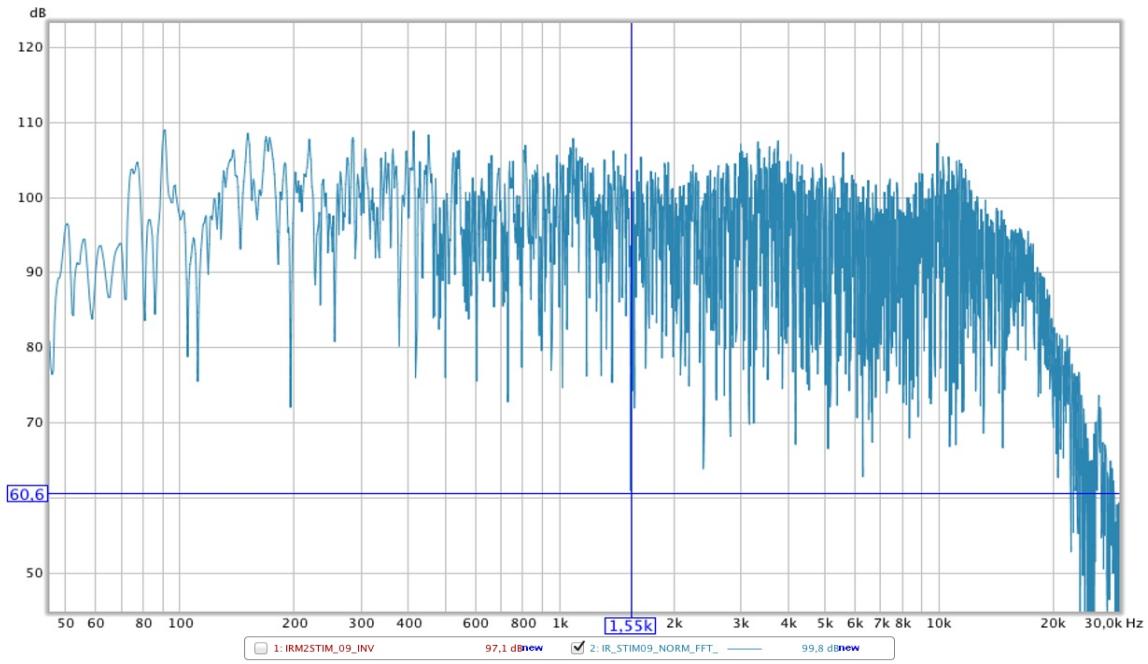


Fig 3-8 : *Réponse en fréquence non lissée de la RI du sweep exponentiel (obtenue par dé-convolution de M2 par rapport au stimulus).*

On voit donc apparaître les problèmes de la division fréquentielle des spectres associés respectivement au signal réverbéré et au signal de référence évoquée dans le chapitre précédent. De plus ces artefacts ne se manifestent pas de manière identique quand on considère le résultat de la déconvolution en fonction du stimulus utilisé pour la mesure. On peut utiliser une réponse impulsionnelle obtenue à l'aide d'un autre stimulus pour réaliser la déconvolution mais cela conduit parfois à l'apparition d'artefacts supplémentaires. En fait, la qualité du résultat de la déconvolution dépend sensiblement de la quantité d'énergie présente dans chaque bande fréquentielle pour le stimulus considéré. On peut remarquer grâce à l'annexe F que si la réponse en fréquence de chaque réponse impulsionnelle est relativement proche, les différences les plus notables apparaissent avec l'étude du retard de groupe. On peut remarquer que les retards de groupe obtenus pour les différents stimuli ne peuvent pas être considérés comme constants. De plus les filtres associés à ces réponses impulsionnelles ne sont pas à phase minimale, c'est ce qu'illustre les figures 3-9 et 3-10 pour l'exemple du premier stimulus.

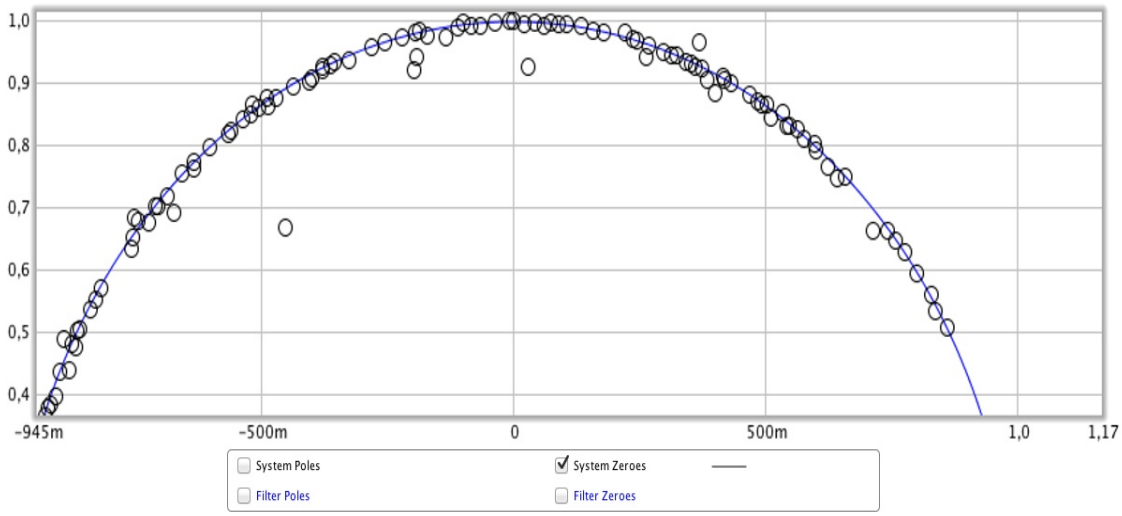


Fig 3-9 : *Représentation sur les deux premiers cadrans du cercle unité des zéros de la réponse impulsionnelle obtenue par déconvolution du premier stimulus capté par M3 (AKG C-451).*

La figure 3-9 montre que le filtre associé possède des zéros hors du cercle unité, il n'est donc pas à phase minimale et son inverse ne sera ni stable, ni causal.

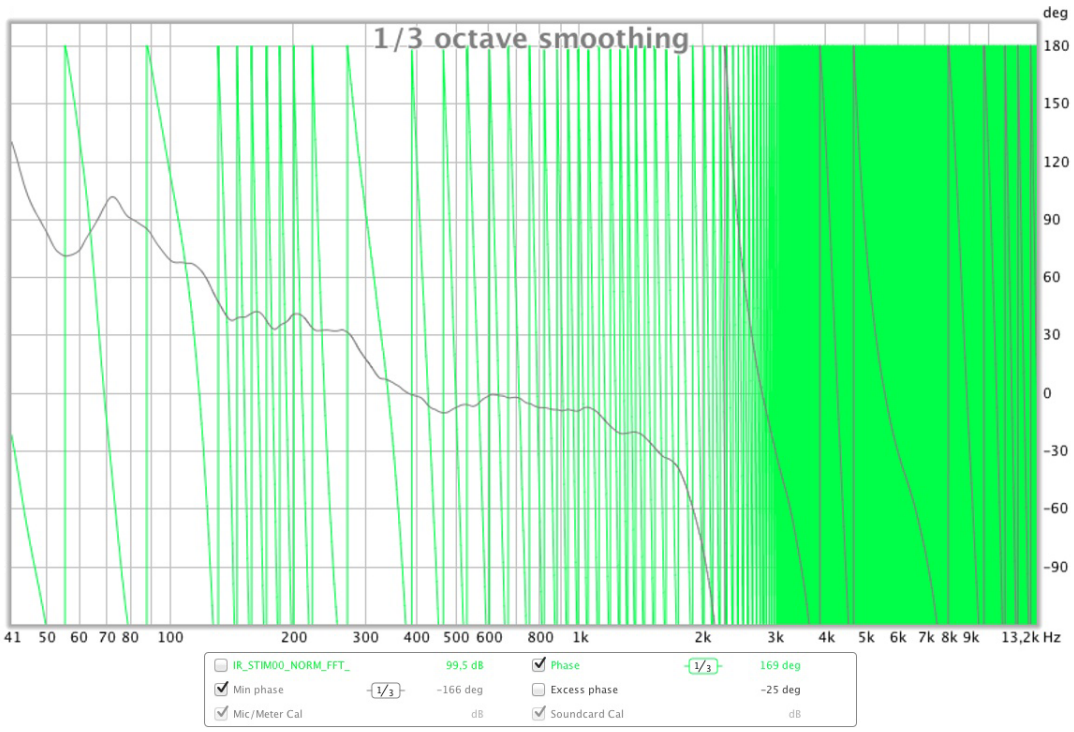


Fig 3-10 : *Phase réelle (en vert) et phase minimale (en gris) de la réponse impulsionnelle obtenue par déconvolution du premier stimulus capté par M3 (AKG C-451).*

La figure 3-10 traduit cette affirmation dans le domaine fréquentiel en représentant la phase réelle et le filtre à phase minimale correspondant sur une représentation à 360°. Cela explique alors la présence d'artefacts lors de la déconvolution. La partie non causale de la réponse impulsionnelle engendre notamment des effets de précédences représentés par des traits horizontaux sur les spectrogrammes des Annexes C et D.

L'annexe D illustre ce qui se passe pour le premier signal vocal de la séquence test une fois dé-réverbéré par les réponses impulsionnelles obtenues pour différents stimuli. Le stimulus 09 est le sweep présenté en annexe C et la réponse impulsionnelle qui lui est associée est celle pour laquelle le retard de groupe est le plus important parmi l'ensemble des stimuli. Le spectrogramme correspondant montre que le signal déconvolué par cette réponse est très bruité.

Nous pouvons aussi chercher une approximation du filtre inverse associé à la réponse impulsionnelle par un filtre à réponse impulsionnelle finie, en utilisant par exemple la

fonction inv_imp.m disponible en annexe. Malheureusement l'efficacité de cette méthode, basée sur la convolution du signal réverbéré par le filtre inverse, semble donner des résultats de qualité limitée car si cette approche permet de réduire les effets de précédences induits par des retards de groupes importants, elle conduit à des résultats présentant toujours une réverbération conséquente. De plus, avec cette méthode, les voix traitées présentent une couleur presque robotique.

3.3.2 Application de l'approche temporelle

La déconvolution polynomiale mise en œuvre dans l'étude théorique pose quelques problèmes pratiques.

Comme le premier échantillon du fichier de référence et du fichier à déconvoluer doivent être non nuls, chaque stimulus commence par un échantillon de valeur -1. Il faut alors resynchroniser l'ensemble des enregistrements réalisés avec chacun des stimuli avant d'entreprendre une division polynomiale. Cette étape peut être réalisée manuellement dans un séquenceur grâce à la forme d'onde ou bien être réalisée de manière automatique à partir de la recherche du maximum d'inter-corrélation entre le signal de référence et le signal réverbéré. Cette remise en phase permet alors de s'affranchir du pré-délai (que l'on peut obtenir aussi directement par déconvolution FFT) et faciliter ainsi le travail de division polynomiale. Il est ensuite possible de réintégrer ce paramètre après le calcul de la division en retardant la réponse impulsionnelle obtenue de la valeur de ce pré-délai.

Le second impératif imposé par l'implémentation de la déconvolution polynomiale réside dans la longueur des fichiers à traiter puisque la durée de l'enregistrement conditionne la longueur de la réponse impulsionnelle. Il faut en effet estimer la durée nécessaire en fonction du temps de réverbération du local. Dans l'étude réalisée nous avons noté que le TR60 obtenu après déconvolution est inférieur à 1,5 s, raison pour laquelle les fichiers enregistrés seront 1,5 s plus longs que les stimuli.

Si la détermination de la réponse impulsionnelle peut être réalisée en quelques minutes à partir d'un enregistrement de 30 s, la déréverbération s'est avérée beaucoup plus problématique puisque nous ne sommes pas parvenus à déconvoluer la réponse impulsionnelle du signal réverbéré. Avec les moyens dont nous disposons, seule l'obtention de la réponse impulsionnelle dans le domaine temporel a pu être réalisée. La dé-réverbération d'une séquence de 30 secondes nécessite plusieurs heures de traitement et donne des résultats incohérents : du silence et quelques échantillons à la pleine échelle. Les essais simulés dans la partie théorique avec une réponse impulsionnelle connue, permettaient de dé-réverbérer complètement le signal avec moins d'artefacts. Cette étape ne paraît pas réalisable pratiquement. Il est probable que le bruit de fond du local perturbe considérablement la dé-convolution. Cette manière d'envisager la déconvolution dans le domaine temporel est avant tout un algorithme très mal conditionné, qui peut très vite diverger lors d'une application sur des signaux long et complexes.

Sans pouvoir mettre en œuvre la dé-réverbération dans le domaine temporel, nous restons pour l'instant confrontés aux mêmes problèmes que ceux précédemment évoqués (artefacts et effets de précédences). Si la déconvolution polynomiale n'offre pas de résultats satisfaisants dans notre étude, nous pourrions en revanche envisager de transposer l'algorithme de déconvolution de J. Usher. L'implémentation de Laurent Millot permet de calculer le résultat d'une déconvolution de deux signaux par détermination d'un filtre adaptatif basé sur la minimisation d'un signal d'erreur.

3.4 Conclusions sur l'approche par dé-convolution

Dans ces conditions il nous est possible, dans une certaine mesure, de dé-réverbérer les sons émis lors de la capture de réponse impulsionnelle. La dé-convolution par rapport à M1 (DPA-4006 placé au droit de la source) ne permet pas d'annuler complètement la réverbération des signaux captés par M2 ou par M3. Cela s'explique par le fait que M1 capte de manière considérable le champ réverbéré (malgré sa proximité avec le haut parleur). Dans ces conditions l'intérêt M1, placé en proximité du

haut parleur s'avère donc très limité. A partir de l'annexe E on remarque des retards de groupes importants pour les microphones omnidirectionnels (M1, M2). De fait, la dé-réverbération des signaux captés en utilisant des réponses impulsionnelles obtenues en mettant en œuvre ce couple de microphones omnidirectionnels entraine plus d'artefacts que ceux induits en utilisant le microphone M3 (directivité cardioïde) c'est pourquoi afin de dé-réverbérer les sons du film on préférera les réponses impulsionnelles obtenues par dé-convolution du microphone M3 et de chacun des stimuli.

Malheureusement, aucune de ces réponses ne permet de dé-reverbérer les sons du tournage. On peut alors remarquer que le notaire n'est jamais tourné vers le micro lors de la prise de son. On suppose alors que ce détimbrage engendre des variations notables dans l'obtention du champ réverbéré. Pour tenter de modéliser cette situation on oriente l'enceinte dans la direction privilégiée par sa voix lors du tournage, puis on réitère l'opération, afin d'obtenir de nouvelles réponses impulsionnelles. Là encore, aucune de ces impulses ne permet de dé-réverbérer la séquence.

Comme le décrit le synoptique, ces enregistrements ont étés doublés sur la caméra afin d'assurer une chaine d'enregistrement identique à celle du tournage. Cependant comme le montre l'annexe F, les différences dues aux étages de pré-amplification et de conversion de la Sony-Z5 sont assez minimes donc on suppose qu'elles ne sont pas directement responsables de cet échec.

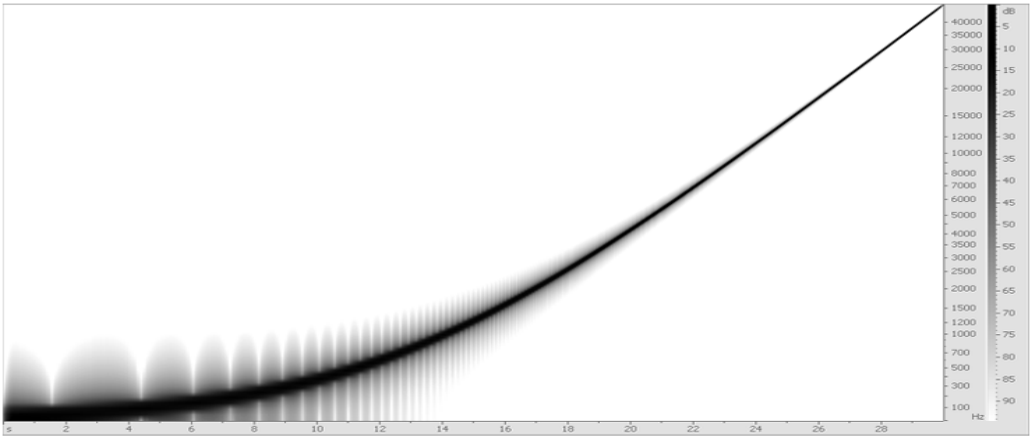
On peut premièrement noter que la fonction « Auto Gain » de la camera était activée lors de la prise en main de l'appareil (cette fonction a été désactivée lors des captures de réponses impulsionnelles). Il est donc très probable que cette dernière ait été utilisée au moment du tournage. Lorsque cette fonction est activée, le système devient fortement non linéaire et la convolution n'a donc plus lieu d'être. Dans l'hypothèse où nous disposerions de l'ensemble des paramètres de compression, nous pourrions envisager une expansion dynamique préalable, en vue de contrer les effets de l'auto gain (à la manière des procédés mis en œuvre dans le Dolby NR). En l'absence de ces données, il nous est difficile de conclure sur le rôle de l'auto gain dans cette étude.

Nous pouvons par contre affirmer l'importance de la nature de la source. En laissant l'ensemble du système de prise de son à la même place, nous avons remplacé l'enceinte par un narrateur. On est alors en mesure de déterminer la réponse impulsionnelle de M2 en prenant M1 en référence. Le faible niveau de la parole entraine une RI très bruitée mais qui permet de retrouver un signal proche de celui capté en M1, avec les limitations que l'on connaît. Il n'est cependant pas possible d'utiliser les réponses issues des stimuli diffusés par l'enceinte pour dé-convoluer la voix du narrateur. On suppose alors que la directivité de la source joue un rôle prépondérant. On constate que la voix parlée n'excite pas le lieu de la même manière. L'empreinte capturée par l'enceinte n'est pas une approximation suffisante du phénomène permettant d'annuler la réverbération de la voix prononcée dans le lieu. Cependant la réponse impulsionnelle obtenue avec cette voix ne permet pas d'annuler la réverbération du tournage. Le faible rapport signal à bruit de cette méthode complique considérablement la résolution du problème.

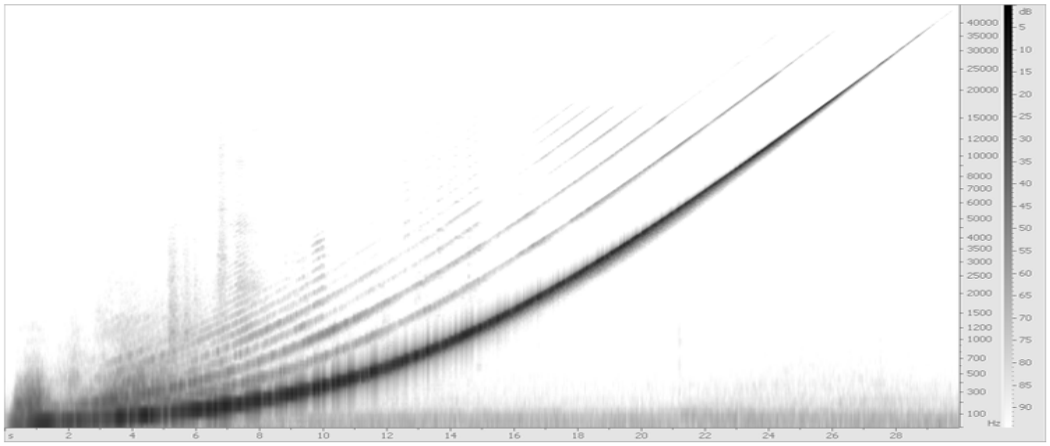
Si les réverbérations à convolution peuvent permettre d'obtenir un placement de source perceptivement proche de celui obtenu par un enregistrement in-situ, la déconvolution nécessite une connaissance beaucoup plus précise du canal acoustique. Néanmoins ce travail pourrait-être exploité dans le but d'orienter plus précisément les traitements visant à estimer le champ réverbéré. Le temps de réverbération et le contenu spectral fournis par ces empreintes sont des données importantes, pouvant orienter par exemple, un réseau neuronal artificiel...

ANNEXE C : EFFETS DE LA DECONVOLUTION

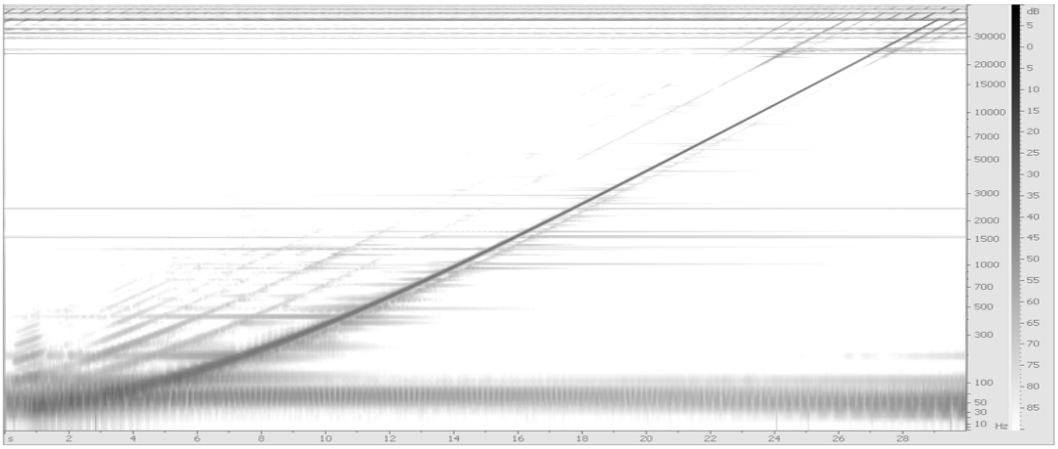
Stimulus 9 - Sweep Exponentiel



Sweep réverbéré - DPA 4006 - Position A3C1

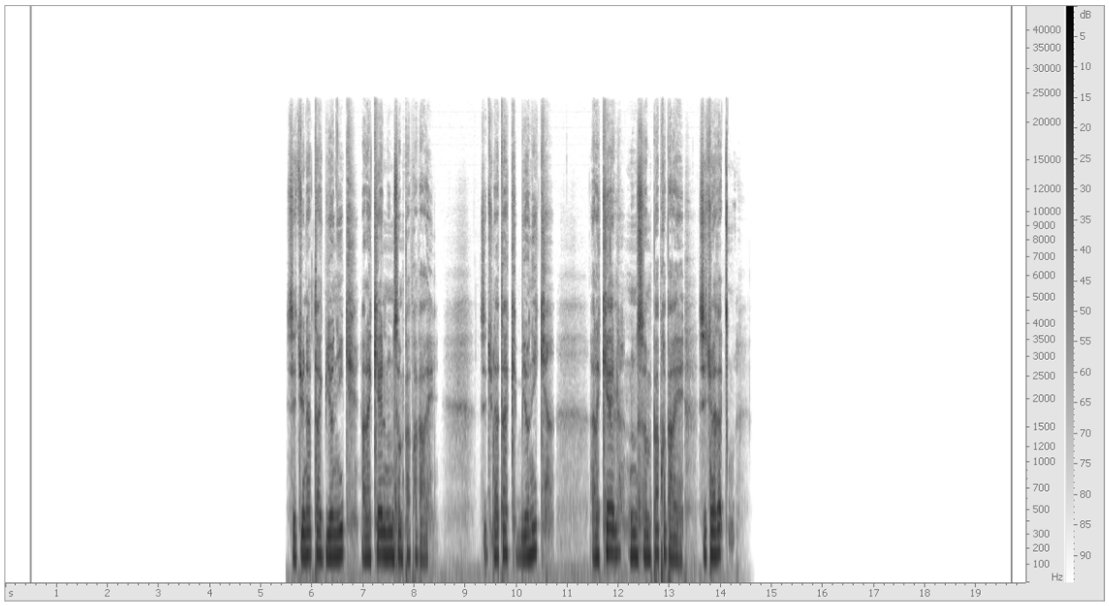


Sweep de-réverbéré - Dé-convolution FFT

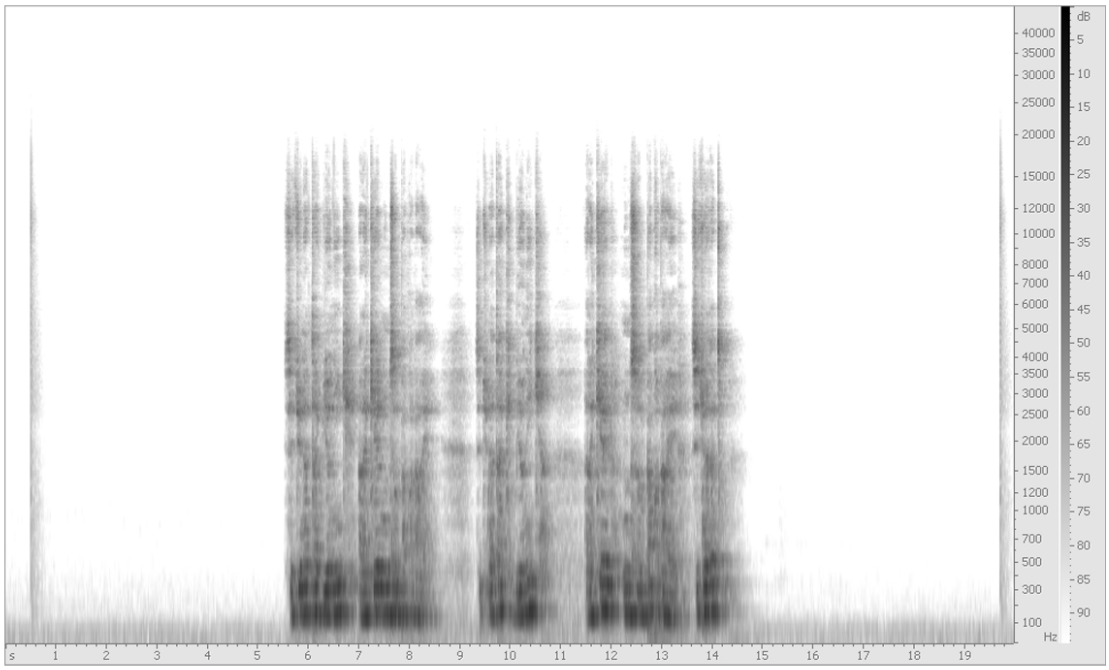


ANNEXE D-1/3: DECONVOLUTIONS CROISEES

Stimulus 0 - Voix Homme

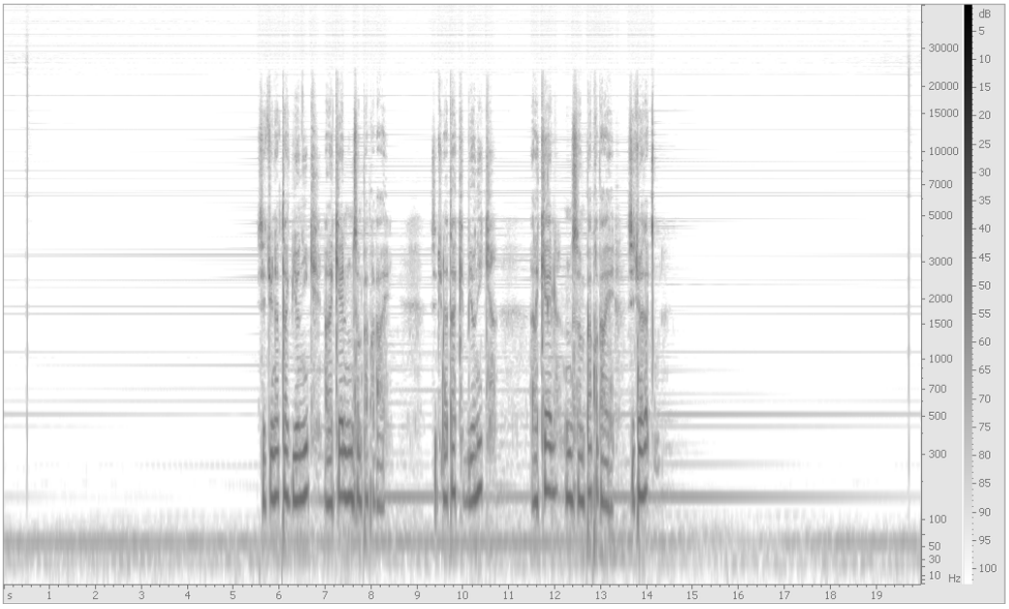


Stimulus 0 réverbéré - DPA 4006 - A3C1

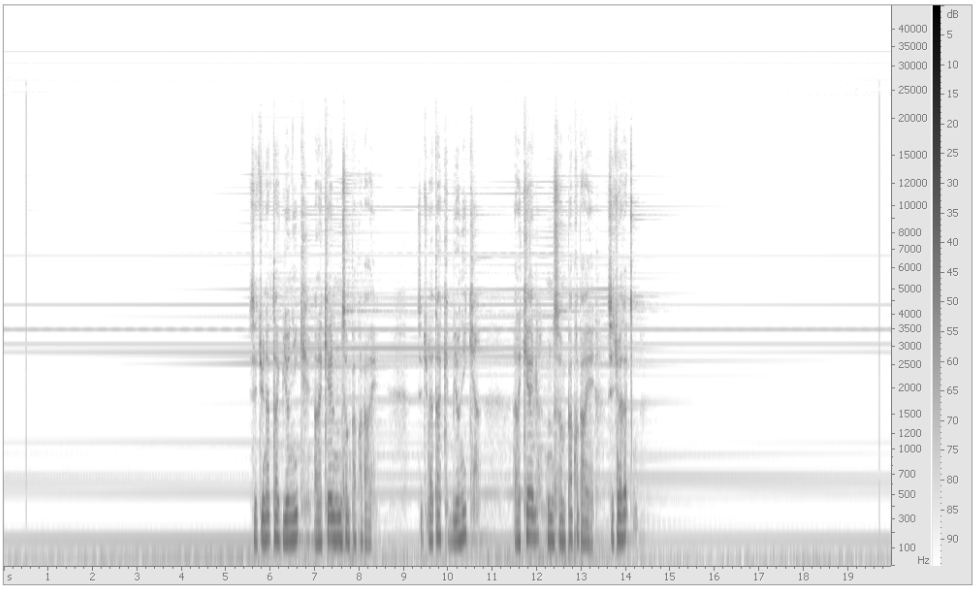


ANNEXE D-2/3: DECONVOLUTIONS CROISEES

Stimulus 0 - Dé-réverbéré par IR STIM0

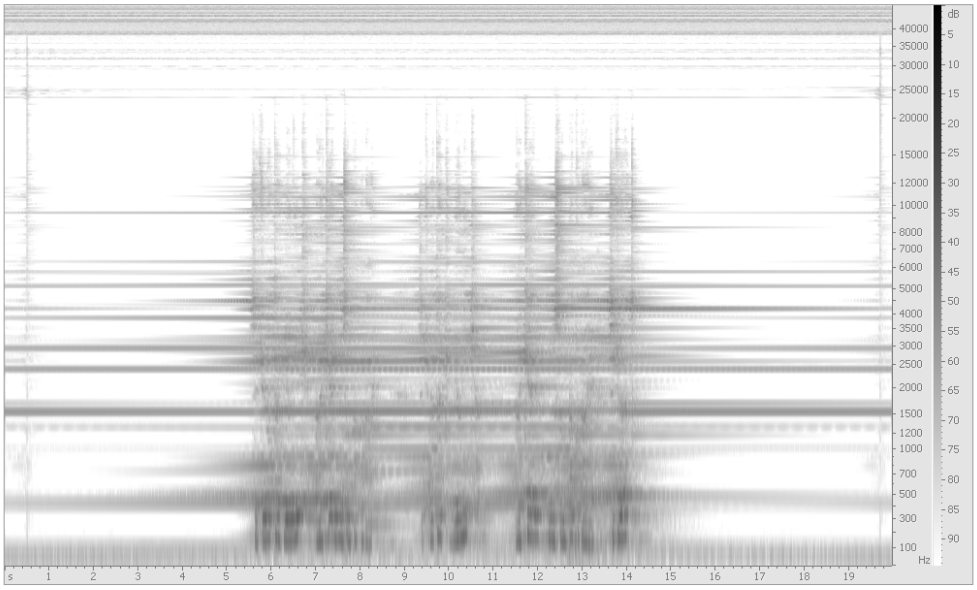


Stimulus 0 - Dé-réverbéré par IR STIM1



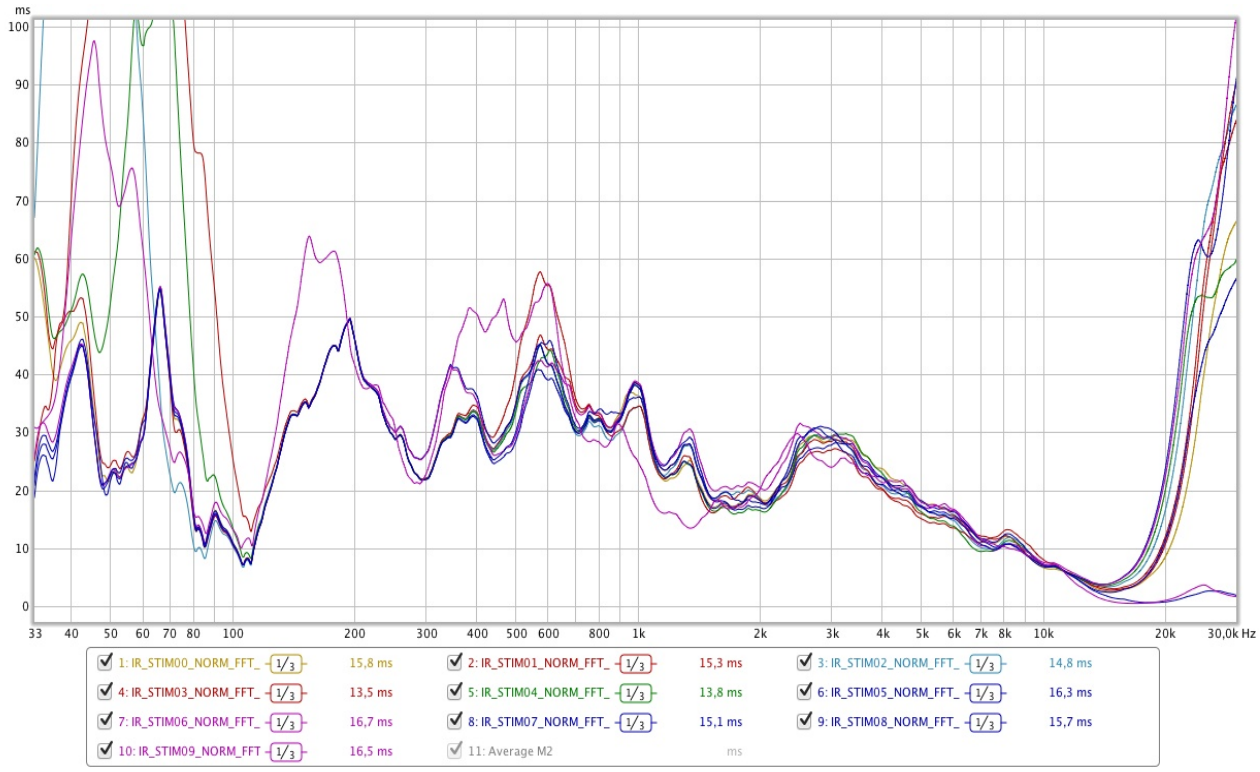
ANNEXE D-3/3 : DECONVOLUTIONS CROISEES

Stimulus 0 - Dé-réverbéré par IR STIM9

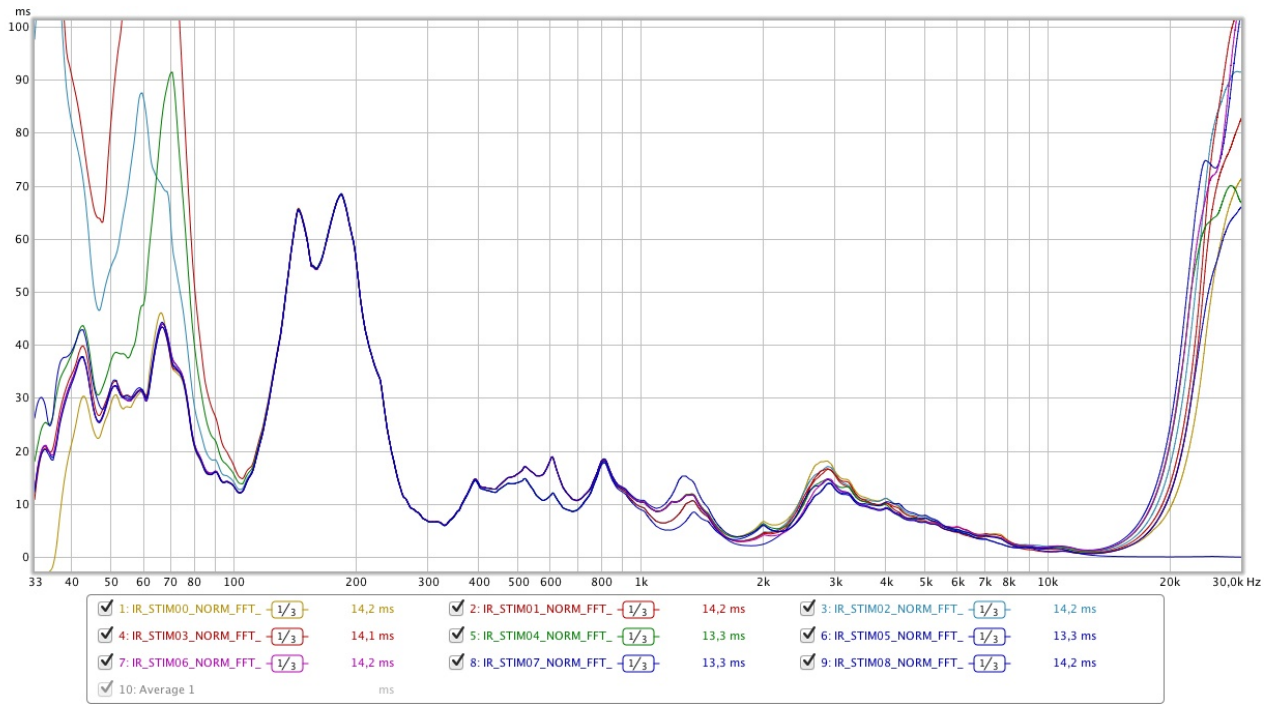


ANNEXE E-1/2: DELAIS DE GROUPE DES RI – Axe Camera

Délais de groupe des différents stimuli – Micro M2 – position : A3C1 – Lissage 1/3 oct

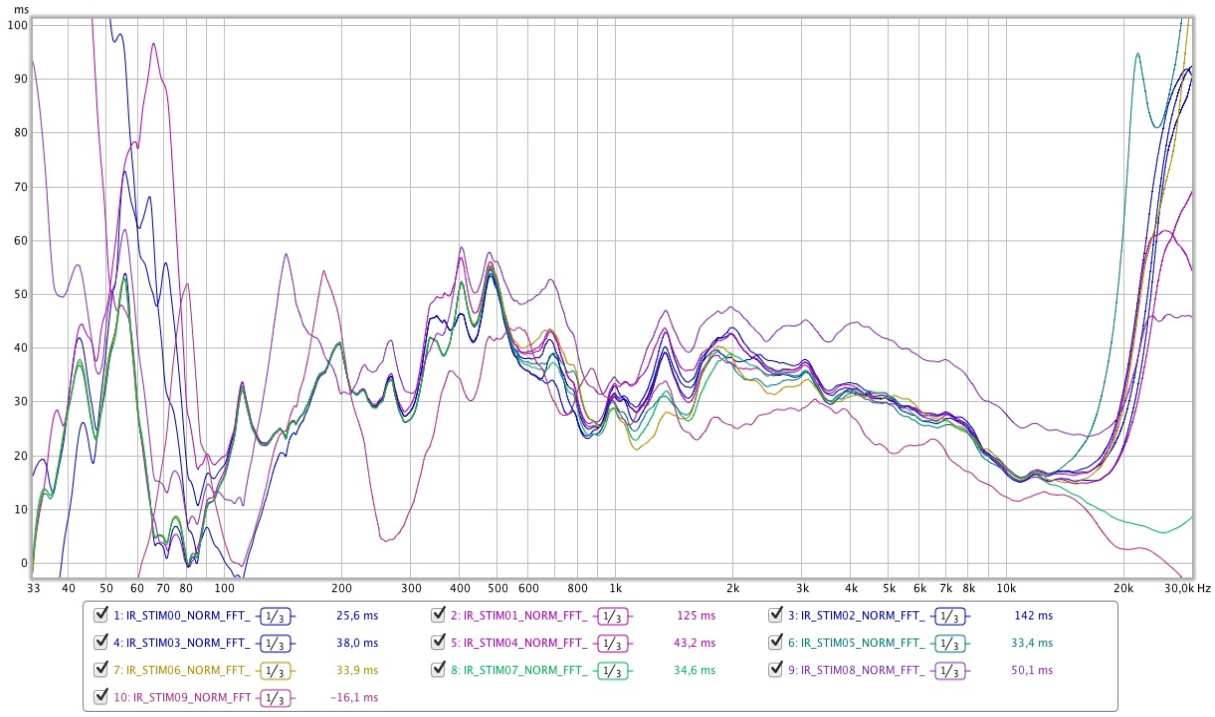


Délais de groupe des différents stimuli – Micro M3 – position : A3C1, - Lissage 1/3 oct

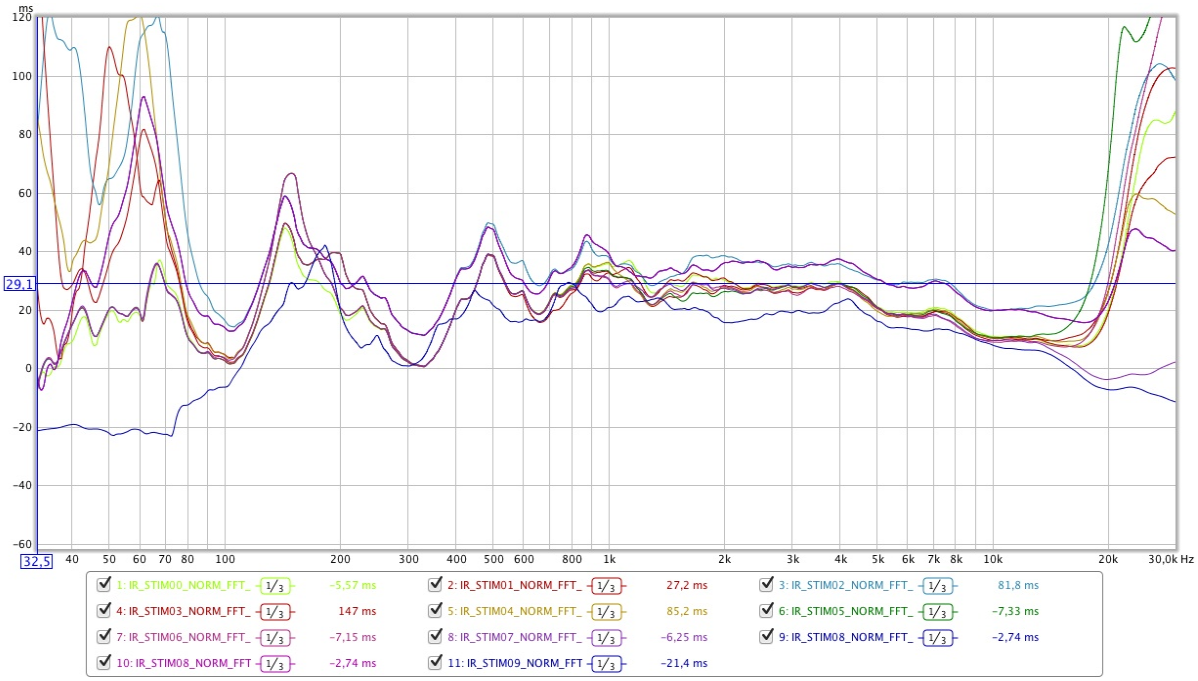


ANNEXE E-2/2: DELAIS DE GROUPE DES RI – Hors axe

Délais de groupe des différents stimuli – Micro M2 – position : A3C1 – Lissage 1/3 oct

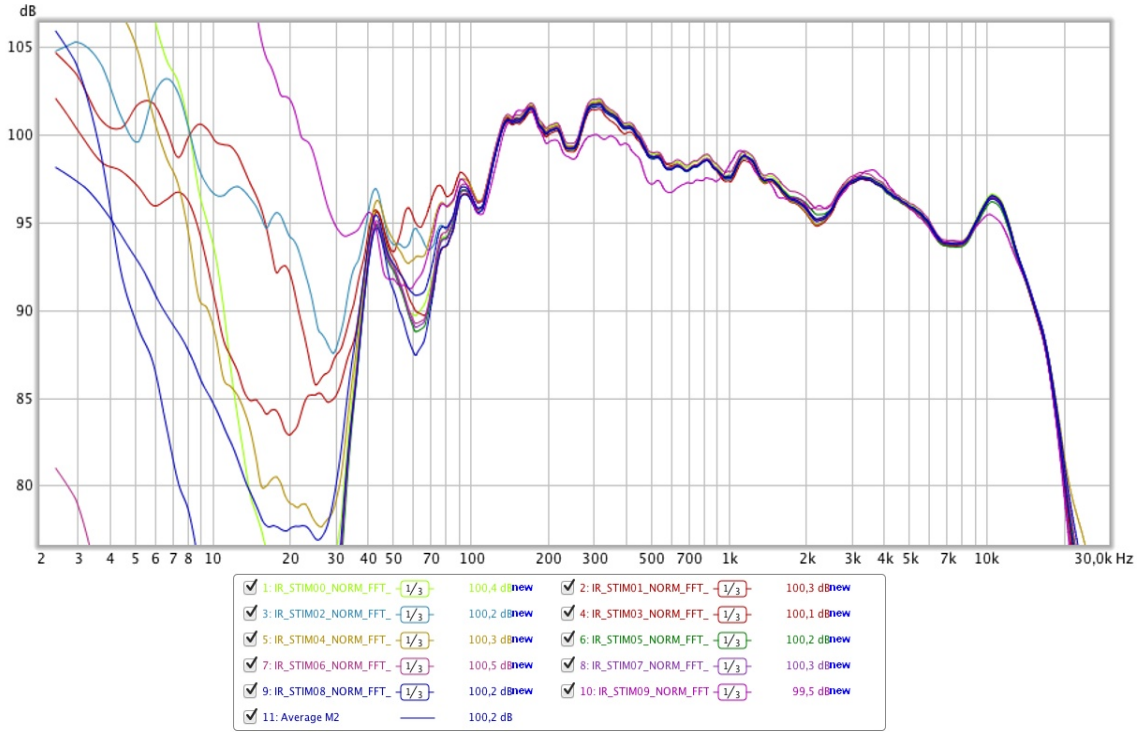


Délais de groupe des différents stimuli – Micro M3 – position : A3C1 – Lissage 1/3 oct

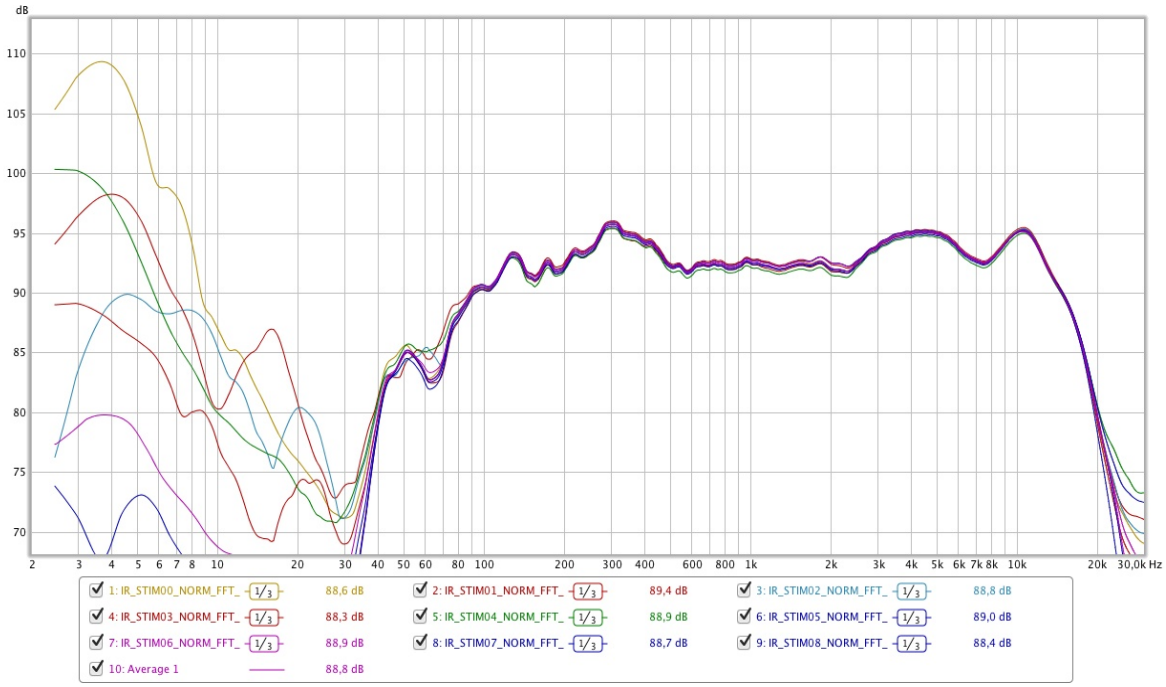


ANNEXE F-1/2: REPONSES FREQUENTIELLES DES RI – Dans l'axe

Micro M2 – position : A3C1 Dans l'axe – Lissage 1/3 oct



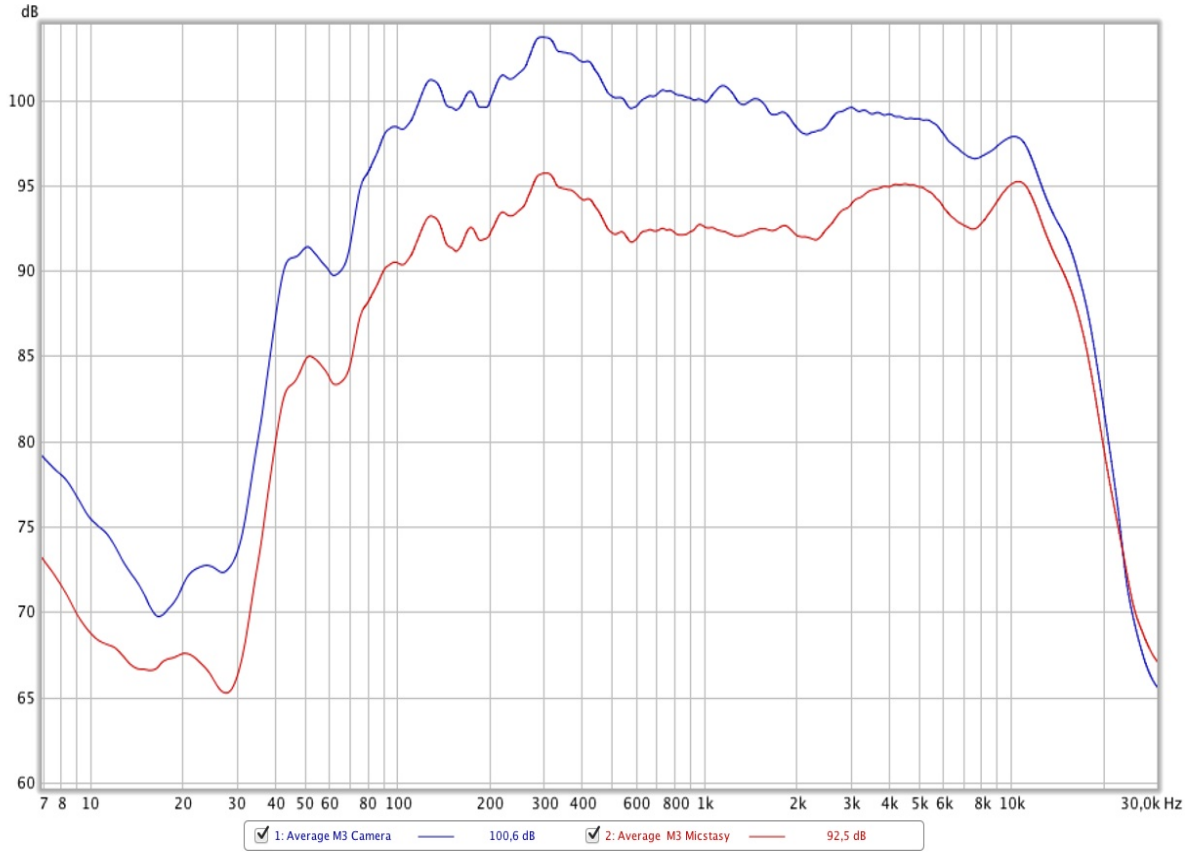
Micro M3 – position : A3C1 Dans l'axe – Lissage 1/3 oct



ANNEXE F-2/2: REPONSES FREQUENTIELLES DES RI – Dans l'axe

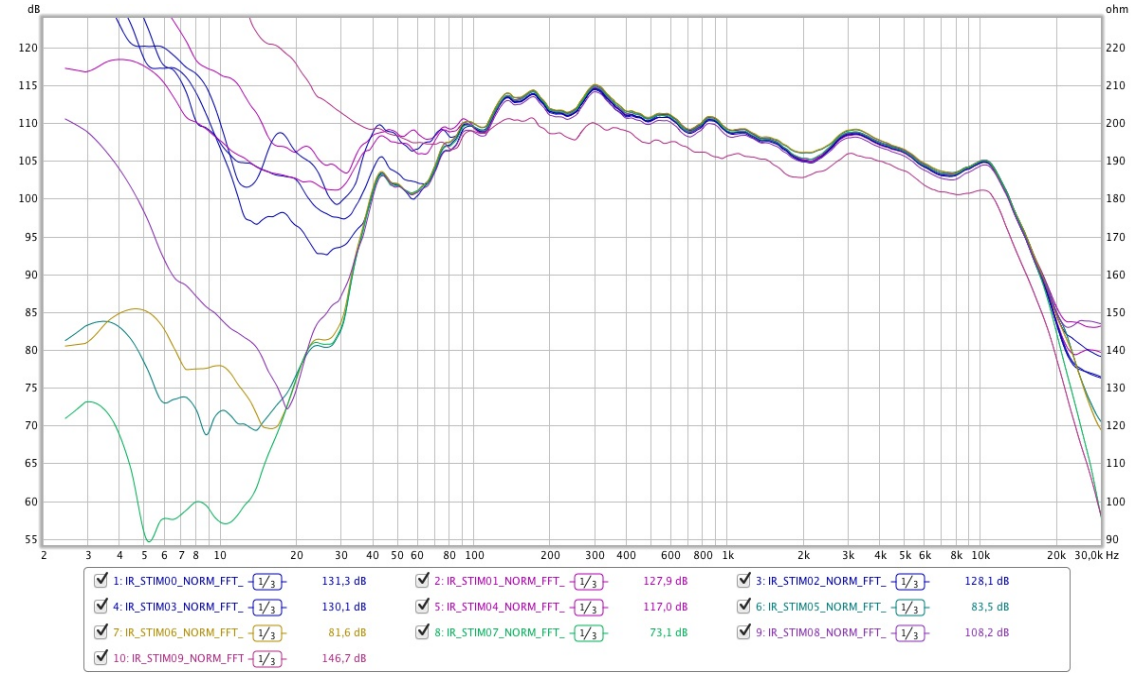
Micro M3 – position : A3C1 Dans l'axe – Lissage 1/3 oct

Moyenne des enregistrements caméra (Bleu) – enregistrement Micstasy (rouge)

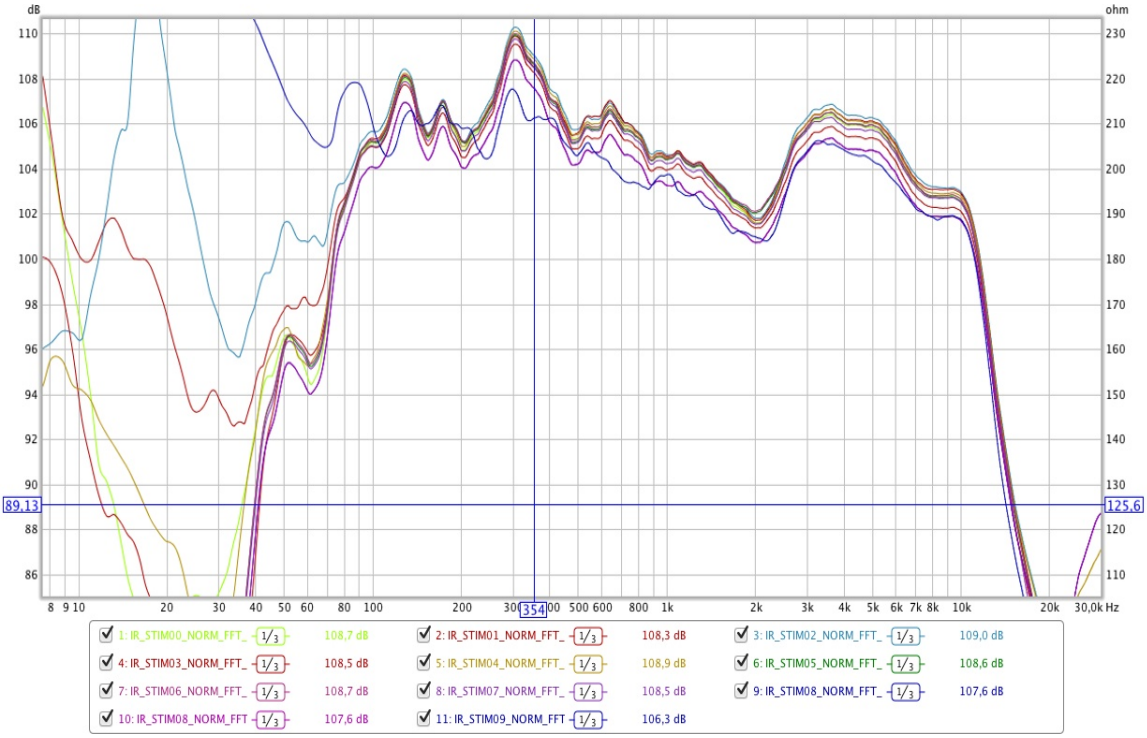


ANNEXE G: REPONSES FREQUENTIELLES DES RI – Hors Axe

Micro M2 – position : A3C1 - Hors axe – Lissage 1/3 oct



Micro M2 – position : A3C1 - Hors axe – Lissage 1/3 oct



ANNEXE H: Implémentation Matlab - déconvolution fréquentielle.

```
function ir = BruteDecon(stim,resp,OPT)
% ir = BruteDecon(stim,resp,OPT)    Deconvolve two signals using
simple FFT
%
% Deconvolve STIMulus from RESPonse of a system
% Created due to lack of success using DECONV
%
% OPT   'fast'  uses power of 2 FFT legnth (zero padding) rather
than true
%               length (seems to cause some minor variations in
IR (need to verify if
%               significant)
%
% Brian FG Katz
%
if nargin < 3,  OPT = 'slow'    ;   end

if OPT == 'fast'
    % find power of 2 just longer than signal (for the speed of
it)
    pfft = ceil(log(length(resp))/log(2))   ;
    nfft = 2^pfft                           ;
elseif OPT == 'slow'
    nfft = length(resp)    ;
else
    error('bad OPT string') ;
end

STIM = fft(stim,nfft)   ;
RESP = fft(resp,nfft)   ;

for loop = 1:size(resp,2),
    IR(:,loop) = RESP(:,loop)./STIM     ; % dont know the syntax
to do this without a loop
    IR(1,loop) = 0                      ; % Set DC component to 0
end % loop

ir = real(ifft(IR,nfft,1))              ;
%ir = real(ir(1:length(stim),:)) ; % remove excess length from
fft padding
%   IR returned is cropped to lenght of stimulus
%   Need to verify that this is not a bad idea

return
```

ANNEXE I-1/2: Implémentation Matlab - inversion de réponse impulsionnelle.

```
function inv=invimplms(den,n,d)
% syntax inv=invimplms(den,n,d)
%         den - denominator impulse
%         n   - length of result
%         d   - delay of result
%         inv - inverse impulse response of length n with delay d
%
% Levinson-Durbin algorithm from Proakis and Manolokis p.865
%
% Author: Bob Cain, May 1, 2001 arcane[AT]arcanemethods[DOT]com

    m=xcorr(den,n-1);
    m=m(n:end);
    b=[den(d+1:-1:1);zeros(n-d-1,1)];
    inv=toepsolve(m,b);



function quo=divimplms(num,den,n,d)
%Syntax quo=divimplms(num,den,n,d)
%       num - numerator impulse
%       den - denominator impulse
%       n   - length of result
%       d   - delay of result
%        quo - quotient impulse response of length n delayed by d
%
% Levinson-Durbin algorithm from Proakis and Manolokis p.865
%
% Author: Bob Cain, May 1, 2001 arcane@arcanemethods.com

    m=xcorr(den,n-1);
    m=m(n:end);
    b=xcorr([zeros(d,1);num],den,n-1);
    b=b(n:-1:1);
    quo=toepsolve(m,b);
```

ANNEXE I-2/2: Implémentation Matlab - inversion de réponse impulsionnelle.

```
function hinv=toepsolve(r,q)
% Solve Toeplitz system of equations.
%    Solves R*hinv = q, where R is the symmetric Toeplitz matrix
%    whos first column is r
%    Assumes all inputs are real
%    Inputs:
%       r - first column of Toeplitz matrix, length n
%       q - rhs vector, length n
%    Outputs:
%       hinv - length n solution
%
%   Algorithm from Roberts & Mullis, p.233
%
%   Author: T. Krauss, Sept 10, 1997
%
%   Modified: R. Cain, Dec 16, 2004 to remove a pair of
transposes
%             that caused errors.

    n=length(q);
    a=zeros(n+1,2);
    a(1,1)=1;

    hinv=zeros(n,1);
    hinv(1)=q(1)/r(1);

    alpha=r(1);
    c=1;
    d=2;

    for k=1:n-1,
       a(k+1,c)=0;
       a(1,d)=1;
       beta=0;
       j=1:k;
       beta=sum(r(k+2-j).*a(j,c))/alpha;
       a(j+1,d)=a(j+1,c)-beta*a(k+1-j,c);
       alpha=alpha*(1-beta^2);
       hinv(k+1,1)=(q(k+1)-sum(r(k+2-j).*hinv(j,1)))/alpha;
       hinv(j)=hinv(j)+a(k+2-j,d)*hinv(k+1);
       temp=c;
       c=d;
       d=temp;
    end
```

3.5 Autres approches, autres outils.

Suite aux résultats peu concluants de l'approche précédemment évoquée, nous portons notre attention sur les outils basés sur les principes évoqués dans de notre deuxième chapitre. Les procédés suivant ne nécessitent pas de connaissances particulières sur le canal de diffusion et son sensés apporter quelques réponses à notre problème. A travers la mise en œuvre de chacun de ces outils, nous tenterons d'évaluer la pertinences de ces approches à travers deux séquences : La scène du notaire évoquée précédemment ainsi qu'un dialogue entre Rémi (le protagoniste) et sa mère. L'ajout de cette deuxième séquence à notre étude permet de confronter ces outils à une autre situation : les personnages sont en mouvements et les voix sont de genre différents.

Notre objectif est de constituer un corpus des différents traitements sur ces deux séquences. Nous proposerons un niveau d'utilisation susceptible d'améliorer le rapport direct à réverbéré dans la limite des artefacts tolérables (si tant est qu'ils le soient). L'appréciation finale concernant l'utilisation de ces procédés sera laissée aux soins du mixeur et du réalisateur.

3.5.1 Mesure de l'efficacité des outils de dé-réverbération

Afin d'évaluer l'apport de chacun de ces procédés, nous souhaitons quantifier de manière objective l'efficacité des outils de dé-réverbération sans connaître précisément la réponse impulsionnelle du canal après traitement. Les outils mis en œuvre utilisent des processus non linéaires, qui ne permettent pas la détermination de cette nouvelle réponse impulsionnelle. La reconnaissance vocale peut être une piste intéressante car le taux d'erreur de mot diminue à mesure que le rapport direct à réverbéré augmente. Cependant nous avons été confronté à un problème pratique : trouver un automate de reconnaissance vocale en français donnant accès au taux erreur de mot.

Pour contourner ce problème il serait possible d'utiliser une autre approche, basée sur le rapport signal à réverbéré, connu sous l'acronyme NSRR (*Normalized Signal* to *Reverberated Ratio*).

$$\mathrm{NSRR} = 20\log_{10}\left(\frac{\|\mathbf{s}_\mathrm{d}\|_2}{\|(1/\hat{\alpha})\hat{\mathbf{s}} - \mathbf{s}_\mathrm{d}\|_2}\right)\mathrm{dB}$$

Fig 3-11 : *Définition de l'indice NSRR.*

sd représente le son direct, s le son traité. Ce dernier est normalisé pour s'affranchir des modifications du niveau induit par le traitement. L'implémentation du NSRR est décrite dans [23]. Malheureusement dans notre cas nous ne disposons pas du son direct, mais du son réverbéré. Deux possibilités s'offrent à nous.

La première consiste à détourner cet indice pour l'adapter à la situation. Comme il nous est impossible d'utiliser le son direct, nous pourrions prendre le son réverbéré comme signal de référence. Nous perdrions alors la correspondance physique avec d'autres critères tels que le Direct to Reverberant Ratio (DRR).

La deuxième méthode consiste à utiliser les sons diffusés lors de la capture d'impulse dans le lieu du tournage. Nous disposons dans ce cas du signal anéchoïc et d'un enregistrement réverbéré relativement proche de celui présent sur la bande son du film. Nous pourrions donc faire subir les mêmes traitements à ces enregistrements que ceux qui seront réalisés sur les séquences étudiées. Les voix utilisées ne sont cependant pas les mêmes et rien ne garantit la cohérence des mesures effectuées sur les sons du films et ceux captés lors de notre parte pratique.

La quantification objective des artefacts est elle aussi un problème de taille. Nous pourrions chercher à qualifier les performances des différents outils pour des niveaux d'utilisations déterminés (NSRR constant). Réverbérations et artéfacts sont tous deux corrélés au signal et variable dans le temps. Il est donc difficile de discriminer de manière objective les artéfacts d'éventuels résidus sur le signal dé-réverbéré.

Ces constatations nous amènent à envisager un test perceptif sur un panel d'auditeurs. Nous pourrions alors évaluer les performances de ces outils en utilisant uniquement les sons diffusés au moment de la capture de réponse impulsionnelle, en

plaçant chaque outil à un NSRR équivalent. Cela nous donnerait alors un avis sur les préférences de l'auditeur. Néanmoins Ces résultats restent peu généralisables car ces traitements visent à être intégrés dans le mixage final. Il est donc possible que certains artefacts soient fortement gênants lorsqu'ils sont entendus séparément mais deviennent masqués lorsqu'ils sont mélangés aux autres sources.

Voilà pourquoi nous nous conterons d'évoquer de manière subjective les effets des dé-réverbérateurs sur les séquences étudiées. Nous détaillerons l'effet des différents paramètres utilisateurs et leurs conséquences sur le signal afin de mieux appréhender les possibilités de chacun de ces outils.

3.5.2 Mise en œuvre de la prédiction linéaire

Nous utilisons ici le plug-in NML RevCon-RR, développé par la société TAC-SYSTEM. Il a été présenté pour la première fois en novembre 2010, lors de la 129ème conférence de l'Audio Engineering Society, à San Francisco. Il a été développé pour le format Real Time AudioSuite (RTAS).

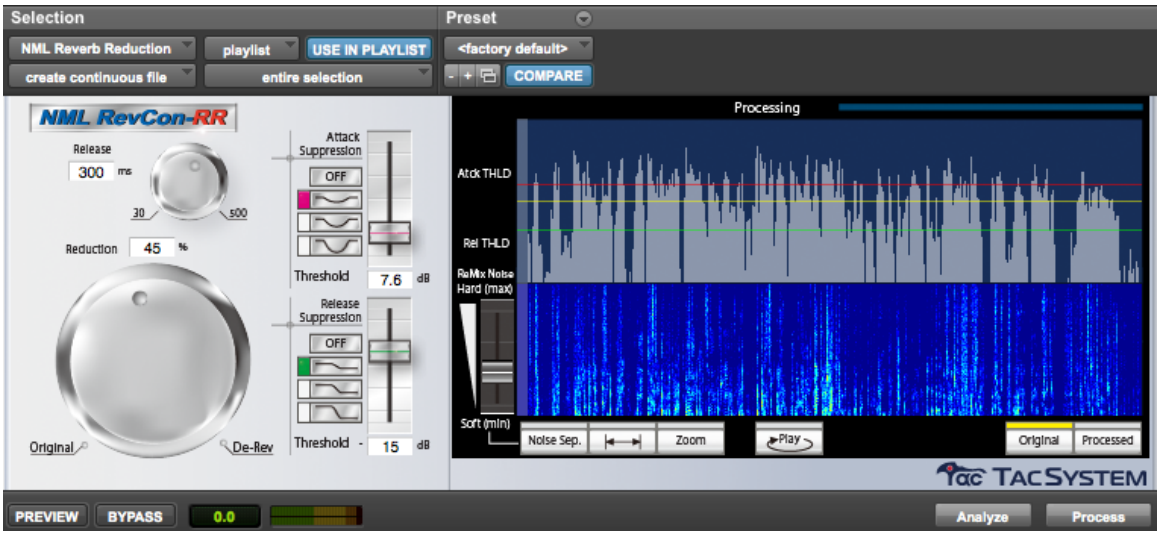


Fig 3-12 : *Interface graphique du plugin NML RevCon-RR.*

Dans la version qui est présentée, peu de paramètres sont laissées à l'utilisateur. Le traitement du champ diffus s'effectue par l'action cumulée des contrôles sur le pourcentage de réduction et le temps release. Un release important permet de traiter une décroissance longue. Cependant lorsque le rythme de la parole est rapide, on voit apparaitre des phénomènes de pompage à la manière d'un compresseur. Ces deux paramètres visent à optimiser la prédiction linéaire multi-pas, algorithme au centre de cet outil.

Comme nous l'avons évoqué dans la partie théorique, il n'est donc pas possible de traiter les réflexions précoces avec cette approche. Dans une première version du logiciel, la durée de ces premières réflexions était laissée à l'appréciation de l'utilisateur. Comme nous pouvons le voir sur cette capture d'écran, ce dernier a disparu pour être figé à une durée de 30 ms.

Afin d'améliorer la prédiction, il est préférable de disposer d'un signal peu bruité, c'est pourquoi cet outil intègre un réducteur de bruit fonctionnant par empreinte du signal stationnaire.

Dans cette dernière version, NML RevCon-RR tente de s'attaquer aux premières réflexions en proposant un traitement de l'enveloppe. Afin de traiter séparément l'attaque et la zone de maintient, trois types d'enveloppes sont disponibles pour des atténuations plus ou moins marquées. Deux curseurs déterminent les seuils d'activation de ces courbes d'expansion. Le but est d'accroitre le facteur de crêtes.

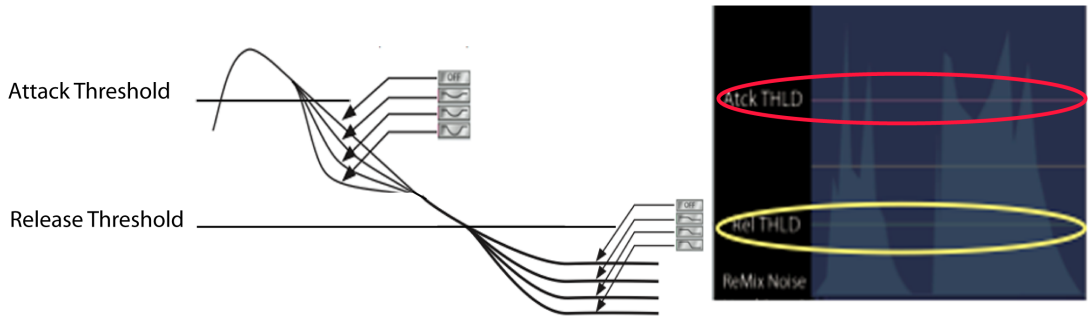


Fig. 3-13 : *Traitements des réflexions précoces dans NML_RevCon-RR*

Deux artefacts audibles découlent directement de ces paramètres. Lorsque l'expansion est trop marquée on entend une modulation d'amplitude appliquée au signal vocal.

Elle se caractérise par un vibrato dépendant du niveau de la voix. Ces seuils fixes imposent donc une faible dynamique du signal entrant.

Lors de l'application à la séquence du notaire nous ne sommes pas parvenus à traiter convenablement les réflexions précoces avec cet outil. Le dialogue entre deux personnages à la dynamique différente donne lieu aux artéfacts précédemment cités.

La suppression du champ diffus est en revanche assez efficace. Le rendu est alors peu naturel et si l'on cherche à accroitre le rapport de réduction, on entend l'effet s'activer. Le champ diffus est brutalement atténué après les premières réflexions qui elles restent intactes. On peu atténuer cet effets en diminuant le rapport direct à réverbéré. A l'inverse, si l'on augmente de façon extrême le rapport de réduction, on voit apparaitre du pompage combiné à des effets de phase comparable dans une certaine mesure au flanger...

Pour résumer, cet outil se prête mal à nos séquences. La décroissance du champ diffus (relativement courte), le rythme rapide de parole et ses variations dynamiques importantes mettent ce procédé à rude épreuve.

3.5.3 Travail sur l'enveloppe

Pour essayer cette approche nous avons utilisé le plug-in Transient Master de Native Instrument. Le fonctionnement de cet outil est semblable à celui décrit dans notre partie théorique. L'interface ci-dessous donne accès aux commandes des VCA contrôlant les zones d'attaque et de résonnance. Le commutateur « Smooth » permet d'allonger l'enveloppe d'attaque.

Fig. 3-14 : *Interface du Transient Master de Native Instrument.*

Cependant même après activation de cette fonctionnalité, l'utilisation de la commande d'attaque nous a paru peu appropriée. L'intervalle de temps, relativement court, concerné par ce traitement semble peu adapté aux signaux vocaux. Cette enveloppe fait émerger des plosives ou certains fragment syllabiques. Elle engendre de l'agressivité sans pour autant réduire (dans les cas étudiés) l'impression de réverbération. Son utilisation combinée à la réduction de la résonnance engendre une modulation de l'amplitude pouvant être très dommageable. Il est probable que ce paramètre trouve plus d'intérêt dans le traitement de percussions, d'instruments à corde pincées ou de sons de synthèse présentant des transitoires plus marquées que celles de la voix.

La commande de « Sustain » s'est pour sa part avérée d'une bonne efficacité. Contrairement à la prédiction linéaire, elle agit de manière continue sur le signal en réduisant considérablement la résonnance. Si le Transient Master ne permet pas d'aller aussi loin dans la suppression du champ diffus, on gagne cependant en homogénéité du rendu. Bien entendu, une utilisation poussée procure une détérioration du signal. Du fait de son placement temporel, ce traitement dynamique affecte les résonnances, donc les parties graves et médium du spectre.

D'autre part, le rythme de parole influe beaucoup sur le degré d'utilisation du traitement. Une réduction importante du sustain combinée à un rythme de parole trop rapide engendre une modification dynamique similaire à celle rencontrée avec NML

RevCon-RR. Il serait alors intéressent de développer un outil donnant accès aux constantes de temps des enveloppes en vue d'adapter le traitement à chaque situation.

Cet outil ne permet donc pas de supprimer complètement la réverbération mais peu, dans nos cas, diminuer son influence aux prix d'artefacts relativement modérés.

3.5.4 Utilisation du CEDAR DNS 3000

Comme l'indique sa dénomination (Dialogue Noise Suppressor), l'objectif de cet outil n'est pas de dé-réverbérer une source mais de supprimer le bruit de fond sur des séquences dialoguées. Cet appareil s'attaque donc aux bruits non corrélés à la voix et offre une grande puissance de traitement dans ces conditions. Qu'en est-il alors concernant la réverbération ? De nombreux praticiens évoquent cet outil comme une arme permettant de lutter contre des réverbérations trop marquées. Voyons alors quels sont les effets du DNS 3000 sur nos deux séquences.

Le DNS 3000 est un processeur de traitement hardware principalement utilisé dans le domaine de la postproduction sonore. Il est capable de traiter deux signaux numériques de manière simultanée. L'interface utilisateur, composée de sept faders. Six donnent accès à un découpage fréquentiel modulable, permettant d'ajuster la précision du traitement aux régions concernées. Le septième permet de doser l'intensité du traitement.

Fig 3-15 : *CEDAR DNS3000.*

Si l'on choisit de travailler en large bande, la totalité du spectre audible est affectées aux six tirettes. L'action sur ces dernières va établir une pondération du traitement. Lorsque les curseurs sont en dessous de la position nominale, on active une atténuation dépendant du signal dans les zones concernées. A l'inverse il est possible de faire émerger certaines composantes en travaillant au dessus de la position nominale.

Nous sommes malheureusement loin de connaître les procédés mis en œuvre dans le DNS3000. Cependant selon les constatations de Jean Rouchouse [24], il semblerait que l'architecture matérielle soit organisée autours d'un : « processeur à champ programmable pour la gestion des signaux d'horloge et la gestion des configurations ainsi que sur l'utilisation de processeurs de traitement de signal (dsp) pour une résolution des traitements en 40 bits flottants".

A l'écoute on pourrait comparer son effet à celui d'un expendeur multi-bandes à paramètres variables. De manière imagée, le taux d'expansion serait fonction de la position de chacun des six potentiomètres linéaire, les temps de déclenchement et de retour différents pour chaque bande de traitement.

L'Utilisation du CEDAR améliore considérablement l'intelligibilité de nos dialogues et son effet est d'autant plus marqué que les voix traitées sont d'une tessiture grave.

Cette réduction de bruit adaptée aux registres graves et médiums démasque les sources de manière efficace. Cependant on ne peut vraiment parler de dé-réverbération. Par le démasquage que nous venons d'évoquer, Le DNS3000 réduit la sensation de réverbération contrôlant l'influence des premiers modes propres de la salle (situés ici à 120Hz, 150Hz et 300Hz environ).

Passé 300 Hz il est plus difficile de contrôler les réflexions les plus précoces tout en préservant l'intégrité du signal. En revanche le Cedar peut avoir un rôle bénéfique dans le traitement des décroissances longues. L'effet est certes moins radical que celui obtenu par prédiction linéaire mais il est beaucoup plus progressif. La faible latence (10 ms) induit par l'architecture DSP est un atout certains puisqu'il permet de conserver le synchronisme du son à l'image. Il autorise alors une utilisation temps réelle et automatisable lors du mixage.

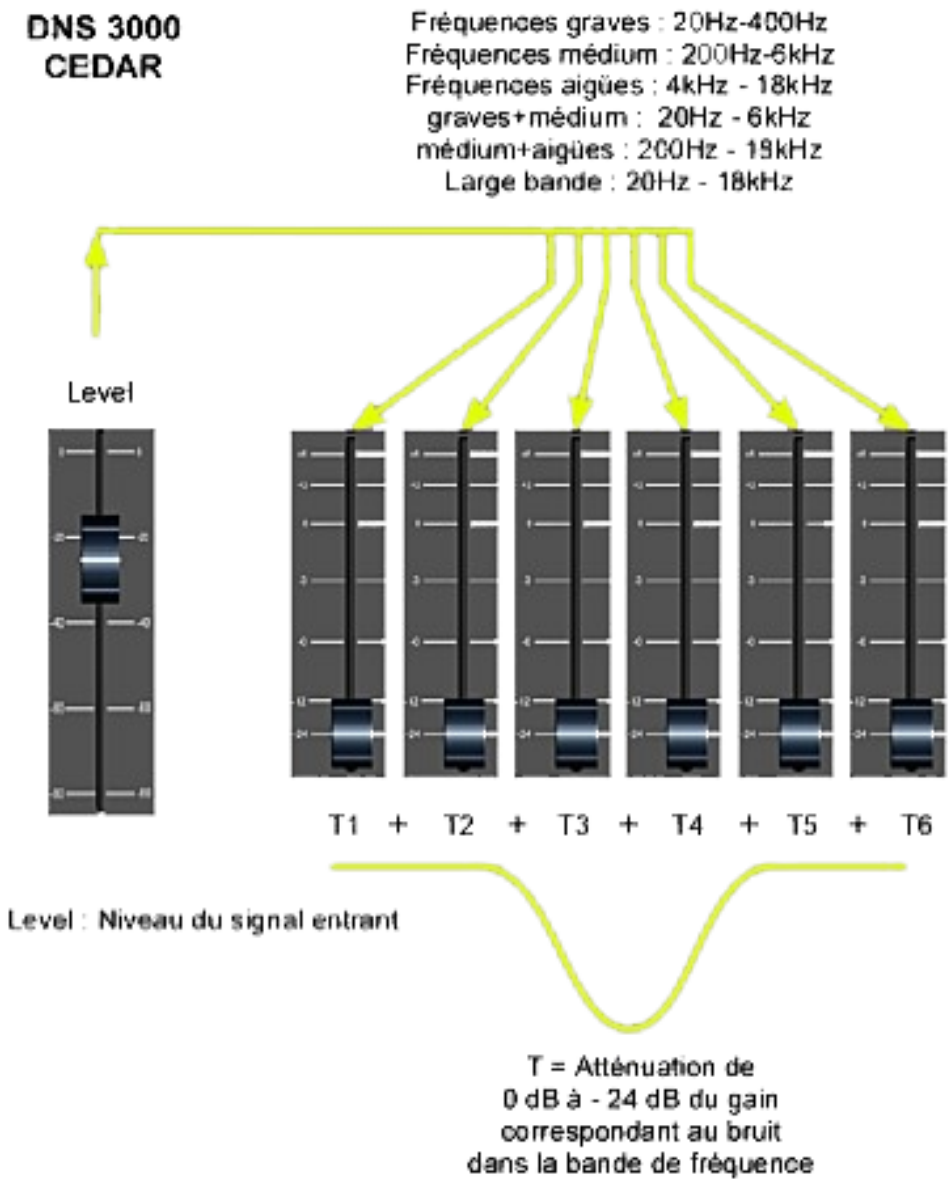


Fig 3-16 : *Principe de fonctionnement du DNS 3000 – d'après [24].*

3.5.5 Paramétrage de Unveil

Ce programme est multiplateformes (Macintosh et Windows). Depuis la version 1.6.0, il est disponible dans divers formats de plug-in (VST, RTAS, AU et AAX). S'il est capable de traiter de 1 à 8 canaux de manière simultanée, il est en revanche très gourmand en ressources de calcul. Il nécessite une architecture multiprocesseur ainsi qu'une optimisation des ressources pour bénéficier des meilleures conditions de

traitement. C'est la raison pour laquelle il requiert des mémoires tampons importantes. Dans le cas d'un traitement à l'image il faudra donc prendre garde à cette latence. Bon nombre de séquenceurs proposent une compensation de ces délais, Cependant il faudra veiller à conserver le synchronisme de la bande son avec l'image.

Il est préférable de réserver un ou plusieurs cœurs processeurs pour réduire le temps de calcul. C'est pourquoi dans le cas d'un traitement stéréo sous Protools, il est préférable de séparer les voies gauches et droites afin de partager ces deux traitements sur deux cœurs de processeur. Protools ne permet pas ce partage des ressources lorsqu'il s'agit d'une piste stéréo.

Si nous avons déjà proposé une explication pour décrire son fonctionnement dans le chapitre 2, nous n'avons pas vraiment détaillé les différents contrôles qui s'offraient à l'utilisateur. Pour aller plus loin dans cette démarche, voici un tour du propriétaire adapté des informations fournies par Denis H. Goekdag, co-fondateur de Zynaptiq :

- **Focus :** la position centrale n'engendre pas de modification du signal. En tournant ce contrôle rotatif dans le sens inverse des aiguilles d'une montre, on favorise le signal réverbéré au détriment du son direct. On agit de manière opposée, en dé-réverbérateur, lorsque l'on tourne ce contrôle dans le sens des aiguilles d'une montre.

- **Focus Bias :** à la manière du Cedar DNS 3000, il est possible de traiter de manière indépendantes les différentes bandes de fréquence. Lorsque le Focus est au maximum, augmenter le Bias n'aura aucun effet. Placer ces tirettes au minimum aura pour conséquence de ne pas traiter la bande concernée. Cela permet de préserver certaines régions sensibles du spectre ou au contraire de favoriser la suppression de fréquences prépondérantes.

- **t/f Localize :** nous l'avions décrit dans la partie théorique par analogie à la taille d'une fenêtre FFT. De faibles valeurs limitent les artefacts de dé-réverbération mais réduisent aussi l'efficacité du traitement. Dans notre cas, à

cause du temps de réverbération relativement court et du faible niveau d'énergie du champ direct, il est nécessaire d'augmenter ce paramètre pour s'attaquer aux réflexions précoces très liées à la source. Néanmoins un usage excessif entraine des modulations d'amplitude importantes ainsi que des phénomènes de pompage. Son utilisation est donc très liée au degré de dé-réverbération matérialisé par le Focus. Ainsi qu'au paramètre t Refract.

- **T Refract** : il caractérise le temps de réaction du réseau neuronal. De faibles valeurs engendrent une estimation plus fine du champ réverbéré mais une moins grande efficacité dans le traitement des réflexions précoces. Il pourrait être considéré comme le temps nécessaire au réseau pour estimer puis, pour modifier sa structure avant d'agir. Augmenter ce paramètre permet un meilleur traitement du champ diffus ainsi qu'une réduction du phénomène de pompage induit par de fortes valeurs de t/f localise.

- **T Adaptation :** laisse l'estimation de temps de réverbération (TR60 ?) à l'utilisateur. Comme l'indique [25], il est néanmoins possible grâce au réseau neuronal de déterminer ce paramètre. On regrette cependant qu'il ne soit pas fait mention d'une fenêtre numérique permettant d'ajuster ce paramètre de manière plus précise.

- **Presence :** en augmentant ce contrôle, on introduit un paramètre aléatoire dans la détection dans la configuration de détection. Cela contribue à diminuer les artéfacts audibles. L'objectif est ici de séparer fréquentiellement la réverbération du son direct. « *Presence* » va rendre la part non significative du signal plus grave et souligner la partie haute du signal utile. Accroitre la ce paramètre peut aussi atténuer la perception des artefacts induits par t/f Localize.

- **Transient threshold :** détermine le seuil de détection des transitoires de manière à les préserver des transformations futures. Cela permet des traitements plus poussés sur le reste du signal. Ce seuil est un paramètre qui prend en considération une dynamique statistique de l'entrée.

Pour optimiser le format de calcul, il est possible de réaliser une normalisation de l'entrée (commande NORM.) et de compenser le gain appliqué en sortie de manière à ne pas engendrer de modification dynamique (Pk link). Pour libérer des ressources on peut aussi choisir de ne pas afficher l'interface (Diagram). Un contrôle de monitoring nous permet d'écouter la part du signal supprimé (I/O Diff).

Nous avons utilisé la version VST de cet outil dans samplitude 11. Le traitement impose un délai de traitement de 4096 samples, soit 92 ms de décalage à 48 kHZ ou 3.68 images de décalages à 25 i/s. Ce décalage est donc perceptible. C'est pourquoi, comme dans le cas du NML_RevCon-RR il est préférable de créer un nouveau fichier à resynchroniser à la session d'origine.

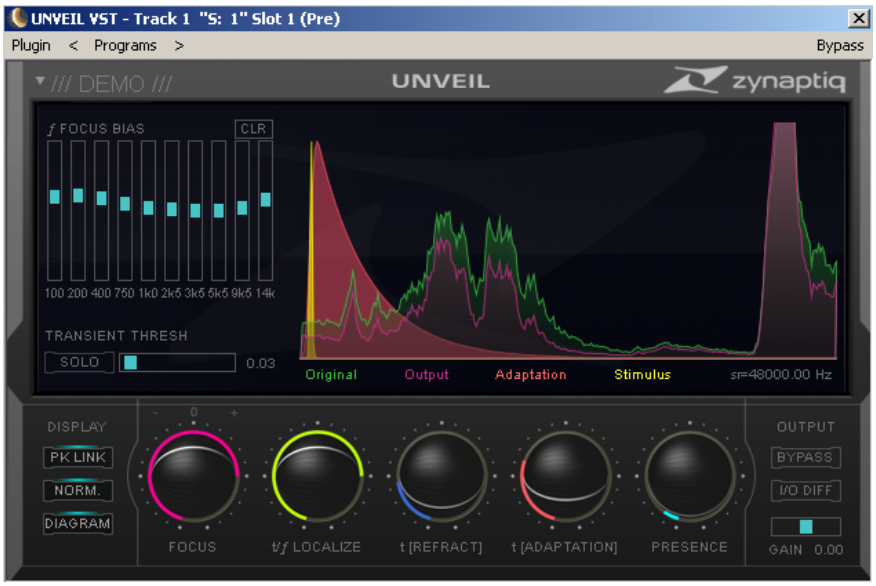


fig X : *Paramétrage de Unveil pour la séquence du notaire.*

Cet outil s'avère d'une efficacité remarquable. Bien qu'elle ne soit pas exempte d'artefact, l'approche mise en œuvre offre un traitement puissant et efficace. La suppression de la réverbération est là encore plus facile à obtenir sur le champ diffus mais Unveil permet de s'attaquer aussi aux réflexions précoces. Les traitements du diffus et du précoce s'effectuent de manière quasi continue. On ne sent donc pas la dé-réverbération entrer en action comme dans le cas du NML_RevConRR. Les modulations

dynamiques sont moins marquées que dans celle rencontrées lors de l'utilisation du Transient Master ou du DNS3000.

Toutefois on ne peut pas dire que cet outil puisse supprimer totalement la réverbération. Il permet simplement d'accroitre la marge de traitement disponible au mixage. Même s'ils sont moins importants que ceux rencontrés lors de l'expérimentation des autres méthodes, Unveil engendre nécessairement des artefacts. Tout l'enjeu réside alors dans l'intégration de ces artefacts dans l'ensemble de la scène sonore. Ce défi est de taille car les modulations, le phasing ou le léger flanging induit par ces outils est loin de nous paraître réaliste. Est-il alors préférable de réduire la réverbération si l'on apporte des artefacts de faibles niveaux mais avec un degré d'intégration plus mauvais ?

3.4 Conclusion de la mise en application

Ces procédés influent donc sur la manière d'envisager la bande son. On pourrait ainsi chercher à masquer ces artefacts par un montage son très présent et étudié de manière à satisfaire à la fois les impératifs de narration et le confort d'écoute. Dans le cas du *Premier Rôle* cet objectif semble difficile à atteindre.

Suites aux multiples difficultés de raccord rencontrées lors du montage son, Renaud Duguet, le monteur son du film, à pris le parti d'assumer le hors champ en recréant une vie en dehors du huit clos. Cette idée souligne aussi la position mimétique des acteurs lors du casting. Les acteurs jouent autours d'objets et de situations fantasmées. La fiction s'installe peu à peu autours d'un processus proche du documentaire. La bande son, elle aussi, évolue au fil de cette transformation. A mesure que le film bascule vers la fiction elle devient de plus en plus figurative, nous pourrions donc utiliser ces éléments pour tenter de masquer les défauts induits par nos traitements.

Malheureusement les scènes concernées par ces problèmes de réverbération sont dans les premières bobines du film. De nombreuses scènes ont été tournées de manière chronologique. Dans ces conditions il est difficile de jouer les ambiances et les bruitages au niveau où nous pourrions le faire à la fin du film. Cela romprait la progression sonore que l'on cherche à installer.

Notre problème est donc double. Si la quantité de réverbération semble effectivement moins importante après utilisation des procédés évoqués précédemment, les problèmes de détimbrage et le mauvais rapport signal à bruit demeurent. A cela s'ajoute les artefacts induits par nous outils. Si le gain en intelligibilité est effectif, les dialoguent manquent de consistance et de vraisemblance.

Pour redonner un semblant de réalisme à ces sons, tout en préservant l'intelligibilité, une piste consiste à envisager un mélange entre un signal fortement dé-réverbéré et celui d'origine en parfait synchronisme. Cette méthode inspirée de la compression parallèle vise à masquer les artefacts par la réverbération elle même. En effet les artefacts comme la réverbérations sont corrélés à la source. Pour que cette technique soit efficace, il faut évidemment que le gain en dé-réverbération soit important que les artefacts comme le bruit de fond soient limités...

Conclusion

Cette étude à la fois théorique et pratique fait entrevoir les limitations des outils de dé-réverbération. Les conditions de tournages donnant naissance au problème traité dans ce mémoire engendre souvent des défauts qui dépassent le simple cadre de la dé-réverbération. Le dé-timbrage et le mauvais rapport signal à bruit, caractéristiques de ce type de captations, demeurent même après dé-réverbération. En ce sens les outils que nous venons d'étudier accroissent la marge de traitement disponible en post-production. Ils ne permettent pas en revanche de re-synthétiser à postériori les conditions d'une captation idéale.

Nous avons envisagé la dé-convolution comme un moyen d'agir sur la réverbération et sur les dé-timbrages du système de captation. Nous sommes forcés de constater que les captures de réponses impulsionnelles réalisées sont loin de fournir les résultats escomptés. Malgré toutes nos précautions, les approximations réalisées sur la source et sa directivité ainsi que sur le positionnement du système, n'ont pas été d'une précision suffisantes pour répondre à nos besoins.
Nous nous sommes aussi heurtés aux limitations théoriques dans la détermination d'une réponse impulsionnelle inverse. La réponse en phase des réverbérations réelles sont la cause d'artefacts importants.
Un échantillonnage spatial élevé du champ acoustique autours de la source pourrait permettre de lever certaines de ces approximations. C'est en tout cas ce que semble suggérer les travaux de Mathias Fink sur le renversement temps acoustique en milieu complexe [26].

Il est néanmoins nécessaire de relativiser ces approches. Ces procédés multi-microphoniques engendrent un surcoût qui va à l'encontre du contexte production responsable des défauts liés à la réverbération. En effet, les problèmes de réverbération sont souvent le fruit d'un défaut de production. Il est évidemment possible et toujours plus efficace de traiter ces questions dès l'enregistrement.

Il faudra alors recourir à une captation de proximité en utilisant le matériel approprié, choisir un lieu adapté (ou le traiter acoustiquement), recruter du personnel qualifié et en nombre suffisant.

Malheureusement les impératifs budgétaires font que l'on assiste à une quête de rentabilité reposant sur le progrès technologique. Les nouveaux outils tendent à la convergence des métiers. Dans le monde du reportage, le journaliste devient cadreur et ingénieur du son ; dans celui de la postproduction sonore, le métier de monteur rencontre parfois celui du mixeur... Si nos professions évoluent aussi en fonction des nos outils, ces derniers ne sont pas infaillibles. Les traitements de dé-réverbération que nous venons d'étudier en sont un bel exemple. Ainsi l'idée de confier la responsabilité d'une prise de son à ces traitements est heureusement loin d'être réalisable. Ces outils apportent un traitement novateur mais leur usage doit s'accorder au niveau d'exigences imposé par le produit. On pourra certainement utiliser ces outils avec plus de facilité sur de courts programmes de flux ou dans des situations extrêmes pour lesquels la postsynchronisation n'est pas envisageable...

Cependant, un réel marché semble s'offrir à ces technologies. Cela fait entrevoir une certaine conception de la post-production sonore. Ces outils viendraient renforcer l'idée qui consiste à considérer la post production non pas comme un espace de création, mais comme le l'étape nécessaire pour pallier aux défauts du tournage. Si ces outils étaient amenés à se développer cela pourrait engendrer une autre remise en cause du métier d'ingénieur du son sur les tournages. Pour l'heure il n'en n'est rien. Même si les recherches à ce sujet sont encore très actives, de nombreuses limitations physiques s'y opposent.

Bibliographie

[1] R. Boite, H. Bourlard, T. Dutoit, J. Hanq, H. Leich, *Traitement de la Parole,* Presse polytechniques et universitaires romandes.

[2] J. R. McCarty, *Timbral Analysis,* https://ccrma.stanford.edu/~jmccarty/formant.html.

[3] A. H. Marshall, J. Meyer, *The Directivity and Auditory Impressions of Singers*, Acta Acustica united with Acustica, Volume 58, N°3, Aout 1985 p. 130 à 140.

[4] E. Ambikairajah, A. G. Davis and W. T. K. Wong, *Auditory Masking and and MPEG-1 audio compression.* ELECTRONICS 8.1 COMMUNICATION ENGINEERING JOURNAL, août 1997.

[5] Herman J.M. Steeneken TNO Human Factors, Soesterberg, *The Measurement of Speech Intelligibility*, the Netherlands.

[6] Guillaume Couturier, *La Réverbération à Convolution - Tentative de simulation acoustique en auditorium de mixage cinema -,* ENS Louis Lumière 2007.

[7] Prem Seetharaman1, Stephen P. Tarzia*, The Hand Clap as an Impulse Source for Measuring Room Acoustics.*

[8] P. Zahorik, *Auditory Display of Sound Source Distance,* Proceedings of the 2002 International Conference on Auditory Display, Kyoto, Japan, July 2-5, 2002.

[9] P. Zahorik*, Assessing auditory distance perception using virtual acoustics*, J. Acoust. Soc. Am. Volume 111, Issue 4, pp. 1832-1846 (2002).

[10] Dolby Laboratories, *5.1 Channel Music Production Guidelines*, Dolby, 2005.

[11] *SPL Transient Designer Model 9842 manual.*

[12] Emanuel Anco Peter Habets, *Single- and Multi-Microphone Speech Dereverberation using Spectral Enhancement.*

[13] K. Kinoshita and T. Nakatani, *Speech Dereverberation Using Linear Prediction.*

[14] BMC Audio Work, http://bcmaudioworks.pbworks.com/f/derevOrc.wav.

[15] John Usher, *Acoustic impulse response measurement using speech and music signals.*

[16] Application Note Polycom Installed Voice Business Group, *How to Choose an Acoustic Echo Canceller.*

[17] S. Nelly et J. Allen, *Invertibility of a room impulse response,* J. Acoust. Soc. Am. Volume 66, Issue 1, pp. 165-169 (1979).

[18] Manuel utilisateur, JBL LSR4326P, http://www.jblpro.com/BackOffice/ProductAttachments/JBL.LSR4326P.v4[1].pdf.

[19] Manuel utilisateur, TANNOY system 600, http://www.tannoy.com/products/158/uman_System600.pdf.

[20] Manuel utilisateur, AKG C-451, http://www.akg.com.

[21] Documentation Technique, DPA 4006 TL, http://www.dpamicrophones.com/en/products.aspx?c=item&category=191&item=24287#description.

[22] J. Lescure, *Simulation temps réel de prise de son multicanale,* ENS Louis Lumière 2011.

[23] Patrick A. Naylor, Nikolay D. Gaubitch, and Emanuel A. P. Habets, Signal-Based Performance Evaluation of Dereverberation Algorithms, Journal of Electrical and Computer EngineeringVolume 2010 (2010), Article ID 127513.

[24] Jean Rouchouse, http://restauration-sonore.over-blog.com/article-le-dns-3000-de-cedar-59632088.html.

[25] Cox, Trevor, J. Li, Francis, Darlington, Paul, *Extracting Room Reverberation Time from Speech Using Artificial Neural Networks,* JAES Volume 49 Issue 4 pp. 219-230; April 2001.

[26] M. Fink, C. Prada, J.L. Thomas, *Le renversement du temps en acoustique*, CNRS, http://www.cnrs.fr/publications/imagesdelaphysique/couv-PDF/imagesphys9596/40-48.pdf.