THÈSE

présentée pour obtenir le titre de Docteur
de l'Ecole Nationale Supérieure
des Télécommunications

Spécialité : Signal et Images

Olivier CAPPÉ

Techniques de réduction de bruit pour la restauration d'enregistrements musicaux.

TELECOM Paris

Soutenue le 27 septembre 1993 devant le jury composé de

Jean-Pierre Tubach	Président
Antoine Chaigne	Examinateurs
Philippe Depalle	
Pierre Duhamel	
Gérard Faucon	Rapporteurs
Jean-Christophe Valière	

Résumé

Depuis une dizaine d'années, l'application de techniques de traitement numérique de signal à la *restauration d'enregistrements musicaux dégradés* a permis d'obtenir des résultats très spectaculaires. Cependant, certaines questions restent ouvertes. En particulier, plusieurs résultats antérieurs indiquent que l'élimination du *bruit de fond* n'est souvent obtenue qu'au prix d'une certaine *distorsion du signal musical*. Actuellement, la réduction du niveau du bruit présent sur les enregistrements musicaux est effectuée grâce à des techniques d'*atténuation spectrale à court-terme*. Le principe du traitement consiste à modifier la transformée à court-terme du signal bruité en utilisant un algorithme de décision désigné par le terme de *règle de suppression de bruit*.

Dans la première partie de l'étude, le fonctionnement de l'atténuation spectrale à court-terme est analysé en considérant d'une part un modèle simplifié des règles de suppression utilisées en pratique, et d'autre part, des signaux-test simples. Les résultats obtenus fournissent des éléments permettant de *caractériser objectivement* les modifications apportées au signal musical lors du traitement. De plus, l'utilisation de résultats classiques de la psychoacoustique permet de prévoir sous quelles conditions les distorsions produites peuvent devenir *audibles*. Les principaux phénomènes mis en évidence sont la modification du timbre, l'apparition d'effets de modulation, ainsi que le lissage des transitoires brusques.

A partir de ces résultats, une nouvelle technique de réduction de bruit de fond, adaptée au cas des enregistrements musicaux, est décrite. Le signal est tout d'abord analysé par une transformée à court-terme de résolution fréquentielle moyenne, puis, les signaux de sous-bande sont traités de deux manières différentes selon leurs caractéristiques locales : lorsque la présence de composantes stationnaires est détectée, la restauration du signal de sous-bande s'effectue par blocs, tandis que dans le cas contraire, celui-ci est traité ponctuellement. Cette *restauration sélective des signaux de sous-bande*, guidée par une procédure de détection, permet d'envisager une amélioration significative par rapport aux résultats des techniques classiques.

Abstract

Noise reduction techniques for the restoration of musical recordings

Recently, very impressive results have been obtained by applying digital signal processing techniques to the restoration of degraded musical recordings. However, some questions remain. In particular, previous results have pointed out that the noise reduction is often obtained at the cost of an alteration of the recorded signal. At present, musical recordings are restored with the help of short-time spectral attenuation techniques. Noise reduction is obtained by modifying the short-time transform of the noisy signal, the spectral modification being determined by a so-called noise suppression rule.

The first part of this dissertation is devoted to the evaluation of short-time spectral attenuation techniques. This evaluation is done with a simplified model of standard noise suppression rules and with elementary test-signals. Signal distortions induced by the restoration process are first evaluated analytically, then their audibility is assessed by use of classic psychoacoustics results. This part of the study brings to light several phenomena observed experimentally in previous studies, such as the modification of timbre, the appearance of modulations and the spreading of transients.

Drawing from these results, we describe a new noise reduction technique intended for enhancing musical signals. In the first step, the noisy signal is analyzed by use of a medium frequency-resolution short-time transform. The restoration then takes place in each sub-band in two different ways according to the nature of the sub-band signal: the processing is carried out block by block when steady signal components are detected, or locally otherwise. This approach has been successfully applied to several musical recordings yielding promising results.

Table des Matières

Liste des Figures

Notations

Divers

- z^* complexe conjugué.
- $\mathrm{Re}\{z\}$, $\mathrm{Im}\{z\}$ partie réelle, partie imaginaire.
- $\mathrm{E}\{X\}$ espérance mathématique de X.
- $\hat{\theta}$ estimation du paramètre θ.
- $\mathbf{X}$ vecteur colonne ou matrice (les deux en gras).
- $\mathbf{X}^T$ transposé, $\mathbf{X}^H$ transposé conjugué.

Signal discret

- Les lettres l, n, m, p, q désignent des indices temporels, k et r des indices fréquentiels.
- $x * y(n)$ convolution.
- $R_{xx}(n)$ fonction d'autocorrélation.
- $P_x(\nu)$ densité spectrale de puissance (ν fréquence réduite $\nu \in]-\frac{1}{2}, \frac{1}{2}]$). On écrit indifféremment $P_x(\omega)$, (ω pulsation normalisée $\omega = 2\pi\nu$).
- $X(\nu)$ (ou $X(\omega)$) transformée de Fourier.
- $X(\omega_k)$ transformée de Fourier discrète (TFD)

  $$X(\omega_k) = \sum_{n=0}^{N-1} x(n)\,\mathrm{e}^{-\mathrm{j}\omega_k n} \quad (\omega_k = 2\pi k/N)$$

  La notation ω_k désigne les N pulsations discrètes de la TFD. On utilise aussi la notation $W_N = \exp(\mathrm{j}\frac{2\pi}{N})$ pour la TFD :

  $$X(\omega_k) = \sum_{n=0}^{N-1} x(n) W_N^{-kn}$$
- $\mathcal{L}(x)$ longueur du signal $x(n)$ (support fini).

Echantillonnage

- F_e fréquence d'échantillonnage.
- $x^{(a)}(t)$ signal analogique correspondant au signal échantillonné $x(n)$.
- $P_x^{(a)}(f)$ densité spectrale de puissance (f fréquence en Hz).

Transformée de Fourier à court-terme (TFCT)

Les notations utilisées pour la transformée de Fourier à court-terme sont présentées au début de l'annexe B. Les principales notations sont :

- $h(n)$ fenêtre d'analyse, $f(n)$ fenêtre de synthèse.
- N durée de la fenêtre, R pas de décalage des fenêtres.
- $X(n, \omega_k)$ TFCT (avec $\omega_k = 2\pi k/N$).

Abréviations

- AR autorégressif.
- DSP densité spectrale de puissance.
- FFT *fast Fourier transform.*
- RSB Rapport signal-à-bruit.
- TFD Transformée de Fourier discrète.
- TFCT Transformée de Fourier à court-terme.

Guide de lecture

Le premier chapitre tient lieu d'introduction générale pour le reste du document. Ce chapitre permet notamment de situer les problèmes abordés par rapport au cadre plus général de la restauration d'enregistrements musicaux (paragraphes 1.1 et 1.2). C'est aussi l'occasion de préciser quels sont les objectifs de la présente étude (paragraphe 1.3). Enfin, ce premier chapitre se conclut par une description approfondie de la dégradation qui est traitée dans toute la suite du document : *le bruit de fond* (paragraphe 1.4). Cette dernière partie permet de mettre en évidence certaines caractéristiques particulières du bruit présent sur les enregistrements anciens (niveau, répartition fréquentielle) ainsi que de revenir sur l'hypothèse classique de stationnarité du bruit.

Les deux chapitres suivants (2 et 3) peuvent être considérés comme une première partie du document consacrée à l'étude des techniques classiques de réduction de bruit de fond fonctionnant selon le principe d'*atténuation spectrale à court-terme*. Le chapitre 2 constitue une présentation, surtout bibliographique, de ces techniques. Le paragraphe 2.3 fait le point sur l'effet des modifications de la transformée de Fourier à court-terme. La présentation relativement originale adoptée dans ce paragraphe 2.3 permet d'obtenir une caractérisation simple de l'effet du débruitage. Le chapitre 3 est consacré à l'analyse des résultats obtenus grâce aux techniques d'atténuation spectrale à court-terme. Cette partie à la fois expérimentale et théorique regroupe la plupart des résultats nouveaux concernant ces techniques classiques de débruitage. Afin de structurer l'analyse, l'effet du traitement de débruitage sur le signal musical est décomposé en trois phénomènes distincts qui correspondent respectivement aux paragraphes 3.1, 3.2 et 3.3. Ce chapitre 3 étant de loin le plus long, il comporte un paragraphe final consacré à la récapitulation des résultats obtenus (paragraphe 3.4) qui conclut la première partie.

Le chapitre 4 constitue la seconde partie du document essentiellement consacrée à la description d'une solution originale, adaptée au cas des enregistrements musicaux, et fondée sur les conclusions du chapitre 3. Le premier paragraphe de ce chapitre (4.1) apporte des éléments concernant le bruit résiduel (c'est à dire, le bruit qui demeure après traitement). En pratique, cet aspect est très important quant à la qualité du résultat mais il n'est évoqué que brièvement ici car les solutions envisagées ne sont pas originales. Le paragraphe 4.2 détaille les différentes étapes de la mise au point de la technique de débruitage par *restauration sélective des signaux de sous-bande*. Cette technique constitue une solution originale visant à réduire les distorsions apportées au signal musical par le traitement de débruitage. Les conclusions de cette partie de l'étude sont présentées au paragraphe 4.2.6.

Chacune des différentes annexes situées en fin de document est associée au chapitre correspondant : l'annexe A avec le chapitre 1, l'annexe B avec le chapitre 2, etc. L'annexe A présente un travail, essentiellement bibliographique, sur une cause de dégradation, très fréquente sur les enregistrements anciens, qui n'est pas spécifiquement étudiée dans le document : les bruits de type impulsionnels. L'annexe B regroupe certains aspects, plus classiques, du point développé au paragraphe 2.3 (modifications de la transformée de Fourier à court-terme). L'annexe C correspond aux détails d'un calcul, relativement fastidieux, dont le résultat est utilisé plusieurs fois au cours des chapitres 3 et 4. Enfin, l'annexe D reprend le texte d'un article (en anglais) qui complète le paragraphe 4.1.2 en précisant la manière dont une technique particulière (règle de suppression de bruit dite d'Ephraim et Malah) permet de contrôler la nature et le niveau du bruit résiduel.

Chapitre 1

Objectifs

La restauration d'enregistrements musicaux est un domaine particulier où, bien que les opérations mises en jeu soient complexes du point de vue scientifique, l'ensemble de la démarche est perçue de manière très intuitive par tout un chacun. Nous avons eu plusieurs fois l'occasion de présenter nos travaux devant un ensemble d'auditeurs relativement varié. Il s'est avéré que ces présentations se déroulaient quasiment toujours de la même façon. En général, la partie technique de la présentation s'effectue dans un silence quasi-religieux où personne n'ose contester la vérité scientifique. Par contre, lors de l'écoute des résultats de traitement, il est rare qu'au moins un des auditeurs ne vienne critiquer les résultats obtenus, mettre en doute la validité de la démarche, où même, saisir l'occasion pour déplorer la suprématie du disque laser ! La discussion qui s'en suit dérive toujours sur des sujets de plus en plus éloignés des aspects techniques de la restauration, pour venir se fixer sur des problèmes musicaux, historiques voir économiques ou philosophiques ("restaurer ou ne pas restaurer ...").

Ces préoccupations, même si certaines sont assez irrationnelles (par exemple, le fameux "son numérique"), indiquent que la problématique de la restauration dépasse largement le cadre des techniques habituelles de l'ingénieur. Une question extrêmement intéressante serait, par exemple, de savoir ce que l'auditeur moyen attend d'un enregistrement restauré. Est-ce la fidélité à l'enregistrement initial ou le confort d'écoute maximal ? Cette question, très importante en pratique pour les utilisateurs potentiels des systèmes de restauration, devrait être abordée au travers d'une démarche expérimentale de type psychoacoustique. Il y a même fort à parier que les résultats d'une telle étude ne pourraient pas être interprétés sans faire appel à des considérations d'ordre sociologique ou musical.

De la même façon, la variété des problèmes posés se traduit en pratique par plusieurs approches différentes de la restauration. Ainsi, dans un article consacré à la restauration d'enregistrements anciens [Schuller 91], au titre explicite de *The ethics of preservation, restoration, and re-issues of historical sound recordings*, Dietrich Schüller distingue sept étapes successives dans un processus complet de restauration. Il est intéressant de constater que les techniques de traitement de signal appliquées à la restauration d'enregistrements ne correspondent qu'à une seule de ces sept étapes. En conséquence, il existe plusieurs manières différentes d'aborder les problèmes posés par la restauration d'enregistrements musicaux, dont la plupart dépassent largement le cadre fixé ici.

Afin de situer notre étude par rapport à l'ensemble de ces préoccupations, ce premier chapitre est consacré aux objectifs que nous nous sommes fixés. Les paragraphes 1.1, 1.2 et 1.3 précisent notamment le type de restauration envisagée, la nature des défauts à éliminer, les objectifs scientifiques visés ainsi que la démarche suivie pour évaluer les résultats. La partie centrale de cette présentation est le paragraphe 1.3 qui justifie la démarche suivie tout au long du document.

Enfin, le dernier paragraphe (1.4) est consacré à une étude approfondie des propriétés du bruit d'enregistrement, à partir des exemples dont nous disposions (qui sont décrits au paragraphe 1.4.1). Cette étude fournit une occasion de vérifier plusieurs hypothèses utilisées dans la suite du document (par exemple, la stationnarité du bruit de fond), ainsi que d'apporter quelques éléments sur certains points peu traités dans la littérature (caractéristiques spectrales, niveau perçu du bruit).

1.1 Restauration objective et fondée sur le signal

1.1.1 Classification des opérations de restauration

La restauration complète d'un enregistrement ancien dégradé implique des opérations de nature très différentes. Ce paragraphe propose une façon de classifier ces différentes étapes du traitement qui s'inspire en partie de la référence [Schuller 91]. Tout d'abord, les divers traitements effectués peuvent être classés en deux catégories selon leur motivation en :

1. Traitements arbitraires
2. Traitements objectifs

Pour le premier type de traitements, c'est uniquement le jugement de l'opérateur du système de restauration qui guide les opérations. Un exemple de cette catégorie de traitement est la pseudo-stéréophonie (obtention d'un signal sur deux canaux à partir d'un enregistrement original monophonique). Dans un tel exemple, il ne s'agit pas de lutter contre un défaut de l'enregistrement, mais plutôt de modifier volontairement certaines caractéristiques de celui-ci. La justification de ce type d'opérations fait en général intervenir des considérations artistiques ainsi que des préoccupations relatives à l'attente des auditeurs potentiels. D'une manière générale, cette catégorie de traitements recouvre l'ensemble des manipulations qui sont habituellement effectuées par l'ingénieur du son. Les techniques mises en œuvre, ne sont pas spécifiques au problème de restauration et correspondent à des manipulations courantes dans le domaine de l'audio (filtrage, compression de dynamique, etc.).

Nous avons choisi de ne pas nous intéresser à ce premier type de traitements, et de ne considérer exclusivement que la seconde catégorie concernant les opérations motivées par des constatations objectives. La restauration d'enregistrements est donc restreinte à la compensation des défauts mesurables, ou tout au moins, caractérisables par une procédure objective. A l'intérieur de cette seconde catégorie, il est encore nécessaire de distinguer deux types d'opérations selon le type de procédure qui permet de caractériser les défauts à traiter :

- **2.a)** Utilisation de connaissances externes à l'enregistrement
- **2.b)** Caractérisation des défauts uniquement à partir de l'enregistrement

Dans le cas des opérations de type **(2.a)**, la connaissance mise en jeu peut être de nature historique comme dans le cas de la recherche de la vitesse de rotation appropriée à un type de disque particulier [Brock 84], musicologique pour le choix de l'enregistrement à restaurer, voire physique quand la conception d'un système de lecture adapté s'avère nécessaire [Meulengracht 76] [Owen 83] [Fesler 83]. Cette catégorie recouvre aussi tout un ensemble de méthodes de restauration fondées sur la modélisation physique des dégradations. Parmi les exemples fournis par la littérature, on peut citer [Fesler 83] qui décrit la reconstitution des conditions originales d'enregistrement, pour un type donné de support. L'intérêt de cette démarche est qu'elle permet d'obtenir des informations sur la nature des défauts présent sur les enregistrements à restaurer. Pour la même raison, l'étude des propriétés du support de l'enregistrement (voir les exemples présentés dans [Roys 78]) peut aussi fournir des renseignements très importants.

Le recours à des informations extérieures est souvent une étape indispensable dans le processus de restauration, dans la mesure ou certaines dégradations ne peuvent pas être caractérisées à partir du signal. C'est en particulier le cas pour toutes les "altérations volontaires" du signal [Schuller 91], c'est à dire les modifications apportées lors de l'enregistrement et destinées à être compensées lors de la restitution. Un exemple simple d'altération volontaire est le standard RIAA qui spécifie la réponse fréquentielle (non plate !) des systèmes de lecture des disques microsillons. Il existe de nombreux standards similaires concernant la lecture des disques 78 tours [Schuller 91]. Dans un cas comme celui-ci, il est extrêmement difficile de faire la part entre les effets de la courbe de réponse non plate, voulue comme telle, et les imperfections éventuelles du système d'enregistrement. Dès lors, le recours à des connaissances sur le système ayant servi à l'enregistrement s'impose. De la même manière, les incertitudes sur la vitesse de rotation ne peuvent, dans certains cas, être résolues que par une recherche historique, car les spéculations sur le diapason sont parfois trompeuses [Brock 84]. Notons aussi que pour mener à bien ce type d'opérations, il est nécessaire de pouvoir disposer de données concernant les enregistrements à traiter. Typiquement, les structures à même de mener ce genre de recherches sont, par exemple, les phonothèques qui possèdent, en principe, des renseignements techniques associés aux archives sonores.

1.1.2 Approche retenue

Nous avons choisi de ne considérer que la dernière catégorie **(2.b)** d'opérations : celles qui ne nécessitent que la connaissance du signal. Un des intérêts de cette approche est qu'elle est extrêmement générale : elle peut a priori s'appliquer à n'importe quel type d'enregistrement sans considération de provenance. Il s'agit, à partir du résultat du transfert d'un enregistrement dans un format numérique adéquat, de caractériser les dégradations présentes sur le signal, puis de mettre en œuvre toutes les techniques de traitement visant à réduire ces défauts. En conséquence, les techniques utilisées ne sont donc pas spécifiquement destinées à un type particulier d'enregistrement.

Toujours pour des considérations de généralité, nous nous sommes cantonnés à la situation la plus fréquente dans le domaine de la restauration d'enregistrements anciens, c'est à dire au cas où l'on ne dispose que d'un seul enregistrement. La possibilité de disposer de plusieurs versions comparables d'un même enregistrement permettrait d'élargir le champ des techniques applicables (voir par exemple [Lim 83]). Pour les lecteurs intéressés par ce point, les références [Vaseghi 88b] et [Vaseghi 89] proposent un exemple intéressant de restauration dans le cas particulier où l'on dispose de deux versions du même enregistrement. Cependant, de l'avis des interlocuteurs qui nous ont fourni des enregistrements (cf. paragraphe 1.4.1), ce cas n'est pas très fréquent. Outre la question de l'utilité pratique d'un traitement spécifiquement adapté à ce cas, ce manque de cas représentatifs rend très difficile l'évaluation d'une telle technique. D'une manière plus générale, la mise au point d'une technique destinée à une situation spécifique requiert toujours un ensemble d'enregistrements suffisamment large et cohérent afin de valider les résultats. Par opposition, la généralité des techniques considérées ici permet de valider les résultats avec des signaux synthétiques très simples, mais néanmoins représentatifs du comportement en situation réelle (cf. chapitre 3).

Le désavantage majeur de ce type de restauration, fondée uniquement sur le signal enregistré, a déjà été cité au paragraphe précédent : certaines dégradations présentes sur les enregistrements ne pouvant pas être caractérisées à partir du seul signal, il est impossible de les traiter par de telles techniques.

1.2 Etat des lieux

1.2.1 Inventaire des dégradations

Le premier exemple de restauration par des techniques de traitement numérique de signal concernait la distorsion due au système d'enregistrement (pour disque datant de 1907), modélisée sous la forme d'un filtrage linéaire [Stockham 75]. Toutefois, le traitement proposé n'entre pas exactement dans la catégorie considérée car il suppose connue une caractéristique du signal avant distorsion. Le terme de déconvolution aveugle est d'ailleurs utilisé de manière un peu abusive, puisque la déconvolution n'est effectuée que lorsque la fonction de transfert du pavillon a été déterminée de manière certaine. Dans [Stockham 75], c'est le "spectre moyen" du signal non-distordu qui est déterminée par une procédure faisant appel à une version moderne de l'enregistrement à restaurer. La procédure proposée est intéressante mais elle repose sur l'hypothèse que deux enregistrements d'une même partition musicale possèdent, en moyenne, les mêmes caractéristiques statistiques. Compte tenu, entre autres, de l'influence des interprètes, de l'effet de salle ou même des propriétés de rayonnement des sources sonores, cette hypothèse semble un peu "optimiste". Le point important ici est que, même avec une modélisation simple de l'effet du pavillon sous la forme d'un filtrage linéaire, il est nécessaire de recourir à des informations externes à l'enregistrement pour caractériser le défaut.

Dans l'exemple précédent, l'obstacle n'est donc pas le traitement à effectuer (déconvolution) mais plutôt la caractérisation du défaut (fonction de transfert du pavillon) à partir du signal. Il en va de même en ce qui concerne les distorsions non-linéaires subies par le signal : dans certains cas, il existe des techniques permettant de corriger ces défauts, par contre on ne dispose pratiquement pas de moyen de les caractériser. Il est, par exemple, possible d'interpoler localement le signal lorsqu'une distorsion non-linéaire survient, du moment que celle-ci n'est pas trop longue (voir l'exemple présenté dans [Godsill 93]). Cette approche pourrait permettre de résoudre les problèmes de saturation très fréquemment rencontrés sur les enregistrements. Cependant, il faudrait mettre au point une technique automatique de détection de la saturation. Dans le cas contraire, c'est un opérateur qui se voit contraint d'effectuer un travail extrêmement fastidieux de marquage. Le problème se pose d'ailleurs exactement dans les mêmes termes pour le traitement des défauts localisés (cf. annexe A). La situation est par contre très différente lorsqu'un argument objectif permet de caractériser les zones de signal distordues comme dans [Abel 91]. Toutefois, la situation écrite par [Abel 91] ne saurait être transposée simplement au cas de la saturation d'un dispositif analogique. De la même manière, le problème du pleurage (distorsion de fréquence) peut sûrement être éliminé grâce à des techniques connues de changement de hauteur (on en trouvera plusieurs dans [Laroche 93a]). Par contre, on ne dispose pas d'une procédure, autre que l'écoute, permettant de distinguer un effet de pleurage d'un simple vibrato appartenant au signal musical. Enfin, la même remarque reste valable pour [Preis 84] qui propose une technique simple pour compenser la non-linéarité d'un système d'enregistrement. Là encore, le traitement n'est possible que si la caractéristique non-linéaire de l'appareil d'enregistrement est connue, c'est à dire, difficilement applicable lorsqu'on ne dispose que du signal enregistré.

Classiquement, les dégradations qu'il est possible de caractériser uniquement à partir de l'enregistrement sont les défauts localisés, ainsi que le bruit de fond [Valiere 91] [Vaseghi 88b]. Les défauts localisés correspondent à des altérations très brèves du signal enregistré (en général pendant moins de 50 ms), par contre, le bruit de fond est un signal perturbateur présent en permanence. Dans le cas des défauts localisés, la dégradation est caractérisée uniquement par sa brièveté (annexe A), tandis que pour le bruit de fond, il est en général possible de mesurer certaines propriétés statistiques du signal perturbateur qui permettent de mettre en œuvre le traitement (chapitre 2).

1.2.2 Problèmes posés par la réduction du bruit de fond

L'annexe A est entièrement consacrée à l'état de l'art des techniques de traitement utilisées pour lutter contre les défauts localisés. Il semble, que dans ce domaine, les techniques connues à ce jour fournissent des résultats relativement satisfaisants. Pour des enregistrements provenant de disques 78 tours dans un état “correct”, l'application des techniques décrites dans l'annexe A permet d'éliminer une grande partie des défauts localisés sans créer pour autant d'artefacts gênants. Ce point est extrêmement important : les techniques actuelles de traitement des défauts localisés, même si elles ne permettent pas toujours l'élimination totale des dégradations, ne causent pas de distorsion audible du signal. Ceci est en grande partie dû à la nature extrêmement ponctuelle des modifications du signal effectuées dans ce type de techniques (cf. annexe A). D'un point de vue scientifique, on peut dire que la connaissance de ce type de technique est assez complète puisqu'il est possible de quantifier, en grande partie, leurs performances (limite de détection, comportement statistique de l'erreur d'interpolation, sensibilité au bruit de fond ...). Il n'en reste pas moins que certaines situations rencontrées en pratique (présence de plusieurs défauts dans un intervalle de temps très bref, défaut de durée importante) ne peuvent actuellement être traitées. C'est ce qui explique le nombre, proportionnellement important, de publications récentes consacrées à ce type de problèmes (cf. annexe A).

Pour le bruit de fond, la situation est assez différente. Tout d'abord, on peut dire que les techniques utilisées pour la restauration d'enregistrements anciens ne sont pas nouvelles. Elles ont, pour la plupart, été décrites dès la fin des années 70 dans le domaine du traitement des signaux de paroles bruités [Lim 79]. L'application de ces techniques aux signaux musicaux a permis d'obtenir des résultats suffisamment satisfaisants pour envisager une utilisation à grande échelle. A l'heure actuelle, il existe au moins deux systèmes commerciaux de restauration d'enregistrements, le *NoNoise* de la société américaine *Sonic Solutions*, et le *CEDAR* développé en Grande-Bretagne par *CEDAR Audio Ltd.* Ces systèmes ont été utilisés pour de nombreux enregistrements, et, à l'heure actuelle, toute personne qui possède une petite collection de disques compacts à de fortes chances de posséder au moins un enregistrement restauré. On peut signaler, par exemple, la collection *Legendary Classics* de la firme *Philips* composée uniquement d'enregistrements restaurés à l'aide du système *NoNoise* [Marzio 88], ainsi que la collection *Références* de la firme *EMI* qui comporte plusieurs enregistrements récemment restaurés avec le procédé *CEDAR*. Cette pratique de la restauration constitue d'une certaine manière une démonstration de l'efficacité de ce type de techniques.

Cependant certains éléments doivent venir modérer cet enthousiasme. Tout d'abord, la position de notre collectionneur de CD ne lui permet pas de faire une comparaison : il ne dispose que de la version restaurée, et pas de l'original. Cette *absence de référence* rend extrêmement difficile toute évaluation de la qualité du traitement. A notre connaissance, il n'a d'ailleurs jamais été fait d'étude psychoacoustique objective visant à déterminer précisément comment sont perçus les signaux restaurés. Il faut souligner que dans ce domaine de la réduction de bruit appliquée aux enregistrements musicaux, les applications sont allées plus vite que la recherche. En effet, les publications les plus anciennes que nous ayons pu trouver datent environ de 1984, il s'agit d'une communication très complète de Lagadec et Pelloni [Lagadec 83][1], et d'un article de Moorer et Berger [Moorer 86]. Ensuite, il faut attendre 1988 avec la parution de deux thèses, [Bourdier 88] et [Vaseghi 88b] pour trouver de nouvelles publications sur le sujet. Entre temps, la société *Sonic Solutions*, fondée par Moorer, a déjà restauré de nombreux disques [Wright 89] ...

Un point très important est que la qualité des résultats obtenus dépend du niveau de bruit de fond. Ainsi, on peut remarquer que pour l'instant la majorité des enregistrement qui ont été restaurés sont postérieurs aux années 40 [Wright 89]. Il s'agit même souvent d'enregistrements pour lesquelles il existe une bande magnétique originale [Marzio 88] (voir aussi les quelques dates rappelées sur le tableau 1.1). Le niveau de bruit présent sur ce type d'enregistrements (souffle de bande) est en général beaucoup plus faible que le niveau du bruit de surface qui affecte les enregistrements sur disques antérieurs aux années 30 (cf. paragraphe 1.4.4). Il est donc légitime de se demander si en cherchant à restaurer des enregistrements de plus en plus en vieux, donc, en général, de plus en plus bruités, les techniques utilisées à l'heure actuelle ne risquent pas de montrer leurs limites.

En fait, il est possible de répondre d'emblée à cette question, en notant que la plupart des publications consacrées aux techniques de réduction de bruit utilisées pour la restauration des enregistrements anciens mentionnent l'apparition de *distorsions* du signal liées au traitement. Cette distorsion était en général mentionnée dans les articles originaux, consacrés au traitement de la parole [Ephraim 84][Lim 86][Vary 85], on la retrouve "amplifiée" dans les publications qui traitent spécifiquement de l'application aux signaux musicaux[2] [Bourdier 88][Valiere 91]. L'importance accordée à ce problème de la distorsion du signal dans les publications spécifiquement consacrées au cas des enregistrements musicaux ne doit pas nous étonner. En effet, dans le cas des signaux de parole, l'évaluation des résultats prend en compte l'intelligibilité du message, notion qui n'a aucun sens pour les enregistrements musicaux. Pour un signal de parole, on peut, à la rigueur, et selon le type d'application envisagée, s'accommoder d'une forte distorsion du signal, du moment que le message reste compréhensible. Par contre, il serait difficile d'expliquer à un utilisateur d'un système de restauration d'enregistrements musicaux que la distorsion due au traitement est tolérable, puisqu'il est encore possible de distinguer quelles notes ont été jouées par le pianiste ! L'exigence de qualité n'est clairement pas la même dans les deux cas. D'après notre expérience, un auditeur pressé de choisir entre un enregistrement bruité et un

[1] Nous n'avons pu nous procurer cet article qu'assez tardivement ce qui explique qu'il n'occupe pas dans le document la place qu'il mérite. En particulier, il décrit déjà, de manière très claire, le lien entre la limite de restauration pour les sons stationnaires et la résolution fréquentielle de la transformée (cf. paragraphe 3.1.1), ainsi que le phénomène de modulation (cf. paragraphe 3.3). A notre connaissance, ces éléments ne figurent dans aucun article plus récent, à l'exception de [Vary 85] concernant l'effet de modulation (de manière moins claire).

[2] A l'exception notable de S. Vaseghi qui qualifie le système de réduction de bruit de très efficace [Vaseghi 88b, Chapitre 7]. On peut toutefois remarquer qu'à la différence des deux autres thèses citées, la réduction du bruit de fond constitue la partie la moins détaillée du travail de S. Vaseghi, qui est plutôt consacré au traitement des défauts localisés, comme en témoignent ses nombreuses publications citées dans l'annexe A.

enregistrement restauré, sur lequel le signal sonore a été distordu de manière notable, préfère, en général, l'enregistrement original. Ceci met en évidence un aspect très important qui est le caractère "naturel" de l'enregistrement. La restauration d'enregistrements musicaux est un domaine où les éventuels effets secondaires du traitement ne peuvent être tolérés que s'ils sont susceptibles de se fondre parmi les dégradations présentes naturellement sur l'enregistrement. L'exemple le plus démonstratif de ce principe est celui de l'artefact décrit au paragraphe 3.2.1 qui, du fait de sa nature pour le moins "synthétique", est totalement inacceptable même à très faible niveau.

En conclusion, il faut retenir qu'à l'heure actuelle, le traitement de réduction de bruit de fond fournit des résultats très contestables pour certains enregistrements fortement bruités [Valiere 91]. De plus, même pour des enregistrements modérément dégradés, l'utilisation d'un système de réduction de bruit de fond reste une opération très délicate. L'opérateur doit effectuer des réglages empiriques des différents paramètres du traitement afin d'obtenir un résultat satisfaisant [Valiere 91][NoNoise 91][CEDAR 91]. Cette étape de réglage n'est pas triviale, car de légères variations de certains paramètres provoquent rapidement l'apparition de distorsions du signal [NoNoise 91][Jacobs 92].

1.3 Objectifs de l'étude

1.3.1 Plan du document

Nous avons choisi de nous intéresser principalement aux problèmes posés par la réduction du bruit de fond. Conformément à ce que nous avons vu au paragraphe précédent, la principale limitation du traitement est liée à l'apparition d'une distorsion du signal, qui est en général inacceptable pour une application musicale. Le premier objectif de notre étude concerne donc la mise en évidence de cette distorsion due au traitement. Pour ce faire, nous nous proposons de caractériser, *de manière objective*, les effets des techniques de réduction de bruit sur le signal enregistré.

Concernant cette distorsion supposée du signal, les principaux éléments dont nous disposions au départ de l'étude nous ont été communiqués par Jean Christophe Valière, du *Laboratoire d'Acoustique de l'Université du Maine (LAUM)*, à l'occasion de plusieurs rencontres à *TELECOM Paris*. En effet, J. C. Valière avait réussi, durant sa thèse, a mettre en évidence plusieurs de ces défauts, en particulier grâce à des test d'écoute effectués avec des professionnels de l'audio [Valiere 91]. Plus récemment, la venue à *TELECOM Paris* de Rémi Jacobs de la société *EMI*, nous a permis de vérifier que certains de ces défauts étaient bien connus des utilisateurs des systèmes de restauration. Les limitations du traitement de réduction de bruit dont nous avons eu connaissance sont les suivantes :

- Le bruit résiduel demeurant après le traitement possède souvent un caractère peu naturel et très dérangeant [Moorer 86] [Bourdier 88] [Valiere 91]. Même lorsque ce n'est pas le cas, l'absence quasi-totale de bruit résiduel s'avère parfois gênante dans le cas d'enregistrements anciens [Jacobs 92].
- Une modification du timbre général de l'enregistrement est souvent constatée, en particulier, l'enregistrement restauré apparaît en général moins "brillant" (synonyme, dans l'acceptation la plus courante, d'une perte de puissance dans la zone des fréquences élevées) que l'original [Valiere 91][Jacobs 92].

- Les transitoires musicaux présents dans le signal enregistré semblent être affectés par le traitement [Valiere 91].

C'est donc en priorité sur ces différents points que nous avons cherché à obtenir une description objective de l'effet du traitement. Le seul point que nous avons mis en évidence et qui ne figure pas dans cette liste, est l'apparition d'un effet de modulation, dont il s'est avéré, après coup, qu'elle était déjà mentionnée dans [Lagadec 83] (cf. note 1). Le chapitre 3 est consacré à cette évaluation des techniques classiques de réduction de bruit de fond. D'une manière plus générale, dans ce document, toute la première partie est consacrée à la présentation puis à l'analyse des techniques utilisées actuellement (qui peuvent être décrite par le terme générique *d'atténuation spectrale à court-terme*).

La seconde partie correspond à un prolongement naturel de cette étude qui est la proposition de solutions adaptées aux différents problèmes mis en évidence. En fait, les résultats du chapitre 3 fournissent avant tout des éléments qui permettent d'évaluer l'intérêt potentiel d'une technique de réduction de bruit appliquée au cas des enregistrements musicaux. La technique originale qui est présentée au paragraphe 4.2 s'appuie donc très fortement sur les conclusions de l'analyse effectuée au chapitre 3.

Hors du cadre strictement “traitement de signal”, il faut remarquer que l'analyse des distorsions liées au traitement présente, en elle-même, un grand intérêt pour les utilisateurs des systèmes de restauration. En effet, cette évaluation objective de la qualité du traitement peut venir aider l'opérateur du système dans les différents choix qu'il a à effectuer. Même si il est peu probable, et peu souhaitable, que l'opérateur lui-même soit remplacé par une procédure automatique, on peut néanmoins imaginer un ensemble d'indicateurs qui permettent de signaler la présence probable de tel ou tel défaut. Le principal avantage de cette démarche est qu'il est beaucoup plus facile de juger auditivement de la qualité d'un résultat lorsque l'on sait d'avance où se situent les problèmes éventuels. L'autre intérêt est que cette analyse des résultats permet, dans certains cas, d'établir une correspondance entre les paramètres du traitement et des grandeurs significatives pour l'utilisateur. Par exemple, le compromis entre, la limite de restauration dans les parties stationnaires, et le lissage des transitoires brusques (cf. chapitre 3), est une notion beaucoup plus utile en pratique que le nombre de points sur lequel est effectuée la transformée de Fourier, or il s'avère qu'il existe une correspondance entre les deux.

Enfin un des prolongements possibles de cette étude est la mise au point de véritables tests psychoacoustiques destinés à préciser comment sont perçus les enregistrements restaurés. En effet, pour constituer la base de données destinée à un tel test, il est nécessaire de savoir quelle est l'influence respective des différents paramètres que l'on se propose de faire varier. L'étude développée ici peut être utilisée pour sérier les problèmes en vue d'une évaluation perceptive. La difficulté posée par cette évaluation perceptive resterait tout de même très importante car la variation d'un seul paramètre du traitement entraîne, en général, de nombreuses conséquences distinctes. En l'occurrence, la démarche la plus efficace serait plutôt d'évaluer la perception des différents effets mis en évidence au chapitre 3 séparément, en utilisant des signaux très simples, avant de passer à des évaluations de véritables enregistrements musicaux dégradés.

Une dernière question qui semble intéresser à juste titre la plupart des utilisateurs potentiels des disques restaurés, est de savoir si les systèmes existants fournissent de bons résultats. Le problème posé par la situation évoquée au paragraphe 1.2.2 est qu'il est extrêmement difficile de savoir précisément comment fonctionnent les systèmes existants, puisqu'il n'existe pas de documents, autres que promotionnels, décrivant les techniques utilisées. Les renseignements dont nous disposons sur ces systèmes [CEDAR 91][NoNoise 91], nous permettent de dire que les techniques utilisées appartiennent bien à la catégorie décrite dans la première partie de ce document (réduction de bruit par atténuation spectrale à court-terme). Pour la plupart, les problèmes évoqués au chapitre 3, sont suffisamment généraux pour qu'on puisse légitimement penser qu'ils surviennent aussi avec les systèmes existants. Cependant, le savoir faire acquis dans les sociétés qui commercialisent ces systèmes a permis de trouver des solutions efficaces à certains problèmes, en particulier, celui du bruit résiduel, qui ne sont pas forcément celles qui sont décrites dans ce document. En l'absence de point de comparaison précis, il est donc très difficile de se prononcer sur les performances des systèmes existants.

1.3.2 Domaine d'application

Comme le souligne la référence [Lagadec 83], il est possible d'envisager deux types d'applications différentes pour les techniques de réduction de bruit de fond. La première consiste à réduire le niveau du bruit présent sur des enregistrements déjà très peu bruités. Il s'agit en général d'enregistrements relativement récents (postérieurs aux années 50), dont ont considère que le niveau de bruit est tolérable à la radio, ou sur un disque 33 tours, mais pas sur un CD. Dans ce cas, le but à atteindre est généralement l'élimination totale du bruit de fond. Comme nous l'avons déjà signalé, cette application représente une grande partie du marché de la restauration d'enregistrement (voir par exemple les enregistrements décrits dans [Marzio 88]). La deuxième application consiste à traiter des enregistrements fortement dégradés où la qualité de l'enregistrement est jugée très gênante.

Le traitement d'un enregistrent peu bruité n'est pas forcément aussi simple qu'il y paraît, car dans le cas d'un "bon" enregistrement, l'exigence de qualité est différente. De plus, les qualités techniques de l'enregistrement (large bande passante, dynamique importante) ont tendance à mettre ponctuellement en évidence des défauts du traitement, difficilement perceptibles sur des enregistrements de très mauvaise qualité. A priori, les deux types d'applications évoquées précédemment nous concernent, cependant le traitement d'enregistrements fortement dégradés apparaît comme plus fondamental puisque la nécessité du débruitage est moins discutable. Par ailleurs, le tableau 1.2 montre que, par la force des choses, c'est aussi le cas auquel nous avons été le plus souvent confronté.

Un dernier point concerne le type de signaux musicaux présents sur les enregistrements traités. Nous avons essayé d'éviter d'imposer des hypothèses supplémentaires concernant la nature des signaux à débruiter. En particulier, nous verrons, à propos du tableau 1.2, que nous disposions de plusieurs enregistrements de piano solo. Dans un cas comme celui-ci, la prise en compte de la nature du signal peut permettre d'envisager une méthode de restauration différente qui passe par la modélisation du signal. Des éléments qui peuvent être pris en compte dans un tel modèle sont, par exemple, la quasi-harmonicité des partiels du signal, ou bien les caractéristiques de décroissance des différents partiels au cours du temps. Récemment, plusieurs type de modèles applicables aux signaux musicaux ont été proposés [Serra 89] [Depalle 91] [Meillier 93]. Pour certains de ces modèles, la restauration fait partie des applications envisagées. Cependant, une des limitations majeure d'une telle approche est qu'elle est pratiquement limitée au cas des signaux monophoniques[3]. Il devient très difficile de modéliser un signal lorsque celui-ci est constitué de plusieurs sources distinctes, surtout en l'absence d'argument permettant de séparer ces différentes sources. Pour l'exemple cité précédemment d'un enregistrement de piano solo, il faut considérer

[3]Une autre difficulté pratique est qu'en présence de niveaux de bruits importants, il est nécessaire de disposer de techniques très robustes pour déterminer les paramètres du modèle. Par exemple, une technique comme la détection de pics effectuée sur le spectre à court-terme, utilisée dans [Serra 89], n'est plus appropriée dans le cas d'un signal fortement bruité.

le fait qu'il s'agit d'un instrument polyphonique. En conséquence, l'utilisation d'une technique automatique de restauration fondée sur un modèle du signal demeure difficilement envisageable dans ce cas.

1.3.3 Evaluation des résultats

1.3.3.a Utilisation de signaux de synthèse

Un point très important, qui peut surprendre le lecteur, est que les enregistrements présentés par le tableau 1.2 n'apparaissent quasiment pas ailleurs dans le document, sauf pour les considérations sur la nature du bruit de fond. La raison en a été déjà mentionnée : nous estimons, qu'en l'état actuel des choses, il nous est très difficile d'analyser les résultats obtenus sur des signaux musicaux complexes. En conséquence, nous avons considéré que des signaux "réels" ne nous permettraient pas de mettre en évidence les effets étudiés au chapitre 3. C'est pourquoi, les signaux utilisés sont en majorité des signaux synthétiques extrêmement simples.

La simplification la plus utilisée dans ce document consiste à considérer qu'il existe des intervalles de temps importants durant lesquels le signal inconnu peut être assimilé à une somme de composantes sinusoïdales faiblement modulées. Il n'est pas question ici d'affirmer que les signaux musicaux sont structurellement sinusoïdaux, mais plutôt de souligner le fait que l'efficacité de plusieurs techniques récemment développées [Serra 89] [George 92] incite à penser que la description sous la forme d'une somme de sinusoïdes est efficace pour une grande variété de sons musicaux. Par ailleurs, plusieurs arguments, tant physiques que perceptifs, permettent de justifier la pertinence d'une telle description [Benade 76] [Hall 80] [Deutsch 82]. En conséquence, une grande partie des calculs du chapitre 3 ont été effectués en considérant une simple sinusoïde noyée dans du bruit. Par suite, nous avons essayé de décrire, surtout qualitativement, la manière dont ce cas simple peut se transposer au cas d'un signal musical, considéré dans une partie quasi-stationnaire.

Une remarque importante est que cette description est particulièrement justifiée dans le cas du débruitage. En effet, pour reprendre la terminologie de Xavier Serra, si on modélise un son musical sous la forme de la somme d'une partie déterministe (somme de sinusoïdes), et d'une partie aléatoire (bruit filtré) [Serra 89], il faut avoir conscience du fait que c'est principalement la partie déterministe qui est susceptible d'être restaurée. Le paragraphe 2.1.1 rappelle en effet, que, de manière fondamentale, la séparation de deux signaux, caractérisés uniquement par leur densité spectrale de puissance, n'est efficace que si leurs supports fréquentiels sont distincts. Dans le cas d'un son musical, c'est donc principalement pour la partie "déterministe" du signal, définie ci-dessus, que l'amélioration est sensible.

Ce point nous donne l'occasion de préciser quelle est la différence fondamentale, en ce qui nous concerne, entre les signaux musicaux et les signaux de parole. Pour un signal musical, la description sous la forme d'une somme de sinusoïdes faiblement modulées est susceptible de rester "valide" sur une durée extrêmement longue. Cette durée peut aller jusqu'à plusieurs secondes dans le cas d'une note tenue, plus communément, elle sera de l'ordre de la centaine de millisecondes. Par contre, dans le cas des signaux de parole, cette description n'est valable que sur des durées très brèves, de l'ordre d'une dizaine de millisecondes. Pour donner un élément de comparaison utile, notons qu'un rythme musical extrêmement rapide, par exemple, des triples croches, exécutées au tempo de 180 à la noire, correspond à des notes d'une durée approximative de 40 ms, ce qui représente la durée d'un événement plutôt long pour le signal de parole. Les exemples sonores présentés avec l'article de Serra et Smith [Serra 90], fournissent d'ailleurs une confirmation de cette constatation. En effet, pour les sons instrumentaux, la partie déterministe (somme de sinusoïdes) est relativement proche des sons originaux à l'exception des parties transitoires qui sont complètement dénaturées. Par contre, pour les sons de parole, c'est quasiment la partie aléatoire (bruit filtré) qui est la plus représentative, en particulier, elle est tout à fait compréhensible. En remarquant que les paramètres du modèle sont déterminés sur des fenêtres d'une durée de l'ordre de 25 ms, on peut en conclure, *que sur cette durée d'observation*, le signal de parole n'est pas décrit de manière satisfaisante par une somme de sinusoïdes, alors que certains signaux musicaux peuvent l'être, en excluant les parties transitoires. Cette différence est très importante, elle justifie, entre autre, pourquoi les mêmes techniques ne sont pas utilisées avec les mêmes paramètres selon que l'on traite un signal de parole ou un enregistrement musical bruité (paragraphe 2.1.3). De même, une technique de traitement intéressante pour un type de signal ne le sera pas forcément pour l'autre, c'est par exemple le cas de la technique décrite au paragraphe 4.2.

Ceci nous amène à parler du traitement des parties transitoires des signaux musicaux, dont ont sait qu'elles jouent un rôle très important du point de vue perceptif, particulièrement pour l'identification des sons instrumentaux [Hall 80, § 6.7] [Deutsch 82, Chap. 2] [Pollard 82]. Pour l'instant, on ne dispose pas d'une méthode de description du signal musical présent pendant les transitoires qui soit aussi pertinente, et aussi générale, que dans le cas de la partie quasi-stationnaire. L'attitude adoptée face à ce problème a consisté à utiliser un signal transitoire idéalisé qui donne une idée du comportement du traitement vis à vis des signaux transitoires, et ce, dans le cas le plus défavorable. Toutefois, les formes d'ondes présentées au paragraphe 3.1.2.c donnent à penser que, *compte-tenu des durées d'observation caractéristiques du traitement de débruitage*, le type de signal utilisé (sinusoïde apparaissant abruptement) est assez représentatif du comportement de certains signaux musicaux dont le transitoire d'attaque est très bref, et la décroissance pas trop rapide. C'est en particulier le cas pour certains sons instrumentaux de type percussif (par exemple le piano ou la guitare).

Une étape intermédiaire, entre les signaux purement synthétiques, et les cas réels, consiste à bruiter un signal musical de bonne qualité. Ce type de démarche est indispensable car elle permet de contrôler complètement les conditions de bruit. En effet, sur un enregistrement réel, une distorsion qui ne se produit que lorsque certaines composantes de signal arrivent au niveau du bruit de fond (cf. chapitre 3) va apparaître de manière extrêmement brève car le signal varie constamment. A l'opposé, avec un signal sur lequel on ajoute artificiellement un bruit, il suffit d'ajuster le niveau du bruit pour que la distorsion se produise à l'endroit voulu. De plus, avec un signal artificiellement bruité, on dispose d'un élément inaccessible dans un cas réel, qui est le signal original. Cette possibilité fournit une référence qui rend l'évaluation auditive beaucoup plus significative. En effet, lorsque la tâche exigée du sujet est complexe, l'absence d'un protocole de test strict, fournit en général des réponses inexploitables. Par contre, lorsque la tâche est simple, s'il s'agit, par exemple, de comparer un aspect précis de deux signaux, on obtient des résultats relativement fiables, même lors d'écoutes informelles.

1.3.3.b Simulation de la perception

Compte tenu de la représentation simplifiée adoptée pour le signal musical, il est légitime de considérer que la perception auditive peut être analysée en utilisant les propriétés connues de l'effet de masquage auditif simultané d'un son pur par un bruit (ou inversement). Là encore, même si cette notion de masquage ne résume pas le fonctionnement réel de la perception auditive, on est en droit d'attendre de bons résultats de cette analyse pour le type de signal considéré. Il s'avère que cet effet de masquage fréquentiel permet effectivement de retrouver certaines constatations empiriques formulées à propos des systèmes de réduction du bruit de fond (voir le chapitre 3). Là encore, on peut considérer que cette vision simplificatrice de la perception auditive reste tout de même valable pour un grand nombre de signaux musicaux, comme en témoigne les performances des techniques récentes de codage de signaux audio fondées sur la simulation de l'effet de masquage [Mahieux 89] [Johnston 88] [Zwicker 91b].

L'avantage de cette démarche que nous avons adoptée est de se ramener à des situations très simples, pour lesquelles il est possible de décrire analytiquement le signal restauré, puis par la suite, d'utiliser des résultats connus sur la perception auditive. Le point délicat est bien sûr la généralisation de ces résultats à des situations plus complexes. On peut considérer qu'elle est parfaitement valable dans le cas des parties quasi-stationnaires à condition de prendre certaines précautions pour l'aspect perceptif (en particulier, en tenant compte du masquage de certaines composantes du signal par d'autres composantes du même signal). Par contre, pour les transitoires, la généralisation à des signaux musicaux complexes reste très approximative, compte tenu de la difficulté de trouver un signal transitoire suffisamment représentatif.

En conclusion, pour en revenir aux enregistrements du tableau 1.2, il faut souligner qu'ils ont tous été traités. Certains d'entre eux ont été utilisés comme signal de test pour comparer différentes techniques. Cependant, nous avons souligné le fait qu'il est très difficile d'interpréter les réactions subjectives des auditeurs pour ce type de signaux complexes. En particulier, ces réactions font appel à des données externes au problème (connaissance musicale, goûts) difficilement quantifiables. Nous avons par ailleurs signalé le fait que même en adoptant un véritable protocole expérimental d'évaluation perceptive, il serait très difficile d'obtenir des informations significatives concernant la restauration de fragments musicaux complets. A fortiori, nous avons donc choisi de pas trop insister sur les remarques qui ont été formulées lors de séances d'écoute informelles.

1.4 Nature du bruit de fond

Le but de cette partie est de présenter plus en détail la dégradation qui est étudiée dans la suite du document, c'est à dire, le bruit de fond. C'est en particulier l'occasion de justifier certaines des hypothèses ou approximations utilisées.

1.4.1 Provenance des enregistrements

1877-1905	Phonographe (cylindre)
1887-1926	Gramophone à disque
1893-1900	Mise en place de l'industrie phonographique
1926	Gravure électrique des disques
1948	Utilisation du magnétophone
1948	Disque microsillon monophonique
1957	Disque microsillon stéréophonique

Tableau 1.1: Quelques dates concernant l'histoire de l'enregistrement. (d'après [Jessel 85])

Le tableau 1.2 présente une description rapide des enregistrements dont nous disposions pour cette étude. Il faut noter que la majorité de ces enregistrements correspondent à des disques assez anciens (antérieurs aux années 40), donc, a priori, au cas d'enregistrement fortement bruités. Pour la série des enregistrements de flamenco, il s'agit même de documents particulièrement anciens comme en témoigne la chronologie sommaire reproduite sur le tableau 1.1. Les restaurations de ces enregistrements, effectuées à *TELECOM Paris*, au cours du second trimestre 1993, doivent d'ailleurs paraître sous la forme d'un CD [Flamenco 93].

Description	Format original	Particularités	Provenance
16 enregistrements historiques de flamenco	Disques enregistrés entre 1907 et 1912	Très dégradés	Peña Juan Breva, Radio Nacional de España (Málaga, Espagne)
2 enregistrements	Disques 78 tours	Piano solo (fortement bruité)	EMI
Version complète du *Requiem* de Fauré	Une dizaine de 78 tours (1929/1930)	Peu de bruit, pleurage important	EMI
5 enregistrements	Disques 78 tours récents (années 40)	-	Collections privées
Extrait	Disque (environ 1929)	Issu de la "base de données" de [Valiere 91] (référence CAR)	J. C. Valière (LAUM)
3 enregistrements	bande magnétique (années 50)	Dont un de piano solo (peu bruité)	Vincent Prost (INA)
Plusieurs signaux de parole dans le bruit	-	Bruit de voiture	Département SIGNAL (TELECOM Paris)

Tableau 1.2: Enregistrements utilisés.

D'après le tableau 1.2, on peut dire que nous disposions d'un ensemble assez représentatif, surtout en ce qui concerne les enregistrements originaux sur disques. Un point intéressant est que nous avons pu obtenir des enregistrements ne comportant qu'un seul instrument musical (en l'occurrence, le piano). Ces enregistrements, plus simples, constituent de très bons signaux-test. En particulier, dans les enregistrements fournis par la société *EMI*, où le niveau de bruit est très important, les attaques relativement marquées du piano (instrument percussif) mettent en évidence, de manière particulièrement flagrante, le phénomène de distorsion des transitoires. Enfin, nous avons eu l'occasion de travailler à partir d'un des enregistrements utilisés par J. C. Valière au cours de sa thèse, cette utilisation de données commune à permis des comparaisons fructueuses avec les résultats obtenus au LAUM.

1.4.2 Caractéristiques et hypothèses

Dans la suite du document, le bruit de fond est systématiquement considéré comme une dégradation, additive, stationnaire, non-corrélée avec le signal. Une remarque importante est que le traitement serait très difficile à mettre en œuvre si l'on excluait l'une des ces hypothèses : la linéarité permet de se retrouver dans le cadre standard du traitement de signal, la non-corrélation est nécessaire car le seul signal accessible est le signal bruité, et enfin, la stationnarité du bruit autorise sa caractérisation par la mesure préalable de la densité spectrale de puissance (ou DSP). Notons de plus, qu'il est impossible de vérifier les hypothèses de linéarité et de non-corrélation dans le cadre que nous nous sommes fixé. En effet, si on ne dispose que du signal dégradé, et en l'absence d'informations sur les signaux présents sur l'enregistrement, il n'est pas possible de vérifier ces hypothèses qui décrivent l'interaction entre le signal et le bruit. Pour vérifier ces deux hypothèses, il faudrait utiliser une approche physique, visant à décrire, voire à simuler, le processus de dégradation pour un type d'enregistrement donné. Par contre, il est possible de vérifier la troisième hypothèse puisqu'elle porte uniquement sur le bruit. Il suffit dans ce cas de s'intéresser aux portions de l'enregistrement qui ne comportent que du bruit de fond. C'est le principe qui est développé au paragraphe 1.4.5.

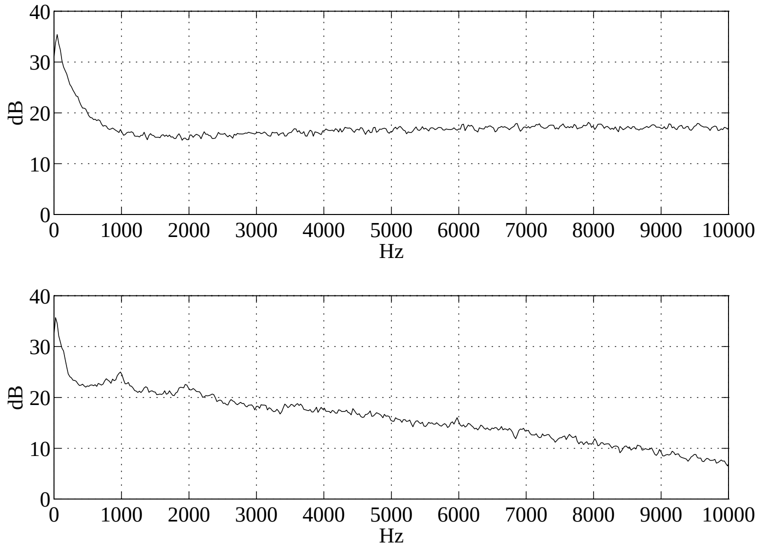

Figure 1.1: Densité spectrale de puissance pour deux bruits de même puissance. En haut, le bruit de souffle présent sur une cassette analogique (sans réducteur de bruit). En bas, le bruit de surface présent sur un disque 78 tours. Calcul par périodogramme moyenné sur 4 secondes de signal (soit environ 95 fenêtres de 1024 points).

La caractéristique du bruit qui est importante pour le type de techniques utilisées est sa densité spectrale de puissance (cf. paragraphe 2.1.1). La figure 1.1 présente deux exemples correspondant à des bruits mesurés sur des supports différents. La puissance de ces deux bruits a été normalisée afin de faciliter la comparaison. La principale caractéristique de ces deux DSP est qu'elles présentent un maximum très prononcé dans le domaine des basses fréquences. On retrouve là une allure typique de la plupart des bruits d'origine physique. Une remarque importante est que selon la nature du bruit considéré, ce maximum ne correspond pas forcément au même phénomène. Nous verrons au paragraphe 1.4.3, pour l'exemple d'un disque 78 tours, que ce maximum traduit la présence d'une résonance de très basse fréquence (autour de 30 Hz). Une différence tout de même entre ces deux bruits est que le bruit de cassette (en haut sur la figure 1.1) possède une DSP quasiment constante au dessus de 2 kHz, alors que la DSP du bruit de disque continue à décroître fortement. L'explication de cette différence est à rechercher dans le fait que les deux systèmes (cassette analogique moderne, et disque 78 tours) ne possèdent pas la même bande passante.

A ce propos, une remarque importante est que pour les enregistrements provenant de disques dont nous disposions, la bande occupée par le signal est extrêmement réduite. Nous avons vérifié pour plusieurs de ces enregistrements, que le signal, une fois filtré passe-haut au dessus de 8 kHz, ne contient plus que du bruit. Il faut noter qu'ici, la procédure inverse, qui consiste à comparer le signal avec un filtrage passe-bas du même signal, ne convient pas, à cause du bruit de fond. En effet, si on filtre un signal issu d'un disque 78 tours par un filtre passe bas de fréquence de coupure 10 kHz, on perçoit bien une modification. Toutefois celle-ci est due à l'élimination d'une partie du bruit (la bande située au dessus de 10 kHz). Pour s'en convaincre il suffit de vérifier que la partie filtrée ne comporte pas de signal détectable dans le bruit. La conséquence pratique de cette limitation de la bande occupée par le signal est qu'une fréquence d'échantillonnage de l'ordre de 20 kHz est largement suffisante pour les enregistrements provenant de disques 78 tours.

Le choix d'une fréquence d'échantillonnage réduite réalise simplement une partie du travail qui serait effectué, de manière plus coûteuse, par l'algorithme de réduction de bruit si on utilisait une fréquence d'échantillonnage plus élevée. Pour des raisons pratiques (conversion de fréquence simple vers la fréquence standard de 48 kHz), c'est souvent à la fréquence de 24 kHz que nous avons échantillonné les signaux dont nous disposions.

Il faut cependant signaler, que même dans le cas où le signal occupe une bande encore plus réduite (pour le cas des enregistrements les plus anciens dont nous disposions la bande occupée par le signal s'arrête plutôt vers 5 kHz !), on n'a pas intérêt à réduire davantage la fréquence d'échantillonnage. En effet la présence de bruit résiduel dans le haut du spectre est, en général, perçu comme un aspect positif qui vient "compenser" la faible bande passante du signal audio (cf. paragraphe 3.1.1.c).

1.4.3 Mesure des caractéristiques spectrales du bruit

Pour les techniques de réduction de bruit étudiées dans ce document, la mesure des caractéristiques spectrales du bruit de fond est en général effectuée par la méthode du périodogramme moyenné [Kay 88]. En effet, d'après le paragraphe 2.1.3, le traitement de débruitage s'effectue, dans la plupart des cas, en utilisant la transformée de Fourier à court-terme (ou TFCT). Dans ce cas, pour mesurer les caractéristiques spectrales du bruit de fond, il suffit d'utiliser la même structure de calcul de TFCT sur une portion de bruit seul, et de moyenner le carré du module des spectres à court-terme obtenus. De plus, la quantité utile pour le traitement n'est, en général, pas l'estimation de la DSP du bruit, mais directement la valeur moyenne du spectre de puissance à court-terme en présence de bruit (cf. paragraphe 2.2). Ce qui explique que, dans le contexte du débruitage, le périodogramme moyenné constitue la méthode d'estimation spectrale la plus naturelle.

Dans la suite du document, nous avons considéré que l'estimation de la DSP du bruit de fond, utilisée lors du traitement, est non biaisée et de variance négligeable. Quelques remarques s'imposent concernant ces deux points :

Biais de l'estimation L'estimation de la DSP obtenue par périodogramme moyenné est asymptotiquement non biaisée [Brillinger 81]. Toutefois, pour des longueurs de trames finies, on observe un effet de lissage qui peut être gênant si la densité spectrale que l'on cherche à estimer présente des variations brusques [Kay 88]. Compte tenu de l'allure des DSP représentées sur la figure 1.1, le seul endroit où l'insuffisance de résolution de l'estimateur spectral peut poser problème est le domaine des basses fréquences (en dessous de 500 Hz).

Afin d'illustrer cet aspect, la figure 1.2 présente une comparaison entre deux estimations spectrales obtenues par périodogramme moyenné, avec des longueurs de trame différentes (en trait plein), et une estimation obtenue par une méthode haute résolution (tirets). Cette comparaison a été effectuée dans le cas d'un bruit provenant d'un disque 78 tours. L'estimation spectrale haute résolution a été obtenue, en deux étapes :

a) **Sous-échantillonnage par un facteur 6** En effet, l'allure des figures 1.1 donne à penser qu'il existe une résonance très basse fréquence. Or, on sait qu'il est impossible d'obtenir une estimation fiable des caractéristiques de cette résonance si sa fréquence est trop faible [Kay 93]. La solution consiste à sous-échantillonner préalablement le signal. La fréquence d'échantillonnage étant ici de 24 kHz, le bruit sous-échantillonné occupe la bande [0,2 kHz].

b) **Estimation AR** L'estimation spectrale est obtenue grâce à un modèle AR du bruit sous-

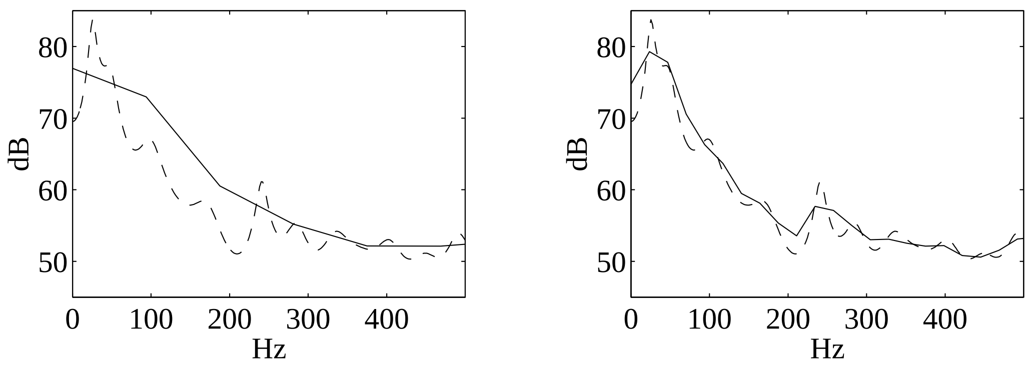

Figure 1.2: Estimations de la densité spectrale de puissance du bruit provenant d'un disque 78 tours. **Tirets,** estimation haute résolution (décimation + modélisation AR). **Trait plein,** estimations obtenues par périodogramme moyenné : **à gauche** moyenne sur des trames de 10 ms, **à droite** moyenne sur des trames de 40 ms.

échantillonné. Afin de mettre en évidence la résonance, nous avons utilisé la méthode de covariance modifiée, avec un ordre important (ici 100) [Kay 88].

L'estimation spectrale obtenue est comparée, dans la zone [0,500 Hz], avec deux estimations obtenues par périodogramme moyenné. L'estimation spectrale haute résolution (tirets) permet bien de mettre en évidence un maximum très marqué de la DSP, à une fréquence de l'ordre de 30 Hz. L'estimation par périodogramme moyenné représentée sur la droite de la figure 1.2 est très proche de cette estimation haute résolution, excepté autour du maximum, qui est légèrement sous-estimé. Par contre, l'estimation obtenue en moyennant les périodogrammes calculés sur des trames plus courtes, représentée sur la gauche de la figure, correspond à une approximation assez grossière de la DSP pour les basses fréquences. L'aspect brisé de la courbe vient du fait que l'écart entre deux valeurs estimées est ici de l'ordre de 100 Hz. En pratique, un comportement tel que celui-ci est assez gênant car la DSP est nettement surestimée dans la zone [100,200 Hz]. Or une telle surestimation risque de se traduire par une distorsion accrue du signal présent dans cette zone de fréquence (cf. chapitre 3).

Pour conclure sur le biais de l'estimation spectrale du bruit de fond obtenue par périodogramme moyenné, il faut retenir que lorsque l'estimation est réalisée sur des trames d'une durée de 40 ms ou plus, on peut considérer que l'estimation obtenue est non-biaisée, même en basse fréquence. Par contre, pour des durées de trame plus courtes, de l'ordre de 10 ms, la DSP estimée en basse fréquence (typiquement en dessous de 300 Hz) est souvent surestimée en raison de la présence d'une résonance très basse fréquence. Cette résonance est en général présente sur tous les enregistrements provenant de disques 78 tours n'ayant pas été préalablement filtrés.

Variance de l'estimation Lorsqu'on utilise le périodogramme moyenné, la variance de l'estimation est inversement proportionnelle au nombre de trames sur lesquelles la moyenne a été effectuée. En utilisant le comportement asymptotique du périodogramme [Brillinger 81], il est possible d'obtenir assez simplement un intervalle de confiance pour les valeurs estimées de la DSP [Kay 88]. Le tableau 1.3 présente les intervalles de confiance à 99% correspondant à différentes valeurs du nombre de trames. D'après ce tableau, il est nécessaire de moyenner au moins sur 20 à 30 trames distinctes pour obtenir une estimation relativement fiable. Il faut souligner que le calcul de l'intervalle de confiance est basé sur l'hypothèse de trames indépendantes. L'intervalle de confiance obtenu est à peu près correct pour des trames de signal successives sans recouvrement (bien que l'hypothèse d'indépendance ne soit strictement respectée dans ce cas que pour un bruit blanc) [Kay 88]. Par contre, l'intervalle est nettement sous-estimé lorsque les trames de signal utilisées pour calculer les périodogrammes se recouvrent fortement. En pratique, on peut considérer que le recouvrement entre les trames ne modifie quasiment pas l'intervalle de confiance, il contribue surtout à obtenir une courbe d'aspect plus "lisse". Par exemple, l'intervalle de confiance associé à un calcul du périodogramme moyenné sur 40 trames se recouvrant à 50% correspond plutôt au chiffre de 20 trames (supposées sans recouvrement) sur le tableau 1.3.

Nombre de trames	**Bornes de l'intervalle de confiance à 99% (en dB)**	
	Inférieure	**Supérieure**
5	-6,5	4
10	-4,5	3
20	-3	2
40	2	1,5

Tableau 1.3: Intervalle de confiance pour les valeurs de la DSP estimées par périodogramme moyenné, en fonction du nombre de trames considérées (trames sans recouvrement). Les bornes sont exprimées en décibels par rapport à la valeur exacte.

En conclusion, il faut retenir que l'estimation des caractéristiques spectrales du bruit de fond requiert une portion de bruit seul de durée assez importante. En effet, nous avons vu qu'il est souhaitable de disposer d'au moins 20 à 30 trames (sans recouvrement) pour garantir un intervalle de confiance raisonnable. De plus, la durée des trames doit être d'au moins une vingtaine de millisecondes afin d'éviter un biais en basse fréquence qui peut être pénalisant. C'est à dire que la portion de bruit seul, utilisée pour mesurer les caractéristiques spectrales, doit avoir une durée d'au moins une demi-seconde (25×20 ms). En pratique, pour simplifier le calcul c'est souvent les mêmes paramètres de TFCT (durée de trame à court-terme, recouvrement) qui sont utilisés lors de la mesure des caractéristiques du bruit de fond et lors du traitement de débruitage[4]. Or, les durées de trames à court-terme utilisées pour la restauration d'enregistrements musicaux sont en général supérieures à 40 ms (cf. paragraphe 2.1.3). Dans ces conditions, la durée de la portion de bruit seul requise pour mesurer la DSP du bruit de fond doit être plus importante (de l'ordre d'une seconde).

Dans la plupart des cas, les enregistrements à restaurer présentent des portions de bruit seul, avant et après la partie musicale, qui durent plusieurs secondes. La mesure des caractéristiques spectrales du bruit de fond ne pose donc pas de problème dans ces conditions. Cependant, nous avons été confronté à des exemples délicats où les parties de bruit seul avaient été artificiellement coupées. Cette situation correspond à des cas où l'on ne dispose pas directement du transfert de l'enregistrement original (par exemple, si l'enregistrement traité provient du repiquage d'un enregistrement plus ancien).

1.4.4 Niveau de bruit de fond

Une caractéristique très importante du bruit de fond est l'intensité sonore avec laquelle il est perçu. En effet, c'est la comparaison de cet attribut avec l'intensité sonore du signal musical qui devrait être utilisé pour quantifier l'importance de la dégradation. D'une manière générale, la mesure de l'intensité sonore perçue est un problème assez compliqué, car de nombreux aspects

[4] Ce choix n'est absolument pas obligatoire : il est possible d'utiliser des trames plus courtes pour la mesure du bruit de fond. Il suffit alors de compléter les trames de signal par *zero-padding* avant le calcul des transformées de Fourier [Kay 88], et d'utiliser la formule de normalisation (C.13) (annexe C) pour garantir l'homogénéité.

doivent être pris en compte (composition fréquentielle du signal, évolution temporelle, niveau absolu d'écoute, etc.) [Botte 88]. Dans ce paragraphe, on se propose simplement de déterminer quelques éléments importants concernant la perception de l'intensité sonore, pour trois exemples de bruits différents, choisis pour leur représentativité. L'idée est de formaliser un peu plus la notion d'enregistrement fortement (ou peu) dégradé en fournissant des points de repères pour quelques cas typiques. Une première idée consiste à utiliser une représentation de type "histogramme de puissance" qui permet une vision plus fine qu'un simple rapport signal-à-bruit. Par référence à la procédure présentée dans [Zwicker 91b], que nous commenterons plus loin, nous avons choisi de représenter un type d'histogramme particulier.

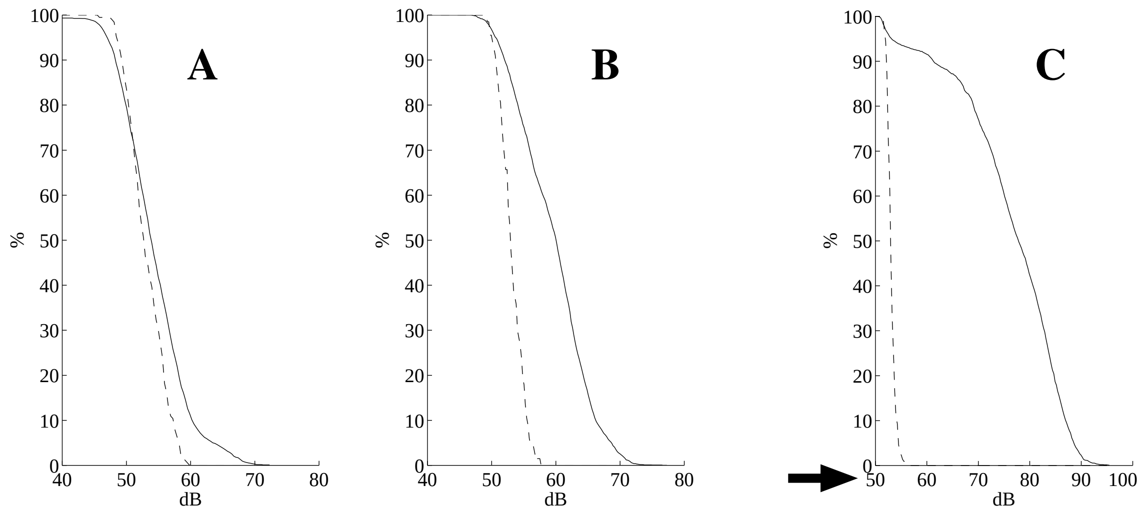

Figure 1.3: Distribution cumulative de la puissance : pourcentage des valeurs mesurées supérieures à une valeur donnée. La puissance du signal est mesurée sur des trames consécutives d'une durée de 40 ms. **En trait plein, distribution cumulative de la puissance du signal bruité. En tirets, distribution cumulative de la puissance du bruit seul.** La flèche signale un changement d'échelle sur les abscisses de la partie **C**.

Les trois enregistrements considérés correspondent aux parties **A B** et **C** de la figure 1.3. La quantité représentée sur chacune de ces figures est la *distribution cumulative de puissance*, c'est à dire la portion de l'enregistrement durant laquelle la puissance mesurée est supérieure à une valeur donnée, sachant que la puissance est mesurée sur des durées d'observation brèves (40 ms). Les trois enregistrements, ont été normalisés, de telle façon que le niveau global du bruit soit identique dans les trois cas. Avant de commenter la figure 1.3, notons que les trois enregistrements sélectionnés possèdent les caractéristiques suivantes :

Enregistrement A Il s'agit de l'enregistrement le plus dégradé de la série des "Flamenco" (cf. tableau 1.2), qui est aussi le plus dégradé dont nous disposions.

Enregistrement B C'est, au contraire, un des enregistrements les moins dégradés de la série des "Flamenco". Par rapport à l'ensemble des enregistrements provenant de disques à notre disposition, il correspond à un enregistrement moyennement à fortement dégradé.

Enregistrement C Correspond à un enregistrement sur bande magnétique datant des années 50. Pour un enregistrement sur bande, le souffle est plutôt important, cependant il correspond à une qualité tout à fait acceptable.

Le tableau 1.4 présente les rapports signal-à-bruit (ou RSB) calculés à partir de l'enregistrement complet, dans chacun des trois cas. Globalement, on peut dire qu'ils confirment notre description sommaire des enregistrements. Cependant, le rapport signal-à-bruit global semble indiquer que la différence entre les enregistrements **A** et **B** est du "même ordre" que celle qui existe entre les enregistrement **B** et **C** (environ 12 dB dans les deux cas). Nous allons voir qu'il n'en est rien, et que l'enregistrement **B** est en fait très proche de l'enregistrement **A**.

Enregistrement A	Enregistrement B	Enregistrement C
1 dB	12 dB	24 dB

Tableau 1.4: Rapport signal-à-bruit estimé. (enregistrements de la figure 1.3)

Le tableau 1.4 nous permet de répondre à une interrogation légitime qui est de savoir pourquoi nous avons choisi de représenter sur la figure 1.3, d'une part la puissance du signal bruité, et d'autre part, celle du bruit seul. Il semblerait en effet plus simple de représenter un rapport signal-à-bruit. Cependant, dans le cas **A**, le niveau de bruit est tel qu'un RSB calculé sur des fenêtres de 40 ms n'a absolument aucune signification, car la variance du RSB estimé est alors du même ordre de grandeur que la valeur du RSB. Dans des cas comme celui-ci, le rapport signal-à-bruit doit forcément être calculé de manière globale.

Si on revient à la figure 1.3, on constate que la distribution cumulative de puissance du signal bruité permet d'établir une différence importante entre les cas **A** et **B**, et le cas **C**. En effet, dans les deux premiers cas, la dynamique du signal bruité (courbe en trait plein) est de l'ordre de 20 à 30 dB tandis qu'elle est d'environ 50 dB dans le cas **C**[5]. Cette différence n'est pas étonnante compte tenu du fait que les deux premiers enregistrements proviennent de disques 78 tours tandis que le troisième correspond à une bande magnétique.

Par la suite, pour chacune des parties de la figure 1.3, la comparaison entre la courbe correspondant au bruit seul (tirets) et celle correspondant au signal bruité permet de se faire une idée des conditions de bruit. Pour la partie **A**, les deux distributions sont quasiment identiques ce qui donne une idée de l'ampleur de la dégradation. Notons d'ailleurs que le fait que la distribution du bruit seul soit supérieure à celle du signal bruité pour les très forts pourcentages est un artefact dû à l'insuffisance du nombre de mesures. En effet, la distribution du signal bruité a été déterminée à partir d'une minute d'enregistrement dans les trois cas. Par contre, pour la distribution du bruit, il a fallu se contenter d'environ quatre secondes de bruit seul prises en début d'enregistrement. Compte tenu de la durée des trames utilisées pour mesurer la puissance (40 ms), cela signifie qu'il y environ 100 mesures distinctes de la puissance du bruit. Ce qui explique que la distribution obtenue soit biaisée pour les pourcentages extrêmes. Il ne faut donc pas accorder trop d'importance à l'allure de la distribution du bruit seul pour les pourcentages inférieurs à 10% ou supérieurs à 90%.

Sur la partie **C** de la figure 1.3, on constate que compte tenu de la dynamique du signal, pendant des portions importantes de l'enregistrement, la puissance du signal bruité est très supérieure à celle du bruit seul. En particulier, le chiffre d'un rapport signal-à-bruit de 24 dB, obtenu par une mesure globale, est souvent largement dépassé. Ceci constitue un point extrêmement important, car intuitivement, il semble que lorsque l'on règle un niveau d'écoute, c'est plutôt le niveau maximal du signal qui sert de guide. La question est donc de savoir comment obtenir à partir des distributions de la figure 1.3 un chiffre représentatif du niveau perçu.

[5] Attention : comme le rappelle la flèche, l'échelle de la puissance (en abscisse) n'est pas la même pour la partie **C** et pour les deux autres parties de la figure 1.3.

Ce problème du calcul de l'intensité sonore perçue a été étudié en détail par E. Zwicker. La méthode du calcul de la sonie, grandeur représentative de l'intensité sonore perçue, proposée par E. Zwicker a d'ailleurs fait l'objet d'une norme internationale [Zwicker 91a]. La référence [Zwicker 91b] rappelle les trois grandes étapes du calcul de la sonie :

1. A partir d'une représentation spectrale du signal sonore, la première étape consiste à évaluer le niveau d'excitation associé au son. Classiquement les niveaux d'excitation se représentent (et se calculent plus aisément) selon une échelle non-linéaire de fréquence, appelée taux de bande critique (graduée en Bark).

2. A partir des niveaux d'excitation, on calcule la sonie spécifique en fonction de la fréquence, en appliquant une fonction non-linéaire qui transforme le niveau d'excitation (homogène à une puissance) en une grandeur proportionnelle à l'intensité perçue. La sonie spécifique est exprimée en sones.

3. Enfin la sonie totale, s'obtient en intégrant la sonie spécifique sur tout le domaine de fréquence audible. Elle est elle aussi exprimée en sones.

Nous avons appliqué cette procédure pour les signaux de la figure 1.3, dans chaque trame de 40 ms, afin d'obtenir une distribution cumulative, analogue à celle de la figure 1.3, mais tracée en fonction de la sonie totale du signal. Dans ce calcul, l'aspect le plus complexe est la détermination des niveaux d'excitation. Dans la procédure proposée par E. Zwicker, ces niveaux sont déterminés de manière approchée en utilisant le niveau du signal mesuré par bandes de tiers d'octave [Zwicker 91a]. Nous avons utilisé la procédure décrite dans [Paillard 92] qui permet de déterminer précisément, avec un coût de calcul modéré, le niveau d'excitation engendré par un son, à partir du spectre à court-terme (voir aussi le paragraphe 3.3.1.c). Pour cette première étape, nous avons utilisé une procédure simplifié en ne tenant compte ni du facteur de transmission de l'oreille externe [Paillard 92], ni des effets de masquage temporels [Zwicker 91b]. De plus, pour l'ensemble du calcul, nous n'avons tenu compte d'aucun effet dépendant du niveau absolu du signal (seuil d'audition, modification de la relation excitation-sonie selon le niveau, etc.). Le fait de négliger ces effets, peut être toléré pour les niveaux d'écoute importants [Zwicker 91b]. Dans notre cas, il est difficile de prendre en compte ces effets puisque le niveau absolu (intensité de la pression sonore) auquel le son est écouté n'est pas fixé. La figure 1.4 représente les distributions cumulatives de la sonie, calculées dans le cas de l'enregistrement **B** de la figure 1.3. La sonie est exprimée en sones, *ici, avec une référence arbitraire* car le niveau d'écoute du signal n'est pas fixé (voir [Zwicker 91b] où [Botte 88] pour la définition du niveau équivalent à 1 sone).

L'intérêt de cette distribution cumulative de la sonie est qu'elle permet d'estimer une valeur représentative de l'intensité sonore réellement perçue par l'auditeur. D'après [Zwicker 91b], cette intensité sonore représentative correspond à l'intersection de la distribution cumulative de la sonie avec la valeur 10%, c'est à dire à l'intensité sonore qui n'est dépassée que sur 10% de l'enregistrement. Ce résultat indique que le niveau considéré comme représentatif, par la majorité des auditeurs, correspond plutôt aux fortes intensités sonores présentes sur l'enregistrement. Le tableau 1.5 présente le rapport entre l'intensité sonore du signal bruité et celle du bruit seul, mesuré au niveau correspondant à 10%, pour les trois enregistrements.

On rappelle que la sonie, exprimée en sones, est *proportionnelle* à l'intensité sonore perçue. Par exemple, pour le cas de l'enregistrement **C**, le résultat obtenu montre que l'intensité perçue du signal bruité est six fois et demi plus importante que celle du bruit seul. Ceci indique que dans ce cas, le bruit ne représente qu'une très faible partie de l'intensité sonore. Par contre, ce n'est plus le cas pour les deux autres enregistrements pour lesquels, l'intensité du bruit seul

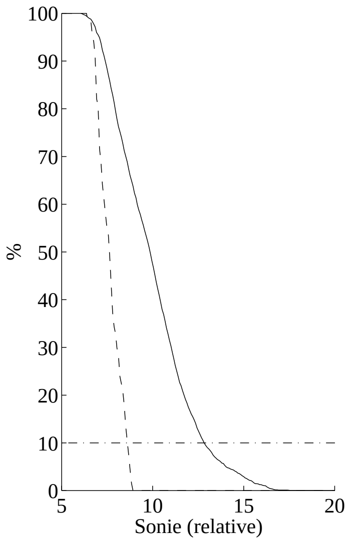

Figure 1.4: Distribution cumulative de la sonie : pourcentage des valeurs mesurées supérieures à une valeur donnée. La valeur de la sonie est exprimée en sones, *avec une référence arbitraire.* **En trait plein, distribution cumulative de la puissance du signal bruité. En tirets, distribution cumulative de la puissance du bruit seul.** Le trait mixte horizontal correspond au niveau 10%.
(le cas représenté est celui de l'enregistrement **B** de la figure 1.3)

Enregistrement A	Enregistrement B	Enregistrement C
1	1,5	6,5

Tableau 1.5: Rapport entre la sonie du signal bruité et celle du bruit seul. La sonie considérée correspond à la valeur dépassée 10% du temps.
(enregistrements de la figure 1.3)

est comparable à celle du signal bruité. Pour le cas **A**, on obtient même une intensité sonore égale pour les deux signaux. Il ne faut pas se méprendre sur le sens de cette égalité, elle signifie simplement que si le volume d'écoute est réglé à un niveau confortable pour le bruit seul, il ne sera pas nécessaire de baisser ce volume lors de l'écoute du signal bruité. Ceci indique que l'on se trouve dans un cas où le signal musical contribue peu à la sensation d'intensité sonore par rapport au bruit de fond. Sur cet exemple, on vérifie très bien cette conclusion lors de l'écoute. Cette constatation n'est pas incompatible avec le fait que le rapport signal-à-bruit global soit égal à 1 dB dans ce cas (cf. tableau 1.5). En effet, lors du calcul de la sonie, les hautes fréquences se voient attribuer un poids plus important à cause de l'élargissement des bandes critiques avec la fréquence [Botte 88] [Zwicker 81]. Un exemple bien connu de ce fait est, qu'à puissance égale, un bruit blanc est perçu comme étant beaucoup plus fort qu'un bruit dont la DSP décroît avec la fréquence [Zwicker 81]. Ici c'est surtout le bruit de fond qui bénéficie de cet effet.

Pour conclure, il faut retenir que la distribution cumulative de puissance (figure 1.3) permet d'obtenir une vision beaucoup plus détaillée du niveau de bruit qu'un simple rapport signal-à-bruit (tableau 1.5). Cependant, si l'on désire vraiment savoir quel est le niveau perçu, il est nécessaire de faire intervenir la notion d'intensité sonore (figure 1.4). Nous avons vu que l'intensité sonore du bruit présent sur un enregistrement "peu dégradé" est faible par rapport à celle du signal. Par contre, pour un enregistrement dégradé l'intensité sonore du signal bruité est en général comparable à celle du bruit seul (dans un rapport de sonie inférieur à 2).

Ceci pose un problème pratique car pour de tels enregistrements l'intensité sonore du signal traité est beaucoup plus faible que celle du signal bruité. L'évaluation du résultat est donc compliquée par le fait qu'écouter le signal original et le résultat au même niveau (physique) revient à écouter le résultat avec une intensité sonore plus faible. Or, l'impression produite par un signal donné dépend beaucoup de l'intensité sonore avec laquelle il est présenté [Moore 82] [Botte 88] [Zwicker 81]. Nous avons souvent été confrontés à ce problème lors du traitement d'enregistrements musicaux fortement dégradés. En général, nous nous sommes contentés de régler empiriquement le niveau du résultat (c'est à dire d'augmenter son volume) pour essayer d'obtenir des intensités sonores comparables. Il est clair que pour une évaluation subjective rigoureuse il serait nécessaire d'ajuster précisément cette correction de niveau sur la base de critères objectifs. Ceci pourrait ce faire, par exemple, en suivant la démarche qui a été présentée dans ce paragraphe.

1.4.5 Le bruit de fond est-il stationnaire ?

Dans notre cas, l'hypothèse d'un bruit stationnaire est très importante. En effet, pour un enregistrement musical, il est très rare de disposer de zones de "silence" (absence de signal) où la mesure des caractéristiques du bruit est possible. En pratique, il faut se contenter de mesurer la DSP du bruit en début et en fin d'enregistrement. Il est donc primordial de vérifier que les caractéristiques du bruit restent relativement stables au cours de l'enregistrement.

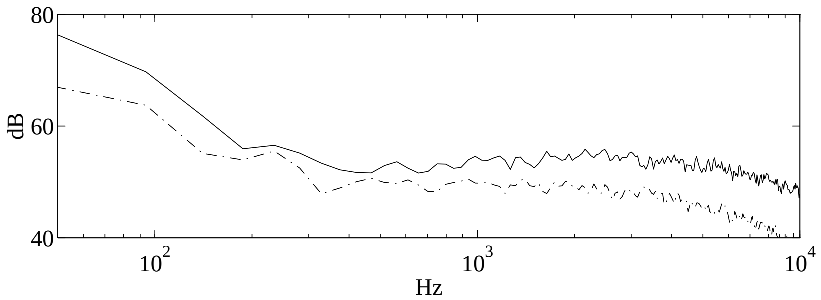

Figure 1.5: Comparaison entre la DSP du bruit de fond estimée en début d'enregistrement *(trait plein)*, et celle estimée en fin *(trait mixte)*. Les fréquences sont représentées selon une échelle logarithmique. Calcul des DSP par périodogramme moyenné sur 5 secondes de signal (soit environ 235 fenêtres de 512 points, ce qui correspond à une résolution fréquentielle de l'ordre de 50 Hz).

Un premier défaut de stationnarité, qui est apparu sur la majorité des enregistrements provenant de disques 78 tours dont nous disposions, est le fait que les caractéristiques du bruit mesurées en début et en fin diffèrent. En général, on constate que la DSP du bruit mesurée juste avant l'enregistrement est supérieure (pour toutes les fréquences) à celle mesurée en fin d'enregistrement. L'explication de ce phénomène semble liée au fait que les disques 78 tours antérieurs aux années 40 ne contiennent, en général, qu'un seul morceau de musique. C'est à dire que le début de l'enregistrement correspond aussi au bord du disque, c'est à dire la zone où vient se poser la cellule de lecture (et éventuellement les doigts du manipulateur). Il semble donc logique que les premiers sillons soient parmi les plus endommagés. Un autre élément qui vient conforter cette hypothèse est le fait que les bruits impulsionnels sont eux aussi beaucoup plus nombreux au début de l'enregistrement.

La figure 1.5 présente un exemple du type de variations observées. Sur cet exemple, on constate que l'allure générale des deux densités spectrales de puissance est identique, par contre, le niveau du bruit en début d'enregistrement est beaucoup plus fort. La différence de puissance mesurée entre le bruit en début d'enregistrement et celui en fin est proche de 8 dB. D'une manière plus générale, les variations constatées sont souvent assez faibles (écart moyen entre les DSP des deux bruits inférieur à 3 dB), par contre, dans plusieurs cas, nous avons observé des écarts moyens pouvant aller jusqu'à 10 dB. Dans ces derniers cas, ne conserver que le bruit de niveau le plus important reviendrait à surestimer fortement la puissance du bruit, c'est à dire augmenter la distorsion du signal (cf. chapitre 3), pendant une bonne partie de l'enregistrement. Une solution élémentaire dans ce type de cas consiste à mesurer les deux bruits (avant/après) et à effectuer une transition entre les deux caractéristiques spectrales mesurées selon la position dans l'enregistrement. Une telle stratégie, même si elle ne peut qu'améliorer les choses, reste très arbitraire car elle n'a aucune raison de correspondre à la variation réelle du niveau de bruit au cours de l'enregistrement.

Un autre point est que l'écoute des différents bruits de fond présents sur les enregistrements suggère que l'hypothèse de stationnarité est plus ou moins bien vérifiée selon la provenance de l'enregistrement (bande magnétique, report de disque 78t, etc.). Le bruit de disque paraît en effet beaucoup moins stable dans le temps que le souffle de bande. Afin de vérifier cette impression, on se propose de comparer l'évolution de la puissance, calculée sur une durée assez brève, pour différents types de bruits.

En pratique, cette procédure n'est pas applicable telle quelle, compte-tenu de l'allure de la densité spectrale des bruits d'enregistrement. En effet, on a vu à propos de la figure 1.1 que la DSP du bruit de fond présente en général un maximum très marqué en basse fréquence. Ceci implique que la décroissance de la fonction d'autocorrélation du bruit est extrêmement lente. Dans ces conditions, on sait que pour obtenir une estimation fiable de la puissance du bruit, il est nécessaire d'utiliser une durée d'observation très importante [Kay 93]. En pratique, pour éliminer les variations non-significatives, il faut estimer la puissance du bruit sur une durée de l'ordre de la seconde, ce qui exclut toute possibilité d'observer des variations temporelles fines.

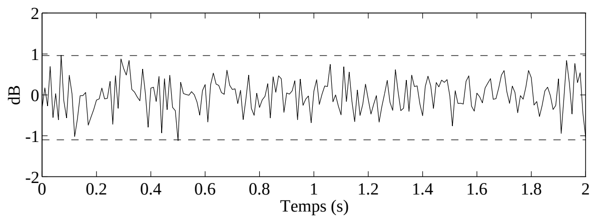

Figure 1.6: Evolution de la puissance, calculée sur des trames de 10 ms, pour un bruit rose généré numériquement. La puissance représentée est celle du bruit blanchi, normalisée par sa valeur moyenne. **En tirets,** l'intervalle de confiance à 99% correspondant au cas d'un bruit blanc gaussien.

Pour remédier à ce problème, nous avons choisi de blanchir préalablement chaque bruit en utilisant un modèle auto-régressif. Les différentes étapes de l'analyse sont donc :

1. Estimation d'un modèle AR du bruit. L'ordre du modèle est fixé à 13, et la modélisation s'effectue par la méthode de corrélation, en estimant les valeurs de la fonction d'autocorrélation sur l'échantillon de bruit complet (soit 2 ou 3 secondes de signal selon les cas).

2. Filtrage par le filtre blanchisseur $A(z)$.

3. Estimation de la puissance du signal blanchi sur des trames de 10 ms.

La figure 1.6 représente le résultat obtenu dans le cas d'un bruit rose gaussien généré numériquement (dont on peut considérer qu'il est strictement stationnaire). Sur cette figure, les traits pointillés délimitent l'intervalle de confiance à 99% pour la valeur de la puissance, *dans l'hypothèse où, le bruit, une fois blanchi, se réduit à un bruit blanc gaussien stationnaire.* Le calcul de l'intervalle de confiance est simple dans ce cas puisque l'estimateur de la puissance ($\{\sum x(n)^2\}/N$) s'obtient comme une somme de variables gaussiennes identiquement distribuées et indépendantes, il suit donc une loi du χ^2 à N degrés de libertés [Lutz 89]. Ici, compte tenu de la valeur de la fréquence d'échantillonnage, la fenêtre de 10 ms correspond à $N = 240$ échantillons. On en déduit un intervalle de confiance à 99% d'environ ± 1 dB. La figure 1.6 montre que pour le bruit rose, cet intervalle de confiance est bien respecté. Compte tenu du mode de génération de ce bruit, ce résultat n'est pas surprenant, il signifie simplement que le filtre blanchisseur utilisé est suffisamment efficace.

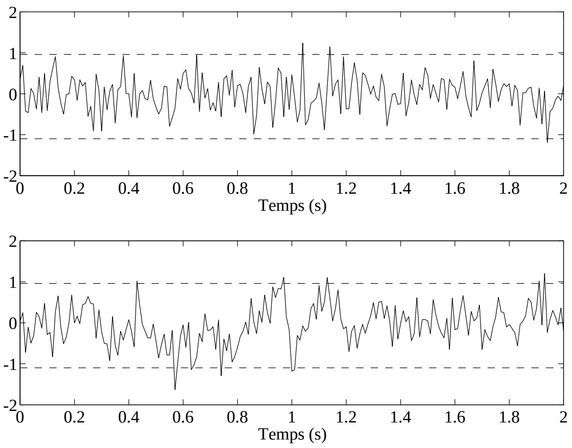

Figure 1.7: Evolution de la puissance, calculée sur des trames de 10 ms. **En haut,** bruit présent sur une cassette analogique récente. **En bas,** bruit présent sur une bande magnétique professionnelle datant des années 50. La puissance représentée est celle du bruit blanchi, normalisée par sa valeur moyenne. **En tirets,** l'intervalle de confiance à 99% correspondant au cas d'un bruit blanc gaussien.

Les cas représentés sur la figure 1.7 correspondent à des bruits de souffle présents sur des bandes analogiques. Par comparaison avec la figure 1.6, les variations de la puissance sont légèrement plus importantes dans le cas des bruits de bande. Cependant, les variations constatées restent compatibles avec l'intervalle de confiance établi dans l'hypothèse d'un bruit blanc gaussien. Ceci signifie que pour l'aspect qui nous intéresse (mesure de la DSP du bruit), il est tout à fait légitime de considérer que le bruit de souffle présent sur une bande est stationnaire. On note toutefois une différence notable entre les évolutions temporelles des deux bruits de la figure 1.7. En particulier, le bruit issu de la bande magnétique (en bas sur la figure 1.7) semble "moins blanc". C'est l'occasion de préciser un point important : ici, le rôle du filtre blanchisseur est de réduire le support de la fonction d'autocorrélation, afin de permettre une évaluation fiable de la puissance à court-terme. La figure 1.8 montre que cette mission est remplie : on vérifie que pour le signal blanchi, la durée de 10 ms est très supérieure au "support" de la fonction d'autocorrélation, ce qui n'était pas le cas pour le bruit original. Cependant, il est faux de dire que le bruit, une fois filtré par $A(z)$, se réduit à un bruit blanc. Sur la figure 1.8, la fonction d'autocorrélation du bruit filtré présente effectivement, après l'impulsion en $n = 0$, des valeurs non négligeables. L'hypothèse d'un bruit blanc gaussien est donc utilisée ici uniquement pour fournir un ordre de grandeur des variations significatives de la puissance estimée.

Pour les cas de bruits provenant de disques 78 tours, représentés sur les figures 1.9 et 1.10, la situation est très différente. L'intervalle de confiance déterminé précédemment est très largement dépassé. Ce qui s'interprète en disant que la puissance de ces bruits présente des variations significatives. Il faut souligner que parmi ces variations d'amplitude significative, les plus brèves

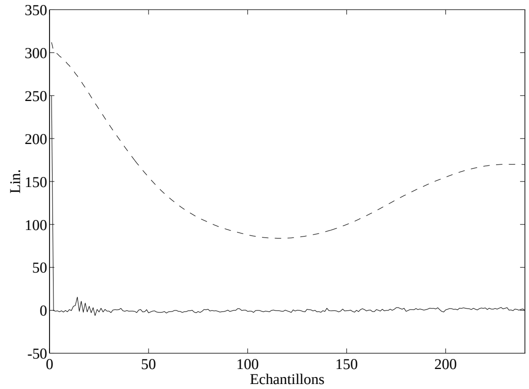

Figure 1.8: Comparaison entre la fonction d'autocorrélation estimé avant **(tirets)**, et après blanchiment **(trait plein)**. Les fonctions d'autocorrélation sont estimées sur les échantillons complets de bruit (2 secondes) pour les 240 premiers indices (soit une durée de 10 ms).

sont dues à la présence de défauts localisés dans le bruit de disque. Cependant il est intéressant de constater qu'il existe aussi des variations significatives de la puissance à plus long terme. En particulier, sur le cas de la figure 1.10 on constate des cycles d'augmentation et de diminution de la puissance, environ deux à trois fois par seconde, qui correspondent bien à la sensation auditive produite par ce type de bruit. En première approximation, on peut considérer que le bruit de disque est un signal stationnaire modulé en amplitude. Mais, avant de conclure, il resterait à refaire la même expérience en mesurant une statistique du bruit permettant de détecter une éventuelle modification de l'allure spectrale au cours du temps. Une statistique simple envisageable est, par exemple, le rapport $R_{xx}(1)/R_{xx}(0)$ qui fournit une indication sur l'allure générale de la DSP [Kay 88].

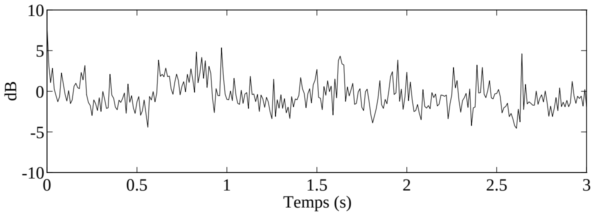

Figure 1.9: Evolution de la puissance, calculée sur des trames de 10 ms, pour un bruit provenant d'un disque 78 tours. La puissance représentée est celle du bruit blanchi, normalisée par sa valeur moyenne.

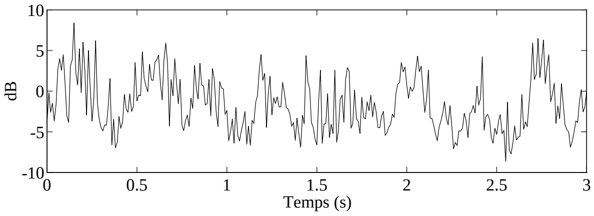

Figure 1.10: Evolution de la puissance, calculée sur des trames de 10 ms, pour un bruit provenant d'un disque 78 tours. La puissance représentée est celle du bruit blanchi, normalisée par sa valeur moyenne.

En pratique, cette possibilité de modélisation plus fine du bruit d'enregistrement, dans le cas des disques, ne nous est pas d'un grand secours car on ne dispose pas forcément de techniques adaptées, et surtout, dès que le signal musical vient se superposer, il devient impossible de suivre les variations du bruit de fond. Pour un cas tel que celui représenté sur la figure 1.9, on peut considérer qu'une fois que les bruits impulsionnels auront été éliminés, les variations du niveau de bruit de fond seront suffisamment faibles (de l'ordre de 2 à 3 dB au dessus de la valeur moyenne), pour être sans grande conséquence. Par contre, dans le cas de la figure 1.10, le niveau de bruit peut aller jusqu'à dépasser de quasiment 5 dB la valeur moyenne. Il faut donc tenir compte de cette éventualité lors du réglage des paramètres de la technique de réduction bruit. Nous verrons en particulier que cet aspect joue un rôle important vis à vis du problème du bruit résiduel présenté au paragraphe 3.2.1.

Chapitre 2

Débruitage par atténuation spectrale à court-terme

Le but de ce second chapitre est de présenter l'ensemble des techniques de réduction de bruit utilisées dans le cas qui nous intéresse : la restauration d'un enregistrement dégradé, sans sources d'information annexes, et lorsque la possibilité de paramétrisation des données est exclue.

Le point important est que la quasi-totalité des techniques proposées dans ce contexte peut être rassemblée de manière unifiée sous l'appellation d'*atténuation spectrale à court-terme*. Le paragraphe 2 présente tout d'abord le principe de fonctionnement général du débruitage par atténuation spectrale à court-terme. Une justification intuitive de l'utilisation du principe d'atténuation spectrale à court-terme pour la restauration est présentée au paragraphe 2.1.1. Par la suite, les choix particuliers propres aux principales techniques proposées dans la littérature sont détaillés aux paragraphes 2.1.3 et 2.2. Le paragraphe 2.2.3 présente une technique un peu particulière (règle de suppression de bruit dite d'Ephraim et Malah) qui présente un grand intérêt dans le cadre de la restauration d'enregistrements musicaux. Nous aurons l'occasion de revenir sur cette technique au paragraphe 4.1.2 (et dans l'annexe D). La présentation adoptée dans ce chapitre vise principalement à mettre en place le cadre général de la réduction de bruit. Il s'agit donc surtout d'une revue de différentes méthodes classiques, l'analyse des résultats obtenus n'étant pour sa part abordée qu'au chapitre suivant.

Enfin, compte tenu du rôle important tenu par la transformée de Fourier à court-terme dans la grande majorité des systèmes de débruitage par atténuation spectrale à court-terme, le paragraphe 2.3 et l'annexe B (pour les aspects plus standards) sont entièrement consacrés aux propriétés de l'ensemble analyse/modifications/synthèse dans le cadre de la transformée de Fourier à court-terme (notée TFCT). Le paragraphe 2.3.2 fournit une description de l'effet, *sur le signal*, d'une modification spectrale effectuée sur la TFCT. A notre connaissance la présentation adoptée dans ce paragraphe est assez originale et s'avère plus utile, dans le cas du débruitage, que la présentation "classique" détaillée au paragraphe B.2 (annexe B).

2.1 Présentation

D'après ce qui a été dit au chapitre 1, la réduction du bruit de fond présent sur un enregistrement dégradé doit se faire dans notre cas uniquement à partir de la donnée de l'enregistrement bruité. La figure 2.1 décrit les notations qui seront utilisées dans toute la suite du document : $x(n)$ représente le signal audio bruité et $d(n)$ le bruit supposé additif. Il faut noter que sur cette figure, toute la partie concernant la dégradation du signal, qui est représentée en pointillés, nous est physiquement inaccessible (cf. paragraphe 1.1). C'est à dire qu'il n'existe pas de source d'information auxiliaire qui pourrait permettre de modéliser le comportement temporel du bruit additif ou du signal inconnu. On suppose tout de même qu'il est possible de mesurer les caractéristiques statistiques du bruit de fond. Pour ce faire, on admet que le début ou la fin de l'enregistrement contient un intervalle de temps où seul le bruit est présent, et suffisamment long pour permettre une estimation correcte de la densité spectrale de puissance du bruit (cf. paragraphe 1.4.3).

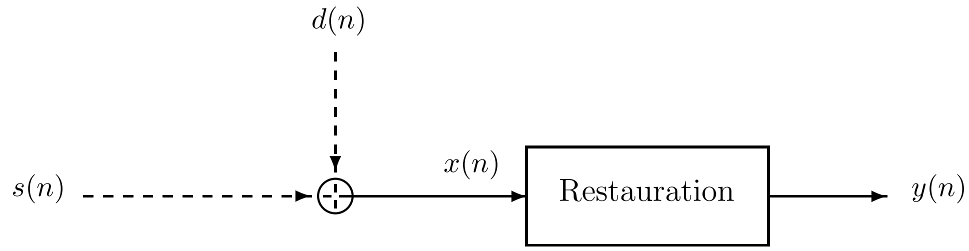

Figure 2.1: Notations utilisées dans le document.

2.1.1 Principe du traitement

Dans ces conditions, les méthodes de traitement utilisent généralement le principe **d'atténuation spectrale à court-terme** (voir [Bourdier 88], [Lim 79] ou [Lim 86] pour une revue de la littérature sur ce sujet). La meilleure description du fonctionnement général de ce type de méthodes est donnée par Mac Aulay et Mallpas dans [Mc Aulay 80]. Une fois traduite en français celle-ci s'énonce à peu près ainsi : "La technique consiste à effectuer une décomposition spectrale d'une trame de signal bruité. Puis, chaque canal du spectre est atténué selon que le niveau mesuré localement, dans le canal, dépasse plus ou moins l'estimation du bruit de fond". La fonction qui permet de déterminer l'atténuation, étant donné la mesure du niveau de puissance ainsi que l'estimation du bruit de fond, constitue la **règle de suppression** (sous-entendu "de bruit"). Le signal débruité $y(n)$ est ensuite obtenu par la transformation spectrale à court-terme inverse. Le schéma de principe d'une telle méthode est représenté sur la figure 2.2. Dans un cadre plus général, il serait utile d'inclure une procédure supplémentaire permettant de détecter en cours de traitement l'absence de signal (voir par exemple le système présenté dans [Mc Aulay 80]). Un intervalle de "silence" (où seul le bruit est présent) peut en effet être utilisé pour actualiser la mesure de la densité spectrale du bruit, ce qui permet de s'adapter aux éventuelles variations de la nature du bruit de fond. Cet ajout s'avère fort efficace dans le cas de la parole où les instants de silence sont assez fréquents. Malheureusement, pour un enregistrement de musique il n'existe en général pas de silence pendant un morceau, il faut donc se contenter de mesurer, avant le traitement, le bruit situé entre deux plages enregistrées (cf. paragraphe 1.4.3).

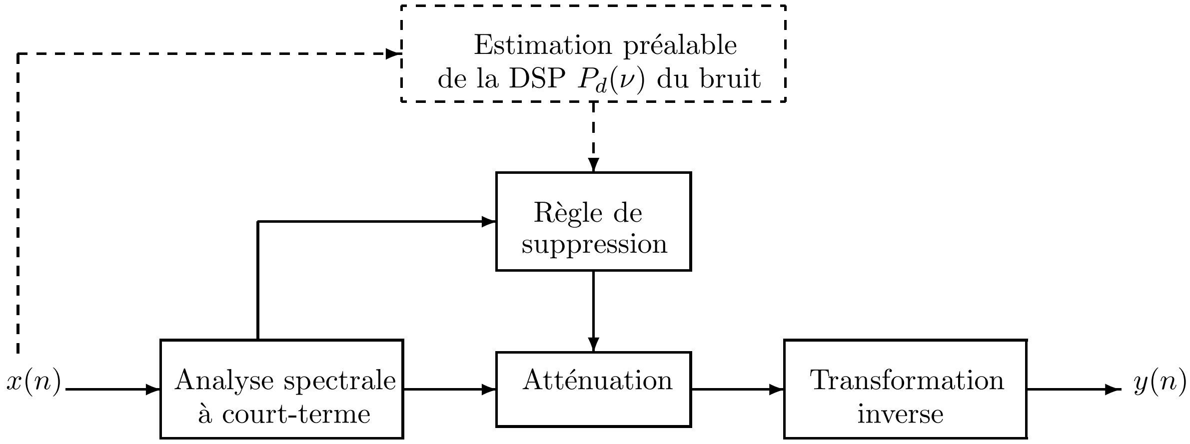

Figure 2.2: Débruitage par atténuation spectrale à court-terme. La partie du schéma représentée en pointillés se fait au préalable à partir d'une portion représentative du bruit $d(n)$. Toutes les opérations représentées en trait plein se font pour chaque fenêtre de signal à court-terme.

Les deux points importants qui vont déterminer l'ensemble des caractéristiques d'une telle méthode sont :

La transformation à court-terme c'est à dire une transformation possédant une interprétation fréquentielle dans laquelle seule une portion temporelle du signal est analysée (on parle de trame de signal analysé).

La règle de suppression ou plus généralement l'ensemble des mécanismes mis en œuvre pour calculer l'atténuation à apporter à chaque canal fréquentiel de la transformée à court-terme (l'atténuation désignant un gain réel compris entre 0 et 1).

En général, la transformation utilisée est la Transformée de Fourier à Court-Terme (notée TFCT) tandis que la règle de suppression vise à obtenir, après application de l'atténuation, une estimation de l'amplitude du spectre à court-terme du signal non bruité (d'où le terme souvent utilisé de **méthodes d'estimation de l'amplitude spectrale à court-terme** [Vary 85]). En effet, comme on applique une atténuation (réelle) à la représentation spectrale à court-terme (qui elle est en général complexe), seule son amplitude est modifiée.

Nous reviendrons en détail aux paragraphes 2.1.3 et 2.2 sur les différents aspects de chacun de ces deux points. Mais il est intéressant de justifier (au moins intuitivement) dès maintenant le choix quasi-systématique de ce type de méthode dans le cadre qui nous intéresse. En général, dans la littérature, le fait que la perception auditive humaine soit essentiellement guidée par le spectre d'amplitude à court-terme est utilisé comme un postulat de départ à partir duquel la technique est élaborée [Lim 79] [Ephraim 84]. L'article de Wang [Wang 82] apporte d'ailleurs une confirmation expérimentale du fait que la connaissance de la phase à court-terme n'apporte rien pour ce type de traitement dans le cas de signaux de parole bruitée[1]. Indépendamment de cette question, on peut se demander, d'une manière plus générale, s'il est bien utile de faire intervenir ici des connaissances sur la perception auditive : d'une part, on sait que l'analogie entre le fonctionnement du système auditif et un simple spectrogramme du signal acoustique est assez réductrice [Moore 82], d'autre part, il semble bien qu'à ce stade le problème posé soit

[1] On reviendra sur ce point au paragraphe 3.3 pour montrer que ce résultat dépend du type de signal traité. En particulier, cette conclusion est fausse dès lors que le signal à restaurer est stationnaire sur une durée bien supérieure à la durée de trame.

essentiellement un problème de traitement de signal qui, même en excluant la question de la perception des résultats, n'est pas forcément simple à résoudre.

2.1.2 Une justification de la méthode

Le but de ce qui suit est de fournir certains éléments qui permettent de justifier le recours à l'atténuation spectrale sans faire intervenir de connaissance a priori sur le fonctionnement du système auditif humain. Il ne s'agit pas de démontrer que l'atténuation spectrale est la seule approche possible du problème mais plutôt d'indiquer comment, en partant d'une méthode très générale, on est conduit plus ou moins naturellement au principe d'atténuation spectrale à court-terme. Cet exemple fournit des éléments concernant l'origine de certaines des caractéristiques de la méthode (atténuation réelle, taille des trames à court-terme, type de signaux traités).

Supposons dans un premier temps que le signal $s(n)$ ainsi que le bruit $d(n)$ soient tous deux stationnaires. L'effet du bruit est supposé être additif, non-corrélé avec le signal. La seule connaissance dont on dispose est celle de la densité spectrale du bruit $P_d(\nu)$ mesurée au préalable. Nous allons chercher à estimer la forme d'onde du signal musical non bruité $s(n)$. A cette fin, nous disposons ici d'un élément supplémentaire : il est possible d'estimer la densité spectrale de puissance du signal inconnu $P_s(\nu)$ par la relation

$$P_s(\nu) = P_x(\nu) - P_d(\nu)$$

où $P_x(\nu)$ est la DSP du signal bruité qui est observable. La relation ci-dessus est valable du fait de la décorrélation entre le signal et le bruit. Pour prendre en compte cette information supplémentaire concernant la quantité à estimer, on a intérêt à se placer dans le cadre de l'estimation bayésienne. La DSP du signal à estimer est donc considérée comme une quantité connue *a priori*. La meilleure stratégie d'estimation (parmi les stratégies usuelles), applicable dans ce cas, consiste à estimer le signal $s(n)$ par un traitement linéaire en minimisant l'erreur quadratique moyenne d'estimation [Kay 93, Fig. 14.1]. Cette stratégie conduit à appliquer le filtre dit de Wiener au signal $x(n)$ afin d'obtenir le signal estimé $\hat{s}(n)$. La réponse fréquentielle du filtre de Wiener est donné par [Kay 93, §12] :

$$G_W(\nu) = \frac{P_s(\nu)}{P_s(\nu) + P_d(\nu)} \tag{2.1}$$

Comme la densité spectrale de puissance du signal bruité peut s'écrire

$$P_x(\nu) = P_s(\nu) + P_d(\nu)$$

le gain du filtre de Wiener s'exprime aussi sous la forme

$$G_W(\nu) = 1 - \frac{P_d(\nu)}{P_x(\nu)} \tag{2.2}$$

C'est à dire que le gain de Wiener réalise bien une atténuation fréquentielle, au sens où nous l'avons définie précédemment, en fonction du rapport $P_d(\nu)/P_x(\nu)$. La méthode d'atténuation spectrale est inspirée directement de ce principe. Le fait que le gain du filtre à appliquer soit réel provient donc fondamentalement de l'hypothèse que toute l'information disponible concernant la dégradation est contenue dans la densité spectrale de puissance $P_d(\nu)$. Ce qui revient à dire qu'on ne dispose pas d'information concernant la forme temporelle du bruit : on ne dispose que de sa fonction d'autocorrélation.

Afin de quantifier l'amélioration qui est susceptible d'être apportée par cette approche, on s'intéresse à la puissance de l'erreur d'estimation résultant de l'application du filtre de Wiener. Celle-ci s'écrit

$$\epsilon_W^2 = \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{P_s(\nu)P_d(\nu)}{P_s(\nu)+P_d(\nu)} d\nu \tag{2.3}$$

Considérons les cas limites suivants :

- On constate que si $P_s(\nu) >> P_d(\nu)$ pour toutes les fréquences, alors la puissance de l'erreur vaut

  $$\epsilon_W^2 = \int_{-\frac{1}{2}}^{\frac{1}{2}} P_d(\nu) d\nu$$

  C'est à dire que *la puissance de l'erreur d'estimation est alors égale à celle du bruit additif.* Ce qui correspond au fait que dans ce cas, l'application du filtrage de Wiener n'apporte aucune modification au signal bruité. Ce qui revient à dire que le filtre de Wiener est alors un filtre passe-tout.

- A l'opposé, quand $P_s(\nu) << P_d(\nu)$ pour toutes les fréquences, alors la puissance de l'erreur vaut

  $$\epsilon_W^2 = \int_{-\frac{1}{2}}^{\frac{1}{2}} P_s(\nu) d\nu$$

  La puissance de l'erreur d'estimation est maintenant égale à celle du signal : le filtrage de Wiener effectue alors une "mise à zéro" du bruit ainsi que du signal. On parle alors de *distorsion du signal.*

- Enfin, l'erreur d'estimation est nulle uniquement dans le cas où $P_s(\nu)$ et $P_d(\nu)$ possèdent des supports disjoints.

Le raisonnement précédent peut être appliqué de la même façon pour une bande de fréquence limitée. En effet, l'erreur d'estimation filtrée par un filtre passe-bande idéal de bande passante $[\nu_1, \nu_2]$ s'écrit simplement

$$\epsilon_{W,[\nu_1,\nu_2]}^2 = \int_{\nu_1}^{\nu_2} \frac{P_s(\nu)P_d(\nu)}{P_s(\nu)+P_d(\nu)} d\nu$$

Le filtrage de Wiener apporte donc d'autant plus d'amélioration (c'est à dire une erreur d'estimation de plus faible puissance) qu'il existe des bandes de fréquence où $P_s(\nu) << P_d(\nu)$. La condition de bon fonctionnement de l'approche du filtrage de Wiener est donc l'existence de zones fréquentielles où la puissance du signal est quasiment nulle. De plus, si la puissance du signal n'est pas rigoureusement nulle dans ces zones fréquentielles, l'amélioration du rapport signal-à-bruit se fait moyennant une certaine distorsion du signal. Dans le cas contraire, où la puissance du signal est toujours supérieure à celle du bruit, le filtrage de Wiener n'apporte pas de diminution de la puissance du bruit.

En pratique les signaux qui sont traités ne sont pas des signaux stationnaires, il est donc nécessaire de modifier la méthode qui vient d'être décrite. L'extension usuelle de l'approche du filtrage de Wiener au cas des signaux non-stationnaires est le filtrage de Kalman [Kay 93, §13]. Malheureusement, cette approche n'est pas utilisable ici car elle suppose un modèle temporel de l'évolution du signal à estimer, possibilité que nous avons exclue (cf. paragraphe 1.3.2). Par conséquent, la solution utilisée consiste à appliquer le principe du filtrage de Wiener en se restreignant à des intervalles (trames à court-terme) sur lesquelles le signal peut être considéré comme stationnaire. On retrouve là exactement le principe de l'atténuation spectrale à court-terme.

L'intérêt de cette présentation de la méthode est essentiellement de souligner que le choix de la durée des trames à court-terme est guidé par la nature des signaux à traiter, et non par d'éventuelles considérations sur l'audition. De plus, nous aurons l'occasion de voir au chapitre 3 que les remarques formulées sur le bon fonctionnement du filtre de Wiener (existence de zones spectrales ne contenant pas de signal) restent valables, d'une manière générale, pour l'atténuation spectrale à court-terme.

2.1.3 Transformation à court-terme

Dans la littérature sur le débruitage, la transformée la plus utilisée est la transformée de Fourier à court-terme (voir les systèmes présentés dans [Lim 79] [Ephraim 84] [Boll 79]), ou son implémentation sous forme de banc de filtres polyphase[2] dans [Vary 85]. On trouve aussi quelques exemples d'utilisation de bancs de filtres non uniformes, c'est à dire où la largeur de bande des filtres dépend de leur position fréquentielle [Mc Aulay 80] [Moorer 86] [Petersen 81].

2.1.3.a Ensemble des transformations à court-terme utilisables

Il est intéressant de caractériser l'ensemble des transformations fréquentielles qui sont susceptibles d'être utilisées dans une application de débruitage à court-terme. D'après le schéma de la figure 2.2, les conditions à remplir sont essentiellement :

1. L'existence d'une transformation inverse qui permet de retrouver sans équivoque le signal analysé (on parle de **procédure d'analyse/synthèse**).
2. La possibilité d'effectuer des modifications dans le domaine spectral.

Ces deux conditions permettent d'éliminer bon nombre de transformations fréquentielles à court-terme. La première condition (1) implique qu'il existe une transformation reliant le signal à sa transformée. Cette condition rend a priori difficile l'utilisation d'une représentation temps-fréquence ne possédant qu'une interprétation énergétique (comme la transformée de Wigner-Ville) [Cohen 89]. Par contre, il est possible d'utiliser une représentation temps-fréquence qui peut s'interpréter sous la forme de la sortie d'un banc de filtres (par exemple une transformée en ondelettes "continue") [Flandrin 89].

La condition (2) vient du fait que le débruitage, c'est à dire l'atténuation de certaines composantes de signal, se fait dans le domaine de la transformée à court-terme. Idéalement, il faudrait être capable d'appliquer une atténuation indépendamment sur chaque composante spectrale au sein d'une même transformée, ainsi qu'indépendamment d'une trame à l'autre. Cette indépendance garantit que la mise en œuvre du débruitage est toujours possible quelle que soit l'atténuation spectrale obtenue par application de la règle de suppression. Il est donc nécessaire de garantir que des points distincts de la transformée à court-terme puissent être modifiés de manière indépendante. On montre que ceci conduit à accepter une certaine *redondance au niveau de la transformation à court-terme.*

[2] Une transformée en bloc, telle que la TFCT, possède une interprétation équivalente sous la forme d'un banc de filtres multi-cadence [Crochiere 83] [Malvar 92]. Pour la TFCT, la formulation sous forme de banc de filtres est plus simple lorsque le nombre de bandes, qui est égal à la longueur du bloc de transformée, est faible (en gros inférieur à 64). L'implémentation efficace se fait alors grâce à l'utilisation des composantes polyphasées du banc de filtres [Crochiere 83][Bellanger 76].

En pratique, ceci veut dire qu'un banc de filtres à décimation maximale ne pourra pas être utilisé pour l'application de débruitage. Le terme "décimation maximale" correspond au fait que chaque voie du banc de filtres est échantillonnée à la fréquence minimale permise par le théorème d'échantillonnage appliqué aux signaux à bande étroite. En contrepartie, les voies d'un tel banc de filtres contiennent des termes dus au repliement spectral. Ces termes ne disparaissent que lors de la synthèse [Crochiere 83]. Cette catégorie comprend des bancs de filtres à structure arborescente comme les bancs de filtres QMF [Vaidyanathan 87] ou les ondelettes discrètes à structure dyadique [Daubechies 88], ainsi que des bancs de filtres uniformes comme les *lapped transforms* [Malvar 92]. Tous ces bancs de filtres sont très intéressants pour le codage car ils minimisent la quantité de données à transmettre [Masson 87] [Crochiere 83]. Cependant l'existence d'un fort repliement spectral interdit de modifier chaque bande indépendamment, ce qui est en contradiction avec le principe de l'atténuation spectrale à court-terme. Toutefois, il faut noter que l'impossibilité d'utiliser de tels bancs de filtres pour le débruitage ne constitue une restriction que du point de vue de l'efficacité de l'implémentation. En effet, il est toujours possible de réaliser un banc de filtres à décimation non maximale possédant les mêmes caractéristiques. Un banc de filtres à structure arborescente comme les ondelettes discrètes peut, par exemple, être réalisé sans le sous-échantillonnage, en utilisant les réponses impulsionnelles équivalentes de chacune des voies. La différence se situe au niveau de l'efficacité de l'implémentation : dans le cas du filtre à structure arborescente, le coût de calcul et le volume de données à traiter de la version non sous-échantillonnée deviennent rapidement prohibitif dès que le nombre de bandes augmente.

2.1.3.b Caractéristiques fréquentielles de la transformée

D'une manière un peu schématique, il est possible de résumer les caractéristiques fréquentielles des transformations à court-terme utilisées pour le débruitage en deux types :

Largeur de bande uniforme Conformément à ce qui a été dit, c'est le type de transformation le plus utilisé pour le débruitage par atténuation spectrale à court-terme. Sa réalisation se fait sous la forme de la transformée de Fourier à court-terme. De plus, dans la plupart des systèmes de restauration spécifiquement destinés aux enregistrements musicaux, c'est la solution qui a été retenue [Lagadec 83] [Bourdier 88] [Valiere 91] [Vaseghi 88b].

Largeur de bande non-uniforme Ici les modes de réalisation rencontrés sont plus variés : groupement de bandes issues d'un banc de filtres TFCT dans [Moorer 86], ou spécification directe des caractéristiques des filtres dans [Mc Aulay 80] (où le nombre de bandes est faible). Dans le cas particulier où l'on désire que la largeur de bande varie proportionnellement à la fréquence centrale (on parle aussi de banc de filtres "à Q constant"), une possibilité intéressante consiste à utiliser une transformation en ondelettes continue [Flandrin 89] pour laquelle des algorithmes rapides existent [Rioul 92] [Combes 89] ainsi que des conditions approchées d'analyse/synthèse [Daubechies 90] [Kronland Martinet 87]. A ce propos, il faut noter que le système proposé en 1981 par Petersen et Boll dans [Petersen 81] utilise une transformation à Q constant pour simuler la perception auditive. Cette transformation décrite dans l'article [Petersen 83] s'avère en fait être une transformation en ondelettes particulière (à la normalisation près).

Une remarque importante est que parmi les systèmes de restauration décrits ci-dessus, ceux qui utilisent une transformation à largeur de bande non uniforme possèdent toujours un très faible nombre de bandes (par exemple 23 bandes pour couvrir la bande [0,4 kHz] dans [Petersen 81]). Ceci est très certainement lié aux difficultés de réalisation d'un tel système. Il devient par exemple difficile d'utiliser la spécification directe de tous les filtres du banc lorsque le nombre de voies est important. Cependant, les auteurs de [Moorer 86] avancent le fait que l'écoute des résultats du traitement indique que les seuls systèmes intéressants sont ceux qui utilisent, soit un très faible nombre de bandes se recouvrant fortement (par exemple 4 bandes non-uniformes), soit un grand nombre de bandes (256 et plus). Entre ces deux solutions, la plus efficace semble être l'utilisation d'un grand nombre de bandes (voir aussi [Lagadec 83]). Cette constatation résume bien les deux tendances présentes dans la littérature :

- Faible nombre de bandes avec, en général, référence à des propriétés auditives ([Petersen 81] ou [Moorer 86]).

- Grand nombre (au moins 256) de bandes uniformes (dans la plupart des autres publications déjà citées).

Pour terminer sur ce point, notons que J-C. Valière indique que dans le cas d'un système de restauration spécifiquement destiné aux enregistrement musicaux, il est préférable d'utiliser une longueur de fenêtre (de TFCT) supérieure à celle couramment utilisée pour le traitement de la parole [Valiere 91]. De manière équivalente ceci conduit à utiliser, à fréquence d'échantillonnage égale, un nombre de bandes supérieur. Cette constatation semble confirmer l'intérêt d'un système possédant un grand nombre de bandes dans la cadre de la restauration d'enregistrement musicaux.

2.1.3.c Caractéristiques temporelles de la transformée

La nature des filtres utilisés a aussi des conséquences sur les aspects temporels. D'une manière générale, un filtre très sélectif ne peut pas avoir un "temps de réponse" aussi court qu'un filtre peu sélectif (la définition mathématique de la durée d'un signal ainsi que la démonstration du principe d'incertitude sont détaillées dans [Papoulis 62]). La durée des filtres réalisées par la transformée de Fourier à court-terme semble avoir des répercussions lorsque l'enregistrement traité contient des sons qui présentent des transitoires brefs [Valiere 90]. Une "durée" trop longue (50 ms et plus) semble provoquer un étalement des transitoires. D'autre part, la question se pose de savoir si il est utile ou non que ces caractéristiques temporelles varient selon la bande de fréquence considérée. Cette discussion est d'autant plus importante, qu'il est de fait possible de réaliser des systèmes hybrides qui se situent entre les deux types de transformations évoquées précédemment. Ainsi le système décrit dans [Valiere 91] utilise un banc de filtres à structure arborescente dyadique pour diviser grossièrement le domaine fréquentiel audible en un faible nombre de bandes (à Q constant). Puis les signaux issus de chaque bande sont traités séparément par atténuation spectrale en utilisant la TFCT (donc avec une largeur de bande uniforme).

Dans la suite du document, la démarche qui a été adoptée consiste à étudier les propriétés objectives du débruitage par atténuation spectrale dans le cas des systèmes à largeur de bande uniforme, et donc de leur réalisation grâce à la TFCT. L'idée étant qu'il est possible de généraliser la plupart des résultats obtenus dans le cas des systèmes à largeur de bande uniforme en considérant les systèmes non-uniformes comme "localement uniformes" (c'est à dire qu'au voisinage d'une fréquence considérée, on admet que la variation de la résolution fréquentielle du système est faible). Nous verrons au chapitre 3 que cette démarche permet effectivement de répondre à plusieurs des questions qui ont été soulevées ici.

2.2 Règles de suppression

2.2.1 Principe général de la suppression de bruit

Sur le schéma de principe de la figure 2.2, la règle de suppression désigne le mécanisme qui permet, pour chaque trame à court-terme, de décider de l'atténuation à apporter à chaque canal fréquentiel de la transformée. La décision s'effectue à partir de deux éléments : d'une part, une *estimation de la densité spectrale de puissance du bruit de fond* $\hat{P}_d(\omega_k)$, et d'autre part, une *estimation "locale" de la densité spectrale de puissance du signal présent dans la trame à court-terme*, notée $\hat{P}_x(p,\omega_k)$, où p désigne l'indice temporel de la trame à court-terme courante.

Peter Vary note à juste titre dans [Vary 85] la différence fondamentale qui existe entre ces deux estimations : $\hat{P}_x(p,\omega_k)$ est estimée à partir des données de la trame à court-terme, tandis que $\hat{P}_d(\omega_k)$ a été estimée au préalable à partir d'un très grand nombre de données. Un ordre de grandeur typique pour la durée de la trame à court-terme est de 50 ms, tandis que la durée d'un "silence", en début ou en fin d'enregistrement, utilisé pour estimer $\hat{P}_d(\omega_k)$ est en général de l'ordre de la seconde (cf. paragraphe 1.4.3). C'est à dire que $\hat{P}_d(\omega_k)$ est estimée à partir d'une observation au moins 10 fois plus longue que la trame à court-terme utilisée pour estimer $\hat{P}_x(p,\omega_k)$. La conséquence est que la variance de l'estimation de la DSP du bruit est négligeable par rapport à celle de $\hat{P}_x(p,\omega_k)$. Dans la suite, on considère donc toujours que la DSP du bruit de fond est connue (où plutôt que l'approximation $\hat{P}_d(\omega_k) \approx P_d(\omega_k)$ est valable). Ceci, suppose tout de même que l'estimation de la DSP du bruit a pu être réalisée dans de bonnes conditions (cf. paragraphe 1.4.3). Par contre, le paragraphe 3.2 montre, que, compte tenu du type d'estimateur utilisé, la variance de $\hat{P}_x(p,\omega_k)$ a des conséquences non négligeables.

On note $G(p,\omega_k)$ l'atténuation spectrale apportée dans la fenêtre à court-terme d'indice p au point fréquentiel ω_k. En fait $G(p,\omega_k)$ est un nombre réel compris entre 0 et 1, c'est à dire qu'il représente le gain apporté à chaque valeur de la TFCT. Etant donné que ce gain est toujours inférieur à 1, la quantité $G(p,\omega_k)$ est souvent qualifiée, par abus de langage, d'atténuation. Toutes les règles de suppression proposées dans la littérature vérifient le principe suivant :

$$\begin{cases} G(p,\omega_k) & = & 1 & \text{quand} & \hat{P}_x(p,\omega_k) \gg \hat{P}_d(\omega_k) \\ G(p,\omega_k) & = & 0 & \text{quand} & \hat{P}_x(p,\omega_k) \leq \hat{P}_d(\omega_k) \end{cases} \tag{2.4}$$

Ce principe correspond à l'idée très intuitive d'un débruitage effectué à partir d'un banc de filtres à gains variables (voir à ce propos la présentation adoptée dans [Moorer 86] ou [Lagadec 83]). En effet, quand la puissance du signal bruité, mesurée dans un canal fréquentiel, est très supérieure à celle du bruit de fond (mesurée dans le même canal), on peut en conclure que la sous-bande considérée contient une composante de signal de niveau non négligeable, qui ne doit donc pas être modifiée. Inversement, lorsque $\hat{P}_x(p,\omega_k)$ est proche de $\hat{P}_d(\omega_k)$, la sous-bande ne contient pratiquement plus que du bruit filtré, d'où la nécessité d'appliquer une forte atténuation.

Mesure du niveau du signal Afin de préciser le fonctionnement des différentes règles de suppression de bruit, on introduit deux grandeurs tout a fait équivalentes qui permettent de quantifier les propriétés énergétiques locales du signal :

Niveau relatif local Sous-entendu du signal bruité par rapport au bruit de fond, défini par

$$\mathcal{Q}(p,\omega_k) = \frac{\hat{P}_x(p,\omega_k)}{\hat{P}_d(\omega_k)} \tag{2.5}$$

Ce niveau relatif est dit local car il ne concerne qu'une trame à court-terme donnée dans un canal fréquentiel particulier.

Rapport signal-à-bruit (ou RSB) local Qui est défini traditionnellement à l'aide du signal non bruité par

$$\mathcal{R}(p,\omega_k) = \frac{\hat{P}_s(p,\omega_k)}{\hat{P}_d(\omega_k)} \tag{2.6}$$

Malheureusement celui-ci est impossible à mesurer directement en pratique. Cependant, en remarquant que si le bruit additif $d(n)$ est non-corrélé avec le signal $s(n)$, alors la quantité $(\hat{P}_x(p,\omega_k) - \hat{P}_d(\omega_k))$ constitue une estimation non biaisée de $\hat{P}_s(p,\omega_k)$, le rapport signal-à-bruit local peut aussi être défini sous la forme

$$\mathcal{R}(p,\omega_k) = \frac{\hat{P}_x(p,\omega_k) - \hat{P}_d(\omega_k)}{\hat{P}_d(\omega_k)}$$

soit, de manière équivalente

$$\mathcal{R}(p,\omega_k) = \mathcal{Q}(p,\omega_k) - 1 \tag{2.7}$$

L'équation (2.7) souligne l'équivalence qui existe entre le niveau relatif et le rapport signal-à-bruit. Du fait de la non-corrélation du signal et du bruit, *la valeur moyenne* du niveau relatif est toujours supérieur à 1. Cependant le niveau relatif mesuré localement peut être inférieur à 1 du fait de la variance de l'estimation $\hat{P}_x(p,\omega_k)$. Dans la suite du document, dès lors qu'il s'agit de grandeurs mesurées localement (au sein d'une trame à court-terme), c'est plutôt le niveau relatif qui est employé car l'interprétation de $\mathcal{R}(p,\omega_k)$, calculé par la relation (2.7), est plus difficile. En effet, la notion de rapport signal-à-bruit perd de son intérêt dès lors que l'on est pas assuré d'obtenir une valeur positive.

Une manière commode de décrire le fonctionnement d'une règle de suppression sous forme graphique consiste à représenter la *caractéristique de suppression* qui relie l'atténuation spectrale au niveau relatif local. Compte tenu du principe général de fonctionnement des règles de suppression décrit par les relations (2.4), on vérifie que les caractéristiques de suppression présentent toujours une allure similaire à celles de la figure 2.3 (page 49).

Choix d'un estimateur spectral Ce dernier point concerne l'estimation spectrale à court-terme du signal bruité $\hat{P}_x(p,\omega_k)$: dans toutes les publications concernant le débruitage par atténuation spectrale à court-terme [Lim 79] [Boll 79] [Ephraim 84], celle-ci est réalisée par le *périodogramme* (module au carré de la transformée de Fourier). L'estimation de la densité spectrale de puissance du signal bruité s'écrit donc

$$\hat{P}_x(p,\omega_k) = |X(p,\omega_k)|^2 \tag{2.8}$$

On ne prend pas en compte ici la constante de normalisation qui assure l'homogénéité avec une densité spectrale de puissance [Brillinger 81]. Il est important de distinguer deux quantités différentes : d'une part la TFCT $X(p,\omega_k)$ qui est modifiée, et d'autre part, l'estimation spectrale locale $\hat{P}_x(p,\omega_k)$ qui est utilisée pour déterminer le gain $G(p,\omega_k)$. Il est possible d'utiliser différentes méthodes pour estimer $\hat{P}_x(p,\omega_k)$, par contre, c'est forcément $X(p,\omega_k)$ qui doit être modifiée. L'utilisation de la relation (2.8) permet d'aboutir, pour certaines règles de suppression, à des expressions beaucoup plus compactes [Boll 79], mais qui ont le défaut de ne pas mettre en évidence ces deux quantités distinctes.

Dans la suite du document, c'est, sauf indication contraire, le périodogramme qui est utilisé pour estimer $\hat{P}_x(p,\omega_k)$. En négligeant la variance de l'estimation de la DSP du bruit de fond, nous utiliserons donc le niveau relatif local sous la forme suivante :

$$\mathcal{Q}(p, \omega_k) = \frac{|X(p, \omega_k)|^2}{\hat{P}_d(\omega_k)} \approx \frac{|X(p, \omega_k)|^2}{\mathrm{E}\left\{|D(p, \omega_k)|^2\right\}} \tag{2.9}$$

Il faut noter que la quantité qui figure au dénominateur de cette expression est bien indépendante de l'indice de la trame à court-terme p du fait de la stationnarité du bruit $d(n)$ (cf. annexe C).

Afin de classer les différentes règles de suppression proposées dans la littérature, nous avons décidé de distinguer deux types de règles. Les *règles de suppression ponctuelles* correspondent au cas où l'atténuation à court-terme $G(p, \omega_k)$ ne dépend que $\mathcal{Q}(p, \omega_k)$. Ce qui revient à dire que l'atténuation à apporter à la TFCT est évaluée *de manière indépendante* pour chaque canal fréquentiel, et dans chaque trame à court-terme. Par contre, la règle de suppression proposée par Ephraim et Malah dans [Ephraim 83] puis détaillée dans [Ephraim 84], ainsi que celles s'inspirant du même principe dans [Ephraim 84] et [Ephraim 85], fonctionnent différemment. En particulier, on montre au paragraphe 2.2.3 que les valeurs de $G(p, \omega_k)$ et de $G(p+1, \omega_k)$ sont fortement liées. Ce dernier type de règle de suppression est en conséquence désigné par le terme générique de *règle de suppression moyennée* (sous entendu, sur l'indice temporel p). Le but de cette présentation des règles de suppression les plus courantes est de mettre en évidence les aspects importants qui seront utilisés par la suite lors de l'évaluation de la qualité du débruitage. Pour une présentation plus détaillée (en particulier sur l'aspect chronologique) des différentes règles de suppression, on peut se reporter à [Lim 79], [Bourdier 88] ou [Vary 85].

2.2.2 Règles de suppression ponctuelles

Parmi les règles de suppression couramment utilisées, certaines n'ont pas de fondement théorique précis mais correspondent néanmoins au principe général de l'équation (2.4). Dans ce cas, le réglage des différents paramètres se fait en général de manière totalement empirique par des tests d'écoute [Moorer 86] [Lim 79]. On ne s'attardera pas ici sur ces règles de suppression qui sont par définition assez difficiles à évaluer.

La plupart des autres règles de suppression sont justifiées, dans un formalisme statistique, par l'estimation d'une caractéristique du signal non bruité. En général, il s'agit du spectre d'amplitude du signal $|S(p, \omega_k)|$, avec l'argument que c'est l'aspect important du point de vue perceptif [Boll 79] [Mc Aulay 80] [Ephraim 84]. Les règles de suppression les plus souvent mentionnées sont principalement :

Soustraction spectrale Cette règle de suppression, introduite par Steven Boll dans [Boll 79], est la plus connue, au point que le terme de soustraction spectrale est souvent utilisé pour désigner, de manière générale, l'ensemble des techniques que nous avons convenu de nommer "atténuation spectrale à court-terme". Le spectre à court-terme obtenu grâce à la soustraction spectrale vérifie [Boll 79] [Lim 79] [Vary 85] :

$$|Y(p, \omega_k)| = |X(p, \omega_k)| - \mathrm{E}\left\{|D(p, \omega_k)|\right\} \tag{2.10}$$

La justification de ce principe est présentée de manière purement intuitive dans [Boll 79]. En utilisant l'expression du niveau relatif local donnée par la relation (2.9), on vérifie que le gain apporté par la soustraction spectrale peut se mettre sous la forme[3] [Vary 85]

$$G_{\mathcal{S}}(p, \omega_k) = 1 - \frac{1}{\sqrt{\mathcal{Q}(p, \omega_k)}} \tag{2.11}$$

On introduit en général la condition supplémentaire

$$G_{\mathcal{S}}(p, \omega_k) = 0 \text{ quand } \mathcal{Q}(p, \omega_k) \leq 1$$

afin d'éviter une atténuation négative qui n'a physiquement aucun sens.

Pseudo Wiener On a déjà vu, au paragraphe 2.1.1, comment il est possible de généraliser grâce à l'atténuation spectrale à court-terme le principe du filtrage de Wiener pour des signaux qui ne sont pas stationnaires. Le filtrage de Wiener permet d'estimer, pour un signal stationnaire, la forme temporelle du signal avec une erreur de puissance minimale. En général la règle de suppression présenté comme "filtrage de Wiener" correspond à l'atténuation spectrale suivante [Lim 79] [Vary 85] [Mc Aulay 80]

$$G_{\mathcal{W}}(p, \omega_k) = \frac{|X(p, \omega_k)|^2 - \mathrm{E}\left\{|D(p, \omega_k)|^2\right\}}{|X(p, \omega_k)|^2} \tag{2.12}$$

Cette expression correspond au gain du filtre de Wiener, défini par (2.2), où *la DSP du signal à traiter a été remplacée par son estimation obtenue par périodogramme* $|X(p, \omega_k)|^2$. C'est ce qui explique que la propriété de puissance d'erreur minimale ne soit absolument pas garantie en pratique. Dans [Lim 79], Jae Lim propose d'ailleurs d'utiliser la même règle avec des estimateurs autres que $|X(p, \omega_k)|^2$. En introduisant l'expression de $\mathcal{Q}(p, \omega_k)$, le gain de l'équation (2.12) se réécrit sous la forme

$$G_{\mathcal{W}}(p, \omega_k) = 1 - \frac{1}{\mathcal{Q}(p, \omega_k)} \tag{2.13}$$

Avec la même condition supplémentaire que précédemment :

$$G_{\mathcal{W}}(p, \omega_k) = 0 \text{ quand } \mathcal{Q}(p, \omega_k) \leq 1$$

Soustraction en puissance L'estimation du spectre à court-terme obtenue par la règle de soustraction en puissance vérifie [Lim 79] [Lim 86] [Vary 85] [Mc Aulay 80]

$$|Y(p, \omega_k)|^2 = |X(p, \omega_k)|^2 - \mathrm{E}\left\{|D(p, \omega_k)|^2\right\} \tag{2.14}$$

En utilisant, l'expression du niveau relatif local, on obtient le gain apporté au spectre à court-terme $X(p, \omega_k)$ sous la forme

$$G_{\mathcal{P}}(p, \omega_k) = \sqrt{1 - \frac{1}{\mathcal{Q}(p, \omega_k)}} \tag{2.15}$$

Avec toujours la condition supplémentaire

$$G_{\mathcal{P}}(p, \omega_k) = 0 \text{ quand } \mathcal{Q}(p, \omega_k) \leq 1 \tag{2.16}$$

[3]Il y a tout de même une petite difficulté ici car dans l'équation (2.10) c'est la quantité $\mathrm{E}\{|D(p, \omega_k)|\}$ qui intervient, alors que c'est $\mathrm{E}\left\{|D(p, \omega_k)|^2\right\}$ qui est utilisée pour définir le niveau relatif. Toutefois, on peut considérer en pratique que $D(p, \omega_k)$ est une variable aléatoire gaussienne complexe (voir les hypothèses concernant l'algorithme d'Ephraim et Malah au paragraphe 2.2.3). Dans ce cas, $|D(p, \omega_k)|$ suit une loi de Rayleigh [Papoulis 91], et on vérifie que $\mathrm{E}\{|D(p, \omega_k)|\} = \sqrt{\pi}/2 \; \mathrm{E}\left\{|D(p, \omega_k)|^2\right\}$. C'est cette dernière relation qui est utilisée pour obtenir l'expression (2.11) en négligeant le facteur de correction $\sqrt{\pi}/2$ (qui vaut environ 0.9).

Grâce à la relation (2.14), on vérifie que la soustraction en puissance fournit une estimation non biaisée du module au carré de la TFCT [Lim 79]. Avec des hypothèses supplémentaires (similaires à celles utilisées au paragraphe 2.2.3 pour l'algorithme d'Ephraim et Malah), il est prouvé dans [Mc Aulay 80] que la DSP du signal après traitement $P_y(p, \omega_k)$, constitue l'estimation au sens du maximum de vraisemblance de la DSP du signal non-bruité $P_s(p, \omega_k)$.

Toutefois, il faut remarquer que dans ces deux démonstrations, seule l'expression (2.15) est prise en compte. La contrainte (2.16) est rajoutée a posteriori pour éviter que l'estimation du spectre de puissance, fournie par (2.14), ne prenne des valeurs négatives. D'une manière plus générale, pour les trois règles de suppression présentées, nous avons souligné qu'il est nécessaire de garantir, par une condition supplémentaire, que $\mathcal{Q}(p, \omega_k) \geq 1$. Cette condition est d'ailleurs aussi utilisée dans l'algorithme d'Ephraim et Malah (cf. paragraphe 2.2.3). On sait *qu'en moyenne* la valeur du niveau relatif est forcément supérieure à 1. Mais du fait de la variance de $\mathcal{Q}(p, \omega_k)$, on observe en pratique des valeurs inférieures à 1. L'attitude adoptée consiste à dire que des valeurs de $\mathcal{Q}(p, \omega_k)$ inférieures à 1 indiquent la présence de bruit seul, donc nécessitent un gain nul. En fait, ceci est inexact, même si $X(p, \omega_k)$ est non nul, on peut observer des valeurs de $\mathcal{Q}(p, \omega_k)$ inférieures à 1 (cf. figure 3.18). La condition $\mathcal{Q}(p, \omega_k) \geq 1$ vient donc corriger les "erreurs" de l'estimateur du niveau relatif afin de garantir la validité de son interprétation en tant que puissance.

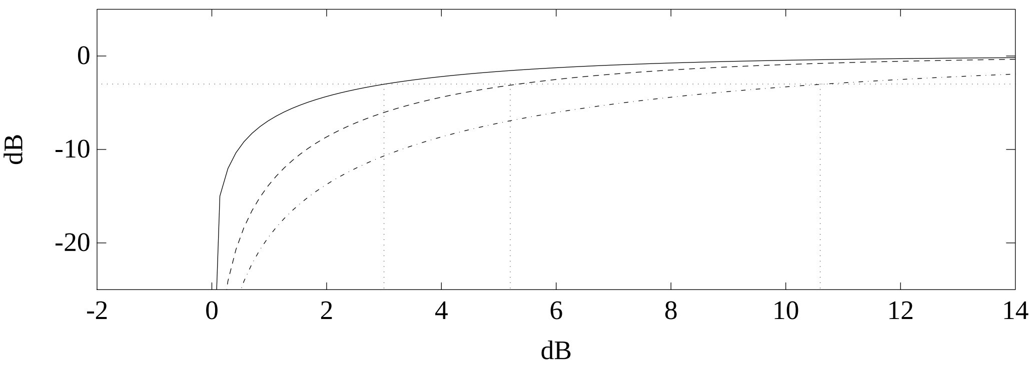

Figure 2.3: Caractéristiques de suppression de trois règles de suppression de bruit : soustraction spectrale (traits mixtes), pseudo Wiener (tirets), soustraction en puissance (trait plein). En ordonnée, l'atténuation $G(p, \omega_k)$ en dB, en abscisse, le niveau relatif local $\mathcal{Q}(p, \omega_k)$ en dB. Les traits pointillés indiquent les niveaux relatifs de coupure (à -3 dB), pour chacune des trois règles.

Comme le niveau relatif local $\mathcal{Q}(p, \omega_k)$ est supérieur à 1, on vérifie grâce aux relations (2.11), (2.15) et (2.13) que les atténuations respectives obtenues par ces trois règles possèdent la propriété suivante

$$\begin{array}{ccccc} G_{\mathcal{S}}(p, \omega_k) & < & G_{\mathcal{W}}(p, \omega_k) & < & G_{\mathcal{P}}(p, \omega_k) \\ \text{soustraction} & & \text{pseudo} & & \text{soustraction} \\ \text{spectrale} & & \text{Wiener} & & \text{en puissance} \end{array} \tag{2.17}$$

Cette propriété peut être représentée graphiquement grâce aux caractéristiques de suppression (atténuation $G(p, \omega_k)$ en fonction du niveau relatif $\mathcal{Q}(p, \omega_k)$) de la figure 2.3. Sur cette figure, on a représenté un *niveau relatif de coupure* (défini traditionnellement à -3 dB) pour chacune de ces trois règles. Le point important est que les comportements de ces trois règles de suppression sont très différents : pour la soustraction en puissance, le niveau relatif de coupure est environ de 3 dB, tandis que pour la soustraction spectrale, ce niveau de coupure est supérieur à 10 dB. Ce qui veut dire que même si la puissance du signal bruité dépasse l'estimation du bruit de fond d'un facteur 10, la soustraction spectrale apporte déjà une atténuation significative. Il faut toutefois noter que la notion de niveau relatif de coupure est d'autant plus pertinente que la caractéristique de suppression présente une variation brusque. Dans le cas de la soustraction en puissance, une variation faible du niveau relatif autour de la valeur de coupure (de l'ordre de 3 dB) suffit à faire passer du cas de l'atténuation totale ($G_{\mathcal{S}}(p, \omega_k) \approx 0$) à celui de la modification négligeable ($G_{\mathcal{S}}(p, \omega_k) \approx 1$). Ceci n'est pas le cas pour le niveau de coupure défini pour la règle de soustraction spectrale, pour laquelle la transition est beaucoup plus longue.

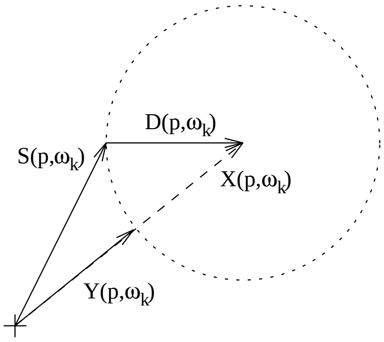

Figure 2.4: Fonctionnement de la soustraction spectrale représenté sous la forme d'un diagramme de Fresnel à la pulsation ω_k. Le cas représenté correspond à une réalisation particulière qui vérifie $|D(p, \omega_k)| = \mathrm{E}\{|D(p, \omega_k)|\}$.

Nous avons vu que la soustraction en puissance permet d'obtenir une estimation non-biaisée du module au carré de la TFCT, par conséquent, la figure 2.3 montre que les deux autres règles (soustraction spectrale et pseudo Wiener) sous-estiment fortement les faibles valeurs du spectre. Pour la soustraction spectrale, ce résultat n'est pas surprenant car le principe même de la relation (2.10) (soustraction des modules) introduit un biais systématique qui est mis en évidence sur la représentation de la figure 2.4. Par contre, la distorsion du spectre est théoriquement nulle avec la soustraction en puissance. En pratique, ceci est faux car compte tenu de la variance de l'estimation spectrale $\mathcal{Q}(p, \omega_k)$, l'estimation du signal $Y(p, \omega_k)$ devient de moins en moins fiable lorsque $\mathcal{Q}(p, \omega_k)$ se rapproche de 1. Il n'en reste pas moins que la règle de soustraction en puissance est la règle de suppression ponctuelle qui implique la plus faible distorsion du spectre du signal. On peut montrer que c'est aussi la règle de suppression pour laquelle le niveau de bruit résiduel est le plus fort, c'est à dire où l'atténuation globale apportée au bruit est la moins importante [Vary 85]. Dans les conditions qui sont celles de la restauration d'enregistrements musicaux fortement bruités, où la distorsion du signal constitue le problème majeur, c'est donc la règle de soustraction en puissance qui est susceptible de fournir les résultats les plus satisfaisants. Le paragraphe 3.1 revient sur l'analyse de la distorsion spectrale du signal, en utilisant la description simplifiée des règles de suppression fournie par les niveaux relatifs de coupure définis sur la figure 2.3.

Une réserve légitime concernant cette présentation consiste à se demander dans quelle mesure la distorsion d'amplitude du spectre à court-terme permet d'évaluer la qualité du traitement. Contrairement à une idée communément admise, il est faux d'affirmer que si le module du spectre à court-terme est correctement estimé, alors le résultat sera forcément perçu comme satisfaisant (cf. paragraphe 3.3). Par contre, le contraire (distorsion du spectre à court-terme) se traduit, pour un son stationnaire, par un filtrage qui peut être perçu comme une modification de timbre. On peut même ajouter que l'ouïe est assez sensible à ce type de modification puisque la variation d'amplitude de l'une des composantes d'un son complexe est détectable environ à partir de ±1 dB [Botte 88].

Une modification de ces règles de suppression très couramment utilisée [Lim 79] [Valiere 91] consiste à surestimer la puissance du bruit. L'intérêt de cette modification est de réduire le niveau du bruit résiduel puisque l'atténuation apportée devient plus forte que ce qui est théoriquement nécessaire. La surestimation peut être réalisée très simplement en multipliant l'estimation du bruit de fond par un coefficient, supérieur à 1, désigné par le terme de *facteur de surestimation.* L'effet quantitatif de cette surestimation dépend bien sûr de la règle de suppression utilisée. La figure 2.5 représente les caractéristiques obtenues en partant de la règle de soustraction en puissance avec différents facteurs de surestimation. Cette figure montre qu'il est nécessaire de limiter au maximum la surestimation, car celle-ci augmente la distorsion apportée au spectre du signal. On peut même dire que l'effet produit est ici particulièrement dévastateur puisque toute valeur du spectre dont le niveau relatif est inférieur au facteur de surestimation est mise à zéro.

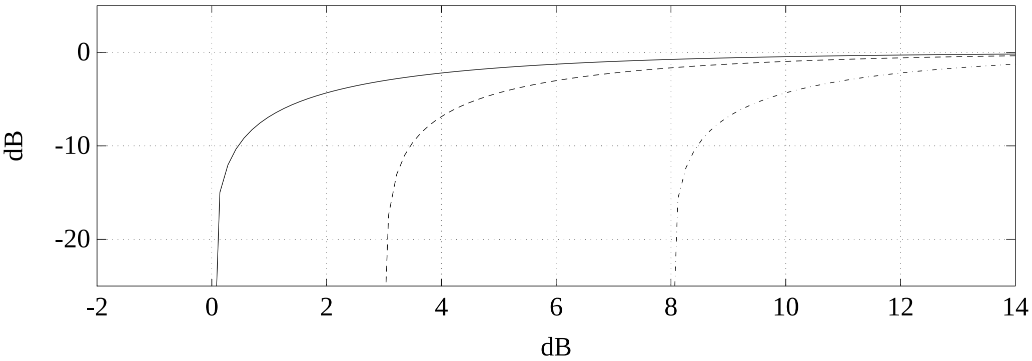

Figure 2.5: Caractéristiques de suppression de la règle de soustraction en puissance pour trois valeurs du facteur de surestimation : 0 dB (trait plein), 3 dB (tirets), 8 dB (en traits mixtes). En ordonnée, l'atténuation $G_{\mathcal{P}}(p, \omega_k)$ en dB, en abscisse, le niveau relatif local $\mathcal{Q}(p, \omega_k)$ en dB.

Une dernière remarque sur ces différentes règles de suppression concerne le problème du *bruit musical* mentionné dans la plupart des publications consacrées au débruitage par atténuation spectrale à court-terme. On montre en effet que l'application directe d'une méthode d'atténuation spectrale génère un bruit résiduel très peu naturel, et en tout cas inacceptable dans un contexte musical. Ce bruit résiduel particulier est en général baptisé *bruit musical* par référence au fait qu'il est constitué de sinusoïdes apparaissant aléatoirement (cf. paragraphe 3.2.1). Cependant cet effet désagréable est avant tout une conséquence de la variance de l'estimateur spectral $\mathcal{Q}(p, \omega_k)$: on montre au paragraphe 3.2.1 que, quelle que soit la règle de suppression ponctuelle utilisée, l'élimination *totale* du bruit musical s'obtient à partir d'un facteur de surestimation fixé (à condition que le bruit vérifie bien les hypothèses de stationnarité et de non-corrélation). En conséquence, on peut considérer que le phénomène de bruit musical et le choix d'une règle de suppression, *parmi les règles ponctuelles*, sont deux aspects indépendants. C'est pourquoi ce phénomène n'a pas été pris en compte dans la discussion.

2.2.3 Règle de suppression d'Ephraim et Malah

Jusqu'à présent, toutes les règles de suppression mentionnées étaient strictement locales, c'est à dire que l'atténuation spectrale $G(p, \omega_k)$ s'exprimait uniquement en fonction du niveau relatif $\mathcal{Q}(p, \omega_k)$ *mesuré dans la trame à court-terme.* Nous allons maintenant détailler le fonctionnement de la règle de suppression proposée par Ephraim et Malah dans [Ephraim 83] (puis reprise de manière plus détaillée dans l'article [Ephraim 84]). La particularité de cette règle provient du fait que la valeur de l'atténuation spectrale $G_{\mathcal{E}}(p, \omega_k)$ *dépend essentiellement* des valeurs du spectre à court-terme mesurées dans les trames précédant la trame courante. Il faut noter qu'Ephraim et Malah ont proposé d'autres règles de suppression en s'inspirant du même principe [Ephraim 84] [Ephraim 85]. Les résultats obtenus étant assez comparables, il nous a semblé préférable de détailler une seule de ces règles de suppression en soulignant les différences significatives qu'elle présente avec les règles de suppression ponctuelles.

L'algorithme d'Ephraim et Malah est fondé sur l'estimation bayésienne du spectre d'amplitude à court-terme du signal non-bruité $|S(p, \omega_k)|$ au sens des moindres carrés. Afin de pouvoir calculer explicitement la valeur de l'estimateur, les auteurs ont ajouté aux hypothèses classiques de stationnarité et de non-corrélation du bruit les hypothèses suivantes :

1. Le bruit est supposé gaussien.
2. Sur le signal, l'hypothèse faite est que la valeur du spectre à court-terme $S(p, \omega_k)$ est une variable gaussienne complexe centrée dont la partie réelle et la partie imaginaire sont indépendantes. Dans ces conditions, le module du spectre à court-terme du signal (non bruité) est distribué selon une loi de Rayleigh [Papoulis 91], tandis que sa phase est uniformément distribuée dans $]-\pi, \pi]$.
3. De plus, la transformation spectrale utilisée est supposée telle que les composantes spectrales $S(p_1, \omega_{k_1})$ et $S(p_2, \omega_{k_2})$ sont statistiquement indépendantes dès que $p_1 \neq p_2$ (indépendance entre trames successives) ou $k_1 \neq k_2$ (indépendances entre valeurs distinctes d'un même spectre à court-terme).

L'hypothèse de bruit gaussien (1) n'est en fait pas nécessaire dès lors qu'on suppose le bruit additif $d(n)$ stationnaire. On sait en effet que la sortie d'un filtre passe-bande très sélectif est quasiment gaussienne. Plus précisément, si $d(n)$ est un processus aléatoire stationnaire dont les moments de tout ordre décroissent suffisamment vite à l'infini, alors les valeurs de sa transformée de Fourier discrète (avec une fenêtre de pondération quelconque) sont asymptotiquement gaussiennes complexes (la partie réelle et la partie imaginaire sont des variables gaussiennes centrées de même variance et indépendantes) [Brillinger 81]. Dans le cas du débruitage par atténuation spectrale, la longueur de la fenêtre de pondération est en général grande (en tout cas, au moins 64 points), on peut donc considérer que le spectre à court-terme du bruit est complexe gaussien dès lors que le bruit est stationnaire (le bruit lui même n'étant pas forcément gaussien). L'hypothèse (2) concerne la densité de probabilité a priori de la grandeur à estimer, elle s'inspire du même type de considérations en supposant cette fois que le signal traité est lui-même un processus aléatoire strictement stationnaire. Cependant, cette situation ne correspond pas forcément au cas des signaux traités. On trouvera des discussion à ce propos pour le cas des signaux de parole dans [Ephraim 84] et [Porter 84]. En pratique l'hypothèse (3) n'est vérifiée que de manière approximative : compte tenu du recouvrement entre trames successives et de la largeur de bande des filtres d'analyse de la TFCT, les valeurs voisines du spectre à court-terme sont corrélées (nous en verrons un exemple au paragraphe 4.2.2.b).

Avec ces hypothèses, l'estimation du spectre d'amplitude à court-terme du signal non-bruité $|S(p, \omega_k)|$ est obtenue sous la forme :

$$|Y(p, \omega_k)| = G_{\mathcal{E}}(p, \omega_k)\, |S(p, \omega_k)| \tag{2.18}$$

Où l'atténuation apportée $G_{\mathcal{E}}(p, \omega_k)$ dépend de deux paramètres:

Le niveau relatif local a posteriori [4] $\mathcal{Q}_{post}(p, \omega_k)$, qui correspond simplement au niveau mesuré dans la fenêtre à court terme tel qu'il est utilisé dans toutes les règles de suppression :

$$\mathcal{Q}_{post}(p, \omega_k) = |X(p, \omega_k)|^2 / \hat{P}_d(\omega_k)$$

Le rapport signal-à-bruit a priori $\mathcal{R}_{prio}(p, \omega_k)$ qui *dépend des trames de signal déjà traitées.* Les auteurs proposent de le déterminer grâce à la procédure suivante

$$\mathcal{R}_{prio}(p, \omega_k) = (1-\alpha) \underbrace{(\mathcal{Q}_{post}(p, \omega_k) - 1)}_{\mathcal{R}_{post}(p, \omega_k)} + \alpha \frac{|Y(p-1, \omega_k)|^2}{\hat{P}_d(\omega_k)} \tag{2.19}$$

Cette équation fait intervenir la valeur du module du spectre à court-terme estimée dans la trame précédente $|Y(p-1, \omega_k)|$ calculée grâce à la formule (2.18). Le terme de gauche (multiplié par $(1-\alpha)$) est un rapport signal-à-bruit local (ou a posteriori dans la terminologie d'Ephraim et Malah), tandis que le terme de droite (en facteur de α) s'interprète comme un rapport signal-à-bruit déduit de la trame précédente. Les auteurs proposent d'utiliser une valeur du paramètre α très proche de 1 (environ 0.98), ce qui équivaut à privilégier très fortement la contribution des trames précédentes. En effet, la formule (2.19) montre que le rapport signal-à-bruit a priori dépend alors essentiellement de l'estimation effectuée dans la trame précédente. Enfin, comme pour les autres règles de suppression, il est nécessaire d'imposer au niveau relatif local a posteriori $\mathcal{Q}_{post}(p, \omega_k)$ (c'est à dire celui qui est mesuré dans la trame) une valeur supérieure à 1, afin d'éviter que le calcul du rapport signal-à-bruit à priori par la formule (2.19) ne fournisse une valeur négative qui n'aurait aucun sens physique.

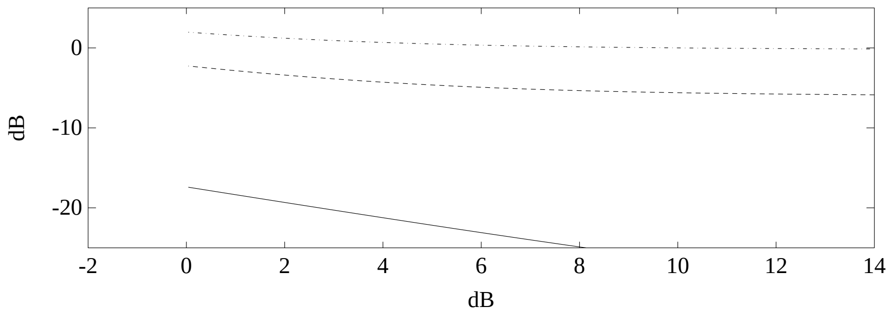

Figure 2.6: Caractéristiques de suppression de l'algorithme d'Ephraim et Malah : atténuation $G_{\mathcal{E}}(p, \omega_k)$ en fonction du niveau relatif a posteriori $\mathcal{Q}_{post}(p, \omega_k)$, pour trois valeurs de niveau relatif a priori $\mathcal{Q}_{prio}(p, \omega_k)$ (0,1 dB en trait plein, 3 dB en tirets, 16 dB en traits mixtes). Cette courbe correspond à la représentation graphique présentée dans les articles originaux d'Ephraim et Malah.

Pour comprendre le fonctionnement de cette règle de suppression, il est nécessaire de décrire le comportement de l'atténuation spectrale $G_{\mathcal{E}}(p, \omega_k)$ en fonction des deux paramètres $\mathcal{Q}_{post}(p, \omega_k)$ (niveau relatif a posteriori) et $\mathcal{R}_{prio}(p, \omega_k)$ (RSB a priori). Ceci peut se faire, par exemple, en traçant un réseau de caractéristiques d'atténuation, puisque l'atténuation dépend maintenant de deux paramètres (la formule explicite du calcul de l'atténuation figure dans [Ephraim 83]

[4] Dans leurs articles originaux Ephraim et Malah utilisent le terme de "rapport signal-à-bruit a posteriori" pour désigner cette quantité. Malheureusement, on a vu au début du paragraphe 2.2 que cette quantité n'est pas homogène à un rapport signal-à-bruit. Ainsi par exemple dans [Ephraim 84], l'absence d'a priori se traduit par la relation "RSB a priori = RSB a posteriori - 1" qui est incompréhensible. Ici, il a été choisi de désigner toutes les quantités conformément à ce qui a été dit au début du paragraphe 2.2.

et [Ephraim 84]). Afin de retrouver des caractéristiques analogues à celles qui existent pour les autres règles de suppression, Ephraim et Malah ont choisi de représenter les courbes donnant le gain de l'algorithme en fonction du niveau relatif a posteriori, les courbes étant paramétrées par le RSB a priori. Ce sont ces caractéristiques qui sont représentées sur la figure 2.6 (le paramètre est le niveau relatif a priori $\mathcal{Q}_{prio}(p,\omega_k)$ défini par $\mathcal{R}_{prio}(p,\omega_k)+1$ ce qui est tout à fait équivalent, mais est plus homogène avec la convention choisie pour le paramètre a posteriori).

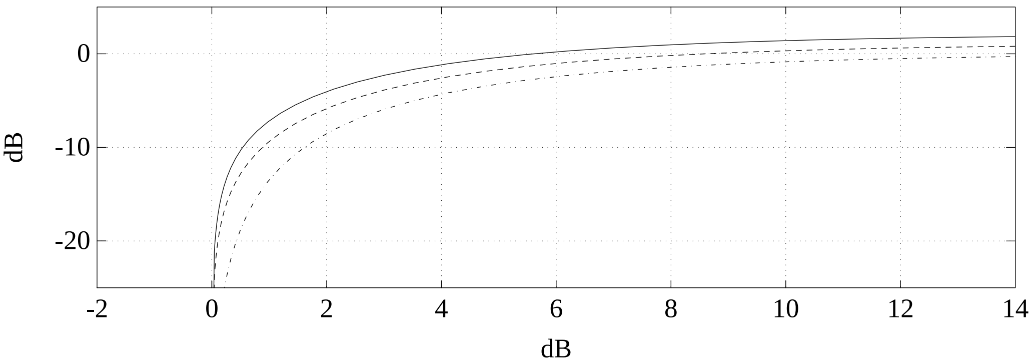

Figure 2.7: Caractéristiques de suppression de l'algorithme d'Ephraim et Malah **en fonction du niveau relatif a priori** : atténuation $G_{\mathcal{E}}(p,\omega_k)$ en fonction du niveau relatif a priori $\mathcal{Q}_{prio}(p,\omega_k)$, pour trois valeurs de niveau relatif a posteriori $\mathcal{Q}_{post}(p,\omega_k)$ (0,1 dB en trait plein, 3 dB en tirets, 16 dB en traits mixtes).

Ce qui est étonnant sur les caractéristiques de la figure 2.6 c'est qu'elles sont relativement horizontales, c'est à dire que la valeur du niveau relatif a posteriori influe peu sur la valeur de l'atténuation apportée. En particulier, une valeur du niveau relatif a posteriori proche de 1 n'implique pas forcément une forte atténuation. Il semble donc que ce soit essentiellement *le niveau relatif a priori qui fixe l'atténuation.* Pour s'en rendre compte, il suffit de tracer les caractéristiques précédentes en échangeant le rôle des deux paramètres. L'examen de la figure 2.7 montre que les courbes reliant l'atténuation au niveau relatif a priori sont tout à fait comparables avec les caractéristiques de suppression de la figure 2.3. De plus, il apparaît clairement sur la figure 2.7 que l'influence du niveau relatif a posteriori est faible par rapport à celle du niveau relatif a priori (en particulier, seule une valeur du niveau relatif a priori proche de 1 est susceptible de provoquer une atténuation importante). Une comparaison plus précise des caractéristiques des figures 2.3 et 2.7 montre que la courbe correspondant à l'atténuation de l'algorithme d'Ephraim et Malah en fonction du niveau relatif a priori est très proche, lorsque le niveau relatif a posteriori n'est pas trop important, de celle de la soustraction en puissance. Ephraim et Malah signalent d'ailleurs qu'en l'absence d'a priori, c'est à dire lorsque le niveau relatif a priori est simplement égal au niveau a posteriori, ce qui peut être simulé simplement en utilisant une valeur de α égale à 1 dans la formule (2.19), l'algorithme se comporte quasiment comme la soustraction en puissance. Un autre point important est que lorsque le niveau relatif a priori possède une valeur faible (proche de 1) l'influence du niveau relatif a posteriori n'est plus négligeable. De plus, on constate dans cette zone (faibles valeurs de $\mathcal{Q}_{prio}(p,\omega_k)$) *une inversion du rôle du niveau relatif a posteriori par rapport au cas des règles de suppression ponctuelles* : l'atténuation apportée est d'autant plus forte que le niveau relatif a posteriori est élevé. Toutefois cette influence du niveau relatif a posteriori n'est sensible que lorsque le niveau relatif a priori est très proche de 1 (en gros inférieur à 3 dB). Enfin, une petite particularité est visible sur la partie droite de la figure 2.7 : dans certaines conditions, l'algorithme d'Ephraim et Malah résulte en un gain légèrement supérieur à 1, ce qui est pour le moins difficile à interpréter.

En conclusion, il faut retenir deux points principaux concernant le fonctionnement de l'algorithme d'Ephraim est Malah :

1. La formule de calcul de l'atténuation qui est proposée est quasiment équivalente à la règle de soustraction en puissance à condition de substituer au niveau relatif mesuré localement un niveau relatif a priori "lissé" temporellement grâce à la formule (2.19).

2. Lorsque ce niveau relatif a priori est proche de 1, le niveau relatif mesuré localement dans la trame (dit a posteriori) agit à l'inverse de ce qui se passe pour les règles de suppression ponctuelles : une valeur élevée du niveau relatif a posteriori implique une atténuation d'autant plus importante.

Le point important est que cette règle de suppression permet d'éviter le phénomène de bruit résiduel "musical" déjà mentionné à propos des règles ponctuelles [Ephraim 84] [Bourdier 88] [Valiere 91]. Or, les tests d'évaluation menés dans [Ephraim 84] permettent de penser que c'est bien la réunion des deux points mentionnés ci-dessus qui produit cet effet. Nous reviendrons au paragraphe 4.1.2 (et dans l'annexe D) sur l'analyse de cette règle de suppression proposée par Ephraim et Malah. Le phénomène de bruit "musical" est quant à lui décrit au paragraphe 3.2.

2.3 Utilisation de la transformée de Fourier à court-terme

D'après ce qui a été dit au paragraphe 2.1.3, l'importance théorique de la transformée de Fourier à court-terme vient de ce qu'elle permet une réalisation efficace d'une transformation à largeur de bande uniforme. En conséquence, dans la plupart des techniques de débruitage par atténuation spectrale à court-terme, c'est la TFCT qui est utilisée. L'annexe B présente une description équivalente des modifications effectuées sur la TFCT sous la forme d'un filtrage linéaire variant dans le temps. Cette description met en évidence plusieurs *conditions que doivent vérifier les paramètres de TFCT* afin de garantir la validité du traitement (paragraphe B.2).

Toutefois, dans le cas particulier du débruitage par atténuation spectrale à court-terme, ces conditions sur les paramètres de TFCT peuvent être notablement relâchées. En particulier le paragraphe 2.3.2, montre pourquoi le phénomène de *repliement temporel*, défini au paragraphe B.2, peut être négligé lorsque le pas de décalage des fenêtres de TFCT est suffisamment faible. La différence majeure, par rapport à la présentation plus classique adoptée dans l'annexe B, vient du fait que la forme de la réponse équivalente à la modification n'est pas supposée connue a priori. A notre connaissance, cet aspect de l'utilisation de la TFCT pour le débruitage n'est pas décrit dans la littérature. Nous avons donc choisi de l'illustrer par un exemple détaillé qui montre le lien existant entre la modification spectrale spécifiée sur la TFCT, et le filtrage effectivement appliqué au signal.

2.3.1 Choix usuels des paramètres

Il est intéressant de constater que les paramètres de TFCT habituellement choisis pour les applications de débruitage par atténuation spectrale ne garantissent pas le respect des conditions évoquées au paragraphe B.2. Une revue de la littérature sur le débruitage par atténuation spectrale à court-terme [Boll 79] [Bourdier 88] [Valiere 91] [Moorer 86] [Ephraim 84] indique que les paramètres de TFCT utilisés sont :

- Fenêtre d'analyse douce : en général une fenêtre de Hamming ou de Hann.
- Recouvrement entre fenêtres successives entre 50% et 75% (c'est à dire que le pas de décalage des fenêtres R est choisi entre $N/2$ et $N/4$).
- Fenêtre de synthèse rectangulaire de même longueur que la fenêtre d'analyse (technique de synthèse OLA, cf. paragraphe B.3.1).

Ce qui semble étonnant c'est qu'avec ces choix de paramètres retenus dans les systèmes de débruitage, les conditions qui permettent d'éviter les termes dus au repliement ne sont pas vérifiées. Pour le repliement temporel, la condition (B.19) portant sur les supports des fenêtres s'écrit, dans le cas où N est la longueur commune des deux fenêtres $h(n)$ et $f(n)$,

$$N + N + \mathcal{L}_{\max_p}(g_p) - 1 \leq 2N$$

soit

$$\mathcal{L}_{\max_p}(g_p) \leq 1 \tag{2.20}$$

Cette dernière condition signifie que dans les implémentations utilisées pour le débruitage, la seule modification qui ne génère pas de repliement temporel est l'atténuation uniforme (de même valeur pour tous les points fréquentiels) ! Quant au repliement spectral, la situation est assez semblable puisque parmi les implémentations utilisées, la seule qui élimine efficacement les composantes de repliement spectral[5] consiste à utiliser un recouvrement d'au moins 75% avec, par exemple, une fenêtre de Hamming [Allen 77]. En particulier, la solution la plus fréquemment employée d'un recouvrement de 50% génère des termes de repliement spectral.

2.3.2 Filtrage équivalent à la modification spectrale

Il semble donc bien que dans le cas du débruitage, le repliement prévu par la théorie n'existe pas, ou en tout cas, reste suffisamment faible pour être inaudible. On peut considérer qu'une des conditions du paragraphe B.2 est quasiment remplie : un recouvrement de 50% ne semble pas tout à fait suffisant, mais il permet tout de même de limiter le repliement spectral. Le paragraphe (2.3.3) présente quelques raisons supplémentaires susceptibles d'expliquer pourquoi ce recouvrement est suffisant en vue du débruitage.

Toutefois, il semble que la seconde condition (élimination du repliement temporel) soit loin d'être remplie. La condition (2.20) est à ce titre assez paradoxale. La suite de ce paragraphe est consacrée à l'étude de ce phénomène de repliement temporel. Notons tout d'abord que la notion de réponse impulsionnelle équivalente à la modification spectrale apportée dans une trame à court-terme, telle qu'elle est définie par la relation B.14 (paragraphe B.2), ne convient pas à notre problème. En effet, pour le débruitage, c'est la modification spectrale $G(p, \omega_k)$ qui est spécifié, il n'y a aucune raison pour que la réponse impulsionnelle $g_p(m)$ obtenue par TFD inverse constitue l'objectif à atteindre. Ce point est très important : dans une application de type convolution rapide, c'est la réponse impulsionnelle du filtre $g_p(m)$ qui est spécifiée, on en

[5] Ou plus précisément, qui assure que le repliement spectral se produit au maximum à un niveau de -40 dB par rapport à l'amplitude des signaux de sous-bande [Crochiere 83].

déduit une modification des spectres à court-terme $G(p, \omega_k)$, par la suite, on cherche à ce que la modification apportée sur la TFCT corresponde exactement au filtrage désiré. C'est ce qui justifie la présentation adoptée au paragraphe B.2.

Au contraire, dans le cas du débruitage, c'est la modification spectrale $G(p, \omega_k)$ qui est spécifié. Le but recherché est que cette modification de la TFCT se traduise par un filtrage équivalent dont la réponse fréquentielle est "proche" de $G(p, \omega_k)$, dans un sens qui reste à définir. Nous allons montrer que tel est bien le cas lorsque le recouvrement entre les fenêtres est suffisant. Rappelons tout d'abord que l'équation (B.16) qui définit le signal obtenu après modification s'écrit

$$y(n) = \sum_{l=-\infty}^{+\infty} \sum_{p=-\infty}^{+\infty} x(l)h(pR-l)f(n-pR) \left\{ \frac{1}{N} \sum_{k=0}^{N-1} G(p, \omega_k) W_N^{k(n-l)} \right\}$$

Le terme entre accolades représente la TFD inverse de la modification spectrale $G(p, \omega_k)$ évaluée à l'indice temporel $(n-l)$. On définit

$$g_p^{\infty}(m) = \frac{1}{N} \sum_{k=0}^{N-1} G(p, \omega_k) W_N^{km} \tag{2.21}$$

Cette *réponse impulsionnelle de durée infinie* est périodique de période N. Notons qu'elle s'obtient par périodisation de la réponse $g_p(m)$ définie par l'équation (B.14). En utilisant la notation $g_p^{\infty}(m)$, le signal obtenu en sortie s'écrit

$$y(n) = \sum_{l=-\infty}^{+\infty} \sum_{p=-\infty}^{+\infty} x(l)h(pR-l)f(n-pR) g_p^{\infty}(n-l) \tag{2.22}$$

En effectuant le changement de variable $m = n - l$, cette expression devient

$$y(n) = \sum_{m=-\infty}^{+\infty} x(n-m) \sum_{p=-\infty}^{+\infty} g_p^{\infty}(m)h(pR-n+m)f(n-pR) \tag{2.23}$$

La comparaison de cette dernière équation avec la relation analogue (B.18), obtenue au paragraphe B.2, montre que les termes de repliement temporel n'apparaissent plus. Dans (B.18), les termes de repliement (obtenus pour $q \neq 0$) sont liés à la périodisation de la réponse $g_p(m)$ effectuée par la relation (B.17). En considérant directement la réponse périodisée de support infini $g_p^{\infty}(m)$, ces termes disparaissent donc naturellement. L'équation (2.23) indique que le lien entre la TFD inverse de la modification spectrale $g_p^{\infty}(m)$ et la réponse impulsionnelle équivalente du filtre variant dans le temps $\tilde{g}_n(m)$ peut toujours se représenter par le schéma B.5, *sans aucune condition supplémentaire, lorsqu'on considère la réponse périodisée* $g_p^{\infty}(m)$. Ce schéma est rappelé ici sur la figure 2.8.

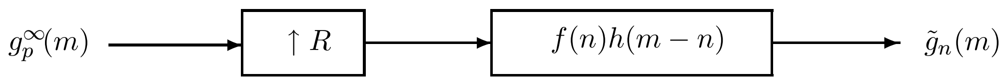

Figure 2.8: Relation entre $g_p^{\infty}(m)$ et $\tilde{g}_n(m)$. L'indice m (rang dans la réponse impulsionnelle variant dans le temps) est supposé fixé.

Compte tenu des supports respectifs des fenêtres d'analyse et de synthèse ($[-(N-1), 0]$ pour $h(n)$, et $[0, N-1]$ pour $f(n)$, d'après le paragraphe B.1), on vérifie que le filtre $f(n)h(m-n)$ est identiquement nul dès que $|m| > (N-1)$. C'est à dire, d'après le schéma 2.8, que *le support de la réponse impulsionnelle* $\tilde{g}_n(m)$ du filtre variant dans le temps équivalent à la modification spectrale *est limité à l'intervalle* $[-(N-1), N-1]$. Ce filtrage non causal correspond simplement à l'idée intuitive que la modification apportée à un échantillon ne dépend que des échantillons situés dans une même trame à court-terme, c'est à dire, distants au plus de $\pm(N-1)$ échantillons. Ceci montre aussi que les seules valeurs de $g_p^\infty(m)$ qui interviennent sont celles qui vérifient $m \in [-(N-1), N-1]$. Il aurait donc été tout aussi légitime de définir $g_p^\infty(m)$ comme étant simplement deux période de la réponse infinie (en excluant l'échantillon correspondant à l'indice $m = -N$). Dans la suite, on considère donc que $g_p^\infty(m)$ correspond uniquement aux deux premières périodes de la réponse infinie.

Cas d'un filtrage invariant Pour étudier le comportement de l'équation (2.23), on considère maintenant le cas plus simple où la modification spectrale $G(p, \omega_k)$ ne dépend pas du temps. Dans, ce cas la réponse impulsionnelle $g_p^\infty(m)$ est aussi invariante dans le temps , elle est donc notée simplement $g^\infty(m)$. L'équation (2.23) se réécrit alors sous la forme

$$y(n) = \sum_{m=-\infty}^{+\infty} x(n-m) g^\infty(m) \left(\sum_{p=-\infty}^{+\infty} h(pR - n + m) f(n - pR) \right) \tag{2.24}$$

Dans cette dernière équation le terme entre parenthèse ne dépend plus que des fenêtre d'analyse et de synthèse. Dans le cas où il n'y a pas de décimation des signaux de sous-bande, c'est à dire où $R = 1$, ce terme s'écrit

$$\sum_{p=-\infty}^{+\infty} h(p - n + m) f(n - p)$$

En posant $l = n - p$, on reconnaît là l'expression du produit de convolution entre $h(m)$ et $f(m)$. Dans le cas où $R = 1$, l'équation (2.24) se simplifie donc en

$$y(n) = \sum_{m=-\infty}^{+\infty} x(n-m) \left\{ g^\infty(m) \left[h * f(m) \right] \right\} \tag{2.25}$$

Ce qui signifie que le signal $x(n)$ subit dans ce cas un simple filtrage linéaire invariant dans le temps. Par conséquent, la réponse impulsionnelle équivalente à la modification spectrale $\tilde{g}_n(m)$ ne dépend pas de l'indice temporel n, elle est donc notée $\tilde{g}(m)$. D'après l'équation (2.25), la réponse impulsionnelle de ce filtre appliqué au signal $x(n)$ s'écrit

$$\tilde{g}(m) = g^\infty(m) \left\{ h * f(m) \right\} \tag{2.26}$$

Une première conséquence de cette relation est que *dans le cas où les signaux de sous-bande ne sont pas décimés* ($R = 1$), *une modification spectrale constante se traduit par un filtrage linéaire (invariant dans le temps) du signal*, et ce, *sans aucune condition sur les fenêtres d'analyse et de synthèse*. Cette propriété correspond simplement au fait que dans le cas où $R = 1$, l'analyse/synthèse par TFCT est une opération linéaire (voir les figures B.1 et B.2 de l'annexe B). La condition (B.19), visant à éliminer le repliement temporel, est une contrainte nécessaire lorsque l'on cherche à ce que le filtre équivalent $\tilde{g}(m)$ soit égal à $g(m)$ (première période de la TFD inverse de $G(\omega_k)$). Cependant, le non-respect de cette condition n'exclue pas la linéarité du traitement. Dans le cas du débruitage, il n'est donc pas utile de chercher à remplir la condition (B.19). Par contre, il faut s'assurer que le filtre $\tilde{g}(m)$ possède bien une réponse fréquentielle proche de de la modification spectrale spécifiée $G(\omega_k)$. Il est possible de décomposer le passage entre $G(\omega_k)$ et $\tilde{g}(m)$ en trois étapes successives :

1. Calcul de $g(m)$ à partir de $G(\omega_k)$ par TFD inverse, en ne conservant qu'une période.

2. Par suite, $g^{\infty}(m)$ s'obtient en périodisant $g(m)$ sur l'intervalle $[-(N-1), -1]$ (on a vu que seules deux périodes de $g^{\infty}(m)$ étaient prises en compte).
3. Enfin, $\tilde{g}(m)$ correspond à la pondération de $g^{\infty}(m)$ par la fenêtre $h * f(m)$.

Ces trois étapes sont représentées dans le domaine temporel sur la figure 2.9 pour la modification spectrale $G(\omega_k)$ représentée sur la partie **A** de la figure 2.10.

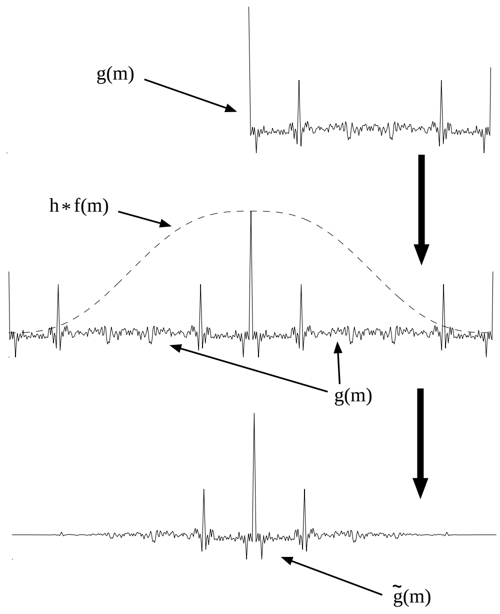

Figure 2.9: Obtention de la réponse impulsionnelle $\tilde{g}(m)$ du filtre équivalent à la modification de la TFCT (où $g(m)$ est la TFD inverse de $G(\omega_k)$). La modification spectrale $G(\omega_k)$ correspondante est représentée sur la partie **A** de la figure suivante.

La modification spectrale représentée sur la partie **A** de la figure 2.10 est un exemple typique des modifications apportées lors du débruitage. Il s'agit ici d'une trame à court-terme de durée 32 ms, issue d'un son de parole voisé, avec un bruit supposé blanc. D'après le principe général de la suppression de bruit, décrit par les relations (2.4), la modification apportée s'apparente alors à un filtrage en peigne puisque que l'atténuation est faible aux fréquences correspondant aux harmoniques du signal, tandis qu'elle est forte entre les partiels. On note sur cette figure un effet qui est décrit au paragraphe 3.1.1 : la largeur des "dents" du peigne est variable car elle est liée à l'amplitude des différents partiels du signal.

La partie **B** de la figure 2.10 présente des oscillations (phénomène dit de Gibbs) qui traduisent le fait que $g(m)$ obtenu par TFD inverse possède une longueur limitée à N échantillons. Les oscillations apparaissent car la réponse fréquentielle spécifiée $G(\omega_k)$ (partie **A**) présente des variations brusques. La partie **C** met en évidence une accentuation des oscillations due à la périodisation de $g(m)$ sur deux périodes. En effet, la réponse périodisée peut s'écrire $g^{\infty}(m) = g(m) + g(m+N)$. On passe donc de $g(m)$ à $g^{\infty}(m)$ par un filtrage par $(\delta(m) + \delta(n+N))$. Pour les réponses fréquentielles, ceci se traduit par une multiplication par

$$1 + \mathrm{e}^{\mathrm{j}\omega N}$$

qui s'écrit aussi

$$2\mathrm{e}^{\mathrm{j}\frac{\omega N}{2}} \cos(\omega N/2)$$

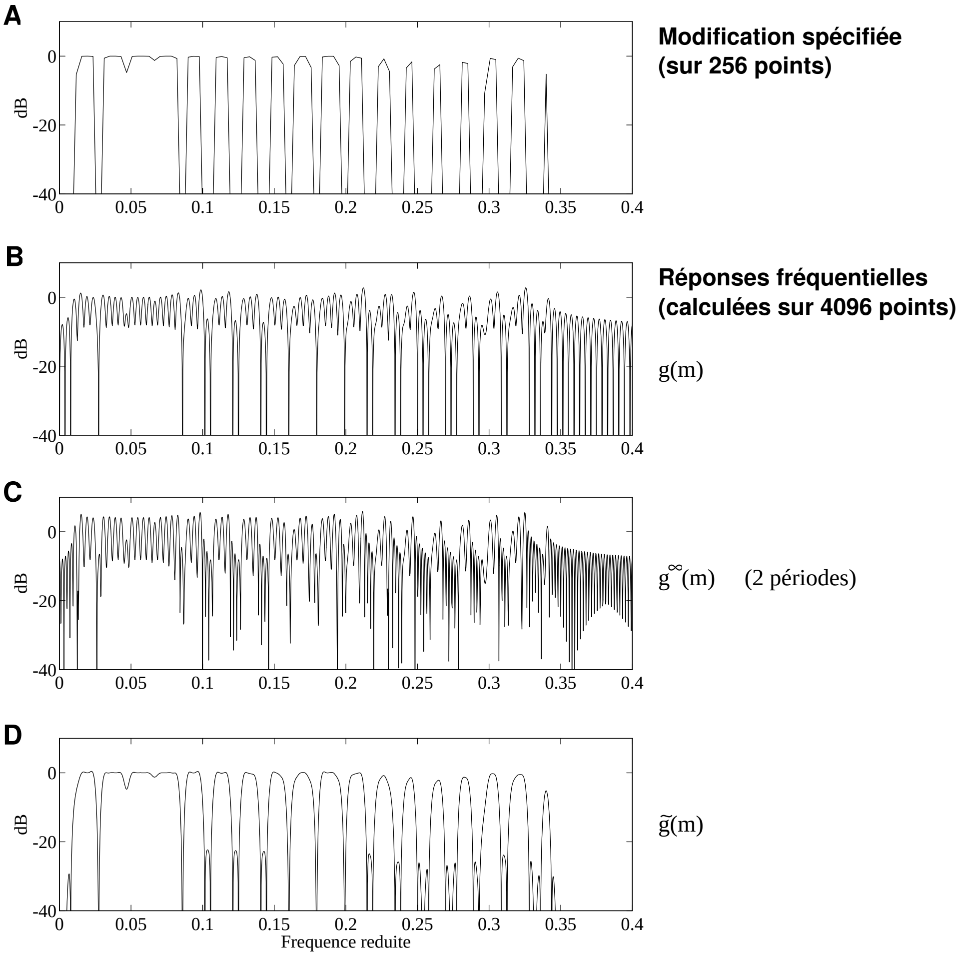

Figure 2.10: Etapes du passage entre la modification spectrale spécifiée sur la TFCT $G(\omega_k)$, et la réponse fréquentielle du filtre équivalent $\tilde{g}(m)$. **Partie A :** modification spécifiée sur la TFCT (sur 256 points) $G(\omega_k)$. **Partie B :** réponse fréquentielle de $g(m)$ obtenu par TFD inverse. **Partie C :** réponse fréquentielle de $g^{\infty}(m)$ (deux périodes). **Partie D :** réponse fréquentielle de $\tilde{g}(m)$ (pondération par $h * f(m)$). La longueur de la fenêtre de TFCT est $N = 256$ (fenêtre de Hann à l'analyse, rectangulaire à la synthèse). Les réponses fréquentielles (**B**,**C** et **D**) sont calculées par TFD sur 4096 points. Seule une partie de l'axe fréquentiel est représentée.

soit encore

$$\{2\,|\cos(\omega N/2)|\}\ \mathrm{e}^{\,\mathrm{j}\left[\frac{\omega N}{2}+\pi \mathrm{neg}\left\{\cos\left(\frac{\omega N}{2}\right)\right\}\right]} \tag{2.27}$$

où $\mathrm{neg}(u) = 1$ si $u < 0$, et $\mathrm{neg}(u) = 0$ sinon. Le terme de gauche (entre accolades) traduit un filtrage en peigne de pas fréquentiel $\Delta\omega = 2\pi/N$. C'est cet effet qui apparaît sur la figure 2.10, en comparant les parties **B** et **C**. Il faut d'ailleurs noter que si on n'observe pas de vrais "zéros" sur la réponse fréquentielle de la partie **C**, c'est uniquement parce que la représentation graphique des réponses impulsionnelles est trop peu précise (calcul des TFD sur 4096 points). La partie droite de l'équation (2.27) (facteur de module 1), correspond quasiment à un terme de phase linéaire traduisant une *avance* de $N/2$ échantillons. Il faut noter qu'à ce stade, la réponse fréquentielle de $g^{\infty}(m)$ (partie **C**) diffère fortement de $G(\omega_k)$ (partie **A**). Cependant, la partie **D** de la figure 2.10 montre que la dernière étape (pondération par $h * f(m)$) permet d'obtenir un filtre $\tilde{g}(m)$ dont la réponse fréquentielle est extrêmement proche de $G(p, \omega_k)$. En effet, la pondération par $h * f(m)$ se traduit dans le domaine fréquentiel par un lissage des réponses par la fonction $H(\omega)F(\omega)$. Dans le cas qui correspond à celui de la figure 2.10 (fenêtre d'analyse de

Hann et fenêtre de synthèse rectangulaire), la largeur à -3 dB du lobe principale de la fonction $H(\omega)F(\omega)$ est environ de $0,8 \times 2\pi/N$, c'est à dire du même ordre de grandeur que la périodicité des "irrégularités" constatées sur les parties **B** et **C** de la figure. Dans ces conditions, le lissage par $H(\omega)F(\omega)$ vient "gommer" à la fois l'effet des phénomène de Gibbs lié à la durée finie de $g(m)$ (partie **B**) et le filtrage en peigne due à la périodisation de $g(m)$ (partie **C**).

La conséquence est que le filtre équivalent $\tilde{g}(m)$ possède une réponse fréquentielle très proche de la modification spécifiée sur la TFCT $G(\omega_k)$. La principale différence observée entre ces deux réponses (**A** et **D** sur la figure 2.10) est l'élargissement de la bande de transition entre les zones fréquentielles non modifiées et celles qui sont atténuées. Cet effet dû au lissage par $H(\omega)F(\omega)$ est particulièrement visible pour le dernier pic à droite de $G(\omega_k)$ (environ à la fréquence réduite 0,34). Cette caractérisation du filtrage passe-bande, équivalent à l'effet du traitement autour de chaque partiel du signal, est utilisée au chapitre 3 pour étudier le comportement du débruitage pour un signal stationnaire.

Quant à la courbe de réponse en phase, notons que par construction $\tilde{g}(m)$ est un filtre non-causal, à phase linéaire, dont le retard de phase vaut 0, puisqu'il est symétrique par rapport à l'origine des temps (cf. figure 2.9). Cette constatation justifie la présence du terme d'avance mentionné à propos de la relation (2.27) : $G(\omega_k)$ étant à phase nulle, $g(m)$ a pour retard de phase $N/2$ (cf. figure 2.9), une avance de $N/2$ échantillons est donc nécessaire pour obtenir un filtre à phase nulle. Au passage, ceci prouve que le lissage de la réponse fréquentielle par $H(\omega)F(\omega)$ a aussi permis d'éliminer les irrégularités de la courbe de phase introduites par le terme $[\exp(\mathrm{j}\pi \mathrm{neg}\{\cos(\omega N/2)\})]$ dans l'équation 2.27.

Modification équivalente dans le cas général Rappelons que le résultat de l'équation (2.26) s'applique dans le cas où, d'une part, la modification spectrale est invariante dans le temps, et de plus, le pas de décalage de fenêtres de TFCT (R) vaut 1.

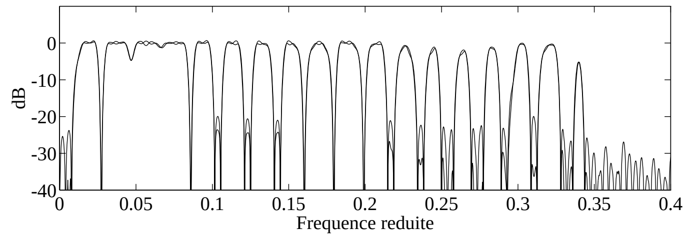

Figure 2.11: Enveloppe des spectres des signaux obtenus par modification de la TFCT d'un signal impulsionnel de position variable. Paramètres de TFCT : longueur $N = 256$, fenêtre d'analyse $h(n)$ de Hann, $f(n)$ rectangulaire, pas de décalage $R = N/8$. La modification apportée est la même que pour la figure précédente. Les spectres sont calculés par TFD sur 4096 points. Seule une partie de l'axe fréquentiel est représentée.

Quand il y a décimation des signaux de sous-bande ($R > 1$), on observe peu à peu des variations de la réponse équivalente au cours du temps. Toutefois, les variations constatées empiriquement restent extrêmement faibles tant que $R < N/8$. La figure 2.11 présente le résultat d'une simulation effectuée avec un décalage des fenêtres $R = N/8$. La simulation en question consiste simplement à appliquer la modification $G(\omega_k)$ à la TFCT d'un signal constitué d'une impulsion numérique. La figure 2.11 représente la réponse en fréquence du signal obtenu après synthèse. Cependant, comme on suspecte la présence d'effets dépendant du temps, la visualisation de la réponse fréquentielle n'est plus vraiment justifiée. En effet, la réponse fréquentielle n'a de sens que si la position temporelle de l'impulsion de départ est spécifiée puisqu'on a affaire à une réponse équivalente qui varie dans le temps. Sur la figure 2.11, la solution adoptée consiste à représenter l'enveloppe (minimum et maximum) des spectres obtenus pour toutes les positions possibles de l'impulsion. En fait, il suffit de décaler la position de l'impulsion R fois de suite pour décrire toutes les réponse impulsionnelles possibles, puisque $\tilde{g}_n(m)$ est périodique (en n), de période R (voir le schéma 2.8).

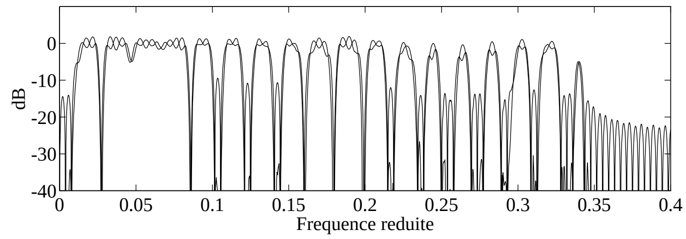

Figure 2.12: Enveloppe des spectres des signaux obtenus par modification de la TFCT d'un signal impulsionnel de position variable. Paramètres de TFCT : longueur $N = 256$, fenêtre d'analyse $h(n)$ de Hann, $f(n)$ rectangulaire, pas de décalage $R = N/2$. Les spectres sont calculés par TFD sur 4096 points. Seule une partie de l'axe fréquentiel est représentée.

En comparant avec la réponse analogue, obtenue par calcul dans le cas où $R = 1$ (partie **D** de la figure 2.10), on constate que les variations observées sur la figure 2.11 restent de très faible amplitude. On observe surtout des différences significatives au niveau des lobes secondaires de la réponse fréquentielle (pour les fréquences réduites supérieures à 0,35). Par contre, le cas d'un décalage de $N/2$ présenté sur la figure 2.12 met en évidence des différences qui ne sont plus négligeables. En particulier, la réjection des zones fortement atténuées est bien moins efficace que dans le cas où $R = 1$ (représenté sur la partie **D** de la figure 2.10). De plus, sur la figure 2.12, on distingue très nettement les deux courbes de l'enveloppe au niveau des zones non modifiées (valeurs de la réponse fréquentielle proche de 0 dB). C'est à dire que selon l'indice temporel considéré, la réponse fréquentielle équivalente à une modification constante, présente, dans ces zones non modifiées, des variations non-négligeables (de l'ordre de 2 dB). Le paragraphe 3.1.1.a revient sur les conséquences de ces effets dépendant du temps dans le cas de signaux sinusoïdaux.

Le cas où la modification spectrale $G(p, \omega_k)$ varie dans le temps est plus complexe puisqu'il n'est plus question de se ramener à un simple filtrage linéaire. Toutefois, un examen plus attentif de l'équation (2.23), dans le cas où $R = 1$, montre qu'il suffit de supposer que $g_p(m)$ est constant sur l'intervalle $[p_0 - (N - 1), p_0 + (N - 1)]$ pour obtenir une relation analogue à (2.25) valable en p_0. Plus généralement, quand $R \neq 1$, si la modification apportée $G(p, \omega_k)$ est localement constante sur un intervalle équivalent à $2N$ échantillons (à la fréquence d'échantillonnage du signal analysé), on obtient une relation similaire à (2.24), valable pour l'échantillon situé au centre de cet intervalle. Les résultats obtenus précédemment peuvent donc être étendus au cas où les variations de la modification spectrale $G(p, \omega_k)$ sont "lentes".

2.3.3 Modification spectrale dans le cas du débruitage

Les arguments exposés au paragraphe précédent permettent de comprendre pourquoi le traitement de débruitage se passe en pratique beaucoup mieux que prévu par la théorie présentée dans l'annexe B. De plus, dans le cas du débruitage, la modification spectrale possède une caractéristique particulière supplémentaire que nous n'avons pas exploitée. Il s'agit du fait que la modification à apporter au spectre à court-terme est évaluée en utilisant ce même spectre à court-terme (voir le schéma de principe de la figure 2.2).

Ainsi, Moorer et Berger [Moorer 86] précisent que la solution qu'ils utilisent pour le débruitage (fenêtre de Hann avec recouvrement de 50%), n'est plus suffisante lorsqu'on désire réaliser un égaliseur. En effet, dans ce dernier cas où c'est l'opérateur qui fait varier comme il le souhaite le gain dans chaque bande, les auteurs remarquent que le système produit des artefacts audibles lorsque la variation temporelle des gains de l'égaliseur est trop rapide. La solution qu'ils préconisent consiste à utiliser un recouvrement beaucoup plus fort. Cet exemple illustre le fait que si le phénomène de repliement ne se produit pas dans le cas du débruitage, c'est précisément parce que les variations temporelles de l'atténuation spectrale sont "lentes". Ceci est dû au fait que l'atténuation à apporter dans chaque fenêtre est calculée en fonction d'une estimation spectrale réalisée à partir des données de la fenêtre (voir le paragraphe 2.2). A partir du moment où les fenêtres se recouvrent fortement, on conçoit que ceci limite les variations temporelles possibles entre deux fenêtres successives. En d'autres termes, les coefficients de la réponse impulsionnelle $g_p(m)$, équivalente à l'atténuation spectrale, varient lentement, ce qui permet de relâcher la contrainte portant sur le filtrage passe-bas réalisé par les fenêtres (cf. schéma 2.8).

Un autre aspect qui intervient dans l'élimination du repliement tient aux caractéristiques fréquentielles de l'atténuation spectrale apportée. En effet, le spectre à court-terme du signal est convolué par la réponse fréquentielle de la fenêtre d'analyse. L'atténuation, qui est fonction de cette estimation spectrale, possède donc toujours un profil fréquentiel "assez régulier" lorsqu'on utilise une fenêtre d'analyse douce. Dans les parties du spectre à court-terme contenant du signal, il n'est donc pas possible de trouver un canal fréquentiel complètement atténué au voisinage direct d'un canal peu atténué (situation dans laquelle le phénomène de repliement est le plus important). Cette remarque sur le rôle de la fenêtre d'analyse pour le lissage de l'estimation spectrale montre au passage pourquoi c'est la fenêtre de synthèse qui est choisie rectangulaire, et non le contraire. Alors que du point de vue de la modification multiplicative de la TFCT, nous savons que ces deux fenêtres jouent des rôles interchangeables (paragraphe 2.3.2).

Ces particularités de la modification spectrale (variations lentes et profil spectral "régulier") permettent de dire que l'analyse effectuée au paragraphe 2.3.2 s'appliquent bien au cas du débruitage. En conséquence, on peut considérer que le respect de la condition (B.19), visant à éviter le repliement temporel (annexe B), ne s'impose pas ici. En particulier, le "zero-padding" du signal pondéré par la fenêtre d'analyse [Crochiere 83], avant calcul de la TFD, ne modifie pas notablement les résultats du débruitage [Bourdier 88] [Boll 79].

Toutefois, d'après le paragraphe 2.3.2, ceci n'est rigoureusement exact que dans le cas d'un recouvrement fort ($R < N/8$). Pour les valeurs de recouvrement utilisés en pratique, on observe des effets dépendant du temps d'amplitude non-négligeable dans le cas d'une modification spectrale constante. Ainsi, le paragraphe 3.1.1.a montre qu'une implémentation du débruitage avec un recouvrement de 50% provoque, dans le cas où le signal traité est un son pur bruité, une modulation qui peut devenir audible dans certaines conditions. Dans des cas comme celui-ci, on constate effectivement que le"zero-padding" permet de réduire la modulation. Cependant, dans ce type de situation où le recouvrement est fixé (en général, pour des considération de charge de calcul), le point le plus important semble être l'utilisation d'au moins une fenêtre (à l'analyse ou à la synthèse) qui possède des valeurs de bord nulles. En pratique, l'utilisation d'une fenêtre dont les valeurs de bord ne sont pas nulles, comme la fenêtre de Hamming, conduit à une très nette aggravation des effets de modulation. En particulier, on obtient alors des "discontinuités" aux bords de chaque trame à court-terme qui viennent élargir le spectre du signal modulant, et donc rendre la distorsion plus audible.

2.3.4 Choix d'une implémentation de la TFCT

En conclusion, il semble bien que, *pour une application de débruitage*, les paramètres proposés dans la littérature soient suffisants. Pour une implémentation où on recherche la rapidité d'exécution, il est possible d'utiliser un recouvrement de 50% avec une fenêtre de Hann (qui présente en plus l'intérêt de permettre une réalisation équivalente du fenêtrage dans le domaine spectral particulièrement peu coûteuse en temps de calcul [Moorer 86]). Cependant, il faut souligner que cette implémentation ne permet pas d'éviter totalement le repliement, en particulier elle n'est plus suffisante dans les cas où l'atténuation est quelconque et non plus calculée à partir d'une estimation spectrale du signal dans la fenêtre, comme c'est le cas pour le débruitage. Les exemples présentés au paragraphe 3.1.1.a montrent d'ailleurs que les effets du repliement sont effectivement audibles, dans certains cas limites, avec les paramètres utilisés pour le débruitage. Pour cette raison, les solutions utilisant un recouvrement inférieur à 50% (en utilisant, par exemple, une fenêtre de type Tukey [Geckinli 78]) ne sont pas recommandées. Par contre, si on désire éliminer totalement le repliement, par exemple dans un cas où le mode de calcul de l'atténuation spectrale est un peu “exotique”, il est nécessaire d'utiliser une recouvrement plus important. Malheureusement, l'élimination des effets dépendant du temps n'est sensible que pour des décalages de fenêtre inférieurs à $N/4$, voire à $N/8$, c'est à dire que l'augmentation du coût de calcul occasionnée est très importante.

Enfin, pour le choix des fenêtres, nous avons souligné, d'une part, le rôle joué par la fenêtre d'analyse $h(n)$ qui influe sur l'allure de l'atténuation apportée en modifiant l'estimation spectrale du signal, et d'autre part, l'importance des valeurs de bords de la fenêtre $h(n)$. Quant à la fenêtre de synthèse $f(n)$, elle n'a pas de fonction spécifique ici puisque c'est le produit de convolution des fenêtres $h*f(n)$ qui intervient vis à vis des modifications de la TFCT. Dans le cadre du débruitage, la charge de calcul supplémentaire que représente l'utilisation d'une fenêtre de pondération (autre que rectangulaire) à la synthèse n'est donc pas vraiment justifiée.

Chapitre 3

Résultats et limites de l'atténuation spectrale à court-terme

Ce chapitre est consacré à l'étude du fonctionnement de la méthode de débruitage par atténuation spectrale à court-terme appliquée à des signaux musicaux bruités. Le but recherché est tout d'abord de vérifier certaines constatations concernant le fonctionnement des systèmes de débruitage que nous avons évoquées au paragraphe 1.3.1. Nous essaierons aussi de mettre en évidence les points qui limitent l'efficacité de ce type de systèmes, afin d'envisager des améliorations possibles qui feront l'objet du chapitre suivant. Ce chapitre s'articule autour de trois grands thèmes qui sont :

- Jusqu'où est-il possible de restaurer des enregistrements bruités compte tenu du principe de fonctionnement de la méthode de débruitage employée ? Cette partie de l'étude permet en particulier de mettre en évidence le fait que si le bruit est trop important, le traitement génère des distorsions du signal musical traité. Les principaux points abordés concernent l'influence de la durée de la fenêtre de traitement sur ces distorsions du signal, à savoir, la modification du timbre des sons stationnaires ainsi que le lissage des transitoires.
- Le deuxième sujet abordé concerne les conséquences de la variance de l'estimation spectrale locale du signal. En particulier, on montre comment apparaît le phénomène bien connu de "bruit musical".
- Le dernier point a trait à l'influence des perturbations apportées par le bruit au spectre de phase à court-terme. On montre en particulier que ces perturbations se traduisent par une modulation des sons stables qui n'est masquée perceptivement que si la fenêtre de traitement est suffisamment longue.

D'un point de vue méthodologique, ces trois thèmes correspondent à des démarches différentes. Pour mettre en évidence la distorsion apportée au signal, on commence par supposer que l'atténuation spectrale apportée par le traitement est une quantité déterministe (paragraphe 3.1). Le traitement est alors équivalent à un filtrage linéaire, éventuellement variant dans le temps, dont on étudie les effets sur le signal. Par la suite, on prend en compte le fait que l'atténuation, qui est fonction de l'estimation spectrale locale, est une grandeur aléatoire (paragraphe 3.2). Enfin, on se place dans une situation où les deux effets précédents peuvent être négligés pour étudier la nature exacte du signal restauré (paragraphe 3.3). Ces distinctions doivent être considérées comme des approximations qui permettent de mener à bien l'étude. L'effet du traitement dans un cas réel correspond bien sûr à la combinaison de tout ces phénomènes.

Dans cette partie, le système de débruitage considéré est toujours, sauf mention contraire, un système de débruitage classique tel qu'il a été présenté au chapitre précédent. C'est à dire que la transformation spectrale à court-terme utilisée est caractérisée par une répartition uniforme des sous-bandes, ce qui fait de la TFCT le mode de réalisation naturel de cette transformation. De plus, la règle de suppression de bruit considérée est toujours une règle ponctuelle (cf. paragraphe 2.2.2). Concernant le comportement de la règle de suppression, on utilise, le plus souvent, la vision simplifiée qui découle de la notion de niveau relatif de coupure présentée au paragraphe 2.2.2. C'est à dire que la caractéristique de suppression est approximée par une caractéristique de type “tout ou rien” :

$$\left\{ \begin{array}{lcll} G(p,\omega_k) & = & 0 & \text{quand} \quad \mathcal{Q}(p,\omega_k) \leq \mathcal{Q}_{lim} \\ G(p,\omega_k) & = & 1 & \text{sinon} \end{array} \right.$$

Où $\mathcal{Q}_{lim}$ désigne le niveau relatif de coupure associé à la règle de suppression considérée. Conformément à ce qui a été dit au paragraphe 2.2.2, cette approximation est plus ou moins justifiée selon le type de règle de suppression considérée.

3.1 Distorsions dues à la modification spectrale

Le but est ici de déterminer, pour un signal musical donné, le niveau de bruit à partir duquel le traitement d'atténuation spectrale à court-terme provoque une distorsion du signal à restaurer. Dans les cas où c'est possible, on cherche à répondre à cette question du point de vue de l'auditeur, c'est à dire, à savoir si la distorsion occasionnée par le traitement est perceptible, compte tenu de la présence du bruit de fond.

Conformément à ce qui a été dit au paragraphe 1.3.3, les calculs sont effectués pour deux signaux-types : le son pur bruité qui représente le cas des parties quasi-stationnaires des signaux musicaux (paragraphe 3.1.1), et l'apparition brutale d'un son pur qui correspond à un cas extrême de signal transitoire (paragraphe 3.1.2). De plus, pour simplifier ces calculs, on considère ici que l'atténuation spectrale à court-terme, apportée par le traitement de débruitage, est une quantité déterministe. C'est à dire que l'on raisonne uniquement à partir de l'*atténuation apportée en moyenne*. On néglige donc la variance de l'estimation spectrale locale, dont les conséquences seront établies au paragraphe 3.2. Enfin, il faut souligner que l'évaluation de l'audibilité de la distorsion due au traitement n'est possible que dans le cas des signaux quasi-stationnaires. Ceci est dû au fait que le signal-test utilisé pour les parties transitoires n'est pas suffisamment représentatif pour permettre une évaluation perceptive significative (cf. paragraphe 1.3.3).

3.1.1 Modification du timbre des signaux stationnaires

Un des reproches qui revient souvent à propos des résultats de traitement est que le processus de restauration semble contribuer à rendre le timbre de l'enregistrement plus terne, moins brillant (cf. paragraphe 1.3.1). Si on considère le spectre d'un son musical stationnaire, on constate, en général, que l'amplitude des partiels du signal décroît globalement avec la fréquence. C'est à dire que les composantes du signal qui sont les plus susceptibles d'être éliminées lors du traitement sont situées dans le haut du spectre. Cette explication semble indiquer que la perte de brillance perçue est directement liée à la distorsion causée par le traitement.

Afin de répondre de manière objective à cet argument, on cherche à déterminer l'ensemble des situations dans lesquelles le traitement se traduit par une *élimination perceptible de certains partiels du signal*. Un point très important est le fait que c'est le signal avant traitement qui sert ici de référence. En effet, la distorsion est considérée comme perceptible uniquement si les composantes du signal musical qui sont éliminées lors du traitement sont audibles à l'écoute du signal bruité (compte tenu de la présence du bruit). Dans le cas contraire, il existe peut-être une distorsion du signal, mais elle n'est pas détectable lors d'une comparaison auditive entre le signal bruité et le résultat du traitement. Pour percevoir auditivement la distorsion dans ce cas, il faudrait disposer du signal musical non-bruité, ce qui ne correspond pas à une situation réelle.

La démarche adoptée dans la suite de cette partie consiste à déterminer tout d'abord la *limite de restauration* pour une composante sinusoïdale du signal. Le terme limite de restauration désigne la puissance de la sinusoïde en dessous de laquelle celle-ci est complètement éliminée par le traitement. Dans une seconde étape, on cherche à évaluer l'audibilité d'une telle composante, située à la limite de restauration, compte tenu de la présence du bruit de fond (paragraphe 3.1.1.c). Enfin, on fournit quelques éléments qualitatifs qui permettent de généraliser les résultats obtenus, dans ce cas simple, à des situations plus complexes.

Le premier paragraphe (3.1.1.a) correspond à une question annexe, mais importante en pratique, qui est celle de savoir s'il est légitime de considérer que le traitement ne provoque pas de distorsion pour un son pur, dès lors que celui-ci se situe au-dessus de la limite de restauration.

3.1.1.a Effet de la troncature du spectre d'une sinusoïde

On suppose ici que le signal à restaurer se réduit à un son pur (signal sinusoïdal) bruité. La présence d'une composante sinusoïdale dans le signal bruité se manifeste au niveau du spectre de puissance à court-terme par la présence d'un pic. Ce qui est sûr c'est que le son pur n'est totalement éliminé que si le niveau relatif local, mesuré au sommet du pic, est en dessous de la valeur de coupure. De même, si la puissance de la sinusoïde est suffisante pour que son spectre à court-terme soit entièrement situé au dessus du niveau local de bruit, il n'y a pas de modification du signal. La question est de savoir ce qui se passe dans les cas intermédiaires.

La figure 3.1 présente un exemple de la situation qui nous intéresse. D'après le paragraphe 2.3, une modification du type de celle qui est représentée sur le bas de la figure 3.1 est équivalente à un filtrage passe-bande du signal, à condition que le recouvrement entre les fenêtres successives soit suffisant. En l'occurrence, le signal étant une sinusoïde, on peut en conclure qu'il n'y a pas de modification du signal dans un cas tel que celui-ci. Malheureusement, pour les applications de débruitage, le recouvrement entre trames successives est en général de 50%. Dans ces conditions, la modification de la TFCT n'est pas strictement équivalente à un simple filtrage, les effets dépendant du temps ne peuvent plus être négligés (cf. paragraphe 2.3).

Afin de donner un ordre de grandeur de ces effets dépendant du temps, nous avons appliqué des modifications spectrales, analogues à celle de la figure 3.1, à la TFCT d'un signal sinusoïdal non-bruité. La figure 3.2 présente l'enveloppe du signal résultant, obtenu après TFCT inverse. L'enveloppe représentée correspond au module de la transformée de Hilbert [Oppenheim 89]. Les trois cas représentés correspondent, de haut en bas, à des niveaux de bruit de plus en plus faibles : sur le haut de la figure 3.2, seuls deux points du spectre à court-terme de la sinusoïde

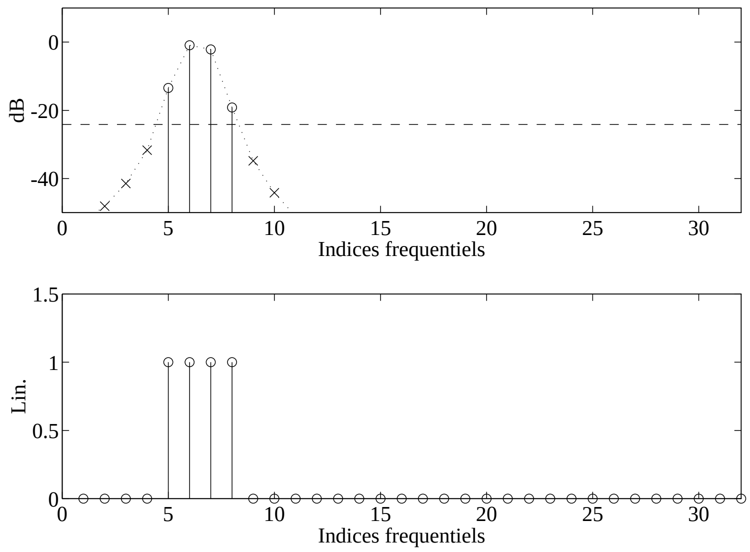

Figure 3.1: Exemple d'atténuation spectrale apportée lors du traitement d'un son pur bruité. **En haut,** le spectre du son pur (en pointillés) et le niveau de bruit (en tirets). Les ronds et les croix distinguent les points du spectre situés en dessous et au dessus du niveau de bruit. **En bas,** l'atténuation spectrale apportée. La fenêtre d'analyse utilisée est une fenêtre de Hann, de longueur 64.

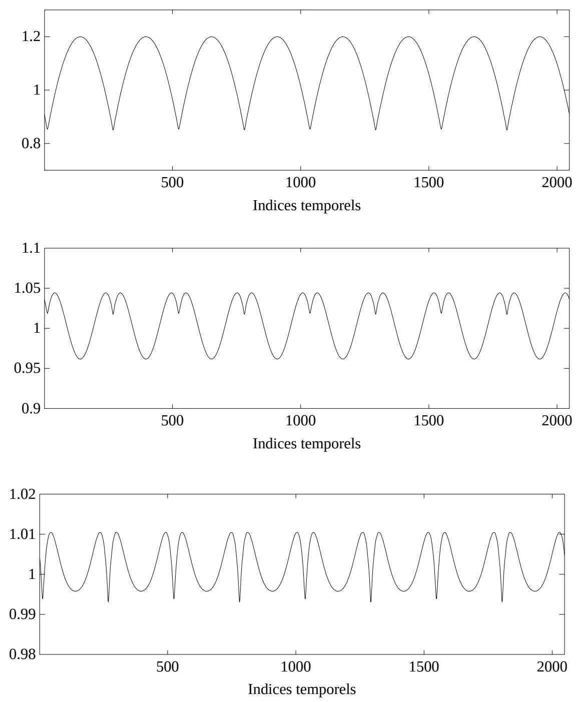

Figure 3.2: Modulation d'amplitude due à la troncature du spectre à court-terme dans le cas d'une sinusoïde. **En haut,** Dans chaque spectre à court-terme, 2 points ont été conservés autour du pic spectral. **Au milieu,** 4 points. **En bas,** 6 points. Les paramètres de TFCT utilisés lors de la modification sont : une fenêtre d'analyse de Hann de longueur 512 points, avec un recouvrement de 50%. La fréquence de la sinusoïde originale se situe entre deux points de discrétisation fréquentielle. **Attention, l'échelle verticale est différente dans les trois cas.**

dépassent du niveau de bruit, au milieu il s'agit de quatre points (ce qui correspond à un cas similaire à celui de la figure 3.1), enfin, en bas, six points ont été conservés.

La figure 3.2 montre tout d'abord que l'amplitude des effets dépendant du temps est loin d'être négligeable lorsqu'on utilise un recouvrement de 50%. Sur une sinusoïde, l'effet produit est celui d'une modulation d'amplitude[1] importante (l'indice de modulation est de l'ordre de 20% pour le cas du haut). De plus, on constate que le nombre de points non-modifiés autour du sommet du pic spectral influe fortement sur le niveau de la modulation : l'indice de modulation est divisé au moins par 10, entre le cas où deux points sont conservés (en haut de la figure 3.2), et celui où six points sont conservés. Intuitivement, ce résultat est naturel car quand le nombre de points non-modifiés autour du pic spectral augmente (c'est à dire lorsque le niveau de bruit diminue), seules les sous-bandes très éloignées de la fréquence de la sinusoïde sont modifiées. Le phénomène de repliement (cf. paragraphe B.2), à l'origine des effets dépendant du temps dans le cas d'une modification spectrale constante, se produit alors sur des signaux d'amplitude très faible, ce qui limite son importance.

Il faut souligner que les résultats de la figure 3.2 dépendent énormément de la position exacte de la sinusoïde par rapport aux points de discrétisation fréquentielle de la TFCT. La même figure, tracée dans le cas d'un point correspondant exactement à une fréquence discrète de la TFCT, donne des indices de modulation 5 à 10 fois plus faibles. Ces écarts sont dus fait que la TFCT ne se comporte pas comme un système linéaire. Le cas représenté sur la figure 3.2 (fréquence de la sinusoïde exactement entre deux points de discrétisation fréquentielle) correspond à la situation où les modulations observées en sortie sont les plus importantes. Par ailleurs, nous avons signalé au paragraphe 2.3.3 qu'une atténuation spectrale dont la variation fréquentielle est "douce" permet de limiter les effets du repliement. En pratique, avec une caractéristique de suppression plus réaliste qu'un simple seuil, on observe aussi une diminution des phénomènes de modulation.

Cependant, les chiffres concernant l'audibilité des phénomènes de modulation fournis au paragraphe B.3.2 (annexe B) indiquent que même si les indices de modulations sont dix fois plus faibles que sur la figure 3.2, la modulation du signal obtenu en sortie est audible (au moins pour les deux cas du haut). Ce qui est surprenant c'est que cet effet n'a jamais été mentionné dans la littérature sur le débruitage. En fait, ce phénomène n'est pas inconnu : on l'entend sur la plupart des systèmes de réduction de bruit par atténuation spectrale pour lesquels le niveau de bruit résiduel est très faible. Il se manifeste auditivement par un effet de grésillement très caractéristique, lié au fait que la modulation due à la troncature du spectre est une modulation par un signal périodique, dont la fréquence fondamentale est fixée par la durée de la fenêtre de TFCT utilisée (voir la figure 3.2). Une autre façon de mettre ce phénomène en évidence consiste à traiter, par atténuation spectrale à court-terme, un signal musical *non-bruité*, en définissant un niveau de bruit fictif. Pour peu que le "niveau de bruit" fixé ne soit pas trop faible, on perçoit très facilement la modulation sur un signal sans bruit.

En pratique, si cette modulation déterministe due à la troncature du spectre de la sinusoïde n'est pas audible, c'est donc qu'elle est masquée par d'autres phénomènes. Nous avons pu vérifier que la principale cause de ce masquage auditif est la présence de bruit résiduel de niveau non-négligeable. En effet, nous verrons au paragraphe 4.1.2 que la règle de suppression d'Ephraim et Malah (cf. paragraphe 2.2.3) permet de faire varier le niveau de bruit résiduel, entre certaines bornes, indépendamment des autres paramètres du traitement. En utilisant cette technique, nous avons constaté que dès que la réduction moyenne du niveau de bruit devient supérieure à environ

[1]Si on observait la phase du signal analytique, on observerait aussi une modulation de fréquence qui n'a pas été représentée ici pour éviter de compliquer la présentation.

25 dB, on obtient une réapparition très nette de cet effet pour une composante sinusoïdale isolée.

En conclusion, **dans des conditions standards de réduction de bruit, on peut considérer que la troncature du spectre d'une composante sinusoïdale, due à l'atténuation spectrale, n'a pas de conséquence audible**. Par conséquent, la seule distorsion envisagée dans la suite est donc la disparition totale de la composante. Cependant, il faut se souvenir que cette conclusion devient fausse lorsque le niveau de bruit résiduel est très faible. C'est en particulier le cas lorsque le niveau de bruit est fortement surestimé. D'après les résultats du paragraphe 2.3, nous savons que, dans un tel cas, la modulation qui est due à la présence de repliement peut être éliminée par une augmentation du recouvrement entre les trames à court-terme.

3.1.1.b Limite de restauration dans le cas d'un son pur bruité

D'après ce qui vient d'être dit, on peut considérer que dans le cas d'un son pur bruité, la recherche de la limite entre restauration et mise à zéro pure et simple du signal revient à déterminer le niveau minimal de la sinusoïde qui garantisse que le sommet du pic spectral dépasse suffisamment le niveau local de bruit. Etant entendu que dans ce paragraphe, on néglige l'action spécifique du bruit sur le spectre à court-terme. De plus, on suppose que le niveau relatif de coupure reste limité à des valeurs raisonnables ce qui nous évite de considérer les cas où la modulation due à la troncature du spectre est audible. Pour fixer l'ordre de grandeur, on considère une valeur de niveau relatif de coupure de l'ordre de 6 dB. C'est à dire que si le niveau relatif mesuré au point correspondant au sommet du pic spectral est inférieur à 6 dB, le signal est totalement éliminé, tandis que dans le cas contraire, le signal n'est pas altéré.

Afin de déterminer la limite de restauration pour un son pur bruité, il est donc nécessaire de relier le niveau relatif moyen, mesuré au sommet du pic de la TFCT, aux paramètres physiques du signal bruité. Le calcul du niveau relatif moyen dans le cas d'un son pur bruité est détaillé dans l'annexe C. Rappelons ici simplement le résultat final donné par la relation (C.19), la **valeur moyenne du niveau relatif mesuré au sommet du pic spectral** s'écrit

$$\mathrm{E}\left\{\mathcal{Q}(p, \omega_{k_0})\right\} = 1 + \frac{\mathcal{P}_s N}{2P_d^{(a)}(F)\, F_e \Delta_h} \tag{3.1}$$

Où $\mathcal{P}_s$ représente la puissance de la sinusoïde, et N la longueur (en nombre d'échantillons) de la fenêtre d'analyse. $P_d^{(a)}(F)$ correspond à la valeur de la densité spectrale du bruit de fond à la fréquence de la sinusoïde, et F_e à la fréquence d'échantillonnage utilisée. Enfin, Δ_h désigne la largeur de bande équivalente de la fenêtre d'analyse de la TFCT vis à vis du bruit. Cette dernière quantité est une constante qui ne dépend que du type de fenêtre choisi.

Interprétation de l'expression du niveau relatif Il est intéressant de noter plusieurs points à propos de ce résultat. Tout d'abord en se souvenant que $\mathcal{Q}(p, \omega_{k_0}) - 1$ est homogène à un rapport signal-à-bruit, on retrouve un résultat bien connu à propos de la transformée de Fourier discrète : quand le signal analysé est une sinusoïde bruitée, plus la taille de la TFD est longue plus le signal émerge du bruit. Plus précisément, une multiplication de la taille de la TFD par un facteur 2 équivaut à un gain en détectabilité de 3 dB (multiplication du rapport des puissances par 2) au niveau du pic de la TFD correspondant à la sinusoïde. Il est possible d'illustrer ce point de manière très simple dans le cadre de l'interprétation de la transformée de Fourier discrète en tant que banc de filtres. Pour ce faire, il est intéressant de réécrire l'équation (3.1) sous la forme

$$\mathrm{E}\left\{\mathcal{Q}(p,\omega_{k_0})\right\} = 1 + \frac{\mathcal{P}_s}{2P_d^{(a)}(F)\left[\frac{F_e}{N}\Delta_h\right]} \tag{3.2}$$

En remarquant que F_e/N est égal à la largeur du pas de discrétisation fréquentielle, le terme $(F_e/N)\Delta_h$ s'interprète comme la largeur (en Hertz cette fois) d'un filtre passe-bande idéal équivalent (au sens défini dans l'annexe C) au filtre de sous-bande propre à l'analyse par TFCT. Le dénominateur dans l'expression (3.2) correspond donc simplement à la puissance du bruit dans le canal fréquentiel de la TFCT d'indice k_0, c'est à dire, après le filtrage passe-bande (le facteur 2 est dû à la prise en compte de la bande de bruit symétrique de fréquence centrale négative). Dès lors, il est clair que si la longueur N de la fenêtre de TFCT est multipliée par 2, la largeur de bande équivalente de chaque canal fréquentiel est elle divisée par 2, ce qui implique que la puissance du bruit filtré dans chaque canal de la TFCT est divisée par 2. La puissance de la sinusoïde filtrée étant inchangée, on obtient donc un rapport signal-à-bruit 2 fois plus grand (en puissance) dans le canal qui contient la sinusoïde. Le point qu'il est important de souligner c'est que la quantité significative, une fois les caractéristiques physiques du signal sinusoïdal et du bruit fixées, est donnée par la largeur de bande équivalente du filtre passe-bande de TFCT exprimée en Hertz. **Le paramètre de contrôle pertinent, le type de la fenêtre d'analyse étant fixé, est la durée de la fenêtre exprimée en secondes** (donnée par N/F_e) et non la longueur de cette fenêtre en nombre de points (N).

L'intérêt d'utiliser des fenêtres longues est donc de diminuer la largeur de bande équivalente du filtre passe-bande de TFCT, et donc de mieux faire ressortir les composantes sinusoïdales du signal par rapport au bruit de fond. Ce raisonnement est valide tant qu'il existe dans le signal musical des composantes sinusoïdales qui sont quasi-stationnaires sur des durées (en seconde) supérieures à celle de la fenêtre d'analyse. Cette conclusion permet de répondre à la question du choix de la durée de la fenêtre tel qu'il a été évoqué au paragraphe 2.1.3. Dans le cas du signal de parole, le signal à restaurer ne présente jamais de composantes sinusoïdales stables sur des durées supérieures à une trentaine de millisecondes, il est donc inutile d'augmenter la taille de la fenêtre au delà de 30ms. Par contre pour un signal musical qui est susceptible de contenir des sons quasi-stationnaires pendant des durées beaucoup plus longues (disons de l'ordre de la demi-seconde), l'augmentation de la durée de la fenêtre permet d'augmenter les possibilités de restauration du système. D'où des choix de durée de fenêtre beaucoup plus longue dans les systèmes de restauration qui sont spécifiquement dédiés aux enregistrements musicaux.

3.1.1.c Audibilité d'un son pur à la limite de restauration

Le calcul précédent nous permet d'évaluer grâce à la formule (3.1) la limite à partir de laquelle la restauration devient impossible (se traduit par une mise à zéro du signal) dans le cas d'un signal constitué d'un son pur bruité. La question posée est maintenant de savoir comment est perçu le signal bruité qui correspond à cette limite. Le but est de déterminer objectivement l'amélioration apportée par le débruitage dans le cas d'un son pur bruité. En particulier, on cherche à savoir si le traitement est susceptible de rendre audible un son pur qui est masqué par le bruit dans l'enregistrement original.

Des études psychoacoustiques ont montré que ce problème peut être décrit simplement en considérant l'effet de masquage du son pur par le bruit. Sans entrer dans les détails, on sait que ceci revient à comparer la puissance du son pur avec la puissance du bruit contenu dans une bande de fréquence située autour de la fréquence F qui est appelée bande critique (centrée en

F) [Zwicker 81]. Le son pur de fréquence F sera masqué par le bruit lorsque sa puissance devient inférieure à la puissance du bruit dans la bande critique centrée en F [2]. Les bandes critiques ont une largeur qui augmentent avec la fréquence centrale selon une loi expérimentale qui est reportée, par exemple, dans [Zwicker 80].

Pour simplifier les calculs, on suppose que bruit est de densité spectrale uniforme dans toute la bande critique centrée sur la fréquence du son pur. La puissance du bruit contenu dans la bande critique s'écrit donc simplement

$$\Delta_{BC}(F) \left\{ 2P_d^{(a)}(F) \right\} \tag{3.3}$$

où $\Delta_{BC}(F)$ désigne la largeur en Hertz de la bande critique centrée en F. En notant $\mathcal{Q}_{lim}$ la valeur du niveau relatif de coupure, la relation (3.2) permet de dire qu'une composante de signal située à la limite de restauration vérifie

$$1 + \frac{\mathcal{P}_s}{2P_d^{(a)}(F)(\frac{F_e}{N}\Delta_h)} = \mathcal{Q}_{lim} \tag{3.4}$$

soit

$$\frac{\mathcal{P}_s}{2P_d^{(a)}(F)} = (\mathcal{Q}_{lim} - 1)\,\frac{F_e}{N}\Delta_h \tag{3.5}$$

D'après la relation (3.3), le son pur est masqué par le bruit dès que le rapport

$$\frac{\mathcal{P}_s}{\Delta_{BC}(F) 2P_d^{(a)}(F)}$$

est inférieur à 1. D'après la relation (3.5), on peut donc dire que **le son pur situé à la limite de restauration est inaudible dans le signal original dès que l'inégalité suivante est vérifiée**

$$\Delta_{BC}(F) > (\mathcal{Q}_{lim} - 1)\,\frac{F_e}{N}\Delta_h \tag{3.6}$$

Cette relation s'interprète simplement si on se souvient que le terme $\Delta_h F_e/N$ représente la largeur de bande équivalente du filtrage passe-bande réalisé par la TFCT (cf. paragraphe 3.1.1.b). L'équation (3.6) signifie donc que le son pur situé à la limite de restauration, est masqué dans le signal bruité, dès que la largeur de la bande critique centrée en F est supérieure à la largeur de bande de la transformée à court-terme (corrigée par un terme qui dépend du niveau relatif de coupure). La correction par le niveau relatif de coupure est nécessaire car plus sa valeur augmente, plus le niveau réel d'une sinusoïde située à la limite de restauration augmente, et donc le masquage de cette composante par le bruit dans le signal bruité devient d'autant moins probable.

Pour effectuer une application numérique, on suppose que le niveau relatif de coupure vaut à peu près 6 dB. La largeur de bande normalisée équivalente à la fenêtre de pondération Δ_h est voisine de $1,5$ pour des fenêtres douces [Harris 78]. Le terme F_e/N est simplement égal à l'inverse de la durée de la fenêtre d'analyse lorsqu'elle est exprimée en secondes. En supposant une fenêtre d'analyse de l'ordre de 40 ms, on obtient

$$(\mathcal{Q}_{lim} - 1)\,\frac{F_e}{N}\Delta_h = \text{ 110 Hz}$$

[2]Attention, dans le cas contraire où c'est la puissance de la sinusoïde qui est supérieure à la puissance de bruit dans la bande critique, on ne peut a priori rien conclure. En général, les deux signaux sont alors audibles car le masquage total du bruit par le son pur nécessite qu'une condition beaucoup plus forte soit remplie [Zwicker 81].

D'après l'équation (3.5), le terme ci-dessus correspond à un rapport d'une puissance par une densité spectrale de puissance, il est donc exprimé en Hertz. Avec les valeurs numériques utilisées, la condition de masquage du son pur par le bruit dans le signal bruité est donc la suivante

$$\Delta_{BC}(F) > \ 110 \text{ Hz}$$

C'est à dire que la largeur de la bande critique centrée en F doit être supérieure à 110 Hz. Cette condition n'est pas tout à fait vérifiée en basse fréquence, puisque la largeur de bande critique en dessous de 500 Hz est de l'ordre de 100 Hz. Par contre, elle est facilement remplie en haute fréquence (la largeur de bande critique à 4 kHz vaut environ 800 Hz). Un point satisfaisant est que ces résultats sont en accord avec la valeur du niveau relatif local, pour un son pur juste audible dans le bruit, telle qu'elle est rapportée dans [Valiere 91]. Il faut toutefois souligner que cette valeur dépend de la fréquence du son pur considéré.

En utilisant une fenêtre de 40 ms pour la TFCT, on est donc en mesure de restaurer des composantes sinusoïdales qui sont situées autour du seuil d'audibilité dans le bas du spectre. Pour le haut du spectre, il est même possible de faire ressortir des composantes qui ne sont pas audibles dans le signal bruité. Pour des valeurs de durées de fenêtre différentes, on rappelle que le sens de variation est le suivant : quand la durée augmente le niveau d'une composante située à la limite de restauration diminue à conditions de bruit constantes (et inversement).

3.1.1.d Perception de la distorsion pour un son musical

L'ordre de grandeur de 60 ms pour la durée de la fenêtre de TFCT est donc une limite très importante au dessus de laquelle on est assuré qu'une composante sinusoïdale de signal qui est mise à zéro par le traitement est inaudible dans le signal bruité. En extrapolant au cas d'un son stationnaire, le signal restauré contient, dans ce cas, plus de partiels du signal original que ce que l'auditeur pouvait percevoir en écoutant le signal bruité. Il est donc légitime de considérer que c'est uniquement dans ces conditions (fenêtre de durée supérieure à 40-60 ms), que la restauration atteint pleinement son but *pour des sons stationnaires.* A l'opposé, la relation (3.6) permet de montrer que si la taille de la fenêtre est de l'ordre de 20 ms, le traitement risque d'éliminer des composantes qui sont audibles dans l'enregistrement bruité. D'après la relation (3.6), ce risque devient d'autant plus grand quand le niveau de bruit est surestimé, puisque qu'on augmente alors la valeur du niveau relatif de coupure $\mathcal{Q}_{lim}$.

Il est nécessaire ici de prendre quelques précautions lors de la généralisation des résultats, obtenus pour une composante sinusoïdale isolée, au cas de la partie stationnaire d'un son musical. En effet, dans un son musical réel, beaucoup des composantes de faible niveau du signal sont masquées, lors de l'audition, par les composantes de signal de niveau plus important. C'est sur ce principe que sont construits la plupart des codeurs de signaux musicaux proposés récemment [Mahieux 89] [Zwicker 91b]. La conséquence du masquage entre les composantes de signal est que, pour un son musical réel, la disparition d'une composante qui devrait être détectable, compte tenu du niveau de bruit, peut être totalement imperceptible du fait de la présence d'autres composantes de signal. Les résultats énoncés précédemment sont donc intéressants surtout dans le cas où il n'existe pas d'effet de masque entre les composantes du signal. Dans le cas de la réduction de bruit, on peut considérer qu'il y a peu de masquage entre les composantes de signal dans le cas d'un niveau de bruit important. En effet, si l'enregistrement est très bruité, les composantes de signal audibles possèdent forcément un niveau important, ce qui limite les possibilités de masquage. De plus, dans le cas d'enregistrement anciens, la bande passante du signal est souvent extrêmement réduite (cf. paragraphe 1.4). Or, le masquage fréquentiel se produit beaucoup moins facilement en basse fréquence (en dessous de 2 kHz), en raison de la diminution de la largeur des bandes critiques [Zwicker 81][Botte 88]. En conclusion, les résultats mentionnés plus haut s'appliquent donc plutôt au cas des enregistrements anciens fortement dégradés (faible bande passante, bruit important).

Une dernière remarque est que même lorsqu'on utilise une durée de fenêtre à court-terme importante, le résultat du traitement peut être perçu comme moins riche que le signal bruité. Cependant il s'agit là d'un effet qui fait intervenir la connaissance des sons musicaux traités, ainsi que l'habitude d'écoute. Par exemple un auditeur averti sera choqué si on lui fait écouter un son de violon filtré passe-bas (disons avec une fréquence de coupure de l'ordre 4 kHz), ce qui traduit la connaissance que l'auditeur a d'un tel son. De manière similaire, on note que si le signal original est très bruité, ou bien s'il est déjà très pauvre (c'est souvent le cas pour de très vieux enregistrements qui possèdent une faible bande passante), le bruit vient combler l'impression de manque, particulièrement dans le haut du spectre. Le bruit semble enrichir le spectre du signal pour lui donner un timbre plus brillant, ce qui rend parfois, par comparaison, le signal restauré assez terne. Une interprétation de cet effet est que le bruit permet de masquer le manque de partiels dans l'enregistrement bruité, c'est à dire que les partiels non audibles sont "synthétisés" par le système auditif à partir de notre connaissance du timbre. Au passage il faut noter que cet exemple souligne le fait que, contrairement à une idée reçue, l'oreille n'est pas toujours capable de séparer les sons musicaux d'un bruit : c'est ici le bruit qui modifie un attribut perceptif du son musical (son timbre) sans que l'auditeur en soit pleinement conscient.

Cependant, on a souligné auparavant que, si la condition sur la durée des fenêtres est vérifiée, il n'est pas possible de trouver une justification objective à cette impression, puisque les partiels qui sont éliminés lors du traitement étaient inaudibles dans le signal bruité. Il n'en reste pas moins que cet effet subjectif existe en pratique. Une solution pour y remédier peut être, par exemple, de conserver un bruit résiduel de niveau relativement important.

3.1.1.e Transformation à largeur de bande non-uniforme

Pour finir, il est intéressant de considérer le même problème dans le cas d'une transformation à court-terme dont la largeur des sous-bandes est non-uniforme. D'après l'interprétation de la formule (3.6), la condition à remplir par la transformation, pour assurer la non-dégradation (au sens où elle a été définie précédemment) du timbre d'un son stationnaire, peut s'énoncer ainsi : pour toute fréquence, il est nécessaire que la largeur de bande équivalente du canal de la transformée (multipliée par le niveau relatif de coupure moins un) soit inférieure à la largeur de bande critique (en supposant que la DSP du bruit est uniforme sur toute la bande critique).

Le choix d'une transformation à largeur de bande non uniforme paraît donc judicieux dans ce cadre puisque la largeur de bande critique augmente avec la fréquence. Toutefois, il faut souligner que la condition de masquage du son pur à la limite de restauration impose l'utilisation d'une transformation possédant une sélectivité fréquentielle relativement importante. Supposons par exemple qu'on s'intéresse à une transformation à Q constant du type transformation en ondelettes. Le cas considéré est celui d'une structure dyadique pour laquelle chaque bande occupe une octave. Pour fixer les idées, on utilise les mêmes paramètres de correction que dans le cas de la TFCT : le facteur multiplicatif qui permet d'obtenir la largeur équivalente d'une bande de transformée vaut $1,5$, et le niveau relatif de coupure de 6 dB. La fréquence d'échantillonnage est supposée être égale à 40 kHz pour couvrir tout le domaine audible. Le tableau 3.1 présente les résultats obtenus sous ces hypothèses pour une transformation par bandes d'octave.

Le tableau 3.1 permet de dire que si on utilise une transformation de ce type, le débruitage risque d'entraîner une dégradation du timbre pour les sons stationnaires puisque pour une composante sinusoïdale, la limite de restauration est nettement plus élevée que le seuil d'audibilité (la relation $\Delta_{BC}(F_c) > \Delta F\,(\mathcal{Q}_{lim} - 1)$ n'est jamais vérifiée, quelle que soit la bande de fréquence). Ceci montre qu'une transformation par bandes d'octave n'est pas assez sélective pour permettre un débruitage dans de bonne conditions. Il est donc nécessaire de rendre le filtrage de sous-bande plus sélectif et donc, en conséquence, d'augmenter le nombre de bandes. Une possibilité pour aller dans ce sens peut être d'utiliser une transformation en ondelettes continue possédant plusieurs voies par octave (étant entendu que chaque voie peut alors être plus sélective). On peut avoir une idée du nombre de voies par octave qui seraient nécessaires en remarquant que dans le tableau 3.1, la voie où le rapport entre la largeur de bande corrigée et la largeur de bande critique est le plus grand est l'octave numéro 6 (de fréquence centrale 1,9 kHz) pour laquelle ce rapport vaut environ 20. C'est à dire que la transformation fréquentielle doit être environ 20 fois plus sélective pour permettre la restauration de composantes sinusoïdales jusqu'au seuil d'audibilité (seuil d'audibilité concernant toujours le signal bruité). Il est donc nécessaire d'utiliser une transformation en ondelette continue possédant au moins une vingtaine de voies par octave. Si on considère, par exemple, une transformation à 32 voies par octave, le nombre total de bandes utilisées pour couvrir tout le spectre est environ de 290.

Octave	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
F_c	60 (Hz)	120	230	470	940	1900	3700	7500	15000
$\Delta F\,(\mathcal{Q}_{lim} - 1)$	180 (Hz)	350	700	1400	2800	5600	11000	22000	45000
$\Delta_{BC}(F_c)$	100 (Hz)	100	100	110	160	280	600	1600	4000
$\Delta F\,(\mathcal{Q}_{lim} - 1)\,/\Delta_{BC}(F_c)$	1,8	3,5	7	13	18	20	18	14	11

Tableau 3.1: Pour une transformation par bandes d'octave, comparaison entre la largeur de chaque bande corrigée par le niveau relatif de coupure $\Delta F\,(\mathcal{Q}_{lim} - 1)$, et la largeur de bande critique correspondant à la fréquence centrale de la bande $\Delta_{BC}(F_c)$ (où F_c désigne la fréquence centrale). Les valeurs sont exprimées en Hertz sauf pour le rapport $\Delta F\,(\mathcal{Q}_{lim} - 1)\,/\Delta_{BC}(F_c)$ qui est sans unité.

Une remarque intéressante est que le point où le manque de sélectivité de la transformation par bande d'octave se fait le plus cruellement sentir se situe autour de 2 kHz. Ce point correspond à la fréquence à partir de laquelle la largeur de bande critique se met à croître significativement : en dessous de 1 kHz, on peut considérer que la largeur de bande critique est à peu près constante, tandis qu'au dessus de 1 kHz, la largeur de bande critique croît nettement avec la fréquence de manière logarithmique. Ceci montre que la transformation fréquentielle qui permet de vérifier la condition recherchée tout en utilisant un minimum de bandes est une transformation de type mixte : transformation à Q constant jusqu'à l'octave 6 (voir le tableau 3.1), 16 voies par octave étant alors suffisantes dans ce domaine de fréquence, le bas du spectre (en dessous de 2 kHz) est quant à lui analysé avec une largeur de bande constante (par une transformation de type TFCT). Un calcul rapide montre qu'on arrive alors à un nombre de bande d'environ 170.

Par comparaison, avec la même fréquence d'échantillonnage (40 kHz), si le traitement est effectué par TFCT, le nombre de bandes utilisées, en ne comptant que les fréquences positives, est de 1000 (en utilisant une fenêtre de 50 ms qui permet de vérifier la condition recherchée). Ceci semble indiquer qu'en ce qui concerne les possibilités de restauration, pour des sons stationnaires (à spectre de raies), le système utilisant la transformée de Fourier à court-terme n'est pas celui qui décompose l'information spectrale de la manière la plus appropriée, du fait des propriétés auditives. Toutefois, il faut nuancer ce résultat en notant que dans le cas d'enregistrements anciens fortement dégradés, la bande occupée par le signal musical est très réduite (cf. paragraphe 1.4). Or, d'après le tableau 3.1, plus on se restreint au domaine des basses fréquences, moins l'intérêt d'une transformation non-uniforme est manifeste.

3.1.2 Lissage des transitoires

Le but recherché ici est de mener une étude analogue à celle du paragraphe précédent, pour le cas des signaux bruités qui présentent des parties transitoires. L'idée est toujours de déterminer une limite de restauration en dessous de laquelle le signal à restaurer subit une distorsion du fait de l'atténuation spectrale apportée par le traitement.

Choix d'un signal transitoire type Par signal transitoire, on désigne ici une portion de signal qui observé à l'échelle de la durée d'une trame à court-terme présente un comportement non-stationnaire. Toutefois, cette définition ne nous permet pas de définir explicitement ce qu'est un signal transitoire, en l'occurrence il s'agit plutôt d'une définition par la négative. Or, pour mener une étude qui ne soit pas purement empirique, il est nécessaire de préciser les caractéristiques de notre signal transitoire. Pour cette raison nous considérons ici un signal transitoire type qui est formé d'un son pur qui apparaît abruptement, ou en retournant l'axe des temps, d'un son pur interrompu. La figure 3.3 présente un exemple représentatif du type de signal étudié. Cette forme de signal peut sembler parfaitement arbitraire, toutefois le choix de ce signal type sera justifié a posteriori au paragraphe 3.1.2.c où on montre comment l'étude de ce cas simplifié permet par généralisation d'aborder le cas de transitoires de signaux musicaux.

Choix d'une référence Le deuxième problème qui se pose est le choix d'une caractéristique qui permette de quantifier objectivement la dégradation du signal bruité. Dans le cas du son pur bruité étudié au paragraphe précédent cette caractéristique était simplement la puissance du son pur : les propriétés du bruit étant fixé, on a vu que l'existence d'une limite de restauration se traduit directement par la condition (3.5) sur la puissance du son pur. Ici la situation est plus complexe car il n'existe pas de mesure unique permettant de caractériser la partie non-stationnaire du signal qui nous intéresse. S'il s'agissait d'un signal transitoire de durée finie, l'énergie du signal pourrait être un indicateur représentatif, mais dans le cas qui nous intéresse, il n'est pas possible de déterminer une quantité représentative du transitoire sans fixer arbitrairement un début et une fin de la partie transitoire. C'est pourquoi nous avons décidé d'utiliser comme référence la puissance de la composante sinusoïdale (considérée dans la partie stationnaire du signal), de la même manière qu'au paragraphe précédent.

La suite de ce paragraphe doit donc nous permettre de déterminer quelle est la distorsion apportée par le traitement, lors de l'apparition brutale d'une composante sinusoïdale dans le bruit, selon la puissance de cette composante. Dans un deuxième temps, les résultats obtenus pour ce signal transitoire particulier sont utilisés pour décrire le comportement du système de restauration dans le cas de transitoires de sons musicaux tels qu'on les rencontre sur des enregistrements réels.

3.1.2.a Mise en évidence du lissage

A partir de maintenant le signal bruité à traiter est donc un signal analogue à celui qui est représenté sur la figure 3.3. La modélisation du traitement effectué par le système de débruitage n'est pas triviale dans le cas d'un tel signal car l'atténuation spectrale apportée varie selon la trame à court-terme considérée. En effet, sur l'exemple de la figure 3.3, les trois positions de la fenêtre d'analyse à court-terme repérées par les lettres A, B et C, correspondent à trois cas distincts :

- Dans la fenêtre A, l'atténuation spectrale apportée est équivalente à une mise à zéro complète du signal puisqu'il s'agit d'une fenêtre qui ne contient que du bruit. Ceci suppose que l'on néglige ici les effets liés à la variance de l'estimation spectrale locale (qui sont étudiés au paragraphe 3.2.1).

- Dans la fenêtre C, le signal fenêtré s'apparente au cas du son pur bruité étudié au paragraphe 3.1.1. D'après les conclusions du paragraphe 3.1.1.a, l'atténuation spectrale apportée est alors équivalente à un filtrage passe-bande autour de la fréquence de la sinusoïde (si on néglige les effets non-linéaires dus au repliement). De plus, on a vu que, pour des conditions de bruit fixées, la bande passante de ce filtre s'élargit lorsque la puissance de la sinusoïde augmente.

- Le cas de la fenêtre B est plus délicat car le signal de trame présente dans cette fenêtre un comportement non-stationnaire très marqué. L'allure générale du spectre à court-terme dans cette fenêtre est cependant assez semblable au cas de la fenêtre C (son pur seul) avec un rehaussement général des zones de faible valeur dû à la présence d'une rupture temporelle dans le signal de trame analysé. La figure 3.4 illustre cette similarité des spectres à court-terme autour de la zone du pic spectral dans les cas des fenêtres B et C. On note que pour la fenêtre centrée sur le transitoire, le niveau du pic spectral diminue tout de même de 6 dB du fait que le signal ne soit présent que sur la moitié de la trame à court-terme. Si le niveau de bruit est suffisant pour masquer le rehaussement des zones de faible puissance du spectre, l'atténuation spectrale apportée dans la fenêtre B sera donc comparable à celle qui est apportée dans le cas de la fenêtre C.

En conséquence, pour les cas qui vont nous intéresser ici, ceux de composantes fortement bruitées, il est légitime de considérer que le traitement réalise une atténuation totale dans toutes les fenêtres situées avant la fenêtre B, tandis qu'il est équivalent à un filtrage passe-bande dans toutes celles qui sont situées après la fenêtre B.

Distorsion pour une composante située à la limite de restauration Considérons le cas d'une composante sinusoïdale qui est située près de la limite de restauration dans la partie stationnaire. D'après les résultats du paragraphe 3.1.1, ceci équivaut à dire que le niveau relatif moyen mesuré au sommet du pic spectral est légèrement supérieur à 1. Le filtrage réalisé par le traitement est alors un filtre passe-bande très sélectif. A la limite, lorsque le niveau relatif mesuré au sommet se rapproche de la valeur 1, il n'existe plus qu'un seul canal de TFCT qui est passant : celui qui correspond au sommet du pic spectral. Etant entendu que l'on considère ici le cas le plus simple d'une exponentielle complexe dont la pulsation Ω correspond à un point de discrétisation fréquentielle de TFCT.

En supposant que la composante de signal apparaît à l'indice temporel $n = 0$[3], la discussion précédente nous permet de dire que pour une composante proche de la limite de restauration, l'atténuation spectrale apportée peut être simplifiée sous la forme suivante :

1. $G(p, \omega_k)$ est nulle si le centre de la fenêtre d'indice p est situé avant l'indice 0 (fenêtres de type A sur la figure 3.4).
2. Dans le cas contraire, $G(p, \omega_k) = 1$ pour $\omega_k = \Omega$ et $G(p, \omega_k) = 0$ pour tous les autres indices fréquentiels (fenêtres de type C sur la figure 3.4).

[3]Cette convention est utilisée afin de simplifier les expressions analytiques. Par contre, sur toutes les figures présentées dans ce paragraphe, l'instant d'apparition de la composante sinusoïdale est fixé arbitrairement à $n = 750$ (ce qui correspond au centre de la figure).

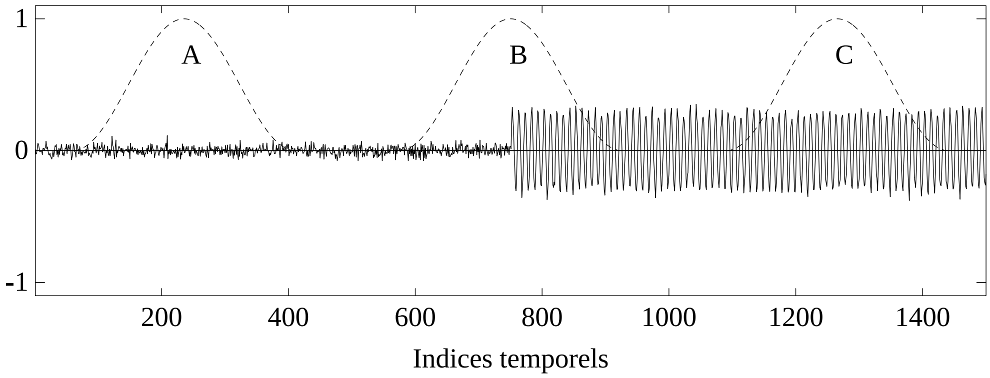

Figure 3.3: Exemple de signal transitoire étudié : apparition brutale d'une composante sinusoïdale dans le bruit. En traits pointillés trois positions typiques de la fenêtre d'analyse.

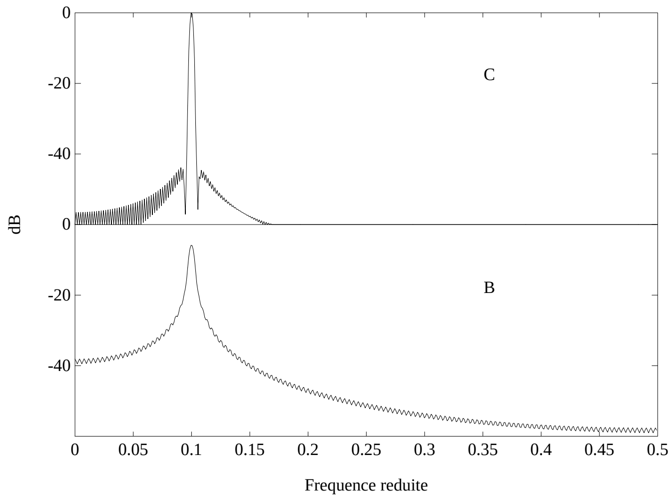

Figure 3.4: Spectre du signal étudié en l'absence de bruit. En haut le cas d'une fenêtre située dans la partie stationnaire (type C). En bas, le cas de la fenêtre centrée sur le transitoire (type B).

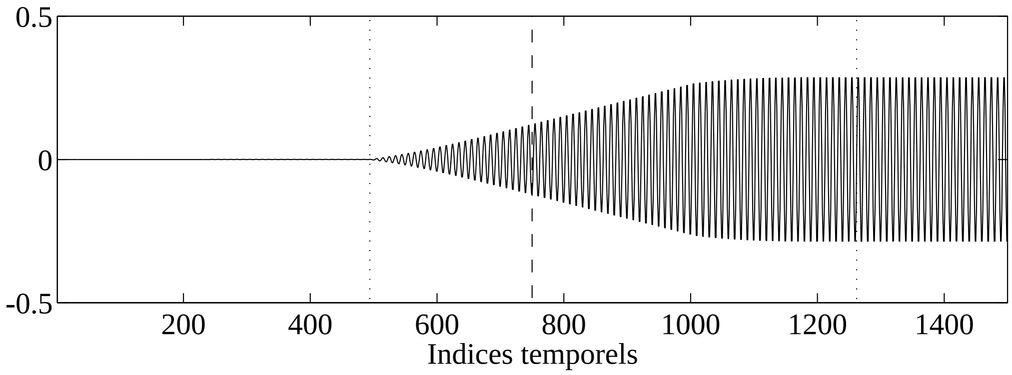

Figure 3.5: Effet du débruitage dans le cas de l'apparition brutale d'un son pur situé à la limite de restauration. En tirets, l'emplacement du transitoire initial. En pointillés, les limites d'étalement ($-N/2$ et $+N$). La fenêtre de TFCT utilisée est une fenêtre de Hann de longueur $N = 512$ points.

Pour simplifier les calculs, on se place dans le cas où le décalage R des fenêtres vaut 1. Avec cette hypothèse, l'équation (2.23) (paragraphe 2.3.2) indique que le signal obtenu en sortie s'écrit :

$$y(n) = \sum_{m=-\infty}^{+\infty} \sum_{p=-\infty}^{+\infty} x(n-m) g_p^{\infty}(m) h(p-n+m) f(n-p) \tag{3.7}$$

avec

$$\begin{cases} x(n) &= A\exp(\mathrm{j}2\pi\Omega n)\Pi(n) \\ g_p^{\infty}(n) &= \frac{1}{N}\exp(\mathrm{j}2\pi\Omega n)\Pi(p+\frac{N}{2}) \end{cases} \tag{3.8}$$

$\Pi(n)$ désignant la fonction créneau ($\Pi(n) = 1$ si $n \geq 0$ et 0 ailleurs). Le facteur $\Pi(n)$ dans l'expression du signal avant traitement $x(n)$ indique que la composante de signal apparaît abruptement à l'indice $n = 0$. Quant à la réponse équivalente à la modification spectrale $g_p^{\infty}(n)$, elle est obtenue d'après l'équation (2.21) par TFD inverse de $G(p, \omega_k)$. Le terme $\Pi(p+\frac{N}{2})$ traduit ici le fait que $G(p, \omega_k)$ est nul quand le centre de la fenêtre est situé avant l'indice $p = 0$, c'est à dire quand $p < -N/2$. En combinant les équations (3.7) et (3.8), on obtient

$$y(n) = A\exp(\mathrm{j}2\pi\Omega n) \times \frac{1}{N} \sum_{p=-\infty}^{+\infty} f(n-p)\Pi(p+\frac{N}{2}) \sum_{m=-\infty}^{+\infty} h(p-n+m)\Pi(n-m) \tag{3.9}$$

soit

$$y(n) = A\exp(\mathrm{j}2\pi\Omega n) \times \frac{1}{N} \sum_{p=-\infty}^{+\infty} f(n-p)\Pi(p+\frac{N}{2})\, h * \Pi(p) \tag{3.10}$$

On définit $M(p) = \Pi(p+\frac{N}{2})\, h * \Pi(p)$. $M(p)$ représente la convolution tronquée de $h(p)$ et $\Pi(p)$. Avec cette notation le signal de sortie s'écrit

$$y(n) = A\exp(\mathrm{j}2\pi\Omega n) \left(\frac{1}{N} f * M(n)\right) \tag{3.11}$$

Le signal obtenu en sortie est donc une sinusoïde (complexe) pondérée par l'enveloppe $\{f * M(n)\}/N$. En utilisant le fait que le support de la fenêtre d'analyse $h(n)$ (resp. fenêtre de synthèse $f(n)$) est l'intervalle $[-(N-1), 0]$ (resp. $[0, (N-1)]$) (cf. annexe B), on montre facilement que l'enveloppe du signal obtenu en sortie $\{f * M(n)\}/N$ est nulle avant $n = -N/2$ et constante (non-nulle) après $n = +N$. La figure 3.5 présente un exemple de signal $y(n)$ calculé grâce à la relation (3.11) pour une fenêtre de TFCT de longueur $N = 512$.

La figure 3.5 montre que pour une composante sinusoïdale située à la limite de restauration, le traitement de débruitage provoque le lissage d'un transitoire abrupt. Cet effet de lissage est non causal et non symétrique : le signal traité croît depuis $N/2$ points avant la position du transitoire jusqu'à N points après la position du transitoire. Il faut noter que dans le cas contraire de l'extinction brutale d'une composante située à la limite de restauration, la conclusion serait la même en inversant *après* et *avant*. Le chiffre a retenir est que, **pour une composante sinusoïdale située à la limite de restauration, le traitement de débruitage se traduit par un lissage des transitoires abrupts, avec un temps de montée de l'ordre de une fois et demi la durée de la fenêtre de TFCT**. Ce lissage est la conséquence directe du filtrage passe-bande autour de la fréquence du son pur réalisé par le traitement.

Afin de vérifier que le modèle du traitement présenté donne une bonne idée de la réalité, malgré les approximations sur lesquelles il est fondé, la figure 3.6 présente un exemple de traitement réel de débruitage qui correspond à un cas où la composante sinusoïdale est de puissance telle qu'elle se situe peu au dessus de la limite de restauration compte tenu des paramètres

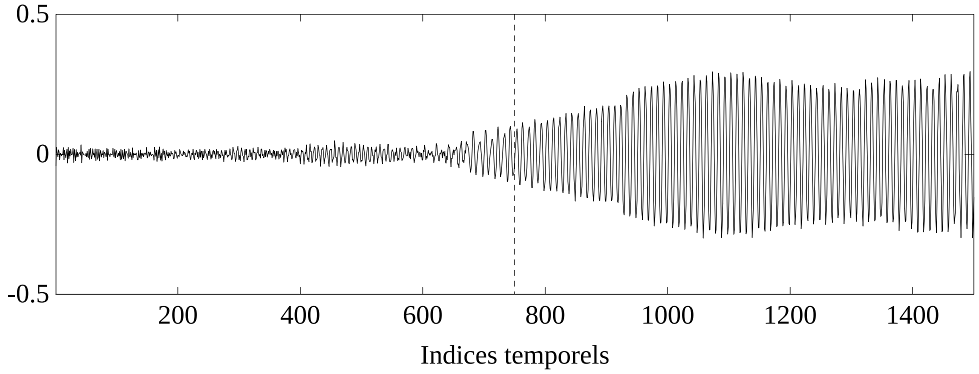

Figure 3.6: Résultat du débruitage dans le cas de l'apparition brutale d'un son pur situé 4 dB au dessus de la limite de restauration. Le trait pointillé vertical indique l'emplacement du transitoire initial. La fenêtre de TFCT utilisée est une fenêtre de Hann de longueur 512 points. La règle de suppression utilisée est la soustraction en puissance avec un facteur de surestimation du niveau de bruit de 8 dB.

utilisés pour la TFCT. On rappelle à ce propos que le fait que la sinusoïde soit située 4 dB au dessus de la limite de restauration signifie que le niveau relatif moyen au sommet du pic spectral correspondant à la sinusoïde est supérieur de 4 dB au niveau relatif de coupure. Etant donné que c'est la règle de soustraction en puissance qui est utilisée, on a vu que le niveau relatif de coupure se situe 3 dB au dessus du facteur de surestimation du bruit de fond (voir le paragraphe 2.2). Ici ce niveau relatif de coupure est donc d'environ 11 dB, ce qui signifie que le niveau relatif moyen mesuré au sommet du pic spectral correspondant à la sinusoïde est de 15 dB.

La figure 3.6 met bien en évidence l'effet de lissage décrit analytiquement par la relation (3.11) et représenté sur la figure 3.5. En particulier, on vérifie que le lissage du transitoire est non causal et qu'il n'est pas symétrique par rapport à la position initiale du transitoire. Un point important est que le résultat théorique représenté par la figure 3.5 correspond au cas où le paramètre de décalage des fenêtres R vaut 1. En effet, dans ce cas, le résultat ne dépend pas de la position du transitoire par rapport aux fenêtres d'analyse de la TFCT. Quand R est supérieur à 1, nous avons observé, sur cet exemple, des variations dépendant de la position des fenêtres qui sont du même ordre que celle mises en évidence pour les composantes sinusoïdales stationnaires au paragraphe 3.1.1.a. Ces variations deviennent très rapidement d'une amplitude non-significatives, en tout cas sur la forme d'onde, dès que le niveau relatif correspondant à la composante sinusoïdale augmente (paragraphe 3.1.1.a), et que le recouvrement entre fenêtres est suffisamment important (paragraphe 2.3). Compte tenu du recouvrement utilisé pour les simulations ($R = N/4$), même sur le cas de la figure 3.6 (composante proche de la limite de restauration), les effets dépendant de la position sont moins importants que ceux dus à la présence du bruit. Dans l'exemple de la figure 3.6, les effets déterministes dépendant de la position de la fenêtre sont tout de même à l'origine des “discontinuités” de l'enveloppe de la composante sinusoïdale (par exemple, au point d'indice 930).

Les principales différences avec le résultat théorique de la figure 3.5 viennent du fait que l'on a négligé l'influence de la variance de l'estimation spectrale. On verra au paragraphe 3.2 que dans un cas réel cette variance est très importante, ce qui explique la présence de bruit résiduel avant le transitoire, ainsi qu'une partie des fluctuations de l'amplitude de la sinusoïde après le transitoire. Il faut d'ailleurs noter que l'exemple de la figure 3.6 correspond à un cas où le niveau de bruit est assez fortement surestimé pour des raisons de lisibilité. En effet, si le niveau de bruit n'était pas surestimé, le bruit résiduel dans la partie gauche de la figure masquerait visuellement l'effet de lissage du transitoire.

Influence du niveau de la composante L'effet de lissage qui vient d'être décrit correspond au cas d'une composante située à la limite de restauration. Il est intéressant d'étudier la manière dont cette distorsion apportée au transitoire évolue selon la puissance de la composante sinusoïdale. On sait déjà que quand le niveau de la sinusoïde augmente, une plus grande partie de son spectre à court-terme dépasse le niveau de bruit, et par conséquent des canaux de TFCT qui étaient totalement atténués deviennent passants. Le filtrage passe-bande effectué par le traitement est donc de moins en moins sélectif lorsque la puissance de la sinusoïde augmente. En conséquence, d'après les résultats du paragraphe précédent, le lissage d'un transitoire abrupt diminue lorsque la puissance de la composante sinusoïdale augmente.

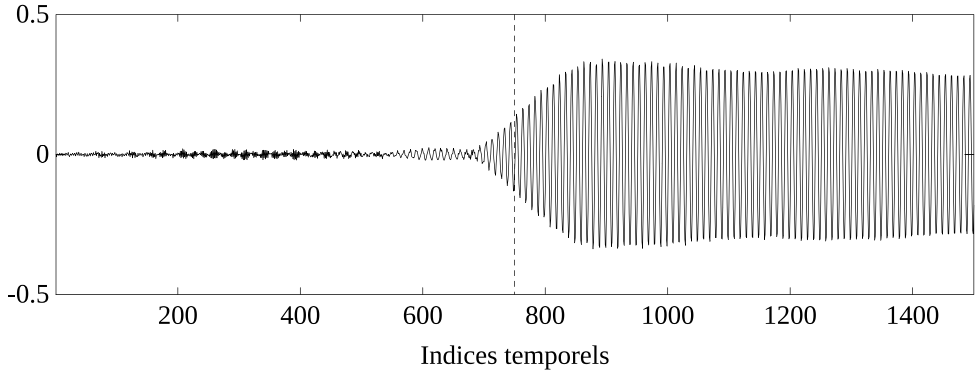

Figure 3.7: Résultat du débruitage dans le cas de l'apparition brutale d'un son pur situé 16 dB au dessus de la limite de restauration. Le trait pointillé vertical indique l'emplacement du transitoire initial. La fenêtre de TFCT utilisée est une fenêtre de Hann de longueur 512 points. La règle de suppression utilisée est la soustraction en puissance avec un facteur de surestimation du niveau de bruit de 8 dB.

Le résultat présenté sur la figure 3.7 correspond à un cas où la sinusoïde se situe 16 dB au dessus de la limite de restauration. Ce qui revient à dire que le bruit qui dégrade le signal traité est plus faible de 12 dB par rapport au cas qui correspond à la figure 3.6. On constate effectivement que le résultat obtenu dans ces conditions est nettement meilleur : le lissage du transitoire se fait sur une longueur beaucoup plus faible. On peut estimer que le temps de montée a été divisé par environ 2,5 par rapport au cas de la figure 3.6. Par ailleurs, il faut noter que les effets de modulation de la composante sinusoïdale (dans la partie stationnaire) ont fortement diminués, ce qui est en accord avec les résultats des paragraphes 3.1.1.a, 3.2.2 et 3.3.1.

Une étude plus précise du spectre à court-terme correspondant à la sinusoïde indique que dans le cas de la figure 3.7, il y a en moyenne trois canaux de TFCT passants lorsque la sinusoïde est présente, alors qu'il n'y en avait qu'un dans le cas d'une composante proche de la limite de restauration (cas de la figure 3.6). Etant donné que la largeur de bande de la réponse fréquentielle équivalente à un canal de TFCT est de l'ordre de grandeur de la distance entre deux canaux successifs (c'est à dire du pas de discrétisation fréquentielle $2\pi/N$), le filtrage passe-bande réalisé par le traitement dans le cas de la figure 3.7 est donc trois fois moins sélectif que celui qui correspond à la figure 3.6. Ce qui explique que le support temporel de la réponse impulsionnelle équivalente soit fortement réduit.

Le point intéressant est de savoir quel est le niveau de la composante sinusoïdale qui garantisse qu'un transitoire brusque ne soit pas lissé par le traitement. Pour obtenir un ordre de grandeur de cette valeur, nous avons déterminé empiriquement la valeur de la puissance de la sinusoïde qui assure que la longueur de lissage est négligeable par rapport à la valeur limite de $1,5\,N$ obtenue dans le cas d'une composante sinusoïdale située à la limite des possibilités de restauration du système. La conclusion est que **pour une composante sinusoïdale située plus de 30 dB au dessus de la limite de restauration, l'effet de lissage d'un transitoire brusque est très fortement limité** puisque le temps de montée est alors de l'ordre d'un dixième de la durée de la fenêtre de TFCT (voir la figures 3.8).

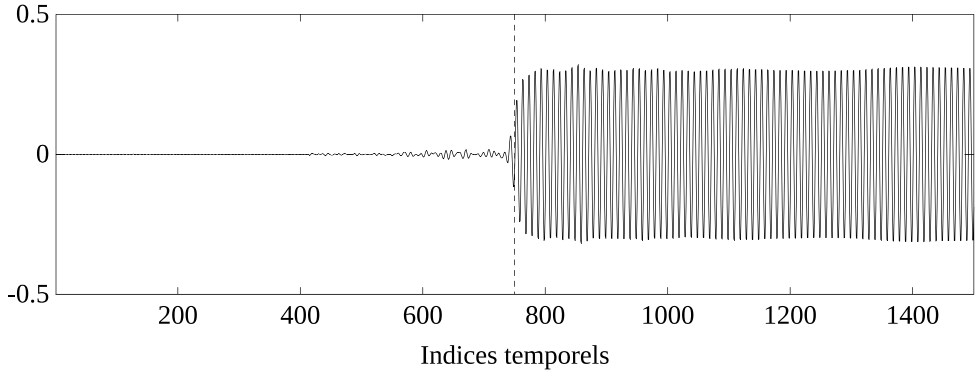

Figure 3.8: Résultat du débruitage dans le cas de l'apparition brutale d'un son pur situé 30 dB au dessus de la limite de restauration. Le trait pointillé vertical indique l'emplacement du transitoire initial. La fenêtre de TFCT utilisée est une fenêtre de Hann de longueur 512 points. La règle de suppression utilisée est la soustraction en puissance avec un facteur de surestimation du niveau de bruit de 8 dB.

3.1.2.b Influence de la durée de la fenêtre de TFCT

Les résultats précédents suggèrent le fait que l'utilisation de fenêtres de TFCT courtes permet de limiter le lissage des transitoires. Plus précisément, les résultats obtenus concernant le lissage du transitoire sont toujours proportionnels à la longueur N de la fenêtre de TFCT. Ce qui semble indiquer que le temps de montée lors d'un transitoire abrupt est directement proportionnel à la durée de la fenêtre utilisée lors du traitement de débruitage. Cependant, il ne faut pas oublier le fait que le niveau relatif local mesuré au sommet du pic spectral correspondant à une sinusoïde dépend de la durée de la fenêtre de TFCT utilisée. D'après l'équation (3.1), l'expression "une sinusoïde située tant de dB au dessus de la limite de restauration" n'a de sens que si les conditions de bruit *et* la durée de la fenêtre de TFCT sont fixées.

Ainsi, on choisit, par exemple, de traiter le même signal que dans le cas de la figure 3.7 en utilisant une fenêtre de TFCT 4 fois plus courte (c'est à dire de 128 points). D'après l'équation (3.1), le niveau relatif moyen mesuré au sommet du pic spectral se trouve divisé par 4, c'est à dire que la composante n'est plus située que 10 dB au dessus de la limite de restauration compte tenu de la nouvelle durée de la fenêtre de TFCT. Ceci implique que le lissage du transitoire est *proportionnellement* plus important que sur la figure 3.7 puisque la composante est plus proche de la limite de restauration. Cependant, une diminution de l'ordre de 6 dB du niveau relatif du pic spectral entraîne une augmentation relative du temps de montée qui est faible par rapport à la division de celui-ci par un facteur 4 du fait de la diminution de taille la fenêtre. C'est ce qui apparaît sur la figure 3.9 qui présente le résultat du traitement *pour le même signal bruité que dans le cas de la figure 3.7* avec une fenêtre de traitement 4 fois plus courte. Conformément à ce qui vient d'être vu, le niveau relatif mesuré au sommet du pic spectral dans la partie stationnaire du signal est alors 6 dB plus faible. Malgré cela, on constate que le temps de montée lors du transitoire est divisé environ par 2 par rapport au résultat présenté sur la figure 3.7.

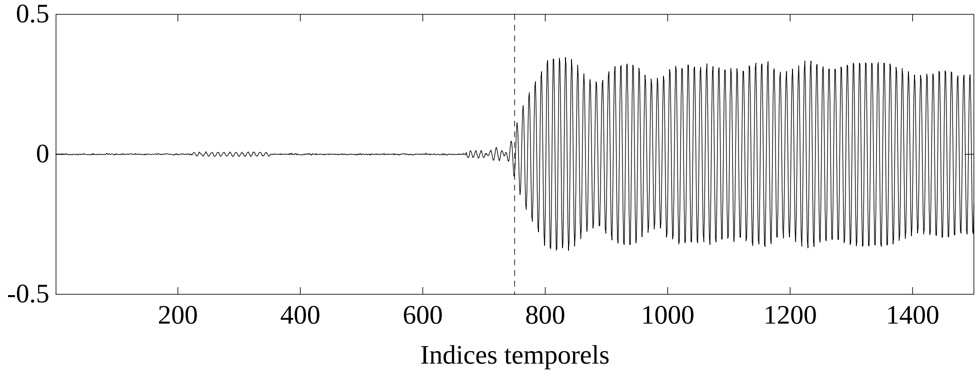

Figure 3.9: Résultat du débruitage dans le cas de l'apparition brutale d'un son pur situé 10 dB au dessus de la limite de restauration. Le trait pointillé vertical indique l'emplacement du transitoire initial. La fenêtre de TFCT utilisée est une fenêtre de Hann de longueur 128 points. La règle de suppression utilisée est la soustraction en puissance avec un facteur de surestimation du niveau de bruit de 8 dB.

Dans le cas que nous avons choisi comme exemple la diminution de 6 dB du niveau relatif du pic spectral s'accompagne d'une augmentation relative du temps de montée qui est inférieure à un facteur 2. Intuitivement, pour que les deux effets se compensent, il faudrait qu'une baisse de 6 dB du niveau relatif local se traduise par un nombre deux fois plus faible de canaux de TFCT passants. Ce n'est jamais le cas sauf si la composante est situé très près de la limite de restauration puisque le nombre de canaux de TFCT passants lors du traitement est alors très faible. En conséquence, pour une composante proche de la limite de restauration, la diminution relative du temps de montée occasionnée par une réduction de la durée de la fenêtre est beaucoup plus faible que pour une composante située bien au dessus de la limite de restauration.

En conclusion, il faut retenir que pour un niveau de bruit fixé, la variation du temps de montée avec la durée de la fenêtre de TFCT utilisée lors du traitement n'est pas tout à fait linéaire. On peut toutefois admettre que le temps de montée est proportionnel à la durée de la fenêtre utilisée si la sinusoïde est situé nettement au dessus de la limite de restauration. Par contre, pour une sinusoïde proche de la limite de restauration, le gain lié à la diminution de la durée de la fenêtre est moins important. Le cas limite est celui d'une composante située à la limite de restauration pour laquelle, une diminution de la durée de la fenêtre d'analyse provoque une élimination totale du signal.

3.1.2.c Transitoires musicaux

Le point intéressant est maintenant de savoir comment les conclusions précédentes, que nous avons obtenues dans un cas de transitoire synthétique très simple, peuvent s'appliquer pour des signaux transitoires réels présents sur des enregistrements musicaux.

La figure 3.10 présente deux exemples de sons musicaux qui présentent des transitoires d'attaque brefs. Afin de pouvoir comparer cette figure avec les précédentes (figures 3.6 à 3.8), on a choisi de fixer la fréquence d'échantillonnage de telle manière que la longueur de la fenêtre utilisée pour ces exemples (512 points) soit équivalente à une durée de fenêtre moyenne couramment utilisée (ici 40 ms). Il est nécessaire de fixer une référence temporelle car la notion de transitoire n'a un sens que vis à vis d'une durée d'observation. Pour une fenêtre de traitement d'une durée de 40 ms, la figure 3.10 correspond donc à une durée d'observation équivalente à celle

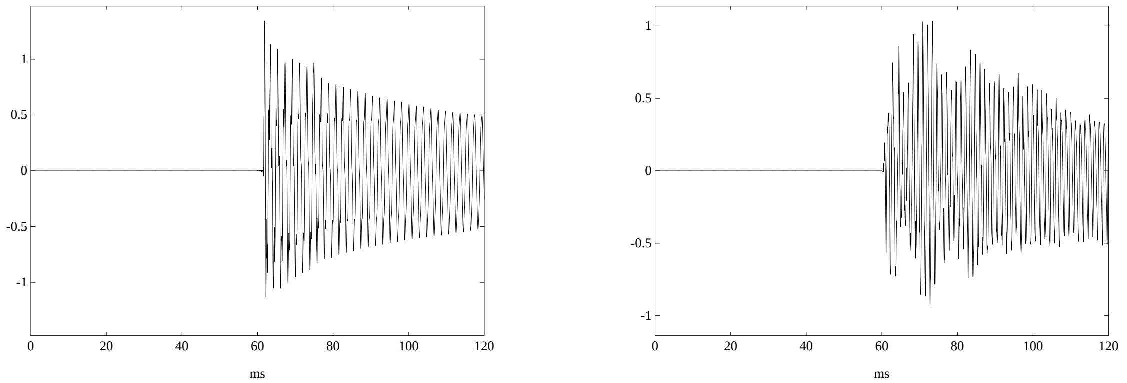

Figure 3.10: Exemples de transitoires d'attaque de sons percussifs. **A gauche,** xylophone (note C5, nuance de jeu forte). **A droite,** piano (note G5, nuance de jeu forte). La durée totale de signal représentée est dans les deux cas de 120 ms.

des figures précédentes. Dans ces conditions, il apparaît que pour ces deux sons, le transitoire est aussi bref que le transitoire synthétique utilisé pour étudier le comportement de la méthode de débruitage. En conséquence, les conclusions sont qualitativement les mêmes, la seule différence étant que les sons musicaux présentés contiennent plusieurs partiels de fréquences distinctes. En admettant que ces composantes du signal sont suffisamment éloignées (séparées d'au moins une dizaine de points de discrétisation fréquentielle associée à la TFCT), on peut appliquer les conclusions précédentes *pour chaque partiel* du signal. On note d'ailleurs que la puissance des partiels décroît en général avec la fréquence de ceux-ci. L'effet de lissage est donc plus important pour les composantes du son de fréquence élevée.

Toutefois, le modèle de transitoire présenté ne correspond pas au cas des composantes spectrales qui s'amortissent très rapidement (en une durée plus faible que celle de la fenêtre de TFCT). Or pour des signaux tels que ceux qui sont représentés sur la figure 3.10, il existe en général des composantes qui s'amortissent très rapidement. Tout ce que l'on peut dire, c'est qu'à niveau maximal comparable, la distorsion produite par le traitement est plus importante pour une composante sinusoïdale qui s'amortit très rapidement après le transitoire, car le spectre à court-terme correspondant est plus "étalé".

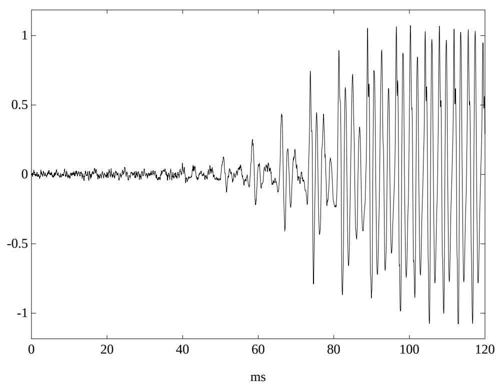

Figure 3.11: Exemple de transitoire d'attaque : son de saxophone (note C3). La durée totale de signal représentée est de 120 ms. Le signal présent à l'extrême gauche de la figure correspond à du bruit de mesure.

Les deux exemples de transitoires d'attaque présentés correspondent à des sons provenant d'instruments percussifs. Ces sons ont justement la particularité de présenter en général une attaque de durée très brève. Beaucoup de sons orchestraux présentent des transitoires beaucoup plus progressifs. Si on considère par exemple le signal représenté sur la figure 3.11, il est clair que l'effet de lissage sera quasiment inexistant dans le cas où une fenêtre de durée 40 ms est utilisée. En effet, il n'existe pratiquement aucune trame à court-terme pour laquelle le signal présente un contenu fréquentiel très étendu comme dans le cas d'un transitoire abrupt. Par conséquent le filtrage réalisé par le traitement de débruitage ne s'accompagne pas d'une distorsion du signal tant que le niveau de ce dernier est suffisant pour placer chacun des partiels au dessus de la limite de restauration.

En conclusion de cette partie consacrée aux limites de la méthode de restauration pour les signaux présentant des parties transitoires, rappelons les principaux résultats obtenus dans le cas du transitoire "sinusoïdal et abrupt" : Pour une composante sinusoïdale, le filtrage réalisé par le traitement de débruitage se traduit par un lissage des transitions abruptes sur une durée à peu près proportionnelle à la durée de la fenêtre d'analyse. Le temps de montée caractéristique de ce lissage est au plus de l'ordre de une fois et demi la durée de la fenêtre d'analyse. Il devient négligeable lorsque la composante se situe environ 30 dB au dessus de la limite de restauration dans sa partie stationnaire.

Pour des sons musicaux, ce lissage est constaté principalement pour des sons présentant des transitoires très brefs comme les sons percussifs. Il affecte principalement les partiels de fréquence élevée qui sont en général de puissance plus faible. La diminution de la durée de la fenêtre d'analyse permet de limiter ce lissage sauf dans le cas où le signal est très bruité, pour lequel une diminution de la durée de la fenêtre d'analyse se traduit par l'élimination totale des partiels de faible puissance.

3.2 Effets dus à la variance de l'estimation spectrale locale

Lors du paragraphe précédent sur les limites du débruitage, tous les raisonnements ont porté sur des quantités moyennes. Par exemple, l'équation (3.1) ne fournit que *l'espérance* du niveau relatif local dans le cas où le signal traité est composé d'une sinusoïde bruitée. Cependant, le niveau relatif mesuré localement dans une trame à court-terme donnée peut différer notablement de cette valeur moyenne du fait de la variance de l'estimation spectrale du bruit. Nous nous intéressons ici aux conséquences de cette variance sur les résultats du traitement.

La première de ces conséquences, qui est aussi la plus connue, concerne le bruit résiduel demeurant après traitement. On montre au paragraphe 3.2.1 que la variance de l'estimation spectrale du bruit se traduit par un bruit résiduel très peu naturel désigné en général sous le nom de *bruit musical*. De plus, l'étude précise des caractéristiques de l'estimation spectrale montre que la solution classique utilisée pour remédier à ce problème (la surestimation du niveau de bruit lors du traitement) provoque obligatoirement une forte distorsion du signal. Le deuxième point abordé (au paragraphe 3.2.2) concerne les résultats du traitement dans le cas d'un son stationnaire dont certains partiels sont de très bas niveau. Dans ce cas, la variance de l'estimation spectrale du bruit de fond se traduit par une fluctuation aléatoire de l'atténuation apportée par le traitement. Cette fluctuation contribue à moduler, de manière aléatoire, des composantes de signal qui sont stationnaires.

3.2.1 Bruit résiduel

3.2.1.a Comportement lors d'un instant de "silence"

Pour aborder l'étude du bruit résiduel, le plus simple est de considérer le cas d'un instant de silence, c'est à dire d'étudier le signal en sortie du système de restauration lorsque le signal d'entrée est constitué uniquement du bruit de fond. Les tests d'écoute concernant les systèmes de restauration par atténuation spectrale rapportés dans la littérature indiquent que le bruit résiduel (qui est dans ce cas simplement la sortie du système de restauration) est jugé comme étant très dérangeant et peu naturel par les auditeurs. Plus précisément les auteurs de [Moorer 86] mentionnent le fait que le traitement remplace le bruit de fond par *a swishing semimusical tonality* (les auteurs de [Ephraim 84] parlent quant à eux de *musical noise*). En français, c'est le terme de bruit musical qui est en général employé pour décrire cet effet [Bourdier 88]. Ce bruit résiduel "musical" apparaît avec toutes les règles de suppression ponctuelles décrites au paragraphe 2.2 [Bourdier 88] [Ephraim 84], et ce dès que le nombre de bandes (qui est équivalent à la longueur de la fenêtre d'analyse) est important [Moorer 86].

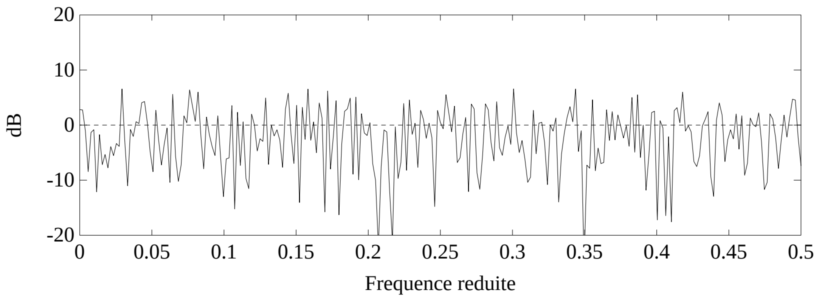

Figure 3.12: Estimation locale de la densité spectrale lors d'un instant de silence (bruit seul). En traits pointillés, le niveau moyen du bruit (ici un bruit blanc). La longueur de la fenêtre d'analyse est de 512 points.

Pour illustrer ce phénomène, la figure 3.12 présente une comparaison entre l'estimation spectrale locale pour une trame à court-terme située dans un instant de silence, et la valeur moyenne de cette estimation spectrale. Cette figure correspond au cas d'un bruit blanc pour des raisons de lisibilité. D'après ce qui a été dit au paragraphe 2.2, on peut considérer que la valeur moyenne de l'estimation spectrale est connue exactement grâce à la mesure du niveau de bruit effectuée avant le traitement proprement dit (cf. paragraphe 1.4.3). On rappelle que la différence entre ces deux estimations vient du fait que la mesure préalable du niveau de bruit se fait par moyennage sur un grand nombre de trames tandis que l'estimation locale se fait par périodogramme simple (module au carré de la transformée de Fourier à court-terme) sur une seule trame à court-terme. Il faut noter que c'est l'échelle en décibels qui fait apparaître visuellement la valeur moyenne comme étant surestimée. La figure 3.12 met en évidence les deux caractéristiques essentielles de l'estimation spectrale locale :

- D'une part, les très forts écarts qui existent par rapport à la valeur moyenne. Sur cet exemple, on note des valeurs allant jusqu'à 8 dB au dessus du niveau moyen (c'est à dire 6 fois supérieures à la moyenne), et inversement certaines valeurs sont inférieures de 20 dB à la valeur moyenne (soit 100 fois inférieures).
- D'autre part, l'estimation locale présente un aspect heurté sous la forme d'une succession de pics et de creux de faible largeur.

Ces deux caractéristiques traduisent des propriétés bien connues de l'estimation spectrale par périodogramme [Kay 88].

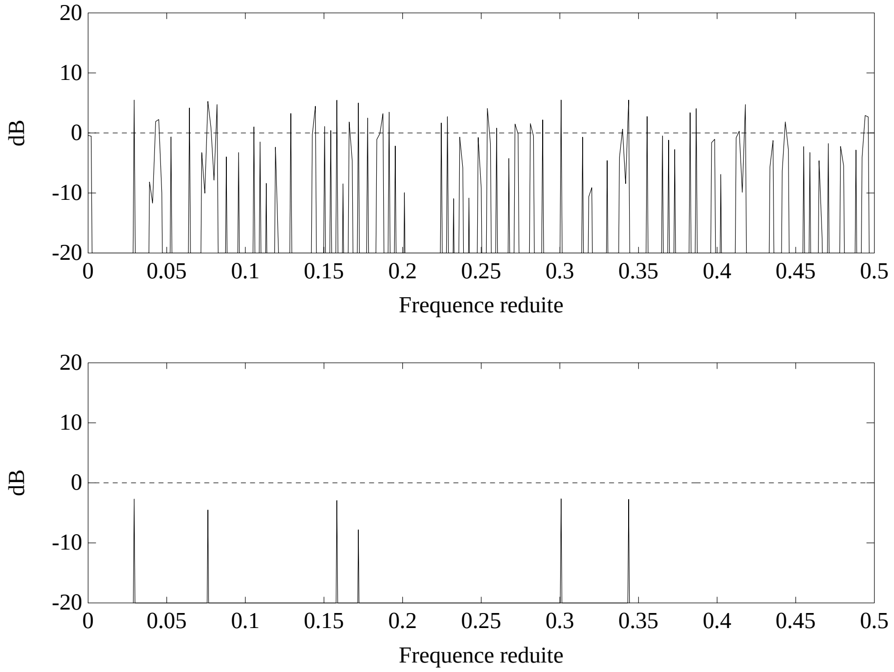

Figure 3.13: Spectre à court-terme après application de l'atténuation spectrale dans le cas d'un instant de silence (bruit seul). La règle de suppression utilisée est la soustraction en puissance. La longueur de la fenêtre d'analyse est de 512 points. **En haut**, résultat sans surestimation du niveau de bruit. **En bas**, résultat avec un facteur de surestimation du bruit de 6 dB.

Observons maintenant l'allure du spectre à court-terme, dans le même cas, mais après application de l'atténuation spectrale. En schématisant le fonctionnement d'une règle de suppression type, dès que la courbe en trait plein dépasse suffisamment la courbe en pointillés sur la figure 3.12, le canal fréquentiel correspondant est laissé inchangé tandis que dans le cas contraire, le canal fréquentiel est mis à zéro. Les résultats du traitement de la trame à court-terme de la figure 3.12 sont présentés par la figure 3.13 dans deux cas, selon que le niveau de bruit est ou non surestimé lors du traitement. La règle de soustraction en puissance est utilisée en tant qu'exemple, mais il faut savoir qu'on obtient qualitativement les mêmes résultats avec n'importe quelle autre règle de suppression ponctuelle. Ce qui est frappant dans le cas où le bruit de fond n'est pas surestimé (en haut de la figure 3.13) c'est qu'il n'y a pas à proprement parler d'élimination du bruit. En effet, du fait de la forte variance de l'estimation locale constatée sur la figure 3.12, il existe de nombreux canaux fréquentiels pour lesquels le niveau mesuré localement est bien supérieur au niveau moyen[4], ceux-ci ne sont donc pas atténués par le traitement. De plus, le spectre à court-terme après traitement présente un aspect particulier puisqu'il met en évidence un grand nombres de raies spectrales distinctement séparées par des zones d'énergies nulles. Ce dernier point est une conséquence directe de l'aspect heurté de l'estimation spectrale locale (figure 3.12). Enfin, on note que la surestimation du niveau de bruit (en bas de la figure 3.13) permet d'éliminer le bruit de fond de manière assez nette sur le plan énergétique (on constate une forte diminution de l'énergie totale du signal à court-terme traité). Cependant, cette surestimation ne fait qu'accentuer le second effet mentionné précédemment puisque le spectre du signal après traitement met très nettement en évidence quelques raies spectrales isolées. En

[4]Statistiquement, dans une trame à court-terme, près de 40% des points fréquentiels présentent un niveau supérieur à la valeur moyenne (voir la courbe 3.14).

conclusion, pour une trame prise lors d'un instant de silence, le traitement substitue au bruit de fond un signal de puissance beaucoup plus faible mais qui possède un spectre de raies.

Pour décrire plus précisément la nature du bruit résiduel lors d'un instant de silence, notons que si le bruit est stationnaire (et moyennant certaines hypothèses sur ses moments), les valeurs du périodogramme, prises en des points fréquentiels distincts, sont asymptotiquement indépendantes [Brillinger 81]. Etant donné que la longueur de la fenêtre d'analyse est dans notre cas assez importante (en tout cas supérieure à 64 points), on peut considérer que dans le cas d'un intervalle où seul le bruit est présent, les valeurs distinctes de l'estimation spectrale locale sont statistiquement indépendantes. C'est ce qui justifie l'allure heurtée de l'estimation spectrale locale (avant traitement) de la figure 3.12, et qui permet de dire que le spectre à court-terme après traitement présente forcément des caractéristiques analogues à celles des résultats de la figure 3.13. Au sein d'une trame à court-terme, le résultat du traitement est donc constitué d'une somme de composantes sinusoïdales. Si on considère maintenant des trames à court-terme successives, les valeurs de l'estimation spectrale locale sont pratiquement indépendantes dès que l'on s'intéresse à deux trames à court-terme qui ne se recouvrent pas. Les composantes sinusoïdales qui apparaissent dans le signal restauré ont donc une durée de vie moyenne qui ne dépasse pas (statistiquement) l'ordre de grandeur de la durée de la fenêtre d'analyse. C'est à dire que lorsque le signal bruité se réduit au bruit de fond, le signal restauré est constitué de composantes sinusoïdales intermittentes dont les fréquences sont distribuées aléatoirement et qui possèdent une durée de vie moyenne de l'ordre de la durée de la fenêtre d'analyse.

C'est exactement cet effet qui a été baptisé "bruit musical" dans la littérature. La sensation produite auditivement par le bruit musical dépend du nombre moyen de composantes sinusoïdales qui sont présentes. Mais, dans tous les cas, cet effet est extrêmement peu naturel, c'est à dire qu'il ne correspond pas du tout à un type de bruit que l'on est habitué à entendre sur un enregistrement de musique conventionnel. De plus il est particulièrement gênant, même à niveau très faible, et surtout dans le contexte d'un enregistrement de musique. Il semble en effet que ce bruit musical attire particulièrement l'attention de l'auditeur du fait de son caractère tonal (il est formé d'une somme de composantes sinusoïdales) et très instable (puisque sa composition varie plusieurs dizaines de fois par seconde). Il est important de souligner que l'apparition de bruit musical est inévitable même si le bruit de fond est parfaitement stationnaire. Ce sont les caractéristiques particulières de l'estimateur spectral à court-terme utilisé (le périodogramme) qui conditionnent la nature du bruit résiduel.

3.2.1.b Elimination du bruit musical par surestimation

Toujours en se plaçant dans le cas d'une règle de suppression ponctuelle, le seul moyen de lutte contre le bruit musical semble être la surestimation du niveau de bruit. D'après ce qui a été dit précédemment, le nombre de composantes sinusoïdales qui constituent le bruit musical décroît statistiquement lorsque le niveau de bruit est surestimé. Par contre, on peut considérer en première approximation que le niveau maximal de ces composantes, *mesuré sur le spectre à court-terme*, varie peu et qu'il se situe autour du niveau du bruit (voir la figure 3.13).

Niveau du bruit musical Une première remarque intéressante, c'est que *le niveau des composantes sinusoïdales qui constituent le bruit musical diminue lorsque la taille de la fenêtre d'analyse utilisée pour la TFCT augmente.* En effet, l' équation (C.14), établie dans l'annexe C, indique que le niveau (en puissance) du bruit vaut

$$P_d(\omega) \sum_n h(n)^2$$

Le niveau du pic associé à une composante sinusoïdale est donné par l'équation (C.9) comme étant

$$\frac{A^2}{4}\left[\sum_n h(n)\right]^2$$

Les composantes de bruit musical apparaissent sur le spectre à court-terme au plus au même niveau que le bruit (voir la figure 3.13). L'amplitude maximale des composantes sinusoïdales du bruit musical est donc donnée par la relation

$$A^2 = 4P_d(\omega)\frac{\sum_n h(n)^2}{\left[\sum_n h(n)\right]^2}$$

Où ω représente la pulsation de la composante de bruit musical considérée. Cette expression peut aussi s'écrire

$$A^2 = 4P_d(\omega)\frac{\Delta_h}{N} \tag{3.12}$$

Où Δ_h représente la largeur de la bande de bruit équivalente à la fenêtre $h(n)$, telle qu'elle est définie par la relation (C.18). On peut donc en conclure que l'amplitude des composantes sinusoïdales qui constituent le bruit musical est proportionnelle à $\sqrt{N}$ (c'est à dire que leur niveau diminue de 3 dB lorsque la taille de la fenêtre double). Par contre lorsque la taille de la fenêtre est multipliée par deux, on observe statistiquement deux fois plus de composantes sinusoïdales dans une trame à court-terme (puisque la propriété d'indépendance statistique des canaux fréquentiels reste valable). L'effet combiné des ces deux points est de fournir un bruit résiduel de même puissance quelle que soit la taille de la fenêtre. Cependant comme le niveau des composantes sinusoïdales qui constituent le bruit musical diminue lorsque la taille de la fenêtre augmente, une fenêtre d'analyse plus longue se traduit souvent par une diminution du niveau du bruit musical lors de l'audition. Ceci est le cas lorsque le facteur de surestimation est suffisamment grand (au moins de 6 dB), car le nombre de composantes sinusoïdales constituant le bruit musical est alors faible, et dans ce cas, l'amplitude de chaque composante est perceptivement plus significative que le nombre de composantes puisqu'elles sont entendues distinctement. Toutefois cette diminution reste assez faible si on s'en tient à des valeurs raisonnables de la taille de fenêtre. De plus cette limitation de niveau ne constitue en général pas une solution efficace au problème du bruit musical qui reste inacceptable, même à niveau réduit, du fait de sa nature particulière.

Effet de la surestimation Expérimentalement, il apparaît que le nombre moyen de composantes sinusoïdales qui constituent le bruit musical diminue lorsque la surestimation du niveau de bruit augmente (figure 3.13). Par contre, on a vu que le niveau maximal de ces composantes n'est pas affecté par la surestimation du bruit. La seule possibilité d'éliminer le bruit musical par surestimation du bruit consiste donc à éliminer totalement tout bruit résiduel en rendant l'apparition d'une composante sinusoïdale impossible. On conçoit aisément qu'au delà d'une certaine limite de surestimation, le bruit résiduel disparaisse totalement, il suffit pour cela de porter le niveau relatif de coupure au dessus des valeurs maximales de l'estimation spectrale locale (les valeurs maximales de la figure 3.12).

Pour définir plus précisément la limite à partir de laquelle l'apparition d'une composante de bruit musical devient extrêmement peu probable, notons que les valeurs du périodogramme d'un bruit stationnaire, normalisées par leur moyenne, sont asymptotiquement distribuées comme des variables de loi $\chi^2(2)/2$ (où $\chi^2(n)$ désigne une loi de χ^2 à n degrés de liberté) [Brillinger 81]. La loi $\chi^2(2)/2$ correspond simplement à une loi exponentielle de moyenne 1 dont la densité de probabilité s'écrit

$$f(x) = \mathrm{e}^{-x} \quad \text{pour } x \in [0, +\infty[$$

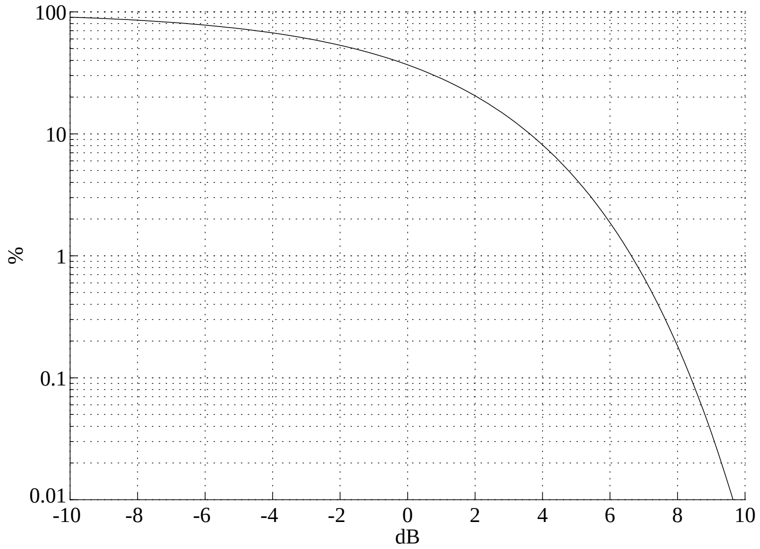

Figure 3.14: Fonction de répartition modifiée du niveau relatif local ($Prob\{\mathcal{Q}(p,\omega_k) \geq x\}$), dans le cas d'un instant de silence (bruit seul). En abscisse, le niveau relatif est indiqué en décibels, en ordonnée, les pourcentages sont représentés selon une échelle logarithmique.

Lorsque qu'il s'agit d'analyser un instant de silence, la valeur de l'estimation spectrale locale normalisée par sa moyenne correspond à ce que nous avons baptisé niveau relatif local $\mathcal{Q}(p,\omega_k)$ défini par l'équation (2.5). *Durant un instant de silence, la densité de probabilité du niveau relatif local est donc donnée par une loi exponentielle.* On en déduit facilement la probabilité de trouver des valeurs du niveau relatif local supérieures à une limite donnée

$$Prob\{\mathcal{Q}(p,\omega_k) \geq \mathcal{Q}_{lim}\} = \mathrm{e}^{-\mathcal{Q}_{lim}}$$

C'est cette fonction de répartition modifiée (en général on considère plutôt les valeurs inférieures à une limite donnée) qui est représentée sur la figure 3.14. Notons que l'aspect convexe de la courbe représentée est simplement due à l'utilisation d'échelles logarithmiques. D'après cette figure, il apparaît qu'une valeur du niveau relatif de coupure d'au moins 8,5 dB (soit une valeur de 7 en échelle linéaire) est nécessaire pour garantir une probabilité de dépassement inférieure à 0.1%. Cependant même une probabilité de dépassement très faible de cette ordre n'est pas suffisante pour éliminer complètement le bruit musical car la longueur de la fenêtre d'analyse intervient aussi : considérons, par exemple, que la durée de la fenêtre est de 1024 points, il existe 513 points fréquentiels pour lesquels le niveau relatif local de coupure est susceptible d'être dépassé. On a vu que ces points fréquentiels sont indépendants, la loi de probabilité qui décrit le nombre de points fréquentiels qui dépassent le niveau de coupure est donc une loi binomiale. Dans ces conditions (niveau relatif de coupure égal à 7 et fenêtre de 1024 points), on obtient encore une probabilité de 40% d'obtenir au moins une composante de bruit musical dans une trame à court-terme donnée.

En pratique il faut se souvenir que le niveau relatif de coupure est toujours légèrement supérieur au facteur de surestimation du bruit (de 3 dB environ pour la règle de suppression en puissance). C'est à dire que si le facteur de surestimation vaut 9 dB, seuls les points fréquentiels dont le niveau relatif est inférieur à 9 dB vont être mis à zéro, cependant, les valeurs de niveau relatif comprises entre 9 et 12 dB provoquent déjà une forte atténuation (voir le paragraphe 2.2). Ceci explique que **pour un bruit stationnaire, la valeur du facteur de surestimation de 9 dB garantit une élimination complète du bruit résiduel** (donc a fortiori de sa qualité "musicale").

3.2.1.c Bruit résiduel en présence de signal

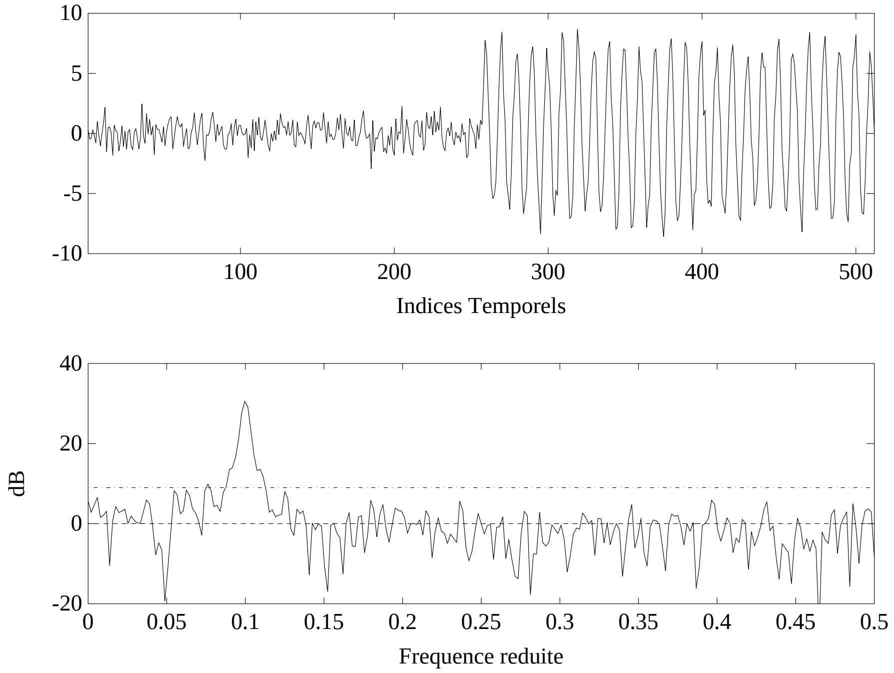

Figure 3.15: Exemple de trame à court-terme contenant un signal transitoire bruité. **En haut**, le signal à court-terme (avant pondération par la fenêtre d'analyse). **En bas**, le spectre à court-terme correspondant. Le trait en pointillés représente le niveau moyen de bruit (ici un bruit blanc). En traits mixtes, la valeur du niveau relatif local correspondant à 9 dB. La fenêtre d'analyse est une fenêtre de Hamming.

En dehors des instants de silence où seul le bruit est présent, le problème reste essentiellement le même : toute zone du spectre à court-terme qui correspond au bruit de fond seul provoque du bruit musical. On observe alors, dans une bande de fréquence donnée, une situation semblable à celle qui a été décrite au paragraphe précédent. Le bruit musical est généralement moins audible en présence de signal du fait des effets de masquage perceptifs (dit d'une autre façon, l'audibilité du bruit musical ne dépend plus uniquement des seuils absolus d'audition dans ce cas). En particulier, la diminution de la sensation de bruit musical lorsque la taille de la fenêtre augmente est plus nette puisqu'une composante qui voit son amplitude diminuer peut devenir totalement masquée par le signal. Toutefois, même en présence d'un signal de fort niveau, le bruit musical constitue une telle gêne auditive qu'il reste nécessaire d'éliminer totalement le bruit musical par surestimation, dès que l'on utilise une règle de suppression ponctuelle. Il faut toutefois souligner que cette élimination du bruit résiduel par surestimation du niveau de bruit est loin d'être entièrement satisfaisante : si on compare, sur la figure 2.5 (paragraphe 2.2), la caractéristique de suppression correspondant à un facteur de surestimation de 8 dB avec la caractéristique non surestimée, il apparaît clairement qu'une surestimation de cet ordre risque de venir fortement atténuer certaines parties du spectre à court-terme du signal musical lui même.

De plus, une surestimation du niveau de bruit de l'ordre de 8 à 9 dB s'avère souvent insuffisante pour dissiper toute sensation de bruit musical. En particulier, il arrive d'entendre des petites bouffées de bruit musical à l'écoute du signal traité, par exemple à l'occasion du transitoire d'attaque d'un son. Ceci souligne le fait que le bruit musical, bien qu'il apparaisse même en l'absence de signal, n'est pas totalement indépendant du signal sonore. Pour illustrer simplement cet effet, considérons un signal analysé du type de celui qui est représenté dans la moitié supérieure de la figure 3.15. Il s'agit d'un signal transitoire sommaire formé d'un son pur qui émerge d'un bruit blanc. Intuitivement, le spectre du signal correspondant occupe une bande fréquentielle large du fait de la "rupture" temporelle créée par la partie transitoire. Le problème est que les valeurs faibles du spectre du signal, qui viennent s'ajouter en moyenne au périodogramme du bruit, contribuent à relever légèrement le spectre du bruit. Ainsi sur le bas de la figure 3.15, la limite correspondant à un niveau relatif local de 9 dB, dont on a vu au paragraphe précédent qu'elle est suffisante dans le cas du bruit seul, est dépassée localement au voisinage de la zone de forte énergie qui correspond au signal. Après application de l'atténuation spectrale, des composantes de bruit musical risquent donc d'apparaître, même avec un facteur de surestimation de 9 dB. Sur la figure 3.15 c'est par exemple le cas autour de la fréquence normalisée 0,008.

3.2.1.d Effet d'une erreur d'estimation du niveau de bruit

En fait, le phénomène précédent est susceptible de se produire dès qu'il existe un signal qui présente sur son spectre à court-terme des zones d'un niveau comparable à la densité spectrale du bruit de fond. En effet, le signal vient alors rehausser le niveau moyen du bruit de fond dans une certaine bande de fréquence, pas suffisamment pour émerger après application de l'atténuation spectrale, mais suffisamment pour faire apparaître localement des composantes de bruit musical. Nous avons en effet constaté lors de l'écoute qu'il arrive qu'un signal de faible niveau et de spectre relativement peu coloré, comme par exemple une respiration dans un signal de parole ou bien un bruit impulsionnel, génère du bruit musical bien que la surestimation utilisée soit théoriquement suffisante pour éliminer complètement le bruit résiduel en l'absence de signal.

Ce phénomène se produit malheureusement dans un autre cas très important : lorsque le niveau estimé du bruit est inférieur à la valeur réelle. Admettons par exemple que le bruit de fond augmente légèrement en cours de l'enregistrement, le facteur de surestimation de 9 dB devient insuffisant puisque le niveau de bruit mesuré localement augmente, et par conséquent, le bruit musical réapparaît. Ce problème se pose très fréquemment puisque les bruits réels présents sur les enregistrements anciens ne sont généralement pas stationnaires. Nous avons vu au paragraphe 1.4.5 que c'est en particulier le cas avec les enregistrements provenant de transferts de disques. Une augmentation locale du niveau de bruit de l'ordre de 3 à 6 dB est tout à fait réaliste (voir par exemple la figure 1.10). Il apparaît dans ces conditions que pour éviter des bouffées locales de bruit musical à chaque irrégularité du bruit de fond, il devient nécessaire de surestimer le niveau de bruit non plus de 9 dB, mais d'environ 13 dB. Cependant il est clair qu'un système de restauration par atténuation spectrale où l'estimation de la densité spectrale du bruit est multipliée par un facteur 20 (13 dB) apporte obligatoirement de très fortes distorsions au signal musical. Un facteur de surestimation inférieur (de l'ordre de 8 dB) peut très bien convenir dans le cas d'un bruit stationnaire. Cependant une simple surestimation de cet ordre ne permet pas d'éviter l'apparition locale de bruit musical qui peut être due soit aux irrégularités du bruit, soit directement au signal enregistré.

La conclusion de ce paragraphe concernant le bruit musical, est que **la surestimation du niveau de bruit ne peut pas constituer à elle seule une solution au problème de la nature du bruit résiduel**. Nous avons vu que la surestimation ne permet pas d'éliminer le caractère "musical" (présence de sinusoïdes) du bruit résiduel. De plus, le facteur de surestimation nécessaire pour éliminer tout bruit résiduel s'avère être trop élevé pour ne pas entraîner de fortes distorsions du signal. Cette situation est une conséquence de la variance trop importante de l'estimation spectrale locale. En pratique, ce problème est aggravé par le fait qu'une règle de suppression ponctuelle ne permet pas de se prémunir contre les irrégularités du bruit de fond. Ces deux points sont abordés de nouveau au paragraphe 4.1.2 dans le cadre de la règle de suppression proposée par Ephraim et Malah.

3.2.2 Estimation des composantes de signal fortement bruitées

Les fluctuations mises en évidence précédemment pour les parties du spectre à court-terme qui correspondent au bruit peuvent aussi se produire dans des parties du spectre qui contiennent des composantes de signal de faible amplitude. Supposons par exemple que le signal à traiter soit formé d'un son pur bruité, situé légèrement au dessus du niveau de bruit. Il faut se souvenir que la formule (3.1) ne fournit qu'une *valeur moyenne* du niveau relatif mesuré au sommet du pic spectral. Dans chaque trame à court-terme, le niveau relatif local diffère notablement de cette valeur moyenne du fait de la forte variance de l'estimation spectrale locale due au bruit. Par conséquent, l'atténuation spectrale apportée au signal n'est pas constante, elle varie aléatoirement d'une trame à l'autre.

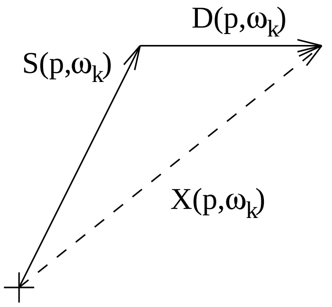

Figure 3.16: Spectre à court-terme du signal bruité. Représentation sous la forme du diagramme de Fresnel pour la pulsation ω_k et l'indice de trame p.

La détermination de la distribution des valeurs du périodogramme en présence de signal est plus difficile que dans le cas du paragraphe 3.2.1.b où seul le bruit est présent. En effet, le diagramme de Fresnel de la figure 3.16 montre que le module au carré du spectre s'écrit

$$|X(p,\omega_k)|^2 = |S(p,\omega_k) + D(p,\omega_k)|^2 \tag{3.13}$$

où $S(p,\omega_k)$ est la partie déterministe associée au signal tandis que $D(p,\omega_k)$ est la perturbation due au bruit. Nous avons indiqué au paragraphe 2.2.3 que l'on peut considérer que la transformée de Fourier du bruit est une variable gaussienne complexe, c'est à dire que sa partie réelle $D_r(p,\omega_k)$ et sa partie imaginaire $D_i(p,\omega_k)$ sont deux variables gaussiennes centrées, indépendantes, et de même variance $\mathrm{E}\left\{|D(p,\omega_k)|^2\right\}/2$ [Brillinger 81]. La propriété que nous avons exploitée au paragraphe 3.2.1.b est le fait que $|D(p,\omega_k)|^2$ suit alors une loi du χ^2 à deux degrés de liberté (somme des carrés de deux variables gaussiennes) [Papoulis 91]. En utilisant la relation (3.13), le niveau relatif $\mathcal{Q}$ défini par (2.9) se met sous la forme suivante

$$\mathcal{Q} = \frac{|S+D|^2}{\mathrm{E}\left\{|D|^2\right\}}$$

Etant entendu que l'on abandonne provisoirement les indices p et ω_k de la TFCT afin de ne pas surcharger les notations. Pour la même raison, la variance de la TFCT du bruit $\mathrm{E}\left\{|D|^2\right\}$ sera simplement notée v dans la suite de cette partie.

Notons qu'il est possible de considérer sans perte de généralité que la *composante de signal S est est réelle et positive* car une rotation de la figure 3.16 n'affecte pas la valeur du module de X. Avec cette hypothèse le niveau relatif s'écrit donc

$$\mathcal{Q} = \frac{(D_r + S)^2 + D_i^2}{v} \tag{3.14}$$

Nous savons que la distribution de probabilité conjointe des variables gaussiennes indépendantes D_r et D_i est donnée par [Kay 93, §15]

$$g(D_r, D_i) = \frac{1}{\pi v} e^{-\left[\frac{D_r^2 + D_i^2}{v}\right]}$$

Afin d'obtenir la densité de probabilité de $\mathcal{Q}$, on effectue le changement de variable classique [Papoulis 91] $(D_r, D_i) \mapsto (\mathcal{Q}, \theta)$, où $\mathcal{Q}$ et θ sont définies par

$$\left\{ \begin{array}{rcl} D_r & = & \sqrt{v}\sqrt{\mathcal{Q}}\cos\theta - S \\ D_i & = & \sqrt{v}\sqrt{\mathcal{Q}}\sin\theta \end{array} \right. \tag{3.15}$$

Le Jacobien de ce changement de variable est défini par

$$\det \begin{pmatrix} \dfrac{\partial D_r}{\partial \mathcal{Q}} & \dfrac{\partial D_i}{\partial \mathcal{Q}} \\ \\ \dfrac{\partial D_r}{\partial \theta} & \dfrac{\partial D_i}{\partial \theta} \end{pmatrix}$$

En utilisant les relations (3.15), on trouve que le Jacobien vaut

$$\det \begin{pmatrix} \dfrac{1}{2}\dfrac{\sqrt{v}}{\sqrt{\mathcal{Q}}}\cos\theta & \dfrac{1}{2}\dfrac{\sqrt{v}}{\sqrt{\mathcal{Q}}}\sin\theta \\ \\ -\sqrt{v}\sqrt{\mathcal{Q}}\sin\theta & \sqrt{v}\sqrt{\mathcal{Q}}\cos\theta \end{pmatrix} = \frac{v}{2}$$

La distribution de probabilité conjointe de $\mathcal{Q}$ et θ s'écrit donc

$$f(\mathcal{Q}, \theta) = \frac{v}{2}\, g(D_r, D_i) = \frac{1}{2\pi} e^{-\left[v\mathcal{Q} - 2\sqrt{v}S\sqrt{\mathcal{Q}}\cos\theta + S^2\right]/v} \tag{3.16}$$

Cette équation peut se réécrire en faisant intervenir la valeur moyenne du niveau relatif local $\bar{\mathcal{Q}}$ qui, du fait de la décorrélation entre le signal et le bruit, vaut

$$\bar{\mathcal{Q}} = \frac{S^2 + v}{v} = S^2/v + 1$$

Avec cette notation, la relation (3.16) devient

$$f(\mathcal{Q}, \theta) = \frac{1}{2\pi} e^{-\left[\mathcal{Q} - 2\sqrt{\mathcal{Q}(\bar{\mathcal{Q}} - 1)}\cos\theta + (\bar{\mathcal{Q}} - 1)\right]}$$

Finalement la densité de probabilité du niveau relatif $\mathcal{Q}$ s'obtient en intégrant sur toutes les valeurs possibles de la variable θ :

$$f(\mathcal{Q}) = e^{-[\mathcal{Q} + (\bar{\mathcal{Q}} - 1)]} \times \frac{1}{2\pi} \int_0^{2\pi} e^{2\sqrt{\mathcal{Q}\,(\bar{\mathcal{Q}} - 1)}\cos\theta}\, d\theta \tag{3.17}$$

On reconnaît dans le terme de droite (intégrale) une fonction de Bessel modifiée d'ordre 0 [Spiegel 68] [Papoulis 91]. La densité de probabilité du niveau relatif local $\mathcal{Q}$ s'écrit donc

$$f(\mathcal{Q}) = e^{-[\mathcal{Q}+(\bar{\mathcal{Q}}-1)]} I_0 \left(2\sqrt{\mathcal{Q}\,(\bar{\mathcal{Q}}-1)} \right) \tag{3.18}$$

où $I_o(x)$ représente la fonction de Bessel modifiée d'ordre 0.

La figure 3.17 présente les densités de probabilité calculées grâce à la relation (3.18) pour différentes valeurs de $\bar{\mathcal{Q}}$. Dans cette représentation *(abscisses représentées en décibels)*, on constate sur cette figure que la densité de probabilité du niveau relatif $\mathcal{Q}$ est d'autant plus concentrée autour de la valeur moyenne $\bar{\mathcal{Q}}$ que cette dernière est plus élevée.

Pour préciser ce dernier point, nous avons évalué les fonctions de répartition correspondant aux densités de probabilité définies par la formule (3.18) par intégration numérique (somme de Riemann). L'évaluation discrète de la fonction de répartition obtenue a ensuite été utilisée pour déterminer l'intervalle de confiance à 99,9% pour la valeur du niveau relatif $\mathcal{Q}$. Les bornes de l'intervalle ont été déterminées comme les abscisses des points d'intersection de la fonction de répartition avec les niveaux 0,05% et 99,95%. Le pas de discrétisation de la fonction de répartition a été fixé dans chaque cas afin d'obtenir une précision sur les bornes de l'intervalle meilleure que ± 1 dB. Les intervalles obtenus sont représentés sur la figure 3.18 en fonction du niveau relatif moyen $\bar{\mathcal{Q}}$.

Ces intervalles de confiance permettent de préciser quelles sont les conséquences, pour le signal traité, des fluctuations du niveau relatif mesuré. Selon le niveau relatif moyen de la composante de signal considérée, on peut distinguer plusieurs cas :

- Pour les composantes dont le niveau relatif moyen $\bar{\mathcal{Q}}$ est inférieur à 10 dB, le niveau relatif mesuré présente la particularité de devenir très faible dans certaines trames. La figure 3.18 montre en effet que tant que $\bar{\mathcal{Q}} \leq 10$ dB (six premiers intervalles), le niveau relatif $\mathcal{Q}$ peut être inférieur à 3 dB avec une probabilité non-négligeable. Lorsque le niveau relatif mesuré est aussi faible, la fréquence correspondant à la composante de signal est totalement atténuée. Ce qui se traduit, dans le signal restauré, par une disparition totale de la composante de signal dans certaines trames à court-terme. Une composante de niveau relatif moyen inférieur à 10 dB n'est donc présente que de manière intermittente dans le signal restauré

- Pour les niveau relatifs moyens entre 10 et 20 dB, le niveau relatif mesuré est dans tous les cas suffisant pour éviter l'atténuation totale de la composante. Cependant, les fluctuations du niveau relatif mesuré ne sont pas encore totalement négligeables. Par exemple sur la figure 3.18, l'intervalle de confiance correspondant à une valeur moyenne de $\bar{\mathcal{Q}} = 14$ dB est [10 dB,17 dB]. Cette écart peut selon la règle de suppression utilisée se traduire par une fluctuation de l'atténuation apportée : la figure 2.3 (chapitre 2) montre que pour la soustraction en puissance la différence entre $G(10$ dB$)$ et $G(17$ dB$)$ est négligeable, par contre elle est de l'ordre de 3 dB pour la soustraction spectrale (figure 2.3) ou pour la soustraction en puissance avec surestimation de 8 dB (figure 2.5). Ces fluctuations, bien que limitées, ne sont pas forcément négligeables car elles peuvent être audibles [Botte 88].

- Enfin, pour les composantes de niveau relatif moyen de l'ordre de 20 dB, on peut négliger les fluctuations du niveau mesuré car on se situe fortement au dessus du niveau de coupure (quelle que soit la règle considérée).

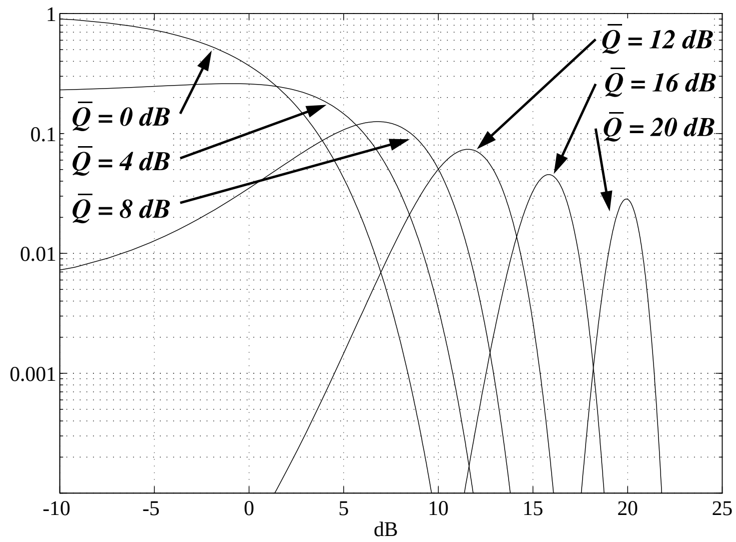

Figure 3.17: Densité de probabilité du niveau relatif local $f(\mathcal{Q})$ pour différentes valeurs moyennes $\bar{\mathcal{Q}}$ (de gauche à droite, 0, 4, 8, 12, 16 et 20 dB). En abscisse, le niveau relatif $\mathcal{Q}$ est représenté en décibels.

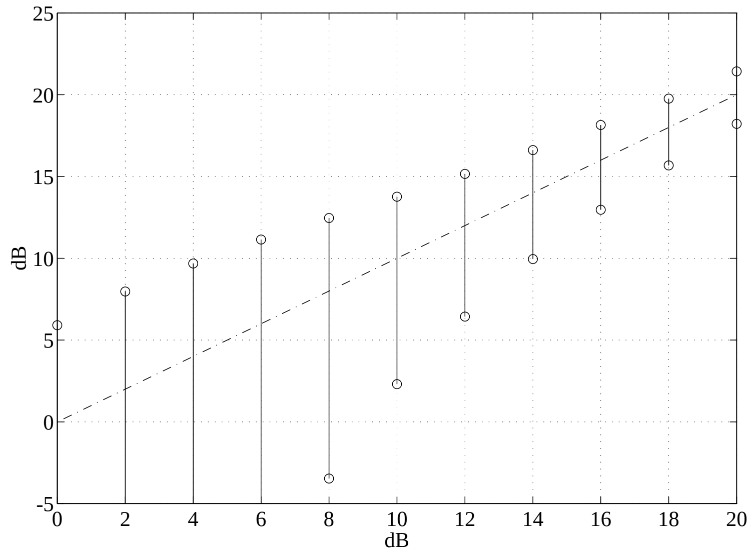

Figure 3.18: Intervalle de confiance à 99,9% pour la valeur du niveau relatif $\mathcal{Q}$ en fonction du niveau relatif moyen $\bar{\mathcal{Q}}$. Le trait mixte rappelle la valeur moyenne $\bar{\mathcal{Q}}$. Pour les quatres premiers intervalles ($\bar{\mathcal{Q}} =$ 0, 2, 4 et 6 dB), la borne inférieure de l'intervalle n'est pas visible.

En pratique, ces fluctuations peuvent altérer le timbre des sons traités même si elles ne sont pas perçues spécifiquement : si on imagine en effet le traitement d'un signal harmonique bruité, seules les composantes de faible niveau (en général dans les hautes fréquences) fluctuent en sortie du système de restauration, tandis que les composantes de niveau plus important restent stables. On sait qu'un tel phénomène influence la manière dont le son complet est perçu [Moore 82, §6].

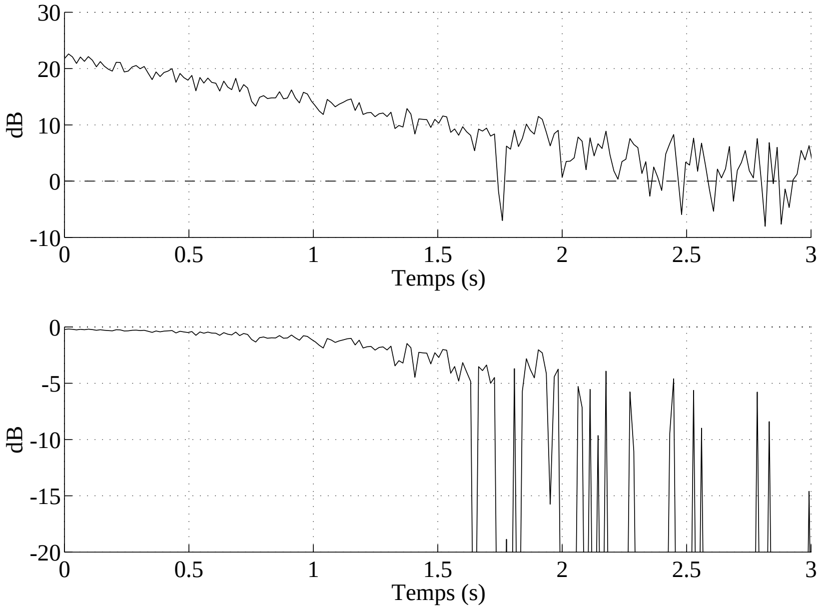

Figure 3.19: Traitement d'une composante sinusoïdale d'amplitude décroissante, noyée dans un bruit : valeurs du niveau relatif **(en haut)**, et de l'atténuation spectrale **(en bas)**, pour la fréquence correspondant à la sinusoïde, en fonction de la position temporelle de la trame à court-terme. La durée des trames à court-terme est de 30 ms (recouvrement 50%). La règle de suppression utilisée est la soustraction en puissance, avec un facteur de surestimation du bruit de 6 dB.

Le cas d'une composante sinusoïdale isolée, de niveau décroissant, représenté sur la figure 3.19 permet bien de mettre en évidence les différents types de fluctuations qui viennent d'être décrits. On distingue nettement sur cette figure trois étapes successives. La première correspond à la zone où le niveau relatif, mesuré à la fréquence de la composante, est compris entre 10 et 20 dB (entre 0,5 et 1,5 secondes sur la figure 3.19). Dans cette zone, l'atténuation apportée à la fréquence correspondant à la sinusoïde présente des fluctuations de l'ordre de quelques décibels, qui deviennent de plus en plus importantes au fur et à mesure que le niveau de la composante diminue. Dans une seconde zone (entre 1,5 et 2 secondes), l'atténuation devient extrêmement chahutée. Cette zone correspond en effet à des valeurs moyennes du niveau relatif inférieures à 10 dB, pour lesquelles nous avons vu que la composante peut être totalement atténuée dans certaines trames. Enfin, la dernière zone correspond au cas où le niveau relatif moyen est inférieur au facteur de surestimation utilisé, ce qui se traduit par une atténuation totale dans la plupart des trames à court-terme. En fait, on peut considérer que le niveau de la composante est négligeable à partir de 2,5 secondes. Les valeurs du gain non nulles, observées entre 2,5 et 3 secondes, correspondent donc au phénomène de bruit musical décrit au paragraphe 3.2.1. Une remarque à ce propos, est que dans le cas où le niveau de bruit n'est pas surestimé, seulement 60% des trames à court-terme correspondant au bruit seul sont fortement atténuées (cf. figure 3.14). Par conséquent, lorsque le bruit de fond n'est pas surestimé, un diagramme tel que celui de la figure 3.19 devient visuellement illisible, ce qui explique la surestimation, assez forte, utilisée ici.

3.3 Influence de la phase à court-terme

Ce dernier paragraphe à propos des résultats du débruitage par atténuation spectrale à court-terme concerne les problèmes posés par le spectre de phase à court-terme. La pertinence de cette question vient du fait que l'atténuation spectrale agit uniquement sur le module du spectre à court-terme. La phase du spectre à court-terme demeure inchangée, or on sait que le bruit implique aussi des perturbations de la phase. Il est donc légitime de se demander quelles sont les conséquences des perturbations de la phase du spectre à court-terme dues au bruit. Notons que l'analogie de la TFCT avec un banc de filtres permet de formuler ce problème d'une autre façon qui s'avère être fort intéressante : le débruitage consiste, dans chaque voie du banc de filtres, à ajuster une *atténuation variable qui compense l'augmentation de puissance due au bruit de fond*. C'est à dire que la modification apportée au signal de sous-bande est caractérisée de manière uniquement énergétique. Ce qui implique que les aspects liés à la *forme temporelle* du signal de sous-bande ne sont pas explicitement traités.

3.3.1 Sons stationnaires

Comme au paragraphe 3.1.1, on commence par s'intéresser au cas le plus simple qui est celui du son pur bruité. On suppose que le niveau de la composante sinusoïdale est suffisant, de telle façon que celle-ci soit située au dessus de la limite de restauration compte tenu des paramètres utilisés pour la TFCT.

3.3.1.a Nature de l'effet de modulation

Dans la littérature, c'est Lagadec et Pelloni qui mentionnent les premiers le fait que, dans cette situation (son pur bruité), *le signal obtenu en sortie du système de restauration est constitué du son pur modulé par un bruit* [Lagadec 83]. Cette constatation est justifiée de manière simple par les auteurs en utilisant le formalisme du banc de filtres. Dans [Vary 85], P. Vary met en évidence le même phénomène en adoptant une présentation différente puisqu'il se place plutôt dans l'analogie de la transformée à court-terme. P. Vary montre que la modulation du son pur en sortie du système est directement liée au bruit qui perturbe le spectre de phase. Pour donner une idée du raisonnement qui conduit à ce résultat, notons les points suivants :

1. On peut considérer que le module du spectre à court-terme de la sinusoïde est correctement estimé dès lors que le niveau de celle-ci la situe suffisamment au dessus de la limite de restauration.
2. Par contre, la phase du spectre à court-terme autour du pic spectral reste perturbée par le bruit. Cette perturbation est d'autant plus importante que le niveau relatif est faible. On peut s'en convaincre en notant sur le diagramme de Fresnel de la figure 3.16 que le déphasage maximal entre $S(p, \omega_k)$ et $X(p, \omega_k)$ augmente quand le rapport $|S(p, \omega_k)| / |D(p, \omega_k)|$ diminue.
3. Or pour une sinusoïde, les variations de la phase du spectre à court-terme au sommet du pic spectral donnent accès, en première approximation, à la loi de variation de la fréquence instantanée (c'est le principe du vocodeur de phase [Moorer 78]).

La conclusion est donc que le signal en sortie est modulé en fréquence, et ce avec un indice d'autant plus grand que le niveau relatif local mesuré au sommet du pic est faible. Ceci reste valable tant que la sinusoïde est située au dessus de la limite de restauration puisque dans le cas contraire la composante sinusoïdale est totalement éliminée.

Malheureusement, cette constatation ne suffit pas pour caractériser complètement l'influence de la phase à court-terme. En effet, les tests d'écoute que nous avons effectués ont mis en évidence deux situations distinctes. D'une part, cet effet de modulation n'est pas audible lorsqu'on utilise une fenêtre de TFCT de durée suffisante (40 ms ou plus). D'autre part, dans les cas où la fenêtre de TFCT est de courte durée (en dessous de 20 ms), la modulation de la sinusoïde est clairement audible dans le signal traité, *même pour des composantes sinusoïdales situées bien au dessus de la limite de restauration*. Pour expliquer ces constatations expérimentales, il ne suffit pas de caractériser la modulation due aux perturbations de la phase à court-terme par un indice de modulation, il est en plus nécessaire de déterminer le contenu spectral du signal modulant [Cappe 91].

Cependant nous n'avons pas poursuivi cette démarche jusqu'au bout, tout d'abord parce que l'écriture analytique exacte du signal modulant s'est avérée très complexe, et surtout du fait que les résultats obtenus ne pouvait pas être reliés simplement à des données perceptives connues. En effet, il apparaît qu'en sortie du traitement la sinusoïde est modulée par un bruit, or nous ne disposions pas de données psychoacoustiques concernant les modulations par des bruits.

Mais cette carence n'est qu'apparente puisqu'il suffit de présenter le même problème de façon différente pour se retrouver dans un des domaines classiques de la psychoacoustique : le masquage d'une bande de bruit par un son pur. En effet, un son pur modulé par un bruit en bande de base est équivalent à un son pur additionné à une bande de bruit centrée autour de la fréquence de la sinusoïde (en utilisant la décomposition en phase et quadrature d'un processus aléatoire à bande étroite [Charbit 90, §III.2.9]). Il faut noter que cette nouvelle présentation du problème induit l'utilisation du formalisme du banc de filtres pour la TFCT. En effet, on vient de dire que si l'entrée du système de restauration est un son pur noyé dans un bruit large bande, la sortie se compose du son pur ainsi que d'un bruit à bande étroite centré autour de la fréquence du son pur. La caractérisation analytique la plus simple de cet effet est obtenue en considérant le filtrage équivalent effectué par le système de restauration. C'est la raison pour laquelle il est plus efficace d'utiliser ici l'analogie du banc de filtres pour la TFCT. Il faut noter que cette présentation, plus simple, qui ne fait pas directement intervenir la notion de phase à court-terme, est celle qu'avaient adoptée les auteurs de [Lagadec 83].

Il faut toutefois faire attention à cette notion de bande de bruit : il ne s'agit pas d'un bruit résiduel stable, ce bruit est fortement liée au signal puisque la bande de fréquence qu'il occupe se déplace quand la fréquence de la sinusoïde varie. C'est pourquoi, dès lors que l'on s'intéresse à des sons musicaux qui ne sont pas complètement stationnaires, le terme de modulation correspond bien à la sensation auditive produite par cet effet. Cependant, on a vu que la description analytique la plus pertinente de ce problème ne fait pas appel à la notion de modulation. D'ailleurs, il faut noter qu'elle ne fait pas non plus explicitement référence à la phase du spectre à court-terme.

Le but de ce qui suit est de confirmer ces diverses conclusions par l'analyse théorique du résultat du traitement dans le cas d'un son pur bruité. La démarche suivie dans de ce paragraphe comporte deux étapes successives. Tout d'abord, la caractérisation du signal obtenu en sortie du traitement, qui s'effectue dans le domaine fréquentiel, en utilisant la réponse en fréquence du filtre équivalent à la modification spectrale. Par suite, la comparaison avec les données psychoacoustiques concernant l'effet de masquage simultané d'une bande de bruit par un son pur permet d'évaluer l'audibilité du phénomène constaté.

3.3.1.b Caractérisation fréquentielle du signal après traitement

Dans le cas où le signal à restaurer est constitué d'un son pur bruité situé au dessus de la limite de restauration, on admet ici que le traitement effectué s'apparente à un simple filtrage linéaire par un filtre passe-bande centré autour de la fréquence de la sinusoïde. On néglige donc ici les fluctuations de l'atténuation spectrale due au bruit (cf paragraphe 3.2.2), ainsi que les effets non-linéaires qui apparaissent à cause de l'insuffisance du recouvrement entre les trames à court-terme (cf. paragraphe 3.1.1.a).

Le problème posé est donc de déterminer la réponse du filtre linéaire équivalent à une modification du spectre à court-terme du type de celle qui est représentée sur la figure 3.1. Du fait de la linéarité du traitement, on commence par s'intéresser à la réponse équivalente à un seul canal de TFCT. C'est à dire le cas où seul le canal de la transformée à court-terme d'indice k_0 est inchangé, tous les autres étant mis à zéro. La modification multiplicative du spectre à court-terme s'écrit alors

$$G(p, \omega_k) = \delta(k_0)$$

La réponse temporelle, équivalente à la modification apportée dans une trame à court-terme, est définie au paragraphe 2.3.2 par la relation (2.21). Celle-ci s'écrit dans notre cas

$$g^{\infty}(m) = \frac{1}{N}\sum_{k=0}^{N-1} G(p, \omega_k) W_N^{km} = \frac{1}{N}\mathrm{e}^{\mathrm{j}2\pi\frac{k_0 m}{N}} \tag{3.19}$$

Il faut noter qu'ici la réponse $g^{\infty}(m)$ n'est pas référencée par l'indice de trame p puisque nous avons supposé que la modification spectrale apportée est identique dans chaque trame à court-terme. On rappelle que la notation $g^{\infty}(m)$ indique que la réponse impulsionnelle équivalente considérée est de support infini (cf. paragraphe 2.3.2). La réponse impulsionnelle du filtre équivalent à la modification du spectre à court-terme est alors donnée par l'équation (B.20) sous la forme

$$\tilde{g}(m) = \sum_{p=-\infty}^{+\infty} g^{\infty}(m)\,\{h * f(m)\}$$

C'est à dire ici

$$\tilde{g}(m) = \mathrm{e}^{\mathrm{j}2\pi\frac{k_0 m}{N}} \left\{\frac{f * h(m)}{N}\right\} \tag{3.20}$$

Cette dernière expression indique donc que le filtre équivalent à un canal de TFCT s'obtient simplement en modulant le filtre composite réalisé par les fenêtres d'analyse et de synthèse $(f * h(m))$ par la fréquence centrale du canal considéré.

Un dernier point à propos de ce filtre équivalent concerne les problèmes de normalisation. D'après le paragraphe 2.3, la transparence de l'analyse/synthèse par TFCT, lorsqu'aucune modification spectrale n'est effectuée, impose une contrainte aux fenêtres utilisées. Cette condition de transparence, donnée par l'équation (B.22), s'écrit

$$\sum_{p=-\infty}^{+\infty} f(n - pR)h(pR - n) = 1$$

Dans le cas où le paramètre de décalage R vaut 1, et la fenêtre de synthèse est rectangulaire (technique dite OLA), la relation précédente devient

$$\sum_{n} h(n) = 1$$

En intégrant cette contrainte de normalisation dans la relation (3.20), l'expression du filtre équivalent à un canal de TFCT devient

$$\tilde{g}(m) = \mathrm{e}^{\mathrm{j}2\pi\frac{k_0 m}{N}} \left\{ \frac{f * h(m)}{N \sum_n h(n)} \right\} \tag{3.21}$$

Où $f(n)$, la fenêtre de synthèse, est supposée être une fenêtre rectangulaire définie par $h(n) = 1$, pour $n \in [0, N-1]$. Il faut noter que ce résultat n'est strictement exact que dans le cas où $R = 1$. Cependant, nous avons vu au paragraphe 2.3.2 que lorsque le recouvrement est suffisant (supérieur à 75%), on peut considérer que le modèle de filtrage linéaire (invariant) est encore valide. Par ailleurs, le cas d'un recouvrement plus faible (50%) donne lieu, pour une composante sinusoïdale, aux effets qui ont déjà été étudiés au paragraphe 3.1.1.a.

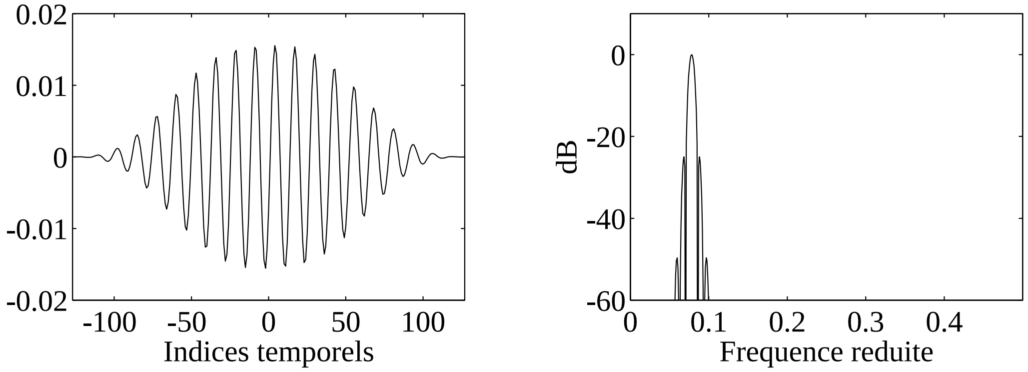

Figure 3.20: Réponse équivalente à un canal de TFCT. **A gauche,** réponse impulsionnelle. **A droite,** réponse fréquentielle. La fenêtre d'analyse utilisée est une fenêtre de Hann de longueur 128 points (méthode OLA pour la synthèse). Le canal de TFCT considéré est le canal d'indice fréquentiel 10, dont la fréquence centrale vaut $7,81.10^{-2}$ (en fréquence réduite).

La figure 3.20 présente un exemple de réponse calculée grâce à la relation (3.21). On note que le support de la réponse impulsionnelle est bien deux fois plus long que la durée de la fenêtre d'analyse ce qui souligne le rôle joué par la fenêtre de synthèse pour le filtrage des modifications apportées sur le spectre à court-terme. Pour la même raison, le filtre passe-bande équivalent à un canal de TFCT est plus sélectif que le filtre d'analyse seul.

Dans le cas d'une atténuation spectrale semblable à celle qui est représentée sur la figure 3.1, le filtrage équivalent est obtenu par simple sommation des réponses impulsionnelles des canaux de TFCT non-atténués (du fait de la linéarité de la TFCT quand $R = 1$). Il faut noter que le nombre de canaux non-atténués, lors du traitement d'un son pur bruité, augmente avec le niveau relatif mesuré au sommet du pic spectral (cf. figure 3.1).

Il est donc possible, grâce à la relation (3.21), de décrire la forme de la densité spectrale du bruit demeurant après traitement. Cependant, pour caractériser complètement le signal obtenu en sortie, il reste nécessaire de déterminer les niveaux respectifs du signal (le son pur) et du bruit. Nous avons vu dans les paragraphes précédents que, dans le cas d'un son pur bruité, la quantité significative pour le traitement d'atténuation spectrale à court-terme est *la valeur moyenne du niveau relatif mesurée au sommet du pic spectral* (notée $\mathrm{E}\{\mathcal{Q}(p,\Omega)\}$). Par conséquent, on choisit ici de déterminer la composition fréquentielle du signal traité en fonction de $\mathrm{E}\{\mathcal{Q}(p,\Omega)\}$. La relation (C.19) (annexe C), fournit l'expression analytique de $\mathrm{E}\{\mathcal{Q}(p,\Omega)\}$ sous la forme

$$\mathrm{E}\{\mathcal{Q}(p,\Omega)\} = 1 + \frac{\mathcal{P}_s N}{2P_d^{(a)}(\frac{\Omega}{2\pi}F_e)\,F_e\Delta_h}$$

Pour le problème qui nous intéresse il n'est pas nécessaire de faire intervenir la densité spectrale de puissance du bruit sous sa forme analogique. L'expression de $\mathrm{E}\{\mathcal{Q}(p,\Omega)\}$ peut donc se réécrire plus simplement comme

$$\mathrm{E}\{\mathcal{Q}(p,\Omega)\} = 1 + \frac{\mathcal{P}_s N}{2P_d(\Omega)\Delta_h}$$

où $P_d(\omega)$ désigne cette fois la densité spectrale de puissance du bruit dans le domaine discret. En utilisant cette relation, la densité spectrale de puissance du bruit de fond, à la fréquence de la sinusoïde, s'écrit

$$P_d(\Omega) = \frac{1}{\mathrm{E}\{\mathcal{Q}(p,\Omega)\}-1} \times \frac{N}{\Delta_h} \times \frac{\mathcal{P}_s}{2} \tag{3.22}$$

On rappelle que le second terme s'interprète comme l'inverse de la largeur de bande (ici exprimée en fréquence réduite) équivalente au filtre d'analyse associé à la TFCT (paragraphe 3.1.1.b). En supposant que la DSP du bruit de fond est quasiment constante au voisinage de la pulsation Ω de la sinusoïde, la **DSP de la bande de bruit qui subsiste en sortie du traitement** s'écrit (formule de filtrage) :

$$\left|\tilde{G}(\omega)\right|^2 \times P_d(\Omega)$$

soit d'après l'équation (3.22)

$$\left|\tilde{G}(\omega)\right|^2 \times \frac{1}{\mathrm{E}\{\mathcal{Q}(p,\Omega)\}-1} \times \frac{N}{\Delta_h} \times \frac{\mathcal{P}_s}{2} \tag{3.23}$$

où $\tilde{G}(\omega)$ désigne la transformée de Fourier de la réponse impulsionnelle $\tilde{g}(m)$ équivalente au filtrage réalisé par le traitement telle qu'elle est définie par l'équation (3.21).

C'est l'équation (3.23) qui est utilisée pour déterminer le spectre de la bande de bruit qui subsiste après traitement. Afin de pouvoir comparer les différents résultats entre eux, le niveau du bruit de fond autour de la fréquence de la sinusoïde est fixé à $P_d(\Omega) = 0$ dB. Les figures 3.21 et 3.22 correspondent donc aux résultats de traitement obtenus pour une composante sinusoïdale de puissance variable noyée dans un bruit de fond dont le niveau est fixé. Nous avons choisit de considérer l'exemple d'une sinusoïde de fréquence 500 Hz (ce choix sera justifié ultérieurement au paragraphe 3.3.1.c). De plus, nous avons supposé que cette fréquence correspond à un point de discrétisation fréquentielle de la transformée de Fourier à court-terme utilisée lors du traitement afin de simplifier la présentation. Pour l'instant, c'est uniquement la partie supérieure des figures 3.21 et 3.22 (qui représente la puissance de la sinusoïde et la DSP de la bande de bruit) qui va être commentée.

La figure 3.21 correspond au cas d'une fenêtre d'analyse de durée 10 ms. Le haut de cette figure présente l'analyse fréquentielle du résultat de traitement pour deux niveaux de la sinusoïde : à gauche (partie **A**), $\mathrm{E}\{\mathcal{Q}(p,\Omega)\} = 5$ dB, et à droite (partie **B**), $\mathrm{E}\{\mathcal{Q}(p,\Omega)\} = 30$ dB. Le trait pointillé représente *la puissance de la sinusoïde* matérialisée à la fréquence de la sinusoïde (500 Hz), tandis que le trait plein correspond à *la densité spectrale du bruit* demeurant après traitement calculé grâce à l'équation (3.23). On vérifie bien qu'entre les parties **A** et **B** de la figure 3.21 le niveau de la sinusoïde s'élève de 25 dB. De plus, on note aussi un élargissement de la bande de bruit subsistant après le traitement lorsque la puissance de la sinusoïde augmente. Nous avons déjà signalé que cet élargissement vient du fait que le filtrage équivalent réalisé par le traitement est d'autant moins sélectif que le niveau relatif de la composante est élevé. Ainsi,

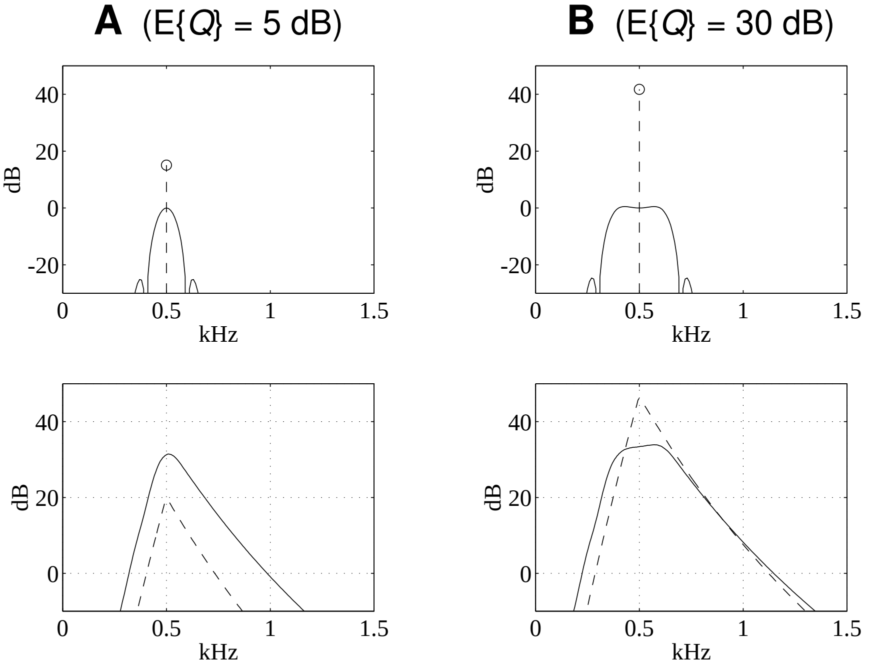

Figure 3.21: En haut, puissance du son pur (pointillés) et densité spectrale de puissance de la bande de bruit subsistant après le débruitage (trait plein). **En bas,** seuil de masquage du son pur (pointillés) et niveau d'excitation de la bande de bruit (trait plein). La fenêtre d'analyse utilisée est une **fenêtre de Hann de durée 10 ms**. Le niveau relatif moyen au sommet du pic spectral pendant le traitement est de **5 dB pour la partie gauche de la figure (A)**, et de **30 dB pour la partie droite (B)**.

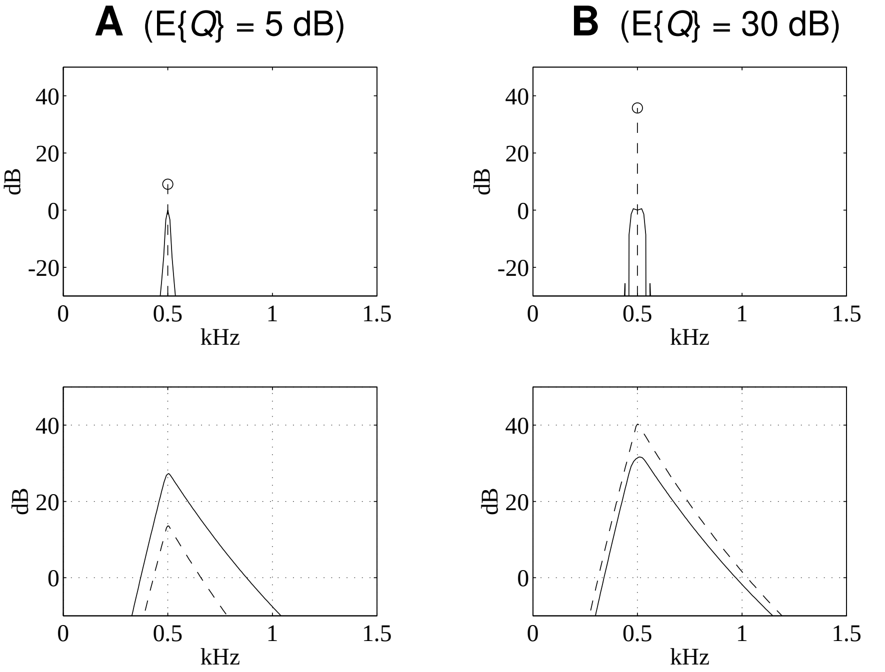

Figure 3.22: En haut, puissance du son pur (pointillés) et densité spectrale de puissance de la bande de bruit subsistant après le débruitage (trait plein). **En bas,** seuil de masquage du son pur (pointillés) et niveau d'excitation de la bande de bruit (trait plein). La fenêtre d'analyse utilisée est une **fenêtre de Hann de durée 40 ms**. Le niveau relatif moyen au sommet du pic spectral pendant le traitement est de **5 dB pour la partie gauche de la figure (A)**, et de **30 dB pour la partie droite (B)**.

dans le cas de la partie **A**, le niveau relatif moyen au sommet du pic spectral étant de 5 dB, c'est à dire proche de la limite de restauration, seul un canal de la TFCT est passant (celui correspondant à la fréquence de la sinusoïde). Par contre, dans le cas de la partie **B**, un niveau relatif moyen de 30 dB correspond à trois canaux de TFCT passants (la fréquence de la sinusoïde ainsi que les deux canaux voisins) [Harris 78].

Les remarques qui viennent d'être exposées restent valables pour la partie supérieure de la figure 3.22 qui correspond à une durée de fenêtre de TFCT de 40 ms. La comparaison entre les figures 3.21 et 3.22 permet de mettre en évidence deux points importants concernant l'influence de la durée de la fenêtre d'analyse :

- Sur la figure 3.22, la bande de bruit subsistant après le traitement est considérablement moins large que sur la figure 3.21. En effet, nous avons vu qu'en augmentant la durée de la fenêtre de TFCT de 10 à 40 ms, le filtrage équivalent effectué par le traitement devient plus sélectif.
- En comparant les parties **A** (ou **B**) des figures 3.21 et 3.22, on note que la puissance de la sinusoïde est plus faible sur la figure 3.22 alors que le niveau relatif est le même dans les deux cas. Cet effet est lié à l'augmentation de la durée de fenêtres : d'après les résultats du paragraphe 3.1.1, une multiplication par 4 de la durée de la fenêtre de TFCT équivaut à diminuer de 6 dB la puissance d'une sinusoïde correspondant à une même valeur de niveau relatif.

3.3.1.c Perception de la distorsion

Les résultats présentés nous permettent d'aborder le problème de l'audibilité de la bande de bruit qui subsiste en sortie du traitement. La question est maintenant de savoir si cette bande de bruit est audible compte tenu de la présence du son pur. Cette situation est plus complexe que que le cas inverse abordé au paragraphe 3.1.1. En effet, il s'agit maintenant de savoir si le son pur est susceptible de masquer toutes les composantes fréquentielles du bruit. C'est à dire que l'effet de masquage concerne toute une bande de fréquence et non plus simplement une fréquence particulière comme dans le cas du paragraphe 3.1.1. Un moyen efficace pour traiter ce type de cas consiste à évaluer les niveaux d'excitation (*excitation pattern* en anglais) engendrés par chacun des deux sons, en fonction de la fréquence. Pour évaluer ces niveaux d'excitation, ainsi que le seuil de masquage associé au signal, nous avons utilisé la méthode présentée dans [Schroeder 79].

Les niveaux d'excitation fournissent un moyen simple de simuler analytiquement les résultats concernant l'effet de masquage simultané de sons stationnaires [Moore 82] [Botte 88] [Zwicker 81]. Ce type de simulation est, depuis quelques années, très souvent utilisé dans le domaine de l'audio, principalement pour le codage [Mahieux 89] [Johnston 88] [Feiten 89], mais aussi, par exemple, pour évaluer la qualité d'un système de reproduction [Paillard 92] [Beerends 92]. Le calcul du niveau d'excitation comporte plusieurs étapes :

1. A partir du spectre de puissance du signal sonore, on détermine la puissance du signal intégrée sur la largeur d'une bande critique. Cette opération se fait en représentant l'information spectrale sur l'échelle fréquentielle du taux de bande critique (en anglais, *critical-band rate scale*) appelée aussi échelle Bark. Cette échelle présente la particularité que la largeur correspondant à une bande critique vaut toujours 1 Bark, quelle que soit la fréquence centrale de la bande critique [Zwicker 81]. La relation qui permet de passer de l'échelle Hertz à l'échelle Bark est linéaire jusqu'à 500 Hz, et logarithmique au dessus de

cette fréquence [Zwicker 80] ([Schroeder 79] et [Zwicker 80] proposent des formules analytiques d'approximation de cette relation Hz-Bark).

2. Le niveau d'excitation se déduit du spectre obtenu à l'étape précédente par lissage. Notons qu'il est possible de réaliser simplement ce lissage à l'aide d'un filtre linéaire très simple (récursif du premier ordre) *lorsque l'échelle Bark est utilisée* [Paillard 92].
3. Enfin, l'application d'un facteur de correction (*sensitivity function* dans [Schroeder 79]) sur le niveau d'excitation *correspondant au signal* permet d'obtenir le seuil de masquage. Le bruit est entièrement masqué par le signal dès que le niveau d'excitation du bruit est situé en dessous du seuil de masquage associé au signal pour toutes les fréquences. Le facteur de correction proposé dans [Schroeder 79] dépend en principe de la fréquence, toutefois, il peut être raisonnablement approximé par un offset constant de -24 dB [Mahieux 89].

Le passage par l'échelle Bark n'est pas indispensable, cependant il simplifie nettement le calcul car il permet d'obtenir une forme de l'effet de masque qui est indépendante de la fréquence. Par contre, la pente supérieure de l'effet de masque augmente légèrement avec le niveau sonore absolu du son étudiée. C'est à dire que le niveau d'excitation obtenu n'est théoriquement valide que pour un niveau fixé du signal lors de l'écoute [Zwicker 81] [Feiten 89]. Par ailleurs, il est aussi souhaitable de tenir compte du filtrage réalisé par le conduit auditif externe [Schroeder 79] [Paillard 92] [Mahieux 89]. Ici, ces deux aspects n'ont pas été pris en compte car les conditions d'écoute (et en particulier, le niveau absolu) ne sont pas fixées.

Cette procédure a été utilisée pour calculer les niveaux d'excitation représentés sur le bas des figures 3.21 et 3.22. Il faut noter que contrairement à la tradition, nous n'avons pas représenté les niveaux d'excitation sur l'échelle Bark. C'est à dire qu'après application de la procédure décrite ci-dessus, nous sommes repassé dans le domaine des fréquences linéaires exprimées en Hertz. Cette représentation a été adoptée afin de rester homogène avec les spectres représentés sur le haut des figures 3.21 et 3.22. On note toutefois que même représenté sur une échelle des fréquences en Hertz, le seuil de masquage associé à la sinusoïde présente l'aspect habituel d'un triangle [Zwicker 81] [Mahieux 89]. Ceci est dû au fait que la relation Hertz-Bark est quasiment linéaire en dessous de 1 kHz [Zwicker 80]. Il faut d'ailleurs souligner que le choix d'une fréquence de 500 Hz pour la sinusoïde nous situe dans la zone où la largeur de bande critique est la plus faible. C'est à dire, dans la gamme de fréquence où l'effet de masquage dû à la sinusoïde est le moins important. Le cas considéré est donc "le pire" dans le sens où le masquage de la bande de bruit présente en sortie du traitement par la sinusoïde est moins probable. La comparaison des figures 3.21 et 3.22 nous permet de préciser l'influence des différents paramètres :

Niveau relatif de la sinusoïde Quelle que soit la durée de la fenêtre de TFCT utilisée, la bande de bruit subsistant après le traitement n'est pas masquée par la sinusoïde lorsque le niveau relatif de cette dernière est trop faible (voir les parties **A** des figures 3.21 et 3.22).

Durée de la fenêtre de TFCT La bande de bruit est masquée (pour une sinusoïde de niveau important) uniquement lorsque la durée de la fenêtre de TFCT est suffisamment longue (comparer les parties **B** des figures 3.21 et 3.22).

Pour éviter que la distorsion due à la présence de la bande de bruit en sortie du traitement ne soit audible, il est donc nécessaire d'utiliser des fenêtres de TFCT suffisamment longues. Cependant, même avec une fenêtre longue, la présence de la bande de bruit peut être perceptible dans le cas de composantes sinusoïdales proches de la limite de restauration. En fait, la forme du niveau d'excitation associé à la bande de bruit nous permet de préciser la manière dont celle-ci est perçue lorsqu'elle est audible :

Quand la durée de la fenêtre de TFCT est inférieure à 30 ms, la présence de la bande de bruit n'est pas masquée, même pour des sinusoïdes de niveau important. Pour l'exemple de la figure 3.21, on vérifie que le masquage de la bande de bruit n'est assuré que lorsque le niveau relatif moyen de la composante est supérieur à 50 dB. De plus, le niveau d'excitation correspondant à la bande de bruit est beaucoup plus "large" que la bande critique centrée sur la fréquence de la sinusoïde (environ 250 Hz pour la partie centrale du niveau d'excitation de la partie **B** de la figure 3.21, à comparer avec 100 Hz pour la largeur de la bande critique centrée en $F = 500$ Hz). En sortie du traitement, le signal sonore perçu est donc celui correspondant à une bande de bruit additionnée à la sinusoïde.

Quand la durée de la fenêtre de TFCT est supérieure à 30 ms, on vérifie que la présence de la bande de bruit n'est plus masquée dès que le niveau relatif de la sinusoïde est inférieur à 25 dB. Cependant, le niveau d'excitation produit par la bande de bruit a la même forme que le seuil de masquage associé à la sinusoïde. Ce qui traduit le fait que la bande de bruit est dans ce cas beaucoup moins large que la bande critique centrée sur la fréquence de la sinusoïde. Dans ces conditions, la présence de la bande de bruit (lorsqu'elle est audible) est perçue plutôt sous la forme d'une variation du niveau sonore perçu [Zwicker 81]. Par conséquent l'effet sonore produit est très différent du cas précédent.

En pratique, on constate que dans le premier cas (durée de la fenêtre de TFCT inférieure à 30 ms), la présence de la bande bruit est en général distinctement audible. De plus, l'effet auditif produit est très dérangeant. D'abord parce qu'une telle bande de bruit de faible largeur est un son peu naturel qui ne correspond à aucun son instrumental classique (en excluant les sons électroniques). De plus, cette bande de bruit est liée à la présence d'une composante de signal : elle est plus ou moins audible selon son niveau, elle se déplace lorsque sa fréquence varie. La présence de cette bande de bruit est donc plutôt perçue comme une distorsion du signal musical, et absolument pas comme une perturbation indépendante du signal. Lorsqu'elle est audible, cette distorsion est totalement inacceptable. En ce sens que, comme le bruit évoqué au paragraphe 3.2.1, elle est suffisamment peu naturelle pour être attribuée sans hésitation à un défaut du traitement. Ce qui limite clairement le champ d'application de la technique de débruitage.

Par contre, dans le cas inverse (durée de la fenêtre de TFCT supérieure à 30 ms), il est très difficile d'isoler l'effet de variation du niveau sonore perçu évoqué précédemment. D'une part, il se confond avec le phénomène de variation de l'amplitude de la composante de signal étudié au paragraphe 3.2.2. D'autre part, cet effet semble peu perceptible dès que l'on se trouve dans une situation réelle (sons musicaux complexes qui varient dans le temps, présence de bruit résiduel large bande de niveau non négligeable[5], etc.). Enfin, il faut souligner que dans le cas de durées de fenêtre de TFCT longues, la variation du niveau perçu de la composante sinusoïdale n'est audible que pour des composantes de faible niveau (niveau relatif en dessous de 25 dB). D'autant plus que la valeur du niveau relatif correspond à une puissance réelle de la composante de signal d'autant plus faible que la durée de la fenêtre de TFCT est importante.

Ces conclusions concernant la durée de la fenêtre de TFCT permettent de justifier les constatations rapportées dans [Moorer 86] et dont on a déjà parlé au paragraphe 2.1.3, à savoir, que les seuls systèmes intéressants sont ceux qui comportent soit très peu de bandes (4 par exemple) soit un grand nombre de bandes. Les résultats que nous avons obtenu montrent d'ailleurs que l'expression "un grand nombre de bande" doit être remplacée par "une résolution fréquentielle suffisante". Quant au premier cas mentionné, celui d'un nombre très faible de bandes, il faut

[5]Voir le paragraphe 4.1

souligner que dans le cas où chaque bande de la transformée à une largeur de plusieurs kHz, la perception auditive de la bande de bruit demeurant après traitement ne se fait plus sous la même forme. On aboutit alors à un effet qui est plus proche de celui du bruit résiduel. Ce qui souligne que dans ce cas il devient alors nécessaire d'avoir des bandes très larges pour obtenir un effet aussi naturel que possible [Moorer 86]. Cependant, nous avons eu l'occasion de voir au paragraphe 3.1.1 que l'utilisation d'une transformation très peu sélective en fréquence se traduit par une forte une distorsion du signal lorsque de bruit est important. Quand le niveau de bruit est élevé, il n'y a donc pas de choix possible : il faut forcément utiliser une transformation sélective (grand nombre de bandes).

Nous avons déjà noté que la distorsion étudiée ici est principalement audible pour des sinusoïdes de fréquence inférieure à 1 kHz. Ceci traduit le fait que dans cette plage de fréquence, la largeur de la bande critique associée à la sinusoïde est faible. La condition qui garantit le masquage auditif de la distorsion dépend donc de la fréquence de la composante de signal considérée. En particulier, nous avons vu que la bande de bruit est peu audible lorsque sa largeur est inférieure à la largeur de la bande critique centrée sur la fréquence de la sinusoïde (cas de la figure 3.22). Ce qui est intéressant c'est que cette dernière condition est analogue à la condition (3.6) qui garantit le respect du timbre dans le cas d'un son stationnaire (paragraphe 3.1.1). Une comparaison plus précise de ces deux relations indique que la condition qui nous intéresse dans ce paragraphe (qui garantit une distorsion inaudible en sortie) impose une largeur de bande pour la transformée à court-terme environ 2 fois moins importante que la condition (3.6). C'est à dire qu'il n'existe pas d'effets audibles dus à la phase à court-terme, pour un son pur bruité, dès que la largeur de bande de la transformée est au moins égale à la moitié de celle du système envisagé au paragraphe 3.1.1.e.

3.3.2 Sons transitoires

La présence de perturbations sur le spectre de phase à court-terme est aussi la cause de distorsions lors des parties transitoires du signal traité. Il s'agit d'un effet bien connu dans le cadre des systèmes de modification de la durée de la parole qui utilisent la transformée de Fourier à court-terme (voir par exemple [Griffin 84]). Récemment, cet effet a aussi été mis en évidence dans le domaine du codage par transformée des signaux audio [Brandenburg 88] [Mahieux 89]. La présence de perturbations de faible amplitude sur le spectre de phase se traduit par l'apparition d'un phénomène de traînage perceptible dans les parties transitoires, dont l'effet est assez similaire à celui d'une réverbération qui serait non causale. La durée de cet effet est limitée au plus à la longueur d'une fenêtre de la transformée à court-terme.

Dans le cas du débruitage, cet effet existe mais il est d'importance secondaire par rapport au lissage des transitoires évoqué au paragraphe 3.1.2. En effet, les points du spectre à court-terme dont la phase est fortement perturbée du fait de la présence du bruit de fond sont ceux pour lequel le niveau relatif local est proche de 1. Pour s'en convaincre il suffit de constater sur le diagramme de Fresnel de la figure 3.16, que plus le module du spectre à court-terme du signal est important plus la déviation de phase due au bruit est faible. Or ces valeurs du spectre à court-terme proche du niveau de bruit sont fortement atténuées lors de l'application de l'atténuation spectrale. Par conséquent, l'influence des distorsions présentes sur le spectre de phase à court-terme est fortement limitée par le fait que les valeurs du spectre corrompues par le bruit sont de toute façon mises à zéro lors du traitement.

Pour illustrer ce propos, on a réalisé une simulation qui consiste à utiliser le spectre de phase à court-terme du signal non-bruité lors de la synthèse du signal restauré : l'application de

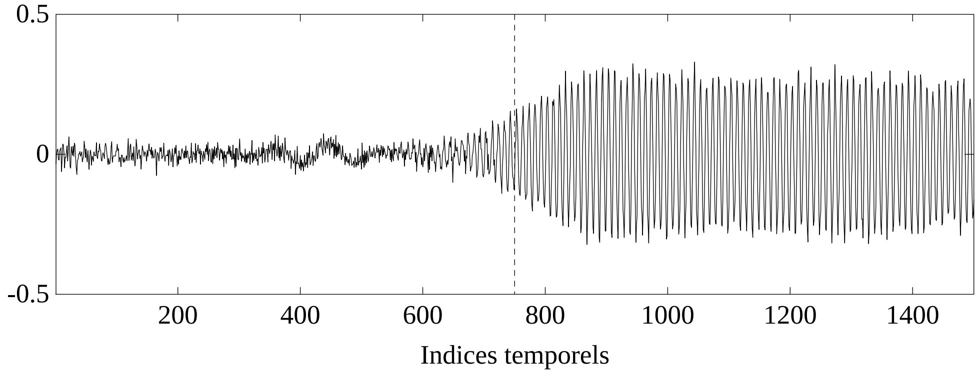

Figure 3.23: Résultat du débruitage dans le cas de l'apparition brutale d'un son pur situé 16 dB au dessus de la limite de restauration. Le trait pointillé vertical indique l'emplacement du transitoire initial. La fenêtre de TFCT utilisée est un fenêtre de Hann de longueur 512 points. La règle de suppression utilisée est la soustraction en puissance avec un facteur de surestimation du niveau de bruit de 3 dB.

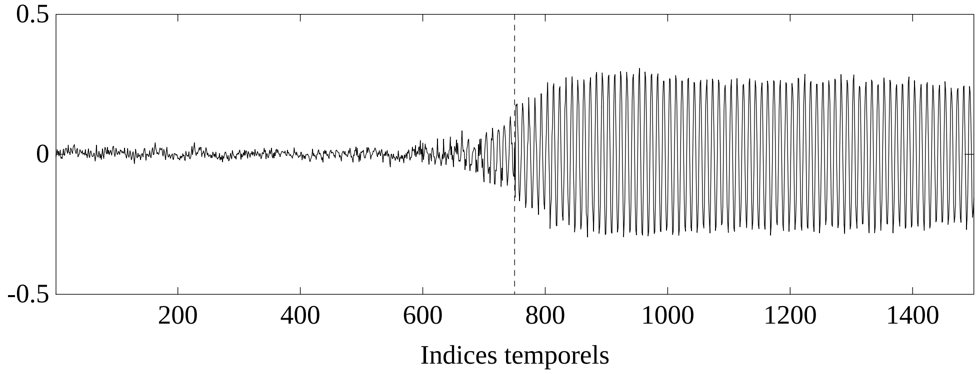

Figure 3.24: Simulation du résultat du traitement incluant l'atténuation spectrale ainsi que l'estimation de la phase à court-terme. Le signal traité est constitué d'un son pur situé 16 dB au dessus de la limite de restauration qui apparaît brusquement. Le trait pointillé vertical indique l'emplacement du transitoire initial. La fenêtre de TFCT utilisée est un fenêtre de Hann de longueur 512 points. La règle de suppression utilisée est la soustraction en puissance avec un facteur de surestimation du niveau de bruit de 3 dB.

l'atténuation spectrale est effectué de manière usuelle, par contre, au moment de la synthèse du signal restauré par TFCT inverse c'est la phase à court-terme du signal non-bruité qui est utilisée. On parle ici de simulation puisque cette substitution n'est possible que si on connaît le signal non-bruité. Le signal traité est constitué, comme au paragraphe 3.1.2, d'un transitoire "abrupt et sinusoïdal" bruité analogue à celui qui est représenté sur la figure 3.3. Le résultat de la simulation est présenté sur la figure 3.24, tandis que le résultat du traitement standard (avec le spectre de phase bruité) correspondant est présenté par la figure 3.23. La comparaison de ces deux figures fait apparaître une légère amélioration dans le cas où la phase non-bruitée est utilisée (figure 3.24) dans la mesure où l'étalement du transitoire du signal est réduit. Cependant, la diminution du temps de montée apportée par l'utilisation du spectre de phase non-bruité est faible par rapport au temps de montée lui-même qui est dû à l'atténuation spectrale (voir le paragraphe 3.1.2). Notons de plus que le cas de la figure 3.24 correspond à l'utilisation d'une règle de suppression avec un facteur de surestimation du bruit modéré de 3 dB. Une surestimation plus importante du niveau de bruit aurait pour conséquence d'atténuer de manière plus forte les faibles valeurs du spectre à court-terme dont la phase est perturbée par le bruit. Les conséquences du bruit sur la phase à court-terme deviennent donc négligeables lorsque le niveau de bruit est fortement surestimé. Pour l'exemple de signal considéré, nous avons vérifié que les figures analogues aux figures 3.23 et 3.24 deviennent complètement indiscernables visuellement dès que le facteur de surestimation est supérieur à 6 dB.

En conclusion, la présence de perturbations dues au du bruit sur le spectre de phase à court-terme se traduit par un étalement lors des transitoires brusques qui est faible par rapport au lissage dû à l'application de l'atténuation spectrale. En particulier, l'influence du bruit sur la phase à court-terme des signaux transitoires devient négligeable lorsque le niveau de bruit est fortement surestimé. En conséquence, les techniques visant à corriger la phase du spectre à court-terme telles que celle qui est décrite dans [Griffin 84] présentent moins d'intérêt dans le cadre du débruitage, puisqu'elles ne permettent pas de résoudre le problème de l'étalement des transitoires brusques.

3.4 Récapitulation

Ce chapitre consacré à l'analyse des résultats du débruitage par atténuation spectrale à court-terme a permis de mettre en évidence plusieurs phénomènes distincts. Les principaux résultats obtenus concernent la caractérisation des modifications apportées au signal musical lors du traitement. La présentation adoptée ici a consisté à analyser séparément les conséquences de chacun des "défauts" du traitement : modification spectrale du signal musical (§ 3.1), caractère aléatoire de l'atténuation spectrale (§ 3.2), et présence de bruit sur la phase à court-terme (§ 3.3). Nous avons déjà signalé en début de ce chapitre que dans un cas réel, tous ces phénomènes se produisent simultanément.

Afin de fournir un éclairage différent sur l'étude menée dans ce chapitre, nous allons résumer les principaux résultats obtenus en utilisant deux points de vues différents : tout d'abord, celui de l'utilisateur d'un système de restauration par atténuation spectrale à court-terme, et ensuite, celui du concepteur d'un système de débruitage d'enregistrements musicaux.

3.4.1 Résumé (I)

Pour l'**utilisateur d'un système de restauration**, il n'est pas question de remettre en cause le principe de la technique utilisée, par contre, il est important de savoir quelle est l'**influence des différents paramètres à disposition**. Pour un tel utilisateur, les résultat principaux concernent la **durée des trames à court-terme** dont l'influence dépend de la portion du signal considérée :

Pour les parties quasi-stationnaires, la durée des trames à court-terme doit être suffisante (supérieure à 30 ms) pour éviter que le phénomène de modulation ne soit audible (§ 3.3.1). Cette condition est impérative car, sauf si l'enregistrement est peu bruité, le phénomène de modulation est très gênant (§ 3.3.1.c). Une deuxième limite importante est la durée de trame de l'ordre de 60 ms qui garantit que le traitement ne peut pas éliminer de composante du signal audibles *compte tenu de la présence du bruit de fond* (§ 3.1.1). Le choix d'une durée de trame inférieure à cette limite de 60 ms se traduit par une modification du timbre du signal musical (§ 3.1.1.c) qui, en général, contribue à rendre le signal restauré moins brillant (§ 3.1.1.d). En ce qui concerne les parties quasi-stationnaires du signal musical, l'influence de la durée de trame à court-terme peut donc être résumée sous la forme du tableau suivant :

Durée de trame à court-terme		
Moins de 30 ms	**De 30 ms à 60 ms**	**Plus de 60 ms**
Phénomène de modulation	Disparition de composantes audibles	-

Pour les transitoires, il faut garder à l'esprit qu'en cas d'apparition brusque de composantes de signal de faible niveau (par rapport à celui du bruit), le traitement provoque un lissage du signal sur une durée quasiment proportionnelle à la durée de trame à court-terme (§ 3.1.2). Par conséquent, si l'enregistrement à restaurer présente des transitoires marqués, l'augmentation de la durée de trame à court-terme se traduit par un lissage accru des transitoires.

Il est donc normal que la durée de trame à court-terme utilisée pour la restauration dépende de l'enregistrement traité : en l'absence de transitoire très violent (voix chantée, instruments à corde frottée), une durée de l'ordre de 60 à 80 ms est réaliste, par contre, si l'enregistrement contient des transitoires brefs (guitare, piano, percussions), une augmentation de la durée au delà de 40 ms risque de se traduire par une distorsion de plus en plus audible au moment des transitoires. Attention, dans ce dernier cas, il ne faut pas en conclure que les résultats obtenus avec des fenêtres courtes sont satisfaisants : nous avons vu que le phénomène de distorsion des transitoires se produit aussi avec des fenêtres courtes (§ 3.1.2). Il est clair que le cas d'un enregistrement très fortement bruité qui comporte des transitoires brusques est un des cas les plus difficiles que l'on puisse rencontrer.

Un second point important pour l'utilisateur est que le **facteur de surestimation du bruit** augmente simultanément les deux distorsions dont nous venons de parler, en augmentant le niveau relatif de coupure à partir duquel le signal est mis à zéro. Pour les composantes stationnaires du signal, une augmentation du niveau relatif de coupure de 3 dB doit être compensée par une multiplication par 2 de la taille de la fenêtre (§ 3.1.1.b). Ceci indique que la surestimation du niveau de bruit, même modérée (de l'ordre de 6 dB), provoque très rapidement une distorsion du signal qui ne peut pas être compensée par le choix d'autres paramètres. Sauf dans le cas d'un enregistrement très faiblement bruité, il est donc impératif de limiter autant que possible la surestimation du bruit de fond.

3.4.2 Résumé (II)

Pour le **concepteur d'un système de restauration**, la situation est un peu différente : il ne s'agit plus seulement de régler les paramètres, mais aussi de **mettre en évidence les limitations du système** afin de pouvoir l'améliorer.

- Pour les transitoires, la cause principale de distorsion est le fait que le filtrage équivalent au traitement est très sélectif pour les composantes de faible niveau relatif. Ce qui se traduit par un fort étalement temporel lors de l'apparition brusque de composantes de signal fortement bruitées (§ 3.1.2 et § 3.3.2).
- Par contre, pour les parties stationnaires, les distorsions apparaissent lorsque la sélectivité fréquentielle est insuffisante. Notons que selon le type de distorsion considérée, il faut prendre en compte, soit la résolution fréquentielle de la partie analyse seule (pour l'élimination

de composantes de signal évoquée au paragraphe 3.1.1.b), soit la sélectivité du filtrage équivalent au traitement complet (pour l'effet de modulation décrit au paragraphe 3.3.1.b).

Sur ce problème de **sélectivité fréquentielle**, les résultats obtenus soulignent l'intérêt potentiel d'une transformation à court-terme légèrement moins sélective pour les fréquences élevées (au dessus de 2 kHz) que pour les fréquences basses du fait des propriétés auditives (§ 3.1.1.e et § 3.3.1.c). Cette conclusion resterait toutefois à confirmer dans le cas de signaux musicaux fortement bruités (§ 3.1.1.e).

Une autre cause de distorsion du signal assez inquiétante est le fait que les composantes de bas niveau voient leur **amplitude fluctuer aléatoirement**. Nous avons vu que les variations d'amplitude dues à la variance de l'estimation spectrale locale peuvent être sensibles jusqu'à des niveaux relatifs d'une vingtaine de décibels (§ 3.2.2). Par ailleurs, nous avons eu l'occasion de souligner qu'*auditivement* ce phénomène de fluctuation peut aussi être dû à la présence, dans le signal restauré, d'une bande de bruit autour des composantes du signal. Il semble que ce dernier phénomène soit perceptible jusqu'à des niveaux relatifs de 25 dB (§ 3.3.1.c). Enfin, nous avons aussi montré qu'avec un recouvrement de 50%, il existe un effet de modulation déterministe du signal lié à la modification de la TFCT qui est à la limite d'être audible. En particulier, si on désire une très forte réduction du niveau de bruit de fond, il est nécessaire d'utiliser un recouvrement plus important pour réduire cette modulation (§ 3.1.1.a). Un point important est que ces phénomènes de fluctuation/modulation concernent les composantes de signal proches du niveau de bruit. Ils deviennent donc moins importants lorsque la taille de la trame à court-terme augmente car ils se se produisent pour des composantes de niveau absolu plus faible (voir l'interprétation de l'équation (3.2)).

Enfin pour terminer par le célèbre phénomène de **bruit musical**, nous avons vu que la surestimation du niveau de bruit ne peut pas constituer une solution réaliste à ce problème. Un point de détail qui peut tout de même être intéressant en pratique est le fait que la composition fréquentielle du bruit musical varie avec la taille de trame à court-terme utilisée (§ 3.2.1.b).

3.4.3 Remarques sur la validité des résultats

Pour conclure ce chapitre, il est intéressant de rappeler quelques remarques concernant le degré de précision des différents résultats obtenus. Pour les effets qui se manifestent durant les parties stationnaires du signal, les résultats présentés sont assez complets puisqu'il a quasiment toujours été possible d'aller jusqu'à évaluer l'audibilité des phénomènes (§ 3.1.1 et 3.3.1). Il faut cependant noter que les résultats fournis concernent des composantes sinusoïdales isolées, la généralisation à un son possédant un spectre complexe doit donc se faire avec précaution (§ 3.1.1.d). Un point qui demeure plus flou est la limite à partir de laquelle les phénomènes de fluctuation aléatoire du niveau des composantes de signal (§ 3.2.2 et § 3.3.1.c) sont gênants. Ce que nous avons constaté, c'est qu'il est assez difficile de déterminer à l'écoute si le niveau de la composante restaurée fluctue ou non à cause de la présence du bruit résiduel. Pour préciser la limite d'audibilité de ces phénomènes, il serait sûrement utile d'effectuer un véritable test psychoacoustique.

La partie, la moins aboutie de ce chapitre concerne le cas des distorsions qui affectent les parties transitoires du signal musical (§ 3.1.2). En effet, il ne nous a été possible que de donner une description analytique de la distorsion, pour un modèle de transitoire précis. En particulier, l'audibilité de l'effet de lissage mis en évidence n'a pas été abordée. Le problème est que le transitoire type utilisé (démarrage abrupt d'une sinusoïde) ne correspond pas vraiment à une modélisation des transitoires réels de signaux musicaux. On peut donc penser que ce signal nous permet de décrire analytiquement l'effet du traitement, par contre, il n'est pas certain que le même signal soit pertinent pour une évaluation perceptive.

Chapitre 4

Solutions adaptées pour les enregistrements musicaux

Ce dernier chapitre est consacré à l'étude de solutions à certains des problèmes mis en évidence au cours du chapitre précédent. La première partie (paragraphe 4.1) concerne le phénomène de bruit musical. Ici l'étude est assez rapide puisque nous disposons déjà d'une technique qui permet de remédier à ce problème : la règle de suppression proposée par Ephraim et Malah (décrite au paragraphe 2.2.3). Notre contribution s'est donc essentiellement limitée à vérifier sous quelles conditions l'algorithme d'Ephraim et Mallah permet de se débarrasser du phénomène de bruit musical. Cette partie de notre travail ayant fait l'objet d'un article détaillé, le texte de celui-ci est joint en annexe (annexe D).

La seconde partie de ce chapitre (paragraphe 4.2) est consacrée à la description d'un système original de réduction de bruit de fond, basé sur les résultats du chapitre 3. Le but recherché est de mettre au point un système de restauration, adapté aux signaux musicaux, qui permette de venir à bout du compromis entre distorsion du signal pendant les parties stationnaires et lissage des transitoires. Le paragraphe 4.2.1 correspond à un premier essai infructueux de traitement qui est décrit brièvement en guise d'introduction.

4.1 Contrôle du bruit résiduel

4.1.1 Règles de suppression lissées, ou moyennées (linéairement)

D'après les résultats du paragraphe 3.2.1, le problème posé par le bruit résiduel est essentiellement lié à l'estimateur spectral utilisé, c'est à dire le périodogramme. Ce sont directement les caractéristiques du périodogramme (forte variance, indépendance entre points voisins) qui sont à l'origine du phénomène de bruit musical. Sachant cela, on peut penser que la solution pour éviter le phénomène de bruit musical consiste à utiliser un estimateur spectral différent, de variance plus faible.

Le choix le plus naturel consiste simplement à utiliser le périodogramme lissé ou moyenné pour l'estimation spectrale locale du signal ($\hat{P}_x(p, \omega_k)$). Ces deux types d'estimateurs correspondent à des solutions classiques pour réduire la variance du périodogramme [Kay 88] [Brillinger 81]. On parle ici de règle de suppression moyennée lorsque l'estimation spectrale locale est obtenue en moyennant les périodogrammes correspondant à *plusieurs trames à court-terme voisines de la trame courante*. Le terme de règle de suppression lissée désigne, au contraire, le cas où le périodogramme obtenu dans la trame courante est lissé, c'est à dire qu'il s'agit alors de cumuler les valeurs obtenues *pour plusieurs indices fréquentiels ω_k successifs*. Il faut souligner que dans un tel traitement c'est uniquement l'étape d'estimation du niveau relatif local qui doit être modifiée. C'est dans ce type de cas que la distinction effectuée au paragraphe 2.2.1 entre, d'une part, le niveau relatif $\mathcal{Q}(p, \omega_k)$ utilisé pour calculer l'atténuation, et d'autre part, la TFCT du signal $X(p, \omega_k)$ qui est modifiée, prend toute son importance. Dans chaque trame à court-terme le déroulement des opérations est donc le suivant

1. Calcul du niveau relatif local, défini par la relation

$$\mathcal{Q}(p, \omega_k) = \frac{\hat{P}_x(p, \omega_k)}{\hat{P}_d(\omega_k)}$$

   où l'estimateur spectral local $\hat{P}_x(p, \omega_k)$ est un estimateur moyenné ou lissé.

2. Utilisation d'une règle de suppression ponctuelle pour obtenir l'atténuation spectrale à partir de la donnée de $\mathcal{Q}(p, \omega_k)$.

3. Application de l'atténuation spectrale au spectre à court-terme du signal $X(p, \omega_k)$.

On trouve des exemples de ce type de démarche dans [Canagarajah 91] où le lissage appliqué sur le périodogramme est inspiré de considérations sur l'audition humaine (les canaux de TFCT sont groupés par bandes-critiques), ainsi que dans [Boll 79] qui propose de moyenner les estimations spectrales obtenues dans des trames à court-terme successives.

Malheureusement, ces solutions, qui ont le mérite d'être simples, ne fournissent pas des résultats très satisfaisants en pratique. La raison principale de cet échec est que pour réduire significativement la variance du périodogramme, le lissage (ou la moyenne) effectué doit porter sur un nombre important de points distincts (cf. tableau 1.3). En dessous de 5 à 6 points cumulés, l'amélioration obtenue est très faible, d'autant plus que les valeurs voisines de la TFCT ne sont pas indépendantes. Le problème est que le cumul de 5 à 6 valeurs distinctes, que ce soit en fréquence ou temporellement, génère des distorsions du signal.

Pour le cas des règles de suppression lissées, on sait que le lissage se traduit par une perte de résolution spectrale de l'analyse fréquentielle. Par conséquent, le lissage augmente les distorsions apportées aux parties stationnaires du signal (paragraphes 3.1.1 et 3.3.1). Si on admet qu'un lissage sur 5 points successifs du spectre à court-terme se traduit par une résolution fréquentielle environ 5 fois plus faible, la condition du paragraphe 3.3.1 ne peut être vérifiée que si la durée de la trame de TFCT est de l'ordre de 150 ms. Ceci indique clairement que le lissage de l'estimation spectrale est impraticable dans des cas réels. Le traitement proposé dans [Canagarajah 91] constitue une exception car le lissage du spectre de puissance à court-terme n'est pas uniforme : il dépend de la fréquence. D'après les paragraphes 3.1.1.c et 3.3.1.c, si le lissage du spectre de puissance simule exactement l'élargissement des bandes critiques avec la fréquence, il ne doit pas y avoir de distorsion supplémentaire du signal. Cependant, nous avons vu au paragraphe 3.1.1.c que la largeur des bandes critiques en dessous de 500 Hz est assez faible (de l'ordre de 100 Hz).

Avec des valeurs de durée de trame usuelles (40 ms), ceci revient à dire que le lissage en dessous de 500 Hz est inexistant. Les essais que nous avons menés en utilisant pour le lissage de l'estimation spectrale la procédure utilisée pour le calcul des niveaux d'excitation (paragraphe 3.1.1.c) ont confirmés exactement ce résultat : au delà d'une certaine fréquence le lissage effectué est suffisant pour fournir un bruit résiduel relativement naturel, par contre en dessous de 2 kHz on obtient un phénomène de bruit musical.

Pour le cas des règles de suppression moyennées, les défauts qui apparaissent sont légèrement différents. Pour illustrer ce point, on suppose, par exemple, que le niveau relatif local est obtenu en moyennant les niveaux relatifs mesurés dans n trames successives, et que l'atténuation calculée est appliquée dans la première trame. On s'intéresse au cas d'un signal qui apparaît brusquement dans le bruit : dès que le signal se présente dans la $n$ième trame à court-terme, le niveau relatif calculé par moyenne sur les n trames augmente très fortement (il faut se souvenir qu'on représente les niveaux de signal en décibels et non sur une échelle linéaire). Par conséquent, l'atténuation spectrale apportée dans la première trame est très faible alors que celle-ci ne contient encore que du bruit. L'effet produit est celui d'une bouffée de bruit de fond qui précède l'apparition du signal. Cet effet (entre autres) est particulièrement audible même lorsque le nombre de trames prises en compte est limité (deux à trois trames suffisent). Ceci indique que les règles de suppression moyennées ne peuvent absolument pas fournir une solution au problème du bruit musical.

4.1.2 Règle de suppression d'Ephraim et Malah

Nous avons dit au paragraphe 2.2 que l'algorithme d'Ephraim et Malah [Ephraim 84] correspond en quelque sorte à une règle de suppression moyennée. Etant donné les résultats du paragraphe précédent, on peut se demander comment une telle règle de suppression peut permettre d'obtenir une élimination du bruit musical sans créer des distorsions audibles du signal.

4.1.2.a Elimination du bruit musical

L'annexe D fournit des justifications précises concernant les simulations effectuées pour évaluer le fonctionnement de la règle de suppression proposée par Ephraim et Malah. Nous nous contentons ici simplement de rappeler deux conclusions importantes :

- L'élimination du bruit musical est surtout liée au fait que le RSB a priori $\mathcal{R}_{prio}(p, \omega_k)$ correspond à un "lissage" des RSB mesurés $\mathcal{R}_{post}(p, \omega_k)$ dans les trames successives. La particularité de ce "lissage" est qu'il est non-linéaire : il correspond effectivement à un lissage important pour les faibles valeurs de RSB mesurés (inférieures à 0 dB), par contre, il devient équivalent à un simple retard d'une trame à court-terme lorsque $\mathcal{R}_{post}(p, \omega_k)$ est important. C'est cette dernière propriété qui garantit que l'utilisation de l'algorithme d'Ephraim et Malah ne génère pas de distorsion temporelle importante du signal traité.

- En pratique, on peut considérer que l'atténuation apportée par la règle de suppression d'Ephraim et Malah correspond à la soustraction en puissance, évaluée en fonction du RSB a priori. Cependant, la formule de calcul de l'atténuation à deux entrées ($\mathcal{R}_{post}$ et $\mathcal{R}_{prio}$), proposée par Ephraim et Malah, ne peut pas être simplement remplacée par la soustraction en puissance, car c'est l'influence corrective du paramètre $\mathcal{R}_{post}$ qui garantit le lissage de $\mathcal{R}_{prio}$.

Par rapport aux autres solutions préconisées pour éliminer le phénomène de bruit musical que nous avons pu trouver dans la littérature, les avantages de l'algorithme d'Ephraim et Malah sont principalement qu'il garantit une élimination complète du bruit musical (à la différence de la solution décrite par [Vaseghi 92]), et qu'il évite la réapparition locale du phénomène si le bruit dépasse légèrement son niveau mesuré (ce qui n'est pas le cas avec la procédure à seuil proposée dans [Boll 79, II.G]).

Toutefois, en utilisant le même modèle qu'au paragraphe 3.1.2 (apparition brusque d'une composante de signal), on montre dans l'annexe D que l'algorithme d'Ephraim et Malah peut provoquer une sur-atténuation au moment du transitoire dans le cas de l'apparition d'une composante de faible niveau. Avec des valeurs usuelles du paramètre α (moins de 0,98), cet effet ne concerne que les composantes de faible niveau (moins de 15 dB au dessus du niveau de bruit). Il s'agit tout de même d'une distorsion supplémentaire qui vient amplifier l'effet de lissage des transitoires décrit au paragraphe 3.1.2. Il faut cependant souligner que même avec cette restriction, les résultats obtenus avec la règle de suppression d'Epraim et Malah présentent une distorsion du signal beaucoup moins importante que ceux obtenus, par exemple, en surestimant le niveau de bruit de 6 dB. Par rapport aux autres techniques mentionnées dans ce document, l'algorithme d'Ephraim et Malah permet d'éliminer efficacement le phénomène de bruit musical dans les situations réelles sans que cela ne se traduise par une distorsion trop marquée du signal. L'annexe D indique d'ailleurs un point qui mériterait d'être fouillé concernant la distorsion transitoire associée à la règle de suppression d'Ephraim et Malah : l'influence du recouvrement entre les fenêtres de TFCT.

4.1.2.b Contrôle du niveau de bruit résiduel

Le dernier point important concernant le fonctionnement de l'algorithme d'Ephraim et Malah est que le paramètre $\mathcal{R}_{(min)}$ fournit un moyen de contrôler le niveau du bruit résiduel qui demeure après le traitement. Ce paramètre $\mathcal{R}_{(min)}$ correspond à la valeur minimale autorisée pour le RSB a priori : en dessous de cette limite la valeur de $\mathcal{R}_{prio}$ doit être forcée à $\mathcal{R}_{(min)}$. On montre que $\mathcal{R}_{(min)}$ est à peu près égal au gain moyen apporté aux parties du spectre qui correspondent au bruit de fond (annexe D).

On constate en pratique que si la valeur de ce paramètre $\mathcal{R}_{(min)}$ est trop faible, on obtient une réapparition du phénomène de bruit musical. Avec les paramètres usuels ($\alpha = 0,98$), cette limite de réapparition du bruit musical est environ de -15 dB. Cette limite s'interprète simplement par le fait que le lissage du RSB a priori est encore trop faible avec $\alpha = 0,98$: le gain calculé dans les parties de bruit seul présente des fluctuations qui sont suffisamment importantes pour se traduire par un effet de bruit musical de faible niveau. Pour éviter d'avoir à augmenter la valeur du paramètre α (ce qui produirait des distorsion accrues au moment des transitoires), on tronque les valeurs les plus basses du RSB a priori (annexe D).

Ce qui est intéressant c'est qu'il est donc possible de fixer la réduction moyenne du niveau de bruit grâce au paramètre $\mathcal{R}_{(min)}$ (avec un maximum de 15 dB si on veut éviter la réapparition du bruit musical). On peut même proposer un raffinement qui consiste à définir un paramètre $\mathcal{R}_{(min)}$ qui dépend de la pulsation discrète ω_k. Avec cette modification, il devient possible de contrôler le niveau de bruit résiduel *en fonction de la fréquence*. Cette modification peut être utile si on désire modifier la balance spectrale du bruit de fond : si $\mathcal{R}_{(min)}$ est constant dans tout le domaine des fréquences le spectre du bruit résiduel et le même que celui du bruit de fond (au facteur multiplicatif $\mathcal{R}_{(min)}$ près), par contre, si $\mathcal{R}_{(min)}$ varie en fonction de la fréquence, il est possible de modifier, dans une certaine mesure, le spectre du bruit résiduel.

Réduction supplémentaire du niveau de bruit Une question importante est de savoir quelle est la démarche à suivre lorsqu'on désire une réduction du niveau de bruit plus importante que 15 dB. Une première réponse peut être de surestimer le niveau de bruit. Toutefois nous savons qu'il faut limiter au maximum la surestimation du bruit pour éviter la distorsion du signal. Une solution plus radicale consiste à remplacer la formule de calcul de l'atténuation proposée par Ephraim et Malah par la règle de suppression dite de Wiener (paragraphe 2.2) évaluée en fonction du RSB a priori $\mathcal{R}_{prio}$. On peut montrer qu'avec ce choix, l'élimination du phénomène de bruit musical est encore garantie [Ephraim 84]. L'intérêt d'une telle substitution est que la règle de suppression de Wiener fournit une atténuation beaucoup plus importante pour les faibles valeurs de RSB que celle de la soustraction spectrale : les relations (2.15) et (2.13) (paragraphe 2.2) indiquent que l'atténuation apporté *en décibels* par la règle dite de Wiener est deux fois plus importante (ce qui équivaut à G^2 en linéaire). Dans les mêmes conditions que précédemment, il devient donc possible de régler, par la valeur de $\mathcal{R}_{(min)}$, l'atténuation moyenne du bruit de fond entre 0 et -30 dB.

Cependant, il faut souligner que la procédure qui vient d'être décrite ne peut en général pas être appliquée dans le cas d'enregistrements fortement bruités. La première raison est liée à l'algorithme d'Ephraim et Malah, on constate en effet que la distorsion transitoire mentionnée au paragraphe 4.1.2.a augmente de manière importante lorsqu'on utilise la règle de Wiener. Ce qui traduit le fait que la sur-atténuation des composantes de bas niveau au moment du transitoire est d'autant plus marquée que l'on utilise une règle de suppression qui provoque une forte atténuation. La seconde raison est beaucoup plus générale, elle s'applique d'ailleurs à toutes les règles de suppression : *les composantes de signal de très faible niveau sont traitées comme les parties qui correspondent au bruit seul.* Nous avons en effet eu l'occasion de voir qu'en dessous d'un certain niveau la présence de signal devient indétectable sur le spectre à court-terme (voir par exemple la figure 3.19). Par conséquent, la limitation de l'atténuation contrôlée par $\mathcal{R}_{(min)}$ ne vaut pas uniquement pour le bruit, mais aussi pour les parties du spectre de signal de faible amplitude. Pour un enregistrement fortement bruité, une élévation de $\mathcal{R}_{(min)}$ n'a donc pas seulement pour effet de relever le niveau du bruit résiduel, elle permet en même temps de réduire la distorsion apportée au signal. En pratique, pour la plupart des enregistrements dont nous disposions, surtout ceux provenant de disque 78 tours, une réduction importante du bruit (plus de 20 dB) se traduisait toujours par une très nette aggravation des distorsions du signal.

En conclusion, il faut retenir que l'algorithme d'Ephraim et Malah fournit un moyen simple de contrôler le niveau du bruit résiduel, éventuellement en fonction de la fréquence, dans certaines limites (réduction maximale de l'ordre de 15 dB). Pour les enregistrements fortement bruités, c'est en général la distorsion du signal qui empêche de réduire le niveau de bruit au delà d'une certaine limite. Il est difficile d'en dire beaucoup plus sur ce sujet car le choix entre le niveau acceptable de bruit résiduel et la distorsion tolérable du signal dépend surtout du cas particulier que constitue chaque enregistrement. De plus un tel choix comprend forcément une grande partie subjective qui dépend, par exemple, de l'objectif recherché lors de la restauration.

4.2 Restauration sélective des signaux de sous-bande

D'après les résultats du chapitre 3, la durée de la trame à court-terme utilisée pour la réduction de bruit de fond doit vérifier deux contraintes incompatibles :

- D'une part, la durée de la fenêtre d'analyse doit être suffisamment importante, c'est à dire obligatoirement supérieure à 30 ms pour éviter que l'effet de modulation ne soit perçu (paragraphe 3.3.1), et de préférence au dessus de 60 ms afin de ne pas éliminer de composantes audibles du signal (paragraphe 3.1.1).
- Inversement, la durée de la fenêtre doit rester aussi courte que possible pour éviter le phénomène de lissage des transitoires (paragraphe 3.1.2).

De plus, pour des enregistrements fortement bruités, on sait que dans les parties où le signal musical est stationnaire, une augmentation de la durée de la fenêtre permet d'abaisser la limite de restauration. Cependant la limite de restauration s'abaisse assez lentement puisqu'une multiplication de la durée de la fenêtre de TFCT par un facteur 2, ne permet de réduire la limite de restauration que de 3 dB. C'est à dire que pour obtenir une réelle augmentation des performances du système il faudrait pouvoir utiliser des fenêtres de TFCT de plusieurs centaines de millisecondes ce qui est impossible à cause des problèmes posés par la présence de transitoires musicaux.

Ce compromis sur le choix de la durée de la fenêtre d'analyse peut d'ailleurs se formuler autrement en mettant en évidence les contraintes que doivent vérifier les *filtres de sous-bande*, c'est à dire les filtres passe-bande équivalents à la transformée à court-terme. Ces filtres de sous-bande doivent à la fois être très sélectifs pour mettre en évidence les partiels des sons stables, et en même temps, avoir une réponse impulsionnelle courte pour éviter le lissage des transitoires abrupts. Or on sait que ces deux propriétés sont incompatibles entre elles.

Une idée qui exploite de manière explicite la distinction entre les parties stationnaires et transitoires du signal consiste à adapter les caractéristiques de la transformation à court-terme utilisée selon la situation locale. L'intérêt de cette approche est qu'un premier test, portant sur la nature du signal, permet de choisir la *durée locale d'observation du signal*. Ce choix doit prendre en compte deux arguments : d'une part, que seule l'augmentation de la durée locale d'observation est susceptible de mettre en évidence des composantes de signal noyé dans le bruit, et d'autre part, que la modification du signal sur une durée trop longue provoque une distorsion dans les parties transitoires. Selon la situation locale c'est l'un ou l'autre des arguments qui doit donc être prépondérant. La caractéristique distinctive de cette approche est l'utilisation de différentes durées d'observation.

Le paragraphe 4.2.1 présente un première application très simple et intuitive de ce principe, où le signal temporel est étiqueté en deux catégories : transitoires ou zones stationnaires. Cependant, cette approche globale, qui ne prend pas en compte la composition fréquentielle du signal, s'avère assez limitée. La suite du paragraphe 4.2 développe l'idée inverse où l'on cherche à détecter la présence de composantes de signal stationnaires. Cette détection s'effectue en observant le comportement à long-terme, c'est à dire, en considérant plusieurs échantillons consécutifs des signaux de sous-bande obtenus par transformée de Fourier à court-terme. La différence majeure par rapport à la technique du paragraphe 4.2.1 est que le traitement n'est plus global : à l'indice temporel p, certains signaux de sous-bandes sont traités en utilisant une durée d'observation longue pour rejeter au maximum le bruit (cas des sous-bandes où une composante de signal a été détectée), tandis que dans les autres sous-bandes, le signal est traité ponctuellement pour éviter le lissage temporel.

4.2.1 Un premier essai : durée de fenêtre variable

4.2.1.a Principe

Le traitement, que nous avons proposé dans [Cappe 92], consiste à effectuer l'atténuation spectrale à court-terme en utilisant, a priori, une durée de fenêtre importante $\mathcal{D}_l$, qui est réduite (durée $\mathcal{D}_c$) uniquement lorsque le signal présente un comportement transitoire. Il faut souligner qu'ici le terme de "transitoire" correspond à une variation importante des caractéristiques du signal *à l'échelle de la durée* $\mathcal{D}_l$. En effet, si la variation constatée sur la durée d'observation $\mathcal{D}_l$ est faible, il n'y a aucun intérêt à utiliser des fenêtres de TFCT de durée plus faible, même si la partie de signal considérée correspond à la partie transitoire d'un son.

La réalisation décrite dans [Cappe 92] revient à effectuer une transition douce, par simple pondération, entre le résultat obtenu avec des fenêtres de TFCT longues et celui obtenu avec des fenêtres de courte durée, et ce, autour du transitoire détecté. La figure 4.1 présente un exemple correspondant à un cas réel d'enregistrement bruité. Sur cette figure, la fenêtre de pondération représentée en pointillés est appliquée au résultat obtenu avec des fenêtres courtes, tandis que la fenêtre complémentaire est appliquée à celui obtenu en utilisant des fenêtres longues. On note $\mathcal{D}_P$ la durée de la fenêtre de pondération, sur la figure 4.1, cette durée vaut environ 30 ms.

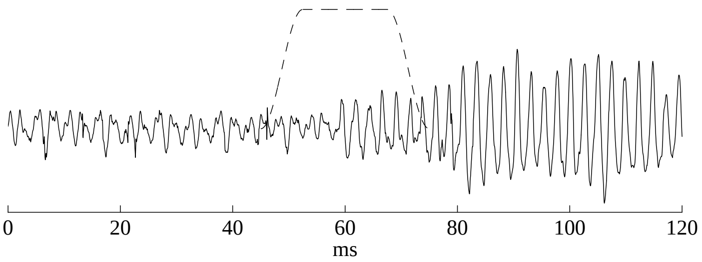

Figure 4.1: Exemple de fenêtre de pondération utilisée autour d'un transitoire. Le résultat obtenu avec des fenêtres courtes est pondéré par la fenêtre représentée en pointillés.

4.2.1.b Mise en œuvre et limitations

La réalisation du traitement proprement-dit ne pose pas de problème. Par contre, l'application à un cas réel nécessite une procédure permettant de détecter automatiquement la présence de parties transitoires dans le signal à traiter. A priori, la conception de cette procédure est assez difficile car nous ne disposons pas d'une caractérisation simple des parties transitoires d'un signal musical.

Dans [Cappe 92] ce problème est résolu en considérant une définition restrictive de la notion de transitoire qui consiste à dire qu'un transitoire est un phénomène qui se manifeste par une variation brusque de la puissance du signal. La procédure proposée consiste donc simplement à détecter des variations brusques de la puissance du signal intégrée sur une durée de l'ordre de $\mathcal{D}_l$. Pour rationaliser cette détection, l'approche utilisée consiste à se ramener à une situation standard, dans le domaine de la détection de rupture : le cas d'un saut dans la moyenne des observations [Basseville 88]. Le test obtenu permet de localiser assez précisément les variations brusques du signal, il présente surtout l'intérêt d'être relativement robuste [Cappe 92]. Cependant, cette procédure reste extrêmement sommaire car la vision du signal obtenue à travers la courbe de puissance est très réductrice. En particulier, dans les cas où le signal musical enregistré est polyphonique, ce type de procédure sous-estime notablement le nombre réel de transitoires présents.

Toutefois, la limitation principale de cette technique n'est pas la difficulté liée au marquage des zones transitoires du signal, mais plutôt le faible degré de liberté dans le choix des différentes durées :

Durée $\mathcal{D}_c$ des fenêtres courtes Pour éviter le lissage des transitoires, la durée des fenêtres courtes $\mathcal{D}_c$ doit être choisie la plus faible possible. En pratique, il est très difficile d'utiliser une durée $\mathcal{D}_c$ inférieure à 10 à 20 ms. En effet, nous avons vu au paragraphe 3.3 que si la durée de la fenêtre de TFCT est trop brève, il y a apparition d'un phénomène audible de modulation. Il s'avère que ce phénomène est tout à fait audible bien qu'il n'apparaisse que transitoirement pendant une durée $\mathcal{D}_P$ (durée de la fenêtre de pondération). L'effet produit est celui d'une bouffée de bruit apparaissant de manière synchrone avec le transitoire.

Durée $\mathcal{D}_P$ de la fenêtre de pondération Avec une durée $\mathcal{D}_c$ de l'ordre de 20 ms, on sait d'après les résultats du paragraphe 3.1.1 que les résultats obtenus sont très mauvais dès lors que l'on a affaire à une partie stationnaire du son. [Cappe 92] rapporte un autre effet secondaire constaté : l'augmentation locale, autour du transitoire, du phénomène de bruit musical. D'après le paragraphe 3.2.1, cet effet peut aussi être relié à la faible valeur de $\mathcal{D}_c$. Par conséquent, afin d'éviter que ces différents défauts ne soient perçus, la fenêtre de pondération doit rester de courte durée. En effectuant des tests d'écoute plus approfondis, il s'est avéré que les résultats les plus satisfaisants sont en général obtenus quand $\mathcal{D}_P$ est inférieure à 30 ms. Au delà, les problèmes associés à la durée de fenêtre courte deviennent assez gênants. Cette valeur est assez variable car elle dépend du type exact de fenêtre de pondération utilisée, ainsi que de la nature du signal transitoire. En plus, il faut souligner que ce type d'évaluation auditive informelle du comportement transitoire est assez difficile à effectuer, et en général, très peu fiable.

Durée $\mathcal{D}_l$ des fenêtres longues Nous avons vu, au paragraphe 3.1.2, que le phénomène de lissage des transitoires peut se produire, dans le pire des cas, sur une durée équivalente à 1,5 fois la durée de la fenêtre de TFCT. Autrement dit, si on désire que ce lissage soit éliminé par la transition vers des fenêtres de traitement courtes, il faut au moins que la valeur $\mathcal{D}_l$ reste plus faible que celle de $\mathcal{D}_P$, c'est à dire inférieure à 30 ms. Dans ce cas, la durée $\mathcal{D}_l$ est du même ordre de grandeur que $\mathcal{D}_c$, et l'ensemble de la procédure est totalement inutile ! En pratique, ce raisonnement basé sur le cas de l'apparition d'une composante de très bas niveau située à la limite de restauration (cf. paragraphe 3.1.2), est un peu pessimiste. On constate en fait une amélioration du comportement transitoire tant que $\mathcal{D}_l$ reste inférieur à 50 ms. Au delà, le lissage dû à l'utilisation de fenêtres de durée longue se manifeste nettement hors de l'intervalle délimité par la fenêtre de pondération.

En conclusion, la limitation de ce traitement à taille de fenêtre variable est que les contraintes à respecter lors du choix des durées de fenêtre ne permettent pas d'obtenir une réelle amélioration : le mieux que l'on puisse faire est de choisir $\mathcal{D}_c$ de l'ordre de 20 ms, et $\mathcal{D}_l$ d'environ 50 ms. C'est à dire que le rapport entre les deux durées de fenêtre n'est que de 2,5. L'analyse simplifiée exposée au chapitre 3 prévoit bien des différences observables entre ces deux cas. Cependant, dans une situation réelle, il devient très difficile de juger de l'efficacité de la technique. Plus précisément, les séances d'écoute informelles effectuées nous permettent de dire que le résultat obtenu est en général discernable des résultats obtenus par atténuation spectrale simple en utilisant, soit des fenêtres courtes de durée $\mathcal{D}_c$, soit des fenêtres longues de durée $\mathcal{D}_l$. Par contre, pour des valeurs intermédiaires de la durée de la fenêtre, c'est à dire celles qui sont utilisées en pratique, la différence n'est pas toujours perceptible. Compte-tenu de l'absence de protocole d'évaluation, on ne peut pas affirmer que la différence est imperceptible, il resterait à mettre en œuvre un véritable test psychoacoustique. Cependant il est certain que le gain obtenu par ce traitement est faible.

4.2.1.c Conclusions

Comme nous venons de le voir, l'utilisation de fenêtres de durée variable ne permet pas d'améliorer significativement les résultats, compte tenu du très faible degré de liberté disponible pour le choix des paramètres (durées respectives des différentes fenêtres). La cause principale de cet échec relatif est la globalité de la technique proposée. En effet, même dans une trame à court-terme qui contient une portion transitoire du signal, le spectre à court-terme présente en général des maxima significatifs. Le paragraphe 3.1.2 a permis de montrer que c'est le comportement de la technique de débruitage *autour de ces maxima* qui cause le lissage des transitoires. Or, la technique proposée ne tient pas compte de cet aspect “local” du problème. La solution qui consiste à modifier globalement la taille de la fenêtre, n'influe donc pas uniquement sur le phénomène de lissage des transitoires. C'est ce qui explique que le gain obtenu (réduction du phénomène de lissage) soit contrebalancé par des effets négatifs (comportement vis à vis des composantes stables du signal).

Ce manque de précision dans le traitement utilisé (diminution de la taille de la fenêtre) caractérise aussi la procédure de détection des zones à traiter. En effet, d'après les résultats du paragraphe 3.1.2, seuls les transitoires de “faible niveau” subissent une distorsion. Le comportement transitoire du signal n'implique donc pas forcément l'effet de lissage. Là encore, ce qui manque ici c'est une procédure permettant de signaler précisément la présence de composantes transitoires du signal de bas niveau susceptibles de subir une distorsion.

Il apparaît donc que le principe retenu pour le traitement nous conduit, tout d'abord, à effectuer une tâche très complexe pour mettre en évidence les défauts éventuels (détection de composantes transitoires fortement bruitées), et par suite, à mettre au point une technique de traitement appropriée visant à limiter l'effet de lissage. A priori, ce dernier objectif est lui aussi très délicat car l'effet de lissage est lié à la sélectivité du filtrage équivalent au traitement de débruitage autour des composantes du signal de bas niveau (cf. paragraphe 3.1.2).

Suite à ce premier essai, nous avons donc été amené à reformuler complètement le principe du traitement. L'idée retenue correspond à une constatation simple : c'est que les deux tâches mentionnées précédemment deviennent beaucoup plus aisées lorsque l'on utilise le principe inverse qui consiste à utiliser, a priori, une durée de fenêtre courte. En effet, les parties du signal qui doivent être détectées, puis traitées spécifiquement sont alors les composantes stables (cf. paragraphe 3.1.1). On se retrouve donc dans un cadre plus classique qui est celui de signaux quasi-stationnaires.

4.2.2 Détection des composantes sinusoïdales du signal

Avant d'exposer le principe complet du traitement de débruitage que nous avons proposé dans [Cappe 93a], nous allons présenter en détail une des parties de ce traitement : la détection de composantes sinusoïdales du signal de bas niveau, à partir du spectre à court-terme du signal. Nous verrons au paragraphe suivant que cette procédure de détection correspond à l'élément le plus important de la technique de débruitage proposée. Nous avons choisi de présenter cette partie en premier, comme un problème séparé, car la démarche suivie pour la mise au point de la procédure de détection fournit naturellement une bonne introduction au traitement proposé pour le débruitage.

4.2.2.a Problème posé par la présence de bruit

Dans cette partie, on suppose donc que le signal musical bruité $x(n)$ est analysé par transformée de Fourier à court-terme (notée $X(p, \omega_k)$). Le problème posé est de déterminer, à partir du spectre à court-terme $X(p, \omega_k)$, la présence éventuelle d'une composante sinusoïdale de signal de pulsation proche d'une pulsation discrète ω_k donnée.

En l'absence de bruit, le problème est assez simple : puisque la TFCT réalise une analyse spectrale du signal contenu dans la trame à court-terme, il suffit de déterminer quels sont les pics significatifs sur le module du spectre à court-terme $|X(p, \omega_k)|$. Cette étape de détection de pics est fréquemment utilisée pour l'analyse des signaux musicaux (voir, par exemple, [Serra 89] ou [Mahieux 89, §4.2.3]). La seule difficulté posée par cette procédure est la mise au point de critères heuristiques permettant d'éviter la détection des lobes secondaires associés à un pic spectral (effet du fenêtrage).

En présence de bruit, la détection de pic ne peut plus être utilisée pour détecter les sinusoïdes de faible niveau. En effet, nous avons vu que le niveau relatif, mesuré à une pulsation discrète ω_k, présente une variance très importante, aussi bien dans le cas où le signal est constitué de bruit seul autour de la pulsation ω_k (paragraphe 3.2.1.b), que dans celui où une composante de signal de faible niveau est présente (paragraphe 3.2.2). Une étude plus précise des intervalles de confiance déterminés aux paragraphes 3.2.1.b et 3.2.2, montre qu'une composante de signal ne peut être détectée "dans de bonnes conditions" (taux d'oublis et taux de fausses détections inférieurs au pourcent) que si son niveau relatif moyen est supérieur à 15 dB.

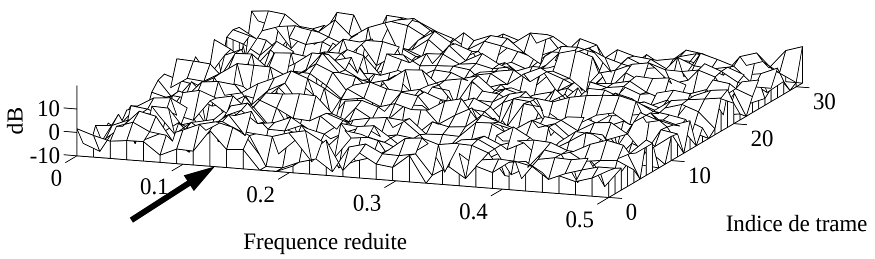

Figure 4.2: Module de la TFCT dans le cas d'une sinusoïde de faible niveau noyée dans un bruit blanc. Les spectres correspondant à 32 trames à court-terme successives ont été représentés. La flèche indique la fréquence de la sinusoïde. Le niveau relatif moyen mesuré à la fréquence de la sinusoïde est de 6 dB.

La figure 4.2 permet de se faire une idée de la difficulté de la détection à partir du module du spectre à court-terme. Pour le cas représenté (sinusoïde située, en moyenne, 6 dB au dessus du niveau de bruit), il déjà presque impossible de repérer visuellement la position de la sinusoïde. Une idée naturelle consiste à essayer de considérer le résultat obtenu dans plusieurs spectres à court-terme successifs. En effet, si on suppose que la composante sinusoïdale présente dans le signal est stable sur une durée importante (c'est à dire bien supérieure à la durée de trame à court-terme), on peut chercher à obtenir un spectre de puissance moyen qui sera plus significatif. C'est ce qui se passe visuellement sur la figure 4.2 où il est tout de même possible de distinguer la présence d'une composante de signal en cherchant la fréquence où les spectres successifs sont les moins chahutés. Cependant, le tableau 1.3 (chapitre 1) rappelle que pour obtenir un spectre de puissance de faible variance il est nécessaire de moyenner sur un nombre de trames successives très important. Si on se fixe par exemple comme objectif de détecter les composantes de signal jusqu'à un niveau relatif moyen de 3 dB, le tableau 1.3 montre qu'il faudrait au moins moyenner les spectres obtenus dans une quarantaine de trames à court-terme successives (sans recouvrement). Si on suppose que la durée d'une trame à court-terme est de l'ordre de 40 ms, ceci implique que la composante sinusoïdale de signal que l'on cherche à détecter soit stable sur une durée d'au moins une seconde. Même pour un signal musical, cette durée est extrêmement longue ce qui limite l'intérêt d'une telle technique de détection.

Nous allons voir qu'ici c'est la manière d'utiliser l'information disponible qui est mauvaise car le module du spectre n'est plus la seule quantité exploitable dès lors que l'on considère plusieurs spectres à court-terme successifs. Par contre, cet exemple permet déjà de mettre en évidence un point très important : plus le nombre de spectres à court-termes successifs pris en compte est grand, plus il est possible de détecter des composantes de bas niveau. Cependant, la durée totale d'observation est alors de l'ordre du nombre T de spectres considérés multiplié par la durée RF_e correspondant au décalage entre deux trames à court-terme successives. La technique de détection ne présente donc un intérêt que si cette durée totale ($T \times RF_e$) est inférieure à la durée "moyenne" des sons musicaux présents dans l'enregistrement à traiter.

4.2.2.b Comportement des valeurs successives de la TFCT

Il est utile de rappeler ici quelques implications de l'interprétation de la TFCT en tant que banc de filtres pour deux exemples de signaux simples (composante sinusoïdale et bruit). Ces constatations sont à l'origine de la procédure de détection proposée.

Analyse d'un signal sinusoïdal On considère dans un premier temps l'analyse par TFCT d'un simple signal sinusoïdal. La relation (C.8) (annexe C) indique que la TFCT d'un signal sinusoïdal de pulsation Ω s'écrit, pour la pulsation discrète ω_k correspondant au sommet du pic spectral

$$S(p, \omega_k) = A\, \mathrm{e}^{\mathrm{j}(\Omega - \omega_k)pR}$$

où le terme complexe A dépend de l'amplitude et de la phase de la sinusoïde analysée. Le signal de sous-bande $S(p, \omega_k)$, considéré comme une fonction de l'indice temporel p, est donc une exponentielle complexe de pulsation $(\Omega - \omega_k)R$ (la valeur de la pulsation dépend du formalisme adopté pour décrire la TFCT, il s'agit ici du résultat pour la convention passe-bas). Cette écriture découle simplement de l'interprétation de la TFCT en tant que banc de filtres complexes sous-échantillonné telle qu'elle est détaillée dans l'annexe B (voir par exemple la figure B.1). Lorsque le signal analysé contient une composante sinusoïdale, les valeurs successives du spectre à court-terme $S(p, \omega_k)$, autour du pic spectral, forment donc un signal cohérent. En particulier, l'observation de plusieurs valeurs successives de ce signal permet, entre autres, d'accéder à une estimation précise de la fréquence du signal. Cette constatation sert de base à la technique du *phase vocoder* utilisée pour modifier les caractéristiques des signaux audio [Moorer 78] [Portnoff 81b].

Analyse d'un bruit Supposons maintenant que le signal analysé est un bruit stationnaire $d(n)$. Le signal de sous-bande $D(p, \omega_k)$, obtenu pour une pulsation discrète ω_k donnée, s'interprète comme le résultat du filtrage du bruit initial par le filtre d'analyse $h(n)$, sous-échantillonné d'un facteur R (figure B.1). Pour déterminer plus précisément les caractéristiques de ce bruit, supposons dans un premier temps que $d(n)$ est un bruit blanc, de variance σ^2. D'après la relation (B.7) (annexe B), la TFCT du bruit s'écrit (en convention passe-bas)

$$D(p, \omega_k) = \sum_{n=-\infty}^{+\infty} h(pR - n)d(n)W_N^{-kn}$$

On en déduit la fonction d'autocorrélation du signal de sous-bande à la pulsation ω_k

$$\begin{aligned} R_{DD}(q) &= \mathrm{E}\left\{D(p+q, \omega_k)D^*(p, \omega_k)\right\} \\ &= \sum_{n=-\infty}^{+\infty} \sum_{m=-\infty}^{+\infty} h[(p+q)R - n]h[pR - m]\mathrm{E}\left\{d(n)d(m)\right\} W_N^{-k(n-m)} \end{aligned}$$

Comme le bruit $d(n)$ est un bruit blanc, cette expression devient

$$R_{DD}(q) = \sigma^2 \sum_{n=-\infty}^{+\infty} h[(p+q)R - n]h[pR - n]$$

En posant $m = pR - n$, cette dernière expression se réécrit sous la forme

$$R_{DD}(q) = \sigma^2 \sum_{m=-\infty}^{+\infty} h(m)h(m + qR) \tag{4.1}$$

La fenêtre d'analyse de la TFCT $h(n)$ étant de longueur N, on vérifie facilement que $R_{DD}(q) = 0$ dès que $|q| > N/R$. La fonction d'autocorrélation du signal de sous-bande est donc limitée à quelques termes. En particulier, quand il n'y a pas de recouvrement entre les fenêtres ($R \geq N$), la fonction d'autocorrélation de $D(p, \omega_k)$ se réduit à un terme non nul : le signal de sous-bande est un bruit blanc. A l'inverse, lorsqu'il n'y a pas de décimation des spectres à court-terme, la fonction d'autocorrélation du signal de sous-bande s'écrit :

$$R_{DD}(q) = \sigma^2 \left\{h(q) * h(-q)\right\}$$

ce qui correspond à la formule du classique de filtrage d'un bruit blanc [Charbit 90] (par $h(n)$, qui joue le rôle de filtre d'analyse de la TFCT). On vérifie par ailleurs que la puissance du signal de sous-bande ($R_{DD}(0)$) vaut $\sigma^2 \sum h^2(m)$ quel que soit le recouvrement entre les fenêtres successives. Ce résultat correspond simplement à la formule (C.14), écrite dans le cas d'un bruit blanc.

La figure 4.3 présente la densité spectrale de puissance correspondant à la fonction d'autocorrélation calculée grâce à la formule (4.1), pour trois valeurs du paramètre de recouvrement. Lorsque le recouvrement est important (cas 75%, en tirets), la DSP du signal de sous-bande est concentrée autour de la fréquence nulle du fait du filtrage réalisé par $h(n)$. Par contre, dans le cas d'un recouvrement de 50% (trait plein), cet effet de filtrage est quasiment inexistant à cause du repliement spectral qui se produit lors de la décimation des voies de TFCT.

Pour le cas, plus général, d'un bruit analysé $d(n)$ de DSP quelconque, l'interprétation de la TFCT en tant que banc de filtres montre que le résultat précédent ne doit pas être modifié si la DSP de $d(n)$ "varie peu" autour de la fréquence ω_k. Sous cette hypothèse, la DSP du signal de sous-bande à la pulsation discrète ω_k peut s'écrire sous la forme approchée suivante

$$R_{D_{(\omega_k)}D_{(\omega_k)}}(q) \approx P_d(\omega_k) \sum_{m=-\infty}^{+\infty} h(m)h(m + qR) \tag{4.2}$$

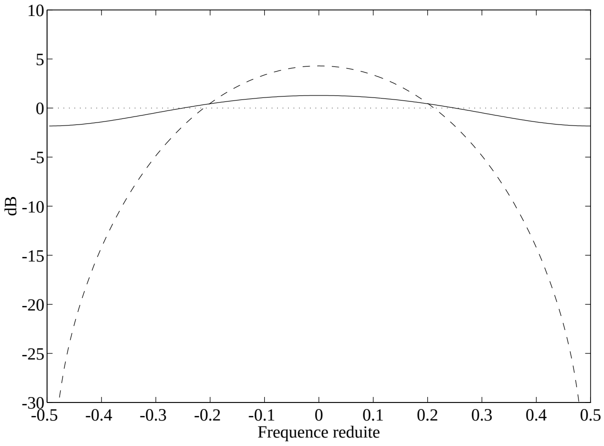

Figure 4.3: Densité spectrale de puissance du signal présent dans une voie de la TFCT d'un bruit blanc. **En pointillés,** pas de recouvrement, **en trait plein,** recouvrement 50%, **en tirets,** recouvrement 75%. La fenêtre d'analyse est une fenêtre Hann. Le niveau 0 dB correspond à la puissance du signal de sous-bande ($R_{DD}(0)$).

En pratique, pour les bruits d'enregistrements, cette formule est bien vérifiée dès que la résolution fréquentielle de la TFCT est suffisante (durée de trame supérieure à 10 ms), sauf pour les très basses fréquences (en dessous de 150 Hz) où l'on ne plus supposer que la DSP du bruit varie peu (cf. paragraphe 1.4.3). En excluant ce cas, nous pouvons donc considérer que la DSP du bruit filtré correspond, dans chaque sous-bande, à celle qui est représentée sur la figure 4.3. Par contre, la puissance totale du bruit filtré varie selon la position de la sous-bande considérée.

4.2.2.c Détection à partir du signal de sous-bande complexe

Revenons maintenant au problème de la détection d'une composante sinusoïdale de signal, de faible niveau, en présence de bruit de fond. Afin de pouvoir formuler simplement une stratégie de détection efficace, nous allons simplifier la situation réelle en émettant les deux hypothèses suivantes :

- **La résolution spectrale de la TFCT est suffisante pour permettre d'isoler les composantes de signal.** Pour un signal musical la résolution nécessaire pour assurer une réelle séparation des composantes du signal est assez importante. Pour un son instrumental isolé, une résolution de l'ordre de 50 Hz est bien suffisante dans la plupart des cas. Par contre, dans le cas d'un enregistrement polyphonique, la résolution nécessaire peut être beaucoup plus importante. Notons que pour la réduction de bruit, cette condition est plus facile à remplir car on ne s'intéresse qu'à des composantes de signal proches du niveau de bruit. L'hypothèse réellement effectuée consiste donc plutôt à dire que dans toutes les voies de la TFCT où le niveau relatif est compris entre 0 et 15 dB (par exemple), il existe au plus une composante de signal. Nous aurons l'occasion de revenir sur la portée réelle de cette hypothèse au cours des paragraphes suivants. Nous avons vu que cette hypothèse permet d'affirmer que la partie "signal" contenue dans une voie de la TFCT est une exponentielle complexe.
- **Le bruit filtré présent dans la sous-bande est un bruit blanc.** D'après les résultats du paragraphe précédent, cette hypothèse n'est vérifiée que pour les faibles valeurs du recouvrement. Nous admettons ici que cette hypothèse reste vérifiée dans le cas d'un

recouvrement de 50%. Ceci revient à négliger la légère variation (de l'ordre de 3 dB) de la DSP du bruit de sous-bande constatée sur la figure 4.3. Cette hypothèse **fixe la valeur du recouvrement à 50%**, ce qui est de toute façon la valeur généralement utilisée pour le débruitage. Par ailleurs, on considère aussi que le bruit filtré présent dans la sous-bande est gaussien. Nous avons vu au paragraphe 2.2.3 que cette supposition est justifiée même si le bruit de fond lui-même n'est pas gaussien.

Avec ces simplifications, la présence d'une composante de signal de bas niveau peut donc se formuler de manière très simple sous la forme d'un test entre deux hypothèses. Etant donné une observation formée de T valeurs successives du signal de sous-bande

$$\mathbf{X} = [X(p_0, \omega_k) \ldots X(p_0 + T - 1, \omega_k)]^T$$

il s'agit de déterminer laquelle des deux hypothèses suivantes est la plus probable

H1: Présence d'une composante Le signal de sous-bande contient un signal (exponentielle complexe), de paramètres (fréquence, amplitude complexe) inconnus, mêlé avec le bruit filtré.

H0: Bruit seul Le signal de sous-bande est constitué uniquement du bruit filtré (qui est un bruit blanc gaussien de puissance connue).

Il s'agit ici d'un *test binaire d'hypothèses composites* [Van Trees 68] car la densité de probabilité des observations n'est pas complètement spécifiée sous l'hypothèse H1, elle dépend de deux paramètres inconnus : l'amplitude et la fréquence de l'exponentielle complexe. Une stratégie de détection applicable dans ce cas consiste à utiliser le rapport de vraisemblance généralisé défini par

$$\Lambda = \frac{\max\limits_{\theta} \{f_{\mathrm{H1}}(\mathbf{X}, \theta)\}}{f_{\mathrm{H0}}(\mathbf{X})} \tag{4.3}$$

où $f_{\mathrm{H0}}(\mathbf{X})$ désigne la vraisemblance du vecteur des observations $\mathbf{X}$ sous l'hypothèse H0 (de même pour H1). La maximisation de la vraisemblance sous l'hypothèse H0 est inutile puisqu'il n'y a pas de paramètre inconnu. Sous l'hypothèse H1, la maximisation de la vraisemblance doit a priori se faire sur les deux paramètres inconnus $\theta = [A\ \phi]$ (amplitude complexe et fréquence). L'hypothèse H1 est considérée comme la plus probable lorsque Λ est supérieur à un seuil donné (et inversement). Compte tenu du caractère gaussien du bruit, et de l'hypothèse d'un bruit blanc, le rapport de vraisemblance s'écrit ici [Van Trees 68] [Kay 93]

$$\Lambda = \frac{\max\limits_{(A,\phi)} \left\{ \exp\left[-\frac{1}{v} (\mathbf{X} - \mathbf{S}_{(A,\phi)})^H (\mathbf{X} - \mathbf{S}_{(A,\phi)}) \right] \right\}}{\exp\left[-\frac{1}{v} \mathbf{X}^H \mathbf{X} \right]} \tag{4.4}$$

où $\mathbf{S}_{(A,\phi)}$ désigne le vecteur correspondant à l'exponentielle complexe $[1\ A\mathrm{e}^{\mathrm{j}\phi} \ldots A\mathrm{e}^{\mathrm{j}\phi(T-1)}]^T$, et v représente la variance du bruit de sous-bande.

En fait, le maximum de la vraisemblance d'une observation formée d'une exponentielle complexe dans un bruit blanc (numérateur) ne dépend pas de l'amplitude de la composante. Le calcul correspondant est assez classique [Kay 88, §13.3.1][Kay 93, Ex. 15.13], il a été détaillé dans [Cappe 93a]. On trouve finalement que

$$\log(\Lambda) = \frac{\max\limits_{\phi} \left\{ \left| \sum_{p=0}^{T-1} \mathbf{X}[p] \mathrm{e}^{-\mathrm{j}\phi p} \right|^2 \right\}}{Tv} \tag{4.5}$$

On reconnaît au numérateur l'expression du périodogramme (module au carré de la transformée de Fourier) de l'observation $\mathbf{X}$. Comme le bruit de sous-bande est un bruit blanc, le terme au dénominateur (Tv) peut s'interpréter comme la valeur moyenne du périodogramme du bruit présent dans la sous-bande (voir l'équation (C.13), annexe C, avec une fenêtre rectangulaire de longueur T). L'expression du logarithme du rapport de vraisemblance s'interprète donc simplement comme le maximum du périodogramme de l'observation divisé par la valeur moyenne du périodogramme du bruit. La stratégie de détection obtenue correspond à l'idée très intuitive que la présence d'une composante de signal est d'autant plus probable que le périodogramme de l'observation présente un pic qui s'élève significativement au dessus du niveau moyen de bruit. Il s'agit là d'une simple analyse spectrale du signal de sous-bande par transformée de Fourier, avec décision sur le spectre d'amplitude obtenu. Attention, ce résultat n'est plus valable dans le cas où la sous-bande est susceptible de contenir plusieurs exponentielles complexes (l'expression de la vraisemblance est alors beaucoup plus complexe [Kay 88]).

Rappelons que le but est de déterminer la présence d'une composante de signal dans une voie donnée de la TFCT (de pulsation centrale ω_k), et autour de l'indice temporel p. Il est donc préférable d'utiliser un vecteur d'observation centré autour de l'indice p de la forme

$$\mathbf{X} = [X(p - T_1, \omega_k) \dots X(p + T_2, \omega_k)]^T$$

où $T_2 + T_1 = (T - 1)$. Par ailleurs, dans le cas où le bruit $d(n)$ présent sur l'enregistrement n'est pas un bruit blanc, la variance du bruit présent dans la sous-bande dépend de la pulsation discrète ω_k considérée. Pour une pulsation discrète ω_k donnée, on a $v = \hat{P}_d(\omega_k)$, où $\hat{P}_d(\omega_k)$ est le niveau moyen de bruit mesuré au préalable. Avec ces notations, le test de présence de la composante dans la voie de pulsation centrale ω_k et à l'indice temporel p s'écrit

$$\Lambda_{(\log)} = \frac{\max\limits_{\phi} \left\{ \left| \sum_{m=p-T_1}^{p+T_2} X(m, \omega_k)\, \mathrm{e}^{-\mathrm{j}\phi m} \right|^2 \right\}}{T\hat{P}_d(\omega_k)} \quad \begin{array}{c} \mathrm{H1} \\ > \\ < \\ \mathrm{H0} \end{array} \quad \mathcal{S} \tag{4.6}$$

où $\mathcal{S}$ désigne la valeur du seuil. La notation $\Lambda_{(\log)}$ rappelle que la quantité évaluée correspond au logarithme du rapport de vraisemblance. Une remarque est que la maximisation du numérateur doit bien être effectuée pour toutes les valeurs de la pulsation $\phi \in [-\pi, \pi[$ car le signal de sous-bande $X(p, \omega_k)$ est un signal complexe.

4.2.2.d Evaluation

La figure 4.4 présente une illustration de la procédure de détection où les deux hypothèses sont représentées à gauche et à droite de la figure. Le test de détection consiste donc à évaluer le maximum du périodogramme du signal de sous-bande (en trait plein) et à diviser par le niveau moyen du bruit (représenté en pointillé). Pour le cas représenté à gauche (absence de composante) le rapport de l'équation (4.6) vaut environ 7 dB, tandis qu'il est de 17 dB pour le cas de droite (présence d'une composante). Ici, la présence d'une composante est très nettement mise en évidence dans le cas représenté à droite. Une remarque est que le niveau de la composante dans ce cas est le même que sur la figure 4.2, la procédure de détection correspondant à l'équation (4.6) procure donc bien une amélioration des possibilités de détection.

Fausses alarmes Notons qu'ici une fausse alarme correspond à la détection erronée d'une composante sinusoïdale dans une voie de la TFCT qui ne contient que du bruit. Compte tenu

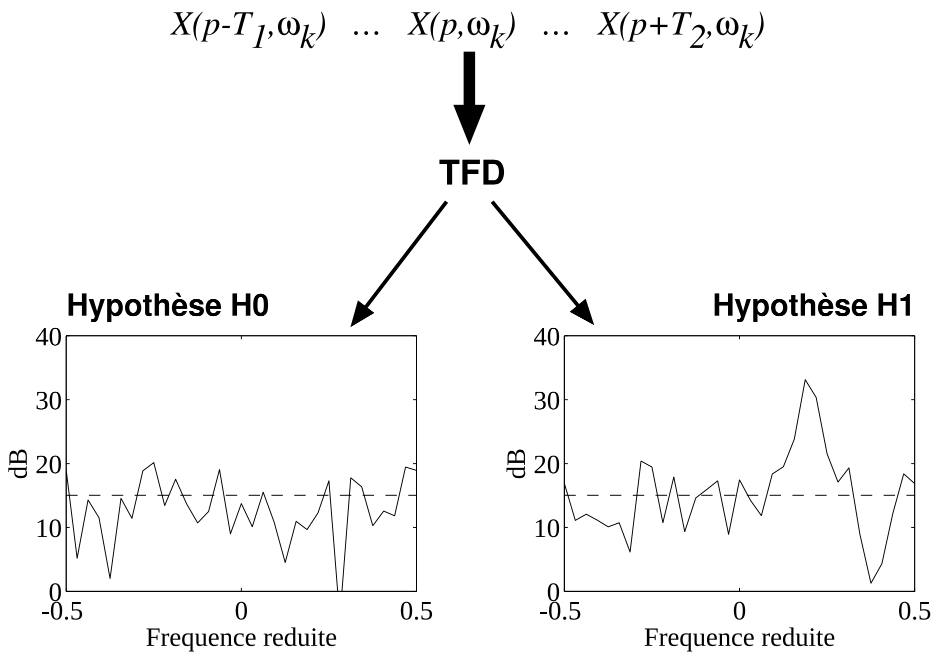

Figure 4.4: Illustration du test de détection de composante. A gauche, absence de composante. A droite, présence d'une composante de niveau relatif 6 dB. En pointillés, la valeur moyenne du périodogramme du bruit de sous-bande. $T = 32$ valeurs successives du signal de sous-bande sont prises en compte.

du type de méthode de restauration employée dans le cas d'une détection (paragraphe 4.2.5), ceci risque de se traduire par la génération de sinusoïdes parasites, c'est à dire par l'effet de bruit musical (cf paragraphe 3.2.1). Comme nous avons vu que cet effet est intolérable pour notre application, nous avons choisi de fixer les paramètres de la procédure de détection de telle façon que le taux de fausses alarmes soit quasiment nul. Pour éviter les fausses alarmes, il est nécessaire que la valeur du seuil $\mathcal{S}$ soit suffisante pour éliminer les pics dus à la variance du périodogramme du bruit présent dans la sous-bande (voir la partie gauche de la figure 4.4). Nous avons vu au paragraphe 3.2.1.b qu'une valeur de l'ordre de 8,5 dB est nécessaire pour assurer une probabilité de dépassement de $0,1\%$. En pratique, une valeur de 8,5 dB pour $\mathcal{S}$ donne un taux de fausses alarmes qui est légèrement plus important. L'explication de ce phénomène est que le bruit de sous-bande n'est pas tout à fait blanc. En effet, la figure 4.3 montre que pour les fréquences centrales de la sous-bande (entre -0,2 et 0,2 en fréquence réduite), le niveau de bruit est légèrement supérieur à la moyenne (1,5 dB au dessus pour la fréquence centrale). Par conséquent, pour obtenir un taux de fausses alarmes inférieur à 0,1%, il faut choisir une valeur du seuil $\mathcal{S}$ d'environ 10 dB.

Limite de détection Dans la pratique, le calcul de périodogramme nécessaire à l'évaluation du rapport de l'équation (4.6) est effectué en utilisant une transformée de Fourier discrète. Nous avons même intérêt à choisir une valeur de T qui soit une puissance de 2 afin de permettre un calcul rapide par FFT. Dans ce cas, la différence par rapport à la procédure décrite par l'équation (4.6), est que la maximisation est effectuée seulement pour les valeurs de la pulsation du type $\phi = 2\pi i/T$ (avec $i = 0, 1, \ldots, T-1$). Si la fréquence exacte de la composante ne correspond pas à un point de discrétisation fréquentielle, on sous-estime donc légèrement la valeur du maximum dans le cas où une composante est présente. Dans le cas où la fréquence de la composante correspond à un point de discrétisation fréquentielle, la valeur du périodogramme au sommet du pic est donnée par $T^2 |A|^2$ où A désigne l'amplitude complexe de la composante. Dans le pire des cas où la fréquence tombe entre deux points de discrétisation il faut compter 4 dB (2,5 en linéaire) en moins pour le sommet du pic [Harris 78]. En se basant sur cette dernière éventualité, la valeur moyenne du périodogramme du bloc de signal mesurée au sommet du pic spectral vaut

$$T^2 \times \frac{1}{2,5} \times |A|^2 + Tv \tag{4.7}$$

Le terme Tv représente la contribution moyenne du bruit de sous-bande de variance v (il s'agit du même calcul que dans l'annexe C, effectué dans le cas complexe). Le rapport de l'équation (4.6) vaut donc en moyenne, dans le cas de la présence d'une composante,

$$\mathrm{E}\left\{\Lambda_{(\log)}\right\} = T \times \frac{1}{2,5} \times \frac{|A|^2}{v} + 1$$

Le terme $|A|^2/v$ correspond au rapport signal-à-bruit local $\mathcal{R}(p, \omega_k)$, mesuré dans la sous-bande où se trouve la composante, tel qu'il a été défini au paragraphe 2.1.1. Pour l'instant, nous avons plutôt utilisé le niveau relatif, mais il se trouve qu'ici c'est le rapport signal-à-bruit qui apparaît le plus naturellement. Dans la suite, nous utiliserons donc plutôt le RSB local $\mathcal{R}(p, \omega_k)$. La valeur moyenne du rapport de vraisemblance s'écrit donc

$$\mathrm{E}\left\{\Lambda_{(\log)}\right\} = T \times \frac{1}{2,5} \times \mathcal{R}(p, \omega_k) + 1 \tag{4.8}$$

Au paragraphe précédent, nous avons dit que la valeur du seuil $\mathcal{S}$ devait être choisie autour de 10 dB. Pour évaluer la limite de détection, il faut donc déterminer les conditions sous lesquelles la valeur du rapport de vraisemblance $\Lambda_{(\log)}$ est supérieure au seuil de 10 dB. A priori, la relation (4.8) ne nous est pas d'un grand secours car elle ne fournit que la valeur moyenne de $\Lambda_{(\log)}$. En fait, l'expression de $\Lambda_{(\log)}$ donnée par l'équation (4.6) indique que le rapport de vraisemblance correspond à un niveau relatif mesuré sur le périodogramme de l'observation $\mathbf{X}$. La densité de probabilité correspondante a donc déjà été étudiée au paragraphe 3.2.2. L'expression analytique de la densité de probabilité (3.18) permet de vérifier deux points importants :

- Lorsque la valeur moyenne $\mathrm{E}\left\{\Lambda_{(\log)}\right\}$ est supérieure à 10 dB, la probabilité que $\Lambda_{(\log)}$ soit supérieur à $\mathrm{E}\left\{\Lambda_{(\log)}\right\}$ est très proche de 50% [1].
- On peut garantir que $\Lambda_{(\log)}$ est supérieur à 10 dB (avec une probabilité d'erreur inférieure à 0,1%) dès que la valeur moyenne $\mathrm{E}\left\{\Lambda_{(\log)}\right\}$ est de l'ordre de 15 dB (cf. figure 3.18).

Ces constatations nous permettent de déterminer deux points particuliers de la courbe donnant la probabilité de détection

$$\begin{cases} 100\% & \text{pour} & \mathrm{E}\left\{\Lambda_{(\log)}\right\} = 15 \text{ dB} \\ 50\% & \text{''} & \mathrm{E}\left\{\Lambda_{(\log)}\right\} = 10 \text{ dB} \end{cases}$$

[1] Cette constatation est loin d'être triviale et elle devient fausse pour les valeurs moyennes du niveau relatif inférieures à 10 dB. On sait, par exemple, que dans le cas limite du bruit seul (niveau moyen de 0 dB), la probabilité de dépassement tombe à 40% (cf. figure 3.14).

En utilisant la relation (4.8), on obtient la probabilité de détection en fonction du rapport signal-à-bruit mesuré dans la sous-bande

$$\begin{cases} 100\% \quad \text{pour} \quad \mathcal{R}(p,\omega_k) = \dfrac{2,5 \times \{(15\text{ dB}) - 1\}}{T} \approx \dfrac{77}{T} \\[2ex] 50\% \qquad \text{''} \qquad \mathcal{R}(p,\omega_k) = \dfrac{2,5 \times \{(10\text{ dB}) - 1\}}{T} \approx \dfrac{22}{T} \end{cases}$$

Pour $T = 16$, les valeurs correspondantes sont 7 et 1 dB, pour $T = 32$, 4 et -2 dB, et enfin, pour $T = 64$, 1 et -5 dB. On rappelle que ces valeurs correspondent au pire des cas où la fréquence de la composante se situe entre deux points de discrétisation fréquentielle.

La stratégie de détection retenue permet d'arriver à une procédure assez efficace puisqu'il est possible de détecter des composantes de signal de très bas niveau (valeurs de RSB inférieures à 0 dB). Cette procédure présente en outre l'intérêt de permettre la détection dans de bonnes conditions à partir d'un nombre raisonnable de spectres à court-terme successifs. En pratique, le choix de $T = 32$ est un bon compromis car il permet de détecter les composantes de signal jusqu'à un rapport signal-à-bruit de -2 dB (avec une probabilité de détection de 50% dans le pire des cas), tout en conservant une durée totale d'observation acceptable. Par exemple, si les fenêtre de TFCT ont une durée de 20 ms avec recouvrement de 50%, la durée totale est de l'ordre de $32 \times 10 = 320$ ms[2]. Il est vrai que cette durée demeure assez importante ce qui peut poser un problème pour un fragment musical assez rapide. Dans un tel cas, il est toujours possible de diminuer la valeur de T mais les possibilités de détection diminuent assez vite : $T = 16$ fournit encore une amélioration assez nette par rapport à la "détection" sur le module de la TFCT, par contre, le choix de $T = 8$ est déjà quasiment inutile.

Un point intéressant est que malgré les approximations effectuées (hypothèse de bruit blanc, calcul du périodogramme par TFD) la procédure de détection obtenue est assez satisfaisante. Il est vrai que l'utilisation de *zero-padding* lors du calcul de la transformée de Fourier du signal de sous-bande est une modification pas trop coûteuse de cette procédure de base qui permettrait d'éviter les variations d'environ 4 dB des limites de détection selon la position de la composante dans la sous-bande. Il serait dans ce cas nécessaire de calculer les TFD au moins sur $4 \times T$ points pour avoir des variations du niveau du maximum négligeables (inférieures à 0,5 dB).

Une dernière remarque est qu'en pratique, l'estimation du niveau moyen de bruit dans chaque sous-bande ne peut en général pas être obtenue avec une précision supérieure à ± 2 dB (cf. paragraphe 1.4.3). Il est donc nécessaire de relever le seuil $\mathcal{S}$ de 2 ou 3 dB afin d'assurer que ces erreurs d'estimation ne se traduisent pas par de fausses détections. Il faut noter que ce rehaussement du seuil de détection peut aussi être utile dans le cas où le bruit d'enregistrement n'est pas tout à fait stationnaire (cf. paragraphe 1.4.5). Par conséquent, le seuil utilisé en pratique est en général de 12 dB. Dans ce cas, la probabilité de détection à 50% correspond à un RSB de 0 dB (élévation de 2 dB par rapport à $\mathcal{S} = 10$ dB). On pourrait vérifier que du fait de la diminution de la variance de $\Lambda_{(\log)}$ lorsque la valeur moyenne $\mathrm{E}\left\{\Lambda_{(\log)}\right\}$ augmente (cf. paragraphe 3.2.2), la probabilité de détection de 100% correspond quasiment à la même valeur que pour le seuil de 10 dB (c'est à dire à un RSB d'environ 4 dB).

Restriction aux niveaux faibles Il est très important de restreindre l'usage de la procédure de détection aux faibles valeurs du spectre à court-terme. En effet, l'hypothèse H0 (absence d'une composante) correspond à la présence du bruit seul dans la sous-bande. Or dans, un cas

[2] On rappelle que le choix d'un recouvrement de 50% est ici nécessaire pour garantir que le bruit présent dans les sous-bandes peut être assimilé à un bruit blanc.

réel, la sous-bande peut très bien contenir une partie de signal qui ne s'apparente pas à une composante sinusoïdale. Une procédure plus réaliste aurait donc consisté à tester l'hypothèse H1 (présence d'une composante) contre une hypothèse H0 du type "présence d'un signal autre qu'une composante sinusoïdale". Malheureusement la stratégie de décision associée à un tel test est très complexe, de plus les performances vis à vis du bruit sont forcément moins bonnes qu'avec l'alternative "bruit seul". La solution utilisée ici consiste à appliquer le test de détection, tel qu'il a été décrit précédemment, uniquement aux points du spectre à court-terme pour lesquels le RSB $\mathcal{R}(p, \omega_k)$ est faible.

Pour obtenir une idée du niveau à partir duquel le test n'est plus significatif notons que dans le cas où le signal présent dans la sous-bande est lui-même un bruit blanc, supposé indépendant du bruit de sous-bande, la valeur moyenne du rapport de l'équation (4.6) est donnée par $\mathcal{R}(p, \omega_k)+1$, quelle que soit la longueur T du bloc de signal de sous-bande. Par conséquent, les valeurs pour lesquelles le RSB $\mathcal{R}(p, \omega_k)$ est de l'ordre du seuil $\mathcal{S}$ choisi pour le test ont de fortes chances d'être incorrectement détectées *même si le signal présent dans la sous-bande s'apparente à un bruit blanc*. En pratique, le test n'est appliqué que pour les points du spectre à court-terme où le RSB, estimé sur tout le bloc de signal de sous-bande, est inférieur d'au moins 3 à 5 dB au seuil $\mathcal{S}$ de détection. Le test de détection présenté ici ne concerne donc que les composantes de très faible niveau qui correspondent à un rapport signal-à-bruit entre 0 dB (limite de détection) et 8 dB (limite des valeurs testées).

Nous reviendrons sur les conséquences de cette limitation au paragraphe 4.2.4.a. En particulier, nous pouvons déjà retenir que la procédure décrite ici est très efficace pour les faibles valeurs du spectre à court-terme, mais qu'elle ne peut pas être appliquée aux valeurs du spectre de niveau plus important. Nous verrons au paragraphe 4.2.4 que le test de détection qui vient d'être présenté ne correspond qu'à une partie de la procédure complète de détection utilisée dans le système de restauration. Toutefois, avant de justifier l'utilité des ajouts à la procédure de détection décrits au paragraphe 4.2.4, il est nécessaire de préciser la structure du système de débruitage, ce qui est l'objet du paragraphe suivant.

4.2.3 Traitement de débruitage

4.2.3.a Principe du traitement

Nous allons maintenant présenter la procédure globale de débruitage telle qu'elle a été proposée dans [Cappe 93a]. Celle-ci peut être résumée par les étapes suivantes :

- **(A)** Décomposition préalable du signal bruité $x(n)$ par une transformation à court-terme de faible résolution fréquentielle (notée $X(p, \omega_k)$). En pratique, c'est la transformée de Fourier à court-terme qui est utilisée avec des fenêtres de durée relativement brève (de l'ordre de 10 à 20 ms).

- **(B)** *Détection, dans chaque sous-bande, de la présence éventuelle d'une composante sinusoïdale du signal*, proche du niveau de bruit et stable sur une durée importante. La procédure de détection utilisée comprend à la fois le test décrit par la relation (4.6) (paragraphe 4.2.2.c) et les ajouts qui seront présentés au paragraphe 4.2.4.

- **(C)** Selon le résultat de cette détection, *la restauration du signal de sous-bande $X(p, \omega_k)$ s'effectue de deux manières différentes* :

Pas de composante détectée $\Rightarrow$ *De manière locale* : $Y(p, \omega_k)$ est obtenu à partir de $X(p, \omega_k)$ par application d'une atténuation.

Présence d'une composante $\Rightarrow$ *En considérant le signal de sous-bande en bloc* : $Y(p, \omega_k)$ est alors estimé à partir du vecteur complexe constitué de T valeurs successives du signal de sous-bande $[X(p-T_1, \omega_k), \ldots, X(p, \omega_k), X(p+1, \omega_k) \ldots, X(p+T_2, \omega_k)]^T$.

- **(D)** Le signal résultant $y(n)$ s'obtient par TFCT inverse à partir de $Y(p, \omega_k)$.

Pour justifier l'intérêt de ce traitement, il faut faire appel aux résultats du chapitre 3. L'utilisation d'une fenêtre de traitement relativement courte pour le calcul de la TFCT $X(p, \omega_k)$ nous assure que le résultat obtenu par atténuation spectrale à court-terme ne subit pas un lissage trop important lors des parties transitoires (paragraphe 3.1.2). Par contre, dans le cas de composantes sinusoïdales de signal de bas niveau, le résultat obtenu par atténuation spectrale directe est inacceptable (paragraphes 3.3.1 et 3.1.1).

Dans la procédure proposée ci-dessus, on commence par évaluer la nature du signal présent dans la sous-bande d'indice ω_k, au voisinage de l'indice temporel p. Dans le cas où cette sous-bande contient un signal autre qu'une composante sinusoïdale du signal de bas niveau, on applique une atténuation locale. La valeur du spectre à court-terme $Y(p, \omega_k)$ obtenue correspond donc exactement au résultat classique de l'atténuation spectrale à court-terme. Par contre, lorsqu'une composante de signal est détectée, on sait que le résultat obtenu par atténuation spectrale locale ne convient pas. Dans ce cas, l'idée retenue consiste à considérer plusieurs valeurs successives du signal de sous-bande. Nous avons vu au paragraphe précédent que le fait de considérer plusieurs valeurs de la TFCT permet d'améliorer les possibilités de détection, il en va exactement de même lorsqu'il s'agit de restaurer la composante détectée (voir le paragraphe 4.2.5).

La partie centrale du traitement est l'étape **(B)** de détection des composantes sinusoïdales stables. C'est cette étape qui fait l'intérêt de la technique puisqu'elle assure que la procédure de restauration utilisée est la mieux adaptée à la nature locale du signal. Si l'on exclut cette étape, aucune des deux techniques de traitement de la partie **(C)** ne présente d'intérêt particulier.

4.2.3.b Mise en œuvre

La mise en œuvre d'un tel traitement est a priori très coûteuse si la procédure de détection est appliquée pour toutes les valeurs du spectre à court-terme de faible niveau. De plus, ce coût de calcul n'est pas vraiment justifié car la procédure de test utilise, dans chaque sous-bande, un périodogramme du signal évalué sur une longueur T. Or, les spectres de deux blocs de signaux de sous-bande décalés d'un seul échantillon ($[X(p-T_1, \omega_k), \ldots, X(p+T_2, \omega_k)]^T$ et $[X(p-T_1+1, \omega_k), \ldots, X(p+T_2+1, \omega_k)]^T$) se "ressemblent" nécessairement. Dans ces conditions, les résultats du test de détection obtenus aux indices temporels p et $p+1$ sont fortement redondants.

Nous avons donc choisi de modifier le principe exposé précédemment *en considérant le signal de sous-bande par blocs successifs de T échantillons.* Pour éviter les "effets de bord" qui apparaissent si les blocs de signaux de sous-bandes sont disjoints, nous avons été conduit à utiliser des blocs de signaux de sous-bande qui se recouvrent. Afin d'obtenir une structure pas trop lourde nous avons utilisé un *recouvrement de $T/2$ échantillons entre les blocs successifs de signaux de sous-bande.* Par analogie avec le cas des modifications de la TFCT, on peut penser qu'un recouvrement de 50% est ici suffisant pour garantir la validité des traitements effectués sur les blocs de signaux de sous-bande.

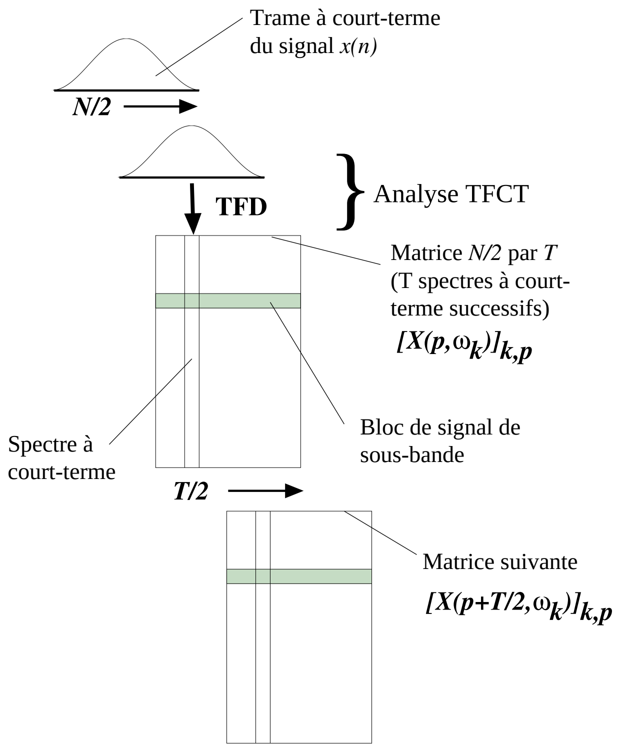

Figure 4.5: Structure du système de traitement **I** – **ANALYSE** : création des blocs de signaux de sous-bande.

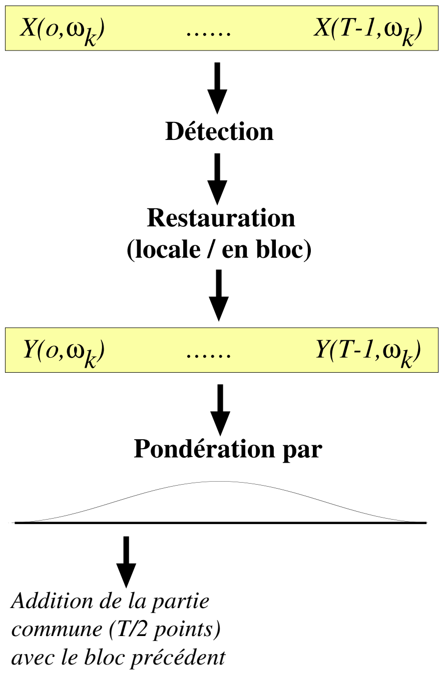

Figure 4.6: Structure du système de traitement **II** – **TRAITEMENT** : traitement de chaque bloc de signal de sous-bande.

La figure 4.5 permet de se faire une idée de la structure adoptée en ce qui concerne le début du traitement (partie analyse). En partant du haut de la figure, on reconnaît l'analyse usuelle par TFCT avec les trames à court-terme de longueur N qui se recouvrent à 50%. Par suite, les spectres à court-terme obtenus (de longueur $N/2$ car le signal $x(n)$ est réel) viennent remplir les colonnes d'une matrice $[X(p, \omega_k)]_{k,p}$. La matrice est complète lorsque T trames à court-termes ont été analysées. Les colonnes de cette matrice correspondent aux spectres à court-terme successifs, tandis que les lignes représentent les blocs de signaux de sous-bande de longueur T. La prochaine matrice de ce type, représentée en bas du schéma, contient les signaux de sous-bande décalés de $T/2$ échantillons, elle possède donc $T/2$ colonnes communes avec la matrice précédente.

Par la suite, pour chaque matrice obtenue, on considère séparément chaque bloc de signal de sous-bande, c'est à dire chaque ligne de la matrice. La suite des opérations effectuées pour chacun des blocs de signal de sous-bande est représentée sur la figure 4.6. La première étape est la procédure de détection de la présence éventuelle d'une composante stable du signal. La procédure complète de détection utilisée pour cette étape est schématisée par la figure 4.11 qui sera présentée plus loin (au paragraphe 4.2.4.c). Si une composante est détectée, il y restauration de tout le bloc de signal (estimation du vecteur $\mathbf{Y}$ à partir du vecteur $\mathbf{X}$). Dans le cas contraire, on applique une atténuation séparément pour chacun des indices temporel : $Y(p, \omega_k) = G(p, \omega_k)X(p, \omega_k)$ pour $p = 0, \ldots, T-1$. On rappelle que ce traitement local est strictement équivalent au traitement par atténuation spectrale à court-terme avec des fenêtres de longueur N. Le choix du mode de calcul du gain $G(p, \omega_k)$ se fait donc parmi les différentes règles de suppression évoquées au paragraphe 2.2, avec une préférence pour la règle d'Ephraim et Malah qui permet d'éviter le phénomène de bruit musical (cf. paragraphe 4.1.2). Enfin, le bloc de signal de sous-bande modifié est pondéré par une fenêtre douce (par exemple de Hann) de longueur T. Cette pondération après modification est nécessaire pour éviter les effets de bords. Ici c'est le traitement en bloc du signal de sous-bande qui peut générer des "discontinuités" entre les blocs successifs. Enfin, on effectue une addition des parties recouvrantes des blocs de signal de sous-bande correspondant à deux matrices modifiées successives.

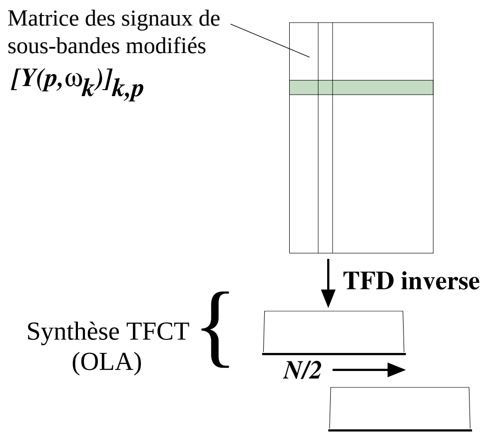

Figure 4.7: Structure du système de traitement **III – SYNTHESE** : synthèse du signal temporel restauré.

La figure 4.7 représente l'obtention du signal restauré à partir de la matrice contenant les signaux de sous-bandes modifiés. Les colonnes de la matrice correspondent aux spectres à court-terme modifiés, il suffit donc à partir de ces spectres d'effectuer la synthèse associée à la TFCT de la manière habituelle. Une remarque est que seuls $T/2$ spectres à court-termes peuvent être utilisés pour synthétiser le signal $y(n)$ puisque la seconde moitié de la matrice des signaux de sous-bande modifiés doit encore être additionnée avec la partie recouvrante de la matrice suivante.

L'ensemble de la structure de traitement décrite par ces trois figures correspond à un coût de calcul modéré. En effet, les parties analyse et synthèse représentées sur les figures 4.5 et 4.7 correspondent simplement à une analyse par TFCT classique avec des fenêtres de longueur N. La seule différence est qu'à l'analyse, au lieu de traiter chaque spectre à court-terme séparément, on les conserve dans une matrice jusqu'à en obtenir T, afin de disposer de blocs de sous-bande de longueur T (figure 4.5). De même à la synthèse, comme le traitement porte sur les blocs signaux de sous-bande de longueur T, on obtient directement une matrice qui contient T spectres à court-termes modifiés (figure 4.7). On réalise alors directement la synthèse par TFCT pour $T/2$ spectres à court-terme successifs (et non T, à cause du recouvrement entre les blocs de signaux de sous-bande). Pour les parties du système représentées sur les figures 4.5 et 4.7 le coût de calcul est exactement le même que pour l'analyse/synthèse par TFCT (fenêtres de longueur N, recouvrement 50%). La différence intervient en terme de stockage des données puisqu'il est maintenant nécessaire de conserver des matrices de T spectres à court-terme successifs (alors que la TFCT simple peut s'effectuer trame par trame).

4.2.4 Ajouts à la procédure de détection

Le dernier élément du système qui reste à préciser est la technique utilisée pour restaurer le signal dans le cas où une composante de signal est détectée (estimation en bloc du signal de sous-bande). Avant de justifier le choix d'une technique particulière, nous allons décrire deux modifications du principe de base de la procédure de détection qui ont été nécessaires pour remédier aux problèmes apparus lors des premiers tests du système de traitement.

4.2.4.a Détection pour les niveaux moyens

Lors des premiers essais du système tel qu'il vient d'être décrit, les résultats se sont avérés plutôt décevants. En particulier, le résultat du traitement présentait encore beaucoup de problèmes associés à la présence de composantes de bas niveau. En particulier, les effets de modulation (paragraphe 3.3.1), où de fluctuation de niveau (paragraphe 3.2.2) étaient encore très audibles. Une analyse plus poussée des résultats obtenus sur des signaux simples (composantes sinusoïdales seules), a permis de montrer que ces effets ne se produisaient pas pour les composantes qui avaient été détectées, mais au contraire pour celles de niveau relatif plus important qui n'étaient pas testées. Nous avons signalé au paragraphe 4.2.2.d que le test de détection ne peut pas être utilisé pour les points du spectre à court-terme de niveau trop important. Ceci pose un problème car nous avons vu au paragraphe 3.3.1 qu'avec une durée trame à court-terme de TFCT de l'ordre de 10 à 20 ms, l'effet de modulation peut être audible même pour des composantes de niveau important (20 dB ou plus). De même, l'effet de fluctuation est sensible jusqu'à des niveaux relatifs de l'ordre de 20 dB (cf. paragraphe 3.2.2). En conclusion, si on désire utiliser une TFCT sur des fenêtres courtes, il n'est pas possible de considérer que le test ne doit être effectué que pour les composantes de niveau faible (inférieur à 10 dB). Il faut trouver un test qui fonctionne pour les valeurs du spectre de niveau moyen (entre 10 et 30 dB).

Pour détecter la présence d'un signal sinusoïdal, nous avons intérêt à utiliser le même principe qu'au paragraphe 4.2.2, à savoir que la décision doit se faire à partir du périodogramme du bloc de signal de sous-bande. La situation est beaucoup plus simple que précédemment car on peut négliger l'influence du bruit si on se limite aux cas où le RSB est important. Dans ce cas, la tâche consiste simplement à déterminer si le périodogramme du signal de sous-bande présente ou non un pic marqué.

Le seul exemple dont nous disposions est la *Spectral Flatness Mesure* (ou SFM) qui est utilisée dans [Johnston 88] et [Mourjopoulos 92] pour déterminer si l'allure globale d'un spectre de puissance est plutôt celle d'un spectre de raies ou bien celle d'un bruit (le but recherché dans ces publications étant d'ajuster les courbes d'effets de masque selon la nature du signal). La SFM est définie comme le rapport entre la moyenne géométrique du spectre de puissance et la moyenne arithmétique de ce spectre, le tout exprimé en décibels [Johnston 88]. Si $S_{\mathbf{X}}(r)$ désigne le module au carré de la TFD du bloc de signal de sous-bande $\mathbf{X}$ défini par

$$S_{\mathbf{X}}(r) = \left| \sum_{p=0}^{T-1} \mathbf{X}[p] \mathrm{e}^{-j2\pi \frac{rp}{T}} \right|^2$$

La SFM est donc calculée par la relation

$$\mathrm{SFM} = \frac{\exp\left[\frac{1}{T} \sum_{r=0}^{T-1} \log\left\{ S_{\mathbf{X}}(r) \right\} \right]}{\frac{1}{T} \sum_{r=0}^{T-1} S_{\mathbf{X}}(r)}$$

Cette mesure permet effectivement de trouver une valeur proche de 0 dB lorsqu'elle est appliquée au spectre d'un bruit, et une valeur très faible pour le spectre d'un signal sinusoïdal. Toutefois, les résultats obtenus dans ce dernier cas dépendent beaucoup du nombre de points du spectre et de la fenêtre de pondération utilisée.

L'idée de la SFM semble assez astucieuse, mais les résultats sont plutôt difficiles à évaluer à cause de l'opération non-linéaire (moyenne géométrique). Nous avons cherché à obtenir un indicateur dont le comportement soit plus simple à prédire. D'après l'équation (4.7), en cas de présence d'une seule composante sinusoïdale de signal, le pic du périodogramme du bloc de signal de sous-bande a pour valeur moyenne (dans le pire des cas où la fréquence est située entre deux points de discrétisation)

$$T^2 \times 0,4 \times |A|^2 + Tv \tag{4.9}$$

La contribution moyenne du bruit Tv peut être négligée puisqu'on s'intéresse ici au cas de RSB importants (10 dB ou plus). Le terme $T\,|A|^2$ représente l'énergie du bloc de signal de sous-bande (puisqu'on néglige le bruit). En utilisant la relation de Parseval, cette énergie du bloc de signal de sous-bande peut aussi s'écrire sous la forme

$$\frac{1}{T} \sum_{r=0}^{T-1} S_{\mathbf{X}}(r)$$

La relation 4.9, indique donc que dans le cas d'un signal sinusoïdal, la valeur maximale du module du spectre au carré est supérieure à

$$0,4 \times \sum_{r=0}^{T-1} S_{\mathbf{X}}(r)$$

Ce qui s'interprète en disant qu'au moins 40% de l'énergie du bloc de signal de sous-bande est concentrée en un seul point quand le signal de sous-bande est une exponentielle complexe. On définit donc une mesure de concentration d'énergie par

$$\mathrm{MCE} = \frac{\max\limits_{r} \left\{ S_{\mathbf{X}}(r) \right\}}{\sum_{r=0}^{T-1} S_{\mathbf{X}}(r)} \tag{4.10}$$

La présence d'une composante sinusoïdale dans la sous-bande est détectée lorsque la valeur de MCE obtenue est supérieure à 0,4.

Les deux procédures de détection qui viennent d'être présentées n'ont pas de fondement théorique très solide, nous nous sommes donc contenté de vérifier que leur comportement était satisfaisant dans trois cas typiques de signaux de sous-bande correspondant à la situation qui nous intéresse :

1. Bruit blanc
2. Exponentielle complexe avec un RSB de 6 dB
3. Deux exponentielles complexes avec un RSB de 10 dB (le RSB total dans la sous-bande est dans ce cas de 13 dB)

Chaque situation a été simulée dix mille fois, pour chacun des deux critères (SFM et MCE). Les performances des deux critères sont assez comparables en ce qui concerne les situations (1) et (2) avec un avantage tout de même pour la SFM qui permet de mieux séparer un spectre correspondant à une exponentielle légèrement bruitée de celui correspondant à un bruit blanc. On trouve que les deux conditions

$$\text{SFM} < \text{-5 dB} \qquad \underline{\text{ou}} \qquad \text{MCE} > 0.35$$

peuvent être utilisées indifféremment pour signaler la présence d'une composante sinusoïdale dans la sous-bande, à condition de se limiter au cas des sous-bandes où le RSB est supérieur à 6 dB. En dessous de cette limite les résultats se dégradent fortement pour les deux techniques. De plus, il faut signaler que ces résultats correspondent au cas où la longueur T du bloc de signal de sous-bande vaut 32. Pour des longueurs inférieures (par exemple $T = 16$), les résultats obtenus sont très mauvais avec les deux techniques. La différence entre les deux critères se manifeste pour l'exemple (3) où la sous-bande contient deux composantes de signal : le critère SFM (appliqué avec le seuil de -5 dB) place de manière certaine le spectre obtenu dans la même catégorie que le spectre d'une seule sinusoïde alors que le critère MCE fournit des résultats proche de ceux obtenus pour le spectre du bruit.

Le choix entre ces deux critères dépend donc de l'objectif recherché : si on cherche à garantir qu'un signal de sous-bande où une détection intervient ne contient pas plus d'une composante de signal, c'est le critère MCE qui convient. A priori, le choix du critère SFM est meilleur car même si la sous-bande contient plusieurs composantes stables du signal il vaut mieux qu'une détection intervienne car la restauration sera alors plus efficace. Nous verrons que ce problème de la composition du signal de sous-bande en cas de détection est lié au choix de la technique de restauration utilisée pour les blocs de signaux de sous-bande. Pour l'instant on peut simplement retenir que le critère MCE est beaucoup plus restrictif que le critère SFM. En particulier, pour des cas où le signal de sous-bande est plus complexe (plusieurs composantes de signal, composante présente seulement sur une partie du bloc), le résultat de détection est en général positif avec le critère SFM et négatif avec MCE. Un dernier point important est que les deux critères sont assez sensibles à la position exacte de la composante à détecter : la détection est facilitée si la fréquence de la composante correspond à un point de discrétisation fréquentielle. Il est vrai que la faible durée du bloc de signal de sous-bande ($T = 32$) et l'absence de fenêtre d'analyse ne simplifient pas la tâche. Là encore le *zero-padding* du bloc de signal de sous-bande avant calcul de la TFD peut permettre de réduire ces variations.

La procédure complète de détection comporte donc une première étape d'estimation du rapport signal-à-bruit à partir du bloc de signal de sous-bande :

$$\hat{\mathcal{R}}(\omega_k) = \frac{\sum_{p=0}^{T-1} |X(p, \omega_k)|^2}{T\hat{P}_d(\omega_k)} - 1$$

Par la suite, on distingue trois cas selon la valeur obtenue : celui des **niveaux faibles** ($\hat{\mathcal{R}}(\omega_k) <$ 8 dB) où la procédure de détection appliquée est celle qui a été décrite au paragraphe 4.2.2, ensuite vient le cas des **niveaux moyens** (8 dB < $\hat{\mathcal{R}}(\omega_k)$ < 25 dB) où l'on utilise l'un des deux critères SFM ou MCE, et enfin celui des **niveaux forts** ($\hat{\mathcal{R}}(\omega_k)$ > 25 dB) pour lequel le test n'est pas effectué. Ce dernier cas correspond à des valeurs importantes du spectre à court-terme du signal pour lesquelles la restauration par atténuation spectrale est dans tous les cas satisfaisante.

4.2.4.b Retard de détection

Un deuxième défaut qui est apparu lors des premiers essais de la technique était la présence de pré-échos audibles dans le signal restauré. Ce phénomène se manifestait de manière très nette lors de l'apparition brusque de composantes de signal. Une étude plus précise de cas simples a permis de montrer que ce phénomène était effectivement lié à la procédure de détection des composantes. Plus précisément, on constate que la procédure de détection (aussi bien la partie destinée aux niveaux faibles du paragraphe 4.2.2, que celle destinée aux niveaux moyen du paragraphe 4.2.4.a) est susceptible de détecter une composante sinusoïdale même si celle-ci n'est pas présente dans tout le bloc de signal de sous-bande. Par suite, lors de la restauration en bloc du bloc de signal de sous-bande, il se produit un effet d'étalement temporel.

Pour comprendre ce qui passe dans un tel cas, il suffit de remarquer que le module du spectre d'une composante sinusoïdale qui apparaît au milieu du bloc de signal est très proche de celui d'une composante qui est présente sur tout le bloc (voir la figure 3.4 au paragraphe 3.1.2). L'étude effectué au paragraphe 4.2.2.d montre que pour le test réservé aux niveaux faibles, la seule différence entre ces deux cas est donc un rehaussement du seuil de détection de 3 dB (pour le cas de la composante qui apparaît au milieu du bloc) lié à l'abaissement de 3 dB du niveau du pic spectral (cf. figure 3.4). Par conséquent, une composante qui apparaît en milieu de bloc dans le signal de sous-bande est détectée quasiment aussi bien que si elle était présente sur tout le bloc (avec tout de même une élévation du seuil de détection). La situation est assez semblable pour les procédure de détection "niveau moyen" du paragraphe 4.2.4.a bien qu'il soit difficile d'évaluer précisément la limite dans le cas de la SFM.

Pour remédier à ce problème, nous avons choisi de retarder la détection d'une composante de signal. Plus précisément, lorsque qu'une composante est détectée dans la voie de pulsation centrale ω_k, le résultat de détection est confirmé uniquement si une composante a été détectée dans la même voie au bloc précédent. L'effet de cette confirmation conditionnelle de la détection est d'éliminer systématiquement la première détection d'un groupe de détections successives correspondant à la même voie. Ce qui revient à retarder la décision de détection de $T/2$ spectres à court-terme dans le cas de l'apparition d'une composante de signal.

En fait, il est apparu que cette procédure éliminait aussi certaines détections correspondant à des composantes stables du signal. En effet, les voies de la TFCT ne sont pas totalement séparées car le filtre d'analyse $h(n)$ possède une bande passante plus large que la distance entre les pulsations centrales de deux sous-bandes voisines. Par conséquent, une composante de signal est forcément présente avec un niveau significatif dans plusieurs sous-bandes voisines. Pour tenir compte de cet aspect, nous avons pris en compte les sous-bandes voisines lors de la confirmation des détections. C'est à dire qu'une détection dans la voie de pulsation centrale ω_k est validée dès qu'une composante a été détectée, au bloc précédent, dans les voies de pulsation centrale ω_{k-1}, ω_k ou ω_{k+1}.

Les figures 4.8, 4.9 et 4.10 illustrent le fonctionnement de la procédure de détection pour le cas d'un signal composé d'un son de piano bruité. La figure 4.8 indique tous les points du spectre à court-terme pour lesquels le module au carré $|X(p, \omega_k|^2$ est supérieur d'au moins 6 dB

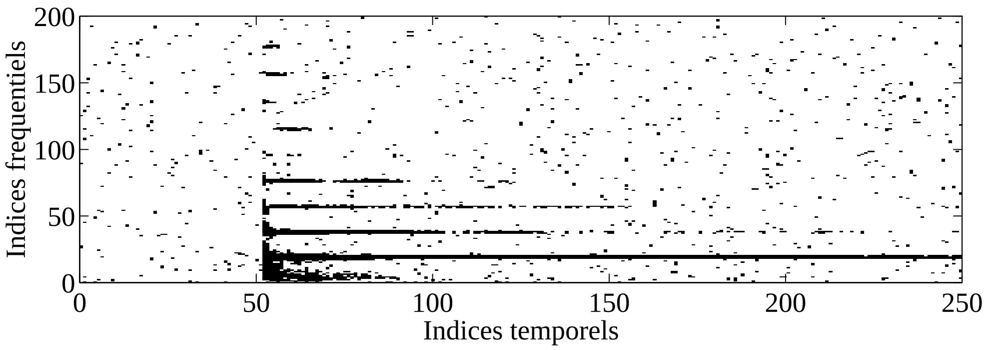

Figure 4.8: TFCT d'un signal correspondant à son de de piano isolé (note A5) dans du bruit blanc : valeurs pour lesquelles le niveau relatif $\mathcal{Q}(p, \omega_k)$ est supérieur à 6 dB. Fenêtre d'analyse de 1024 points (soit 21 ms). Seuls les 200 premiers indices fréquentiels sont représentés (bande de 0 à 9,4 kHz).

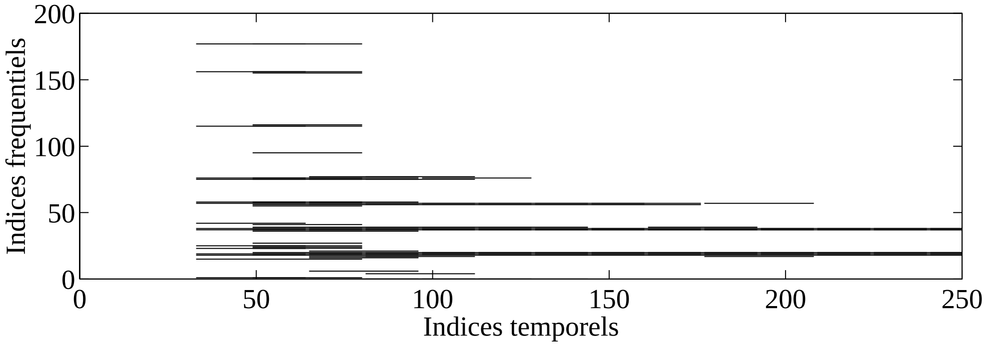

Figure 4.9: Résultat de la détection : blocs de signaux de sous-bandes où la présence d'une composante a été détectée. Longueur des blocs de signal de sous-bande $T = 32$.

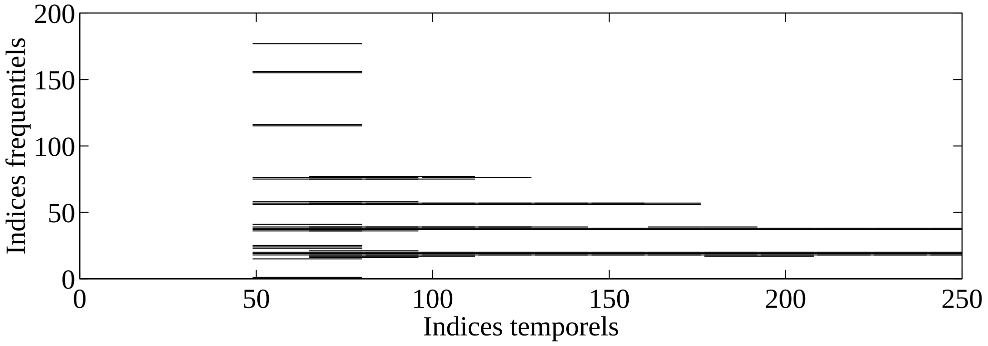

Figure 4.10: Résultat après confirmation conditionnelle des détections.

au niveau de bruit. L'indice temporel p de la trame à court-terme est représenté en abscisse (de 0 à 250) tandis que les 200 premiers indices fréquentiels ω_k sont représentés en ordonné. Les points isolés qui apparaissent sur cette figure sont dus à la variance du spectre de puissance à court-terme (d'après la figure 3.14, ces points correspondent à environ 2% du nombre total d'indices (p, ω_k) où seul le bruit est présent). On distingue sur la figure 4.8 l'attaque du son de piano, autour de l'indice temporel $p = 52$, les harmoniques du signal (très nettement les quatre premiers, plus faiblement ceux de rang 6, 8 et 9, et à peine les cinquième et septième). Certains de ces harmoniques disparaissent beaucoup plus rapidement que d'autres, mais comme il n'y a pas d'indication d'amplitude sur cette représentation il est impossible de faire la part entre le taux de décroissance et l'amplitude initiale des harmoniques. La figure 4.9 présente le résultat de la procédure complète de détection. Ici le critère de concentration d'énergie (MCE), utilisé pour les niveaux moyens, a été aussi appliqué dans le cas des niveaux forts (RSB > 25 dB) afin de permettre la comparaison avec la figure 4.8. Le mode de représentation n'est pas le même que sur la figure 4.8 car le test de présence d'une composante n'est effectué qu'une fois tous les $T/2 = 16$ indices temporels (cf. paragraphe 4.2.3.b). Pour indiquer ce "sous-échantillonnage" des indices temporels, une détection est représentée sur la figure 4.9 par un trait reliant les indices temporels extrêmes du bloc de signal de sous-bande considéré (donc un trait de longueur T). Par exemple, le trait représenté à l'indice fréquentiel 95 (qui correspond à la fréquence du cinquième harmonique du signal) sur la figure 4.9 ne correspond qu'à un seul résultat de détection positif. La figure 4.10 présente les résultats de détection qui ont été confirmés après application de la procédure décrite précédemment.

La comparaison entre les figures 4.9 et 4.10 confirme le fait que la procédure de confirmation des résultats de détection a essentiellement pour effet de retarder (de $T/2$ indices temporels) la détection dans le cas de l'apparition d'une composante. Il est vrai que dans le cas de composantes qui disparaissent rapidement, cette procédure peut aussi conduire à éliminer une détection "valide". Ici c'est le cas pour le cinquième harmonique de signal, qui est détecté sur la figure 4.9, et qui disparaît de la figure 4.10. Toutefois, le résultat obtenu sur la figure 4.10 reste très satisfaisant puisque la procédure de détection a permis de mettre en évidence tous les harmoniques du signal visibles sur la figure 4.8, tout en éliminant les effets liés à la présence de bruit. De plus on constate sur la figure 4.10 que les composantes de signal sont détectées beaucoup plus longtemps que sur la figure 4.8, ce qui est cohérent avec les résultats du paragraphe 4.2.2.d concernant la limite de détection. Cet effet est particulièrement visible sur la figure 4.10 pour les harmoniques du signal qui décroissent lentement (surtout pour le second).

Enfin, il est important de noter que même sur la figure 4.10 la détection de la présence de signal se fait de manière légèrement anticipée. En effet, les composantes de signal sont détectées à partir du quatrième bloc de signal de sous-bande qui débute à l'indice temporel $p = 3 \times T/2 = 48$. Il n'y a aucune raison pour que l'attaque du signal soit synchronisée avec le début d'un bloc de signal de sous-bande : ici celle-ci se produit légèrement plus tard autour de l'indice temporel $p = 52$. Pour éliminer totalement cette anticipation, il faudrait retarder la détection d'une durée équivalente à un bloc complet de signal de sous-bande, c'est à dire appliquer de nouveau la procédure de confirmation sur le résultat de la figure 4.10 (on rappelle que les blocs successifs de signaux de sous-bande se recouvrent à 50%). Pour cet exemple, ceci reviendrait à éliminer complètement les détections correspondant aux harmoniques de rangs 6, 8 et 9 qui ne sont détectés que pour un bloc de signal de sous-bande. Dans la pratique, nous avons constaté qu'une seule application de la procédure de confirmation des détections est suffisante pour éliminer la sensation auditive de pré-échos, même dans des cas particulièrement sensibles comme celui de la figure 4.8 (attaque d'un signal isolé). Ceci n'est pas surprenant compte tenu du fait que dans toute la première moitié d'un bloc de signal de sous-bande, le résultat obtenu après traitement est pondéré avec celui du bloc précédent. Dans le cas d'une première détection pour une sous-bande donnée, le lissage dû à la restauration en bloc du signal est donc compensé, dans la première moitié du bloc, par le résultat obtenu dans le bloc précédent.

4.2.4.c Procédure de détection complète

En tenant compte des deux modifications de la procédure de détection présentées dans ce paragraphe (test pour les niveaux moyens et retard de détection), la partie détection du système de débruitage peut être représentée par le schéma de la figure 4.11.

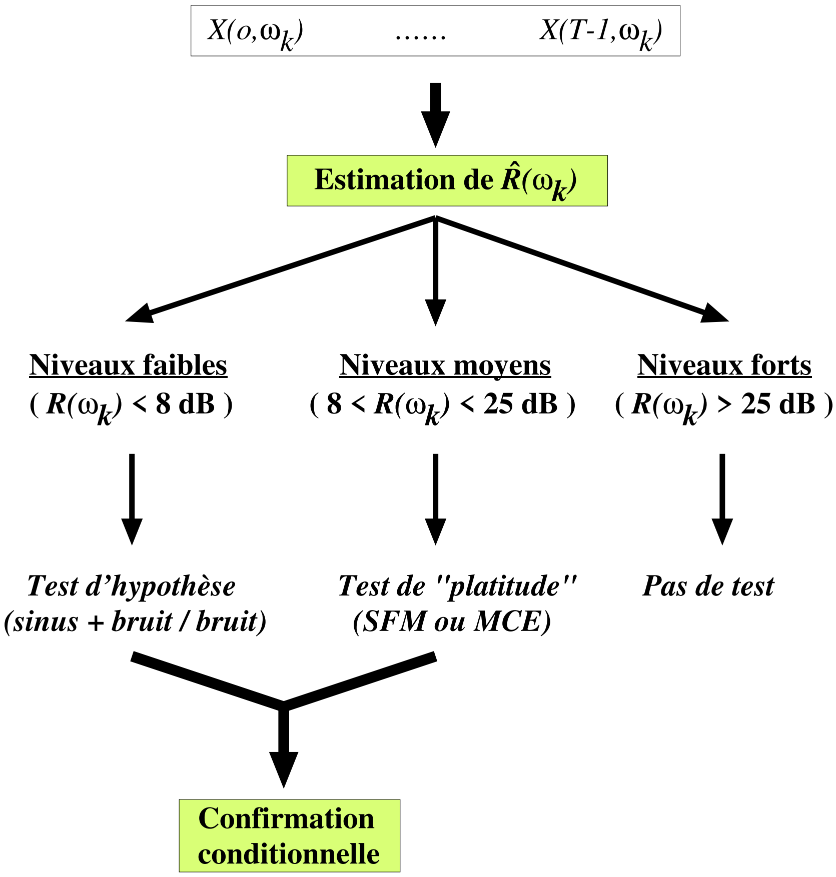

Figure 4.11: Procédure de détection complète.

Sur cette figure, la première étape de la détection consiste à estimer le rapport signal-à-bruit $\hat{\mathcal{R}}(p, \omega_k)$ sur le bloc de signal de sous-bande. Selon la valeur de $\hat{\mathcal{R}}(p, \omega_k)$ on distingue trois cas : les **niveaux faibles** pour lesquels la procédure de détection appliquée est celle qui a été décrite au paragraphe 4.2.2 ; les **niveaux moyens** pour lesquels la détection s'effectue avec une des techniques présentées au paragraphe 4.2.4.a ; les **niveaux forts** pour lesquels aucune procédure de détection n'est employée. On considère en effet que dans le cas des niveaux forts, le traitement direct (atténuation indépendante de chacun des points du bloc de signal de sous-bande) donne de toute façon des résultats satisfaisants. Il est donc inutile de mettre en œuvre une procédure de détection dans ce cas. Enfin, dans une troisième étape, la procédure de confirmation conditionnelle (paragraphe 4.2.4.b) est appliquée dans les deux premiers cas (niveaux faibles et moyens). Notons d'ailleurs qu'à ce point du schéma la distinction entre niveaux faibles et moyens n'est plus effectuée : pour confirmer une détection, tous les résultats de détection des blocs de signaux de sous-bande précédents sont considérés, qu'ils correspondent à des niveaux moyens ou forts.

4.2.5 Restauration en bloc du signal de sous-bande

Ce dernier paragraphe est consacré à un aspect important de la technique proposée qui est la méthode de restauration en bloc du signal de sous-bande, utilisée en cas de détection. Cette partie introduit une approche nouvelle dans le document qui est l'utilisation d'une technique paramétrique dans le cadre de la restauration. Nous avons indiqué au paragraphe 1.3.2 pourquoi une technique de restauration reposant sur un modèle du signal complet ne nous paraissait pas envisageable. Ici, la situation est différente car le signal présent dans la sous-bande en cas de détection est a priori très simple. C'est pourquoi on envisage l'utilisation de techniques de restauration paramétriques. Cependant, il s'est avéré que les résultats les plus satisfaisants sont en général obtenus avec une technique non-paramétrique qui s'inspire directement de l'atténuation spectrale à court-terme (c'est le principe qui a été présenté dans [Cappe 93a]). Nous reviendrons dans une seconde partie sur les raisons qui permettent d'expliquer cette constatation.

4.2.5.a Restauration paramétrique

Dans le cas où une composante de signal a été détectée, l'étude menée pour la mise au point du test de détection au paragraphe 4.2.2 suggère directement le principe de la technique de restauration à utiliser. En effet, si on utilise ici les mêmes hypothèses qu'au paragraphe 4.2.2.c (au plus une composante, bruit blanc), la restauration du signal de sous-bande en cas de détection se réduit à un problème d'estimation des paramètres d'une exponentielle complexe noyée dans un bruit blanc. C'est un problème très classique, sa caractéristique principale est qu'on ne dispose pas d'une statistique exhaustive pour la détermination des paramètres complets $\theta = [A\ \phi]$, ou même du paramètre de fréquence seul (ϕ). Par conséquent, on ne sait pas construire simplement un estimateur efficace de ces paramètres [Kay 93, §5, Ex. 5.9]. La démarche adoptée en général consiste donc à estimer tout d'abord le paramètre de fréquence ϕ, puis à estimer le paramètre d'amplitude (complexe) par la relation de démodulation

$$\hat{A} = \frac{1}{T} \sum_{p=0}^{T-1} \mathbf{X}[p] \mathrm{e}^{-\mathrm{j}\phi p} \tag{4.11}$$

On montre que $\hat{A}$ obtenu par la relation ci-dessus est l'estimateur linéaire à variance minimale de l'amplitude complexe, lorsque ϕ est supposée connue [Kay 88, §11.3]. Pour estimer la pulsation ϕ il est possible d'utiliser l'estimateur du maximum de vraisemblance [Kay 93, Ex. 15.13] qui présente l'intérêt d'être asymptotiquement (lorsque T est grand) efficace [Kay 93, §7]. Nous avons déjà vu au paragraphe 4.2.2.c, à propos de l'équation (4.5), que le maximum de vraisemblance est donné par la pulsation ϕ pour laquelle le périodogramme de l'observation est maximal.

La démarche la plus cohérente pour restaurer le signal de sous-bande en cas de détection est donc d'utiliser l'estimation au sens du maximum de vraisemblance pour obtenir les paramètres de la sinusoïde. L'estimation de la fréquence est en fait fournie directement par la procédure de détection de la composante effectuée au préalable (maximum du périodogramme). L'amplitude de la composante s'obtient par l'équation (4.11), qui correspond d'ailleurs à la valeur de la transformée de Fourier de l'observation à la pulsation ϕ. Il est indispensable de calculer la TFD du signal sur un grand nombre de points par *zero-padding* pour obtenir une précision suffisante dans la détermination des paramètres.

Malheureusement, cette idée assez élégante fournit des résultats très décevants dans des situations réelles. La raison de cet échec a déjà été évoquée au paragraphe 4.2.4.a : la procédure de détection ne garantit absolument pas que le signal présent dans la sous-bande se réduit à une exponentielle complexe. Nous avons indiqué au paragraphe 4.2.4.a que dans le cas des niveaux moyens du spectre, le choix d'une procédure appropriée (en l'occurrence la concentration d'énergie MCE) permet de se rapprocher de l'hypothèse d'une unique composante de signal. Par contre, pour les niveaux faibles, il est très difficile de préciser la nature du signal. Le raisonnement du paragraphe 4.2.2.d montre, par exemple, que si le signal de sous-bande est constitué de deux exponentielles complexes, le seuil de détection ne s'abaisse au pire que de 3 dB. On ne peut donc absolument pas exclure le cas où le signal est formé de plusieurs composantes. Pour l'exemple de la figure 4.8, on sait que les harmoniques du son de piano présentent des phénomènes de battements (d'une fréquence de un à quelques Hertz, selon le rang de l'harmonique, et la qualité de l'accord [Benade 76, §17.3]). Or, un bloc de signal de sous-bande correspond à une durée de plusieurs centaines de millisecondes ($T/2 \times NF_e$), c'est à dire de l'ordre de la période des battements. Par conséquent, le signal présent dans les sous-bande qui correspondent aux harmoniques du signal ne peut pas se réduire à une simple exponentielle complexe, ce qui n'empêche pas la détection (cf. figure 4.10).

L'hypothèse de présence d'une composante au plus dans chaque sous-bande ne pose pas de problème tant qu'il s'agit de concevoir une procédure de détection efficace. Par contre, il est naïf de penser que cette hypothèse correspond suffisamment à la réalité pour permettre une restauration de bonne qualité.

Une idée naturelle consiste à étendre le principe de cette restauration paramétrique en utilisant une technique de modélisation qui permette de déterminer simultanément les paramètres de plusieurs composantes de signal dans la sous-bande. Lorsqu'on cherche à estimer les paramètres de plusieurs exponentielles complexes, l'estimateur du maximum de vraisemblance devient très difficile à déterminer [Kay 88, §13]. En l'absence de solution “optimale” pour ce problème, un grand nombre de techniques différentes ont été proposées [Scharf 91, §11].

Nous avons fait plusieurs essais en utilisant une technique de modélisation de type ESPRIT [Hua 90]. Cette technique a été retenue car elle présente de bonnes performances vis à vis du bruit [Hua 90]. De plus, elle fournit directement les paramètres recherchés sans nécessiter une procédure supplémentaire de tri des valeurs obtenues [Scharf 91, §11]. Comme précédemment la technique de modélisation ne fournit que les fréquences (ou fréquences et taux d'amortissements) des exponentielles complexes, l'amplitude des composantes est estimée dans une seconde étape par l'équivalent de la relation 4.11 dans le cas de plusieurs composantes [Laroche 93b]. Les résultats obtenus en supposant a priori la présence de deux à quatre composantes de signal sont meilleurs que ceux obtenus précédemment. Une grande partie des problèmes liés à la présence de battements dans le signal musical sont résolus. Toutefois, certains effets peu naturels continuent à se manifester. Tout d'abord, pour des enregistrements complexes (signal musical polyphonique, variations temporelles rapides) il apparaît ponctuellement des distorsions importantes du signal qui se manifestent par des résonances parasites. De plus, le signal restauré présente souvent un effet de grésillement audible, similaire à celui qui se produit avec l'atténuation spectrale à court-terme lorsque le recouvrement entre trames à court-terme est insuffisant (cf. paragraphe 3.1.1.a), localisé autour des transitoires du signal. Ce dernier phénomène semble indiquer que dans le cas de signaux variant rapidement, et en présence de bruit, le recouvrement de 50% entre les blocs de signaux de sous-bande est insuffisant.

D'une manière plus générale, on peut remarquer que l'utilisation de ce type de techniques dans le cadre de la restauration pose un problème important. En effet, il s'agit en fait de méthodes de modélisation et non de restauration. En particulier, la présence de bruit n'est pas explicitement prise en compte (autrement que par l'hypothèse d'un bruit de mesure blanc). La conséquence est que si on applique la technique qui vient d'être décrite dans les sous-bandes correspondant à l'analyse d'un signal non-bruité, on obtient, pour la plupart des signaux musicaux, une distorsion nettement audible du signal. Le moins que l'on puisse attendre d'une méthode de restauration est qu'en l'absence de bruit le signal restauré soit identique au signal original.

La conclusion de ces différents essais est que malgré le tri effectué par la procédure de détection, le signal présent dans une sous-bande en cas de détection reste trop compliqué pour pouvoir se prêter facilement aux contraintes du modèle sinusoïdal. De plus, il est nécessaire de prendre en compte plus explicitement la présence de bruit lors la restauration du signal de sous-bande, et en particulier, d'exploiter le fait que la puissance du bruit présent dans la sous-bande est connue.

4.2.5.b Justification de l'efficacité de la technique non-paramétrique

La solution retenue pour la restauration en bloc du signal de sous-bande s'inspire directement de la figure 4.4 qui indique qu'en cas de détection le spectre du bloc de signal de sous-bande contient un pic qui dépasse d'au moins une dizaine de décibels (la valeur du seuil $\mathcal{S}$) la valeur moyenne du spectre du bruit. La partie qui correspond au signal est donc mise en évidence sur le spectre du bloc de signal de sous-bande. Dans ces conditions, on sait que l'application d'une atténuation spectrale calculée par une règle de suppression standard permet d'éliminer efficacement le bruit.

La procédure de restauration en bloc retenue consiste donc à appliquer une atténuation spectrale sur le spectre du bloc de signal de sous-bande puis à synthétiser le bloc de signal $\mathbf{Y}$ par TFD inverse. L'atténuation spectrale à apporter aux T points du spectre de $\mathbf{X}$ est calculée par exemple par la règle de soustraction en puissance dont le gain est donné par la relation (2.15) qui s'écrit ici

$$G(r) = \sqrt{1 - \frac{T\hat{P}_d(\omega_k)}{S_{\mathbf{X}}(r)}} \tag{4.12}$$

Le terme $T\hat{P}_d(\omega_k)$ correspond à la valeur moyenne du module au carré de la TFD sur T points du bruit présent dans la sous-bande de pulsation centrale ω_k. Cette valeur est la même pour tous les indices r du spectre car le bruit de sous-bande est supposé blanc (cf. paragraphe 4.2.2.d). La notation $S_{\mathbf{X}}(r)$ désigne toujours le carré du module du spectre du signal de sous-bande. Avec ces notations, la formule (4.12) correspond bien au gain de la soustraction en puissance tel qu'il a été vu au paragraphe 2.2.2. Ici on peut même se permettre de surestimer assez fortement le niveau de bruit de fond puisqu'on sait qu'en cas de détection le spectre du signal se compose (idéalement) d'un pic qui dépasse largement le niveau de bruit. Un facteur de surestimation du niveau de bruit ($T\hat{P}_d(\omega_k)$) de l'ordre de 8 dB est utilisé en pratique afin d'éviter les problèmes de bruit musical (cf. paragraphe 3.2.1).

On remarque que si le résultat de la détection est toujours positif dans une sous-bande donnée, le signal présent dans la sous-bande est traité par atténuation spectrale à court-terme avec des fenêtres de longueur T. Les seules différences par rapport au traitement tel qu'il a été étudié au chapitre 3 sont l'application de la fenêtre "douce" à la synthèse (après traitement des blocs de signal de sous-bande), et le fait que le signal de sous-bande traité est un signal complexe. Tous les résultats du paragraphe 3 restent donc valables *pour le signal de sous-bande* dans le cas où une composante est constamment détectée. Il faut toutefois se souvenir que les signaux de sous-bande sont décimés d'un facteur R par rapport au signal traité (annexe B). Par exemple, pour évaluer la sélectivité fréquentielle du traitement dans le cas où une composante est détectée, on peut utiliser la relation (3.20) avec des fenêtres de longueur T en n'oubliant pas de compresser l'axe des fréquences d'un facteur R. Cette interprétation montre qu'en négligeant le recouvrement entre les voies de la TFCT, on peut considérer que les composantes qui sont détectées sont traitées comme elles le seraient par atténuation spectrale à court-terme directe, avec des fenêtres de longueur $T \times R$ (soit $TN/2$ car le recouvrement est ici de 50%). A l'opposé, nous avons souligné au paragraphe 4.2.3.b que pour les voies "non détectée", le traitement effectué correspond à l'atténuation spectrale à court-terme avec des fenêtres de longueur N.

Par comparaison avec les techniques de modélisation évoquées au paragraphe précédent, la restauration des blocs de signaux de sous-bande par atténuation spectrale présente deux avantages :

- Les résultats obtenus ne présentent pas d'artefacts liés au traitement en bloc.
- Dans les cas où les techniques de modélisation donnent des résultats satisfaisants, les résultats obtenus par modélisation ou par atténuation spectrale sont en général indiscernables.

Le premier point est simplement dû au caractère non-paramétrique de la méthode de restauration : comme on ne fait pas d'hypothèse contraignante sur le signal à restaurer, les résultats restent plus "uniformes". De plus, il s'agit maintenant d'une méthode de restauration qui prend en compte la présence de bruit : comme les valeurs proches du niveau moyen de bruit sont fortement atténuées, il est par exemple impossible d'obtenir des composantes "parasites" dues à la présence de bruit.

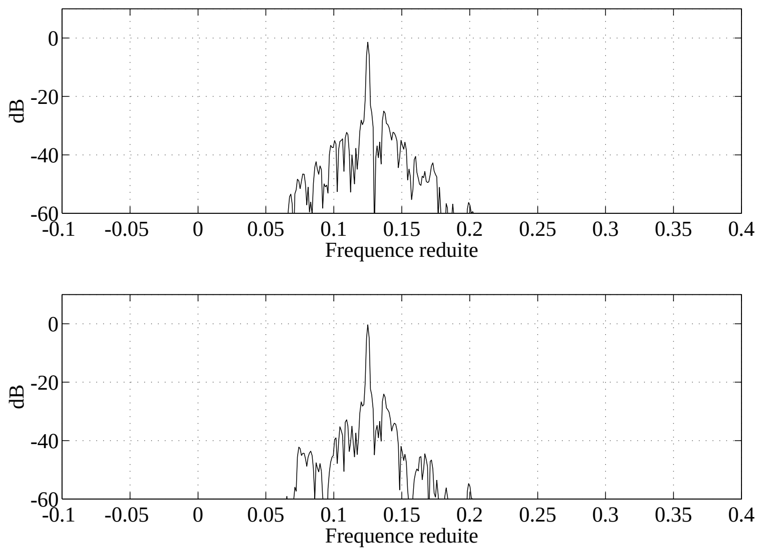

Figure 4.12: Spectre du signal de sous-bande après traitement en bloc : cas d'une exponentielle complexe dans le bruit filtré (RSB 6 dB). **En haut,** estimation des paramètres de la composante au sens du maximum de vraisemblance (avec calcul des TFD sur $32 \times T$ points). **En bas,** atténuation spectrale. La longueur des blocs de signal de sous-bande est de $T = 32$. La fréquence de la composante de signal correspond à un point de discrétisation fréquentielle ($4/T = 0,125$). Les spectres sont calculés par TFD sur 1024 points du signal de sous-bande. Seule une partie de l'axe fréquentiel est représentée.

Le second point semble plus étonnant car on pourrait s'attendre à ce que les techniques de modélisation fonctionnent mieux que l'atténuation spectrale lorsque les hypothèses du modèle sont respectées. Cependant, si on se place dans l'hypothèse où la sous-bande ne contient effectivement qu'une seule composante de signal mêlée au bruit filtré, nous avons vu au paragraphe 4.2.5.a que l'estimateur du maximum de vraisemblance revient à synthétiser un signal sinusoïdal à partir du point correspondant au maximum du périodogramme du bloc de signal de sous-bande. Dans une telle situation, nous avons eu l'occasion d'indiquer au paragraphe 3.1.1.a que le résultat obtenu par atténuation spectrale à court-terme correspond quasiment au même signal. Les figures 4.12 et 4.13 illustrent ce point pour un signal de sous-bande idéal ne contenant qu'une seule composante de signal (avec un RSB de 3 dB). Pour éviter tout problème de discrétisation, le périodogramme des blocs de sous-bande nécessaire à l'estimation du maximum de vraisemblance (cf. paragraphe 3.1.1.a) est calculé par *zero-padding* sur $32 \times T$ points. La figure 4.12 représente le cas où la fréquence de la composante de signal correspond à un point de discrétisation (du type $2\pi r/T$). Dans ce cas, on constate que les signaux de sous-bande obtenus après modifications ont des spectres quasiment identiques. Nous avons déjà vu au paragraphe 3.1.1.a que pour l'atténuation spectrale les effets liés à la discrétisation sont minimaux lorsque la fréquence de la sinusoïde correspond à un point de discrétisation fréquentielle. Dans ce cas particulier, on peut même dire que le signal obtenu par atténuation spectrale correspond exactement au maximum de vraisemblance car il n'y a pas de fenêtre de pondération à l'analyse : le spectre de la sinusoïde est donc réduit à un point. Une remarque importante est que le spectre du signal modifié n'est pas réduit à la seule composante de signal : on distingue nettement du bruit autour de la composante. Ceci est dû au fait que la fréquence estimée (par le maximum de vraisemblance) n'est pas une quantité déterministe : elle possède une certaine variance (qui dépend du RSB). Par conséquent, les paramètres estimés de la sinusoïde ne sont pas semblables dans toutes les blocs de signal de sous-bande, ce qui se traduit par une modulation de la composante restaurée clairement visible sur la figure 4.12.

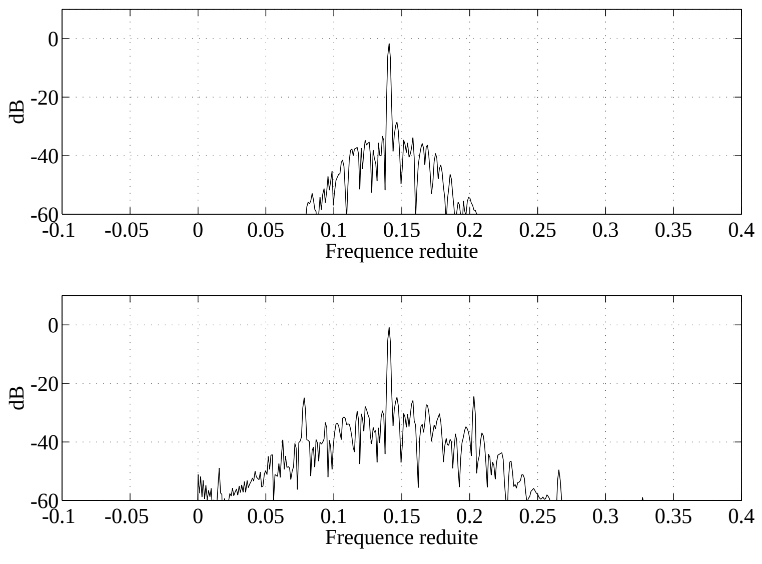

Figure 4.13: Spectre du signal de sous-bande après traitement en bloc : cas d'une exponentielle complexe dans le bruit filtré (RSB 6 dB). **En haut,** estimation des paramètres de la composante au sens du maximum de vraisemblance (avec calcul des TFD sur $32 \times T$ points). **En bas,** atténuation spectrale. La longueur des blocs de signal de sous-bande est de $T = 32$. La fréquence de la composante de signal est située exactement entre deux points de discrétisation fréquentielle ($4,5/T \approx 0,141$). Les spectres sont calculés par TFD sur 1024 points du signal de sous-bande. Seule une partie de l'axe fréquentiel est représentée.

La figure 4.13 présente les spectres des résultats obtenus dans le cas où la fréquence de la composante se situe exactement entre deux points de discrétisation fréquentielle. Compte tenu du facteur de *zero padding* très important utilisé pour calculer les paramètres, le résultat de l'estimation au sens du maximum de vraisemblance (représenté en haut de la figure) est comparable à celui de la figure 4.12. Par contre le résultat obtenus par atténuation spectrale (en bas de la figure) présente une modulation beaucoup plus importante que dans le cas précédent. Cette modulation traduit simplement les effets dépendant du temps qui apparaissent, lors de l'atténuation spectrale, du fait de la troncature du spectre de la sinusoïde (cf. paragraphe 3.1.1.a). Toutefois nous avons vu au paragraphe 3.1.1.a que dans les situations usuelles, cette modulation n'est pas audible. Ici l'argument du bruit résiduel (paragraphe 3.1.1.a) est encore valable bien que le facteur de surestimation utilisée pour la restauration du signal de sous-bande par atténuation spectrale soit important. En effet, une sous-bande où une composante de signal est détectée est en général entourée de sous-bandes "non détectées" pour lesquelles le niveau de bruit résiduel est important.

En conclusion, on peut retenir que dans le cas particulier d'une seule composante de signal, les résultats de l'atténuation spectrale sont identiques à ceux obtenus par l'estimation au sens du maximum de vraisemblance, aux problèmes de discrétisation du spectre près. De plus, ces problèmes de discrétisation ne se traduisent pas par des effets audibles. Ceci montre que lorsque le signal de sous-bande se conforme au modèle, les résultats obtenus par modélisation sinusoïdale ne sont pas meilleurs que ceux obtenus par atténuation spectrale.

Un dernier point concerne le problème de résolution spectrale : on sait qu'une modification effectuée par atténuation spectrale possède une résolution fréquentielle limitée (cf. paragraphes 2.3.2 et 3.3.1.b) de l'ordre de quelques fois le pas de discrétisation fréquentielle. Dans le cas de composantes de signal de fréquences très proches, le résultat obtenu par atténuation spectrale ne peut donc pas être équivalent à celui obtenu par une méthode de modélisation sinusoïdale pour laquelle le problème de résolution ne se pose pas dans les mêmes termes [Laroche 93b] (il ne s'agit pas d'une limite fixe mais d'une dégradation progressive des performances vis à vis du bruit lorsque les fréquences à estimer se rapprochent [Hua 90]). Toutefois nous n'avons pas réussi à construire des exemples permettant de mettre en évidence des différences audibles. Il faut en effet se souvenir que les signaux de sous-bande issus de la TFCT sont sous-échantillonnés par un facteur R (cf. figure B.1), par conséquent une différence de fréquence de l'ordre de quelques points sur le spectre du signal de sous-bande se traduit par une différence de fréquence R fois plus faible à la fréquence d'échantillonnage du signal traité. Par exemple, l'intervalle de $4/T$, mesuré sur le spectre du bloc de signal de sous-bande, correspond à un écart de fréquence de l'ordre de 5 Hz compte tenu des valeurs utilisées ($T = 32$, N équivalent à une vingtaine de millisecondes). On conçoit que dans ces conditions, des éventuelles différences liées à la présence ou non de bruit entre les composantes de signal soient difficilement perceptibles.

4.2.6 Conclusion

4.2.6.a Résultats

Pour illustrer les résultats obtenus par le système de traitement qui vient d'être décrit, la figure 4.14 présente une comparaison avec les résultats de l'atténuation spectrale à court-terme classique. Le signal traité est ici le même que sur la figure 4.8 (son de piano isolé). La quantité représentée correspond à la puissance du signal dans une bande située autour de la fréquence correspondant au troisième harmonique du signal (à l'indice fréquentiel 57 sur la figure 4.8). Cette courbe d'évolution de la puissance au cours du temps à été obtenue par TFCT avec les

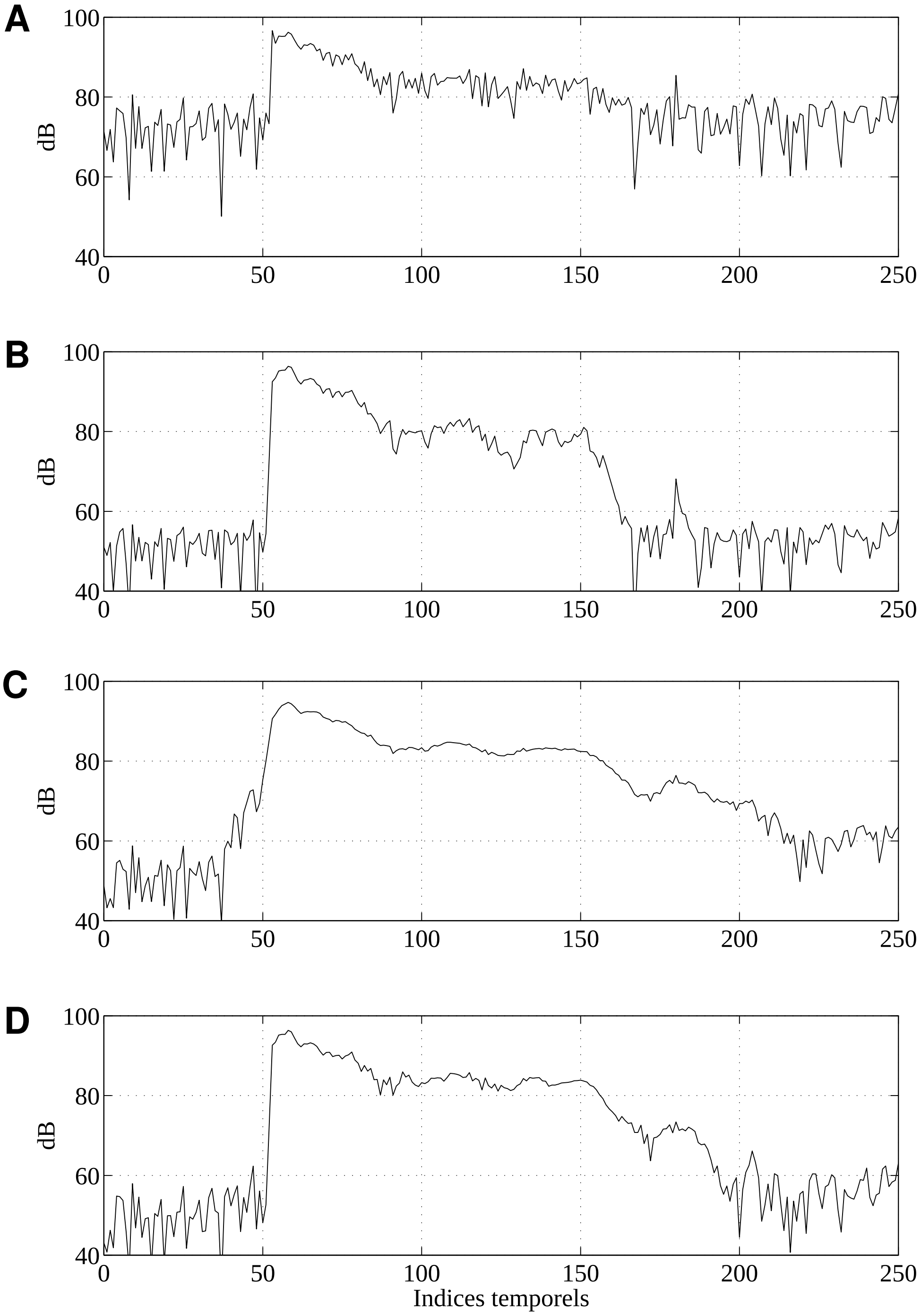

Figure 4.14: Comparaison de différents traitements d'un signal bruité : puissance en fonction du temps pour la fréquence correspondant à un des partiels du signal. **A,** Signal bruité. **B,** Restauration par atténuation spectrale à court-terme avec une trame de $N = 1024$ points (21 ms). **C,** Restauration par atténuation spectrale à court-terme avec une trame de $N = 16384$ points (340 ms). **D,** Restauration sélective des signaux de sous-bande ($N = 1024$, $T = 32$). Calcul de la courbe d'évolution de la puissance par TFCT sur 1024 points, avec un recouvrement de 50% : seul le point d'indice fréquentiel 57 est représenté, les indices temporels des trames de TFCT sont indiqués en abscisse.

(le signal traité est le même que sur la figure 4.8) mêmes paramètres que sur la figure 4.8 (trame de 21 ms) : la quantité représentée est donc $|X(p, \omega_{57})|^2$. Les indices temporels sont identiques sur les figures 4.8 et 4.14.

La partie **A** de la figure 4.14 correspond au signal bruité. On distingue sur cette représentation l'apparition (indice $p \approx 53$) puis la décroissance de la composante de signal. Compte tenu des paramètres de TFCT utilisés pour l'analyse, la présence de la composante est difficilement détectable au delà de l'indice $p = 170$.

La partie **B** représente l'analyse du résultat obtenu par atténuation spectrale à court-terme avec des trames de longueur 1024 (durée de 21 ms). On constate effectivement que le niveau du bruit de fond (présent avant l'indice $p = 50$) a été réduit (d'environ 20 dB). A partir de l'indice $p = 150$, on observe une disparition progressive de la composante de signal due au fait que, compte tenu des paramètres utilisés pour la TFCT, celle-ci est en train de passer sous le niveau de bruit.

La partie **C** correspond à l'analyse du résultat obtenu par atténuation spectrale à court-terme avec des trames de longueur 16384 (environ 340 ms). On constate cette fois que la composante est présente beaucoup plus longtemps (de manière nette, jusqu'à l'indice $p = 200$), ce qui traduit le fait qu'une multiplication par 16 de la durée de la fenêtre équivaut à un abaissement de 12 dB de la limite de restauration (cf. paragraphe 3.1.1.b). On retrouve bien cette différence d'environ 12 dB entre les niveaux correspondant aux indices $p = 150$ (début de la disparition de la composante, avec $N = 1024$) et $p = 200$ (même phénomène pour $N = 16384$). On vérifie de plus que les fluctuations de l'amplitude de la composante, constatées sur la partie **B**, entre les indices $p = 100$ et $p = 150$, ont disparu. Ceci confirme que ces fluctuations ne sont pas caractéristiques du signal mais bien dues au traitement. Ces fluctuations disparaissent sur la partie **C** grâce à l'abaissement de la limite de restauration (cf. paragraphes 3.2.2 et 3.3.1). Par contre, on constate qu'au moment de l'apparition de la composante, le traitement avec une fenêtre aussi longue se traduit par un étalement très sensible du transitoire. On vérifie au passage que cet effet de lissage est non-causal (cf. paragraphe 3.1.2).

Enfin, la partie **D** présente le résultat obtenu par la technique de restauration sélective des signaux de sous-bande qui vient d'être décrite. La longueur des trames à court-terme de la TFCT utilisée pour le traitement est de $N = 1024$, et les blocs de sous-bande ont une longueur de $T = 32$. Sur cet exemple, on peut dire que la technique atteint pleinement son objectif : la partie transitoire (avant l'indice $p = 80$) est identique au résultat obtenu par atténuation spectrale à court-terme avec $N = 1024$ (partie **B** de la figure), tandis que pour la partie quasi-stationnaire du signal (indices entre 80 et 200), le résultat est très proche de ce qui est obtenu avec des trames de longueur $R \times T = 16384$ (cas correspondant à la partie **C**). Pour être plus précis, on peut se reporter à la figure 4.8 qui indique que la composante de signal correspondant au troisième harmonique du signal est détecté entre les indices $p = 48$ et $p = 176$. En comparant les parties **C** et **D** de la figure 4.14, on constate effectivement que sur la partie **D** le niveau de la composante "décroche" aux alentours de l'indice $p = 180$, ce qui correspond au moment où la composante n'est plus détectée et où la restauration du signal de sous-bande recommence à se faire de manière ponctuelle (paragraphe 4.2.3.b).

Pour donner des indications plus qualitatives sur les résultats obtenus, on peut dire que dans un cas tel que celui de la figure 4.14, l'amélioration obtenue est assez spectaculaire. Le timbre du signal restauré est beaucoup plus brillant que celui obtenu par atténuation spectrale à court-terme avec des fenêtres courtes. De plus, l'effet de modulation audible dans le cas de fenêtres courtes est totalement éliminé[3], et les phénomènes de fluctuations sont grandement réduits. De

[3]Il faut se souvenir que c'est un point que la figure 4.14 ne peut pas "montrer" : le résultat présenté sur la

plus, l'attaque du signal, pour autant que l'on puisse en juger, reste identique à ce qu'elle est avec des fenêtres courtes. Par rapport à un résultat obtenu par atténuation spectrale à court-terme avec une durée de trame moyenne (de l'ordre de 50 ms), on vérifie bien que le résultat du traitement proposé présente à la fois un timbre plus brillant, une décroissance plus naturelle et une attaque beaucoup plus marquée.

D'une manière plus générale, pour des enregistrements qui se prêtent bien aux caractéristiques de la méthode, les résultats obtenus sont très encourageants. Par exemple, pour les enregistrements de piano solo dont nous disposions, on peut dire que les résultats obtenus sont très distinctement meilleurs que ceux de l'atténuation spectrale à court-terme usuelle, *quelle que soit la durée de trame utilisée*. Pour des enregistrements polyphoniques plus complexes, il semble que la différence entre le traitement proposé et l'atténuation spectrale à court-terme reste assez nette. Toutefois, ceci est plus ou moins vrai selon la nature exacte de l'enregistrement traité (niveau de bruit, type de sons instrumentaux, etc.) : lors des tests d'écoute que nous avons effectués, il s'est avéré parfois nécessaire de faire subir à l'auditeur un "apprentissage" (écoute commentée des signaux à comparer) avant qu'il lui soit possible de séparer les résultats des deux traitements. Il est clair qu'ici pour pouvoir avancer des résultats plus précis sur l'apport du traitement il faudrait effectuer une évaluation perceptive avec un protocole plus élaboré.

4.2.6.b Améliorations

Pour nuancer ces résultats plutôt positifs, il faut tout de même signaler que le traitement de restauration sélective des signaux de sous-bande, tel qu'il a été décrit, est susceptible de générer des artefacts dans certaines situations. En dehors de l'aspect d'évaluation perceptive qui vient d'être mentionné, c'est principalement sur ces différents points qu'il serait nécessaire d'apporter des solutions :

Réverbération Il arrive que le signal restauré présente ponctuellement des effets de réverbération peu naturels. Nous avons vu au paragraphe 4.2.4.b qu'il était nécessaire de prendre des précautions particulières pour éviter l'apparition de pré-échos. Il s'agit ici exactement du même phénomène qui se manifeste à la fin du signal. Pour la plupart des sons instrumentaux, ce phénomène est inaudible car la décroissance du signal est beaucoup plus lente que l'attaque (c'est par exemple le cas pour les enregistrements de piano). De plus, ce phénomène peut être masqué par la présence d'un effet naturel de réverbération. Toutefois pour des sons qui présentent des variations temporelles rapides, il s'est avéré que cet effet de réverbération pouvait être audible et gênant. Ici la mesure à prendre est assez simple (dans le principe) puisqu'il suffirait d'appliquer la procédure de confirmation de détection, décrite au paragraphe 4.2.4.b, dans le sens temporel inverse pour éliminer les détections finales. Cependant, outre la difficulté pratique d'un traitement non-causal, nous avons signalé au paragraphe 4.2.4.b qu'une seconde application de la procédure de confirmation se traduit en général par une perte importante d'information.

Création de composantes parasites Nous avons signalé au paragraphe 4.2.5.a que l'utilisation de techniques de modélisation sinusoïdale pour la restauration en bloc du signal de sous-bande pouvait générer des composantes sinusoïdales parasites. Il s'avère que dans des situations particulières, ceci peut aussi se produire même avec la technique de restauration non-paramétrique décrite au paragraphe 4.2.5.b. Comme pour les techniques de modélisation, la cause de ce phénomène est toujours la sélectivité insuffisante du test de détection : il peut arriver dans des situations complexes (enregistrement polyphonique, variations rapides du signal)

partie **A** est inacceptable car le bruit subsistant autour de la composante de signal est audible (paragraphe 3.3.1).

qu'une sous-bande où une composante est détectée contienne en fait un signal relativement compliqué. Dans un tel cas, compte tenu de la forte surestimation utilisée lors du traitement en bloc (paragraphe 4.2.5.b), le spectre du bloc de signal de sous-bande après modification spectrale risque de se réduire à quelques points non-nuls, ce qui se traduit par la présence de sinusoïdes parasites. Pour remédier à ce problème, il serait possible d'utiliser pour la restauration en bloc du signal de sous-bande une technique d'atténuation spectrale à faible distorsion de type Ephraim et Malah (paragraphe 4.1.2). On peut aussi songer à modifier le principe du test de détection de composantes afin de le rendre plus sélectif, mais intuitivement il semble difficile de pouvoir trouver une procédure aussi robuste au bruit que celle décrite au paragraphe 4.2.2.

Pour conclure, on peut dire que le traitement tel qu'il a été présenté ne doit pas être considéré comme un système définitif mais plutôt comme un prototype qui démontre les potentialités du principe retenu : décomposition fréquentielle du signal à traiter par un banc de filtres à faible résolution, suivie d'une restauration de chaque signal de sous-bande, soit locale (bonne localisation temporelle), soit en bloc (bonne résolution fréquentielle), selon la nature du signal présent dans la sous-bande. Le point important est que pour éviter un compromis global, portant sur tout le signal (et fixé par l'utilisateur), c'est un test de détection, effectué pour chaque bloc de signal de sous-bande, qui permet de choisir la technique de restauration la plus appropriée.

Nous avons détaillé ici la mise au point d'un système fonctionnant selon ce principe pour le cas de signaux musicaux bruités. Un des points les plus importants que cette démarche a permis d'illustrer est le fait que pour obtenir un test de détection pertinent, il est nécessaire de considérer des blocs de signaux de sous-bande d'une taille relativement importante. Pour des développements futurs, et éventuellement des applications à d'autres types de signaux, on peut dire que c'est là que réside le problème clef : comment améliorer le fonctionnement de la partie détection, tout en conservant une durée globale d'observation qui soit acceptable compte tenu des signaux traités ? On peut penser qu'une solution pour avancer sur cette voie consiste à abandonner l'hypothèse de présence d'une seule composante de signal par sous-bande qui conduit à effectuer, dès le départ, une décomposition relativement sélective par TFCT. Toutefois, sans cette hypothèse simplificatrice nous ne sommes pas assurés de pouvoir formuler une stratégie de détection performante. Nous avons tout de même vu au paragraphe 4.2.4.a un exemple de stratégie de détection qui pourrait être applicable dans le cas de plusieurs composantes (SFM). Cependant, il faudrait vérifier que cette procédure possède bien les performances souhaitées (en particulier en ce qui concerne la robustesse vis à vis du bruit). Cette évaluation précise de la partie détection est d'autant plus importante que nous avons vu au paragraphe 4.2.5 que les caractéristiques de la technique utilisée pour la détection influent fortement sur les performances de la restauration effectuée sur les blocs de signaux de sous-bande.

Perspectives

Le document comportant deux parties relativement distinctes (chapitres 2 et 3 d'un côté, et 4 de l'autre), les conclusions proprement dites de chacune de ces parties ont été reportées en fin des chapitres correspondants : paragraphe 3.4 pour l'analyse des résultats de l'atténuation spectrale à court-terme, et paragraphe 4.2.6 en ce qui concerne la technique de restauration sélective des signaux de sous-bande. D'autre part, plusieurs applications directes de l'évaluation des techniques de débruitage effectuée au chapitre 3 ont été mentionnées au paragraphe 1.3.1. Nous allons simplement donner ici quelques indications concernant les prolongements possibles de cette étude.

Nous avons déjà eu l'occasion de signaler que l'aspect le plus délicat du travail d'évaluation de la qualité du débruitage mené au chapitre 3 est celui qui concerne les parties transitoires des signaux traités. En effet, pour la partie quasi-stationnaire des signaux musicaux, il nous a été possible, d'une part, de caractériser très précisément de manière analytique l'effet du traitement de débruitage, et d'autre part, d'évaluer l'audibilité des phénomènes mis en évidence. Par contre, pour le traitement des parties transitoires nous avons seulement réussi à obtenir une description précise dans le cadre d'un "modèle" assez grossier du comportement transitoire des signaux musicaux. Les résultats obtenus sont assez encourageants car ils permettent déjà de proposer une explication à certaines constatations connues sur le fonctionnement du débruitage par atténuation spectrale à court-terme. Cependant, une telle description est insuffisante pour fournir, par exemple, des tolérances perceptives vis à vis du phénomène de lissage mis en évidence.

Il faut d'ailleurs souligner que la difficulté ne réside pas uniquement ici dans le fait que le transitoire "idéal" que nous avons considéré peut difficilement être utilisé pour modéliser complètement un transitoire musical réel. En effet, à travers les différents tests d'écoute effectués au cours de ce travail, il est apparu que l'évaluation auditive du comportement durant les parties transitoires est extrêmement difficile à effectuer. Même dans le cas du signal transitoire simple utilisé au chapitre 3, il est très délicat de distinguer à l'écoute les différents stades du lissage mis en évidence au paragraphe 3.1.2. En l'état actuel des choses, il nous serait par exemple très difficile de préciser quelle est la variation minimale d'un paramètre de la méthode (par exemple la durée des trames à court-terme) qui risque d'entraîner une modification perceptible. Pour progresser sur ce point, il faudrait mettre en œuvre un véritable test psychoacoustique. Outre la nécessaire réflexion sur le protocole de test et l'exploitation des résultats, il serait nécessaire de porter ici une attention toute particulière au choix des signaux-test. On peut en effet penser que compte tenu de la brièveté des phénomènes que l'on cherche à détecter, il serait par exemple préférable d'utiliser des séries de répétitions du même signal.

Dans le domaine de l'amélioration des techniques utilisées pour la réduction de bruit de fond, il nous semble important de souligner que la technique de restauration sélective des signaux de sous-bande présentée au chapitre 4, même si elle n'est pas applicable à tous les signaux telle quelle, présente des potentialités très intéressantes. Les résultats obtenus pour certains signaux représentent une amélioration très nette par rapport aux techniques habituelles de soustraction spectrale. D'une manière plus générale, les tests d'écoute que nous avons effectués nous permettent de dire que même si les résultats obtenus ne sont pas toujours complètement satisfaisants, cette technique est celle qui permet d'obtenir les améliorations les plus sensibles. Nous avons indiqué au paragraphe 4.2.6.b quels sont les points sur lesquels il nous paraît nécessaire de perfectionner la technique de restauration sélective des signaux de sous-bande.

Enfin, pour conclure, nous souhaitons mentionner deux autres aspects plus "prospectifs" qui nous paraissent pouvoir donner lieu à des développements très fructueux. Tout d'abord, nous avons vu au chapitre 3 que l'utilisation de propriétés connues de l'audition fournit des résultats très intéressants : même en ne faisant intervenir que le phénomène de masquage fréquentiel simultané, il est déjà possible de justifier une bonne partie des constatations empiriques formulées à propos des systèmes de débruitage (paragraphes 3.1.1 et 3.3.1). Il semble donc logique de chercher à exploiter une propriété perceptive comme le masquage fréquentiel dans le cadre de la restauration. Il ne faut pas s'attendre à ce que la prise en compte de ce type de propriété fournisse une amélioration "objective" (au sens mise en évidence d'un signal indétectable). Cependant, il est clair que l'évaluation de l'effet de masquage fréquentiel présente un grand intérêt pour le contrôle et la décision. Il est par exemple possible d'envisager que certains paramètres (niveau de bruit résiduel, surestimation) soient ajustés constamment au cours du traitement afin de garantir que les distorsions du signal restent inaudibles. Une difficulté pratique est que dans le cas de la restauration d'enregistrements, le niveau d'écoute absolu n'est pas fixé. Or, la perception du signal restauré dépend beaucoup du niveau d'écoute : il n'est pas rare qu'à faible niveau, le bruit de fond semble totalement disparaître parce qu'il est passé en dessous des seuils absolus d'audition. Une idée qui peut tout de même fournir une estimation approchée du niveau d'écoute probable consiste à mesurer l'intensité sonore perçue par l'auditeur (cf. paragraphe 1.4.4) et à se baser sur la différence entre l'intensité maximale et l'intensité minimale au cours de l'enregistrement. En effet, dans la plupart des situations, l'auditeur va ajuster le volume d'écoute afin que le niveau maximal (perçu) soit raisonnablement fort et qu'en même temps, le niveau minimal soit distinctement audible.

Le dernier point concerne l'utilisation de la synthèse dans le cadre de la restauration. A priori, nous avons banni l'utilisation de la synthèse en excluant au paragraphe 1.3.2 la possibilité de modéliser le signal à restaurer. Les résultats du paragraphe 4.2.5.a permettent d'ailleurs de confirmer ce que nous avions annoncé au départ : la modélisation d'un signal audio complexe reste une tâche très complexe. A l'heure actuelle, pour des enregistrements musicaux polyphoniques, l'approche modificative (modification du signal bruité) utilisée ici reste beaucoup plus réaliste que l'approche par synthèse (détermination des paramètres du modèle à partir du signal bruité, puis synthèse). Cependant, il faut souligner qu'il existera toujours des cas qui résisteront à une approche purement modificative. En effet, la mise en évidence du signal dans le bruit n'est possible qu'au dessus d'une certaine limite. L'utilisation d'une technique appropriée peut permettre d'abaisser cette limite (c'est l'objet du chapitre 4), mais non de la supprimer complètement. On peut penser que l'utilisation ponctuelle d'une technique de synthèse, dans les cas où l'existence de cette limite pose problème, est une bonne solution. Nous avons par exemple constaté qu'un point très désagréable, qui limite notablement les possibilités de restauration, est la variation du timbre des sons en fonction de leur nuance. En effet, pour un enregistrement fortement bruité le spectre du signal musical est notablement modifié, et ce d'autant plus qu'il s'agit d'un signal de faible niveau. Au cours d'une phrase musicale où la nuance de jeu varie, on peut donc obtenir des variations considérables du timbre d'un même instrument. Dans ce cas particulier, seule l'utilisation locale d'une technique de synthèse qui prend en compte la "continuité" (dans un sens qui reste à préciser !) du timbre de l'instrument pourrait permettre de surmonter le problème.

Annexes

Annexe A

Traitements des défauts localisés

Les enregistrements anciens présentent généralement un certain nombre de défauts qui ne sont pas assimilables à du bruit de fond. Parmi ceux-ci, nous avons choisi de désigner par le terme de “défauts localisés” les dégradations qui n’affectent que de courts instants de l’enregistrement. Il s’agit par exemple des craquements qui se manifestent lors de l’écoute d’un disque 78 tours dont la surface a été rayée.

Cette annexe principalement bibliographique a pour but de présenter l’état de l’art dans le domaine du traitement de ces défauts localisés. Le paragraphe A.1 est consacré à la description des divers défauts localisés fréquemment présents sur les enregistrements anciens. On distingue en particulier deux types de défauts localisés, les craquements et les bruits impulsionnels. Les seconds diffèrent des premiers par leur extrême brièveté (durée de l’ordre d’une milliseconde), ce sont aussi les défauts les plus importants quantitativement. Les paragraphes A.2 et A.3 présentent les techniques utilisables, dans le cadre d’un système de restauration automatique, afin d’éliminer les bruits impulsionnels. L’accent est mis sur les techniques qui semblent être à ce jour les plus robustes et les plus efficaces qui impliquent, aussi bien pour la détection que pour la correction des bruits impulsionnels, l’utilisation d’un modèle AR du signal. Les points A.1.2 et A.2.2.c illustrent des problèmes peu développés dans la littérature auxquels nous avons eu à faire face lors de la mise au point d’un système de traitement des bruits impulsionnels. Le système que nous utilisons n’est pas décrit ici car il ne présente pas d’innovation technique majeure. A ceci près que la mise en œuvre d’une modélisation AR en blocs successifs implique plusieurs changements par rapport à la technique décrite dans [Vaseghi 88b] (le lecteur intéressé par ce point trouvera la description de cette technique dans [Le May 91]).

A.1 Caractérisation des défauts localisés

Le but est ici de définir les critères permettant de caractériser les défauts qui doivent être traités. A priori, ces critères correspondent d’une part, aux caractéristiques physiques des défauts localisés présents sur les enregistrements à traiter, et d’autre part, à l’audibilité de ce type de défauts.

A.1.1 Caractéristiques physiques

Les défauts localisés apparaissent principalement lors du traitement d'enregistrements provenant de disques analogiques. Quantitativement, ce cas est très important puisqu'il recouvre au moins l'ensemble des enregistrements réalisés avant la généralisation de l'enregistrement original sur bande magnétique, c'est à dire antérieurs à la fin des années quarante [Jessel 85]. Sur les disques analogiques les défauts localisés proviennent de détériorations ponctuelles du support dues, par exemple, à l'usure ou à des dépôts de poussières. Il existe de nombreuses références consacrées au fonctionnement des différents types de disques analogiques, par exemple, le chapitre 8 de la référence [Blair Benson 88] qui fournit une bibliographie très détaillée, [Roys 78] qui réuni un ensemble d'articles sur ce sujet, et enfin [Aes 77] surtout pour l'aspect historique. Malheureusement, il nous a été impossible de trouver des données concernant la nature des bruits apparaissant lors de la lecture. L'étude la plus complète dont nous disposions a été effectuée par J-C. Valière sur un échantillon de cinq enregistrements provenant soit de disques 78 tours, soit d'enregistrements sur cylindres datant du début du siècle [Valiere 91].

Avant d'exposer les principaux résultats de cette étude, notons tout d'abord que l'on distingue en général deux types de défauts localisés [Vaseghi 88b] [Valiere 91] :

Les bruits impulsionnels (*impulsive-noise* ou *clicks* en anglais) qui ont une durée extrêmement brève, de l'ordre de la milliseconde. Les bruits impulsionnels constituent l'essentiel des défauts localisés présents sur les disques analogiques.

Les craquements (*scratches* en anglais) qui perturbent le signal durant des durées beaucoup plus importantes de l'ordre de la vingtaine de millisecondes. Les craquements étant dus à des dégradations importantes du support, ils sont souvent d'amplitude très importante par rapport au signal enregistré.

La figure A.1 présente quelques exemples de craquements issus d'un même enregistrement, on note que le support temporel des défauts varie ici entre 5 et 10 ms. Il est difficile de trouver des méthodes de traitement élaborées applicables au cas des craquements car le signal enregistré est en général complètement perdu sur des durées très importantes pouvant aller jusqu'à une cinquantaine de millisecondes. Dans ces conditions, il est illusoire de vouloir retrouver le signal musical original, et il est alors plus réaliste d'essayer de dissimuler le défaut. Le type de technique utilisé dans ce cas s'apparente plus au montage, c'est à dire la coupure de la zone de signal corrompue avec raccordement progressif des deux tronçons (*cross-fade* en anglais) [Blair Benson 88]. Il faut noter qu'en exploitant la quasi périodicité du signal, il est possible d'effectuer cette modification sans défaut audible lorsque le signal musical enregistré s'y prête [Goodman 86]. Un autre argument est que les craquements étant liés à des défauts importants du support, ils sont en général peu nombreux. Dans le cas d'une seule rayure sur un disque 78 tours, il apparaît un craquement environ toutes les 750 millisecondes, une technique de traitement très simple est donc suffisante car les artefacts éventuels sont ponctuels et donc plus difficilement perceptibles. En pratique, une coupure de quelques dizaines de millisecondes de signal pratiquée une ou deux fois par seconde est quasiment toujours inaudible.

Le seul exemple de traitement quasi automatique des craquements est décrit par S. Vaseghi [Vaseghi 88b] [Vaseghi 92]. Cependant, le traitement proposé repose sur des hypothèses contraignantes puisque l'auteur avance que les craquements sont relativement semblables et que l'effet du craquement est en grande partie additif. Les justifications fournies par S. Vaseghi ne permettent pas de savoir dans quelle mesure ces hypothèses sont vérifiées pour un enregistrement quelconque. Cependant, il semble que la première hypothèse (répétabilité des craquements) soit relativement

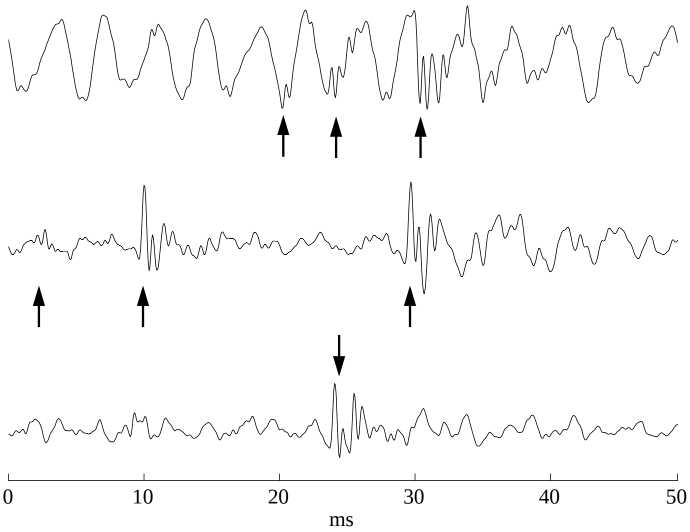

Figure A.1: Exemples de craquements : trois trames de 50 ms provenant du même enregistrement (transfert depuis un enregistrement datant du début du siècle). Les positions marquées par les flèches correspondent à un marquage manuel du fichier.

bien vérifiée en pratique (voir l'exemple de la figure A.1), ce qui permet d'envisager une détection automatique des craquements. La justification physique de cette propriété consiste à dire qu'un craquement correspond en première approximation à la réponse mécanique du système de lecture à une déformation très brève [Valiere 91].

Le cas des bruits impulsionnels a donné lieu a plus de développements théoriques puisque la distorsion du signal étant très brève, il est possible d'envisager des méthodes d'interpolation du signal enregistré afin de corriger le défaut. Dans le cas des bruits impulsionnels, la seule hypothèse généralement utilisée est que le support temporel du défaut est très faible, c'est à dire au maximum de quelques millisecondes [Vaseghi 88b] [Valiere 91]. Pour l'étude mentionnée précédemment, J-C. Valière a utilisé une technique de détection automatique reposant sur un filtrage passe-haut du signal (voir le paragraphe A.2.1) pour obtenir les caractéristiques moyennes des bruits impulsionnels. Les résultats obtenus soulignent plusieurs points importants :

1. Le nombre moyen de bruits impulsionnels détectés par seconde est de plusieurs dizaines. Pour les enregistrements provenant de cylindres, les bruits impulsionnels semblent plus nombreux que pour les enregistrements provenant de 78 tours (jusqu'à 80 détections en moyenne par seconde).

2. La durée moyenne d'un bruit impulsionnel est comprise entre 0,5 et 1 ms, et la quasi totalité de ces bruits impulsionnels ont une durée inférieure à 1,5 ms.

Ceci confirme bien le fait que les bruits impulsionnels sont des événements à support temporel très bref nettement différenciés du signal musical (voir l'exemple de la figure A.2). Par contre, cette étude met en évidence le fait que le nombre de bruits impulsionnels à détecter est très important, ce qui implique que la technique de détection doit obligatoirement être complètement automatique. Pour donner une idée de la charge représentée par le traitement, notons que pour un morceau musical complet issu d'un disque 78 tours, le nombre d'interventions nécessaires est en moyenne supérieur à 2000. D'autre part, le système de correction utilisé pour éliminer

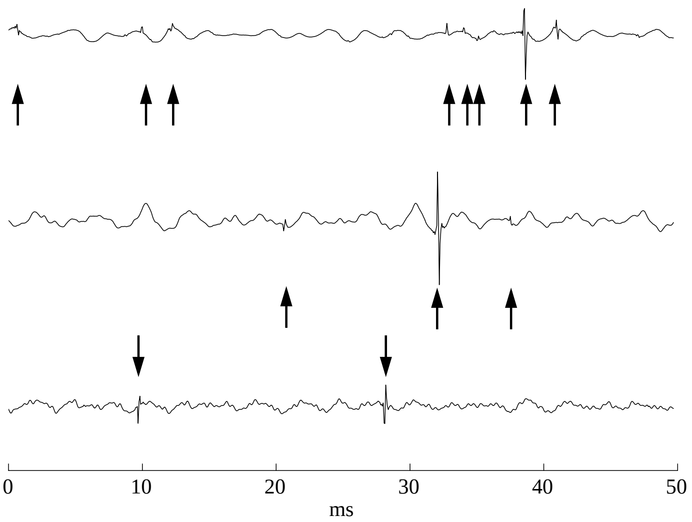

Figure A.2: Exemples de bruits impulsionnels : trois trames de 50 ms provenant du même enregistrement (transfert depuis un disque 78 tours). Les positions marquées par des flèches ont été détectées automatiquement en utilisant la modélisation AR du signal. Ces détections correspondent toutes à des défauts audibles.

les bruits impulsionnels étant amené à intervenir plusieurs dizaines de fois par seconde, il est nécessaire de choisir une technique de bonne qualité afin d'éviter des artefacts audibles.

Pour conclure sur la caractérisation physique des défauts localisés présents sur les enregistrements anciens, notons qu'il serait très intéressant de vérifier les conclusions de l'étude expérimentale effectuée par J-C. Valière pour un nombre plus important d'enregistrements et en quantifiant l'influence de la technique de détection utilisée. Nous n'avons pas effectué d'étude systématique comparable, mais il faut noter qu'avec la technique de détection que nous utilisons (détection à partir d'une modélisation AR en blocs successifs décrite dans [Le May 91]), les résultats obtenus sont assez semblables quant au nombre de bruits impulsionnels détectés : en moyenne le nombre de détection par seconde varie entre 5 et 20 pour les enregistrements provenant de disques analogiques. Exceptionnellement, on observe plus de 50 détections par seconde dans le cas d'enregistrements fortement dégradés. Ce résultat dépend du seuil de détection associé à la méthode de détection par modèle AR. Toutefois, les arguments exposés au paragraphe A.2.2 donnent à penser que ce seuil de détection est inférieur au seuil d'audibilité des bruits impulsionnels. Il semble donc légitime de considérer que le nombre de bruits impulsionnels audibles est de l'ordre de plusieurs dizaines par seconde.

La figure A.3 présente l'allure typique du nombre de détections effectuées au cours du temps (ici pour un disque 78 tours datant de 1911). Pour cet exemple, nous avons vérifié grâce à un éditeur de signal que toutes les détections correspondent à des défauts visibles sur la forme d'onde. Par contre, il demeure un nombre assez important de défauts visibles mais non détectés. Toutefois, à l'audition du résultat du traitement, il ne subsiste pratiquement plus de bruits impulsionnels audibles. Au vu des résultats présentés dans les paragraphes A.2.2 et A.1.2, les deux constatations précédentes peuvent s'expliquer par la présence d'un bruit de fond de niveau très important.

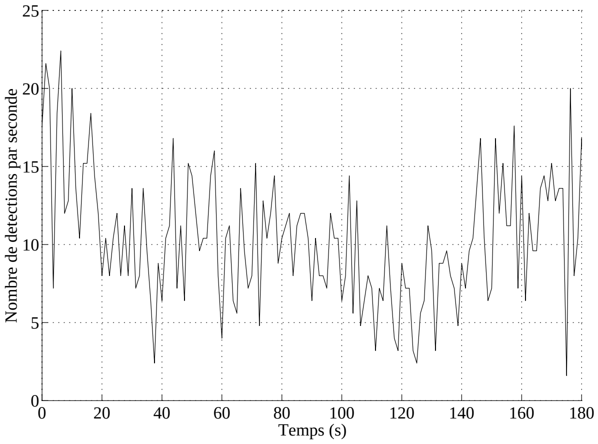

Figure A.3: Nombre de bruits impulsionnels détectés par seconde au cours d'un enregistrement (technique de détection automatique utilisant la modélisation AR).

A.1.2 Audibilité des défauts localisés

Ce second point fixe les performances que doit atteindre la technique de traitement des défauts localisés : idéalement, il faut pouvoir détecter puis corriger avec une précision suffisante tous les défauts qui sont audibles. Le point important à souligner est que ce seuil de détection des défauts localisés varie beaucoup selon la composition spectrale du signal musical contaminé.

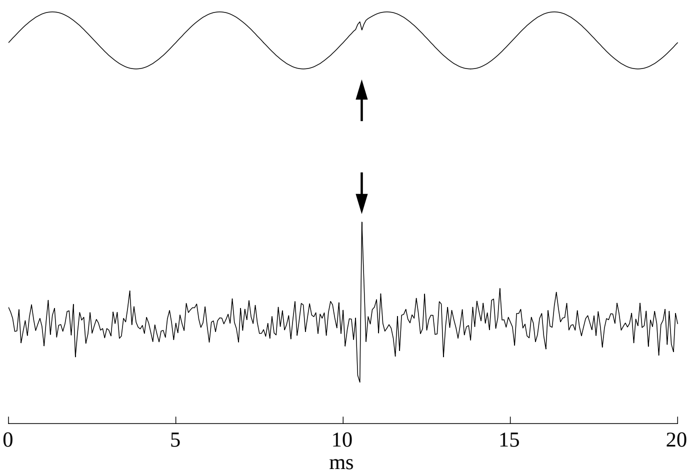

Figure A.4: Deux exemples de bruits impulsionnels : sur le haut de la figure, le bruit impulsionnel qui affecte un signal sinusoïdal est clairement audible (NRB $\approx$ -10 dB), tandis que le bruit impulsionnel superposé au bruit blanc (sur le bas de la figure) est inaudible (NRB $\approx$ 17 dB) .

Cette situation est illustré sur la figure A.4 par un exemple synthétique : le même bruit impulsionnel d'une durée d'environ 0.3 ms, extrait d'un enregistrement réel, a été ajouté à deux signaux test. Le Niveau Relatif du Bruit impulsionnel (ou NRB) est défini ici comme étant le *rapport entre l'énergie moyenne du bruit impulsionnel et la puissance du signal*. Les seuils d'audibilité du bruit impulsionnel ont été évalués lors d'un test d'écoute informel par trois auditeurs différents. Il s'avère que lorsque le bruit impulsionnel affecte un bruit blanc il est nécessaire que le niveau relatif du bruit soit supérieur à 18 dB pour que le bruit impulsionnel soit audible. Par contre, lorsque le même bruit impulsionnel affecte un son pur de fréquence 200 Hz, il est audible même pour des niveaux relatifs inférieurs à -40 dB. Ceci indique d'ailleurs qu'une détection "visuelle" des bruits impulsionnels, effectuée en inspectant la forme d'onde, peut s'avérer insuffisante pour certains types de signaux.

Pour retrouver cette constatation expérimentale à partir de résultats psychoacoustiques connus, nous considérerons que les bruits impulsionnels correspondent à des impulsions de bruit de très faible durée. Ceci est justifié puisque des expériences psychoacoustiques ont montré que pour des durées aussi brèves (inférieures à 1 ms), la forme de l'impulsion influe peu sur la sensation sonore produite [Zwicker 81, §75]. En admettant cette approximation, le seuil d'audibilité d'un bruit impulsionnel d'une durée de 2 ms masqué par un bruit large bande est obtenu pour une valeur de NRB d'environ 3 dB [Zwicker 81, §63]. Pour une durée de 0.2 ms, le seuil d'audibilité s'élève jusqu'à une valeur du NRB de 12 dB. Ceci montre que la quantité significative est ici *l'énergie totale du bruit impulsionnel*, c'est à dire le niveau relatif multiplié par la durée du bruit, et ce à cause des propriétés d'intégration temporelle de l'ouïe pour des sons de durée brève (inférieure à 100 ms) [Zwicker 81, §63]. Il semble que ces seuils soient légèrement sous-estimé dans le cas de conditions standards d'écoute (écoute sur haut-parleurs, dans une pièce moyennement réverbérante) puisque, pour la même expérience, nous obtenons des seuils d'audibilité en moyenne 5 dB au dessus de ceux qui viennent d'être mentionnés. Ces résultats mériteraient d'être vérifiés avec un protocole expérimental plus strict, cependant, on peut retenir l'ordre de grandeur suivant : **pour un bruit impulsionnel de durée moyenne (entre 0,5 et 1 ms) affectant un bruit large bande, le seuil d'audibilité est atteint pour un niveau relatif d'environ 10 dB**.

Pour le cas d'un bruit impulsionnel affectant un son pur, le seuil d'audibilité est beaucoup plus bas. Intuitivement ceci est dû au fait que le bruit impulsionnel possède un spectre très large, il produit une impression sonore qui sera difficilement masquée par un son possédant un spectre très étroit. Pour pouvoir pousser ce raisonnement plus loin, il est nécessaire de considérer la puissance du bruit stationnaire équivalent au bruit impulsionnel, c'est à dire le bruit blanc perçu avec la même sensation de force sonore. D'après les résultats présentés dans [Zwicker 81, §74], un bruit impulsionnel d'une durée de 0,5 à 1 ms provoque une sensation de force sonore environ 25 dB plus faible qu'un bruit blanc stationnaire de même niveau (c'est à dire de puissance égale à l'énergie moyenne du bruit impulsionnel). On peut donc penser que le seuil d'audibilité d'un bruit impulsionnel affectant un son pur est supérieur de 25 dB au seuil de détection d'un bruit blanc masqué par le même son pur. Le masquage total d'un bruit blanc par un son pur est extrêmement difficile à obtenir, surtout si l'écoute se fait à niveau élevé. Notons par exemple qu'à fort niveau d'écoute, un bruit blanc additionné à un son pur avec un rapport signal-à-bruit de 70 dB est audible, ce qui indique que les bruits impulsionnels peuvent être audibles jusqu'à un niveau relatif de -45 dB lorsqu'ils affectent un son pur. En pratique, ce seuil d'audibilité est toujours nettement sous-estimé à cause de la présence du bruit de fond. En effet, si en plus du son pur l'enregistrement contient du bruit large bande stationnaire avec un rapport signal-à-bruit de -30 dB (ce qui est peu pour un enregistrement ancien), c'est le bruit de fond qui fixe le seuil d'audibilité des bruits impulsionnels et non plus le son pur. Un autre aspect qui contribue à élever le seuil d'audibilité est que les signaux présents sur les enregistrements musicaux sont nettement plus complexes que des sons purs dans le sens où ils possèdent de nombreuses composantes fréquentielles. Les bruits impulsionnels sont donc plus facilement masqués par un son musical que par un son pur. En conclusion, notons que **pour un enregistrement comportant un bruit de fond de niveau important (rapport signal-à-bruit de 20 dB ou moins), le seuil d'audibilité des bruits impulsionnels est fixé par le bruit de fond, il est donc supérieur à -10 dB**. Ceci justifie le fait que les bruits impulsionnels qui sont audibles sont en général visibles sur la forme d'onde. Par contre, **dans des passages ou le niveau du signal musical est nettement supérieur à celui du bruit de fond, il est possible qu'il existe des bruits impulsionnels audibles mais non visibles sur la forme d'onde**.

Un dernier point sur les bruits impulsionnels concerne la manière dont ils sont effectivement perçus lorsqu'ils sont très nombreux. Nous avons déjà évoqué le fait que sur les disques 78 tours, le nombre moyen de bruits impulsionnels détectés est de l'ordre de 20 à 30 par seconde. Or, on sait que la force sonore produit par des impulsions répétées croit de manière significative dès lors que que la fréquence de répétition dépasse 5 Hz. Ce qui revient à dire que l'ouïe intègre les énergies d'événements séparés temporellement de moins de 200 ms lors de la formation de la sensation de force sonore. D'après les données présentées dans [Zwicker 81, §74], la sensation de force sonore produite par une trentaine de bruits impulsionnels par seconde est 10 à 15 dB plus importante que celle produite par un bruit impulsionnel isolé. Ceci indique que dans la plupart des cas, **même si les bruits impulsionnels ne sont pas audibles individuellement, ils provoquent une sensation sonore perceptible du fait de leur grand nombre**. Cette dernière propriété est très nette dans le cas des disques 78 tours fortement bruités. En effet, nous avons vu précédemment que lorsque le bruit de fond est important, il est rare que les bruits impulsionnels considérés individuellement soient audibles sauf si ils ont une amplitude importante. Par contre, à l'écoute du résultat du traitement des bruits impulsionnels, on note quasiment systématiquement une modification du timbre du bruit de fond, avec une perte audible d'énergie dans le haut du spectre, qui est en fait due à l'élimination des bruits impulsionnels.

Pour évoquer brièvement le cas des craquements, notons que d'après ce que nous venons de voir, plus une impulsion est de durée importante, plus elle est audible. Cependant, ce résultat est vrai pour une impulsion de bruit large bande mais il devient faux pour une impulsion de forme régulière dont la durée est supérieure à 1 ms. En effet, à partir de cette limite, le centre de gravité du spectre de l'impulsion se déplace vers les basses fréquences où l'oreille est moins sensible. On note que pour l'exemple d'une impulsion de forme gaussienne, le seuil d'audibilité est minimal lorsque la durée de l'impulsion est proche de 1 ms [Zwicker 81, §64]. Il est difficile de généraliser ce résultat, mais on peut tout de même remarquer que les craquements ayant en général une forme relativement régulière, la règle du plus long, donc plus audible n'est pas forcément vérifié. En particulier un craquement d'une dizaine de millisecondes n'est pas forcément plus audible qu'un bruit impulsionnel de même niveau relatif. Par contre, la différence entre les deux est très importante en terme de timbre : le craquement est perçu comme un son beaucoup plus grave que le bruit impulsionnel.

A.2 Détection des bruits impulsionnels

Dans la suite, nous allons passer en revue les méthodes de traitement applicables au cas des bruits impulsionnels. D'après ce que nous venons de voir, le bruit impulsionnel sera caractérisé comme un événement très bref (moins de 2 ms) dont il n'est utile de considérer ni la forme exacte, ni le mode d'interaction avec le signal. Par contre il est nécessaire de prendre en compte le fait que le nombre de bruits impulsionnels peut être extrêmement élevé (de l'ordre de plusieurs dizaines par seconde). Enfin, nous avons vu que le système de détection des bruits impulsionnels doit être au moins aussi précis que l'inspection visuelle de la forme d'onde, sachant qu'il devrait par contre être beaucoup plus précis pour certains signaux peu bruités.

Nous avons choisi de présenter tout d'abord les techniques de détection des bruits impulsionnels, puis dans une seconde partie, les solutions au problème de la correction de ce type de défauts.

Cette présentation souligne le fait qu'il est toujours possible d'utiliser des techniques différentes pour ces deux étapes de traitement [Valiere 91]. Toutefois, dans certains cas les méthodes de détection et de correction des défauts étant fortement liés, il est assez difficile de dissocier ces deux étapes (voir l'exemple de [Le May 91]).

A.2.1 Filtrage passe-haut

L'idée la plus ancienne pour détecter les bruits impulsionnels semble être de travailler sur une version du signal préalablement filtré par un filtre passe-haut. La justification de cette procédure est que le bruit impulsionnel étant très bref, il conservera une énergie importante une fois filtré passe-haut, par contre, si la fréquence de coupure du filtre est choisie suffisamment haute, le signal audio sera fortement atténué par le filtrage. L'intérêt de travailler sur le signal filtré est qu'une technique de détection assez simple est suffisante : par exemple, dans [Montresor 90, Valiere 91], un simple seuillage de l'enveloppe du signal filtré est utilisé. La référence [Kinzie 73] présente d'ailleurs un exemple intéressant de cette simplicité puisque que le système de détection est entièrement réalisé avec des composants d'électronique analogique. Là encore, la technique de détection utilisée est assez simple puisqu'il s'agit de comparer une "enveloppe de décroissance", d'une durée d'environ 6 ms, associée à chaque pic détecté dans le signal, avec l'amplitude instantanée du signal afin de décider si le pic correspond ou non à un bruit impulsionnel. Cette seconde méthode s'avère d'ailleurs plus efficace si il subsiste un résidu du signal audio d'amplitude non négligeable dans le signal filtré.

La limitation théorique du filtrage passe-haut est qu'il implique implicitement de travailler avec un signal "suréchantillonné" dans le sens où il faut être capable de trouver dans le haut du spectre une bande de fréquence où le signal est peu présent mais où l'énergie du bruit impulsionnel est suffisamment importante. Toutefois, dans le cas d'enregistrements provenant de disques 78 tours ce problème ne pose pas car le système de lecture moderne est en général plus performant que ne l'était le système d'enregistrement qui a servi à produire le disque original. Les bruits impulsionnels apparaissant à la lecture possèdent donc de l'énergie à des fréquences que le signal audio enregistré n'atteint pas (typiquement au delà de 8 kHz). Il suffit dans ce cas de choisir une fréquence d'échantillonnage un peu plus élevé que la fréquence de Nyquist correspondant au signal (par exemple 24 kHz). Il n'en reste pas moins que les performances de cette méthode sont assez difficiles à évaluer car elles dépendent essentiellement du contenu fréquentiel du signal en haute fréquence. Dans le cas de la technique étudiée par S. Montresor [Montresor 91] [Valiere 91] les résultats obtenus en pratique semble être assez satisfaisants.

S. Montresor [Montresor 91] décrit aussi une technique d'estimation de la durée du bruit impulsionnel à partir du signal filtré passe-haut. Il faut noter que dans ce cas, il est nécessaire de prendre en compte la réponse impulsionnelle du filtre utilisé pour le filtrage passe-haut : l'estimation de la durée du bruit impulsionnel n'est efficace que si cette dernière est nettement plus importante que celle de la réponse du filtre passe-haut. Par ailleurs, S. Montresor montre aussi que l'utilisation d'un banc de filtres (à la place d'un simple filtre passe-haut) permet d'obtenir des résultats intéressants dans le cadre de la détection de bruits impulsionnels. Il semble que le banc de filtres réalisé par certaines transformées en ondelettes soit particulièrement bien adapté pour cette utilisation [Montresor 91][Grossmann 89]. Toutefois, à notre connaissance, cette possibilité n'a pas été explorée de manière plus précise.

A.2.2 Modélisation autorégressive

Une technique apparue plus récemment consiste à travailler à partir du signal résiduel (ou erreur de modélisation) issu de la modélisation autorégressive (ou AR) du signal. L'utilisation de cette méthode pour le cas des bruits impulsionnels a été étudié par S. Vaseghi dans sa thèse [Vaseghi 88b], puis détaillée dans diverses publications [Vaseghi 88a] [Vaseghi 92], la référence la plus complète étant une publication commune avec P. Rayner [Vaseghi 90]. S. Vaseghi propose de réaliser le traitement en deux étapes successives :

1. Calcul du signal résiduel (modélisation AR du signal, puis passage au signal résiduel)
2. Détection par filtrage adapté (filtrage adapté puis détection à seuil mobile)

A.2.2.a Evaluation théorique

La justification de ce principe découle d'une simplification de la situation réelle [Vaseghi 88b] [Vaseghi 88a] [Le May 91] [Laroche 93a] : on suppose, d'une part que le signal utile $s(n)$ (signal audio et bruit de fond) est un processus stationnaire issu d'un modèle autorégressif supposé connu (on notera $1/A(z)$ sa fonction de transfert, et p son ordre), et d'autre part, que le bruit impulsionnel peut être assimilé à une impulsion de Dirac additionnée au signal utile.

"Amplification" du bruit impulsionnel par passage au résiduel Avec les hypothèses ci-dessus, on montre que le bruit impulsionnel est toujours amplifié (par rapport au niveau du signal) par passage au signal résiduel de la modélisation AR [Vaseghi 88b] [Laroche 93a]. Plus précisément,

$$\frac{\text{niveau relatif du bruit impulsionnel dans le signal résiduel}}{\text{niveau relatif du bruit impulsionnel dans le signal original}} \geq \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{|A(\mathrm{e}^{\mathrm{j}\omega})|^2}d\omega \qquad \text{(A.1)}$$

Le niveau relatif du bruit étant défini ici comme la *puissance maximale du bruit impulsionnel divisée par la puissance du signal qu'il affecte.*

Le terme de droite de l'équation (A.1) correspond à l'énergie totale de la fonction de transfert du modèle AR, dont on montre qu'elle est forcément supérieure à 1 (pour les propriétés des modèles AR, se reporter, par exemple, à [Kay 88]). Ce terme d'amplification provient de la relation de filtrage qui existe entre le signal $s(n)$ et le signal résiduel. L'intérêt de travailler sur le signal résiduel $e(n)$ tient essentiellement au fait que dès que la fonction de transfert du modèle AR présente des zones assez résonantes, le gain de l'équation (A.1) devient extrêmement important [Vaseghi 88b]. L'équation (A.1) indique donc que le bruit impulsionnel à détecter est nettement amplifié lors du passage au signal résiduel. En fait, il faut noter qu'il s'agit d'une amplification relative, c'est à dire que le bruit impulsionnel conserve à peu près son amplitude, par contre le signal dans lequel il est noyé voit son amplitude fortement diminuer.

Une interprétation équivalente de ce résultat fait intervenir la notion de prédiction linéaire associé à la modélisation AR [Vaseghi 88b]. En effet, Le signal résiduel représente aussi l'erreur de prédiction à un pas compte tenu des valeurs passées du signal [Kay 88], cette erreur est donc forcément très importante lorsque l'on rencontre un échantillon contaminé par un bruit impulsionnel. Pour un échantillon de signal, l'erreur de prédiction est d'autant plus faible que la corrélation entre valeurs successives de signal est importante [Kay 88], c'est à dire que la réponse du modèle présente des zones fortement résonantes. Par contre, pour un échantillon contaminé, l'erreur de prédiction est quasiment égale à l'amplitude du bruit impulsionnel puisque celui-ci est complètement non-prédictible.

Localisation du bruit impulsionnel par filtrage adapté Lors du passage au signal résiduel, le bruit impulsionnel est convolué par la réponse impulsionnelle du filtre inverse $A(z)$. Ce nouveau bruit impulsionnel est noyé dans le signal résiduel qui est un bruit blanc gaussien dans l'hypothèse ou le signal de départ provient d'un processus autorégressif. Dans ces conditions, la meilleure stratégie de détection du bruit impulsionnel, au sens du rapport de vraissemblance, consiste à utiliser le filtre dit adapté. Dans le cas réel, ce filtre adapté s'obtient simplement en inversant l'ordre temporel des coefficients de la réponse impulsionnelle du filtre $A(z)$ [Van Trees 68] [Charbit 90]. On montre que le gain en détectabilité procuré par filtrage adapté s'écrit[1] [Vaseghi 88b] [Laroche 93a] :

$$\frac{\text{NRB dans le résiduel après filtrage adapté}}{\text{NRB dans le signal résiduel}} = \frac{\sum_i a_i^2}{\max_i \left\{a_i^2\right\}} \tag{A.2}$$

Cette équation traduit donc bien un gain, d'autant plus important que la réponse du filtre inverse $A(z)$ est étalée. En pratique, le principal intérêt du filtrage adapté est de permettre une localisation plus précise du défaut. En effet, l'étalement du bruit impulsionnel dû au filtrage par $A(z)$ lors du passage au résiduel peut rendre difficile la localisation du bruit impulsionnel. Le filtrage adapté permet dans ce cas de déterminer sans ambiguïté la position du bruit impulsionnel [Vaseghi 88b].

A.2.2.b Mise en œuvre et limitations

L'analyse théorique précédente montre que l'utilisation du signal résiduel du modèle AR est une technique extrêmement efficace pour la détection des bruits impulsionnels. En particulier, cette technique est pleinement cohérente avec les résultats concernant l'audibilité des bruits impulsionnels obtenus au paragraphe A.1.2 : dans le cas d'un bruit impulsionnel qui affecte un bruit blanc le gain total de la méthode est nul puisque le modèle est trivial ($A(z) = 1$), la limite de détection est donc obtenue pour une valeur du niveau relatif du bruit impulsionnel de l'ordre de 9 dB (en utilisant la règle des "3σ" [Charbit 90]). Par contre, dans le cas d'un bruit impulsionnel affectant un son pur, le gain de la méthode est théoriquement très important. En pratique, nous verrons que dans ce cas, la limite de détection dépend essentiellement de la méthode de modélisation AR utilisée. Néanmoins, les bruits impulsionnels affectant un son pur avec un niveau relatif supérieur à -40 dB sont toujours aisément détectés. De plus, l'intérêt de la modélisation AR par rapport au filtrage passe-haut est qu'elle fournit des résultats suffisants du point de vue auditif quelle que soit la bande de fréquence occupée par le signal. Il n'est plus nécessaire de travailler sur un signal suréchantillonné comme au paragraphe A.2.

Toutefois, il faut avoir conscience du fait que la situation réelle a été grossièrement simplifié lors de l'analyse théorique précédente, plusieurs points viennent limiter l'efficacité de la méthode lorsqu'elle est appliquée à des signaux issus d'enregistrements réels :

$\rightarrow$ **La puissance du signal résiduel est inconnue** D'après ce qui a été dit précédemment, le gain apporté par la méthode est relatif, c'est à dire que le bruit impulsionnel est mis en évidence essentiellement par un abaissement du niveau du signal qui l'entoure lors du filtrage inverse. Par la suite, l'équation (A.2) indique que l'étape de filtrage adapté entraîne une amplification relative du bruit impulsionnel qui ne dépend que de la réponse impulsionnelle du filtre inverse. Le seuil de détection doit donc être fixé de manière relative par rapport au niveau du

[1] La notation NRB désignant toujours le niveau relatif du bruit impulsionnel, c'est à dire sa puissance maximale divisée par la puissance du signal qu'il affecte.

signal résiduel, qui est inconnu dans une situation réelle. Il est donc nécessaire d'estimer la puissance d'un signal sachant qu'il risque d'être corrompu par des bruits impulsionnels de niveau très important compte tenu du gain de la méthode. En pratique une estimation directe de la puissance donne des résultats toujours très nettement surestimés ce qui réduit les possibilités de détection. Il est donc nécessaire de réduire l'influence des échantillons du signal résiduel affectés par des bruits impulsionnels, c'est à dire en pratique, des valeurs du signal résiduel anormalement grandes [Vaseghi 88b] [Le May 91].

→ **Le bruit impulsionnel n'est pas une impulsion de Dirac** Pour fixer un ordre de grandeur, nous avons vu que la durée moyenne d'un bruit impulsionnel est de l'ordre de 0,5 ms, ce qui représente une dizaine d'échantillons à une fréquence d'échantillonnage de 20 kHz. L'hypothèse d'un défaut strictement impulsionnel est donc assez nettement erronée. Le gain de la première étape (A.1) n'est pas remis en cause, par contre, l'intérêt du filtrage adapté est sujet à caution puisque la forme du bruit impulsionnel dans le signal résiduel ne correspond pas simplement à la réponse impulsionnelle du filtre inverse $A(z)$. Cet effet rend assez difficile l'estimation précise du support du bruit impulsionnel. De même, la séparation de deux bruits impulsionnels proches, c'est à dire distants d'un nombre d'échantillons inférieur à l'ordre du modèle AR peut devenir impossible. La référence [Godsill 92] présente une méthode de détection plus complexe que le simple filtrage adapté qui prend en compte un modèle du bruit impulsionnel plus réaliste que l'impulsion de Dirac.

→ **Le modèle AR du signal doit être estimé** Dans une situation réelle, il est nécessaire d'estimer le modèle AR du signal à partir des données bruitées, le problème étant justement que la présence éventuelle de bruits impulsionnels peut fortement biaiser l'estimation du modèle. Pour les bruits impulsionnels, ce phénomène est assez gênant car il est fréquent de rencontrer des bruits avec un NRB de 10 dB ou plus (voir par exemple la figure A.2). La présence de bruits impulsionnels tend à aplatir le spectre du modèle estimé en limitant la dynamique des zones résonantes et en relevant le niveau des zones de faible niveau. D'après la relation (A.1), ceci limite fortement les possibilités de détection du système. Une solution consiste à pondérer les erreurs de prédiction lors de la modélisation selon que l'on a affaire ou non à un échantillon corrompu par un bruit impulsionnel [Di Claudio 91]. Cette procédure est assez simple à mettre en œuvre lorsqu'une modélisation AR récursive est utilisée puisque la détection des bruits impulsionnels a déjà été effectuée sur tous les échantillons qui sont pris en compte lors de la modélisation [Vaseghi 88b]. Par contre, la situation est plus difficile lorsqu'une modélisation AR en blocs est utilisée car la position des bruits impulsionnels n'est pas connue à priori.

Indépendamment de ce problème lié à la présence des bruits impulsionnels, la nécessité d'effectuer une modélisation AR du signal implique des choix (ordre du modèle, méthode de modélisation) qui modifient les performances théoriques. Pour citer des exemples, dans [Vaseghi 88b] la technique retenue est la version adaptative de l'algorithme des moindres carrés récursif (*Recursive Least-Square* en anglais) qui correspond plutôt à la méthode de modélisation dite de covariance (*Least-Square* en anglais), tandis que la méthode dite de corrélation (*Yule-Walker* en anglais) implémentée en blocs est utilisée dans [Le May 91]. Comme nous le verrons au paragraphe A.2.2.c, l'évaluation de ces différents systèmes n'est donc plus uniquement régie par les équations (A.1) et (A.2) mais elle implique également les caractéristiques propres des méthodes de modélisation.

→ **Le signal n'est pas AR !** Pour la plupart des signaux audio, l'hypothèse d'un signal AR n'est elle-même pas vérifiée. En particulier, pour des sons de parole voisés, il est bien connu que le signal résiduel n'est en général pas un bruit blanc mais plutôt une suite d'impulsions quasi périodiques : c'est le principe de la LPC [Makhoul 75]. Pour beaucoup de sons musicaux, considérés dans leur partie quasi stationnaire, la situation est assez semblable, voir pire puisque les sons musicaux possèdent en général un très grand nombre de composantes spectrales. Seuls les sons de fréquence fondamentale très élevée donnent un signal résiduel relativement exempt de structure périodique [Depalle 91]. En pratique, la structure exacte du signal résiduel est donc assez complexe ce qui explique que l'estimation du seuil de détection des bruits impulsionnels soit en fait un point assez délicat et particulièrement sensible. D'ailleurs, une remarque sur ce point est qu'un seuil réglé à un niveau trop bas se traduit en général par de fausses détections quasi périodiques du fait de la structure du résiduel. Après interpolation (voir le paragraphe A.3), le résultat auditif produit est assez caractéristique est peut être utilisé pour régler empiriquement le seuil de détection.

Des résultats malgré tout Compte tenu de tout ce qui vient d'être dit, il est légitime de se demander dans quelle mesure la détection de bruits impulsionnels à partir d'une modélisation AR peut fonctionner. En fait, dans la pratique cette méthode s'avère extrêmement efficace et surtout assez robuste. Ce succès vient essentiellement du fait que le gain obtenu par filtrage inverse est assez important, quelle que soit la nature du signal, du moment que la densité spectrale du signal analysé présente des pics suffisamment marqués. Le résultat est donc conditionné par l'estimation spectrale associé à la modélisation AR plus que par l'adéquation du modèle au signal. En particulier, le choix de l'ordre du modèle est un paramètre assez secondaire : du moment que les principaux pics spectraux du signal sont mis en évidence, la variation de l'ordre n'entraîne plus de modification importante, particulièrement compte tenu de l'effet décrit au paragraphe suivant.

A.2.2.c Influence du filtre adapté

Une conséquence assez importante du fait que le signal résiduel ne soit pas un bruit blanc est que l'étape de filtrage adapté peut produire des résultats bien différents de la théorie telle quelle est décrite par la relation (A.2). A notre connaissance cet effet n'a pas été mentionné dans la littérature, nous allons donc le détailler assez brièvement sur un exemple. Tout d'abord, la limite de détection associée au filtre adapté ne dépend théoriquement que de l'énergie du signal à détecter, ici le bruit impulsionnel, et de la puissance du signal dans lequel il est noyé, ici le signal résiduel [Van Trees 68]. Malheureusement, ce résultat devient faux lorsque le signal résiduel n'est pas un bruit blanc, on constate alors que la réponse fréquentielle du filtre adapté, ici $|A(\mathrm{e}^{\mathrm{j}\omega})|$, se met à avoir une influence. Pour illustrer cette propriété, nous avons effectué la simulation suivante :

- Partant d'une portion de son de piano issu d'une note du registre médium-grave (A3 en notation américaine, soit une fréquence fondamentale à environ 220 Hz), on y additionne un bruit impulsionnel idéal, ici une impulsion numérique, de NRB -40 dB. Ce bruit impulsionnel n'est donc pas visible sur la forme d'onde.

- Le traitement de détection complet (passage au résiduel puis filtrage adapté) est réalisé en utilisant deux méthodes de modélisation AR différentes : la méthode de corrélation (notée COR), et la méthode de covariance (notée COV). Les paramètres retenus pour la modélisation sont une durée de trame de 500 échantillons, soit 20 ms compte tenu de la fréquence d'échantillonnage, et un ordre de 20 pour le modèle. Ces valeurs correspondent à des choix usuels de paramètres [Vaseghi 88b] [Le May 91].

A partir des modèles estimés, on calcule les gains théoriques grâce aux relations (A.1) et (A.2) dans les deux cas. Les résultats sont :

Pour la méthode de corrélation	**Pour la méthode de covariance**
23 dB pour le passage au résiduel	52 dB pour le passage au résiduel
4 dB pour le filtrage adapté	11 dB pour le filtrage adapté
soit un total de 27 dB	soit un total de 63 dB

Ces résultats sont confirmés par les réponses fréquentielles des modèles AR estimés représentées sur la figure A.5. On constate bien que le modèle obtenu avec la méthode de covariance possède des résonances beaucoup plus marquées, ce qui est une caractéristique connue de la méthode [Kay 88]. Par conséquent le gain obtenu lors du passage au signal résiduel est beaucoup plus important, ici de près de 30 dB, avec la covariance qu'avec la corrélation. On note au passage que le gain obtenu lors du filtrage adapté est aussi légèrement supérieur avec la méthode de covariance, mais cette différence reste secondaire. D'après ces résultats théoriques, le bruit impulsionnel devrait être largement détecté lorsque la covariance est utilisée, et par contre, complètement indétectable avec la méthode de corrélation.

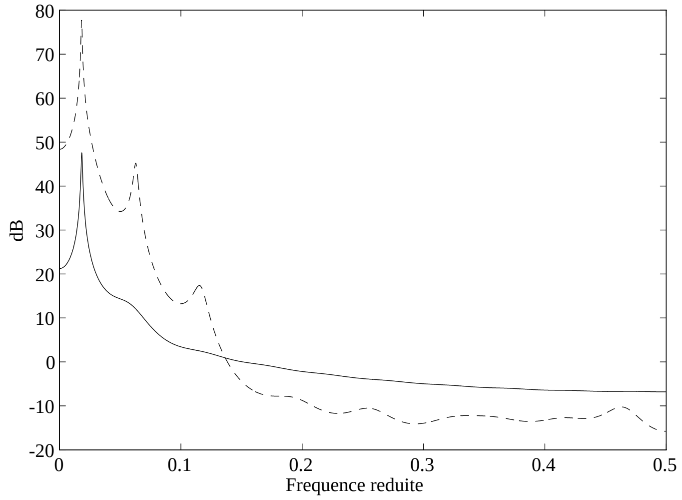

Figure A.5: Réponse en fréquence de deux modèles AR du même signal. **En trait plein,** la méthode de corrélation. **En pointillés,** la méthode de covariance.

Le signal résiduel ainsi que le résultat du filtrage adapté sont représentés pour chacune des deux méthodes sur la figure A.6. **La colonne de droite** qui concerne le cas de la covariance est relativement conforme aux prévisions de la théorie : le niveau du signal résiduel se situe environ 60 dB au dessous de celui du signal original ce qui permet largement de détecter l'impulsion, et le résultat du filtrage adapté traduit bien une légère augmentation relative du niveau du bruit impulsionnel. Par contre le cas de **la colonne de gauche** qui représente les résultats obtenus avec la méthode de corrélation échappe complètement à la théorie. Pour le signal résiduel le résultat est à peu près correct puisque le niveau de celui-ci est bien 20 à 30 dB en dessous de celui du signal analysé. Par contre, on constate que le filtrage adapté produit un gain tout à fait étonnant qui permet même de détecter le bruit impulsionnel, c'est à dire que le gain relatif du filtrage adapté est ici d'au moins une vingtaine de décibels.

Ces résultats étonnants sont liés au fait que l'hypothèse de signal AR est tout à fait erronée, surtout dans le cas du modèle obtenu par la méthode de corrélation. On constate en effet que le signal résiduel associé au modèle COR n'est absolument pas blanc (**colonne de gauche,** en

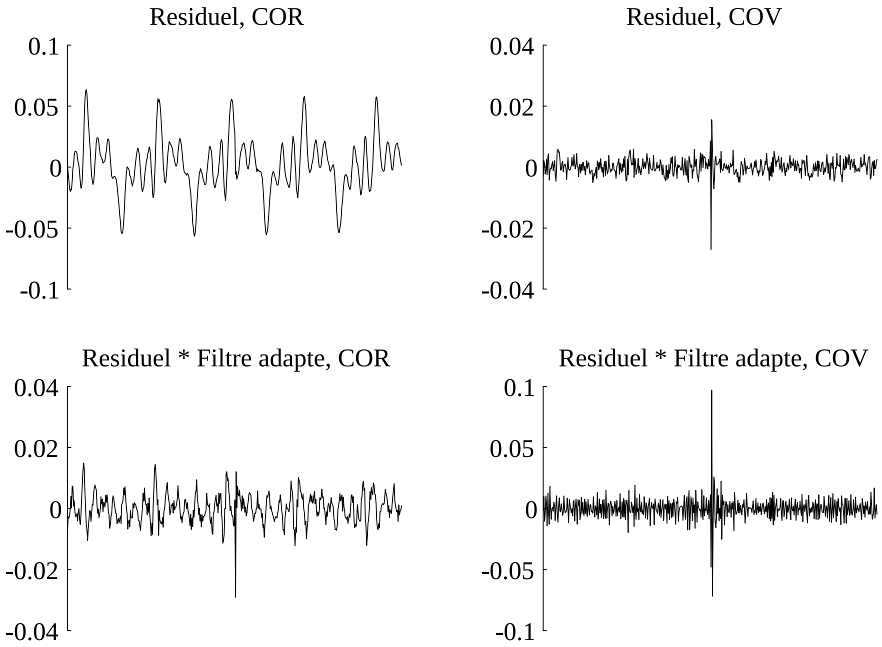

Figure A.6: Signal résiduel (en haut), et résultat du filtrage adapté (en bas). **A gauche,** le cas de la méthode de corrélation. **A droite,** le cas de la méthode de covariance. L'amplitude de tous les signaux est normalisée par rapport à l'écart type du signal analysé.

haut sur la figure A.6) : la corrélation ne permettant pas de mettre en évidence tous les pics du spectre de niveau important, le signal résiduel est en fait assez proche du signal original, et en particulier il est clairement périodique. **Le rôle du filtre adapté est ici complètement détourné : il agit comme un filtre passe-haut** qui pourrait d'ailleurs être remplacé par un filtre indépendant du signal comme dans la méthode du paragraphe A.2.1. La preuve en est que le niveau *absolu* du signal résiduel diminue lors du filtrage adapté alors qu'il augmenterait forcément dans le cas du bruit blanc (toujours car l'énergie totale du filtre inverse est supérieure à 1). Le gain obtenu lors du filtrage adapté est donc très important uniquement parce que le filtre adapté permet ici d'éliminer une grande partie des résidus du signal analysé. Le caractère fortement passe-haut du filtre adapté (filtre inverse “retourné”) est garanti en pratique par le fait que les signaux audio ont en général le maximum de leur énergie dans la bande basse du spectre.

La conséquence du phénomène illustré par la figure A.6 est que l'utilisation de la méthode de corrélation fournit toujours des résultats assez proches de ceux de la méthode de covariance, bien qu'en général plutôt inférieurs. La différence entre ces deux méthodes tend d'ailleurs à s'estomper lorsque le niveau de bruit de fond augmente, puisque les modèles obtenus avec la covariance tendent alors à devenir nettement moins résonants. Pour illustrer l'influence du bruit de fond sur les résultats, nous avons réalisé la même simulation, avec la méthode de covariance, en additionnant au préalable du bruit blanc au signal analysé avec un rapport signal-à-bruit de -40 dB. Les paramètres de modélisation étant inchangés, on constate sur la figure A.7 que la présence de bruit, même à niveau faible, a fortement modifiée l'estimation spectrale obtenue par la méthode de covariance. L'aplatissement du spectre estimé se traduit par une forte baisse du gain lors du passage au résiduel. Par rapport au cas de la figure A.5, on obtient une élévation du seuil de détection d'environ 14 dB.

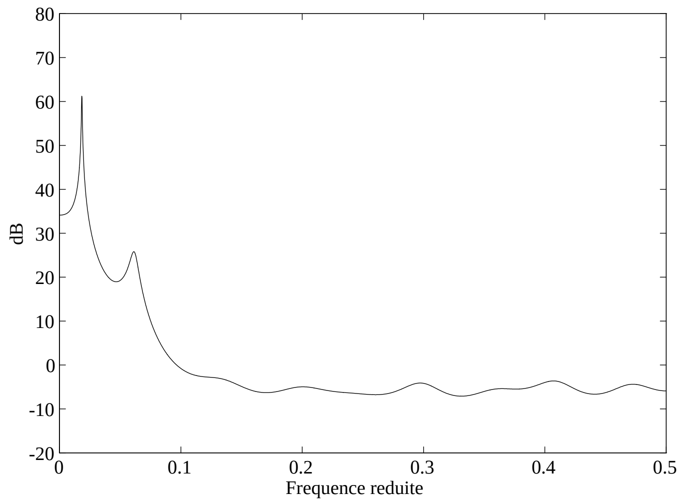

Figure A.7: Réponse en fréquence du modèle AR du signal bruité estimé par la méthode de covariance (rapport signal-à-bruit de -40 dB).

L'idée à retenir de cet exemple est que le filtre adapté joue un rôle très différent de celui qui lui est assigné par la théorie lorsque les hypothèses du modèle AR deviennent ouvertement erronées. Nous avons utilisé comme exemple le cas de la méthode de corrélation avec un signal de fréquence fondamentale assez basse. Toutefois on obtient le même type de comportement avec la covariance lorsque l'ordre du modèle est très faible ou quand la fréquence fondamentale du son analysé est très basse. L'aspect positif est que dans ces cas, le filtre adapté améliore nettement les possibilités de détection en agissant comme un filtre passe-haut qui élimine les résidus de signal. Une conséquence pratique est que la méthode de corrélation peut donc être utilisée, avec l'intérêt d'un coût de calcul nettement plus faible [Kay 88]. Par rapport à la méthode de covariance, qui demeure toutefois mieux adaptée à cet usage, on ne constate qu'une dégradation modérée des performances, et ce du fait de l'action du filtre adapté. De plus, la différence entre ces deux méthodes de modélisation devient de moins en moins importante lorsque l'on traite des enregistrements fortement bruités.

A.3 Correction des bruits impulsionnels

A.3.1 Aperçu des méthodes d'interpolation

A notre connaissance, la référence la plus complète sur ce sujet est la thèse de R. Veldhuis [Veldhuis 88] intitulée *"Adaptive restoration of unknown samples in discrete time signals and digital images"*[2]. Dans son introduction, R. Veldhuis effectue une revue de la littérature qui met en évidence quelques grandes classes de méthodes connues :

- La substitution de forme d'onde, approche heuristique qui consiste à exploiter la périodicité supposée du signal pour dupliquer le signal dans les parties manquantes.
- La reconstruction de signaux à bande-limitée qui a donné lieu a beaucoup de littérature

[2]Il existe aussi une version remaniée de la thèse parue sous forme d'un livre [Veldhuis 90], ainsi qu'une publication très détaillée pour le cas des signaux autorégressifs [Jansen 86].

mais qui s'applique plutôt mal aux cas des signaux audio qui occupent en général une large bande de fréquence.

- La restauration statistique qui exploite des propriétés statistiques du signal (par exemple un modèle d'état) supposées connues.

L'apport de R. Veldhuis est de présenter dans un cadre unifié (l'estimation linéaire à variance minimale) plusieurs méthodes d'estimation des échantillons manquants selon les hypothèses qu'il est légitime de faire sur le signal. L'expression obtenue pour l'estimateur général des données manquantes fait intervenir la matrice d'autocorrélation du signal observé. L'estimation de cette matrice complète étant irréalisable (l'estimateur obtenu n'est pas consistant), les autres estimateurs proposées correspondent à des cas particuliers de signaux pour lesquels la matrice de corrélation présente des propriétés particulières qui rendent son estimation complète inutile [Veldhuis 88]. Pour le cas des signaux audio, la méthode préconisée consiste à utiliser l'hypothèse d'un signal auto-régressif, cette méthode qui a été reprise par la suite [Vaseghi 88b] [Valiere 91], est actuellement la plus utilisée dans les systèmes de restauration d'enregistrements. Il faut noter que R. Veldhuis présente aussi une simplification de cette méthode pour le cas des signaux de parole où l'hypothèse supplémentaire effectuée est que le signal est quasi-périodique. Enfin, un point intéressant concerne la méthode de restauration reposant sur l'hypothèse d'un signal constitué d'une somme de sinusoïdes. Dans ce cas, la matrice d'autocorrélation du signal est de rang incomplet, ce qui permet théoriquement d'obtenir une interpolation des données sans erreur [Veldhuis 88]. Malheureusement, les performances de cette méthode se dégradent très vite en présence de bruit de fond, et sont très sensibles au nombre supposé de sinusoïdes dans le signal. Ceci rejoint, des observations connues concernant la modélisation sinusoïdale des signaux musicaux [Laroche 89].

Un article récent [Godsill 93] présente une autre approche extrêmement intéressante de ce problème d'interpolation. Le principe est que si l'on dispose d'une estimation de la densité spectrale du signal, il est possible, sous certaines hypothèses sur le signal, de spécifier la distribution de probabilité de la transformée de Fourier discrète d'une portion du signal (c'est exactement le même modèle qui est utilisé pour la réduction du bruit de fond dans [Ephraim 84]). L'interpolation consiste donc à utiliser cette information a priori dans un cadre bayésien, afin d'obtenir l'estimation des données manquantes qui maximise la densité de probabilité a posteriori. Cette méthode d'interpolation se traduit par une charge de calcul strictement équivalente à la méthode utilisant un modèle AR du signal qui est décrite au paragraphe A.3.2.a. Par contre, les auteurs indiquent que cette nouvelle méthode fournit de meilleurs résultats lorsque la zone à interpoler est très longue. Toutefois, il reste un point délicat non précisé dans l'article qui est la sensibilité de l'estimateur vis à vis des erreurs d'estimation de la densité spectrale de puissance du signal. En effet, on peut penser que la qualité de l'estimation de la DSP (liée à la taille de l'observation), ainsi que le type d'estimateur utilisé (en particulier la résolution spectrale) risquent d'influencer fortement les résultat de l'interpolation.

Enfin, il est assez souvent fait référence à des techniques de "filtrage non-linéaire" dont la plus célèbre est le filtrage médian [Gallagher 81]. Ces techniques connues depuis une vingtaine d'années sont utilisées avec succès, en particulier dans le domaine du traitement d'image, pour éliminer les nuisances impulsionnelles (on trouvera un exemple d'application dans [Pasian 84]). Malheureusement, l'efficacité du filtrage médian repose sur le fait qu'il élimine les bruits impulsionnels tout en préservant les contours (*edges* en anglais) du signal. Or cette notion de contour s'applique mal aux signaux audio qui ne présentent en général pas de transitions abruptes. En conséquence, l'utilisation du filtrage médian sur des signaux audio produit des distorsions très importantes du signal. Les publications [Nieminen 87b] et [Nieminen 87a] présentent une solution dans laquelle le filtrage médian est appliqué sur des données composites provenant à la fois du signal corrompu par les bruits impulsionnels et de la modélisation de ce signal obtenue grâce à un filtre AR adaptatif. Les auteurs revendiquent de bons résultats pour les signaux de parole ainsi que pour les signaux musicaux. Malheureusement, compte tenu de la complexité du système due à la présence de la non-linéarité, l'évaluation des performances du système demeure purement empirique. Une remarque est que l'efficacité de ce système repose sur le fait que le filtre adaptatif permet de réaliser une prédiction linéaire des valeurs du signal sauf dans le cas où celles-ci sont corrompues par un bruit impulsionnel. On retrouve là le principe de la détection par modélisation AR du paragraphe A.2.2.

A.3.2 Modélisation autorégressive

Cette partie décrit la mise en œuvre de la technique qui apparaît, à ce jour, comme étant la plus efficace pour la correction des défauts localisés sur les signaux audio.

A.3.2.a Formule d'interpolation

Les hypothèse utilisées sont les suivantes :

- Comme au paragraphe A.2.2, on suppose que le signal observé $s(n)$ est issu d'un modèle AR connu $A(z)$, d'ordre p, dont les coefficients s'écrivent $a_0, a_1, \ldots, a_p$ (avec $a_0 = 1$). On note b_n la fonction d'autocorrélation des coefficients du modèle définie par

$$b_n = \sum_{i=-p}^{p} a_i a_{i+n}$$

- De plus, on suppose que les positions des l échantillons à interpoler sont connues, on note $n(1), n(l) \ldots, n(l)$ les indices de ces échantillons. Par ailleurs, on désigne par $x_{\mathrm{mod}}(n)$ le signal dans lequel les échantillons à interpoler ont été remplacés par des valeurs nulles.

Avec ces notations, l'estimation du vecteur $\hat{\mathbf{s}} = (s(n_1), \ldots, s(n_l))^t$ des échantillons à interpoler s'écrit [Veldhuis 88] [Jansen 86]

$$\mathbf{G}\hat{\mathbf{s}} = -\mathbf{z} \tag{A.3}$$

$$\text{où} \quad g_{i,j} = b_{n(j)-n(i)} \quad \text{pour } i, j = 1, \ldots, l \tag{A.4}$$

$$\text{et} \quad z_i = \sum_{k=-p}^{p} b_k x_{\mathrm{mod}}\left(k - n(i)\right) \quad \text{pour } i = 1, \ldots, l \tag{A.5}$$

L'erreur quadratique d'interpolation $\mathrm{E}\left\{|s(n) - \hat{s}(n)|^2\right\}$ est minimale pour l'estimateur donné par (A.3) car il correspond à l'estimation linéaire à variance minimale des données manquantes. Cette erreur est d'autant plus faible que le modèle AR du signal présente des zones fortement résonantes. Malheureusement il n'existe pas de relation simple, du type de (A.1), permettant de mettre en évidence ce phénomène pour un modèle AR quelconque [Jansen 86] [Veldhuis 88].

Dans le cas général, la procédure d'interpolation décrite par la relation (A.3) est assez lourde à mettre en œuvre puisque la résolution du système implique l'inversion d'une matrice de dimension $l \times l$. Toutefois, on remarque que si tous les échantillons à interpoler sont consécutifs, les coefficients de la matrice $\mathbf{G}$ décrite par l'équation (A.4) vérifient

$$g_{i,j} = b_{j-i}$$

c'est à dire que la matrice $\mathbf{G}$ possède alors une structure de Toeplitz [Veldhuis 88] [Vaseghi 88b]. On sait que l'inversion de $\mathbf{G}$ peut alors être réalisée de manière beaucoup moins coûteuse grâce à l'algorithme de Levinson [Press 92]. Une autre remarque importante est que dans l'équation (A.5), seuls interviennent les p échantillons situés avant le premier point à interpoler (d'indices $n(1)-p$ à $n(1)-1$) et ceux situés après le dernier point à interpoler (d'indices $n(1)+1$ à $n(1)+p$). Par conséquent, **dès que les zones de signal à interpoler sont séparées d'un nombre d'échantillons au moins égal à l'ordre du modèle AR, il est plus rapide de réaliser séparément les interpolations sur chacune des zones** puisque que pour chaque zone à interpoler, on se ramène alors au cas d'échantillons consécutifs. Il n'en reste pas moins que dans le cas où plusieurs bruits impulsionnels sont très proches, il est nécessaire d'utiliser une technique générale de résolution du système de l'équation (A.3) (qui passe par une décomposition de Cholevsky de la matrice $\mathbf{G}$) [Veldhuis 88] [Valiere 91]. On note que cette éventualité est prise en compte dans [Valiere 91] mais pas dans [Vaseghi 88b]. On peut en effet se demander si ce cas est bien réaliste dans le cadre de l'application à la correction des bruits impulsionnels. Il faut savoir que le travail de Veldhuis était guidé par une application assez différente[3] pour laquelle la position des zones à interpoler est parfaitement connue. Par contre, dans le cas des bruits impulsionnels, nous avons vu au paragraphe A.2.2.b que la séparation de deux bruits impulsionnels proches n'est pas une tâche aisée. Il faudrait donc faire une étude plus précise pour savoir si une mauvaise localisation des zones à interpoler n'est pas plus gênante que la présence éventuelle de bruits impulsionnels au voisinage de la zone à interpoler. Une autre solution consiste à utiliser une technique permettant d'obtenir une localisation plus précise des bruits impulsionnels [Godsill 92].

A.3.2.b Application au cas des bruits impulsionnels

Ici encore, la technique d'interpolation repose sur l'hypothèse que le modèle AR du signal est connu. En pratique, il est à la fois nécessaire d'estimer les paramètres du modèle AR et les valeurs du signal dans les zones à interpoler. La solution étudiée par R. Veldhuis consiste en une procédure itérative où on réalise, à chaque étape, la modélisation AR du signal suivie de l'estimation des données manquantes [Veldhuis 88]. Toutefois, pour le traitement des bruits impulsionnels, une seule itération est en général appliquée [Vaseghi 88b] [Valiere 91]. Il s'agit là encore d'une distinction importante entre le traitement des bruits impulsionnels et l'application originelle des travaux de R. veldhuis : **dans le cas de la restauration une erreur d'interpolation de l'ordre de l'écart-type du signal résiduel est largement acceptable**. En effet, le principe de la détection par modélisation AR est justement que seuls les bruits impulsionnels de niveau supérieur à l'écart-type du signal résiduel peuvent être détectés. D'après les résultats du paragraphe A.2.2, il semble bien que cette limite de détection soit compatible avec les seuils d'audibilité des bruits impulsionnels dans la plupart des cas. Nous pouvons donc conclure qu'une erreur d'interpolation de l'ordre de l'écart-type du signal résiduel est inaudible du fait de la différence d'échelle très importante qui existe entre le signal analysé et le signal résiduel. Lorsque le nombre de points successifs à interpoler reste faible (dans [Veldhuis 88] la limite empirique de $l < p/3$ est proposée, ce qui correspond en général à une dizaine d'échantillons), on

[3]La thèse de R. Veldhuis ayant été réalisée dans les laboratoires de la compagnie Philips à Eindhoven, l'application consistait, entre autre, à mettre au point un système performant pour le masquage (*concealement* en anglais) des erreurs non-corrigées sur les disques compacts. Dans ce cas, la position exactes des zones à interpoler est connue, puisqu'elle est transmise par le code correcteur d'erreur.

constate que la variance de l'erreur d'interpolation est très faible, et en général inférieure à la puissance du signal résiduel, même à la première itération.

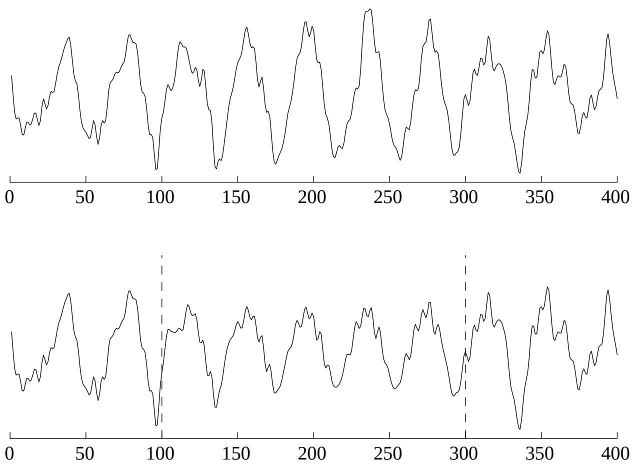

Figure A.8: Interpolation d'un signal AR sur 200 échantillons. **En haut,** le signal original. **En bas,** le signal interpolé, les pointillés indiquent la zone de signal interpolée.

En pratique, on ne se trouve que très rarement dans ce cas puisque 10 échantillons à une fréquence d'échantillonnage de 20 kHz ne représentent que 0,5 ms. Pour des longueur de zones à interpoler plus importante, il semble que la convergence de l'algorithme itératif ne soit pas assurée [Veldhuis 88]. De plus, dès que la longueur l de la zone interpolée atteint 30 à 50 échantillons, on obtient des valeurs de la variance de l'erreur d'interpolation comparables à la puissance du signal lui-même. Toutefois le succès de cette méthode d'interpolation vient de ce que la minimisation de l'erreur quadratique de modélisation équivaut aussi à minimiser une distance spectrale (liée à la distance d'Itakura [Basseville 89]) entre le signal inconnu et le signal interpolé [Veldhuis 88]. En pratique, **même si l'énergie de l'erreur d'interpolation est importante, cette dernière demeure inaudible en vertu du phénomène de masquage fréquentiel simultané car elle à la même composition fréquentielle que le signal audio**. C'est ce qui explique que l'interpolation de zones de signal allant jusqu'à 100 échantillons successifs soit quasiment inaudible pour beaucoup de signaux audio [Jansen 86] [Rayner 91]. Les figures A.8 et A.9 illustrent cette propriété pour l'interpolation longue (sur 200 échantillons) d'un signal AR dont le modèle est connu (le modèle en question est d'ordre 30 et présente des zones assez résonantes puisque le facteur de l'équation (A.1) vaut environ 35 dB). La figure A.8 montre bien que l'erreur d'interpolation est très importante dans ces conditions puisque la forme d'onde du signal interpolé diffère fortement de celle du signal original. Toutefois, la figure A.9 indique que le spectre de l'erreur de prédiction est proche de celui du signal, et ce au sens de la distance d'Itakura, c'est à dire que les zones fréquentielles de forte puissance sont les mêmes pour les deux spectres, et de même pour les zones de faible puissance.

La valeur de 100 échantillons représente 5 ms de signal à une fréquence d'échantillonnage de 20 kHz, c'est à dire qu'elle elle largement suffisante pour les bruits impulsionnels, par contre elle peut ne pas suffire pour le cas des craquements. Lorsque le signal est interpolé sur une zone encore plus longue, il apparaît souvent un "trou" de faible puissance au milieu de la zone interpolée, surtout si le modèle est peu résonant [Rayner 91]. Cette baisse locale de puissance du signal audio devient perceptible, et il est alors nécessaire de modifier légèrement la technique d'interpolation pour remédier à ce phénomène [Rayner 91]. La référence [Vaseghi 90] présente

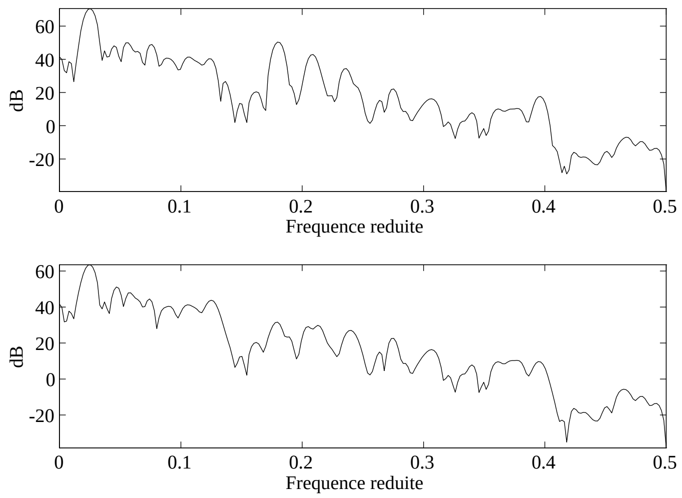

Figure A.9: Comparaison entre le spectre du signal et le spectre de l'erreur d'interpolation pour un signal AR. **En haut,** le spectre du signal original. **En bas,** le spectre de l'erreur d'interpolation. Les deux spectres sont calculés par transformée de Fourier sur les 200 échantillons correspondant à la zone interpolée.

une autre solution à ce problème, utilisable dans le cas des signaux de parole, qui incorpore explicitement l'information de fréquence fondamentale du signal.

Annexe B

Transformée de Fourier à court-terme

Cette annexe est consacrée aux propriétés théoriques de l'ensemble analyse/modification/synthèse dans le cadre de la transformée de Fourier à court-terme (TFCT). Le but est d'obtenir une description équivalente du traitement par atténuation spectrale à court-terme pour des situations simples. Les résultats obtenus sont, entre autres, utilisés au chapitre 3 lors de l'évaluation des techniques de débruitage.

Nous avons choisi de présenter deux conventions d'écriture de la TFCT, en insistant à chaque fois sur les différentes interprétations de la procédure d'analyse/synthèse. Le paragraphe B.2 propose une description équivalente de la modification spectrale effectuée sur les transformées à court-terme, sous la forme d'un *filtrage linéaire variant dans le temps* appliqué au signal. Enfin, le paragraphe B.3.1 présente quelques remarques sur la réalisation pratique d'un système de traitement à court-terme utilisant la TFCT.

B.1 Définition(s) de la TFCT

Les deux écritures de la transformée de Fourier à court-terme considérées sont :

La convention passe-bas (ou **référence temporelle fixe**) qui est la notation de la TFCT se prêtant le plus facilement aux calculs du fait de sa simplicité. C'est aussi l'écriture la plus classique. Cependant cette convention présente le défaut de ne pas correspondre exactement à la manière dont la TFCT est réalisée en pratique.

La convention passe-bande (ou **référence temporelle glissante**) qui traduit fidèlement la manière dont la transformation est implémentée en pratique, mais devient par contre beaucoup plus lourde lorsqu'il s'agit d'effectuer des calculs formels.

Comme il existe une correspondance assez simple entre les deux conventions, il est intéressant d'utiliser l'une ou l'autre selon les cas. Bien que la convention passe-bande soit la plus "correcte" puisqu'elle correspond au traitement réellement effectué, la notation passe-bas est utilisée, lorsque c'est possible, afin de simplifier les calculs.

D'autre part, quelle que soit la convention adoptée pour définir la TFCT, il existe deux *interprétations* qui permettent d'illustrer différemment le fonctionnement de la transformée de Fourier à court-terme :

L'analogie avec une transformée en blocs est celle qui apparaît naturellement en considérant la TFCT comme une suite de spectres à court-terme.

L'analogie avec un banc de filtres qui permet de décrire de manière simple le comportement des **signaux de sous-bande** constitués par les valeurs successives de la transformée à court-terme en un point fréquentiel fixé.

Bibliographie et notations

Parmi les références bibliographiques concernant la transformée de Fourier à court-terme, la plus complète est le livre de Crochiere et Rabiner [Crochiere 83] qui traite du problème plus général des bancs de filtres multi-cadence (c'est à dire où les signaux de sous-bande ne sont pas échantillonnés à la même fréquence que le signal analysé). Pour une présentation moins détaillée de la TFCT, on peut se reporter à [Nawab 88] ou à [Allen 77]. Enfin, l'interprétation du spectre de Fourier à court-terme (ou plutôt du *spectrogramme*) en tant que représentation temps-fréquence est présentée dans [Cohen 89].

Les principales notations utilisées pour la TFCT sont les suivantes :

- **h**(**n**) Fenêtre (ou filtre) d'analyse, de support temporel $[-(N-1), 0]$.
- **N** Longueur (en d'échantillons) de la fenêtre d'analyse.
- **f**(**n**) Fenêtre (ou filtre) de synthèse.
- **R** Pas de décalage des fenêtres (facteur de sous-échantillonnage des signaux de sous-bande).
- Enfin, pour simplifier l'écriture des formules, on note
$$\omega_k = 2\pi k/N \ \text{ pour } k = 0, \ldots, N-1$$
et
$$W_N = \mathrm{e}^{\mathrm{j}\frac{2\pi}{N}} = \mathrm{e}^{\mathrm{j}\omega_1}$$

B.1.1 Convention passe-bas

Avec cette convention, les voies du banc de filtres équivalent sont des signaux complexes en bande de base (d'où l'appellation). Dans l'analogie de la transformée en bloc, cette convention se traduit par une **référence fixe** (ou absolue) [Crochiere 83]. C'est à dire que toutes les trames de signal sont référencées par rapport à l'origine du signal analysé. De manière équivalente, on peut dire que c'est la fenêtre d'analyse qui glisse (dans le sens des indices temporels croissants) le long du signal analysé (qui lui reste fixe). C'est en général sous cette forme que la transformée de Fourier à court-terme est définie [Crochiere 83] [Portnoff 81a] [Bourdier 88]. La définition générale de la TFCT d'un signal à temps discret s'écrit dans cette convention [Nawab 88] :

- A l'analyse,

$$X(m,\omega) = \sum_{n=-\infty}^{+\infty} h(m-n)x(n)\exp(-\mathrm{j}\ \omega n) \tag{B.1}$$

- A la synthèse,

$$y(n) = \sum_{m=-\infty}^{+\infty} f(n-m)\frac{1}{2\pi}\int_{-\pi}^{\pi} Y(m,\omega)\exp(\mathrm{j}\ \omega n)d\omega \tag{B.2}$$

Où $h(n)$ représente la fenêtre d'analyse, tandis que $f(n)$ est appelé fenêtre de synthèse. La portion de signal pondéré par la fenêtre d'analyse, $h(m-n)x(n)$, est appelée **trame de signal à court-terme** d'indice m. L'équation (B.1) indique que le spectre à court terme référencé par m est la transformée de Fourier de la trame de signal à court-terme d'indice m.

Discrétisation de l'axe fréquentiel Par suite, on s'intéresse aux cas où la fenêtre d'analyse est de longueur finie N (en nombre d'échantillons). Par convention, le support de la fenêtre $h(n)$ est ici choisi égal à $[-(N-1), 0]$. D'après la relation (B.1), avec cette convention le premier spectre à court-terme $X(0,\omega)$ résulte de l'analyse du signal sur l'intervalle $[0, N-1]$. Par contre, la fenêtre $f(n)$ est supposé causale : la première valeur synthétisée $y(0)$ ne dépend que des spectres à court-terme précédents $[Y(m,\omega)]_{m\leq 0}$. Cette convention sur le support des fenêtres permet de retrouver la solution généralement utilisée en pratique où les spectres à court-terme sont référencés par l'indice du début de la trame de signal analysée [Bourdier 88] [Serra 89].

L'équation (B.1) montre que la transformée de Fourier porte alors sur un signal de durée finie N, on sait qu'il suffit dans ce cas d'évaluer la transformée de Fourier pour les N pulsations discrètes $\omega_k = 2\pi k/N$ pour $k = 0,\ldots,N-1$ (c'est le principe de la transformée de Fourier discrète [Delmas 91]). Il est tout à fait possible de réaliser la TFCT avec un nombre de points d'échantillonnage de la pulsation numérique différent de la longueur de la fenêtre d'analyse [Allen 77] [Crochiere 83]. Cependant, seul le cas où $\omega_k = 2\pi k/N$ nous intéressera, car il correspond à l'échantillonnage minimal sans perte d'information. La TFCT échantillonnée en fréquence s'écrit alors[1] :

- A l'analyse,

$$X(m,\omega_k) = \sum_{n=-\infty}^{+\infty} h(m-n)x(n)W_N^{-kn} \tag{B.3}$$

- A la synthèse,

$$y(n) = \sum_{m=-\infty}^{+\infty} f(n-m)\frac{1}{N}\sum_{k=0}^{N-1} Y(m,\omega_k)W_N^{kn} \tag{B.4}$$

Il faut noter que toutes les sommes ne portent en fait que sur un nombre fini de points (puisque les fenêtres sont de durée finie). L'écriture sous forme de sommes infinies évite de surcharger les notations avec des indices inutiles. Pour mettre en évidence l'analogie de la TFCT avec un banc de filtres, les équations (B.3) et (B.4) peuvent se mettre sous la forme :

- Pour l'analyse,

$$X(m,\omega_k) = [x(m)\exp(-\mathrm{j}\ \omega_k m)] * h(m) \tag{B.5}$$

[1] En utilisant la notation $W_N^k = \exp(\mathrm{j}\ \omega_k)$.

- Et à la synthèse,

$$y(n) = [Y(n, \omega_k) * f(n)] \exp(\mathrm{j}\ \omega_k n) \tag{B.6}$$

L'équation (B.5) montre que le signal démodulé, $x(m)\exp(-\mathrm{j}\ \omega_k m)$, est filtré par $h(n)$ pour obtenir le signal de sous-bande $X(m, \omega_k)$. Si on souhaite que la TFCT puisse s'interpréter comme le résultat d'une analyse fréquentielle, il est nécessaire que la fenêtre $h(n)$ soit la *réponse impulsionnelle d'un filtre passe-bas*. La valeur de $X(m, \omega_k)$ rend alors compte uniquement de ce qui se passe "au voisinage" de la fréquence centrale de la sous-bande ω_k. Pour une interprétation plus rigoureuse du rôle de la fenêtre d'analyse, il faudrait faire intervenir la notion de représentation temps-fréquence [Cohen 89]. Du fait du caractère passe-bas de $h(n)$, le signal de sous-bande $X(m, \omega_k)$ est bien un signal en bande de base.

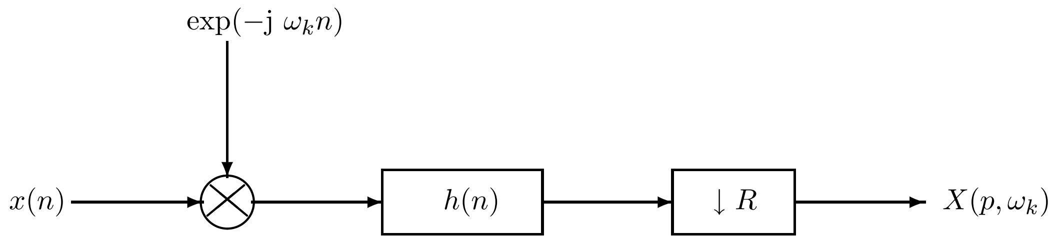

Figure B.1: Analyse par TFCT dans la convention passe-bas pour la voie d'indice k: Démodulation par l'exponentielle complexe correspondant à la fréquence centrale de la sous-bande, suivi d'un filtrage passe-bas par $h(n)$, puis d'un sous-échantillonnage par un facteur R.

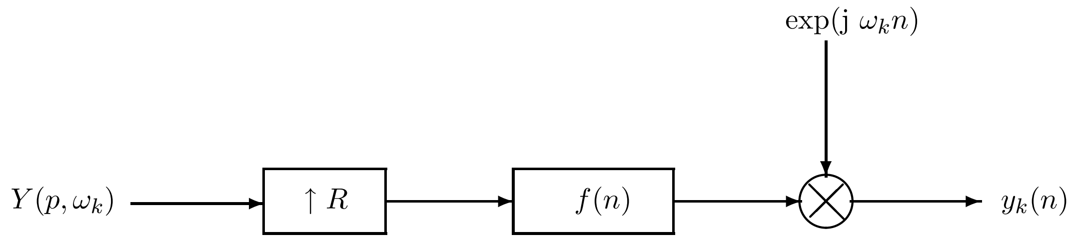

Figure B.2: Synthèse par TFCT dans la convention passe-bas pour la voie k du banc de filtres : Suréchantillonnage par le facteur R, suivi d'un filtrage passe-bas par $f(n)$, puis modulation par la fréquence centrale de la sous-bande. Par suite toutes les sorties des voies du banc de filtres $y_k(n)$ sont additionnées pour synthétiser le signal complet $y(n)$ (partie non-représentée sur le schéma).

Sous-échantillonnage des signaux de sous-bande La partie gauche de la figure B.1 représente sous la forme d'un schéma bloc l'analyse par TFCT telle quelle est décrite par l'équation (B.3). Sur cette figure, le bloc supplémentaire (à droite du schéma) représente la décimation par un facteur R du signal de sous-bande. De même, sur le schéma de synthèse B.2 qui correspond à l'équation (B.4), une opération d'interpolation par un facteur R a été représentée sur la gauche de la figure. La décimation consiste à ne conserver qu'un échantillon sur R tandis que l'interpolation revient à insérer $(R-1)$ valeurs nulles entre deux échantillons successifs [Crochiere 83]. Si l'on excepte la démodulation (multiplication par $\exp(-\mathrm{j}\ \omega_k n)$), la figure B.1 représente une opération de sous-échantillonnage, tandis que la figure B.2 correspond à un sur-échantillonnage (suivi d'une modulation par $\exp(\mathrm{j}\ \omega_k n)$). En effet, chaque voie du banc de filtres de TFCT ayant été filtrée par un filtre passe-bas, il est possible de décimer les signaux de sous-bande. C'est à dire que l'on peut abaisser la fréquence d'échantillonnage des signaux de sous-bande par un facteur R. Cette décimation, si le facteur R est choisi en accord avec le théorème d'échantillonnage, permet de diminuer le volume de données sans perdre d'information. Dans l'analogie transformée en bloc le facteur R s'interprète comme le pas de décalage des fenêtres [Crochiere 83]. La TFCT telle qu'elle est représentée par les schémas B.1 et B.2 s'écrit :

- A l'analyse,

$$X(p, \omega_k) = \sum_{n=-\infty}^{+\infty} h(pR - n)x(n)W_N^{-kn} \tag{B.7}$$

- A la synthèse,

$$y(n) = \sum_{p=-\infty}^{+\infty} f(n - pR)\frac{1}{N}\sum_{k=0}^{N-1} Y(p, \omega_k)W_N^{kn} \tag{B.8}$$

Pour un ω_k fixé, le signal de sous-bande $X(p, \omega_k)$ indicé par p est bien échantillonné à une fréquence R fois plus faible que le signal analysé $x(n)$ (indicé par n).

B.1.2 Convention passe-bande

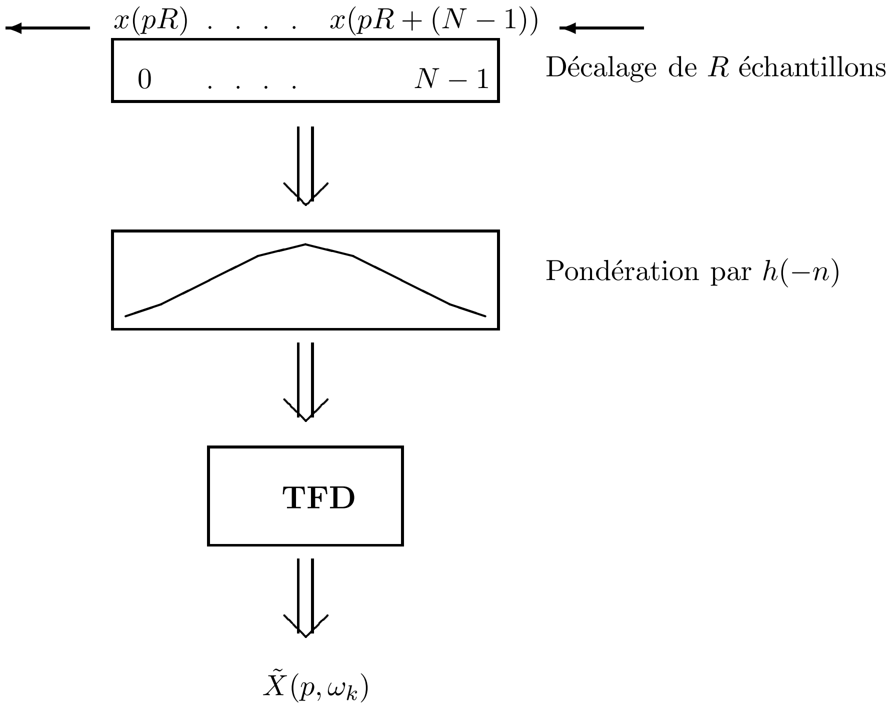

Figure B.3: Analyse par TFCT dans la convention passe-bande pour spectre à court-terme d'indice p. Le formalisme adopté est celui d'une transformée en blocs.

C'est la notation de la TFCT qui correspond à la manière dont s'effectue en pratique le calcul lorsqu'on utilise une implémentation sous la forme de calculs de TFD de blocs successifs. Cette fois, on parle de notation à **référence temporelle glissante** (ou relative) [Crochiere 83]. C'est à dire que les trames à court-terme sont maintenant référencées par rapport au début de la fenêtre courante. Le signal glisse (dans le sens des indices décroissants) par rapport à la fenêtre dont la position reste fixe. Dans l'analogie du banc de filtres, cette convention se traduit par le fait que les signaux de sous-bande sont des signaux à bande étroite. Dans cette convention, la TFCT s'écrit :

- A l'analyse,

$$\tilde{X}(p,\omega_k) = \sum_{n=0}^{N-1} h(-n)x(pR+n)W_N^{-kn} \tag{B.9}$$

- A la synthèse,

$$y(n) = \sum_{p=-\infty}^{+\infty} f(n-pR)\frac{1}{N}\sum_{k=0}^{N-1}\tilde{Y}(p,\omega_k)W_N^{k(n-pR)} \tag{B.10}$$

La notation $\tilde{X}(p,\omega_k)$ sera utilisée pour indiquer que la transformée à court-terme utilisée est à origine des temps glissante. Dans la formule (B.9), l'écriture $h(-n)$ permet de maintenir la cohérence avec la convention concernant le support de la fenêtre d'analyse. L'équation d'analyse (B.9) peut se représenter sous la forme du schéma de la figure B.3.

Lien entre les deux conventions de TFCT Le changement de variable $n' = n - pR$ dans la formule (B.7) permet de démontrer facilement la relation qui existe entre les transformées dans les deux notations [Crochiere 83] :

$$X(p,\omega_k) = W_N^{-kpR}\tilde{X}(p,\omega_k) \tag{B.11}$$

L'équation (B.11) indique que **la transformée de Fourier à court-terme dans la notation passe-bas se déduit de la transformée en notation passe-bande par une multiplication par un terme de phase dépendant linéairement du temps** (ou autrement dit par une modulation par une exponentielle complexe). On en déduit simplement le schéma équivalent de la TFCT en notation passe-bande sous la forme d'un banc de filtres (voir figure B.4).

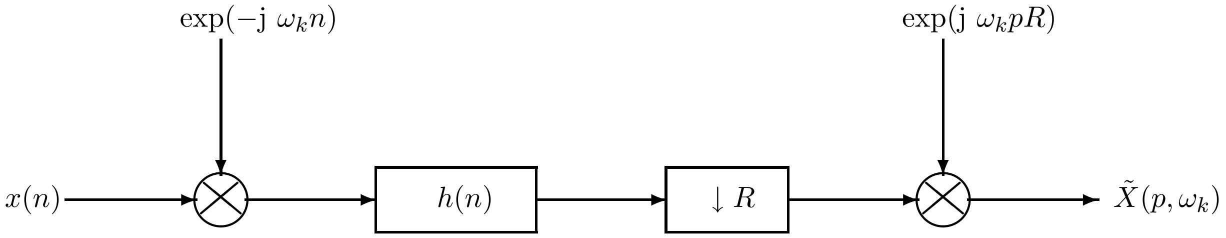

Figure B.4: Analyse par TFCT dans la convention passe-bande pour la voie d'indice k.

Le signal de sous-bande $\tilde{X}(p,\omega_k)$ est bien un signal passe-bande du fait de la modulation par l'exponentielle complexe de pulsation $\omega_k R$ (voir la figure B.4). Cependant, il ne faut pas oublier l'effet du sous-échantillonnage : sur la figure B.4, la pulsation centrale du signal de sous-bande normalisée par rapport à la fréquence d'échantillonnage de ce signal vaut

$$\omega_k R = \frac{2\pi k}{N}R \ \textbf{modulo}\ 2\pi$$

Ainsi, sauf dans le cas particulier où le pas de décalage de la fenêtre d'analyse R vaut 1 (voir l'utilisation de la TFCT dans [Moorer 86]), la fréquence centrale du signal de sous-bande se trouve translatée à cause du repliement lié à la décimation par R (cf. théorème d'échantillonnage pour les signaux à bande étroite [Crochiere 83]).

En conclusion, il faut retenir que dans chacune des conventions d'écriture de la transformée de Fourier à court-terme (passe-bas ou passe-bande), il existe deux analogies équivalentes : un banc de filtres multi-cadence ou une transformée en blocs successifs. L'analogie avec un banc de filtres est naturellement adaptée au cas où on s'intéresse aux valeurs successives d'un même canal fréquentiel $X(p,\omega_{k_0})$ pour k_0 fixé (voir par exemple les figures B.1 et B.4). A l'opposé, la forme d'une transformée en blocs successifs est plus naturelle lorsqu'on s'intéresse aux spectres à court-terme. Enfin, les deux conventions de notation de la TFCT ne diffèrent que par un terme de phase linéaire (modulation par une exponentielle complexe). La convention passe-bande est celle qui correspond à l'implémentation pratique de la TFCT par transformée de Fourier discrète (avec utilisation de l'algorithme rapide FFT). Tandis que la convention passe-bas se prête beaucoup plus simplement aux calculs théoriques (comparer par exemple les schémas B.1 et B.4). Dans la suite, et en particulier lorsqu'on ne s'intéresse qu'au module de la TFCT, c'est la convention passe-bas qui est utilisée pour effectuer les calculs.

B.2 Effet des modifications de la TFCT

Le débruitage par atténuation spectrale à court-terme revient à effectuer des modifications sur les transformées de Fourier à court-terme. Le but de ce paragraphe est de décrire l'effet de l'atténuation spectrale sous la forme d'une modification équivalente portant directement sur le signal temporel. Il s'avère que c'est la notion de *filtrage linéaire variant dans le temps* qui permet de traduire naturellement l'effet d'une modification multiplicative de la TFCT.

Il semble naturel de chercher à relier l'atténuation effectuée sur le spectre à court-terme avec un filtrage linéaire du signal. Cependant la modification spectrale apportée étant différente à chaque fenêtre à cause du caractère non-stationnaire du signal traité, il est nécessaire de faire intervenir la notion de *filtre linéaire variant dans le temps*. Nous allons donc étudier comment une modification multiplicative de la TFCT se traduit sous la forme d'un filtrage linéaire variant dans le temps. Les difficultés rencontrés dans cette démarche proviennent essentiellement de deux causes. D'une part, la modification multiplicative d'une transformée de Fourier discrète correspond en général à une convolution circulaire et non à une convolution simple [Delmas 91]. L'équivalence avec un filtre linéaire (même variant dans le temps) ne peut donc être obtenue qu'au prix de certaines conditions portant sur les paramètres du traitement ainsi que sur la modification spectrale (phénomène de *repliement temporel*). Par ailleurs, la modification spectrale n'est spécifiée qu'une fois par trame à court-terme, c'est à dire à une cadence R fois plus faible que la fréquence d'échantillonnage du signal traité. L'opération de conversion de fréquence qui est fait implicitement lors de la synthèse va elle aussi imposer des conditions supplémentaires (phénomène de *repliement spectral*).

Tous les résultats présentés dans cette partie sont adaptés du chapitre 7.4 de [Crochiere 83]. Par ailleurs, on suppose toujours, comme dans l'annexe B.1, que le nombre de voies du banc de filtres de TFCT est N (c'est à dire égal à la longueur de la fenêtre d'analyse). On se limite donc à un cadre plus restreint que celui de la référence citée ci-dessus ce qui peut expliquer quelques différences dans les formules.

Pour étudier cet aspect, on se place dans le cadre de la convention passe-bas, où la TFCT est définie par les deux équations (B.7) (à l'analyse) et (B.8) (à la synthèse). La modification apportée à la transformée à court-terme est modélisée sous la forme d'une multiplication :

$$Y(p, \omega_k) = X(p, \omega_k) G(p, \omega_k) \tag{B.12}$$

Dans le cas du débruitage, $G(p, \omega_k)$ représente l'atténuation spectrale apportée dans la fenêtre à court-terme d'indice p au point fréquentiel ω_k. Le but est d'exprimer l'effet de cette modification multiplicative du spectre à court-terme sous la forme d'un filtrage linéaire dépendant du temps.

Le modèle du filtre variant dans le temps est donné par

$$y(n) = \sum_m x(n-m)\tilde{g}_n(m) \tag{B.13}$$

Où $\tilde{g}_n(m)$ représente la réponse impulsionnelle du filtre variant dans le temps à l'indice n. Les bornes de la sommation ne sont pas précisées, car le support de la réponse impulsionnelle $\tilde{g}_n(m)$ peut dépendre de n (les bornes de la sommation varient alors avec n).

Une des difficultés posées par cette représentation sous la forme d'un filtrage linéaire variant dans le temps vient du fait que *les modifications dans le domaine spectral sont effectuées sur les signaux en sous-bande $X(p, \omega_k)$ échantillonnés à une fréquence R fois plus basse que la fréquence d'échantillonnage du signal analysé $x(n)$*. Ainsi, si on définit la réponse impulsionnelle $g_p(m)$, équivalent de la modification multiplicative de l'équation (B.12) dans le domaine temporel, comme étant la première période de la TFD inverse de $G(p, \omega_k)$ définie par

$$g_p(m) = \frac{1}{N}\sum_{k=0}^{N-1} G(p, \omega_k) W_N^{km} \quad \text{pour } m = 0, \ldots, N-1 \tag{B.14}$$

cette réponse impulsionnelle variant dans le temps $g_p(m)$ présente des similarités avec la réponse recherchée $\tilde{g}_n(m)$ définie par l'équation (B.13), cependant leurs fréquences d'échantillonnage ne sont pas semblables.

D'après la relation de synthèse (B.7), le signal résultant de la modification multiplicative de la TFCT s'écrit

$$y(n) = \sum_{p=-\infty}^{+\infty} f(n-pR)\frac{1}{N}\sum_{k=0}^{N-1} X(p, \omega_k) G(p, \omega_k) W_N^{kn} \tag{B.15}$$

où $G(p, \omega_k)$ représente la modification spectrale définie par la relation (B.12). En se souvenant que la relation d'analyse par TFCT (B.8) s'écrit

$$X(p, \omega_k) = \sum_{l=-\infty}^{+\infty} h(pR-l)x(l)W_N^{-kl}$$

la relation (B.15) se met sous la forme

$$y(n) = \sum_{l=-\infty}^{+\infty}\sum_{p=-\infty}^{+\infty} x(l)h(pR-l)f(n-pR)\left\{\frac{1}{N}\sum_{k=0}^{N-1} G(p, \omega_k) W_N^{k(n-l)}\right\} \tag{B.16}$$

D'après l'expression (B.14) qui définit la réponse impulsionnelle équivalente $g_p(m)$, le terme entre accolades dans l'équation ci-dessus correspond *à la périodisation* de $g_p(n-l)$ sur une durée infinie (propriété classique de la TFD inverse [Delmas 91]). Plus précisément cette périodisation peut s'écrire sous la forme

$$\frac{1}{N}\sum_{k=0}^{N-1} G(p, \omega_k) W_N^{k(n-l)} = \sum_{q=-\infty}^{+\infty} g_p(n-l+qN) \tag{B.17}$$

Avec cette notation, l'équation (B.16) devient

$$y(n) = \sum_{l=-\infty}^{+\infty}\sum_{p=-\infty}^{+\infty} x(l)h(pR-l)f(n-pR)\left\{\sum_{q=-\infty}^{+\infty} g_p(n-l+qN)\right\}$$

En effectuant le changement de variable $m = (n-l+qN)$, l'expression du signal obtenu en sortie se met sous la forme suivante

$$y(n) = \sum_{m=-\infty}^{+\infty} \left\{ \sum_{q=-\infty}^{+\infty} \left(x(n-m+qN) \sum_{p=-\infty}^{+\infty} g_p(m) h(pR-n+m-qN) f(n-pR) \right) \right\} \quad \text{(B.18)}$$

L'équation (B.18) ne définit pas en général un filtrage linéaire dépendant du temps : la sommation sur l'indice q traduit une convolution cyclique de période N. Ceci est dû au fait que l'équation (B.12) correspond, dans le cas général, à une convolution cyclique et non à une convolution simple. Pour retrouver un filtrage analogue à celui de l'équation (B.13), une comparaison entre les premiers termes des expressions (B.13) et (B.18) indique qu'il est nécessaire que la somme sur q se réduise au terme $q = 0$. Cette condition n'est réalisée que si les fenêtres d'analyse et de synthèse éliminent les termes dus au repliement temporel c'est à dire la contribution des termes $x(n-m-qN)$ pour $q \neq 0$. On montre qu'une **condition générale portant sur les fenêtres pour que les termes de repliement temporel disparaissent** est [Crochiere 83]

$$\mathcal{L}(h) + \mathcal{L}(f) + \mathcal{L}_{\max}(g_p) - 1 \leq 2N \quad \text{(B.19)}$$

Où $\mathcal{L}(h)$ désigne la longueur de la fenêtre $h(n)$ en échantillons. D'après l'équation (B.14), la longueur de la réponse impulsionnelle $g_p(m)$ est par définition inférieure à N. Cette longueur peut néanmoins varier au cours du temps selon que $g_p(m)$ comporte plus ou moins de termes nuls. La notation $\mathcal{L}_{\max}(g_p)$ correspond donc à la longueur maximale de la réponse équivalente sur l'ensemble des trames à court-termes (c'est à dire pour toutes les valeurs de p).

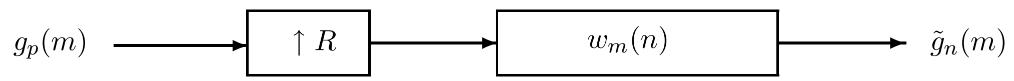

Figure B.5: Conversion de fréquence entre $g_p(m)$ (échantillonné à la fréquence des signaux de sous-bande) et $\tilde{g}_n(m)$ (échantillonné à la fréquence du signal analysé c'est à dire R fois plus grande). L'indice m (rang dans la réponse impulsionnelle variant dans le temps) est supposé fixé.

Lorsque la condition (B.19) portant sur les supports des fenêtres $h(n)$ et $f(n)$ est vérifiée, une comparaison entre les relations (B.13) et (B.18, avec $q = 0$) indique que la réponse impulsionnelle du filtre variant dans le temps, équivalent à la modification multiplicative $G(p, \omega_k)$ apportée sur la transformée à court-terme, s'écrit

$$\tilde{g}_n(m) = \sum_{p=-\infty}^{+\infty} f(n-pR) h(pR-n+m) g_p(m) \quad \text{(B.20)}$$

En notant $w_m(n) = f(n)h(m-n)$ le filtre composite obtenu par produit décalé des fenêtres d'analyse et de synthèse[2], la relation (B.20) devient

$$\tilde{g}_n(m) = \sum_{p=-\infty}^{+\infty} w_m(n-pR) g_p(m) \quad \text{(B.21)}$$

Cette dernière expression, si on la considère à m fixé (c'est à dire que l'on se place à un rang donné de la réponse impulsionnelle variant dans le temps), s'interprète comme une opération de conversion de fréquence (plus précisément de suréchantillonnage par un facteur R) selon le schéma de la figure B.5. Celle-ci montre que le filtre $w_m(n)$ sert de filtre de conversion de

[2] Les signes opposés de l'indice temporel n pour h et f sont dus à la convention sur les supports respectifs de ces deux fenêtres.

fréquence. C'est à dire qu'il doit théoriquement se comporter comme un filtre passe-bas idéal de fréquence de coupure $\frac{\pi}{R}$ (quel que soit l'indice m considéré dans le support de $g_p(m)$). Pour préciser le rôle du filtre réalisé par les fenêtres de la TFCT, on peut noter les points suivants :

- Si le filtre $w_m(n)$, ne coupe pas suffisamment la bande $[\frac{\pi}{R}, \pi]$, il apparaît des termes de **repliement fréquentiel** dans la réponse impulsionnelle équivalente $\tilde{g}_n(m)$. Cet effet est très gênant car il implique que l'atténuation effectivement apportée à une bande de fréquence peut différer notablement de celle qui a été spécifiée (sur la TFCT) à cause du repliement spectral. Pour éviter ce phénomène, $w_m(n)$ doit donc être un filtre passe-bas efficace. Pour l'indice $m = 0$, la condition à remplir est que le produit des fenêtres $h(-n)f(n)$ doit avoir une réponse aussi faible que possible dans la bande $[\frac{\pi}{R}, \pi]$. Intuitivement, il est très difficile d'assurer le caractère passe-bas de $w_m(n)$ pour tous les indices m allant de 0 à $(N-1)$. Le problème de repliement spectral se pose donc de manière d'autant plus importante que la longueur de la réponse impulsionnelle équivalente $g_p(m)$ est grande. Dans le cas où le support de $g_p(m)$ n'est pas trop important ($\mathcal{L}_{\max}(g_p) \ll N$), on peut considérer que la condition en $m = 0$ (caractère passe-bas de $h(-n)f(n)$) permet d'éviter le repliement spectral.
- Dans les cas où la bande $[\frac{\pi}{R}, \pi]$ est efficacement coupée, la modification apportée à la TFCT n'en est pas moins filtrée par $w_m(n)$ (voir aussi sur ce point les références [Allen 77] et [Nawab 88]). La variation temporelle de la réponse impulsionnelle équivalente (c'est à dire le signal $\tilde{g}_n(m)$ pour un rang m fixé) subit un filtrage passe-bas. En particulier, même si on spécifie une modification $G(p, \omega_k)$ qui varie très rapidement d'une fenêtre à l'autre, la modification effectivement apportée $\tilde{g}_n(m)$ aura une variation temporelle beaucoup plus lente.

En conclusion, la modification multiplicative de la transformée de Fourier à court-terme se traduit par une convolution avec un filtre variant dans le temps. Chacun des coefficients de ce filtre est obtenu par une opération de sur-échantillonnage à partir des réponses impulsionnelles successives correspondant aux modifications apportées dans chaque fenêtre. Les filtres passe-bas appliqués lors du suréchantillonnage de chacun des coefficients de la réponse impulsionnelle, sont obtenus par produit des fenêtres d'analyse et de synthèse décalées. Notons enfin que ce résultat n'est exact que lorsque le phénomène de repliement temporel peut être négligé, c'est à dire que la contrainte (B.19) sur les supports des fenêtres est vérifiée.

B.3 Choix des paramètres de TFCT

B.3.1 Techniques de synthèse

A la lecture de certaines publications, on a l'impression qu'il existe plusieurs techniques de synthèse différentes associées à la TFCT, désignées par leurs abréviations (en anglais) OLA, OLS, FBS ... En fait, la relation (B.8) permet bien d'unifier toutes les techniques de synthèse rencontrées dans la littérature. Les termes OLA, OLS (etc.) correspondent simplement à différents choix des paramètres de TFCT (en particulier, de la fenêtre de synthèse $f(n)$). Ces différentes techniques de synthèse se répartissent en deux grands groupes selon les buts recherchés :

Convolution rapide par TFCT On montre que la réalisation efficace du filtrage par un filtre FIR de réponse impulsionnelle longue peut se faire grâce à la TFCT. Le rôle des fenêtres

$h(n)$ et $f(n)$ est alors uniquement de sélectionner les parties de signal valides du point de vue de la convolution. Il suffit, par exemple, de choisir des fenêtres rectangulaires. On distingue deux types de techniques selon les positions respectives des supports de $h(n)$ et de $f(n)$: **OLA**(pour Overlap-Add) et **OLS** (pour Overlap-Save) [Crochiere 83].

Modification variant dans le temps C'est le cas qui correspond à l'application de débruitage par atténuation spectrale à court-terme. Les différentes techniques de synthèse proposées dans la littérature sont :

FBS (pour *Filter-Bank-Summation*) Dans laquelle il n'y a pas de filtre de synthèse : pour chaque trame à court-terme, seul le point correspondant au milieu de la trame est synthétisé (c'est à dire que $f(n) = \delta(n - N/2)$) [Nawab 88] [Allen 77]. Par principe cette technique de synthèse est limitée au cas où $R = 1$, donc appropriée uniquement lorsque le nombre N de voies de la TFCT est faible.

OLA (pour *Overlap-Add*)[3] Dans laquelle, la fenêtre de synthèse $f(n)$ est une fenêtre rectangulaire. La réalisation efficace de cette technique, lorsque le nombre de bandes N est grand, se fait en utilisant l'algorithme de FFT [Crochiere 83].

WOLA (pour *Weighted-Overlap-Add*) Qui correspond à une forme plus générale, où $f(n)$ est a priori quelconque [Nawab 88].

En pratique, pour le débruitage par atténuation spectrale à court-terme, c'est toujours la seconde technique (OLA) qui est utilisé (cf. paragraphe 2.3.1).

B.3.2 Transparence de l'analyse/synthèse

Une condition nécessaire au bon fonctionnement de l'analyse/synthèse par TFCT est la transparence du système. Cette contrainte correspond simplement à l'idée intuitive que le signal obtenu en sortie doit être identique au signal analysé en l'absence de modification. Mais cette condition ne garantie en aucune façon que les modifications se feront dans de bonnes conditions. En effet, en adoptant le formalisme du paragraphe B.2, la condition de transparence peut s'écrire

$$g_p(m) = \delta(m) \;\Rightarrow\; \tilde{g}_n(m) = \delta(m)$$

D'après le schéma B.5, ceci revient à dire que la conversion de fréquence, pour le coefficient d'ordre $m = 0$, doit se passer correctement *pour un signal continu* (constant pour tous les indices temporels p). Cette condition n'est donc absolument pas suffisante pour garantir le comportement du système vis à vis de modifications plus compliquées. Et même dans le cas où $g_p(m)$ est réduite à un seul coefficient pouvant éventuellement varier au court du temps (gain variable selon les trames à court-terme), le phénomène de repliement fréquentiel peut apparaître puisque seul le comportement vis à vis d'un signal continu est spécifié par la condition de transparence [Laroche 93a]. Dans ce paragraphe, on ne se préoccupe que des conditions pratiques qui garantissent la transparence du système, néanmoins il faut se souvenir qu'il existe d'autres conditions nécessaires au bon fonctionnement du débruitage (celles-ci seront étudiées en détail au paragraphe 2.3).

[3] Cette désignation n'est pas judicieuse car elle entretient la confusion avec la méthode de convolution rapide du même nom. Mais les buts recherchés dans les deux cas étant distincts, les conditions imposées aux paramètres de TFCT sont très différentes. En particulier, le choix de fenêtres rectangulaires n'est pas adapté au cas d'une modification variant dans le temps.

En reprenant l'équation (B.20), avec l'hypothèse $g_p(m) = \delta(m)$ pour toutes les valeurs de p, la condition d'analyse/synthèse équivalente à l'identité s'écrit

$$\sum_{p=-\infty}^{+\infty} f(n-pR)h(pR-n) = 1 \;\text{ quel que soit } n \tag{B.22}$$

Dans le cas où on utilise un fenêtre de synthèse $f(n)$ rectangulaire, de même support que la fenêtre d'analyse $h(n)$, cette dernière condition se simplifie sous la forme

$$\sum_{p=-\infty}^{+\infty} h(pR-n) = 1 \;\text{ quel que soit } n \tag{B.23}$$

Il faut noter que dans cette équation, la sommation porte sur un nombre fini de points puisque seul un nombre limité de fenêtres d'analyse se recouvrent (la fenêtre d'analyse est de durée finie N). De plus il suffit que la condition soit vérifiée pour $n = 0 \ldots R-1$, car le problème est inchangé si n est remplacé par $n + qR$ (où q est un entier quelconque).

Choix du pas de décalage des fenêtres C'est cette condition d'analyse/synthèse équivalente à l'identité qui impose des valeurs de recouvrement discrètes [Bourdier 88] [Crochiere 83]. A ce propos, on peut donner une petite règle qui permet de sélectionner les paramètres de recouvrement pour lesquels cette condition sera vérifiée. Il suffit de remarquer que la plupart des fenêtres utilisées couramment, s'obtiennent, par construction, en échantillonnant (sur N points) une fenêtre à temps continu du type

$$h_{\mathrm{cont}}(t) = \sum_{p=0}^{P_h} a_p \cos(2\pi pt) \;\text{ avec } t \in [-\frac{1}{2}, \frac{1}{2}] \tag{B.24}$$

Par exemple, pour la fenêtre de Hamming, P_h vaut 1 tandis que pour la fenêtre de Blackman P_h vaut 2 [Harris 78]. On montre facilement en utilisant les relations trigonométriques que si le taux de recouvrement s'écrit sous la forme $(1 - 1/K)$ (c'est à dire, si le pas de décalage des fenêtres divise la longueur de la fenêtre), il est nécessaire de choisir $\underline{K > P_h}$ pour que $h_{\mathrm{cont}}(t)$ vérifie la condition (B.23) au facteur d'échelle Ka_0 près. Par exemple, pour une fenêtre de Hamming les valeurs de recouvrement $\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \ldots$ conviennent, tandis que pour une fenêtre de Blackman, les valeurs $\frac{2}{3}, \frac{3}{4}, \ldots$ sont correctes.

Pour la fenêtre discrète $h(n)$ de longueur N, obtenue par échantillonnage de la fenêtre à temps continu $h_{\mathrm{cont}}(t)$, un paramètre de décalage R qui s'écrit $R = N/K$ avec $K > P_h$ permet de vérifier *approximativement* la relation (B.23). La condition n'est qu'approximativement vérifiée à cause des problèmes d'échantillonnage. Un exemple simple de ce type de problèmes est qu'il est impossible de réaliser exactement un recouvrement de 2/3 lorsque la longueur de la fenêtre est une puissance de 2 (pour permettre l'utilisation de l'algorithme de FFT !). Par ailleurs, cet aspect est aussi lié à la manière dont est échantillonnée la fenêtre $h_{\mathrm{cont}}(t)$ [Harris 78] [Serra 89]. On vérifie par exemple que la fenêtre de Hann échantillonnée de manière symétrique

$$h(n) = \frac{1}{2}\left\{1 + \cos\left(2\pi\frac{2n + (N-1)}{2(N-1)}\right)\right\} \;\text{ pour } n = -(N-1),\ \ldots\ ,-1, 0$$

ne vérifie la condition (B.23) que de manière approximative pour un pas de décalage R de $N/2$ (en supposant la longueur de la fenêtre N paire). Par contre, l'échantillonnage dissymétrique de la même fenêtre

$$h(n) = \frac{1}{2}\left\{1 + \cos\left(2\pi\frac{2n + N}{2N}\right)\right\} \;\text{ pour } n = -(N-1),\ \ldots\ ,-1, 0$$

permet de remplir exactement la condition (B.23), à la précision de la représentation arithmétique près. Enfin, ces problèmes liés à l'échantillonnage de la fenêtre $h_{\text{cont}}(t)$ sont d'autant plus important que la longueur de la fenêtre N est faible : intuitivement, plus la fenêtre est longue plus on se rapproche du cas de la fenêtre à temps continu.

Modulation due à l'analyse/synthèse Toutefois, il est légitime de tolérer de légers écarts par rapport à la condition (B.23) du moment que les modifications apportées lors de l'analyse/synthèse par TFCT (toujours sans modifications spectrales volontaires) restent inaudibles. On remarque que la quantité suivante (définie dans le cas d'une fenêtre de synthèse rectangulaire)

$$a_h(n) = \sum_{p=-\infty}^{+\infty} h(pR - n) \tag{B.25}$$

est constante quand la condition d'analyse/synthèse équivalente à l'identité (B.23) est vérifiée. Avec cette notation, l'équation (B.20), considérée dans le cas de l'analyse/synthèse sans modification ($g_p(m) = \delta(m)$), s'écrit

$$\tilde{g}_n(m) = a_h(n)\delta(m) \tag{B.26}$$

C'est à dire que lorsque qu'on effectue une analyse/synthèse par TFCT sans modification, le résultat correspond au signal analysé multiplié par le facteur $a_h(n)$. Par exemple pour un signal stationnaire constitué d'une somme de composantes sinusoïdales, chaque composante est modulée en amplitude[4] par le signal $a_h(n)$. On peut donc considérer que $a_h(n)$ représente la modulation d'amplitude due à l'analyse/synthèse. Serra [Serra 89] propose d'utiliser un l'indice de modulation calculé à partir de $a_h(n)$ pour évaluer l'audibilité du phénomène. Cet indice de modulation peut s'écrire sous la forme

$$m_h = \frac{\max(a_h(n)) - \min(a_h(n))}{\bar{a}_h(n)}$$

Cependant, il est nécessaire de connaître la composition fréquentielle du signal modulant pour évaluer l'audibilité d'une modulation [Zwicker 81]. Ici tout ce que l'on peut dire de manière générale c'est que $a_h(n)$ est périodique de période R. Comme on ne connaît pas le nombre d'harmoniques qui interviennent dans la décomposition spectrale de $a_h(n)$, il est naturel de se référer à la valeur limite du taux de modulation d'amplitude qui garantit un effet inaudible quelles que soit les fréquences du signal modulant et de la porteuse. D'après les résultats de tests psychoacoustiques reportés dans [Zwicker 81], cette valeur limite de m_h se situe autour de 0,02% ce qui correspond à la valeur donnée dans [Serra 89].

Cette valeur de 0,02%, qui garantit que l'effet de modulation reste inaudible, est en général trop contraignante. Il semble que même avec des valeurs de m_h environ 10 fois supérieures à cette limite, le résultat reste satisfaisant. Pour expliquer ce phénomène, on peut noter que pour une fenêtre d'analyse $h(n)$ de *durée* 40 ms ou plus (c'est à dire que $NF_e > 40$ ms) avec un paramètre R de l'ordre de $N/2$, la fréquence fondamentale du signal modulant est inférieure à 50 Hz ($1/R$). Or la limite d'audibilité des phénomènes de modulations est beaucoup plus haute lorsque la fréquence du son modulant est faible (inférieure à 500 Hz) [Zwicker 81]. Si on retient comme fréquence maximale du signal modulant environ 250 Hz (en considérant que le spectre de $a_h(n)$ contient au plus 5 harmoniques de puissance significative), on obtient comme valeur limite du taux de modulation 0,16%. Cette dernière valeur est plus proche de ce qui semble être en pratique la valeur limite du taux de modulation audible. Cependant, ces remarques n'ont pas valeur de démonstration. En particulier, on a supposé qu'une modulation par un signal

[4] Il n'est pas question ici de la définition rigoureuse du signal modulant en amplitude. On sait en effet que pour un signal qui s'écrit $a(n)\exp(\mathrm{j}\ \omega n)$, le signal modulant en amplitude (défini grâce à la transformée de Hilbert) n'est égal à $a(n)$ que si ce dernier vérifie certaines conditions [Picinbono 83].

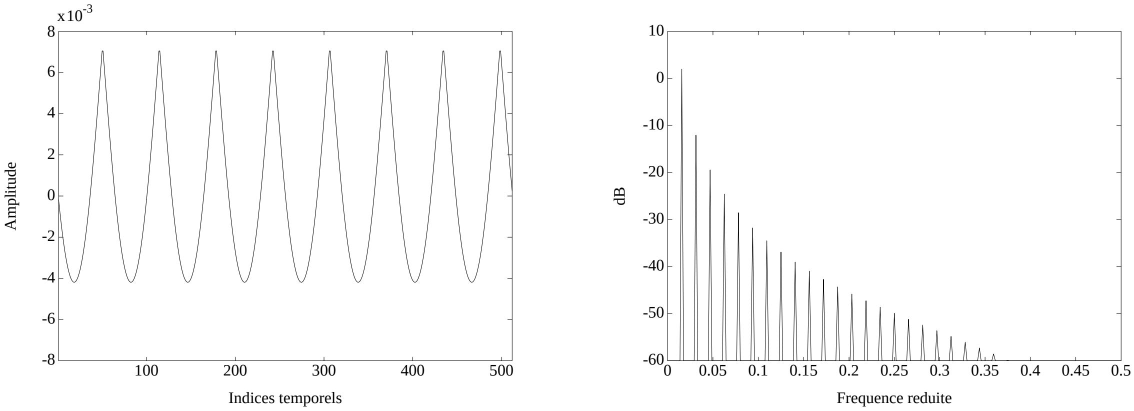

Figure B.6: Modulation d'amplitude due à l'analyse/synthèse, avec une fenêtre de Hamming de longueur 128 (échantillonnée de manière symétrique) et un pas de décalage de 64 points. **A gauche,** forme temporelle de la modulation relative d'amplitude $\{a_h(n) - \bar{a}_h(n)\}/\bar{a}_h(n)$. **A droite,** spectre de la modulation d'amplitude.

harmonique est moins audible qu'une modulation par la composante de plus haute fréquence, avec le même indice. Ceci semble assez logique au regard des courbes reportées dans [Zwicker 81], cependant il est certain qu'il serait nécessaire de le vérifier par des expériences psychoacoustiques spécifiques.

Annexe C

Niveau relatif moyen d'un son pur bruité

On supposera ici que le signal à traiter est constitué d'un son pur noyé dans du bruit. Les arguments exposées aux paragraphe 3.1.1.a indiquent que, dans ce cas, la qualité du débruitage par atténuation spectrale à court-terme dépend essentiellement du **niveau relatif moyen mesuré au sommet du pic spectral correspondant à la sinusoïde**. C'est cette quantité que l'on se propose de déterminer en fonction des caractéristiques du signal sinusoïdal et du bruit.

Caractéristiques du signal étudié Le modèle du signal bruité dans le domaine analogique (avant l'échantillonnage) s'écrit sous la forme

$$x^{(a)}(t) = s^{(a)}(t) + d^{(a)}(t) \tag{C.1}$$

où t représente la variable temporelle, et l'exposant (a) est utilisé pour indiquer qu'il s'agit de quantités analogiques (à temps continu). La composante sinusoïdale (c'est à dire le signal non-bruité) s'écrit

$$s^{(a)}(t) = A\cos(2\pi Ft + \phi^{(a)}) \tag{C.2}$$

La puissance de la composante sinusoïdale est donnée par $\mathcal{P}_s = A^2/2$. La valeur de la densité spectrale de puissance du bruit additif à une fréquence f est notée $P_d^{(a)}(f)$. Après échantillonnage (supposé idéal), et en négligeant les aspects liés à la quantification, le signal bruité s'écrit

$$x(n) = s(n) + d(n)$$

La composante de signal est donnée par

$$s(n) = A\cos(\Omega n + \phi) \tag{C.3}$$

où Ω représente la pulsation réduite correspondant à la fréquence de la sinusoïde, définie par $\Omega = 2\pi F/F_e$ (F_e désignant la fréquence d'échantillonnage). Quant au bruit $d(n)$ issu de l'échantillonnage du processus aléatoire à temps continu $d^{(a)}(t)$, sa densité spectrale de puissance s'écrit [Charbit 90]

$$P_d(\omega) = F_e P_d^{(a)}(\frac{\omega}{2\pi}F_e) \tag{C.4}$$

pour toutes les pulsations réduites $\omega \in]-\pi, \pi]$.

La prochaine étape consiste à expliciter la transformée de Fourier à court-terme du signal bruité $x(n)$. Conformément à ce qui a été dit dans l'annexe B.1 , puisque l'on s'intéresse seulement au module de la TFCT, on utilise la convention passe-bas qui est plus simple pour les calculs. L'expression de la TFCT est alors donnée par la formule (B.7). Par linéarité, la TFCT du signal bruité peut s'écrire

$$X(p, \omega_k) = S(p, \omega_k) + D(p, \omega_k)$$

L'espérance du spectre de puissance à court-terme du signal bruité s'écrit donc

$$\begin{aligned} \mathrm{E}\left\{|X(p, \omega_k)|^2\right\} &= S(p, \omega_k)S(p, \omega_k)^* + \mathrm{E}\left\{S(p, \omega_k)S(p, \omega_k)^*\right\} + \\ &\quad \mathrm{E}\left\{S(p, \omega_k)D(p, \omega_k)^*\right\} + \mathrm{E}\left\{D(p, \omega_k)S(p, \omega_k)^*\right\} \end{aligned}$$

Comme la TFCT s'obtient à partir du signal par une opération linéaire, et que le bruit est supposé non corrélé avec le signal, les deux derniers termes de l'expression ci-dessus sont nuls. C'est à dire que

$$\mathrm{E}\left\{|X(p, \omega_k)|^2\right\} = |S(p, \omega_k)|^2 + \mathrm{E}\left\{|D(p, \omega_k)|^2\right\} \tag{C.5}$$

Dans la suite, on considère séparément les contributions des deux parties du signal, composante sinusoïdale et bruit additif.

Pour la composante sinusoïdale La transformée de Fourier à court-terme s'écrit

$$S(p, \omega_k) = \sum_{n=-\infty}^{+\infty} h(pR - n)s(n)\mathrm{e}^{-\mathrm{j}\omega_k n}$$

où $h(n)$ représente la fenêtre d'analyse (voir l'annexe B.1). En décomposant la sinusoïde de l'équation (C.3) sur ses composantes complexes en phase et en quadrature, l'expression de la TFCT devient

$$S(p, \omega_k) = \frac{A}{2}\sum_{n=-\infty}^{+\infty} h(pR - n)\mathrm{e}^{\mathrm{j}(\Omega n + \phi - \omega_k n)} + \frac{A}{2}\sum_{n=-\infty}^{+\infty} h(pR - n)\mathrm{e}^{-\mathrm{j}(\Omega n + \phi + \omega_k n)}$$

soit

$$S(p, \omega_k) = \frac{A}{2}\mathrm{e}^{\mathrm{j}\phi}\sum_{n=-\infty}^{+\infty} h(pR - n)\mathrm{e}^{-\mathrm{j}(\omega_k - \Omega)n} + \frac{A}{2}\mathrm{e}^{-\mathrm{j}\phi}\sum_{n=-\infty}^{+\infty} h(pR - n)\mathrm{e}^{-\mathrm{j}(\omega_k + \Omega)n}$$

En effectuant le changement de variable $m = pR - n$ afin de faire apparaître la transformée de Fourier de la fenêtre d'analyse, cette relation se met sous la forme

$$S(p, \omega_k) = \frac{A}{2}\mathrm{e}^{\mathrm{j}(\phi + (\Omega - \omega_k)pR)}\sum_{m=-\infty}^{+\infty} h(m)\mathrm{e}^{-\mathrm{j}(\Omega - \omega_k)m} + \frac{A}{2}\mathrm{e}^{-\mathrm{j}(\phi + (\Omega + \omega_k)pR)}\sum_{m=-\infty}^{+\infty} h(m)\mathrm{e}^{\mathrm{j}(\Omega + \omega_k)m}$$

c'est à dire

$$S(p, \omega_k) = \frac{A}{2}\mathrm{e}^{\mathrm{j}(\phi + (\Omega - \omega_k)pR)}H(\Omega - \omega_k) + \frac{A}{2}\mathrm{e}^{-\mathrm{j}(\phi + (\Omega + \omega_k)pR)}H^*(\Omega + \omega_k) \tag{C.6}$$

où $H(\omega)$ désigne la transformée de Fourier de la fenêtre d'analyse $h(n)$.

Par la suite, seules les fréquences positives de la transformée de Fourier discrète sont considérées. On suppose de plus que la fenêtre d'analyse $h(n)$ est un filtre passe-bas suffisamment sélectif pour limiter l'influence du second terme de la relation (C.6) qui correspond aux fréquences négatives. Plus précisément, si la pulsation de coupure de $h(n)$ est très inférieure à 2Ω, la relation (C.6) se simplifie au voisinage de la fréquence de la sinusoïde sous la forme

$$S(p,\omega_k) \approx \frac{A}{2} \mathrm{e}^{\mathrm{j}(\phi+(\Omega-\omega_k)pR)} H(\Omega-\omega_k) \tag{C.7}$$

Cette dernière relation est valable lorsque l'écart fréquentiel $(\Omega-\omega_k)$ est faible devant la fréquence de la sinusoïde (Ω). Il faut noter que cette condition n'est plus vérifiée quand Ω est de l'ordre du pas de discrétisation des pulsations, c'est à dire de $\omega_1 = 2\pi/N$ où N désigne la longueur de la fenêtre. Pour des durées usuelles de la fenêtre d'analyse (supérieures à 20ms), ce pas de discrétisation est de l'ordre de 50 Hz. Il n'est donc pas utile de se préoccuper des fréquences de cet ordre car elles sont à la limite inférieure du domaine audible (et en tout cas, pour la plupart des anciens enregistrements, celles-ci se trouvent en dehors de la bande passante !). Toutefois, il peut être nécessaire de prendre en compte ce point lorsque des fenêtres de très courte durée sont utilisées. L'équation (C.6) montre d'ailleurs que dans ce cas, la valeur du module de la TFCT au voisinage de la pulsation de la sinusoïde varie dans le temps (elle dépend de l'indice temporel p).

Dans notre cas, une dernière simplification découle du fait que le canal fréquentiel auquel on s'intéresse est essentiellement celui dont la pulsation centrale ω_k est la plus proche de la pulsation de la sinusoïde Ω. L'indice de cette sous-bande est noté k_0. La transformée de Fourier de la fenêtre de pondération $H(\omega)$ présente en général un lobe principal large par rapport au pas de discrétisation $2\pi/N$. On montre en effet que pour les fenêtres de pondération douces (par exemple, Hamming, Hann ou Blackman), le rapport $H(\Omega-\omega_{k_0})/H(0)$ est supérieur à -2 dB (voir dans [Harris 78] les chiffres concernant le *Scalloping Loss*). Cet écart est suffisamment faible pour qu'il ne soit pas utile de le considérer ici. Ce qui revient à dire que la différence de hauteur du pic spectral, selon que la pulsation de la sinusoïde correspond ou non à un point de la TFD, est négligeable. La valeur de la TFCT au point fréquentiel correspondant au sommet du pic s'écrit donc

$$S(p,\omega_{k_0}) \approx \frac{A}{2} \mathrm{e}^{\mathrm{j}(\phi+(\Omega-\omega_{k_0})pR)} H(0) \tag{C.8}$$

En ne s'intéressant qu'au module de ce terme, on obtient le niveau du pic

$$|S(p,\omega_{k_0})| = \frac{A}{2} H(0) \tag{C.9}$$

Le terme réel $H(0)$ représente le gain en continu de la fenêtre qui s'écrit

$$H(0) = \sum_n h(n)$$

Pour finir, il reste à réécrire cette relation définissant le niveau du pic du spectre à court-terme, en fonction de la puissance de la composante sinusoïdale

$$|S(p,\omega_{k_0})| = \sqrt{\frac{\mathcal{P}_s}{2}} \sum_n h(n) \tag{C.10}$$

Pour le bruit additif La quantité à évaluer est l'espérance mathématique du spectre de puissance à court-terme pour chaque point fréquentiel. Ce qui s'écrit

$$\mathrm{E}\left\{|D(p,\omega_k)|^2\right\} = \mathrm{E}\left\{\left|\sum_{n=-\infty}^{+\infty} h(pR-n)d(n)\mathrm{e}^{-\mathrm{j}\omega_k n}\right|^2\right\}$$

En posant $m = n - pR$ dans l'expression du spectre de puissance à court-terme, cette relation peut être réécrite comme suit

$$\mathrm{E}\left\{|D(p,\omega_k)|^2\right\} = \mathrm{E}\left\{\left|\sum_{m=-\infty}^{+\infty} h(-m)d(pR+m)\mathrm{e}^{-\mathrm{j}\omega_k m}\right|^2\right\} \tag{C.11}$$

Cette expression montre que la quantité recherchée ne dépend pas de l'indice de trame p. En effet, si on suppose le bruit $d(n)$ stationnaire, le processus $d(pR+m)$ possède les mêmes propriétés statistiques que le processus $d(m)$ quel que soit la valeur de l'indice temporel p. En particulier, il est possible, sans changer le résultat d'écrire $d(m)$ à la place de $d(pR+m)$. Dans le même ordre d'idée, il est légitime pour plus de simplicité d'utiliser le processus $d(-m)$ dans l'expression (C.11) (du fait de la stationnarité, celui-ci possède aussi les mêmes propriétés) ce qui simplifie encore l'écriture sous la forme suivante

$$\mathrm{E}\left\{|D(p,\omega_k)|^2\right\} = \mathrm{E}\left\{\left|\sum_{m=-\infty}^{+\infty} h(-m)d(-m)\mathrm{e}^{-\mathrm{j}\omega_k m}\right|^2\right\}$$

Dans cette relation, il suffit de poser $n = -m$ pour retrouver dans le membre de droite l'expression classique du périodogramme (en se souvenant que le module de la TFD d'un signal réel est pair). Le terme recherché s'écrit alors

$$\mathrm{E}\left\{|D(p,\omega_k)|^2\right\} = \mathrm{E}\left\{\left|\sum_{n=-\infty}^{+\infty} h(n)d(n)\mathrm{e}^{-\mathrm{j}\omega_k n}\right|^2\right\} \tag{C.12}$$

Ce résultat montre que la valeur moyenne du spectre de puissance à court-terme, pour un point fréquentiel donné, est constante dans le temps, et que de plus, celle-ci correspond simplement à l'espérance mathématique d'un périodogramme (avec une fenêtre de pondération) du processus aléatoire $d(n)$. Le périodogramme (défini comme le module au carré de la TFD) est un estimateur de la densité spectrale de puissance du processus très classique. En particulier, on montre que le périodogramme est un estimateur asymptotiquement non biaisé de la DSP du processus étudié [Brillinger 81], c'est à dire que

$$\mathrm{E}\left\{\left(\sum_m h(m)^2\right)^{-1}\left|\sum_{n=-\infty}^{+\infty} h(n)d(n)\mathrm{e}^{-\mathrm{j}\omega n}\right|^2\right\} = P_d(\omega) \tag{C.13}$$

Ce résultat est vérifié asymptotiquement, lorsque la longueur de la fenêtre N est assez grande (ce qui correspond aux cas qui nous intéressent). De plus, la démonstration de ce résultat est obtenue en supposant que le bruit $d(n)$ est gaussien (et stationnaire) et que sa fonction d'autocorrélation décroît"suffisamment vite" à l'infini. Dans le cas général (bruit non gaussien), ce résultat est aussi vérifié dès lors que les moments de tout ordre de $d(n)$ vérifient certaines contraintes [Brillinger 81]. L'espérance du spectre de puissance à court-terme, pour la partie bruit, est donc donnée par

$$\mathrm{E}\left\{|D(p,\omega_k)|^2\right\} = P_d(\omega)\left(\sum_n h(n)^2\right) \tag{C.14}$$

C'est à dire, en introduisant la densité spectrale de puissance du processus sous sa forme analogique

$$\mathrm{E}\left\{|D(p,\omega_k)|^2\right\} = F_e\, P_d^{(a)}(\frac{\omega_k}{2\pi}F_e)\left(\sum_n h(n)^2\right) \tag{C.15}$$

Pour évaluer cette valeur au point d'indice k_0 (celui qui correspond au pic de la sinusoïde), on suppose qu'autour de la fréquence F de la sinusoïde, la densité spectrale de puissance du bruit est localement uniforme. En particulier, on admet que la DSP du bruit est quasiment constante autour de F sur une largeur de bande au moins égale à la bande passante du filtre d'analyse de la TFCT. Cette hypothèse se traduit par la relation

$$P_d^{(a)}(\frac{\omega_{k_0}}{2\pi}F_e) \approx P_d^{(a)}(F)$$

D'où l'expression finale pour la partie bruit,

$$\mathrm{E}\left\{|D(p,\omega_{k_0})|^2\right\} = F_e\, P_d^{(a)}(F) \left(\sum_n h(n)^2\right) \tag{C.16}$$

Calcul de la valeur moyenne du niveau relatif La dernière étape du calcul consiste à évaluer la valeur moyenne du niveau relatif au point correspondant au sommet du pic spectral dû à la sinusoïde. Celui-ci s'écrit

$$\mathrm{E}\{\mathcal{Q}(p,\omega_{k_0})\} = \mathrm{E}\left\{\frac{|X(p,\omega_k)|^2}{\mathrm{E}\left\{|D(\omega_k)|^2\right\}}\right\}$$

En supposant que le signal et le bruit sont décorrélés, l'expression du niveau relatif moyen se met sous la forme

$$\mathrm{E}\{\mathcal{Q}(p,\omega_{k_0})\} = 1 + \frac{|S(p,\omega_k)|^2}{\mathrm{E}\left\{|D(\omega_k)|^2\right\}}$$

D'où l'expression finale

$$\mathrm{E}\{\mathcal{Q}(p,\omega_{k_0})\} = 1 + \frac{\mathcal{P}_s\left[\sum_n h(n)\right]^2}{2P_d^{(a)}(F)\, F_e\left[\sum_n h(n)^2\right]} \tag{C.17}$$

Dans cette dernière expression, il est intéressant de regrouper tous les termes qui dépendent de la fenêtre d'analyse. Plus précisément, on choisit de poser

$$\Delta_h = \frac{N\left[\sum_n h(n)^2\right]}{\left[\sum_n h(n)\right]^2} \tag{C.18}$$

On montre que *cette quantité ne dépend que du type de la fenêtre*, et non de la longueur N de la fenêtre [Harris 78]. De plus, la valeur de Δ_h est toujours supérieure à un 1 dans le cas d'une fenêtre dont les coefficients sont positifs, la valeur $\Delta_h = 1$ correspondant au cas de la fenêtre rectangulaire.

Cette quantité est baptisée *Equivalent Noise Bandwidth* dans [Harris 78], c'est à dire en français, **largeur de bande équivalente de la fenêtre $h(n)$ vis à vis du bruit**. Plus précisément Δ_h représente la largeur de bande d'un filtre passe-bas idéal, possédant le même gain en continu que $h(n)$, et qui, si on l'applique à un bruit blanc, fournit un signal de même puissance que si il avait été filtré par $h(n)$. La réponse fréquentielle de ce filtre passe-bas idéal équivalent à la fenêtre $h(n)$ est représenté en pointillés sur la figure C.1 pour une fenêtre $h(n)$

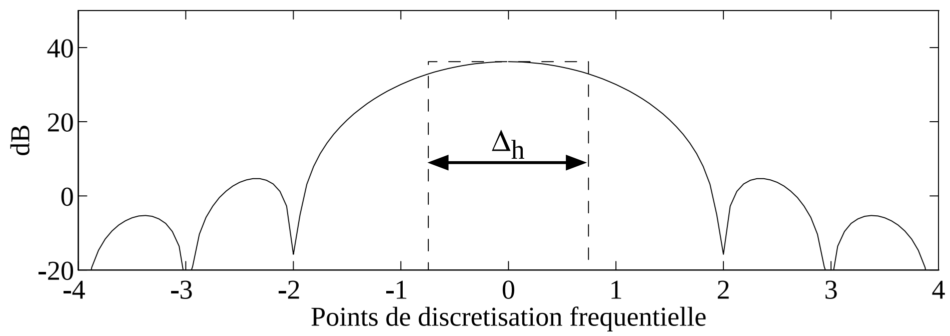

Figure C.1: Largeur de bande équivalente Δ_h d'une fenêtre de Hann (de longueur $N = 128$) vis à vis du bruit. En trait plein, la réponse fréquentielle de la fenêtre de Hann, et en traits pointillés, celle du filtre passe-bas idéal équivalent. Les abscisses correspondent à des points de discrétisation fréquentielle de pas $1/N$ (soit à $N\nu$ où ν représente la fréquence réduite).

de Hann. Il faut noter que Δ_h représente un nombre de points, en supposant une discrétisation fréquentielle de pas $1/N$. C'est à dire que Δ_h/N représente bien une largeur de bande exprimée en fréquence réduite. La figure C.1 montre qu'il est légitime de substituer (de manière fictive) à la fenêtre d'analyse $h(n)$, le filtre dont la réponse fréquentielle est représentée en pointillés : la hauteur du pic associé à la sinusoïde est identique dans les deux cas puisque ces deux filtres possèdent le même gain cohérent en puissance ($H(0)^2 = [\sum_n h(n)]^2$), de plus, le niveau moyen du spectre de puissance du bruit est le même car les deux filtres ont le même gain incohérent ($[\sum_n h(n)^2]$). La quantité Δ_h est donc une mesure commode de la manière dont la fenêtre $h(n)$ filtre un bruit. Une dernière remarque est que si Δ_h paraît visuellement un peu faible par rapport à la largeur de bande de $h(n)$ sur la figure C.1, c'est uniquement à cause de la représentation de la réponse en fréquence en décibels.

En introduisant le terme Δ_h dans la relation (C.17), la valeur moyenne du niveau relatif local, dans le canal fréquentiel où se trouve le pic correspondant à la sinusoïde, peut s'écrire

$$\mathrm{E}\left\{\mathcal{Q}(p, \omega_{k_0})\right\} = 1 + \frac{\mathcal{P}_s N}{2 P_d^{(a)}(F)\, F_e \Delta_h} \tag{C.19}$$

L'interprétation de cette expression est détaillée, à propos de la limite de restauration, au paragraphe 3.1.1.b.

Annexe D

Règle de suppression d'Ephraim et Malah (texte en anglais)

Cette annexe reproduit le texte de l'article [Cappe 94] à paraître (courant 1994) dans la revue *IEEE Transactions on Speech and Audio Processing*. Le but de cet article est de montrer comment, et à quel prix, l'utilisation de la règle de suppression d'Ephraim et Malah permet d'éviter le phénomène de bruit musical. Les deux premiers paragraphes se recoupent largement avec la présentation générale de l'algorithme d'Ephraim et Malah effectuée au paragraphe 2.2.3. Enfin, pour éviter un va-et-vient pénible entre deux conventions, nous avons choisi, pour cet article, de n'utiliser que les rapports signal-à-bruit $\mathcal{R}_{post}(p, \omega_k)$ et $\mathcal{R}_{prio}(p, \omega_k)$. Pour le paramètre a posteriori, la convention adoptée est donc différente de celle du paragraphe 2.2.3 qui correspondait à la présentation originale de l'article d'Epraim et Malah.

Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor

This paper presents a study of the noise suppression technique proposed by Y. Ephraim and D. Malah. This technique has been used recently for the restoration of degraded audio recordings because it is free of the frequently encountered 'musical noise' artifact. It is demonstrated how this artifact is actually eliminated without bringing distortion to the recorded signal even if the noise is only poorly stationary.

D.1 Introduction

At present, the noise reduction techniques used for the restoration of degraded audio recordings are based on *short-time spectral attenuation*. In such techniques the attenuation that is to be applied to each one of the short-time Fourier transform coefficients is estimated by the *noise suppression rule* [Lim 79] [Mc Aulay 80] [Vary 85].

One artifact that has been widely reported concerning the use of short-time spectral attenuation techniques is that the noise remaining after the processing has a very unnatural disturbing quality [Boll 79] [Moorer 86] [Valiere 91] [Vaseghi 92]. This comes from the fact that the magnitude of the short-time spectrum $|X(p,\omega_k)|$ exhibits strong fluctuations in noisy areas which is a well-known feature of the periodogram [Brillinger 81]. After application of the spectral attenuation, the short-time magnitude spectrum in the frequency bands that originally contained noise now consists in a succession of randomly spaced spectral peaks corresponding to the maxima of $|X(p,\omega_k)|$. In between these peaks the short-time spectrum values are strongly attenuated because they are close to or below the estimated average noise spectrum. As a result, the residual noise is composed of sinusoidal components with random frequencies that come and go in each short-time frame [Boll 79] [Moorer 86]. This artifact is known as the 'musical noise phenomenon', the term 'musical' being a reference to the presence of pure tones in the residual noise.

Some modifications of the basic suppression rules have been proposed in order to overcome this problem [Boll 79] [Vaseghi 92], but these techniques *only reduce* the musical noise without completely eliminating it. The complete elimination of the musical noise phenomenon is generally only obtained by a crude overestimation of the noise average spectrum. An unwanted consequence is that the short-time spectrum is attenuated much more than would be necessary, a fact which can generate audible distortions in the audio signal [Cappe 93b].

It has been reported that the noise suppression rule proposed by Ephraim and Malah (that will be referred to as the EMSR in the following) [Ephraim 83] [Ephraim 84] makes it possible to obtain a significant noise reduction while avoiding the musical noise phenomenon described above. This feature explains why this suppression rule is an excellent choice for the restoration of musical recordings where the musical noise artifact is to be strictly avoided [Valiere 91].

In the original papers by Ephraim and Malah, this aspect of the suppression rule was only mentioned as an experimental finding. In this correspondence, we investigate the mechanisms that counter the musical noise phenomenon.

D.2 Description of the EMSR

The EMSR was proposed by Ephraim and Malah in [Ephraim 83] and developed in [Ephraim 84], two other suppression rules along the same principle were introduced later by the authors in [Ephraim 84] and [Ephraim 85]. Here we will focus only on the EMSR, the fundamental mechanism that counters the musical noise effect being basically the same in all these suppression rules.

The EMSR can be expressed as a spectral gain $G(p,\omega_k)$ to be applied to each short-time spectrum value $X(p,\omega_k)$, this gain is given by [Ephraim 83] [Ephraim 84]

$$G = \frac{\sqrt{\pi}}{2}\sqrt{\left(\frac{1}{1+\mathcal{R}_{post}}\right)\left(\frac{\mathcal{R}_{prio}}{1+\mathcal{R}_{prio}}\right)} \times \mathbf{M}\left[(1+\mathcal{R}_{post})\left(\frac{\mathcal{R}_{prio}}{1+\mathcal{R}_{prio}}\right)\right] \tag{D.1}$$

where $\mathbf{M}$ stands for the function

$$\mathbf{M}\left[\theta\right] = \exp(-\frac{\theta}{2}) \left[(1+\theta)\, I_0\left(\frac{\theta}{2}\right) + \theta I_1\left(\frac{\theta}{2}\right) \right]$$

I_0 and I_1 being the modified Bessel functions of zero and first order, respectively [Ephraim 84].

In (D.1), the time and frequency indexes p and ω_k have been omitted for reasons of compactness. The spectral gain depends on two parameters, $\mathcal{R}_{post}(p, \omega_k)$ and $\mathcal{R}_{prio}(p, \omega_k)$ evaluated in each short-time frame and for all spectral bins. These two parameters are interpreted as follows:
The a posteriori Signal-to-Noise Ratio (or a posteriori SNR) $\mathcal{R}_{post}(p, \omega_k)$ given by

$$\mathcal{R}_{post}(p, \omega_k) = \frac{|X(p, \omega_k)|^2}{v(\omega_k)} - 1 \tag{D.2}$$

Where $v(\omega_k)$ denotes the noise power at frequency ω_k. Eq. (D.2) indicates that $\mathcal{R}_{post}(p, \omega_k)$ is a local estimate of the SNR computed from the data in the current short-time frame. Note that in the original papers by Ephraim and Malah, the definition of the a posteriori parameter is slightly different [Ephraim 84]. The definition of Eq. (D.2) was preferred because it allows a simpler interpretation of $\mathcal{R}_{post}(p, \omega_k)$.
The so-called a priori Signal-to-Noise Ratio (or a priori SNR) $\mathcal{R}_{prio}(p, \omega_k)$ represents the information on the unknown spectrum magnitude gathered from previous frames, and is evaluated in the "decision-directed" approach [Ephraim 84] by

$$\begin{aligned} \mathcal{R}_{prio}(p, \omega_k) = (1-\alpha)\, P\left[\mathcal{R}_{post}(p, \omega_k)\right] + \\ \alpha\, \frac{|G(p-1, \omega_k) X(p-1, \omega_k)|^2}{v(\omega_k)} \end{aligned} \tag{D.3}$$

Where $P[x] = x$ if $x \geq 0$ and $P[x] = 0$ otherwise. As $\mathcal{R}_{post}(p, \omega_k)$ defined by (D.2) is not necessarily positive, the operator P guarantees that $\mathcal{R}_{prio}(p, \omega_k)$ is always non-negative or equivalently, that the expression of the gain given by (D.1) is valid. On the second line of (D.3), $G(p-1, \omega_k) X(p-1, \omega_k)$ corresponds to the noiseless signal spectrum value as estimated in the previous frame. The term $|G(p-1, \omega_k) X(p-1, \omega_k)|^2 / v(\omega_k)$ thus corresponds to an estimation of the SNR in the frame of index $p-1$. $\mathcal{R}_{prio}(p, \omega_k)$ is therefore an estimate of the SNR that takes into account the current short-time frame, with weight $(1-\alpha)$, and the result of the processing in the previous frame, with weight α. On the basis of simulations, the parameter α was set by the authors to about 0.98.

For standard suppression rules, the gain applied to each short-time spectral coefficient depends only on the signal level $|X(p, \omega_k)|^2$ measured in the current frame: The gain can be expressed as a function of $\mathcal{R}_{post}(p, \omega_k)$. Fig. D.1 displays such suppression characteristics for the power subtraction and the so-called Wiener suppression rules [Mc Aulay 80] [Vary 85]. The two curves of Fig. D.1, although they correspond to different strategies, illustrate the same intuitive principle that those points where the SNR is close to $-\infty$ dB are the ones that should be attenuated. These two curves are strongly related because the Wiener gain is the square of the power subtraction gain [Mc Aulay 80].

The connection between the EMSR and more standard suppression rules is made clearer by plotting the gain of the EMSR versus the a priori SNR (in their original papers [Ephraim 83] [Ephraim 84] the authors used a reverse representation). The alternate representation of Fig. D.2 highlights the respective influence of the two parameters of the EMSR:

1. The a priori SNR is the dominant parameter: Strong attenuations are obtained only if $\mathcal{R}_{prio}(p, \omega_k)$ is low (left half of Fig. D.2), and low attenuations are obtained only if

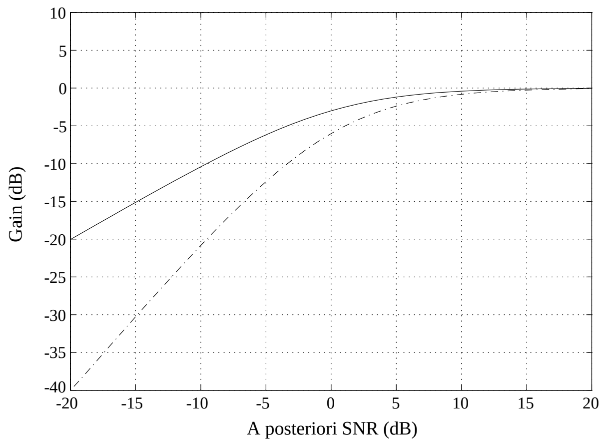

Figure D.1: Gain versus a posteriori signal-to-noise ratio; solid line: Power subtraction; dashed line: Wiener.

$\mathcal{R}_{prio}(p,\omega_k)$ is high (right half of Fig. D.2). Moreover, note that the overall shape of the gain is similar in Figs. D.2 and D.1 (although it must be stressed that the abscissa corresponds to $\mathcal{R}_{post}$ in Fig. D.1 and to $\mathcal{R}_{prio}$ in Fig. D.2).

2. The a posteriori SNR acts as a correction parameter whose influence is limited to the case where the a priori SNR is low (left half of Fig. D.2). The surprising point is that this correction effect acts to the opposite of what is intuitively expected: The larger $\mathcal{R}_{post}(p,\omega_k)$, the stronger the attenuation. This over-attenuation is a consequence of the disagreement between the a priori and the a posteriori SNR's. Why this counter-intuitive behavior is actually useful will be explained later.

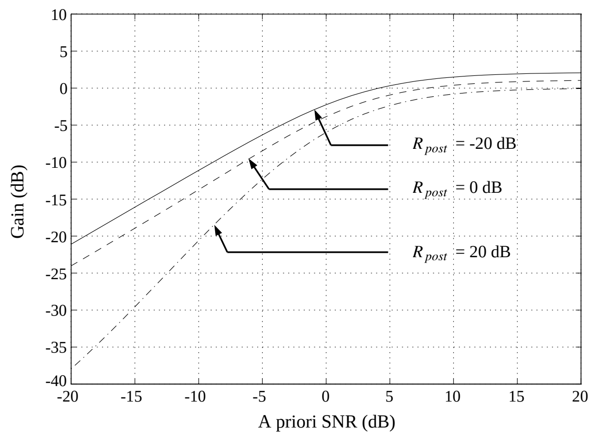

Figure D.2: EMSR gain versus a priori SNR, for different values of the a posteriori SNR; topmost curve: $\mathcal{R}_{post}(p,\omega_k) =$ -20 dB; middle curve: $\mathcal{R}_{post}(p,\omega_k) =$ 0 dB; bottom curve: $\mathcal{R}_{post}(p,\omega_k) =$ 20 dB.

Comparison between Figs. D.1 and D.2 indicates that the EMSR is very close to the Wiener suppression rule, *evaluated as a function of* $\mathcal{R}_{prio}(p,\omega_k)$, when $\mathcal{R}_{post}(p,\omega_k)$ is 20 dB (bottom curves in the two figures). This remains true for values of $\mathcal{R}_{post}(p, \omega_k)$ above 20 dB. Conversely when $\mathcal{R}_{post}(p, \omega_k)$ is -20 dB, the EMSR gets very close to the power subtraction suppression rule evaluated as a function of $\mathcal{R}_{prio}(p, \omega_k)$ (top curves in the two figures). This is actually true for values of $\mathcal{R}_{post}(p, \omega_k)$ below -5 dB. In practice, it can be considered that the EMSR corresponds to a smooth transition between the two suppression rules of Fig.D.1, the a priori SNR $\mathcal{R}_{prio}(p, \omega_k)$ controls the x-coordinate along the suppression characteristics, while the a posteriori SNR $\mathcal{R}_{post}(p, \omega_k)$ controls the transition between the two asymptotic curves.

D.3 Elimination of the musical noise

D.3.1 The smoothing effect in the EMSR

The a priori SNR is evaluated by the non-linear recursive relation of (D.3). An experimental study of (D.3) indicates two different behaviors for the a priori SNR:

1. When $\mathcal{R}_{post}(p, \omega_k)$ stays below or is sufficient close to 0 dB, the a priori SNR corresponds to a highly smoothed version of the a posteriori SNR over successive short-time frames. As a consequence the variance of $\mathcal{R}_{prio}(p, \omega_k)$ is much smaller than that of $\mathcal{R}_{post}(p, \omega_k)$.

2. On the contrary when $\mathcal{R}_{post}(p, \omega_k)$ is much larger than 0 dB, the a priori SNR follows the a posteriori SNR with a simple delay of one short-time frame. To see that, note that when the a priori SNR is high, the attenuation brought to the spectrum is negligible (right part of Fig. D.2). Then, (D.3) reduces to

$$\mathcal{R}_{prio}(p, \omega_k) \approx (1 - \alpha)\mathcal{R}_{post}(p, \omega_k) \; + \; \alpha \frac{|X(p-1, \omega_k)|^2}{v(\omega_k)}$$

   As $\mathcal{R}_{post}(p, \omega_k) \gg 1$, this can be written as

$$\mathcal{R}_{prio}(p, \omega_k) \approx (1 - \alpha)\mathcal{R}_{post}(p, \omega_k) \; + \; \alpha \mathcal{R}_{post}(p-1, \omega_k)$$

   Finally, the parameter α being generally chosen very close to 1, we can make the following approximation

$$\mathcal{R}_{prio}(p, \omega_k) \approx \alpha \mathcal{R}_{post}(p-1, \omega_k) \tag{D.4}$$

These two different behaviors of $\mathcal{R}_{prio}(p, \omega_k)$ are visible on the example of Fig. D.3. Notice how in the left-hand part of the figure, the variance of $\mathcal{R}_{prio}(p, \omega_k)$ is much lower than that of $\mathcal{R}_{post}(p, \omega_k)$, while on the right hand part, $\mathcal{R}_{prio}(p, \omega_k)$ follows $\mathcal{R}_{post}(p, \omega_k)$ with a one frame delay.

The smoothness of the a priori SNR helps reducing the musical noise effect. In the parts of the short-time spectrum corresponding to noise only, the a posteriori SNR is $-\infty$ dB in average, which corresponds to the case 1 above: Due to the smoothing behavior, the a priori SNR has a significantly reduced variance. Because the attenuation of the EMSR depends mainly on the value of the a priori SNR, the attenuation itself does not exhibit large variations over successive frames. As a consequence, the musical noise (sinusoidal components appearing and disappearing rapidly over successive frames) is reduced.

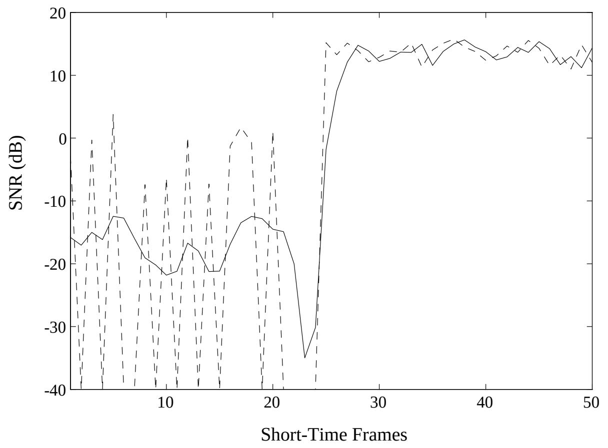

Figure D.3: Signal-to-noise ratios in successive short-time frames; dashed curve: A posteriori SNR; solid curve: A priori SNR. For the first 25 short-time frames, the analyzed signal contains only noise at the displayed frequency, for the next 25 frames, a component with 15 dB SNR emerges at the displayed frequency. Parameter α is set to 0.98.

The idea of calculating the attenuation from the short-time spectrum averaged over successive frames was also exploited in [Boll 79]. However the superiority of the EMSR lies in the non-linearity of the averaging procedure. When the signal level is well above the noise level, (D.3) becomes equivalent to a mere one-frame delay and $\mathcal{R}_{prio}(p, \omega_k)$ no longer is a smoothed SNR estimate which is important in the case of non-stationary signals.

D.3.2 Protection from local overtaking

The preceding results remain true if the EMSR gain function G in (D.3) is replaced by the Wiener suppression rule, evaluated as a function of $\mathcal{R}_{prio}(p, \omega_k)$ [Ephraim 84]. However, simulations show that this is not the case when the power subtraction rule is used. Because the power subtraction attenuation is too small for values of the SNR around 0 dB (about -3 dB), the a priori SNR undergoes less smoothing and still exhibits important fluctuations.

In the EMSR, another effect helps eliminating the musical noise. In the frequency bands containing only noise, we have seen that the a priori SNR is about -15 dB in average (see Fig. D.3). In that case, improbable high values of the a posteriori SNR are assigned an increased attenuation: In the left half of Fig. D.2, the attenuation increases for high values of the a posteriori SNR (values above 0 dB). This over-attenuation is all the more important as $\mathcal{R}_{prio}(p, \omega_k)$ is small. Thus, values of the spectrum higher than the average noise level are 'pulled down'.

This feature of the EMSR is particularly important for the recordings where the background noise is non stationary (e.g. recordings of old analog disks). The use of the EMSR avoids the appearance of local bursts of musical noise whenever the noise exceeds its average characteristics.

D.4 Influence of the parameters

D.4.1 Influence of α

The choice of the value of parameter α is guided by a trade-off between the degree of smoothing of parameter $\mathcal{R}_{prio}(p, \omega_k)$ in noisy areas, and the acceptable level of transient distortion brought to the signal.

Simulations show that when the analyzed signal contains only noise at a given frequency, both the average value and the standard deviation of the a priori SNR are proportional to $(1 - \alpha)$ when α is sufficiently close to one (above 0.9). As a result, in order to counter the musical noise effect one will choose values of α as close to one as possible.

On the other hand, when a signal component appears abruptly, the EMSR reacts immediately by raising the gain from a low value to a value close to 1, only if the SNR of the signal component is larger than $1/(1-\alpha)$. For signal components with lower SNR, simulations show that $\mathcal{R}_{prio}(p, \omega_k)$ takes a longer time to reach its final value. This results in an unwanted attenuation of low amplitude signal components during transient parts. The approximate limit of $1/(1 - \alpha)$ is found by considering the study case where the a posteriori SNR is a deterministic quantity which equals zero before frame index p_0 and has a fixed value of $\mathcal{R}$ for short-time frames with index $p \geq p_0$. As the gain of the EMSR is null before p_0, we have from (D.3)

$$\mathcal{R}_{prio}(p_0, \omega_k) = (1 - \alpha)\mathcal{R}$$

if this first value satisfies $\mathcal{R}_{prio}(p_0, \omega_k) \gg 1$, the gain of the EMSR evaluated at frame index p_0 is already close to 1 (see Fig. D.2). The condition that guarantees that there is no signal attenuation during the transient is thus $(1 - \alpha)\mathcal{R} \gg 1$.

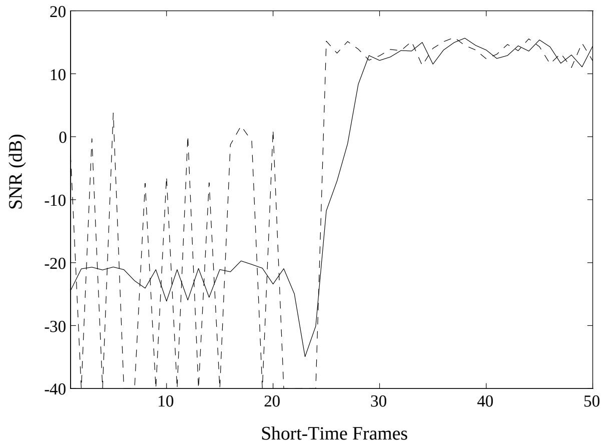

Figure D.4: Signal-to-noise ratios in successive short-time frames; dashed curve: A posteriori SNR; solid curve: A priori SNR. The analyzed signal is the same as in Fig. 3. Parameter α is set to 0.998.

The influence of parameter α appears clearly when comparing Figs. D.3 and D.4. In Fig. D.4, the factor $(1 - \alpha)$ is divided by 10 compared to the case of Fig. D.3. The average value of $\mathcal{R}_{prio}(p, \omega_k)$ when noise is present drops from approximately -15 dB for the case of Fig. D.3 to

-25 dB for Fig. D.4. The variance of $\mathcal{R}_{prio}(p, \omega_k)$ is also strongly reduced in Fig. D.4. But there is now an important delay between the appearance of the transient component and the time when $\mathcal{R}_{prio}(p, \omega_k)$ raises significantly above 0 dB. As a consequence, the signal component is incorrectly attenuated in the first short-time frames following the transient. In practice, the use of such a value of parameter α results in audible modifications of the signal transients.

It should be noted that a more important overlap between successive windows reduces the transient distortion as the same number of short-time frame results in a shorter time delay. As a consequence, an overlap of 2/3 or more is sometimes preferred to the standard 50 % setting [Valiere 91]. However, the variation of the overlap factor gives only slight perceptual differences because only the low level transient components are distorted when reasonable values of α are used; for example with $\alpha = 0.98$, the limit of $1/(1-\alpha)$ results in a SNR value of 15 dB.

D.4.2 Residual noise level

In the original paper by Ephraim and Malah, the gain function of (D.1) is tabulated for values of both signal-to-noise ratios between -15 dB and 15 dB [Ephraim 84]. The lower bound of this table is in fact a key parameter for the a priori SNR. Despite the smoothing performed by the procedure of (D.3), $\mathcal{R}_{prio}(p, \omega_k)$ still has some irregularities that can generate a perceptible low level musical noise. A simple solution to this problem consists in constraining the a priori SNR to be larger to a threshold $\mathcal{R}_{(min)}$. In practice, the value of $\mathcal{R}_{(min)}$ is chosen to be larger than the average a priori SNR in the frequency bands containing noise only. As a consequence, in the frequency bands containing noise only, the average value of the constrained a priori SNR is close to $\mathcal{R}_{(min)}$. Furthermore, in the same frequency bands, most values of the a posteriori SNR are below 0 dB, and the gain function of the EMSR is close to the power subtraction whose squared gain can be shown to be equal to the SNR for low SNR values [Mc Aulay 80]. As a result, in the frequency bands containing noise only, the average squared gain is close to $\mathcal{R}_{(min)}$. $1/\mathcal{R}_{(min)}$ can therefore be interpreted as the average noise power reduction.

When α equals 0.98, the average value of $\mathcal{R}_{prio}(p, \omega_k)$ is of -15 dB, and a value of $\mathcal{R}_{(min)}$ around -15 dB is sufficient to eliminate the musical noise phenomenon. But $\mathcal{R}_{(min)}$ could as well be set to a larger value with the effect of raising the level of the residual noise. The possibility to control the level of the residual noise is important for old recordings where the preservation of a certain amount of background noise is often judged as a positive aspect.

D.5 Conclusion

We have presented an analysis of the different mechanisms that counter the musical noise effect in the suppression rule proposed by Ephraim and Malah. The major factor was found to be the non-linear smoothing procedure used to obtain a more consistent estimate of the SNR. With an appropriate choice of parameter α, the use of the smoothing procedure doesn't generate audible distortion in the signal. However, low level signal components actually undergo a measurable over-attenuation during abrupt transients. This transient distortion is hardly perceptible and more precise listening tests would be necessary to decide wether it is useful or not to use an overlap factor larger than 50%. Finally it was shown that the attenuation function proposed by Ephraim and Malah avoids the appearance of the musical noise phenomenon even when the background noise is poorly stationary.

Bibliographie

[Abel 91] J. S. Abel and J. O. Smith. Restoring a clipped signal. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1991.

[Aes 77] The phonograph and sound recording after one-hundred years. *J. Audio Eng. Soc., Centennial Issue*, vol. 25 (10/11), October/November 1977.

[Allen 77] J. B. Allen and L. R. rabiner. A unified approach to short-time Fourier analysis and synthesis. *Proc. IEEE*, vol. 65 (11), pp. 1558–1564, November 1977.

[Basseville 88] M. Basseville. Detecting changes in signals and systems–a survey. *Automatica*, vol. 24 (3), pp. 309–326, 1988.

[Basseville 89] M. Basseville. Distance measures for signal processing and pattern recognition. *Signal processing*, vol. 18 (4), pp. 349–369, 1989.

[Beerends 92] J. G. Beerends and J. A. Stemerdink. A perceptual audio quality measure based on a psychoacoustic sound representation. *J. Audio Eng. Soc.*, vol. 40 (12), December 1992.

[Bellanger 76] M. G. Bellanger, G. Bonnerot, and M. Coudreuse. Digital filtering by polyphase network: Applications to sample-rate alteration and filter bank. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24 (2), pp. 109–114, 1976.

[Benade 76] Arthur H. Benade. *Fundamentals of musical acoustics.* Oxford University Press, 1976.

[Blair Benson 88] K. Blair Benson, editor. *Audio engineering handbook.* McGraw-Hill, 1988.

[Boll 79] S. F. Boll. Suppression of acoustic noise in speech using spectral substraction. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27 (2), pp. 113–120, 1979.

[Botte 88] M. C. Botte, G. Canevet, and L. Demany. *Psychoacoustique et perception auditive.* INSERM, Paris, 1988.

[Bourdier 88] R. Bourdier. *Analyse temps / fréquence, filtrage et synthèse numériques de signaux de parole. Application au filtrage, à la réduction de bruit, et à la restauration d'enregistrements anciens.* PhD thesis, Université du Maine, Le Mans, 1988.

[Brandenburg 88] K. Brandenburg. High quality sound coding at 2. 5 bit/sample. *Preprints of the AES 84th Convention*, Paris, 1988.

[Brillinger 81] D. R. Brillinger. *Time Series Data Analysis and Theory.* Holden-Day, expanded edition, 1981.

[Brock 84] G. Brock and Nannestad. A concerted approach to historical recordings. *Gramophone*, vol. January, pp. 925–927, 1984.

[Canagarajah 91] C. N. Canagarajah and P. J. W. Rayner. A single-input hearing aid based on the auditory perceptual features to improve speech intelligibility in noise. *IEEE ASSP Workshop on applications of signal processing to audio and acoustics*, Mohonk, 1991.

[Cappe 91] O. Cappé and A. Chaigne. Perceptual effects of noise disturbances on phase spectrum in stft analysis/synthesis procedures. application to restoration processes. *IEEE ASSP Workshop on applications of signal processing to audio and acoustics*, Mohonk, October 1991.

[Cappe 92] O. Cappé. On the use of variable length windows in the restoration of audio recordings. *Colloque C1, supplément au Journal de Physique III, Volume 2*, pp. 101–104. Deuxième Congrès Français d'Acoustique, Avril 1992.

[Cappe 93a] O. Cappé. Enhancement of musical signals degraded by background noise, using long-term behavior of the short-term spectral components. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. I–217–I–220, 1993.

[Cappe 93b] O. Cappé and J. Laroche. Evaluation of short-time spectral attenuation techniques for the restoration of musical recordings. Submitted for publication in IEEE Trans. Speech and Audio Processing, 1993.

[Cappe 94] O. Cappé. Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. To appear in IEEE Trans. Speech and Audio Processing, April 1994.

[CEDAR 91] Système CEDAR. Présentation du système lors de la 90ème convention de l'AES, Paris, Février 1991.

[Charbit 90] M. Charbit. *Eléments de théorie du Signal: les signaux aléatoires.* Collection Pédagogique de Télécommunication. Ellipse, Paris, 1990.

[Cohen 89] L. Cohen. Time-frequency distributions – A review. *Proc. IEEE*, vol. 77 (7), pp. 941–981, 1989.

[Combes 89] J. M. Combes, A. Grossmann, and Ph. Tchamitchian. *Wavelets: Time-frequency methods and phase space.* Springer-Verlag, 1989.

[Crochiere 83] R. E. Crochiere and L. R. Rabiner. *Multirate digital signal processing.* Prentice-Hall, 1983.

[Daubechies 88] I. Daubechies. Orthogonal bases of compactly supported wavelets. *Commun. on Pure and Applied Mathematics*, vol. 41, pp. 909–996, 1988.

[Daubechies 90] I. Daubechies. The wavelet transform, time-frequency localisation and signal analysis. *IEEE Trans. Inform. Theory*, vol. 36 (5), pp. 961–1005, 1990.

[Delmas 91] J. P. Delmas. *Eléments de théorie du Signal: les signaux déterministes.* Collection Pédagogique de Télécommunication. Ellipse, Paris, 1991.

[Depalle 91] P. Depalle. *Analyse, modélisation et synthèse des sons basées sur le modèle source-filtre.* PhD thesis, Université du Maine, Le Mans, 1991.

[Deutsch 82] Diana Deutsch, editor. *The psychology of music.* AP series in cognition and perception. Academic Press, 1982.

[Di Claudio 91] E. D. Di Claudio, G. Orlandi, F. Piazza, and A. Uncini. Optimal weighted ls ar estimation in presence of impulsive noise. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 3149–3152, 1991.

[Ephraim 83] Y. Ephraim and D. Malah. Speech enhancement using optimal non-linear spectral amplitude estimation. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 1118–1121, Boston, 1983.

[Ephraim 84] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32 (6), pp. 1109–1121, 1984.

[Ephraim 85] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33 (2), pp. 443–445, 1985.

[Feiten 89] B. Feiten. Spectral properties of audio signals and masking with aspect to bit data reduction. *Preprints of the AES 86th Convention*, Hamburg, 1989.

[Fesler 83] J. C. Fesler. Electrical reproduction of acoustically recorded cylinders and disks, Part 2. *J. Audio Eng. Soc.*, vol. 31 (9), September 1983.

[Flamenco 93] Flamenco, restauration historique. Disque publié à l'occasion du XXIème Congrès d'Art Flamenco, Flamenco en France (FLAM 9309), 1993.

[Flandrin 89] P. Flandrin. Ondelettes, spectrogrammes et lissages de la distribution de wigner-ville. *Douzième colloque GRETSI*, pp. 5–7, 1989.

[Gallagher 81] N. C. Gallagher and G. L. Wise. A theoretical analysis of the properties of median filters. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29 (6), pp. 1136–1141, 1981.

[Geckinli 78] N. C. Geckinli and D. Yavuz. Some novel windows and a concise tutorial comparison of window families. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 26 (6), pp. 501–507, 1978.

[George 92] E. B. George and M. J. T. Smith. Analysis-by-synthesis/Overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones. *J. Audio Eng. Soc.*, vol. 40 (6), pp. 497–516, 1992.

[Godsill 92] S. J. Godsill and P. J. W. Rayner. A bayesian approach to the detection and correction of error bursts in audio signals. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. II–261—II–264, 1992.

[Godsill 93] S. J. Godsill and P. J. W. Rayner. Frequency-based interpolation of sampled signals with applications in audio restoration. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. I–209—I–212, 1993.

[Goodman 86] D. J. Goodman, O. G. Jaffe, G. B. Lockhart, and W. C. Wong. Waveform substitution techniques for recovering missing speech segments in packet voice communications. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 105–108, 1986.

[Griffin 84] D. W. Griffin and J. S. Lim. Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32 (2), pp. 236–242, April 1984.

[Grossmann 89] A. Grossmann, R. Kronland-Martinet, and J. Morlet. Reading and understanding continuous wavelet transforms. J. M. Combes, A. Grossmann, and Ph. Tchamitchian, editors, *Wavelets. Time-frequency methods and phase space.* Springer-Verlag, 1989.

[Hall 80] Donald E. Hall. *Musical acoustics: An introduction.* Wadsworth, 1980.

[Harris 78] F. J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proc. IEEE*, vol. 66 (1), pp. 51–83, 1978.

[Hua 90] Y. Hua and T. K. Sarkar. Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38 (5), pp. 814–824, May 1990.

[Jacobs 92] Rémi Jacobs. Communication personnelle, Novembre 1992.

[Jansen 86] A. J. E. M. Jansen, R. N. J. Veldhuis, and . B. Vries. Adaptive interpolation of discrete-time signals that can be modelled as autoregressive processes. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34 (2), pp. 317–330, 1986.

[Jessel 85] M. Jessel. Enregistrement et reproduction des sons. *Impact : science et société (Revue publiée par l'UNESCO)*, no. 138/139, 1985.

[Johnston 88] J. D. Johnston. Transform coding of audio signals using perceptual noise criteria. *IEEE J. Selec. Areas. Commun.*, vol. 6 (2), pp. 314–323, 1988.

[Kay 88] Steven M. Kay. *Modern spectral estimation.* PH signal processing series. Prentice Hall, 1988.

[Kay 93] Steven M. Kay. *Fundamentals of statistical signal processing: Estimation theory.* PH signal processing series. Prentice-Hall, 1993.

[Kinzie 73] G. R. Kinzie and D. W. Gravereaux. Automatic detection of impulse noise. *J. Audio Eng. Soc.*, vol. 21 (3), pp. 181–184, 1973.

[Kronland Martinet 87] R. Kronland-Martinet, , J. Morlet, and A. Grossmann. Analysis of sound patterns through wavelets transforms. *Int. J. Patt. Recog. Art. Int.*, vol. 1 (2), pp. 97–126, 1987.

[Lagadec 83] R. Lagadec and D. Pelloni. Signal enhancement via digital signal processing. *Preprints of the AES 74th Convention*, New York, 1983.

[Laroche 89] J. Laroche. *Etude d'un système d'analyse et de synthèse utilisant la méthode de Prony. Application aux intruments de musique de type percussif.* PhD thesis, Ecole Nationale Supérieure des Télécommunications, Paris, 1989.

[Laroche 93a] J. Laroche. Signal audio et parole. Document du module SAP, TELECOM Paris, Département SIGNAL, 1993.

[Laroche 93b] J. Laroche. The use of the matrix pencil method for the spectrum analysis of musical signals. *J. Acoust. Soc. Am.*, October 1993.

[Le May 91] P. Le May. Restauration d'enregistrements audio dégradés par des clics. Dossier long de l'option SIGNAL, TELECOM Paris, Juin 1991.

[Lim 79] J. S. Lim and A. V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proc. IEEE*, vol. 67 (12), pp. 1586–1604, December 1979.

[Lim 83] J. S. Lim, editor. *Speech enhancement.* Prentice-Hall signal processing series. Prentice-Hall, 1983.

[Lim 86] J. S. Lim. Speech enhancement (preconference lecture). *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 3135–3142, 1986.

[Lutz 89] P. Lutz. Statistiques expérimentales. Cours ESE 3306, Gif-sur-Yvette, 1989.

[Mahieux 89] Y. Mahieux. *Algorithmes de codage par transformée pour la réduction de débit des voies son haute qualité.* PhD thesis, Université de Rennes 1, Rennes, 1989. Thèse 329.

[Makhoul 75] J. Makhoul. Linear prediction: A tutorial review. *Proc. IEEE*, vol. 63 (4), pp. 561–580, 1975.

[Malvar 92] H. S. Malvar. *Signal processing with lapped transforms.* Artech House, 1992.

[Marzio 88] Y. Marzio. Restauration en numérique. *HIFI VIDEO*, Juin 1988.

[Masson 87] J. Masson. Bancs de filtres numériques pour l'analyse et la synthèse des signaux. *Onzième colloque GRETSI*, pp. 17–20, 1987.

[Mc Aulay 80] R. J. Mc Aulay and M. L. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28 (2), pp. 137–145, April 1980.

[Meillier 93] Jean-Louis Meillier. *Analyse/Synthèse de signaux percussifs par un modèle source-filtre.* PhD thesis, Université du Maine, 1993.

[Meulengracht 76] H. Meulengracht and Madsen. On the transcription of old phonograph wax records. *J. Audio Eng. Soc.*, vol. 24 (1), pp. 27–32, 1976.

[Montresor 90] S. Montresor, J. C. Valière, and M. Baudry. Détection et suppression des bruits impulsionnels appliqués à la restauration d'enregistrement anciens. *Colloque de physique C2, supplément au n 2, tome 51*, pp. 757–760. Premier Congrès Français d'Acoustique, Février 1990.

[Montresor 91] S. Montresor. *Etude de la transformée en ondelettes dans le cadre de la restauration d'enregistrements anciens et de la détermination de la fréquence fondamentale de la parole.* PhD thesis, Université du Maine, Le Mans, 1991.

[Moore 82] B. C. J. Moore. *An introduction to the psychology of hearing.* Academic Press, second edition, 1982.

[Moorer 78] J. A. Moorer. The use of the phase vocoder in computer music applications. *J. Audio Eng. Soc.*, vol. 26 (1/2), pp. 42–45, 1978.

[Moorer 86] J. A. Moorer and M. Berger. Linear-phase bandsplitting: Theory and applications. *J. Audio Eng. Soc.*, vol. 34 (3), pp. 143–152, 1986.

[Mourjopoulos 92] J. Mourjopoulos, G. Kokkinakis, and M. Paraskeyas. Noisy audio signal enhancement using subjective spectra. *Preprints of the AES 92nd Convention*, Vienna, 1992.

[Nawab 88] S. H. Nawab and T. F. Quatieri. Short-time Fourier transform. J. S. Lim and A. V. Oppenheim, editors, *Adanced topics in signal processing*, chapter 6. Prentice-Hall, 1988.

[Nieminen 87a] A. Nieminen, P. Heinonen, and Y. Neuvo. Suppression and detection of impulse type interference using adaptive median hybrid filter. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 117–120, 1987.

[Nieminen 87b] A. Nieminen, M. Miettinen, P. Heinonen, and Y. Neuvo. Music restoration using median type filters with adaptive filter substructures. V. Cappellini and A. Constrantinides, editors, *Digital signal processing-87*. North-Holland, 1987.

[NoNoise 91] Système NoNoise. Présentation du système lors de la 90ème convention de l'AES, Paris, Février 1991.

[Oppenheim 89] A. V. Oppenheim and R. W. Schafer. *Discrete-time signal processing.* Prentice-Hall signal processing series. Prentice-Hall, international edition, 1989.

[Owen 83] T. Owen. Electrical reproduction of acoustically recorded cylinders and disks, Part 1. *J. Audio Eng. Soc.*, vol. 31 (4), April 1983.

[Paillard 92] B. Paillard, P. Mabilleau, S. Morisette, and J. Soumagne. Perceval: Perceptual evaluation of the quality of audio signals. *J. Audio Eng. Soc.*, vol. 40 (1/2), pp. 21–31, 1992.

[Papoulis 62] A. Papoulis. *The Fourier integral and its applications.* McGraw-Hill electronic sciences series. McGraw-Hill, 1962.

[Papoulis 91] A. Papoulis. *Probability, random variables, and stochastic processes.* McGraw-Hill, New York, 3rd edition, 1991.

[Pasian 84] F. Pasian and A. Crise. Restoration of signals degraded by impulse noise by mean of a low-distorsion non-linear filter. *Signal Processing*, vol. 6 (1), pp. 67–76, 1984.

[Petersen 81] T. L. Petersen and S. F. Boll. Acoustic noise suppression in the context of a perceptual model. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, volume 1086–1088, 1981.

[Petersen 83] T. L. Petersen and S. F. Boll. Critical band analysis-synthesis. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 31 (3), pp. 656–663, 1983.

[Picinbono 83] B. Picinbono and W. Martin. Représentation des signaux par amplitude et phase instantanées. *Ann. Télécommun.*, vol. 38 (5-6), pp. 179–190, 1983.

[Pollard 82] H. F. Pollard and E. V. Jansson. Analysis and assessment of musical starting transients. *Acustica*, vol. 51 (5), pp. 249–262, 1982.

[Porter 84] J. E. Porter and S. F. Boll. Optimal estimator for spectral restoration of noisy speech. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 18A.2.1—18A.2.4, 1984.

[Portnoff 81a] R. Portnoff. Short-time Fourier analysis of sampled speech. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29 (3), pp. 364–373, 1981.

[Portnoff 81b] R. Portnoff. Time-scale modifications of speech based on short-time Fourier analysis. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29 (3), pp. 374–390, 1981.

[Preis 84] D. Preis and H. Polchlopek. Restoration of nonlinearly distorted magnetic recordings. *J. Audio Eng. Soc.*, vol. 32 (1/2), pp. 26–30, 1984.

[Press 92] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical recipes in C : the art of scientific computing.* Cambridge University Press, second edition, 1992.

[Rayner 91] P. J. W. Rayner and S. J. Godsill. The detection and correction of artefacts in degraded gramophone recordings. *IEEE ASSP Workshop on applications of signal processing to audio and acoustics*, Mohonk, 1991.

[Rioul 92] O. Rioul and P. Duhamel. Fast algorithms for discrete and continuous wavelet transforms. *IEEE Trans. Inform. Theory*, vol. 38 (2), pp. 569–586, March 1992.

[Roys 78] H. E. Roys, editor. *Disc recording and reproduction*, volume 12 of *Benchmark papers in Acoustics.* Dowden, Hutchinson & Ross Inc., 1978.

[Scharf 91] Louis L. Scharf. *Statistical signal processing: detection, estimation, and time series analysis.* Addison-Wesley, 1991.

[Schroeder 79] M. R. Schroeder, B. S. Atal, and J. L. Hall. Optimising digital speech coders by exploiting masking properties of the human ear. *J. Acoust. Soc. Am.*, vol. 66 (6), pp. 1647–1652, 1979.

[Schuller 91] D. Schüller. The ethics of preservation, restoration, and re-issues of historical sound recordings. *J. Audio Eng. Soc.*, vol. 39 (12), pp. 1014–1017, 1991.

[Serra 89] X. Serra. *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition.* PhD thesis, CCRMA Department of Music, Stanford University, Stanford, California, 1989. Report No. STAN-M-58.

[Serra 90] X. Serra and J. Smith. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music J.*, vol. 14 (4), pp. 12–24, Winter 1990.

[Spiegel 68] M. R. Spiegel. *Mathematical Handbook of Formulas and Tables.* Schaum's outline series. McGraw-Hill, 1968.

[Stockham 75] T. G. Stockham, T. M. Canon, and R. B. Ingebretsen. Blind deconvolution through digital signal processing. *Proc. IEEE*, vol. 63 (4), pp. 678–692, 1975.

[Vaidyanathan 87] P. P. Vaidyanathan. Quadrature mirror filter banks, m-band extensions and perfect-reconstruction techniques. *IEEE ASSP Magazine*, vol. 4 (3), pp. 4–20, 1987.

[Valiere 90] J. C. Valière, S. Montresor, and J. F. Allard. Présentation d'une méthode de suppression des bruits de surface sur les anciens enregistrements de musique. *Colloque de physique C2, supplément au n 2, tome 51*, pp. 761–764. Premier Congrès Français d'Acoustique, Février 1990.

[Valiere 91] J. C. Valière. *La restauration des enregistrements anciens par traitement numérique – Contribution à l'étude de quelques techniques récentes.* PhD thesis, Université du Maine, Le Mans, 1991.

[Van Trees 68] H. L. Van Trees. *Detection, Estimation, and Modulation Theory, Part I.* J. Wiley & Sons, 1968.

[Vary 85] P. Vary. Noise suppression by spectral magnitude estimation – Mechanism and theoretical limits. *Signal Processing*, vol. 8 (4), pp. 387–400, 1985.

[Vaseghi 88a] S. Vaseghi and P. J. W. Rayner. A new application of adaptive filters for the restoration of archived gramophone recordings. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 2548–2551, 1988.

[Vaseghi 88b] S. L. Vaseghi. *Algorithms for restoration of archived gramophone recordings.* PhD thesis, University of Cambridge, Department of engineering, 1988.

[Vaseghi 89] S. L. Vaseghi and P. J. W. Rayner. The effects of non-stationary signal characteristics on the performance of adaptive audio restoration systems. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 377–380, 1989.

[Vaseghi 90] S. Vaseghi and P. J. W. Rayner. Detection and suppression of impulsive noise in speech communication systems. *IEE Proceedings, Part I*, vol. 137 (1), pp. 38–46, February 1990.

[Vaseghi 92] S. Vaseghi and R. Frayling-Cork. Restoration of old gramophone recordings. *J. Audio Eng. Soc.*, vol. 40 (10), pp. 791–801, 1992.

[Veldhuis 88] R. N. J. Veldhuis. *Adaptive restoration of unknown samples in discrete-time signals and digital images*. PhD thesis, Katholieke Universiteit te Nijmegen, Eindhoven, 1988.

[Veldhuis 90] R. Veldhuis. *Restoration of lost samples in digital signals*. Prentice-Hall international series in acoustics, speech, and signal processing. Prentice-Hall, New York, 1990.

[Wang 82] D. L. Wang and J. S. Lim. The unimportance of phase in speech enhancement. *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30 (4), pp. 679–681, 1982.

[Wright 89] M. Wright. Putting the byte on noise. *AUDIO*, March 1989.

[Zwicker 80] E. Zwicker and E. Terhardt. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am.*, vol. 68 (5), pp. 1523–1525, 1980.

[Zwicker 81] E. Zwicker and R. Feldtkeller. *Psychoacoustique, l'oreille récepteur d'information*. Masson, 1981.

[Zwicker 91a] E. Zwicker, H. Fastl, U. Widmann, K. Kurakata, S. Kuwano, and S. Namba. Program for calculating loudness according to din 45631 (iso 532b). *J. Acoust. Soc. Jpn. (E)*, vol. 12 (1), pp. 39–42, 1991.

[Zwicker 91b] E. Zwicker and U. T. Zwicker. Audio engineering and psychoacoustics: Matching signals to the final receiver, the human auditory system. *J. Audio Eng. Soc.*, vol. 39 (3), pp. 115–126, 1991.