Traitements pour la réduction de bruit. Application à la communication parlée.

Cyril Plapous

N° d'ordre: 3261

THÈSE

présentée

devant l'université de Rennes 1

pour obtenir

le grade de : DOCTEUR DE L'UNIVERSITÉ DE RENNES 1
Mention : TRAITEMENT DU SIGNAL ET TÉLÉCOMMUNICATIONS

par

Cyril PLAPOUS

Équipe d'accueil : France Télécom R&D - TECH/SSTP
École doctorale : MATISSE
Composante universitaire : UNIVERSITÉ DE RENNES 1 - IRISA/ENSSAT

Titre de la thèse :

Traitements pour la réduction de bruit.
Application à la communication parlée.

soutenue le 12 Décembre 2005 devant la commission d'examen

Mme :	Régine	LE BOUQUIN-JEANNÈS	Présidente
MM. :	Olivier	CAPPÉ	Rapporteurs
	Dirk	VAN COMPERNOLLE	
MM. :	Gaël	MAHÉ	Examinateurs
	Claude	MARRO	
	Pascal	SCALART	

Remerciements

C’est par ces quelques dernières lignes que je clos ma thèse ainsi que ces trois années passées à France Télécom R&D. Ainsi, je souhaite tout d’abord remercier Dominique Massaloux et Jean-Pierre Petit (merci pour la bouteille !) pour m’avoir accueilli dans leur équipe. Je remercie également André Gilloire pour l’intérêt qu’il a porté à mes travaux et pour son soutien scientifique tout au long de ma thèse.

Je tiens tout particulièrement à remercier Claude Marro pour avoir encadré ma thèse durant ces trois années ainsi que pour tout ce qu’il a fait pour moi. Je remercie également Pascal Scalart qui m’a fourni l’opportunité de réaliser cette thèse et qui a assuré le rôle de directeur de thèse. Tous deux ont su me faire profiter de leur expérience et de leurs conseils avisés et ce dans une très bonne ambiance que j’ai vraiment énormément appréciée.

Un grand merci à Olivier Cappé et Dirk Van Compernolle pour avoir accepté la lourde tâche de rapporteur. Merci également à Régine le Bouquin-Jeannès pour avoir accepté de faire partie du jury en tant que présidente et à Gaël Mahé en tant qu’examinateur. Merci à tous les membres du jury pour l’intérêt qu’ils ont porté à mon travail ainsi que pour leurs remarques constructives qui ont contribué à améliorer la qualité de cette thèse.

J’aimerais exprimer ma gratitude à Stéphane Ragot et Alexandre Guérin, Stéphane pour m’avoir suggéré une idée très intéressante qui a été exploitée avec succès et Alexandre pour m’avoir aimablement autorisé à m’inspirer d’une partie de son très bon travail de synthèse pour une partie de ma thèse.

Un merci général à toute l’équipe TPS pour leur très bon accueil et pour avoir supporté la période intensive de tests subjectifs que je leur ai fait subir, aidé dans cette tâche par Jean-Luc Garcia (merci d’avoir été solidaire !). Merci en particulier à Janine Denmat et à Noëlle Mevelle qui lui a succédé pour leur soutien logistique, leur efficacité et surtout leur gentillesse.

Un grand merci également à tous ceux, nombreux, qui ont partagé mon bureau où il a toujours régné une bonne ambiance, du moins je l’espère ! Je termine en remerciant tous ceux que j’ai eu le plaisir de côtoyer à FT que ce soit d’un point de vue professionnel et/ou personnel et qui ont su rendre ces années très agréables de par les moments que nous avons partagés. Il serait trop long de tous les citer mais je suis certain qu’ils se reconnaîtront !

Résumé

Avec l'avènement des télécommunications mobiles grand public, le besoin d'améliorer la prise de son, notamment en réduisant la gêne due au bruit, s'est fait de plus en plus présent. Les techniques de réduction du bruit sont soumises à un compromis entre le niveau effectif de réduction et la distorsion qui affecte le signal de parole. Au vu des performances actuelles, il est souhaitable de supprimer plus de bruit tout en conservant un niveau de dégradation acceptable du signal restauré, ceci en particulier lorsque le niveau de bruit est important. Les techniques qui ont suscité le plus d'intérêt au cours de ces 30 dernières années sont les approches par atténuation spectrale à court terme qui consistent à modifier une transformée à court terme du signal bruité en utilisant une règle de suppression. L'essor de cette famille de techniques s'explique essentiellement par le fait qu'elles permettent de respecter les contraintes de temps réel et de complexité inhérentes aux applications de communication parlée.

La première partie de ce document est consacrée à l'analyse des techniques majeures de réduction du bruit par atténuation spectrale à court terme. Ce sera notamment l'occasion d'identifier les limitations, points de blocage et autres défauts de ces méthodes ainsi que de montrer qu'il existe une marge de progression intéressante en terme de qualité par rapport à ces différents points clés. La seconde partie est essentiellement consacrée à la description et l'analyse de solutions originales proposées en réponse aux limitations identifiées dans la première partie. Un soin particulier a également été apporté à la mise en œuvre qui fait partie intégrante des techniques de réduction de bruit et qui conditionne la qualité du signal restauré.

L'analyse des limitations des techniques de réduction du bruit actuelles a permis de dégager plusieurs approches originales permettant de résoudre tout ou partie des problèmes identifiés. Ainsi, l'introduction de nouveaux modèles statistiques, adaptés aux signaux de parole et de bruit, pour déterminer l'expression d'une règle de suppression permet d'obtenir des résultats sensiblement meilleurs qu'en utilisant le modèle gaussien classique. Un problème d'ordre plus général concerne les défauts des estimateurs du rapport signal à bruit, paramètre fondamental qui conditionne les performances des techniques de réduction de bruit. La suppression de ces défauts conduit effectivement à une limitation des distorsions de la parole. Cependant, le signal restauré souffre toujours de certaines dégradations dues notamment aux erreurs d'estimation du bruit et à l'impact de la phase. En effet, l'estimation du bruit, qui constitue une étape clé des techniques de réduction de bruit, souffre de nombreuses limitations surtout lorsque le bruit n'a pas un caractère stationnaire. Dans une moindre mesure, la phase, qui est souvent négligée, a aussi une influence importante dans l'estimation du signal de parole, en particulier lorsque le niveau de bruit est élevé. Une approche originale qui tire parti de la structure voisée du signal de parole pour limiter les distorsions harmoniques engendrées par les techniques

classiques est proposée et permet de dépasser les limites de performances des techniques classiques.

Outre ces nouvelles approches, leur mise en œuvre conditionne également la qualité finale du signal restauré. Plusieurs points sensibles sont donc soulevés et des solutions sont données qui permettent d'éviter de nombreux artefacts ("clics", nasalisation, bruit musical) désagréables. Les approches proposées sont évaluées en utilisant des critères objectifs dont les résultats sont au besoin validés par des tests subjectifs. Les résultats obtenus montrent des améliorations significatives par rapport aux performances des techniques de référence.

Abstract

Because of the growing importance of mobile telecommunication, the need to enhance the sound pick-up, especially by reducing the inconvenience due to the noise, becomes more important. Short-term spectral attenuation techniques are widely used since they make it possible to respect real time constraints imposed by conversational speech applications. However, Noise reduction techniques are subject to a compromise between the effective level of noise reduction and the distorsion of the enhanced speech. It would be interesting to overcome this limitation and remove more noise while keeping an acceptable level of distorsion, particularly when the noise level is high.

In the first part of this thesis, short-term spectral attenuation techniques are analyzed and their limitations are underlined. This analysis shows that it is possible to outperform state of the art techniques. In the second part, several new approaches are described and fully analyzed. We show that these techniques are able to solve partly or completely the limitations identified in the first part. We also take care of the design of noise reduction techniques that govern the global quality of the enhanced signal.

The analysis of limitations in current noise reduction techniques allowed to find several new approaches able to solve them. The use of statistical models matching noise and speech signals to determine a noise reduction rule gives better results than using the classical Gaussian model. A more important concern is the limitations of the signal-to-noise ratio estimators which is a very important parameter that governs the performance of noise reduction techniques. The solutions proposed to overcome these limitations effectively succeed to limit the speech distorsions. However, the enhanced signal still suffers from degradations due to noise estimation errors and the impact of the phase. In fact, the noise estimation used to compute the signal-to-noise ratio suffers from many limitations especially for non-stationary noise. The phase parameter, though it is usually neglected, is also important for speech estimation particularly when the noise level is high. An original approach that takes advantage of the harmonic structure of speech to limit the harmonic distorsion has been proposed and outperforms the classical techniques.

Besides, the design of these new approaches governs the global quality of the enhanced signal. Several solutions are proposed to avoid unpleasant artifacts such as “clicks” and musical noise. The performance of the proposed techniques is evaluated using objective measures completed, when required, by subjective tests. The obtained results show a significant improvement with respect to the reference techniques.

Table des matières

Notations

Divers

- $\mathrm{E}[X]$: espérance mathématique de X
- $\Re[X]$: partie réelle de la quantité complexe X
- $[.]$: partie entière
- $\mathrm{p}(X(p,k) \mid A,\theta)$: densité de probabilité de $X(p,k)$ sachant A et θ
- $\max(.,K)$: maximum par rapport à K
- $\min(.,K)$: minimum par rapport à K
- $\hat{\theta}$: estimation du paramètre θ
- $\|X\|$: norme L_2 de X
- $< X,Y >$: produit scalaire des vecteurs X et Y
- $x * y$: convolution de x et y
- C^p : caractérise une fonction continûment dérivable d'ordre p
- p : densité de probabilité
- δ : distribution de Dirac

Signal discret

- Fe : fréquence d'échantillonnage
- F_0 : fréquence fondamentale ou pitch
- $x(n)$: signal temporel discret
- p : indice temporel de trame
- $x(p,n)$: trame p du signal $x(n)$
- σ_x^2 : variance de x
- μ_x : moyenne de x
- f : fréquence
- k : indice fréquentiel de la fréquence discrète f_k
- $X(p,k)$: TFCT de $x(p,n)$
- $|X(p,k)|$: module de $X(p,k)$
- $\phi_X(p,k)$: phase de $X(p,k)$
- $G(p,k)$: gain spectral
- $\gamma_x(f)$: DSP de x
- $\gamma_x(k) = \mathrm{E}[|X(p,k)|^2]$: DSP de X

Abréviations

- 3G : 3ème génération de télécommunications mobiles
- 3GPP : the 3rd generation partnership project agreement
- AB ou ABX : test subjectif CCR avec une échelle réduite
- ACR : absolute category rating (type de test subjectif)
- CCR : comparison category rating (type de test subjectif)
- CMOS : comparative mean opinion score (note obtenue à l'issu d'un test subjectif CCR)
- DAV : détection d'activité vocale
- dB : décibel
- DC : distance cepstrale
- DCR : degradation category rating (type de test subjectif)
- DSP : densité spectrale de puissance
- ECG : électrocardiogramme
- EEG : électroencéphalogramme
- EQMM : erreur quadratique moyenne minimum
- FFT : fast Fourier transform (transformée de Fourier rapide)

- GMM : Gaussian mixture model (modèle de mélange de Gaussiennes)
- Hz : Hertz
- IFFT : inverse fast Fourier transform (transformée de Fourier inverse rapide)
- LPC : linear predictive coding (codage par prédiction linéaire)
- MMSE : minimum mean square error (*cf.* EQMM)
- OLS : technique de synthèse par overlap-and-save
- OLA : technique de synthèse par overlap-and-add
- RSB : rapport signal à bruit
- RTC : réseau téléphonique commuté
- SPL : sound pressure level (niveau de pression du son)
- SRI : système de référence intermédiaire
- TCD : transformée en cosinus discret
- TF : transformée de Fourier
- TFCT : transformée de Fourier à court terme
- TFCTI : transformée de Fourier à court terme inverse
- TFD : transformée de Fourier discrète
- TFDI : transformée de Fourier discrète inverse
- UIT : union internationale des télécommunications

Guide de lecture

Le premier chapitre du présent mémoire tient lieu d'introduction générale pour le reste du document. Il explore la problématique de la réduction de bruit et cible ses applications potentielles (partie 1.1). C'est aussi l'occasion de présenter les différentes familles de techniques monovoie existantes et de décrire plus précisément les techniques dites d'atténuation spectrale à court terme qui connaissent un succès important. L'essor de ces techniques est porté par l'existence d'algorithmes optimisés et de processeurs dédiés permettant une mise en œuvre qui respecte les contraintes imposées par les applications de communication parlée (partie 1.2), *i.e.* faible retard et complexité numérique raisonnable. Les objectifs concernant cette étude seront également exposés et coïncident bien sûr avec les limitations des techniques actuelles (partie 1.2). La dernière partie (1.3) de ce chapitre permet de souligner certaines caractéristiques du signal de parole et du mécanisme d'audition qui seront exploitées dans les chapitres suivants.

Les deux chapitres suivants (2 et 3) forment la première partie du présent document et sont consacrés à l'étude des techniques classiques de réduction du bruit par atténuation spectrale à court terme. Le chapitre 2 présente une vision unifiée ce ces techniques. Les parties 2.1 et 2.2 exposent le principe général et les paramètres communs à toutes ces approches, puis les grandes lignes de leur mise en œuvre sont présentées dans la partie 2.3. Les parties 2.4 et 2.5 constituent l'étude bibliographique à proprement parler des principales méthodes d'atténuation spectrale à court terme et des techniques d'estimation de la densité spectrale de puissance (DSP) du bruit.

Le chapitre 3 propose quant à lui une analyse des limitations majeures des approches classiques. Quatre limitations ont été identifiées qui laissent entrevoir une marge de progression importante en terme d'amélioration de la qualité des techniques de réduction de bruit. Ainsi, la partie 3.1 montre qu'il existe des modèles statistiques mieux adaptés aux signaux traités que celui utilisé couramment. Ensuite, la partie 3.2 met en avant les défauts (et avantages) des principaux estimateurs du rapport signal à bruit (RSB). La partie 3.3 montre que le fait d'estimer le bruit à long terme constitue en soi une limitation des techniques de réduction de bruit. Finalement, la partie 3.4 explore l'impact de la phase, généralement négligé, dans les méthodes par atténuation spectrale à court terme.

Les chapitres 4, 5 et 6 constituent la seconde partie de ce document essentiellement consacrée à la description et l'analyse de solutions originales basées sur les conclusions du chapitre 3. Le chapitre 4 regroupe les différentes solutions proposées. La partie 4.1 présente des approches utilisant des modèles statistiques adaptés au signal de parole et au bruit qui permettent d'obtenir des résultats sensiblement meilleurs qu'avec le modèle gaussien classique. Les parties 4.2, 4.3 et 4.4 correspondent à 3 approches se proposant de supprimer les défauts de l'estimateur du rapport signal à bruit le plus

couramment utilisé. Cet estimateur corrigé, le signal restauré souffre toujours de distorsions dues notamment à l'estimation de la DSP du bruit et à l'impact de la phase, deux des limitations identifiées dans le chapitre 3. Une approche est proposée pour résoudre ces problèmes. Elle n'améliore pas l'estimation du bruit ou de la phase mais tire parti de la structure voisée du signal de parole pour limiter les distorsions engendrées par les techniques classiques.

Le chapitre 5 est entièrement consacré à la mise en œuvre des techniques de réduction de bruit qui, si elle est bien réalisée, permet d'éviter certaines dégradations du signal restauré. Avec cet objectif, l'étape de synthèse sera détaillée permettant de souligner les contraintes à respecter pour éviter d'introduire des artefacts dans le signal restauré (partie 5.1). Une application particulière qui en découle sera présentée dans la partie 5.2. La nécessité et l'intérêt de conserver une partie du bruit original pour masquer les distorsions du signal de parole seront également évoqués dans la partie 5.3. La partie suivante (5.4) est consacrée à une solution permettant de contrôler la résolution du filtre de réduction de bruit et par ce biais le compromis entre artefacts et qualité du signal. La limitation de la résolution du filtre limite la présence d'artefacts mais entraîne un phénomène d'étouffement de la parole qui peut être efficacement supprimé en introduisant l'approche psychoacoustique dans les techniques de réduction de bruit classiques (partie 5.5). Finalement, la partie 5.6 présente une approche basée sur cette maîtrise de la résolution du filtre mais qui aurait également eu sa place dans le chapitre 4 comme amélioration de l'approche proposée dans la partie 4.5.

Le 6ème et dernier chapitre est consacré aux évaluations objectives et subjectives (l'oreille étant le juge final de la qualité de la restauration) des techniques proposées dans le chapitre 4 ainsi que des approches proposées dans les parties 5.5 et 5.6 du chapitre 5. En premier lieu, le corpus de test est détaillé dans la partie 6.1 puis les mesures objectives et les tests subjectifs utilisés pour quantifier la performance des différentes approches sont détaillés dans la partie 6.2. Finalement, les résultats sont présentés et analysés dans la partie 6.3. Nous verrons à ce propos qu'il existe une certaine hiérarchie dans les performances des approches proposées qui découle naturellement de leur interdépendance.

Chapitre 1

Problématique et contexte de la réduction de bruit

Avec l'avènement des télécommunications mobiles, le besoin d'améliorer la prise de son s'est fait de plus en plus présent. La partie 1.1 explique pourquoi la perturbation qualifiée de bruit est si gênante pour celui qui la subit et par conséquent pourquoi la réduction de bruit est nécessaire dans les communications modernes ainsi que dans bien d'autres applications d'ailleurs. Ensuite, la partie 1.2 présente les différentes familles de techniques de réduction de bruit et explique l'essor des techniques dites d'atténuation spectrale à court terme. Les défauts de cette famille de techniques sont également soulignés faisant ressortir par là même les nouveaux défis à relever pour améliorer leurs performances. Pour ce faire, il est indispensable de connaître les caractéristiques de la parole. Ainsi, dans la partie 1.3 nous verrons que ce signal est très structuré et qu'il est possible de le modéliser. Le mécanisme de l'audition sera aussi présenté car il est bien entendu que dans les applications de communication parlée le juge final de la qualité du signal restauré est l'oreille humaine.

1.1 Problématique de la réduction de bruit

L'essor fantastique des télécommunications de ces 20 dernières années a permis au grand public de bénéficier d'outils de communication mobiles. Il est désormais devenu possible et courant de téléphoner de partout (ou presque) dans des environnements aussi divers et variés que la rue, une gare ou bien encore une voiture. Cependant, tous ces lieux ne bénéficient pas du calme du salon où le téléphone fixe était autrefois cantonné. La gêne occasionnée par la perturbation qualifiée de bruit est généralement source d'inconfort et de fatigue pour les correspondants quand ce n'est pas l'intelligibilité même du message qui est remise en cause. De plus, la volonté de dématérialiser la prise de son (système mains-libres) va favoriser l'émergence du bruit du fait de l'augmentation de la distance entre la bouche et le microphone. Dans ces conditions, la conversation téléphonique peut s'avérer rapidement fastidieuse ce qui justifie le besoin d'un traitement à même de réduire la gêne des utilisateurs.

1.1.1 La gêne due au bruit

La gêne due au bruit est de nature différente pour la personne plongée dans l'ambiance bruitée (le locuteur) et pour celle qui subit les perturbations par l'intermédiaire de son téléphone (l'auditeur) [Beaugeant 1999b]. L'auditeur est le plus gêné car, contrairement au locuteur, il n'a aucun contrôle sur le milieu acoustique. Le locuteur garde un certain contrôle sur le bruit ambiant, il a la possibilité de hausser la voix (effet Lombard) ou de se réfugier dans un endroit moins bruyant. Dans le cas d'une communication mains-libres, il peut aussi focaliser son oreille sur le signal utile grâce aux capacités de localisation spatiale de l'oreille alors que l'auditeur subit complètement la perturbation sonore. Ce dernier est le plus pénalisé dans la mesure où la prise de son par microphone et sa restitution sont ponctuelles. L'ensemble du champ sonore (des sons provenant de toutes les directions) est donc intégré et restitué en une somme des perturbations. L'information de spatialisation ayant disparu, l'auditeur ne peut donc pas séparer l'information utile des différentes sources de bruit bien que l'oreille humaine en soit normalement capable. De plus, la superposition du bruit et de la parole réduit l'intelligibilité du message ce qui demande un effort constant de la part de l'auditeur et le fatigue rapidement.

Il faut souligner aussi que, dans tout réseau de communication, il y a une étape de codage du signal de parole qui est largement basée sur ses propriétés. En présence de bruit le signal à coder s'éloigne du modèle utilisé pour l'opération de codage-décodage et il en résulte une dégradation d'autant plus gênante que le niveau de bruit est élevé. Le timbre de la voix est altéré et le bruit est codé de façon peu naturelle ce qui le rend encore plus désagréable [Beaugeant 1999b]. Il y a donc tout intérêt à réduire les perturbations liées au bruit avant la phase de codage afin de faciliter sa tâche.

1.1.2 Les applications

Afin d'améliorer la qualité du signal transmis au correspondant distant, d'accroître son intelligibilité et de réduire la fatigue de ce dernier, il s'avère important de développer des systèmes de réduction de bruit dont le but consiste à extraire l'information utile en effectuant un traitement sur le signal d'observation bruité. En plus de ces applications de communication parlée, l'amélioration de la qualité du signal de parole s'avère également nécessaire pour la reconnaissance vocale, dont les performances sont fortement altérées lorsque l'utilisateur est plongé dans un environnement bruyant.

La réduction de bruit s'applique principalement au domaine du traitement du signal sonore (parole ou musique) dont voici une liste non exhaustive :

- téléconférence et visioconférence en milieu bruité (en salle dédiée ou bien à partir d'un ordinateur multimédia, *etc.*),
- téléphonie : traitement au niveau du terminal (terminaux classiques et terminaux mobiles, *etc.*) et au sein du réseau de transport,
- terminaux mains-libres (bureau, terminaux mobiles, terminaux fixes embarqués en véhicules, *etc.*),
- prise de son dans les lieux publics (gare, aéroport, rue, *etc.*),
- prise de son mains-libres dans les véhicules,
- reconnaissance de parole robuste à l'environnement acoustique,

– restauration d'enregistrements anciens,
– prise de son pour le cinéma et les médias (radio, télévision, par exemple pour le journalisme sportif ou les concerts, *etc.*).

Le principe de la réduction de bruit est également applicable à tous les domaines où l'on cherche à extraire une information utile à partir d'une observation bruitée. On peut notamment envisager les domaines suivants : imagerie sous-marine, télédétection sous-marine, traitement des signaux biomédicaux (EEG, ECG, imagerie biomédicale, *etc.*).

1.2 Trois décennies de travaux dans le domaine de la réduction de bruit

Ce qui ressort de près de 30 ans de travaux dans le domaine de la réduction de bruit est la suprématie des techniques dites d'atténuation spectrale à court terme. Pourtant il existe de nombreuses autres familles d'approches mais qui n'ont pas reçu autant d'attention. Nous allons donc exposer le principe général des différentes approches formant le domaine de la réduction de bruit.

1.2.1 Un tour d'horizon des différentes familles de techniques

Le choix d'une technique de réduction de bruit dépend en premier lieu du nombre d'observations disponibles. Les travaux menés au cours de cette thèse se restreignent aux techniques monovoies, sans doute le cas le plus courant mais aussi le plus critique. En effet, les approches multivoies disposent d'au moins deux observations, avec dans un cas idéal une référence de bruit, ce qui facilite d'autant la tâche des techniques de réduction de bruit. La faisabilité n'est pas toujours évidente dans la mesure où il faut placer un second micro suffisamment loin du locuteur pour obtenir une référence valide de bruit. Mais quand bien même cela est possible, le fait d'utiliser des microphones supplémentaires coûte cher. Dans le cas monovoie, seule l'observation bruitée est disponible. Une étape commune et indispensable à la majorité des techniques de réduction de bruit consiste alors à extraire des connaissances statistiques sur le bruit qui seront ensuite exploitées pour le supprimer. Ceci peut d'ailleurs se révéler particulièrement difficile lorsque le bruit est non-stationnaire et que son niveau est important. Sans être exhaustif, voici une courte présentation de ces familles de réduction de bruit monovoies.

– Des techniques de filtrage adaptatif, normalement bivoies, ont vu le jour en utilisant des astuces permettant d'obtenir une référence de bruit seul [Sambur 1978, Oppenheim 1994]. Malgré tout ces approches restent très minoritaires. On peut également citer des approches basées sur du filtrage de Kalman [Gabrea 2002, Deng 2005].
– Il est aussi possible de réduire le bruit en projetant le signal de parole bruité sur le sous-espace associé au signal utile [Ephraïm 1995b, You 2005]. Toutefois, ces approches sont complexes dans la mesure où elles requièrent la décomposition en éléments simples de matrices de corrélation. Pour limiter cette complexité et pour respecter la stationnarité du signal de parole, ces matrices doivent rester de taille raisonnable ce qui en pratique limite la résolution fréquentielle équivalente de ces techniques.
– Des techniques à base de modèle de Markov cachés (HMM pour hidden Markov model) ont également vu le jour [Ephraïm 1989, Ephraïm 1992]. Cependant, tout comme les approches à

base de mélange de gaussiennes (GMM pour gaussian mixture model) [Ding 2005] leurs performances sont dépendantes de l'adéquation entre le corpus d'apprentissage et les conditions réelles d'utilisation.

– Certaines approches basées sur les réseaux de neurones ont également été proposées [Wan 1998], cependant, on peut voir ces techniques comme des boîtes noires entraînées sur un corpus donné. Si les performances sont bonnes sur un corpus équivalent, dès que l'on s'en écarte il faut refaire l'entraînement du réseau de neurones.

– Dans les approches par atténuation spectrale à court terme, le traitement se fait dans un domaine transformé qui est généralement celui de Fourier. D'autres transformées sont exploitables, entre autres la TCD (transformée en cosinus discret) [Hasan 2002] et la transformée en ondelettes [Abry 1997, Donoho 1994], cette dernière autorisant une analyse multirésolution du signal. Cependant la TFD (transformée de Fourier discrète) reste la plus utilisée dans la mesure où des algorithmes optimisés de calcul très efficaces existent (FFT : transformée de Fourier rapide) et que les autres transformées n'ont pas démontré de gain plus significatif en performance.

1.2.2 Les contraintes de réalisation

La famille des approches par atténuation spectrale à court terme regroupe pratiquement l'ensemble des solutions utilisées dans les équipements industriels en raison de la simplicité des concepts mis en jeu et de la grande disponibilité d'outils de base (notamment la FFT) nécessaires à la programmation de ces techniques. La mise en œuvre de ces dispositifs conduit généralement à effectuer l'estimation d'une fonction de transfert du filtre de réduction de bruit, puis à réaliser le traitement de filtrage à partir d'une multiplication dans le domaine spectral. Il ne fait aucun doute que l'essor de ces techniques repose en premier lieu sur la possibilité d'effectuer facilement ces traitements en temps réel sur un processeur de traitement du signal. De plus, il faut respecter des contraintes de complexité assez drastiques. En effet, ce type de traitement peut être embarqué dans un terminal mobile auquel cas **sa complexité doit rester raisonnable pour pouvoir l'intégrer dans l'architecture correspondante et pour en limiter le coût global.** Par ailleurs, si le traitement est fait dans le réseau, moins il sera complexe et plus le nombre de voies traitées simultanément pour une puissance de calcul donnée sera important.

Il faut distinguer le temps de calcul, lié à la notion de temps réel, et le retard inhérent au traitement du signal sonore. Dans le cas de la réduction de bruit, ce retard s'ajoute à celui du codeur de parole, de l'annuleur d'écho acoustique (éventuellement et selon la stratégie utilisée) et bien sûr à celui de la transmission. **Pour que la communication reste interactive, il est nécessaire que le retard total soit le plus faible possible [Bouteille 2002].** Au delà de 200ms il devient très perturbant. Ainsi, on comprend bien que l'on cherche à le minimiser de façon à ne pas pénaliser la communication. Pour un système de réduction de bruit, un retard de plus de 40ms est déjà considéré comme excessif ce qui doit être gardé à l'esprit lors de la conception de ces techniques.

Bien entendu, même s'il faut toujours garder à l'esprit ces contraintes de complexité et de retard, un autre critère de choix concerne la qualité sonore disponible en sortie de traitement. En définitive, un compromis doit souvent être réalisé entre ces différents paramètres : complexité, retard et qualité.

1.2.3 Objectifs pour les nouvelles techniques de réduction de bruit

De façon générale, les techniques de réduction de bruit sont soumises à un compromis. En effet, **le fait de supprimer le bruit génère des distorsions du signal de parole. Plus on réduit le niveau de bruit et plus elles sont importantes.** Il faut donc trouver un compromis acceptable entre le niveau de réduction et la dégradation du signal. Le bruit ne peut donc pas être complètement supprimé, cela n'est d'ailleurs pas souhaitable dans la mesure où il fait aussi partie de l'information à transmettre. Cependant, au vu des performances actuelles, il est indispensable de supprimer un niveau plus important de bruit en conservant un niveau de dégradation acceptable. Ceci est vrai en particulier lorsque le niveau de bruit est important et que celui-ci est non-stationnaire.

Outre ce compromis, le signal restauré peut aussi être entaché de nombreux artefacts. Entre autres, **on a généralement le "choix" entre un bruit résiduel peu naturel et désagréable, qualifié de bruit musical, et un phénomène de réverbération qui touche le signal de parole** [Cappé 1993, Scalart 1996a]. Ce dernier est certainement moins gênant que le bruit musical mais reste tout de même un effet indésirable. Une mauvaise gestion de la mise en œuvre peut également causer de nombreux artefacts tels que des "clics", une nasalisation de la parole ou encore une certaine rugosité. Autant de défauts qu'il est donc souhaitable de supprimer en respectant des règles strictes pour la mise en œuvre.

1.3 La parole, vecteur de communication

La parole est un système structuré qui permet aux êtres humains de communiquer entre eux. L'information d'un message parlé est transmise par les fluctuations de la pression de l'air qui sont émises par l'appareil phonatoire, c'est le signal vocal. Ce signal est analysé par l'oreille et les informations résultantes sont transmises au cerveau qui les interprète. Au sens strict, le contenu d'un signal vocal est représenté uniquement par son intelligibilité. Dans un sens plus large, il faut aussi tenir compte de tout ce qui représente la qualité du signal vocal, c'est-à-dire les intonations, la prosodie et les perturbations du milieu ambiant. Nous allons ci-après décrire certaines propriétés de l'appareil phonatoire humain. Cette partie est inspirée de [Boite 1987] où le lecteur pourra, si nécessaire, trouver de plus amples informations.

1.3.1 Le mécanisme de phonation

Le signal vocal est le résultat de plusieurs actions conjuguées. En effet, la parole résulte de l'action coordonnée des appareils respiratoires et masticatoires contrôlés par le système nerveux central. L'appareil respiratoire fournit l'air qui est expiré par la trachée artère. L'air passe ensuite par le larynx où sa pression est modulée grâce aux cordes vocales qui déterminent la taille de l'ouverture (la glotte) par laquelle il peut passer. Finalement, l'air transite par le conduit vocal qui s'étend du pharynx aux lèvres pour devenir le signal vocal qui est émis par le locuteur. La figure 1.1 représente un schéma général de l'appareil phonatoire humain[1]. On peut distinguer deux grandes classes de sons émis par

1. `http://www.iict.ch/Tcom/Laboratoires/digivox2000/chap/chap5/synthese.htm`

l'appareil phonatoire : les sons voisés et les sons non voisés.

– Les sons voisés, de forte énergie, résultent de l'excitation du conduit vocal par des impulsions périodiques de pression liées aux oscillations des cordes vocales. À chaque impulsion la glotte s'ouvre brusquement et libère la pression accumulée en amont puis elle se referme plus graduellement.

– Les sons non voisés, beaucoup moins énergétiques que les sons voisés, résultent quant à eux de l'écoulement libre de l'air par la glotte qui reste ouverte. Le chuchotement est un mode de phonation particulier car la glotte reste ouverte en permanence, il n'y a donc pas de production de sons voisés pendant le chuchotement.

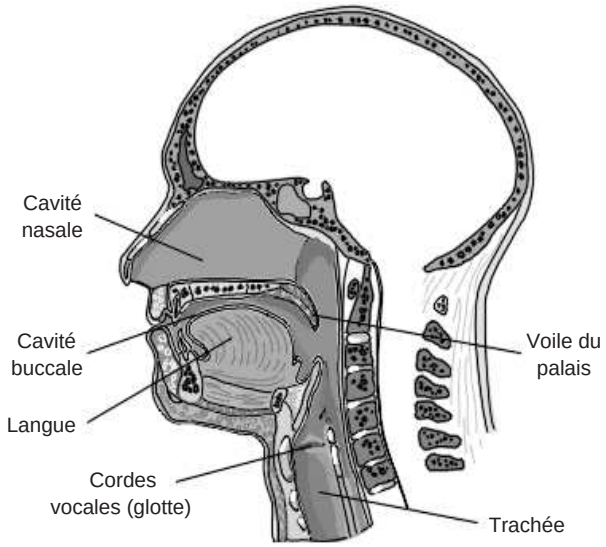

FIG. 1.1 – *Représentation de l'appareil phonatoire humain.*

Il est important de remarquer que le signal vocal n'est pas stationnaire, son évolution suit les déformations du conduit vocal. Cependant, **ces déformations sont suffisamment lentes pour que le signal vocal puisse être considéré comme stationnaire sur des périodes allant de 20 à 40ms**. On considère donc ce signal comme quasi-stationnaire sur de tels intervalles ce qui motive le choix du traitement par trames (de 20 à 40ms) du signal vocal pour de nombreuses applications, notamment pour la réduction de bruit et le codage de la parole.

1.3.2 Caractéristiques du signal de parole

L'analyse spectrale des sons voisés et non voisés donne des informations fondamentales dans le cadre du traitement de la parole et en particulier pour la réduction de bruit. Nous allons donc illustrer les principales caractéristiques spectrales des sons voisés et non voisés.

1.3.2.1 Les sons voisés

La hauteur d'un son voisé est liée à la fréquence de vibration des cordes vocales. Cette fréquence est appelée fréquence fondamentale ou pitch. Un son voisé est par définition un signal quasi-périodique qui possède un spectre fréquentiel très caractéristique, comme on peut le voir sur l'exemple de la figure 1.2. La première raie de ce spectre correspond au fondamental (F_0) et les raies suivantes à ses harmoniques (multiples de F_0). L'enveloppe de ces raies possède des maxima locaux appelés formants (localisés en F_i avec $i = 1,\ldots,4$). Si les trois premiers formants sont indispensables pour assurer l'intelligibilité du signal vocal, les formants d'ordres supérieurs jouent quant à eux un rôle important pour la qualité du signal vocal. Deux sons de même intensité et de même hauteur se distinguent par le timbre qui est déterminé par les amplitudes relatives des harmoniques du fondamental. La fréquence fondamentale (F_0) peut varier :

- de 80 à 200 Hz pour une voix masculine,
- de 150 à 450 Hz pour une voix féminine,
- de 200 à 600 Hz pour une voix d'enfant.

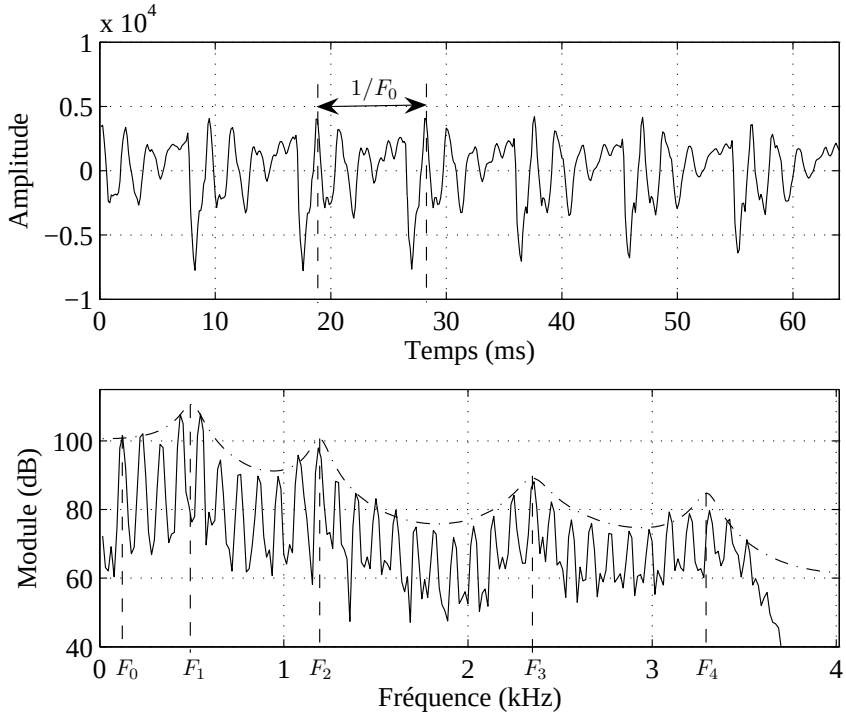

FIG. 1.2 – *Forme d'onde d'une trame de signal vocal voisé et son spectre d'amplitude. Les positions du fondamental (F_0) et des formants (F_1, F_2, F_3 et F_4) sont indiquées.*

1.3.2.2 Les sons non voisés

Un son non voisé ne présente pas de structure périodique, il peut être considéré comme un bruit blanc filtré par le conduit vocal. Son spectre ne possède donc pas de structure particulière et c'est

souvent dans les hautes fréquences que le spectre est le plus énergétique comme l'illustre la figure 1.3.

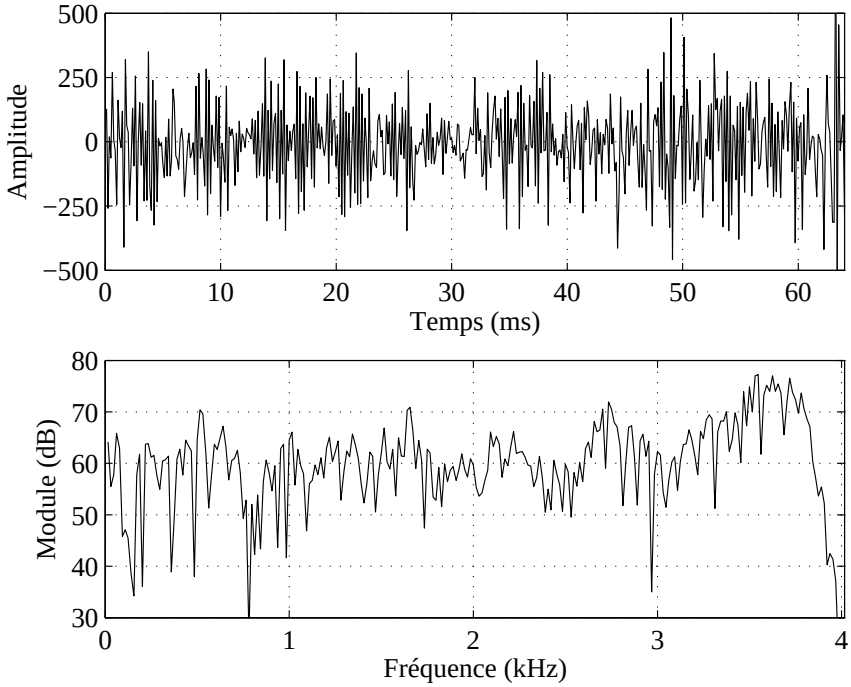

FIG. 1.3 – *Forme d'onde d'une trame de signal vocal non voisé et son spectre d'amplitude.*

1.3.3 Propriétés statistiques du signal de parole

À court terme, le signal vocal peut être considéré comme la réalisation particulière d'un processus aléatoire non-stationnaire. Ses statistiques moyennes doivent être estimées à long terme (au moins plusieurs secondes) et moyennées pour plusieurs locuteurs afin d'être fiables. Le signal de parole étant quasi-stationnaire on peut aussi définir des statistiques à court terme sur la durée d'une trame. Ces deux types de statistiques sont couramment utilisées dans les techniques de réduction de bruit, les statistiques à court et long terme portant des informations complémentaires.

Pour approcher les statistiques à long terme du signal vocal temporel, les fonctions de densité de probabilité suivantes peuvent être utilisées (la moyenne μ_x est supposée nulle) :

– loi de Gauss :

$$\mathrm{p}(x) = \frac{1}{\sqrt{2\pi}\sigma_x} exp\left(-\frac{x^2}{2\sigma_x^2}\right), \tag{1.1}$$

– loi de Laplace :

$$\mathrm{p}(x) = \frac{1}{\sqrt{2}\sigma_x} exp\left(-\frac{\sqrt{2}|x|}{\sigma_x}\right), \tag{1.2}$$

– loi Gamma :

$$p(x) = \left(\frac{\sqrt{3}}{8\pi\sigma_x|x|}\right)^{\frac{1}{2}} exp\left(-\frac{\sqrt{3}|x|}{2\sigma_x}\right), \tag{1.3}$$

où σ_x^2 est la variance de x. La figure 1.4 représente ces trois densités de probabilité, dans l'hypothèse d'une moyenne nulle et d'une variance unité, ainsi que la densité de probabilité expérimentale obtenue pour environ 50s de parole (prononcée par 4 locuteurs, 2 femmes et 2 hommes). **La loi Gamma**

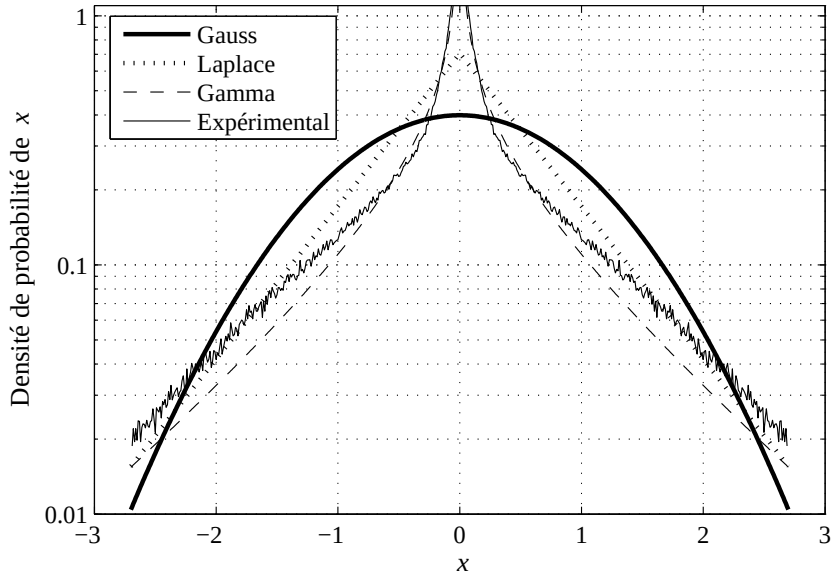

FIG. 1.4 – *Densité de probabilité à long terme du signal vocal (trait fin) et densités de probabilité utilisées pour l'approcher : loi de Gauss (trait fort), loi de Laplace (pointillé) et loi Gamma (tirets).*

approche fidèlement la loi expérimentale et la loi de Laplace en est aussi relativement proche. Cependant, dans la pratique, la loi de Gauss est très souvent choisie à cause des nombreuses simplifications que son utilisation apporte, et ce bien que cette loi soit relativement éloignée de la réalité. On peut noter que la valeur importante en $x = 0$ de la densité de probabilité à long terme du signal vocal s'explique par la présence de silence dans la parole continue.

1.3.4 Mécanisme de l'audition

Les ondes sonores sont recueillies par l'appareil auditif qui les transforme en influx nerveux alors transmis au cerveau provoquant ainsi les sensations auditives. L'oreille, représentée par la figure 1.5, est constituée de trois régions [2] :

– L'oreille externe : le conduit auditif s'ouvre au centre du pavillon qui reçoit, concentre et guide les ondes sonores.
– L'oreille moyenne : elle est séparée du conduit de l'oreille externe par une fine membrane, le tympan. Ses vibrations, dues aux ondes sonores, sont transmises mécaniquement par une chaîne de 3 osselets (le marteau, l'enclume et l'étrier) à la fenêtre ovale.

2. `http://ctn.ffessm.fr/oreilsch.html`

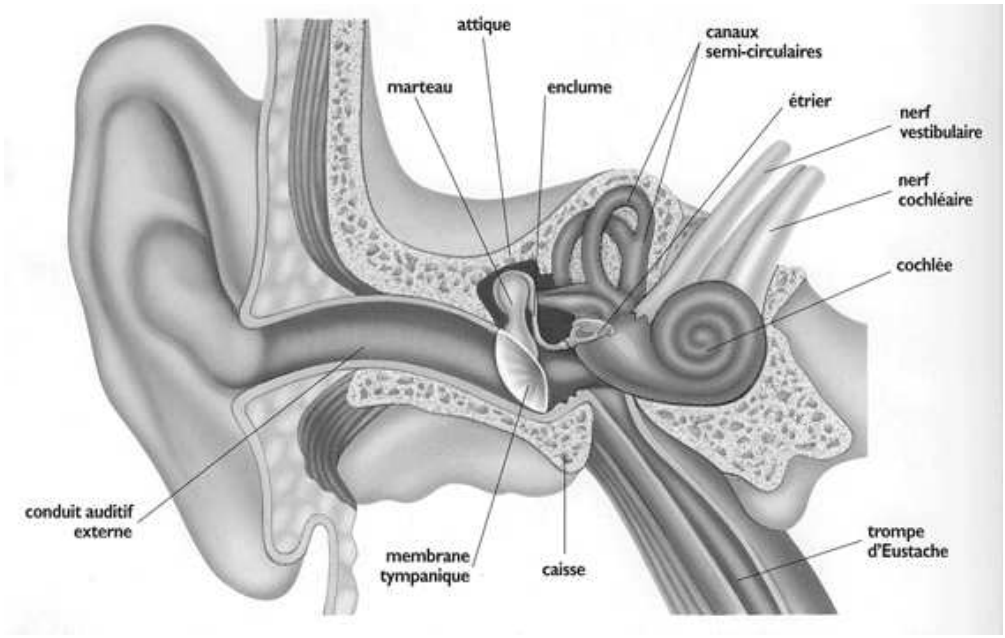

FIG. 1.5 – *Appareil auditif humain.*

– L'oreille interne : La cochlée ou limaçon contient les récepteurs auditifs qui baignent dans le liquide cochléaire dont le mouvement est imposé par celui de la fenêtre ovale. L'excitation des récepteurs est alors transformée en un influx nerveux qui est transmis au cerveau par le nerf cochléaire. En outre, les canaux semi-circulaires renferment des récepteurs sensibles aux mouvements de la tête et jouent un rôle dans le mécanisme de l'équilibre.

L'organe majeur de l'audition est la membrane centrale de la cochlée appelée membrane basilaire. Elle renferme l'organe de Corti qui est constitué d'environ 25000 cellules ciliées raccordées au nerf auditif et baignant dans le liquide cochléaire. Chacune de ces cellules possède une réponse en fréquence qui dépend de sa position sur la membrane. On peut donc dire qu'une transformation fréquence-espace s'effectue le long de cette membrane. La figure 1.6 représente la distribution des fréquences le long de la membrane basilaire d'une cochlée humaine[3]. En fonction de sa fréquence, la vibration a un effet maximal (résonance) en un point différent de la membrane basilaire : c'est la tonotopie passive. Quelques fréquences caractéristiques sont indiquées et permettent de remarquer que cette distribution est non-linéaire (la tendance est logarithmique). L'échelle des Barks correspond à une division de la membrane basilaire en intervalles de longueur constante (1,3mm). Un Bark correspond à une bande critique dont la bande passante équivalente varie en fonction de la fréquence. Cette échelle a été construite de façon à ce que chaque bande critique contribue de façon équivalente à la perception auditive. Le tableau 1.1 [Virag 1996] donne la bande passante fréquentielle correspondant à chaque bande critique. La relation entre l'échelle des Barks et celle des Hertz peut être approchée par la relation analytique approximative suivante [Schroeder 1979] :

$$B = 13 \operatorname{Arctan}(0{,}76 f) + 3{,}5 \operatorname{Arctan}\left(\frac{f}{7{,}5}\right)^2 \tag{1.4}$$

3. `http://www.iurc.montp.inserm.fr/cric/audition/`

où B s'exprime en Barks et f en kHz.

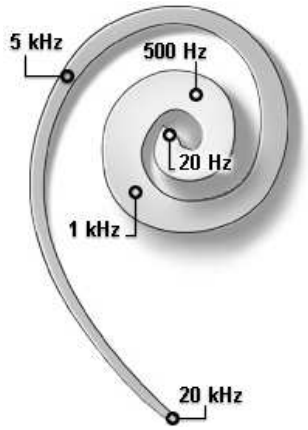

FIG. 1.6 – *Distribution des fréquences le long de la membrane basilaire.*

TAB. 1.1 – *Liste des bandes critiques.*

Bande critique numéro	Fréq. de coupure basse (Hz)	Fréquence centrale (Hz)	Fréq. de coupure haute (Hz)
1	0	50	100
2	100	150	200
3	200	250	300
4	300	350	400
5	400	450	510
6	510	570	630
7	630	700	770
8	770	840	920
9	920	1000	1080
10	1080	1170	1270
11	1270	1370	1480
12	1480	1600	1720
13	1720	1850	2000
14	2000	2150	2320
15	2320	2500	2700
16	2700	2900	3150
17	3150	3400	3700
18	3700	4000	4400
19	4400	4800	5300
20	5300	5800	6400
21	6400	7000	7700
22	7700	8500	9500
23	9500	10500	12000
24	12000	13500	15500
25	15500	19500	

1.3.5 Le masquage psychoacoustique

La notion de bande critique est liée au masquage psychoacoustique. Pour définir ce dernier, il est nécessaire d'introduire la notion de seuil absolu d'audition qui correspond au niveau minimum de pression du son (SPL pour sound pressure level) nécessaire pour détecter un son tonal pur dans un environnement calme. La figure 1.7 représente le seuil absolu d'audition, en dB SPL, en fonction de la fréquence. Les limites extrêmes du système auditif humain sont environ de 20 et 20000 Hz. **Le**

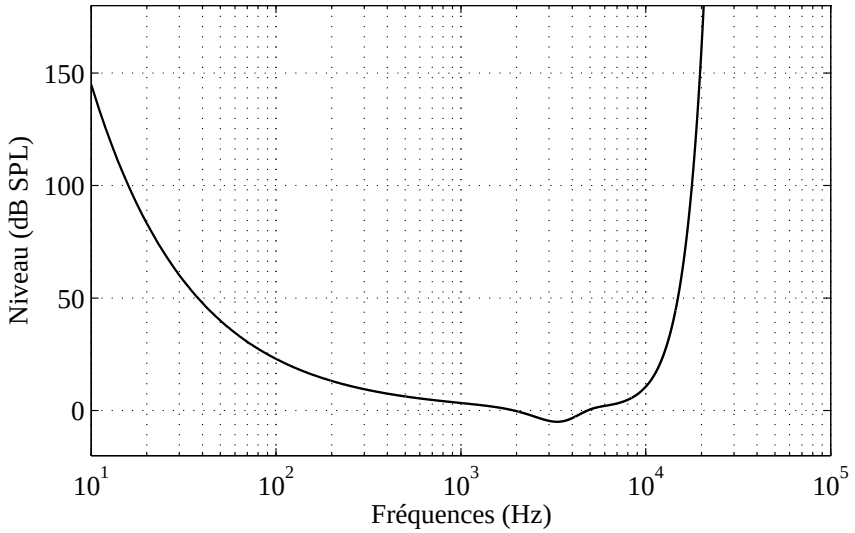

FIG. 1.7 – *Seuil absolu d'audition.*

masquage psychoacoustique intervient quand un signal dit masquant rend inaudible un signal d'énergie plus faible appelé signal masqué. Ce phénomène s'explique par une augmentation du seuil d'audition due au signal masquant. Le masquage peut se produire entre deux sons simultanés ou non [Painter 2000, Akbari Azirani 1995b].

1.3.5.1 Masquage simultané

Le masquage simultané apparaît lorsque deux sons émis en même temps possèdent des fréquences identiques ou relativement proches. En effet, lorsqu'un son crée une excitation d'un niveau suffisant sur la membrane basilaire dans une bande critique donnée, cela bloque la détection d'un son de plus faible niveau. L'effet d'un signal masquant est prépondérant dans la bande critique à laquelle il appartient. Cependant, s'il possède un niveau suffisant il peut aussi modifier le seuil d'audition dans les bandes adjacentes. Selon la nature des sons masquant et masqué (une tonale ou un bruit) il est possible de distinguer trois scénarios :

- tonale masquant un bruit,
- bruit masquant une tonale,
- bruit masquant un bruit.

Pour résumer, on peut dire que le masquage dépend de la différence d'énergie des deux signaux, de leur différence de fréquence et de leur distribution spectrale (tonale ou bruit). Le seuil de masquage fréquentiel résultant est toujours asymétrique. En effet, pour un son masquant à une fréquence donnée, le masquage est toujours plus important pour les fréquences supérieures que pour celles qui lui sont inférieures. Il est aussi d'autant plus important et couvre une plage plus grande de fréquences que l'intensité du son masquant est importante. Un son aigu de faible niveau est donc facilement masqué par un son plus grave de niveau élevé. On peut noter qu'il est plus difficile de masquer un bruit par une tonale que l'inverse, c'est pour cette raison que le calcul de la courbe de masquage d'un signal complexe (parole ou musique) prend en compte la nature du signal masquant. Il n'est pas nécessaire de détailler plus précisément ces trois scénarios car cela n'apporterait rien par rapport à l'utilisation que nous faisons du masquage psychoacoustique (*cf.* parties 2.4.3 et 5.5). Une présentation complète des phénomènes mis en jeu est disponible dans [Painter 2000]. Différentes techniques permettent de calculer la courbe de masquage d'un signal complexe, on peut notamment citer les plus courantes : la méthode de Johnston [Johnston 1988] et les deux méthodes de la norme ISO MPEG [ISO MPEG 1992].

1.3.5.2 Masquage non simultané

Ce type de masquage est beaucoup moins exploité car moins bien maîtrisé que le masquage simultané. Un son masquant peut en effet masquer un autre son qui le précède (masquage antérieur sur 1 à 2ms) ou qui le suit (masquage postérieur sur 50 à 300ms). Dans le cadre des techniques de réduction de bruit ce type de masquage ne sera pas pris en compte.

1.4 Conclusion

Ce premier chapitre a permis d'exposer les enjeux actuels des techniques de réduction de bruit et en particulier la nécessité d'améliorer la qualité de ces techniques pour faire face à des conditions bruitées toujours plus critiques de façon à assurer un certain confort de communication. Les applications de réduction de bruit ne se limitent d'ailleurs pas à la communication parlée et on peut par exemple citer le domaine de la reconnaissance vocale où un tel traitement est nécessaire pour améliorer la robustesse des moteurs de reconnaissance. La suprématie des techniques dites d'atténuation spectrale à court terme est fortement liée au fait que les outils mis en jeu (FFT notamment) permettent des réalisations peu coûteuses en calcul ce qui est évidemment intéressant. De cet avantage il découle la possibilité de réaliser le traitement en temps réel mais il faut en plus assurer un retard algorithmique faible pour garantir une communication interactive. Le chapitre suivant propose une revue des principales techniques qui forment le socle de l'atténuation spectrale à court terme.

Références

[Abry 1997] P. Abry, "Ondelettes et Turbulence. Multirésolutions, Algorithmes de Décomposition, Invariance d'Échelles," *Diderot Éditeur*, Paris, 1997.

[Akbari Azirani 1995b] A. Akbari Azirani, R. Le Bouquin Jeannes, et G. Faucon, "Optimizing Speech Enhancement by Exploiting Masking Properties of the Human Ear," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Detroit, États-Unis, Vol. 1, pp. 800–803, Mai 1995.

[Beaugeant 1999b] C. Beaugeant, "Réduction de Bruit et Contrôle d'Écho pour les Applications Radiomobiles," *Thèse de l'Université de Rennes 1*, 1999.

[Boite 1987] R. Boite, et M. Kunt, "Traitement de la Parole," *Presses Polytechniques Romandes, Complément au Traité d'électricité, Première édition*, 1987.

[Bouteille 2002] F. Bouteille, "Traitement de la Parole dans les Ponts de Conférence à Accès Hétérogènes Synchrones (RNIS, RTC) et Asynchrones (IP)," *Thèse de l'Université de Rennes 1*, 2002.

[Cappé 1993] O. Cappé, "Techniques de Réduction de Bruit pour la Restauration d'Enregistrements Musicaux," *Thèse de l'École Nationale Supérieure des Télécommunications*, Paris, Septembre 1993.

[Deng 2005] J. Deng, M. Bouchard, et T. Yeap, "Speech Enhancement using a Switching Kalman Filter with a Perceptual Post-Filter," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Philadelphie, États-Unis, Vol. 1, pp. 1121–1124, Mars 2005.

[Ding 2005] G.-H. Ding, X. Wang, Y. Cao, F. Ding, et Y. Tang, "Speech Enhancement Based on Speech Spectral Complex Gaussian Mixture Model," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Philadelphie, États-Unis, Vol. 1, pp. 165–168, Mars 2005.

[Donoho 1994] D. L. Donoho, et I. M. Johnstone, "Threshold Selection for Wavelet Shrinkage of Noisy Data," *IEEE Intl. Conf. On Engineering in Medicine and Biology Society*, Vol. 1, pp. A24-A25, Novembre 1994.

[Ephraïm 1989] Y. Ephraïm, D. Malah, et B.-H. Juang, "On the Application of Hidden Markov Models for Enhancing Noisy Speech," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 37, No. 12, pp. 1846–1856, Décembre 1989.

[Ephraïm 1992] Y. Ephraïm, "A Bayesian Estimation Approach for Speech Enhancement using Hidden Markov Models," *IEEE Trans. Speech Audio Processing*, Vol. 40, Issue 4, pp. 725–735, Avril 1992.

[Ephraïm 1995b] Y. Ephraïm, et H. L. Van Trees, "A Signal Subspace Approach for Speech Enhancement," *IEEE Trans. Speech Audio Processing*, Vol. 3, No. 4, pp. 251–266, Juillet 1995.

[Gabrea 2002] M. Gabrea, “Speech Signal Recovery in Colored Noise using an Adaptative Kalman Filtering,” *IEEE Canadian Conf. on Electrical and Computer Engineering*, Vol. 2, pp. 12–15, Mai 2002.

[Hasan 2002] M. K. Hasan, M. S. A. Zilany, et M. R. Khan, “DCT Speech Enhancement with Hard and Soft Thresholding Criteria,” *Electronics Lett.*, Vol. 38, No. 13, pp. 669–670, Juin 2002.

[ISO MPEG 1992] ISO/IEC, JTC1/SC29/WG11 MPEG “Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media at Up to about 1.5Mbit.s - Part 3 : Audio,” IS11172-3, 1992.

[Johnston 1988] J. D. Johnston, “Transform Coding of Audio Signals Using Perceptual Noise Criteria,” *IEEE J. on Select. Areas Commun.*, Vol. 6, No. 2, pp. 314–323, Février 1988.

[Oppenheim 1994] A. V. Oppenheim, E. Weinstein, K. C. Zangi, M. Feder, et D. Gauger, “Single Sensor Active Noise Cancellation,” *IEEE Trans. Speech Audio Processing*, Vol. 2, No. 2, pp. 285–290, Avril 1994.

[Painter 2000] T. Painter, et A. Spanias, “Perceptual Coding of Digital Audio,” *IEEE proc.*, Vol. 88, No. 4, Avril 2000.

[Sambur 1978] M. R. Sambur, “Adaptive Noise Cancelling for Speech Signals,” *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-26, No. 5, pp. 419–423, Octobre 1978.

[Scalart 1996a] P. Scalart, et J. Vieira Filho, “Speech Enhancement Based on a Priori Signal to Noise Estimation,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Atlanta, États-Unis, Vol. 2, pp. 629–632, Mai 1996.

[Schroeder 1979] M. R. Schroeder, B. S. Atal, et J. L. Hall, “Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear,” *J. Acoustical Society of America*, 66(6), pp. 1647–1652, Décembre 1979.

[Virag 1996] N. Virag, “Speech Enhancement Based on Masking Properties of the Human Auditory System,” *Thèse de l’École Polytechnique Fédérale de Lausanne*, 1996.

[Wan 1998] E. A. Wan, et A. T. Nelson, “Handbook of Neural Networks for Speech Processing,” *Ed. S. Katagiri, Artech House, Première édition*, 1998.

[You 2005] C. H. You, S. N. Koh, et S. Rahardja, “Signal Subspace Speech Enhancement for Audible Noise Reduction,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Philadelphie, États-Unis, Vol. 1, pp. 145–148, Mars 2005.

Chapitre 2

Atténuation spectrale à court terme

Ce chapitre est consacré à la présentation d'une vision unifiée des principales techniques de réduction de bruit par atténuation spectrale à court terme (limitée aux systèmes monovoie) dont l'esprit sera conservé tout au long de ce mémoire. Tout d'abord, le principe général de ces approches est exposé dans la partie 2.1, à cette occasion les notations et les hypothèses classiques seront introduites. Ensuite, la partie 2.2 permet de définir les quantités utilisées dans le calcul des fonctions de réduction de bruit. Ces quantités sont fondamentales dans la mesure où elles conditionnent la qualité du traitement. Les principales étapes de la mise en œuvre de ces techniques, communes à la majorité des approches, sont ensuite exposées dans la partie 2.3. Au vu des très nombreuses approches qui existent, l'inventaire proposé dans la partie 2.4 ne se veut pas exhaustif, seules sont présentées et analysées les approches majeures et il faut garder à l'esprit qu'il existe de nombreuses variantes pour chacune d'entre elles. Finalement, une sélection de techniques d'estimation du bruit est présentée dans la partie 2.5. Cette étape est en effet fondamentale car nécessaire au calcul des paramètres présentés dans la partie 2.2.

2.1 Principe de l'atténuation spectrale à court terme

Dans le cas monovoie considéré, **l'objectif consiste à estimer le signal de parole utile** $s(n)$**, celui-ci étant perturbé par un bruit additif** $b(n)$ **supposé indépendant du signal de parole, à partir du seul signal observé** $x(n)$:

$$x(n) = s(n) + b(n). \tag{2.1}$$

Les approches basées sur l'atténuation spectrale à court terme réalisent la réduction de bruit dans le domaine fréquentiel (ou spectral). Si les signaux sont stationnaires alors à partir de la relation temporelle (2.1) on peut écrire :

$$\gamma_x(f) = \gamma_s(f) + \gamma_b(f), \tag{2.2}$$

où $\gamma_x(f)$, $\gamma_s(f)$ et $\gamma_b(f)$ représentent les densités spectrales de puissance (DSP) respectives des signaux $x(n)$, $s(n)$ et $b(n)$. Cette représentation sous forme de DSP n'est malheureusement pas exploitable à cause de la non-stationnarité du signal de parole. En effet, s'il est acceptable de considérer le

bruit stationnaire (on verra dans la partie 3.3 qu'en réalité cette hypothèse est trop forte), la parole ne peut être considérée comme telle que sur de courtes durées. **Il devient alors possible d'exploiter la quasi-stationnarité de la parole sur des trames d'une durée de l'ordre de 20 à 40 ms.** C'est entre autres pour cette raison qu'une majorité des techniques de réduction de bruit par atténuation spectrale sont basées sur une analyse à court terme du signal traité. De nombreuses transformées à court terme sont disponibles mais si l'on prend en compte la nécessité qu'une transformée inverse existe ou bien encore la possibilité de modifier le signal bruité dans le domaine transformé alors le panel des transformées envisageables se réduit considérablement. La transformée de Fourier (TF) à court terme (TFCT) remplit les conditions requises et possède des mises en œuvre rapides et numériquement peu coûteuses ce qui explique qu'elle est très souvent utilisée, ce qui sera également le cas dans la suite de ce document. Chaque trame du signal temporel $x(n)$ peut donc être représentée dans le domaine fréquentiel par son module $|X(p,k)|$ et sa phase associée $\phi_X(p,k)$, où p représente l'indice temporel de la trame d'analyse courante et k le canal fréquentiel d'indice k ou autrement dit la fréquence discrète f_k. Dans le domaine fréquentiel, l'équation (2.1) considérée à la trame p peut donc s'exprimer ainsi :

$$|X(p,k)|e^{i\phi_X(p,k)} = |S(p,k)|e^{i\phi_S(p,k)} + |B(p,k)|e^{i\phi_B(p,k)}. \tag{2.3}$$

L'objectif de l'atténuation spectrale consiste alors à estimer le spectre à court terme du signal de parole $\hat{S}(p,k)$. On suppose qu'il est toujours possible d'estimer la DSP du bruit, soit sur une partie du signal d'observation ne contenant que du bruit soit de façon continue (*cf.* partie 2.5). Sans cette connaissance *a priori* de la DSP du bruit, généralement supposé stationnaire, il ne serait pas possible de réaliser une technique de réduction de bruit pour un système monovoie.

Sans faire aucune hypothèse supplémentaire sur les caractéristiques des signaux, il est possible d'obtenir des estimateurs du signal utile à partir d'une optimisation de certains critères. Il est aussi possible de poser des hypothèses supplémentaires sur les statistiques du signal de parole ou du bruit pour obtenir d'autres types d'estimateurs.

2.2 Définition des rapports signal à bruit (RSB)

En réalité, il n'existe pas de solution simple à l'estimation spectrale de $S(p,k)$, donc généralement un gain spectral $G(p,k)$ qui dépend du RSB est obtenu puis est appliqué au spectre bruité $X(p,k)$:

$$\hat{S}(p,k) = G(p,k)X(p,k). \tag{2.4}$$

Le gain spectral $G(p,k)$ a toujours le comportement asymptotique suivant :

- Une valeur importante du RSB indique qu'une forte composante de parole est présente par rapport au niveau du bruit, le gain $G(p,k)$ doit donc être proche de 1 pour préserver cette composante.
- Une valeur faible du RSB indique que la parole est absente ou très faible par rapport au niveau du bruit, le gain $G(p,k)$ doit donc apporter une atténuation importante ($G(p,k) \ll 1$) pour réduire l'effet du bruit.

Tout le problème consiste donc à estimer ce RSB. Selon les hypothèses choisies pour exprimer le gain spectral, deux types de RSB sont utilisés, le RSB *a posteriori* et le RSB *a priori* [McAulay 1980,

Ephraïm 1984] :

$$RSB_{post}(p,k) = \frac{|X(p,k)|^2}{\gamma_b(k)} = \frac{|X(p,k)|^2}{\mathrm{E}[|B(p,k)|^2]}, \tag{2.5}$$

$$RSB_{prio}(k) = \frac{\gamma_s(k)}{\gamma_b(k)} = \frac{\mathrm{E}[|S(p,k)|^2]}{\mathrm{E}[|B(p,k)|^2]}. \tag{2.6}$$

La quantité $RSB_{post}(p,k)$ représente le RSB de la trame courante en tenant compte du module carré du signal bruité et dépend donc du temps. La quantité $RSB_{prio}(k)$ quant à elle ne dépend pas du temps car elle exprime le RSB à long terme en supposant les statistiques du signal de parole utile connues *a priori*. **En pratique, la DSP du signal de parole est bien évidemment amenée à évoluer au cours du temps, on fera donc par la suite référence à la grandeur $RSB_{prio}(p,k)$ tout en sachant que ses caractéristiques évoluent lentement par rapport à celles du $RSB_{post}(p,k)$.**

À partir du RSB *a posteriori*, on peut également définir le RSB *instantané* qui correspond à une estimée locale (ou à court terme) du RSB *a priori* par soustraction directe de la DSP du bruit au module carré du signal bruité [Cappé 1994, Plapous 2004] :

$$RSB_{inst}(p,k) = \frac{|X(p,k)|^2 - \gamma_b(k)}{\gamma_b(k)} = \frac{|X(p,k)|^2 - \mathrm{E}[|B(p,k)|^2]}{\mathrm{E}[|B(p,k)|^2]} = RSB_{post}(p,k) - 1. \tag{2.7}$$

Ces trois expressions de RSB restent théoriques dans la mesure où seule la quantité $|X(p,k)|^2$ est connue. D'une part, la quantité $\mathrm{E}[|B(p,k)|^2]$ doit être estimée à partir du signal bruité. Ceci est décrit dans la partie 2.5 qui est consacrée aux techniques d'estimation de la DSP du bruit. D'autre part, il faut aussi estimer la quantité $\mathrm{E}[|S(p,k)|^2]$. Cette estimation est assez problématique et donne lieu à diverses techniques d'estimation du RSB *a priori* qui seront exposées dans la section 3.2. **Un des enjeux majeurs des systèmes de réduction de bruit par atténuation spectrale à court terme est d'obtenir une bonne estimation du RSB *a priori* ou du RSB *a posteriori*.**

2.3 Mise en œuvre de l'atténuation spectrale à court terme

D'un point de vue très général, les méthodes de réduction de bruit par atténuation spectrale à court terme sont mises en œuvre de la façon suivante :

- Le signal temporel bruité, $x(t)$, est découpé en trames pouvant se chevaucher puis chaque trame est fenêtrée. Nous verrons qu'un recouvrement est en fait nécessaire, mais pas suffisant, pour éviter certains artefacts liés au traitement par blocs. Chaque trame d'analyse est alors transformée dans un domaine où l'étape de réduction de bruit est réalisable. Généralement une TFD est utilisée ce qui implique que le signal est exploité dans le domaine fréquentiel. Cette première étape constitue la TFCT.
- Une estimation de la densité spectrale de puissance du bruit à long terme est réalisée soit en période de bruit seul ce qui nécessite alors une detection d'activité vocale (DAV), soit de façon continue même pendant l'activité vocale.
- Une atténuation spectrale à court terme ou autrement dit un gain spectral évalué à partir d'une règle de suppression des composantes spectrales de bruit est appliqué au module du signal bruité.

Le calcul du gain spectral requiert l'estimation du RSB (*a posteriori* et/ou *a priori*). Ces quantités sont estimées à partir du module carré du signal bruité $|X(p,k)|^2$ et de la DSP du bruit $\mathrm{E}[|B(p,k)|^2]$.

– Le module du signal de parole ainsi estimé $|\hat{S}(p,k)|$ et la phase du signal bruité sont alors utilisés pour revenir dans le domaine temporel en utilisant une TFD inverse (TFDI). Le signal de sortie est finalement synthétisé à partir d'une technique de type OLS (pour overlap and save) ou OLA (pour overlap and add) qui seront détaillées dans la partie 5.1. Cette dernière étape constitue la TFCT inverse (TFCTI).

Le schéma de principe de la figure 2.1 résume le fonctionnement général de la réduction de bruit par atténuation spectrale à court terme. Comme le montre ce schéma, seul le module de la TFD subit un

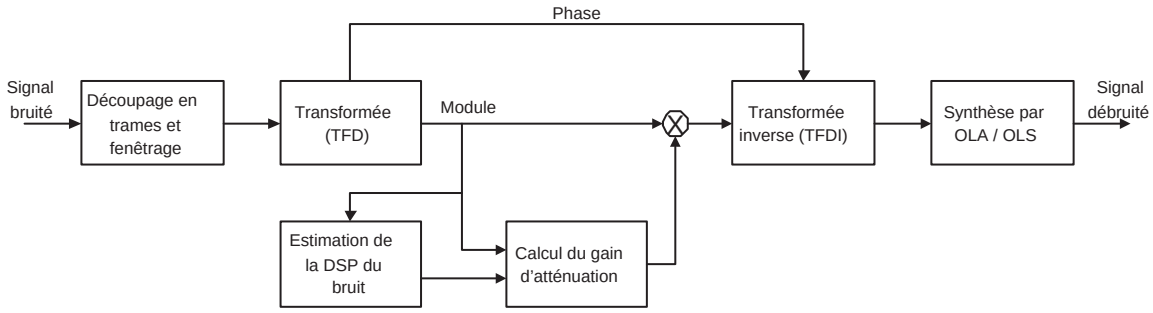

FIG. 2.1 – *Schéma de principe des techniques de réduction de bruit par atténuation spectrale à court terme.*

traitement, la phase étant réutilisée sans modification. Ceci est principalement dû au fait que, d'un point de vue pratique, la phase du signal original est très difficile à estimer et que l'oreille humaine est peu sensible aux distorsions de phase [Wang 1982].

2.4 Principales méthodes d'atténuation spectrale à court terme

Les techniques de réduction de bruit par atténuation spectrale à court terme peuvent être classées de différentes manières. Dans cet état de l'art, nous allons différencier les approches ne nécessitant pas de modèle statistique pour les signaux traités de celles qui en requièrent. Une partie supplémentaire sera consacrée aux techniques basées sur une approche psychoacoustique.

2.4.1 Approches ne nécessitant pas de modèle statistique

2.4.1.1 Soustraction spectrale

Le principe de la soustraction spectrale a été introduit dans [Boll 1979]. L'utilisation de ce type d'algorithme est très répandue bien que sa justification soit purement intuitive. En effet, dans la soustraction spectrale d'amplitude (SSA) le module du spectre à court terme est estimé comme ceci :

$$|\hat{S}(p,k)| = |X(p,k)| - \sqrt{\mathrm{E}[|B(p,k)|^2]}. \tag{2.8}$$

On ne peut pas garantir que la valeur de $|\hat{S}(p,k)|$ soit toujours positive car on soustrait une valeur moyennée $\sqrt{\mathrm{E}[|B(p,k)|^2]}$ au module d'un spectre instantané $|X(p,k)|$ dont la variance est beaucoup plus importante. Toutefois, le module du spectre du signal estimé doit rester positif ou nul, une valeur négative n'ayant pas de signification physique, cette contrainte est satisfaite par un simple seuillage :

$$|\hat{S}(p,k)| = \begin{cases} |X(p,k)| - \sqrt{\mathrm{E}[|B(p,k)|^2]} & \text{si } |X(p,k)| \geq \sqrt{\mathrm{E}[|B(p,k)|^2]}, \\ 0 & \text{sinon.} \end{cases} \tag{2.9}$$

L'équation ci-dessus nous donne directement accès à l'estimée du module du spectre du signal utile. La phase du signal bruité est recombinée avec ce module pour donner l'estimation du spectre du signal utile :

$$\hat{S}(p,k) = |\hat{S}(p,k)| e^{i\phi_X(p,k)}. \tag{2.10}$$

Toutefois dans un souci d'homogénéisation des écritures avec les autres techniques présentées, la SSA peut aussi être écrite sous forme d'un gain dépendant du RSB *a posteriori* (2.5) défini dans la partie 2.2. Ainsi, le spectre estimé du signal utile peut être obtenu par l'équation suivante :

$$\hat{S}(p,k) = G_{SSA}(p,k) X(p,k) \tag{2.11}$$

avec

$$G_{SSA}(p,k) = \begin{cases} 1 - \sqrt{\frac{1}{RSB_{post}(p,k)}} & \text{si } RSB_{post}(p,k) \geq 1, \\ 0 & \text{sinon.} \end{cases} \tag{2.12}$$

La figure 2.2 représente le gain de la SSA en fonction du RSB *a posteriori*.

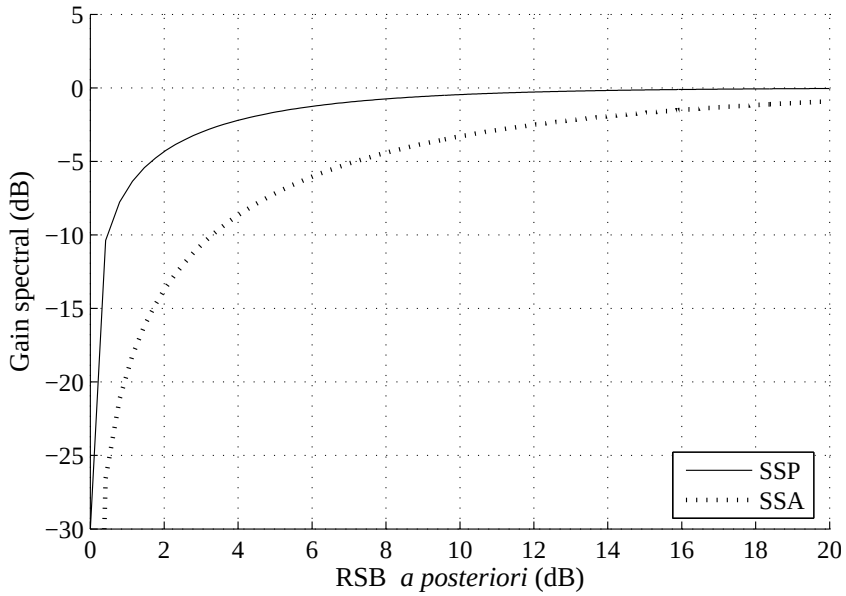

FIG. 2.2 – *Gain spectral de la SSP (trait plein) et de la SSA (pointillé) en fonction du RSB* a posteriori.

D'autres règles basées sur le principe de la soustraction spectrale ont été proposées, l'une des plus connues étant la soustraction spectrale en puissance (SSP) [Lim 1979, Vary 1985] :

$$|\hat{S}(p,k)|^2 = |X(p,k)|^2 - \mathrm{E}[|B(p,k)|^2]. \tag{2.13}$$

De la même façon que pour la SSA, les valeurs négatives du module carré $|\hat{S}(p,k)|^2$ sont à proscrire ce qui amène à contraindre le résultat :

$$|\hat{S}(p,k)|^2 = \begin{cases} |X(p,k)|^2 - \mathrm{E}[|B(p,k)|^2] & \text{si } |X(p,k)|^2 \geq \mathrm{E}[|B(p,k)|^2], \\ 0 & \text{sinon.} \end{cases} \tag{2.14}$$

On peut anticiper sur la partie suivante en remarquant qu'en utilisant les hypothèses de l'approche Ephraïm et Malah (les composantes spectrales sont indépendantes et le bruit et la parole ont une distribution gaussienne, *cf.* partie 2.4.2.2) alors la SSP correspond à l'estimateur au sens du maximum de vraisemblance (MV) de la DSP du signal utile [McAulay 1980] assurant ainsi une estimation non biaisée du module carré du signal utile [Cappé 1993]. La SSP peut aussi s'exprimer en fonction du RSB *a posteriori* :

$$G_{SSP}(p,k) = \begin{cases} \sqrt{1 - \frac{1}{RSB_{post}(p,k)}} & \text{si } RSB_{post}(p,k) \geq 1, \\ 0 & \text{sinon.} \end{cases} \tag{2.15}$$

L'avantage de la SSP (et de la soustraction spectrale en général) est sans aucun doute la simplicité de sa mise en œuvre, cependant cette approche possède un inconvénient majeur car elle engendre un bruit résiduel dit "bruit musical" qui se révèle très gênant à l'écoute. Ce bruit musical est dû au fait que le spectre moyen d'énergie du bruit $\mathrm{E}[|B(p,k)|^2]$ est soustrait à un spectre d'énergie instantané $|X(p,k)|^2$ possédant une forte variance par rapport à cette moyenne (en période de bruit seul). **Le spectre d'énergie résiduel possède donc des pics fréquentiels localisés de façon aléatoire en fréquence et qui ne dépassent pas statistiquement la durée d'une trame. Bien que la puissance du bruit résiduel soit inférieure à celle du bruit original, sa structure tonale fait qu'il est souvent plus gênant que le bruit original**, ce qui est inadmissible. Cette explication est transposable au gain spectral, c'est alors bien entendu la forte variance du RSB *a posteriori* qui génère le bruit musical. La figure 2.2 représente également le gain de la SSP en fonction du RSB *a posteriori*. On a vu que la SSP assure une estimation non biaisée du module carré de la parole, or le gain de la SSA est toujours inférieur à celui de la SSP, la parole va donc subir des distorsions (sous-estimation) par contre le niveau du bruit musical sera sensiblement diminué (il reste malgré tout très audible). Des solutions ont été proposées pour limiter le bruit musical résiduel, elles sont englobées sous l'appellation de soustraction spectrale généralisée.

2.4.1.2 Soustraction spectrale généralisée

Certaines améliorations ont été apportées à la soustraction spectrale avec principalement pour but de supprimer le bruit musical généré par cette approche. Leur regroupement conduit à la technique de soustraction spectrale généralisée (SSG) [Berouti 1979] qui s'exprime de la manière suivante :

$$|\hat{S}(p,k)|^2 = \begin{cases} D(p,k)^{\frac{1}{\gamma}} & \text{si } D(p,k)^{\frac{1}{\gamma}} \geq \beta \mathrm{E}[|B(p,k)|^2], \\ \beta \mathrm{E}[|B(p,k)|^2] & \text{sinon} \end{cases} \tag{2.16}$$

avec

$$D(p,k) = (|X(p,k)|^2)^{\gamma} - \alpha(\mathrm{E}[|B(p,k)|^2])^{\gamma} \tag{2.17}$$

où $\alpha \geq 1$, $0 \leq \beta \leq 1$ et $\gamma \geq 0$. Ces trois paramètres ont chacun un rôle bien précis :

- α est le coefficient de surestimation, il permet de surestimer la DSP du bruit et ainsi de réduire le niveau de bruit musical résiduel. Cependant, cette surestimation introduit une distorsion du signal de parole restauré. Pour limiter cet effet, le coefficient de surestimation peut être rendu adaptatif en fonction de la fréquence et suivant le type de bruit [Le Bouquin 1991].
- β correspond au plancher spectral, ce plancher permet de conserver une certaine quantité du bruit original ce qui va permettre de masquer le bruit musical. Cependant, pour supprimer totalement le bruit musical il faut utiliser une valeur de β suffisamment grande ce qui a pour effet de réintroduire une grande partie du bruit.
- γ détermine le type de soustraction spectrale ou autrement dit "l'agressivité" du gain spectral.

De la même façon que pour la SSA et la SSP, la SSG peut s'exprimer sous forme de gain :

$$G_{SSG}(p,k) = \begin{cases} \left(1 - \alpha \frac{1}{RSB_{post}^{\gamma}(p,k)}\right)^{\frac{1}{2\gamma}} & \text{si } \left(RSB_{post}^{\gamma}(p,k) - \alpha\right)^{\frac{1}{\gamma}} \geq \beta, \\ \sqrt{\frac{\beta}{RSB_{post}(p,k)}} & \text{sinon.} \end{cases} \tag{2.18}$$

Notons que si $\alpha = 1$ et $\beta = 0$ alors la SSG se résume à la SSA lorsque $\gamma = \frac{1}{2}$ et à la SSP dans le cas où $\gamma = 1$.

La surestimation du niveau de bruit est couramment utilisée. Son impact est illustré par la figure 2.3 en se basant sur l'approche SSP ($\beta = 0$ et $\gamma = 1$) et en faisant varier le paramètre α de surestimation : $\alpha = 1,2,3,4,5$. Quand $\alpha = 1$, il n'y a pas de surestimation et il s'agit donc de la SSP

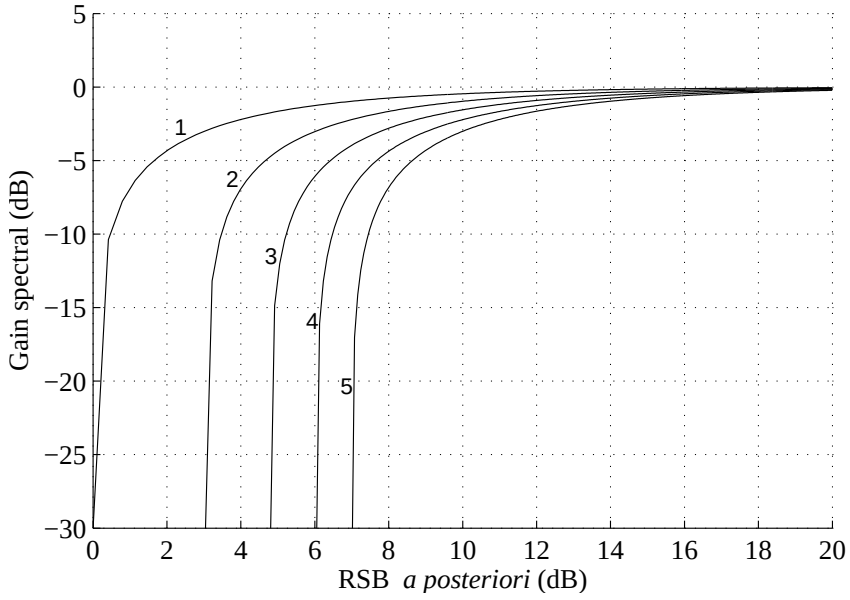

FIG. 2.3 – *Gain spectral de la SSG en fonction du RSB* a posteriori. $\beta = 0$, $\gamma = 1$ *et* α *prend les valeurs 1, 2, 3, 4, 5.*

classique qui sert de référence pour comparer les quatre autres courbes. Il apparaît clairement sur cette figure que surestimer le bruit revient à décaler la courbe de gain vers les forts RSB *a posteriori*.

Ainsi, plus on surestime le bruit (α important) et plus le niveau de bruit musical résiduel sera faible, par contre la distorsion de la parole augmentera aussi rapidement. Il s'agit donc d'ajuster au mieux le compromis entre ces deux tendances.

2.4.1.3 Filtrage de Wiener

Le filtre de Wiener [Scalart 1996a] est le filtre linéaire optimal au sens de l'erreur quadratique moyenne minimum (EQMM), c'est-à-dire qu'il minimise la fonction d'erreur suivante :

$$\mathrm{E}[(S(p,k) - G_W(p,k)X(p,k))^2]. \tag{2.19}$$

Avec les hypothèses classiques posées dans la partie 2.1, le filtre de Wiener peut s'exprimer ainsi :

$$G_W(p,k) = \frac{\mathrm{E}[|S(p,k)|^2]}{\mathrm{E}[|X(p,k)|^2]} = \frac{\mathrm{E}[|S(p,k)|^2]}{\mathrm{E}[|S(p,k)|^2] + \mathrm{E}[|B(p,k)|^2]}. \tag{2.20}$$

Les quantités intervenant dans l'expression du filtre de Wiener étant calculées à long terme, on peut l'exprimer en fonction du RSB *a priori* :

$$G_W(p,k) = \frac{RSB_{prio}(p,k)}{1 + RSB_{prio}(p,k)}. \tag{2.21}$$

En pratique, il est possible de remplacer la DSP du signal utile $\mathrm{E}[|S(p,k)|^2]$ par une estimée obtenue directement à partir du spectre bruité instantané selon le principe de la SSP : $|X(p,k)|^2 - \mathrm{E}[|B(p,k)|^2]$, bien que cela n'ait pas de fondement théorique. On s'éloigne alors des hypothèses de stationnarité nécessaires au filtre de Wiener. C'est pourquoi on parle de filtre pseudo-Wiener selon l'appellation proposée dans [Cappé 1993]. Il est alors possible d'exprimer le filtre pseudo-Wiener en fonction du RSB *a posteriori* :

$$G_W(p,k) = 1 - \frac{1}{RSB_{post}(p,k)}. \tag{2.22}$$

On peut d'ailleurs remarquer que dans ce cas $G_W(p,k) = G^2_{SSP}(p,k)$ attestant du fait que ce filtre fait partie de la famille des méthodes de soustraction spectrale.

La figure 2.4 permet de situer le filtre pseudo-Wiener par rapport à la SSP et la SSA. Les gains correspondant à ces trois approches y sont représentés en fonction du RSB *a posteriori*. Il apparaît que le filtre pseudo-Wiener réalise un compromis entre la SSP et la SSA limitant ainsi la distorsion de la parole par rapport à la SSA mais ne permettant pas de supprimer beaucoup plus de bruit musical que la SSP.

On a vu que le véritable filtre de Wiener s'écrit en fonction du RSB *a priori*. La figure 2.5 représente son gain en fonction du RSB *a priori*. Cette quantité varie théoriquement plus lentement que le RSB *a posteriori*, le signal restauré ne souffre donc pas du bruit musical, sa cause (la variance importante du RSB) ayant disparu. Par contre les non-stationnarités du signal de parole ne peuvent pas être prises en compte ce qui crée un autre type de distorsion : **par analogie avec une image, le signal restauré semble "flou" et perd de sa dynamique. En pratique, le calcul du RSB *a priori* est très souvent basé sur l'approche "decision-directed", présentée dans la partie 3.2.5, et l'effet**

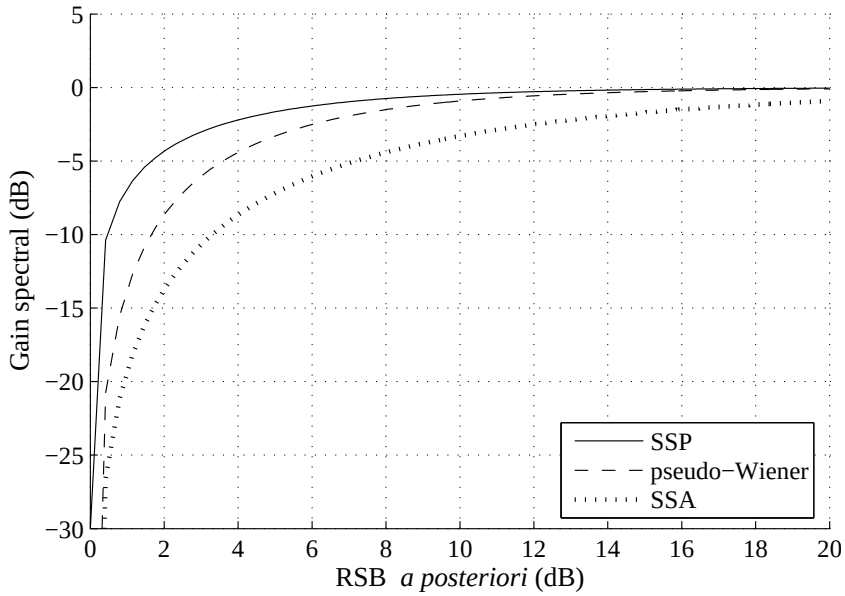

FIG. 2.4 – *Gain spectral du filtre pseudo-Wiener (tirets), de la SSP (trait plein) et de la SSA (pointillé) en fonction du RSB* a posteriori.

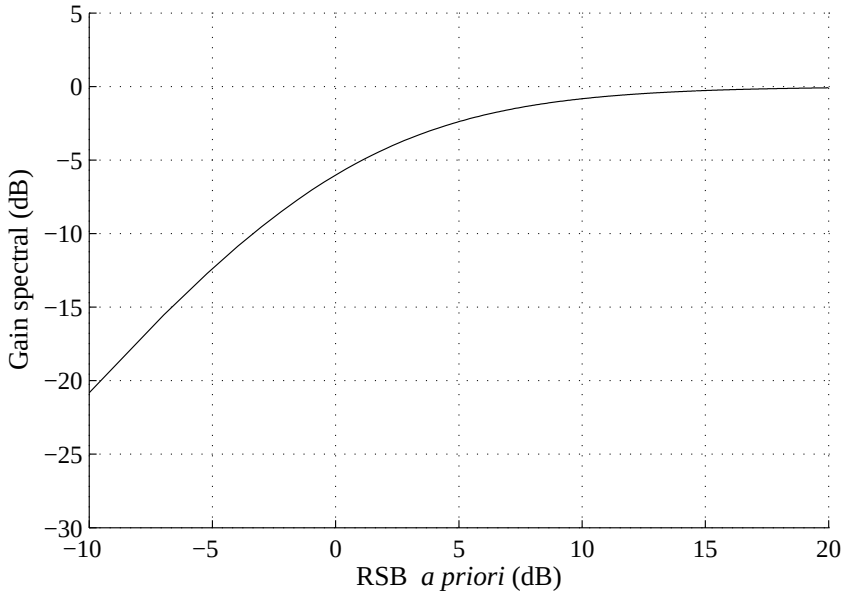

FIG. 2.5 – *Gain spectral du filtre de Wiener en fonction du RSB* a priori.

de flou s'interprète alors comme de la réverbération [Cappé 1993]. Une alternative (peu usitée, car moins performante) consiste à soustraire la DSP du bruit au périodogramme lissé du signal bruité afin d'obtenir une estimée de la DSP du signal de parole [Guérin 2002].

2.4.1.4 Formulation paramétrique de la soustraction spectrale généralisée

Cette approche a été proposée dans [Sim 1998] et a pour but d'améliorer les résultats de la SSG tout en conservant une faible complexité. Le point de départ de cette méthode est de considérer une formulation paramétrique de la SSG (SSGP) :

$$|\hat{S}(p,k)|^{\alpha} = a_{\alpha}(p,k)|X(p,k)|^{\alpha} - b_{\alpha}(p,k)\mathrm{E}[|B(p,k)|^{\alpha}]. \tag{2.23}$$

L'estimateur optimal du module de la parole est ensuite obtenu au sens de l'EQMM en minimisant :

$$\mathrm{E}[(|S(p,k)|^{\alpha} - |\hat{S}(p,k)|^{\alpha})^2]. \tag{2.24}$$

On peut obtenir deux estimateurs selon que les paramètres optimaux $a_{\alpha}(p,k)$ et $b_{\alpha}(p,k)$ sont obtenus avec ou sans contrainte. Sans contrainte on obtient l'estimateur optimal suivant :

$$G_{SSGP}(p,k) = \left[\left(\frac{RSB^{\alpha}_{prio}(p,k)}{RSB^{\alpha}_{prio}(p,k)+1}\right)\right. \tag{2.25}$$

$$\left.\left(1-\left(1-RSB^{-\frac{\alpha}{2}}_{prio}(p,k)\right)\Gamma\left(\frac{\alpha}{2}+1\right)\left(\frac{1}{RSB_{post}(p,k)}\right)^{\frac{\alpha}{2}}\right)\right]^{\frac{1}{\alpha}} \tag{2.26}$$

où $\Gamma(.)$ représente la fonction Gamma. Avec la contrainte $a_{\alpha}(p,k) = b_{\alpha}(p,k)$, le gain optimal devient :

$$G_{SSGP}(p,k) = \left[\left(\frac{RSB^{\alpha}_{prio}(p,k)}{RSB^{\alpha}_{prio}(p,k)+\beta_{\alpha}}\right)\left(1-\Gamma\left(\frac{\alpha}{2}+1\right)\left(\frac{1}{RSB_{post}(p,k)}\right)^{\frac{\alpha}{2}}\right)\right]^{\frac{1}{\alpha}} \tag{2.27}$$

où β_{α} est une constante pour une valeur fixée de α :

$$\beta_{\alpha} = \frac{\Gamma(\alpha+1) - \Gamma^2\left(\frac{\alpha}{2}+1\right)}{\Gamma(\alpha+1)}. \tag{2.28}$$

La figure 2.6 représente le gain spectral de l'approche SSGP avec contrainte en fonction du RSB *a priori* et paramétré par le RSB *a posteriori* : $RSB_{post} = 0, 5, 15$dB. On peut tout d'abord remarquer que le paramètre α et le RSB *a priori* imposent l'allure générale du gain ajustant ainsi le compromis entre dégradation de la parole et réduction du bruit. Chaque jeu de courbes en comporte en effet 3 qui sont peu ou prou des versions décalées d'une seule et même courbe. C'est le RSB *a posteriori* qui permet de régler l'atténuation globale du gain, *i.e.* l'offset qui fixe la position de la courbe. En effet, plus le RSB *a posteriori* est faible et plus le gain se trouve décalé vers les atténuations importantes. Ce comportement permet de supprimer efficacement le bruit musical résiduel et supprime aussi l'effet de réverbération obtenu avec le filtre de Wiener. Ceci est rendu possible grâce au rôle du RSB *a posteriori*. Cependant, en contrepartie de l'action bénéfique du RSB *a posteriori* on note des atténuations importantes du gain lorsque le RSB *a priori* est important. Le gain de l'approche SSGP est donc souvent largement inférieur à celui du filtre de Wiener ce qui engendre une distorsion du signal de parole.

L'approche SSGP sans contrainte ne sera pas détaillée car elle introduit plus de bruit musical et de distorsion que l'approche avec contrainte.

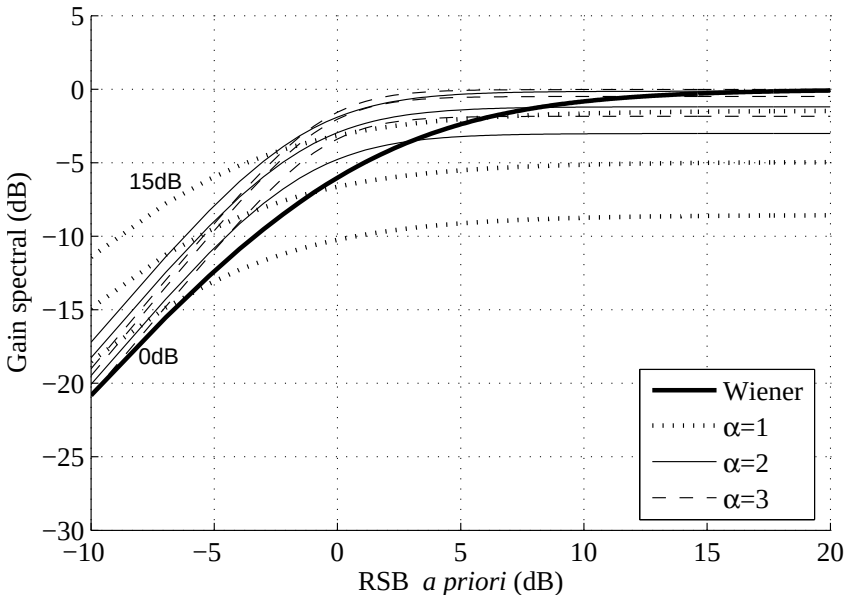

FIG. 2.6 – *Gain spectral de l'approche SSGP avec contrainte en fonction du RSB* a priori *et paramétré par le RSB* a posteriori *: $RSB_{post} = 0, 5, 15dB$. 3 jeux de courbes sont tracés selon la valeur de* α*, en pointillé pour* $\alpha = 1$*, en trait plein pour* $\alpha = 2$ *et avec des tirets pour* $\alpha = 3$*. La courbe en trait fort correspond au filtre de Wiener.*

2.4.2 Approches nécessitant des modèles statistiques

2.4.2.1 Méthode de MacAulay et Malpass

Estimateur du maximum de vraisemblance

En vue de déterminer l'estimateur au sens du maximum de vraisemblance, MacAulay et Malpass posent les deux hypothèses suivantes [McAulay 1980] :

- Le bruit est un processus gaussien indépendant du signal de parole.
- La parole est caractérisée par un signal déterministe d'amplitude et de phase inconnues.

L'expression de l'équation (2.3) reste donc la même à ceci près que la TFCT du signal de parole devient :

$$S(p,k) = Ae^{i\theta} \tag{2.29}$$

où A représente l'enveloppe du signal et θ sa phase. Le bruit étant supposé gaussien, et le signal de parole déterministe, la densité de probabilité du signal bruité s'exprime ainsi :

$$\mathrm{p}(X(p,k) \mid A,\theta) = \frac{1}{\pi \mathrm{E}[|B(p,k)|^2]} \cdot exp\left[-\frac{|X(p,k)|^2 - 2A\Re[e^{-i\theta}X(p,k)] + A^2}{\mathrm{E}[|B(p,k)|^2]}\right] \tag{2.30}$$

où $\Re$ est l'opérateur permettant d'obtenir la partie réelle d'une grandeur complexe et $\mathrm{p}(X(p,k) \mid A,\theta)$ s'interprète comme la densité de probabilité de $X(p,k)$ sachant A et θ. Pour obtenir l'estimateur de l'enveloppe A au sens du maximum de vraisemblance, il suffit de maximiser $p(X(p,k) \mid A,\theta)$. Le

terme de phase θ est un paramètre gênant qu'il est possible d'éliminer en choisissant de maximiser la fonction de vraisemblance moyennée suivante :

$$\overline{\mathrm{p}(X(p,k) \mid A)} = \int_0^{2\pi} \mathrm{p}(X(p,k) \mid A,\theta)\mathrm{p}(\theta)\mathrm{d}\theta, \tag{2.31}$$

où $p(\theta)$ est la densité de probabilité de la phase qui est supposée uniformément distribuée sur $[0,2\pi[$. Ainsi l'estimateur au sens du maximum de vraisemblance de l'enveloppe du signal utile de parole est :

$$\hat{A} = \frac{1}{2}\left[|X(p,k)| + \sqrt{|X(p,k)|^2 - \mathrm{E}[|B(p,k)|^2]}\right]. \tag{2.32}$$

En utilisant la phase du signal de parole bruitée, on obtient alors l'estimée suivante pour le spectre du signal de parole :

$$\hat{S}(p,k) = \hat{A}\frac{X(p,k)}{|X(p,k)|} = \frac{1}{2}\left[1 + \sqrt{\frac{|X(p,k)|^2 - \mathrm{E}[|B(p,k)|^2]}{|X(p,k)|^2}}\right] X(p,k) \tag{2.33}$$

On peut également exprimer le gain de cet estimateur en fonction du RSB *a posteriori* :

$$G_{MV}(p,k) = \frac{1}{2}\left[1 + \sqrt{1 - \frac{1}{RSB_{post}(p,k)}}\right]. \tag{2.34}$$

Introduction du filtre à deux-états

Les règles d'atténuation spectrale que nous avons vues jusqu'ici ne considéraient pas l'intermittence du signal de parole (sinon dans la façon d'estimer le bruit). MacAulay et Malpass ont donc proposé de prendre en compte cet aspect "deux-états" de la parole par le biais d'une approche appelée "soft-decision (SD) filter" [McAulay 1980]. Ce filtre est basé sur les hypothèses suivantes :

- H_0 : absence de parole.
- H_1 : présence de parole.

L'estimateur au sens de l'EQMM du module du spectre de parole peut alors se mettre sous la forme suivante :

$$\begin{aligned}|\hat{S}(p,k)| = {} & \mathrm{E}\left[|S(p,k)| \mid |X(p,k)|,\mathrm{H}_0\right]\mathrm{p}(\mathrm{H}_0 \mid |X(p,k)|) \\ & + \mathrm{E}[|S(p,k)| \mid |X(p,k)|,\mathrm{H}_1]\mathrm{p}(\mathrm{H}_1 \mid |X(p,k)|).\end{aligned} \tag{2.35}$$

Sous l'hypothèse H_0 la parole est absente et par conséquent l'espérance du module du signal utile est nulle, l'estimateur se simplifie donc de la manière suivante :

$$|\hat{S}(p,k)| = \mathrm{E}[|S(p,k)| \mid |X(p,k)|,\mathrm{H}_1]\mathrm{p}(\mathrm{H}_1 \mid |X(p,k)|). \tag{2.36}$$

L'espérance $\mathrm{E}[|S(p,k)| \mid |X(p,k)|,\mathrm{H}_1]$ est alors remplacée par l'estimateur du maximum de vraisemblance de l'équation (2.33) ce qui conduit à :

$$G_{MV}^{SD}(p,k) = \frac{1}{2}\left[1 + \sqrt{1 - \frac{1}{RSB_{post}(p,k)}}\right]\mathrm{p}(\mathrm{H}_1 \mid |X(p,k)|). \tag{2.37}$$

La probabilité *a posteriori* de présence de la parole $\mathrm{p}(\mathrm{H}_1 \mid |X(p,k)|)$ s'exprime ainsi :

$$\mathrm{p}(\mathrm{H}_1 \mid |X(p,k)|) = \frac{\Lambda(p,k)}{1+\Lambda(p,k)} \tag{2.38}$$

avec

$$\Lambda(p,k) = e^{-RSB_{prio}} I_0 \left[2\sqrt{RSB_{prio} RSB_{post}(p,k)} \right] \tag{2.39}$$

où I_0 est la fonction de Bessel modifiée du premier genre et d'ordre zéro. La quantité $\mathrm{p}(\mathrm{H}_1 \mid |X(p,k)|)$ est obtenue en supposant que, *a priori*, les présences de parole et de bruit sont équiprobables ce qui n'est pas une hypothèse très réaliste. Le RSB *a priori*, RSB_{prio}, qui intervient dans l'équation (2.39) est une constante et est perçu comme un facteur de suppression qui permet d'ajuster l'atténuation du filtre et donc d'obtenir un compromis entre réduction du bruit et distorsion de la parole. Typiquement, si ce facteur croît, le bruit de fond sera de plus en plus atténué, par contre la distorsion du signal de parole augmentera. En fait, le RSB *a posteriori* permet de prendre en compte le comportement dynamique de la TFCT du signal d'observation alors que les informations statistiques des signaux sont prises en compte par le biais du RSB *a priori*.

Pour illustrer l'apport de l'approche SD, la figure 2.7 représente le gain spectral de l'approche MV classique (2.34) et associée à l'approche SD (2.37). On voit que l'on peut obtenir un jeu de courbes selon la valeur choisie pour le $RSB_{prio} = 2,\ 5,\ 10,\ 15,\ 30$ (en valeur naturelle). Il apparaît sur la

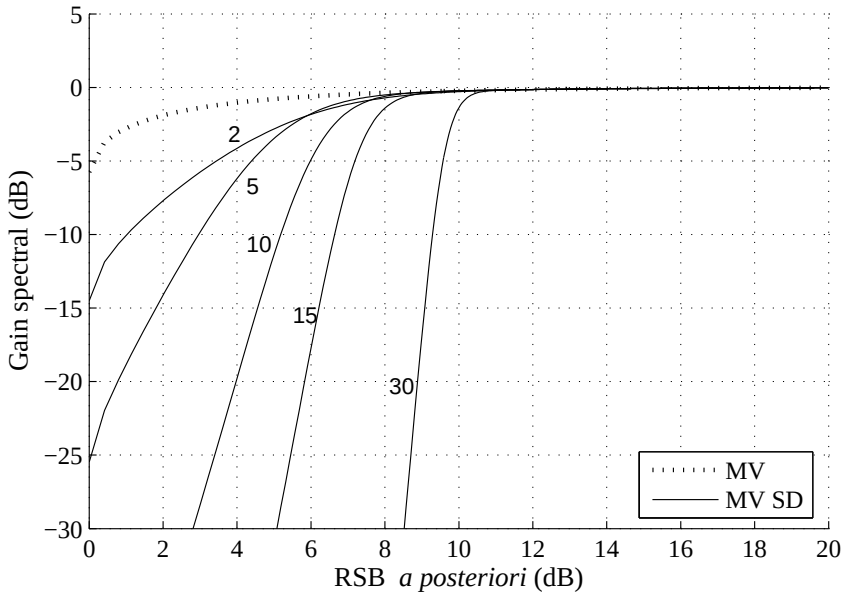

FIG. 2.7 – *Gain spectral de l'approche MV seule (pointillé) et associée à l'approche SD en fonction du RSB* a posteriori *et paramétré par le* $RSB_{prio} = 2,\ 5,\ 10,\ 15,\ 30$ *(trait plein).*

figure 2.7 que l'augmentation de la valeur du RSB_{prio} fait tendre le gain spectral vers un comportement de tout ou rien. Ainsi, par exemple, si le $RSB_{prio} = 30$ alors le gain spectral conserve les composantes de parole correspondant à un RSB *a posteriori* supérieur à environ 10dB et supprime les autres. Le bruit musical est donc bien supprimé mais au détriment de certaines composantes de parole. Ce

comportement correspond donc peu ou prou au gain suivant :

$$G(p,k) = \begin{cases} \sqrt{1 - \frac{1}{RSB_{post}(p,k)}} & \text{si } RSB_{post}(p,k) \geq \alpha, \\ 0 & \text{sinon} \end{cases} \tag{2.40}$$

où α correspond à un seuil de 10dB environ si $RSB_{prio} = 30$. Cette approche ressemble donc à la SSG (avec $\beta = 0$, $\gamma = 1$ et une surestimation du bruit par le paramètre α, *cf.* partie 2.4.1.2), cependant, au lieu d'utiliser une valeur biaisée du RSB *a posteriori* $\left(\alpha \frac{1}{RSB_{post}(p,k)}\right)$ dans l'équation (2.18) on calcule le gain spectral à partir du RSB *a posteriori* classique ce qui permet de limiter les dégradations de la parole comparé à la SSG tout en apportant la même réduction du niveau de bruit musical.

2.4.2.2 Méthodes d'Ephraïm et Malah

Estimateur MMSE STSA

Cette approche réalise une estimation optimale de l'amplitude spectrale à court terme du signal de parole au sens de l'EQMM ce qui lui a d'ailleurs donné son nom : Minimum Mean Square Error Short Time Spectral Amplitude ou de façon plus concise MMSE STSA. Ephraïm et Malah proposent d'approcher le spectre à court terme du signal de parole par des composantes gaussiennes aléatoires et indépendantes les unes des autres (au sein d'une même trame ainsi que d'une trame à l'autre) [Ephraïm 1984]. Ce modèle trouve sa justification dans le théorème central limite même si le signal temporel n'est pas gaussien [Godsill 1998]. Le bruit est ici aussi considéré gaussien et indépendant de la parole. L'estimateur du module du spectre de la parole est obtenu ainsi :

$$|\hat{S}(p,k)| = \mathrm{E}[|S(p,k)| \mid |X(p,k)|] = \mathrm{E}[A(p,k) \mid |X(p,k)|]. \tag{2.41}$$

Suivant le modèle proposé, on peut écrire :

$$\mathrm{E}[A(p,k) \mid |X(p,k)|] = \frac{\int_0^\infty \int_0^{2\pi} a_k \mathrm{p}(X(p,k) \mid a_k, \alpha_k) \mathrm{p}(a_k, \alpha_k) \mathrm{d}\alpha_k \mathrm{d}a_k}{\int_0^\infty \int_0^{2\pi} \mathrm{p}(X(p,k) \mid a_k, \alpha_k) \mathrm{p}(a_k, \alpha_k) \mathrm{d}\alpha_k \mathrm{d}a_k}. \tag{2.42}$$

ce qui conduit à l'expression du gain de l'estimateur MMSE STSA suivante :

$$G_{STSA}(p,k) = \frac{\sqrt{\pi}}{2} \sqrt{\left(\frac{1}{RSB_{post}(p,k)}\right)\left(\frac{RSB_{prio}(p,k)}{1 + RSB_{prio}(p,k)}\right)} M(\nu(p,k)), \tag{2.43}$$

avec

$$\nu(p,k) = \frac{SNR_{prio}(p,k)}{1 + SNR_{prio}(p,k)} SNR_{post}(p,k) \tag{2.44}$$

et où $M(.)$ représente la fonction hypergéométrique confluente suivante :

$$M(x) = exp\left(-\frac{x}{2}\right)\left[(1+x) I_0\left(\frac{x}{2}\right) + x I_1\left(\frac{x}{2}\right)\right], \tag{2.45}$$

où I_0 et I_1 sont, respectivement, les fonctions de Bessel du premier genre d'ordre zéro et d'ordre un.

La figure 2.8 représente le gain spectral de l'approche MMSE STSA en fonction du RSB *a priori* et paramétré par le RSB *a posteriori*. Le filtre de Wiener et celui de la SSP (dans la mesure où on

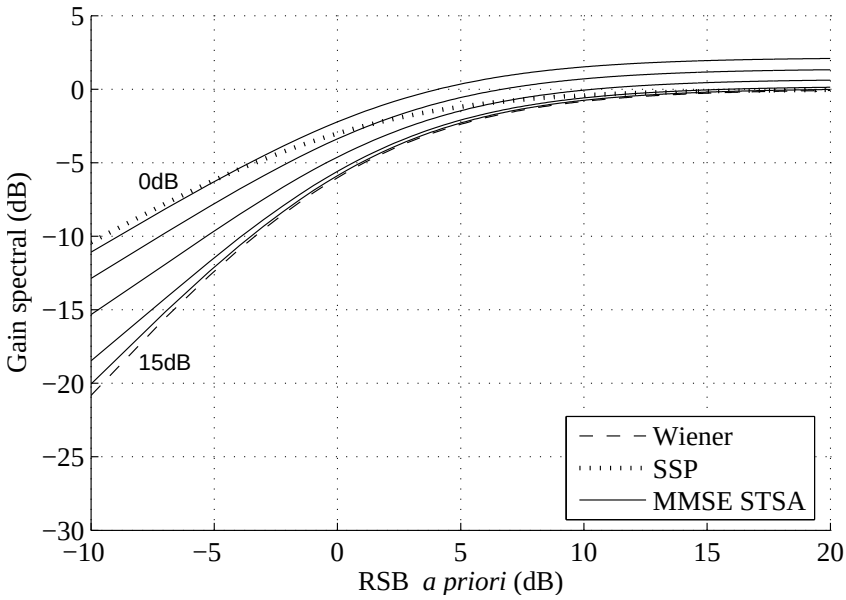

FIG. 2.8 – *Gain spectral de l'approche MMSE STSA en fonction du RSB* a priori *et paramétré par le RSB* a posteriori *: $RSB_{post} = 0, 2, 5, 10, 15$dB (trait plein). Les gains des filtres de Wiener (tirets) et de la SSP (pointillé) sont également représentés.*

s'autorise à l'exprimer en fonction du RSB *a priori*) sont également représentés. Avant tout, on peut remarquer que c'est le RSB *a priori* qui impose l'allure générale du gain de l'approche MMSE STSA. Le RSB *a posteriori* apparaît alors comme un paramètre permettant de régler l'agressivité du filtre. En effet, le RSB *a posteriori* permet d'assurer la transition entre le filtre de Wiener (RSB_{post} élevé) et la SSP (RSB_{post} faible). On peut également remarquer que si le RSB *a priori* est égal au RSB *a posteriori* alors l'approche MMSE STSA devient très proche de la SSP ce qui limite donc son intérêt mais permet de situer ses performances globales.

Dans cette règle de suppression, l'influence du RSB *a posteriori* va à l'encontre de ce que l'on observe dans les règles dites ponctuelles (celles qui s'expriment seulement en fonction du RSB *a posteriori*). **En effet, plus ce paramètre est élevé et plus l'atténuation apportée est importante. C'est cette propriété qui permet de limiter le bruit musical en faisant remonter le niveau de bruit résiduel autour des pics fréquentiels isolés.** Par contre le niveau de bruit résiduel (non musical donc) sera toujours plus élevé qu'en utilisant le filtre de Wiener. **Bien entendu, le fait que le RSB *a priori* soit un paramètre prépondérant et lissé contribue aussi fortement à la diminution du bruit musical. En contrepartie, du fait de ce lissage, le signal restauré souffre d'un effet de réverbération plus important encore que celui introduit par le filtre de Wiener (*cf.* partie 2.4.1.3).** On peut observer que quand le RSB *a posteriori* est faible et que le RSB *a priori* est élevé le gain devient supérieur à 0dB ce qui est difficile à interpréter. Cependant si l'on considère que le bruit s'est ajouté en opposition de phase avec le signal de parole et qu'il a détruit l'information utile ($RSB_{post} < RSB_{prio}$), il est alors possible d'y trouver une interprétation. Cependant, compte tenu du risque d'amplifier ce signal on peut choisir d'appliquer un seuil au gain.

Il est possible d'associer à la technique MMSE STSA l'approche "deux-états" ou SD proposée

dans [McAulay 1980]. Pour ce faire le terme $\Lambda(p,k)$ qui apparaît dans l'équation (2.38) est obtenu en utilisant le rapport de vraisemblance généralisé :

$$\Lambda(p,k) = \frac{1-q_k}{q_k} \frac{exp(\nu(p,k))}{1+RSB_{prio}(p,k)}, \tag{2.46}$$

où q_k représente la probabilité d'absence du signal de parole utile dans la bande de fréquence k. Finalement le gain de l'approche MMSE STSA SD s'écrit ainsi :

$$G_{STSA}^{SD}(p,k) = G_{STSA}(p,k) \frac{\Lambda(p,k)}{1+\Lambda(p,k)}. \tag{2.47}$$

Le RSB *a priori* utilisé dans cette variante SD peut être modifié pour prendre en compte l'aspect "deux-états" :

$$RSB_{prio}^{SD} = \frac{RSB_{prio}}{1-q_k}. \tag{2.48}$$

La figure 2.9 représente le gain spectral de l'approche MMSE STSA SD. La probabilité d'absence du signal de parole dans chaque bande de fréquence q_k est choisie constante et égale à 0,2 (choix empirique [Ephraïm 1984]). En prenant comme référence l'approche MMSE STSA (sans SD), illustrée

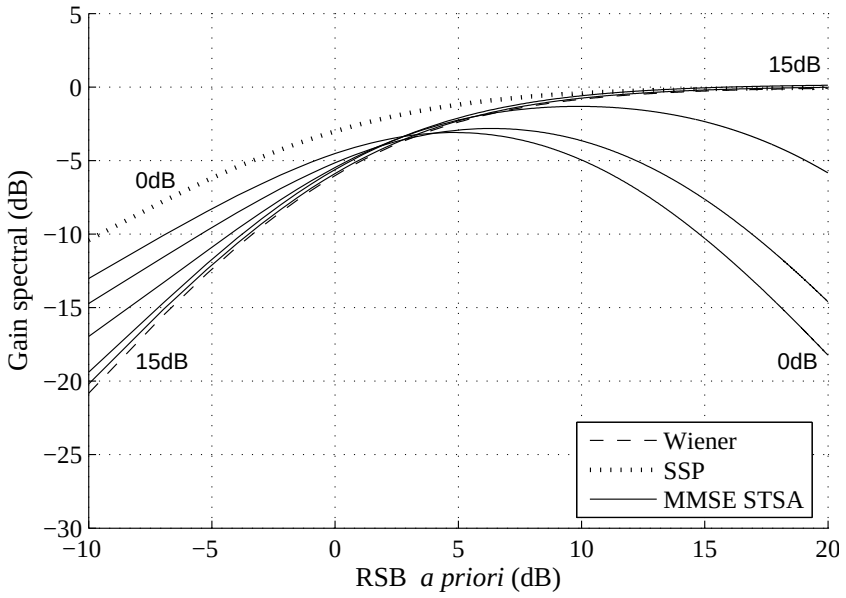

FIG. 2.9 – *Gain spectral de l'approche MMSE STSA SD ($q_k = 0{,}2$) en fonction du RSB* a priori *et paramétré par le RSB* a posteriori *: $RSB_{post} = 0, 2, 5, 10, 15$dB (trait plein). Les gains des filtres de Wiener (tirets) et de la SSP (pointillé) sont également représentés.*

par la figure 2.8, on peut remarquer que l'approche SD a une influence importante quand le RSB *a priori* est élevé et que le RSB *a posteriori* est faible. Ce désaccord entre les deux types de RSB est alors réglé par le terme $\frac{\Lambda(p,k)}{1+\Lambda(p,k)}$ dans l'équation (2.47), qui est fortement dépendant du RSB *a posteriori* et qui corrige donc la courbe de gain en l'atténuant fortement (très marqué dans la partie droite de la courbe et sensible aussi dans la partie gauche). Dans ce cas, le RSB *a posteriori* agit de façon "naturelle" en limitant le gain quand cette quantité est faible. Cela va donc à l'encontre de

ce que l'on a observé pour l'approche MMSE STSA où le RSB *a posteriori* agit paradoxalement à l'inverse en augmentant le gain, permettant ainsi de limiter le bruit musical. L'utilisation de l'approche SD limite donc cet effet contre nature ce qui permet de diminuer le niveau du bruit de fond résiduel (sans augmenter le niveau de bruit musical) ainsi que l'effet de réverbération lié au lissage de l'estimateur du RSB *a priori* lorsque $q_k = 0{,}2$. Cependant, l'efficacité de l'approche SD dépend fortement de cette valeur. Ainsi, lorsque q_k est proche de 0 l'amélioration n'est pas très importante, par contre si q_k est proche de 1 alors le niveau de bruit résiduel est très atténué et la réverbération est supprimée mais au prix d'une résurgence de certaines composantes de bruit musical. On peut noter que l'utilisation de l'approche SD supprime les occurrences où le gain MMSE STSA est supérieur à 0.

Estimateur MMSE LSA

Par la suite, Ephraïm et Malah ont proposé d'estimer le logarithme de l'amplitude du signal de parole plutôt que l'amplitude elle-même [Ephraïm 1985]. Cette technique est appelée Minimum Mean Square Error Log-Spectral Amplitude (MMSE LSA). Cette approche est justifiée par le fait que l'oreille est plus sensible aux variations du logarithme de l'amplitude qu'à celles de l'amplitude. L'estimateur optimal au sens de l'EQMM est obtenu de la façon suivante :

$$|\hat{S}(p,k)| = exp(\mathrm{E}[ln(|S(p,k)|) \mid X(p,k)]). \tag{2.49}$$

En utilisant les mêmes hypothèses que pour l'estimateur MMSE STSA, le gain correspondant à l'estimateur MMSE LSA peut se mettre sous la forme :

$$G_{LSA}(p,k) = \frac{SNR_{prio}(p,k)}{1+SNR_{prio}(p,k)} exp\left[\frac{1}{2}\int_{\nu(p,k)}^{\infty} \frac{exp(-t)}{t}\mathrm{d}t\right] \tag{2.50}$$

et $\nu(p,k)$ est défini par l'équation (2.44). Il est tout à fait possible d'utiliser également pour cet estimateur l'approche "deux-états", le gain devient alors :

$$G_{LSA}^{SD}(p,k) = G_{LSA}(p,k)\frac{\Lambda(p,k)}{1+\Lambda(p,k)}, \tag{2.51}$$

le terme $\Lambda(p,k)$ étant défini par l'équation (2.46). Cette approche MMSE LSA ne sera pas détaillée ici, son comportement étant relativement proche de celui de la technique MMSE STSA. Toutefois, on peut noter que certaines améliorations ont été apportées à l'approche MMSE LSA "deux-états". Ainsi, dans [Malah 1999] la probabilité d'absence du signal de parole, q_k, utilisée dans l'équation (2.46) est rendue adaptative afin de suivre les non-stationnarités du signal de parole. On peut également citer l'approche OM LSA [Cohen 2002b] où l'estimation du RSB est optimisée en fonction de l'approche "deux-états".

2.4.2.3 Modèles statistiques non gaussiens

Méthode de Porter et Boll

Porter et Boll ont proposé dans [Porter 1984] une méthode basée sur l'EQMM qui est originale par rapport à celles déjà présentées. En effet, plutôt que de poser des hypothèses assez éloignées de la réalité pour le modèle statistique de la parole (modèle gaussien classique), sa distribution est apprise à partir d'une base de données de parole. Le bruit est quant à lui toujours supposé gaussien et indépendant du signal de parole. À la manière de l'approche MMSE LSA [Ephraïm 1985], les auteurs proposent d'estimer une fonction $f(.)$ de l'amplitude spectrale à court terme. Ainsi, l'estimateur optimal au sens de l'EQMM de cette fonction est :

$$f(\hat{S}(p,k)) = \int f(s_k)\mathrm{p}(s_k \mid X(p,k))\mathrm{d}s_k. \tag{2.52}$$

Dans le cas où la densité de probabilité $\mathrm{p_b}$ du bruit est connue, la relation ci-dessus devient :

$$f(\hat{S}(p,k)) = \frac{\int f(s_k)\mathrm{p_b}(X(p,k)-s_k)\mathrm{p_s}(s_k)\mathrm{d}s_k}{\int \mathrm{p_b}(X(p,k)-s_k)\mathrm{p_s}(s_k)\mathrm{d}s_k}. \tag{2.53}$$

La densité de probabilité du signal de parole $\mathrm{p_s}$ peut être approchée à partir de l'histogramme obtenu sur une base de données. Cet estimateur peut alors être écrit en fonction des RSB *a posteriori* et *a priori* selon la fonction $f(.)$ choisie [Porter 1984].

Modèles statistiques super-gaussiens

L'approche précédente permet de modéliser fidèlement le signal de parole mais cela nécessite un apprentissage. Les approches proposées dans [Martin 2002, Guédon 2002] sont moins générales car les modèles statistiques utilisés restent théoriques mais ils sont tout de même plus proches de la réalité que le modèle gaussien classique. Ainsi, plusieurs estimateurs des parties réelle et imaginaire du signal de parole sont obtenus à partir de modèles laplaciens (pour le bruit) et Gamma (pour la parole) qui permettent d'obtenir des résultats sensiblement meilleurs que dans le cas gaussien. Ces approches ne sont pas développées ici car les parties 3.1, 4.1 et 6.3.2 leur sont consacrées avec une analyse et des résultats détaillés. Il existe bien sûr d'autres possibilités pour exploiter des modèles super-gaussiens. On peut citer par exemple l'approche proposée dans [Breithaupt 2003] qui permet d'estimer par une approche EQMM le module carré du signal de parole avec des résultats proches de ceux obtenus dans [Martin 2002]. On trouve également dans [Chen 2005] un estimateur MMSE STSA (*cf.* partie 2.4.2.2) où le signal de parole est modélisé par une loi de Laplace. Il est rapporté une amélioration par rapport à l'approche MMSE STSA classique.

2.4.3 Approches basées sur un modèle psychoacoustique

Il existe des techniques de réduction de bruit qui exploitent le principe de masquage psychoacoustique. Il est en effet possible de calculer le seuil de masquage d'un signal de parole qui indique le niveau de bruit qu'il est capable de masquer pour chaque bande de fréquence (*cf.* partie 1.3.5). **Plusieurs des techniques existantes sont basées sur le principe qu'il est inutile de supprimer les composantes de bruit qui sont de toute façon masquées (rendues inaudibles) par le signal de parole [Lin 2002]**. L'intérêt de cette approche est qu'ainsi la distorsion du signal utile est minimisée.

En contrepartie celle du bruit augmente mais cet effet est inaudible car le bruit non supprimé est masqué. Le problème majeur consiste donc à estimer un seuil de masquage valide étant donné que seul le signal bruité est disponible [Virag 1996]. La solution consiste à réaliser une première réduction de bruit classique de façon à obtenir une estimation du signal de parole propre et d'en déduire le seuil de masquage. Lorsque le RSB du signal bruité n'est pas trop critique le seuil estimé est proche de celui obtenu directement à partir du signal de parole propre.

À partir du seuil de masquage il est possible de déterminer le niveau de bruit audible qui pourra ensuite être utilisé pour exprimer le filtre de réduction de bruit ou pour le contraindre. Soit $\hat{S}(p,k)$ l'estimée du spectre de parole qui a servi à calculer le seuil de masquage $T(p,k)$. Alors, selon [Tsoukalas 1993], le module carré du bruit audible peut s'exprimer ainsi :

$$|\hat{B}^{aud}(p,k)|^2 = \begin{cases} |X(p,k)|^2 - |\hat{S}(p,k)|^2 & \text{si } |X(p,k)|^2 \geq T(p,k) \text{ et } |\hat{S}(p,k)|^2 \geq T(p,k), \\ |X(p,k)|^2 - T(p,k) & \text{si } |X(p,k)|^2 \geq T(p,k) \text{ et } |\hat{S}(p,k)|^2 < T(p,k), \\ 0 & \text{sinon} \end{cases} \tag{2.54}$$

où l'exposant *aud* signifie qu'il s'agit de la partie audible du bruit. Cette estimateur est repris dans [Akbari Azirani 1995b] où il est proposé de conserver le signal bruité lorsque le bruit est inaudible, *i.e.* $|\hat{B}^{aud}(p,k)|^2 = 0$, et d'utiliser le filtre de Wiener avec l'approche "deux-états" lorsque le bruit reste audible, *i.e.* $|\hat{B}^{aud}(p,k)|^2 > 0$. Dans ce dernier cas, une partie plus ou moins importante du bruit peut malgré tout être masquée. Il est possible d'exploiter plus avant le masquage psychoacoustique en conservant cette quantité inaudible de bruit de façon à limiter la distorsion de la parole. Cette approche a été exploitée dans [Lin 2002] où la DSP du bruit audible est définie comme ceci :

$$\mathrm{E}[|B^{aud}(p,k)|^2] = \max(\mathrm{E}[|B(p,k)|^2] - T(p,k), 0), \tag{2.55}$$

où $\mathrm{E}[|B(p,k)|^2]$ est la DSP de bruit classique qui peut être estimée en utilisant, par exemple, une des approches décrites dans la partie 2.5 et $\max(.,0)$ est le maximum par rapport à 0. La DSP du bruit audible, qui est à supprimer, est alors utilisée pour modifier l'expression du filtre de Wiener :

$$G_W^{aud} = \frac{\mathrm{E}[|S(p,k)|^2]}{\mathrm{E}[|S(p,k)|^2] + \max(\mathrm{E}[|B(p,k)|^2] - T(p,k), 0)} \tag{2.56}$$

Ainsi, seule la partie audible du bruit est supprimée tout en conservant la partie masquée ce qui limite les distorsions de la parole.

Il existe des approches qui s'écartent quelque peu du cadre qui vient d'être décrit. Ainsi, Virag a proposé dans [Virag 1996, Virag 1999] d'adapter les paramètres α et β de la soustraction spectrale généralisée (*cf.* partie 2.4.1.2) en fonction du seuil de masquage ce qui permet de limiter les distorsions de la parole ainsi que le niveau de bruit musical. Gustafsson quant à lui a proposé dans [Gustafsson 1998] d'utiliser un gain permettant de masquer les distorsions du bruit. Le gain résultant ne dépend que du seuil de masquage et de la DSP du bruit. Ces deux quantités étant très lissées en fréquence, le signal restauré ne souffre pas du bruit musical. Par contre, pour des RSB faibles les distorsions de la parole sont importantes.

Les méthodes mettant à profit le masquage psychoacoustique permettent donc de limiter la distorsion du signal utile (exceptée celle proposée par Gustafsson dont le but est de réduire la distorsion du bruit) et de diminuer l'effet de bruit musical résiduel notamment lorsqu'elles sont associées à des approches de type soustraction spectrale.

2.5 Techniques d'estimation du bruit

2.5.1 Introduction

Nous avons vu que l'efficacité des techniques d'atténuation spectrale à court terme dépend du gain choisi et de la qualité des estimateurs du RSB. **Les différents RSB utilisés requièrent tous l'estimation de la DSP du bruit. Ainsi, les performances des systèmes de réduction de bruit sont-elles très dépendantes de la qualité de l'estimation de cette DSP.** Si l'hypothèse de stationnarité du bruit est valide alors l'estimation de sa DSP peut être faite à long terme pendant les périodes d'inactivité vocale. Cette approche nécessite évidemment une détection d'activité vocale (DAV) robuste ce qui est pour le moins problématique lorsque le RSB global est faible. Cependant, dans nombre de cas, le bruit peut être amené à évoluer plus ou moins rapidement et peut présenter les mêmes caractéristiques de non-stationnarité que la parole (bruit de foule par exemple). Il est alors indispensable de suivre les évolutions du bruit aussi fidèlement que possible de façon continue. Des techniques répondant à ce problème existent et permettent de poursuivre les non-stationnarités du bruit pendant l'activité vocale aussi bien que pendant l'inactivité vocale. Il faut noter que, même dans ce cas, l'estimation de la DSP du bruit reste lissée pour éviter une variabilité trop importante qui pourrait être une source de dégradation en cas d'erreur d'estimation. Ces approches d'estimation en continu ne nécessitent pas de DAV ce qui est un avantage compte tenu de leur piètre robustesse à faible RSB.

Bien que la DSP du bruit soit estimée à long terme, elle est calculée à partir de trames à court terme et l'indice de trame p sera conservé pour faciliter la présentation des différents estimateurs. Dans cette partie, l'estimée de cette quantité sera donc notée $\hat{\gamma}_b(p,k)$. Par contre, dans le reste du document la notation $\hat{\gamma}_b(k)$ sera conservée.

2.5.2 Estimation du bruit nécessitant une DAV

Dans ce cas de figure, la DSP du bruit est estimée à long terme et uniquement pendant l'inactivité vocale, c'est-à-dire qu'une DAV [Van Gerven 1997, Beritelli 2002] renseigne l'estimateur sur la présence ou non de parole pour la trame courante. Les périodes d'inactivité vocale doivent être suffisamment longues pour obtenir une estimée de la DSP du bruit possédant une faible variance ce qui ne pose pas de problème en pratique. Si la trame courante contient uniquement du bruit alors l'estimée de la DSP du bruit est obtenue par un lissage exponentiel sinon, *i.e.* la trame courante contient de la parole, l'estimation de la DSP du bruit est figée [Scalart 1996a] :

$$\begin{cases} \hat{\gamma}_b(p,k) = \lambda_B \hat{\gamma}_b(p-1,k) + (1-\lambda_B)|X(p,k)|^2 & \text{si la trame } p \text{ contient uniquement du bruit,} \\ \hat{\gamma}_b(p,k) = \hat{\gamma}_b(p-1,k) & \text{sinon.} \end{cases} \tag{2.57}$$

Le paramètre λ_B contrôle le degré de lissage désiré et peut éventuellement varier au cours du temps. En général $0{,}50 \leq \lambda_B \leq 0{,}99$ et peut se calculer comme ceci :

$$\lambda_B = exp(-R/(Fe\tau_b)) \tag{2.58}$$

où R est le nombre de points de recouvrement entre deux trames successives, Fe la fréquence d'échantillonnage et τ_b une constante de temps fixée selon le degré de stationnarité du bruit.

2.5.3 Estimation du bruit en continu (sans DAV)

Des approches permettant de poursuivre les non-stationnarités du bruit en continu, *i.e.* indifféremment pendant l'inactivité ou l'activité vocale, ont été développées. L'aperçu qui va suivre n'est pas exhaustif mais permet de donner un éventail des différentes approches possibles.

2.5.3.1 Estimation bornée

La méthode la plus simple a été proposée dans [Arslan 1995]. L'estimation se fait par lissage exponentiel à chaque trame selon la première expression de l'équation (2.57) mais également pendant l'activité vocale. Cependant, l'évolution de la DSP du bruit est contrainte de la manière suivante :

$$1{,}006\sqrt{\hat{\gamma}_b(p-1,k)} \leq \sqrt{\hat{\gamma}_b(p,k)} \leq 0{,}978\sqrt{\hat{\gamma}_b(p-1,k)}. \tag{2.59}$$

Ainsi, l'estimée de la DSP du bruit ne peut pas augmenter de plus de 3dB par seconde et diminuer de plus de 12dB par seconde avec les paramètres de traitement donnés dans [Arslan 1995]. L'estimée de la DSP du bruit augmente lentement pendant l'activité vocale mais a la capacité de retourner rapidement à une valeur correcte pendant l'inactivité vocale. Cette approche a l'avantage d'être très simple, cependant, l'estimée de la DSP du bruit est biaisée pendant l'activité vocale ce qui engendre des distorsions du signal restauré.

2.5.3.2 Estimation par histogramme

Deux méthodes ont été proposées par Hirsch. La première proposée dans [Hirsch 1995] présente l'avantage d'être très simple. Lorsque le module de l'observation $|X(p,k)|$ est supérieur à un seuil donné par

$$Seuil(p,k) = \beta\sqrt{\hat{\gamma}_b(p-1,k)} \tag{2.60}$$

avec β prenant une valeur entre 1.5 et 2.5, alors la composante fréquentielle est considérée comme étant de la parole et est ignorée. Sinon la DSP de bruit est mise à jour par lissage exponentiel selon l'équation (2.57).

La seconde méthode qui est décrite également dans [Hirsch 1995] est basée sur ce même principe. Cependant, lorsque le module de l'observation $|X(p,k)|$ est inférieur au seuil (2.60), la DSP du bruit est estimée à partir de l'histogramme des valeurs passées, sur 400ms, de $|X(p,k)|$ et ce dans 40 bandes de fréquences. Dans chacune d'elles, l'amplitude correspondant au maximum de la distribution est choisie comme estimée de l'amplitude du bruit. Cette estimée est ensuite lissée pour éviter les pics qui pourraient survenir. Cette approche permet d'obtenir des résultats plus précis que la technique de lissage exponentiel classique.

2.5.3.3 Approche minimum statistics

L'approche "minimum statistics" proposée par Martin dans [Martin 1993, Martin 1994] est basée sur la poursuite du minimum de la DSP de la parole bruitée. Cette DSP est lissée en utilisant

l'estimateur récursif suivant :

$$\hat{\gamma}_x(p,k) = \lambda_X \hat{\gamma}_x(p-1,k) + (1-\lambda_X)|X(p,k)|^2 \tag{2.61}$$

où λ_X est un facteur d'oubli tel que $\lambda_X \leq 0{,}9$. Ce paramètre doit idéalement rester relativement faible étant donné la nature non-stationnaire du signal de parole. La recherche du minimum de $\hat{\gamma}_x(p,k)$ se fait sur une fenêtre temporelle de longueur D trames. Cette recherche est basée sur l'hypothèse que les pics de la DSP $\hat{\gamma}_x(p,k)$ correspondent à de la parole alors que les vallées correspondent au spectre lissé du bruit. L'estimée de $\hat{\gamma}_b(p,k)$ est alors mise à jour à partir du minimum de $\hat{\gamma}_x(p,k)$ pondéré par un facteur de correction calculé de façon à obtenir une estimation non biaisée. Les performances de cette technique dépendent du choix des paramètres, ainsi la longueur D de la fenêtre de recherche est soumise au compromis suivant : si cette fenêtre est trop courte alors la DSP du bruit aura tendance à être surestimée par contre si elle est trop longue alors le suivi des non-stationnarités devient moins réactif (temps de réponse à une variation du bruit pouvant dépasser une à deux secondes). Ainsi pour un bruit stationnaire, cette approche donne des résultats équivalents à une technique utilisant une DAV robuste. Par contre, si le bruit est non-stationnaire, cette approche permet un bon suivi des évolutions du bruit pendant l'activité vocale ce qui lui permet de se démarquer des techniques à base de DAV.

2.5.3.4 Approche de Doblinger

Doblinger inscrit sa technique d'estimation de la DSP du bruit [Doblinger 1995] comme une alternative moins complexe à l'approche "minimum statistics". À partir de la DSP du signal bruité $\hat{\gamma}_x(p,k)$ obtenue en utilisant l'équation (2.61), l'estimée de la DSP du bruit est calculée ainsi :

$$\begin{cases} \hat{\gamma}_b(p,k) = \alpha\hat{\gamma}_b(p-1,k) + \frac{1-\alpha}{1-\beta}[\hat{\gamma}_x(p,k) - \beta\hat{\gamma}_x(p-1,k)] \\ \qquad \text{si } \hat{\gamma}_x(p,k) > \hat{\gamma}_b(p-1,k), \\ \hat{\gamma}_b(p,k) = \hat{\gamma}_x(p,k) \\ \qquad \text{sinon.} \end{cases} \tag{2.62}$$

En fait, le terme $\hat{\gamma}_x(p,k) - \beta\hat{\gamma}_x(p-1,k)$ qui apparaît dans cette équation permet de réaliser un suivi automatique et peu complexe du minimum. Les paramètres utilisés sont $\lambda_X = 0{,}7$, $\alpha = 0{,}998$ et $\beta = 0{,}96$ ce qui correspond à une période d'adaptation de la DSP du bruit comprise entre 200 et 400ms pour les paramètres de traitement donnés dans [Doblinger 1995].

2.5.3.5 Approche MCRA

Une approche récente appelée "minima controlled recursive averaging" (MCRA) a été proposée dans [Cohen 2002a]. En se basant sur l'approche "deux-états" introduite dans [McAulay 1980] il est possible d'écrire l'estimée de la DSP du bruit en fonction de la probabilité conditionnelle de présence de la parole $\mathrm{p}'(p,k)$:

$$\hat{\gamma}_b(p,k) = \tilde{\lambda}_B \hat{\gamma}_b(p-1,k) + (1-\tilde{\lambda}_B)|X(p,k)|^2 \tag{2.63}$$

où

$$\tilde{\lambda}_B = \lambda_B + (1-\lambda_B)\mathrm{p}'(p,k) \tag{2.64}$$

est un paramètre de lissage variant dans le temps. Ce paramètre est dépendant du paramètre de lissage constant λ_B et est fonction de la probabilité conditionnelle $p'(p,k)$ qui doit être estimée. Pour cela, à partir de la DSP du signal bruité $\hat{\gamma}_x(p,k)$ obtenue en utilisant l'équation (2.61), le minimum local de cette DSP est poursuivi grâce à la procédure suivante :

$$\begin{cases} \hat{\gamma}_x^{min}(p,k) = \min(\hat{\gamma}_x^{min}(p-1,k), \hat{\gamma}_x(p,k)), \\ \hat{\gamma}_x^{tmp}(p,k) = \min(\hat{\gamma}_x^{tmp}(p-1,k), \hat{\gamma}_x(p,k)). \end{cases} \tag{2.65}$$

Quand une fenêtre de L trames a été traitée, *i.e.* p divisible par L, alors la variable temporaire $\hat{\gamma}_x^{tmp}(p,k)$ est utilisée ainsi :

$$\begin{cases} \hat{\gamma}_x^{min}(p,k) = \min(\hat{\gamma}_x^{tmp}(p-1,k), \hat{\gamma}_x(p,k)), \\ \hat{\gamma}_x^{tmp}(p,k) = \hat{\gamma}_x(p,k), \end{cases} \tag{2.66}$$

puis la recherche de minimum se poursuit en utilisant l'équation (2.65). Ceci permet de remettre régulièrement à jour le minimum. Ensuite, la quantité $\hat{\gamma}_x(p,k)/\hat{\gamma}_x^{min}(p,k)$ est comparée à un seuil empirique δ. Le résultat binaire de cette comparaison $I(p,k)$ est alors utilisé pour estimer la probabilité conditionnelle $p'(p,k)$:

$$\hat{p}'(p,k) = \lambda_P \hat{p}'(p-1,k) + (1-\lambda_P) I(p,k) \tag{2.67}$$

où $0 < \lambda_P < 1$ est un paramètre de lissage.

Cette approche est soumise au même type de compromis que l'approche "minimum statistics". En effet la longueur L de la fenêtre de recherche du minimum doit être suffisamment grande pour assurer un suivi fiable et suffisamment courte pour pouvoir poursuivre les non-stationnarités du bruit. Selon le compromis choisi, le temps de réponse à une variation du bruit peut aller jusqu'à 2 secondes. L'auteur a par la suite proposé une amélioration de cette approche nommée "improved MCRA" qui améliore les performances de la technique MCRA notamment pour les faibles RSB et les bruits non-stationnaires [Cohen 2003].

2.5.3.6 Estimation par pondération du spectre bruité

Dans cette technique proposée dans [Kato 2003], le RSB *a posteriori* de la trame courante est utilisé pour calculer un facteur de pondération non-linéaire qui lui est inversement proportionnel. Le spectre du signal bruité qui est pondéré par ce facteur est alors utilisé pour estimer la DSP du bruit en utilisant une moyenne des valeurs pondérées sur une mémoire de plusieurs trames. Les composantes possédant de forts RSB ne sont pas utilisées pour estimer la DSP du bruit ce qui permet d'éviter les surestimations. Cette technique donne de bons résultats et a été intégrée dans un mobile de 3G au Japon dont le système de réduction de bruit satisfait aux spécifications du 3GPP (The 3rd generation partnership project agreement).

2.5.4 Conclusion

Il ressort de cette revue des techniques d'estimation de la DSP du bruit que les méthodes d'estimation en continu présentent deux avantages par rapport aux techniques à base de DAV. D'une part, il n'est pas nécessaire de créer une DAV robuste, ce qui est très difficile à faible RSB. D'autre part,

elles permettent de suivre les non-stationnarités du bruit pendant les périodes d'activité vocale. Ceci peut s'avérer indispensable dans certains cas où la DAV est inutilisable. On peut noter par exemple les applications sans instant de silence (*e.g.* messagerie vocale, musique) ou encore plus généralement les applications où le RSB est très faible. Néanmoins, aucune technique d'estimation de la DSP du bruit ne permet de réaliser un suivi fin des évolutions du bruit trame par trame. En effet, il y a toujours un délai de réponse plus ou moins long à une variation du niveau de bruit.

2.6 Conclusion

Comme on l'a vu dans la partie 2.3, la structure générale des techniques d'atténuation spectrale à court terme répond à des caractéristiques précises. Il existe toutefois une grande diversité dans ces approches qui s'exprime essentiellement par le biais du gain spectral et de l'estimation de la DSP du bruit. Au vu de l'analyse des techniques présentées dans ce chapitre, on ressent au delà du gain spectral la forte influence des RSB *a posteriori* et *a priori*. En effet, le gain spectral s'exprime en fonction du RSB et dépend donc fortement de la qualité de son estimateur. Le RSB est en réalité le paramètre clé des techniques de réduction de bruit. Le gain spectral peut quant à lui être vu comme un moyen de régler le compromis entre la réduction effective de bruit et la dégradation du signal de parole. Ainsi, pour un gain spectral donné, le seul moyen d'améliorer la qualité du signal restauré, et il existe une marge de progression intéressante, est d'améliorer l'estimation du RSB. Le chapitre suivant traite de ce sujet et est plus généralement consacré aux diverses limitations des techniques de réduction de bruit.

Références

[Akbari Azirani 1995b] A. Akbari Azirani, R. Le Bouquin Jeannes, et G. Faucon, "Optimizing Speech Enhancement by Exploiting Masking Properties of the Human Ear," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Detroit, États-Unis, Vol. 1, pp. 800–803, Mai 1995.

[Arslan 1995] L. Arslan, A. McCree, et V. Viswanathan, "New Methods for Adaptive Noise Suppression," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Detroit, États-Unis, Vol. 1, pp. 812–815, Mai 1995.

[Beritelli 2002] F. Beritelli, S. Casale, G. Ruggeri, et S. Serrano, "Performance Evaluation and Comparison of G.729/AMR/Fuzzy Voice Activity Detectors," *IEEE Signal Processing Lett.*, Vol. 9, No. 3, Mars 2002.

[Berouti 1979] M. Berouti, R. Schwartz, et J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Washington, États-Unis, pp. 208–211, Avril 1979.

[Boll 1979] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-27, No. 2, pp. 113–120, Avril 1979.

[Breithaupt 2003] C. Breithaupt, et R. Martin, "MMSE Estimation of Magnitude-Squared DFT Coefficients with Supergaussian Priors," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Vol. 1, pp. 896–899, 2003.

[Cappé 1993] O. Cappé, "Techniques de Réduction de Bruit pour la Restauration d'Enregistrements Musicaux," *Thèse de l'École Nationale Supérieure des Télécommunications*, Paris, Septembre 1993.

[Cappé 1994] O. Cappé, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor," *IEEE Trans. Speech Audio Processing*, Vol. 2, No. 2, pp. 345–349, Avril 1994.

[Chen 2005] B. Chen, et P. C. Loizou, "Speech Enhancement using a MMSE Short Time Spectral Amplitude Estimator with Laplacian Speech Modeling," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Philadelphie, États-Unis, Vol. 1, pp. 1097–1100, Mars 2005.

[Cohen 2002a] I. Cohen, et B. Berdugo, "Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement," *IEEE Signal Processing Lett.*, Vol. 9, No. 1, pp. 12–15, Janvier 2002.

[Cohen 2002b] I. Cohen, "Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator," *IEEE Signal Processing Lett.*, Vol. 9, Issue 4, pp. 113–116, Avril 2002.

[Cohen 2003] I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging," *IEEE Trans. Speech Audio Processing*, Vol. 11, No. 5, pp. 466–475, Septembre 2003.

[Doblinger 1995] G. Doblinger, "Computationally Efficient Speech Enhancement by Spectral Minima Tracking in Subbands," *Eurospeech*, Madrid, Espagne, Vol. 2, pp. 1513–1516, Septembre 1995.

[Ephraïm 1984] Y. Ephraïm, et D. Malah, "Speech Enhancement Using a Minimum Mean Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-32, No. 6, pp. 1109–1121, Décembre 1984.

[Ephraïm 1985] Y. Ephraïm, et D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-33, No.3, pp. 443–445, Avril 1985.

[Godsill 1998] S. J. Godsill, P. J. W. Rayner, et O. Cappé, "Digital Audio Restoration," *Appl. of Digital Signal Processing to Audio and Acoust.*, Kluwer Academic Publishers, pp. 133–193, 1993.

[Guédon 2002] L. Guédon, "Mise en œuvre de Nouvelles Hypothèses dans les Algorithmes de Réduction de Bruit par Atténuation Spectrale," *Document interne FT R&D*, Août 2002.

[Guérin 2002] A. Guérin, "Rehaussement de la Parole pour les Communications mains-libres. Réduction de Bruit et Annulation d'Écho Non Linéaire," *Thèse de l'Université de Rennes 1*, 2002.

[Gustafsson 1998] S. Gustafsson, P. Jax, et P. Vary, "A Novel Psychoacoustically Motivated Audio Enhancement Algorithm Preserving Background Noise Characteristics," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Seattle, États-Unis, pp. 397–400, Mai 1998.

[Hirsch 1995] H. G. Hirsch, et C. Ehrlicher, "Noise Estimation Techniques for Robust Speech Recognition," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Detroit, États-Unis, Vol. 1, pp. 153–156, Mai 1995.

[Kato 2003] M. Kato, M. Serizawa, N. Toki, U. Mori, Y. Morishita, et K. Hayashi, "Noise Suppression with High Speech Quality Based on Weighted Noise Estimation for 3G Handsets," *NEC Res. & Develop. Special issue on Device and Systems for Mobile Communications*, Vol. 44, No. 4, pp. 340–348, Octobre 2003.

[Le Bouquin 1991] R. Le Bouquin, "Traitements pour la Réduction du Bruit sur de la Parole. Application aux Communications Radio-Mobiles," *Thèse de l'Université de Rennes 1*, 1991.

[Lim 1979] J. S. Lim, et A. V. Oppenheim, "Enhancement and Bandwith Compression of Noisy Speech," *Proc. IEEE*, Vol. 67, No. 12, pp. 1586–1604, Décembre 1979.

[Lin 2002] L. Lin, W. H. Holmes, et E. Ambikairajah, "Speech Denoising Using Perceptual Modification of Wiener Filtering," *IEEE Electronics Lett.*, Vol. 38, No. 23, pp. 1486–1487, Novembre 2002.

[Malah 1999] D. Malah, R. V. Cox, et A. J. Accardi, "Tracking Speech-Presence Uncertainty to Improve Speech Enhancement in Non-Stationary Noise Environments," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Phoenix, États-Unis, pp. 789–792, Mars 1999.

[Martin 1993] R. Martin, "An Efficient Algorithm to Estimate the Instantaneous SNR of Speech Signals," *Eurospeech*, Berlin, Allemagne, pp. 1093–1096, Septembre 1993.

[Martin 1994] R. Martin, "Spectral Subtraction Based on Minimum Statistics," *Eusipco*, Edinburgh, Royaume-Uni, pp. 1182–1185, Septembre 1994.

[Martin 2002] R. Martin, “Speech Enhancement Using MMSE Short Time Spectral Estimation with Gamma Distributed Speech Priors,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Orlando, États-Unis, Vol. 1, pp. 253–256, Mai 2002.

[McAulay 1980] J. McAulay, et M. Malpass, “Speech Enhancement Using a Soft-Decision Noise Suppression Filter,” *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, No. 2, pp. 137–145, Avril 1980.

[Plapous 2004] C. Plapous, C. Marro, L. Mauuary, et P. Scalart, “Two-Step Noise Reduction Technique,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Montréal, Canada, Vol. 1, pp. 289–292, Mai 2004.

[Porter 1984] J. E. Porter, et S. F. Boll, “Optimal Estimators for Spectral Restoration of Noisy Speech,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, San Diego, États-Unis, Vol. 2, pp. pp. 18A2.1–2.4, Mars 1984.

[Scalart 1996a] P. Scalart, et J. Vieira Filho, “Speech Enhancement Based on a Priori Signal to Noise Estimation,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Atlanta, États-Unis, Vol. 2, pp. 629–632, Mai 1996.

[Sim 1998] B. L. Sim, Y. C. Tong, J. S. Chang, et C. T. Tan, “A Parametric Formulation of the Generalized Spectral Subtraction Method,” *IEEE Trans. Speech Audio Processing*, Vol. 6, No. 4, pp. 328–337, Juillet 1998.

[Tsoukalas 1993] D. Tsoukalas, M. Paraskevas, et J. Mourjopoulos, “Speech Enhancement Using Psychoacoustic Criteria,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Minneapolis, États-Unis, Vol. 2, pp. 359–362, Avril 1993.

[Van Gerven 1997] S. Van Gerven, et F. Xie, “A Comparative Study of Speech Detection Methods,” *Eurospeech*, Vol. 3, pp. 1095–1098, Grèce, Septembre 1997.

[Vary 1985] P. Vary, “Noise Suppression by Spectral Magnitude Estimation - Mechanism and Theoretical Limits,” *Signal Processing*, Vol. 8, pp. 387–400, 1985.

[Virag 1996] N. Virag, “Speech Enhancement Based on Masking Properties of the Human Auditory System,” *Thèse de l’École Polytechnique Fédérale de Lausanne*, 1996.

[Virag 1999] N. Virag, “Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System,” *IEEE Trans. Speech Audio Processing*, Vol. 7, No. 2, pp. 126–137, Mars 1999.

[Wang 1982] D. L. Wang, et J. S. Lim, “The Unimportance of Phase in Speech Enhancement,” *Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-30, No. 4, pp. 679–681, Août 1982.

Chapitre 3

Limitations des techniques de réduction de bruit

Avant même de penser à améliorer les performances des techniques de réduction de bruit, la première étape consiste à identifier les limitations, points de blocage et autres défauts des méthodes antérieures. Ce chapitre y est entièrement consacré et montre qu'il existe une marge de progression intéressante sur différents points clés des approches par atténuation spectrale à court terme. La première limitation, évoquée dans la partie 3.1, concerne les techniques de réduction de bruit nécessitant un modèle statistique pour les signaux de parole et de bruit (*cf.* partie 2.4.2). En effet, le modèle gaussien est choisi pour sa simplicité d'utilisation et non pas pour son adéquation avec les signaux réels. La seconde limitation, des plus fondamentales, concerne l'estimation du RSB et est détaillée dans la partie 3.2. En pratique il existe deux types de RSB différents qui possèdent chacun des défauts et des avantages qu'il serait intéressant de concilier de façon à obtenir un estimateur du RSB beaucoup plus efficace. Une autre limitation, exposée dans la partie 3.3, concerne l'estimation du bruit qui est loin d'être parfaite mais qui conditionne malgré tout les performances des techniques de réduction de bruit. Finalement, le rôle et l'impact important de la phase, généralement négligé, sera mis en avant dans la partie 3.4. Les limitations ainsi identifiées et analysées serviront de base aux axes d'amélioration proposés dans la suite du mémoire.

3.1 Adéquation entre les modèles statistiques théoriques et réels

Dans la partie 1.3.3 il est rappelé qu'une loi Gamma ou encore de Laplace est plus appropriée que la loi de Gauss classique pour modéliser la densité de probabilité d'un signal temporel de parole. Les techniques de réduction de bruit impliquent généralement de travailler sur des trames d'une durée de quelques dizaines de ms (< 50ms) pour des raisons d'une part de stationnarité à court terme du signal de parole et d'autre part de mise en œuvre et de complexité. Dans le domaine fréquentiel, quand un modèle statistique est nécessaire, la loi de Gauss est très largement utilisée dans la mesure où elle simplifie nombre de calculs, en témoignent les techniques présentées dans la partie 2.4.2. Cependant, à l'instar du domaine temporel, nous allons vérifier que dans le domaine fréquentiel aussi il existe des

modèles mieux adaptés au signal de parole que la loi de Gauss [Martin 2002]. Dans la suite de cette partie, pour alléger les écritures, la quantité fréquentielle complexe $Y(p,k)$ sera notée Y et ses parties réelle et imaginaire seront respectivement notées Y_R et Y_I. De la même façon, la variance de Y (évaluée sur chaque composante spectrale d'analyse) sera notée γ_y. L'énergie des signaux est supposée répartie de façon égale entre les parties réelle et imaginaire. Ainsi, la variance des signaux Y_R et Y_I sera $\gamma_y/2$. La moyenne ne portant aucune information, celle-ci est toujours supposée nulle. Ainsi, dans la suite on s'intéressera seulement à la densité de probabilité de la partie réelle Y_R du signal Y, les résultats pouvant être généralisés à la partie imaginaire. Sous ces hypothèses, les différents modèles statistiques considérés pour le signal Y_R s'écrivent ainsi :

– loi de Gauss :

$$\mathrm{p}(Y_R) = \frac{1}{\sqrt{\pi\gamma_y}} exp\left(-\frac{Y_R^2}{\gamma_y}\right), \tag{3.1}$$

– loi de Laplace :

$$\mathrm{p}(Y_R) = \frac{1}{\sqrt{\gamma_y}} exp\left(-\frac{2|Y_R|}{\sqrt{\gamma_y}}\right), \tag{3.2}$$

– loi Gamma :

$$\mathrm{p}(Y_R) = \frac{\sqrt[4]{3}}{2\sqrt{\pi}\sqrt[4]{2\gamma_y}} |Y_R|^{-\frac{1}{2}} exp\left(-\frac{\sqrt{3}|Y_R|}{\sqrt{2\gamma_y}}\right). \tag{3.3}$$

Les figures 3.1 et 3.2 représentent la densité de probabilité expérimentale de la partie réelle S_R des coefficients complexes de parole utile (*i.e.* sans silence) S pour un intervalle de fréquence compris entre 1900 et 2100Hz. La fréquence d'échantillonnage est de 8kHz et les trames analysées ont une longueur de 32ms. On peut noter que la méthodologie utilisée ici est quelque peu différente de celle présentée dans [Martin 2002] où les coefficients complexes sont sélectionnés sur une base énergétique. D'autres intervalles de fréquence ont été testés avec des résultats équivalents. Cette densité expérimentale correspond à l'histogramme obtenu pour environ 50s de parole prononcée par 4 locuteurs différents, 2 femmes et 2 hommes. Les densités de probabilité associées aux différents modèles statistiques proposés sont également représentées. La figure 3.2 correspond à un zoom de la figure 3.1 pour mieux apprécier les différences entre l'histogramme expérimental et les densités de probabilité théoriques. Ces deux figures montrent que les lois de Laplace et surtout Gamma modélisent mieux les parties réelle et imaginaire (par extension) du signal utile que la loi de Gauss classiquement utilisée.

La mesure de divergence de Kullback [Kullback 1958, Breithaupt 2003] permet de corroborer cette observation. Il s'agit d'une mesure utilisée pour quantifier l'adéquation entre deux densités de probabilité, soit dans le cas présent entre la densité expérimentale $\mathrm{p}_{exp}(\nu)$ et la densité théorique $\mathrm{p}_{th}(\nu)$:

$$J(exp:th) = \sum_{\nu=1}^{N} (\mathrm{p}_{exp}(\nu) - \mathrm{p}_{th}(\nu)) \log\left(\frac{\mathrm{p}_{exp}(\nu)}{\mathrm{p}_{th}(\nu)}\right) \tag{3.4}$$

où N est le nombre de canaux des deux densités discrètes. Le tableau 3.1 contient les mesures de divergence de Kullback pour un signal de parole. Elles sont calculées pour chacune des 3 lois discrètes théoriques exprimées par les équations (3.1) à (3.3). Ces mesures sont normalisées par celle obtenue pour la loi de Gauss afin d'en faciliter l'interprétation. Ainsi, la valeur obtenue pour la loi de Laplace est environ 2 fois plus faible que la référence (loi de Gauss) et environ 8 fois plus faible dans le

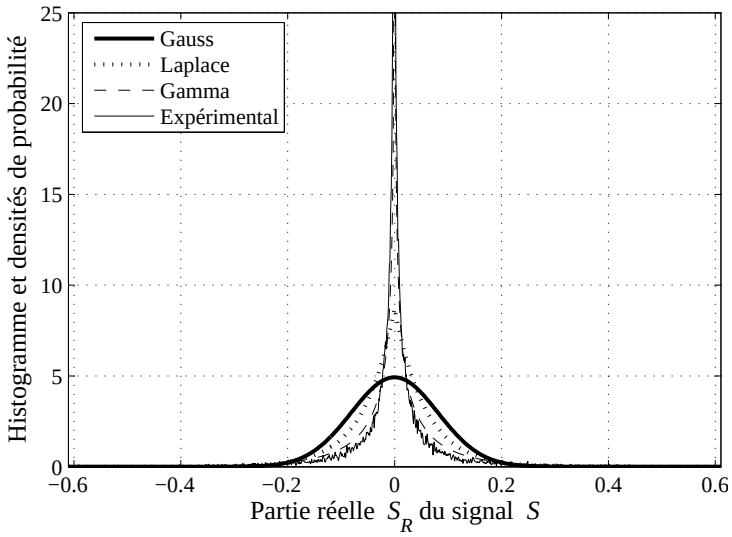

FIG. 3.1 – *Loi de Gauss (trait fort), de Laplace (pointillé) et Gamma (tirets) ainsi que la densité de probabilité expérimentale de la partie réelle d'un signal de parole pour les fréquences comprises entre 1900 et 2100Hz.*

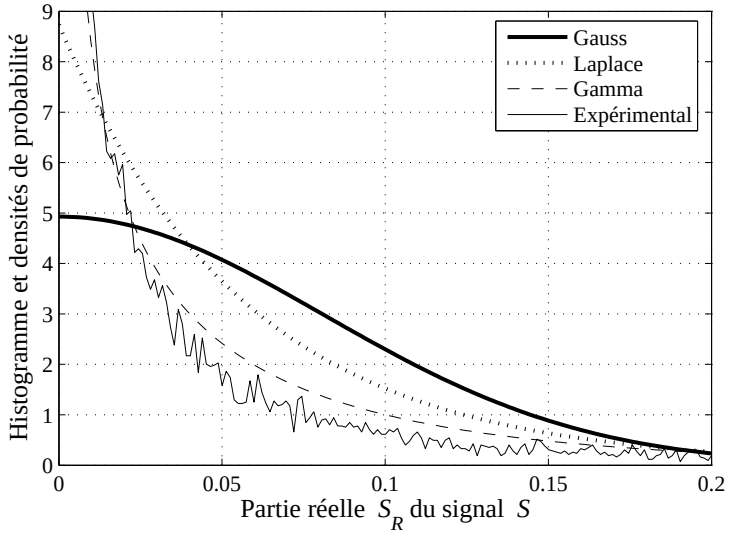

FIG. 3.2 – *Zoom de la figure 3.1.*

cas de la loi Gamma. **Ces deux modèles et en particulier le modèle Gamma constituent donc de meilleures hypothèses que le modèle gaussien classique pour représenter le signal de parole dans le domaine fréquentiel.**

Il est possible d'adopter la même démarche pour déterminer le meilleur modèle applicable au signal de bruit mais dans ce cas le résultat n'est pas aussi tranché que pour la parole. Ceci est illustré par la figure 3.3 qui est basée sur le même principe que la figure 3.1 mais cette fois c'est la densité de probabilité de la partie réelle B_R du signal de bruit B qui est évaluée toujours pour un intervalle de

TAB. 3.1 – *Mesure de divergence de Kullback entre les 3 modèles (loi de Gauss, de Laplace et Gamma) et la densité de probabilité expérimentale (trait fin) pour un signal de parole.*

Loi théorique (th)	$J(exp:th)/J(exp:Gauss)$
Gauss	1
Laplace	0,48
Gamma	0,12

fréquences compris entre 1900 et 2100Hz. Les figures (a) et (b) correspondent respectivement aux cas où le bruit est stationnaire (bruits Bureau et Voiture présentés dans la partie 6.1) ou non (bruits Rue et Foule). Seules les densités de probabilité associées à loi de Gauss et de Laplace sont représentées, la loi Gamma étant trop éloignée de la réalité. Dans les deux cas illustrés par les figures 3.3.(a) et 3.3.(b)

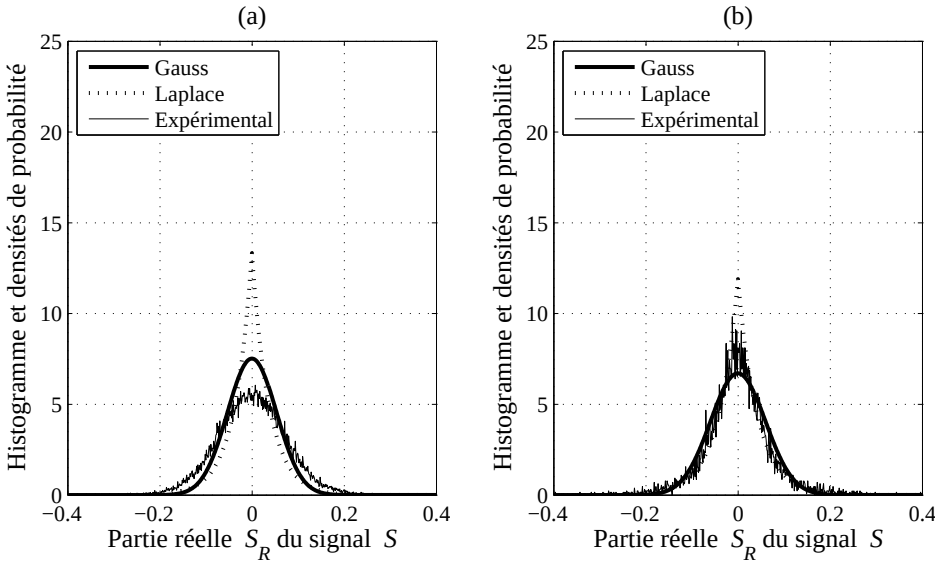

FIG. 3.3 – *Loi de Gauss (trait fort) et de Laplace (pointillé) ainsi que la densité de probabilité (trait fin) de la partie réelle d'un signal de bruit stationnaire (Voiture) (a) et non-stationnaire (Foule) (b) pour les fréquences comprises entre 1900 et 2100Hz.*

il est difficile de trancher de visu entre la loi de Gauss et de Laplace. Il faut donc avoir recours à la mesure de divergence de Kullback (*cf.* tableau 3.2) calculée pour les lois discrètes théoriques exprimées par les équations (3.1) et (3.2). À partir des résultats du tableau 3.2, **il apparaît que la loi de Laplace est un meilleur candidat que la loi de Gauss pour modéliser les bruits non-stationnaires mais elle est en revanche moins bonne pour modéliser les bruits stationnaires.**

Dans la partie 4.1, des estimateurs MMSE des parties réelle et imaginaire du signal de parole sont obtenus à partir des nouvelles hypothèses validées ici. Les densités de probabilités des signaux de parole et de bruit (non-stationnaire) étant mieux modélisées par des modèles super-gaussiens que par le modèle gaussien classique, il est légitime d'attendre de ces estimateurs de meilleurs résultats

TAB. 3.2 – *Mesure de divergence de Kullback entre 2 des 3 modèles (loi de Gauss et de Laplace) et la densité de probabilité expérimentale pour un signal de bruit stationnaire (Voiture) ou non (Foule).*

	$J(exp:th)/J(exp:Gauss)$	
Loi théorique (th)	Stationnaire	Non-stationnaire
Gauss	1	1
Laplace	1,24	0,41

qu'avec l'estimateur MMSE classique (modèle gaussien pour la parole et le bruit). On peut d'ailleurs remarquer que ce dernier n'est autre que le filtre de Wiener qui ne fait pourtant pas d'hypothèse sur les statistiques des signaux.

3.2 Limitations des estimateurs du RSB

On a vu que le RSB est un (sinon le) paramètre clé qui gouverne la qualité des techniques de réduction de bruit (*cf.* chapitre 2). Ses différents estimateurs sont toutefois soumis à certaines limitations. Pour s'en convaincre, il suffit d'un test simple : réaliser la réduction de bruit en utilisant un estimateur idéal du RSB (en connaissant bien sûr tous les signaux). Le résultat est saisissant de qualité. Si l'on n'atteint pas celle du signal propre on s'en rapproche toutefois beaucoup. **Ce test permet de confirmer qu'il existe une marge de progression énorme dans l'estimation du RSB.**

3.2.1 Signal test servant de fil rouge

Dans la suite de ce chapitre et dans les suivants, le discours sera (sauf exception) illustré à partir d'une phrase française dégradée par un bruit de voiture (décrit dans la partie 6.1) avec un RSB global de 12dB (*cf.* figure 3.4). Des résultats complets, généralisés à d'autres RSB et types de bruit, seront donnés dans la partie 6.3.

3.2.2 Prépondérance des estimateurs du RSB sur le gain spectral

Depuis maintenant près de 30 ans de nombreuses recherches ont porté sur la création de différents gains spectraux. Comme le montre la partie 2.4, le gain spectral joue un rôle important dans les techniques de réduction de bruit dans la mesure où il détermine le compromis entre la réduction de bruit et la dégradation de la parole. Ce compromis se matérialise sous la forme de la courbe de gain qui est fonction du RSB. Cependant, le paramètre réellement clé dans les techniques de réduction de bruit est justement le RSB [Scalart 1996a, Ephraïm 1984, Cappé 1994]. Pour s'en convaincre on peut prendre le cas du filtre de Wiener (*cf.* partie 2.4.1.3) qui peut s'exprimer en fonction du RSB *a priori* ou du RSB *a posteriori* (filtre pseudo-Wiener). Ces deux RSB ont des comportements très différents et selon que le filtre de Wiener est exprimé en fonction de l'un ou de l'autre, les signaux restaurés seront très différents. Ainsi, un filtre ne dépendant que du RSB *a posteriori* générera du

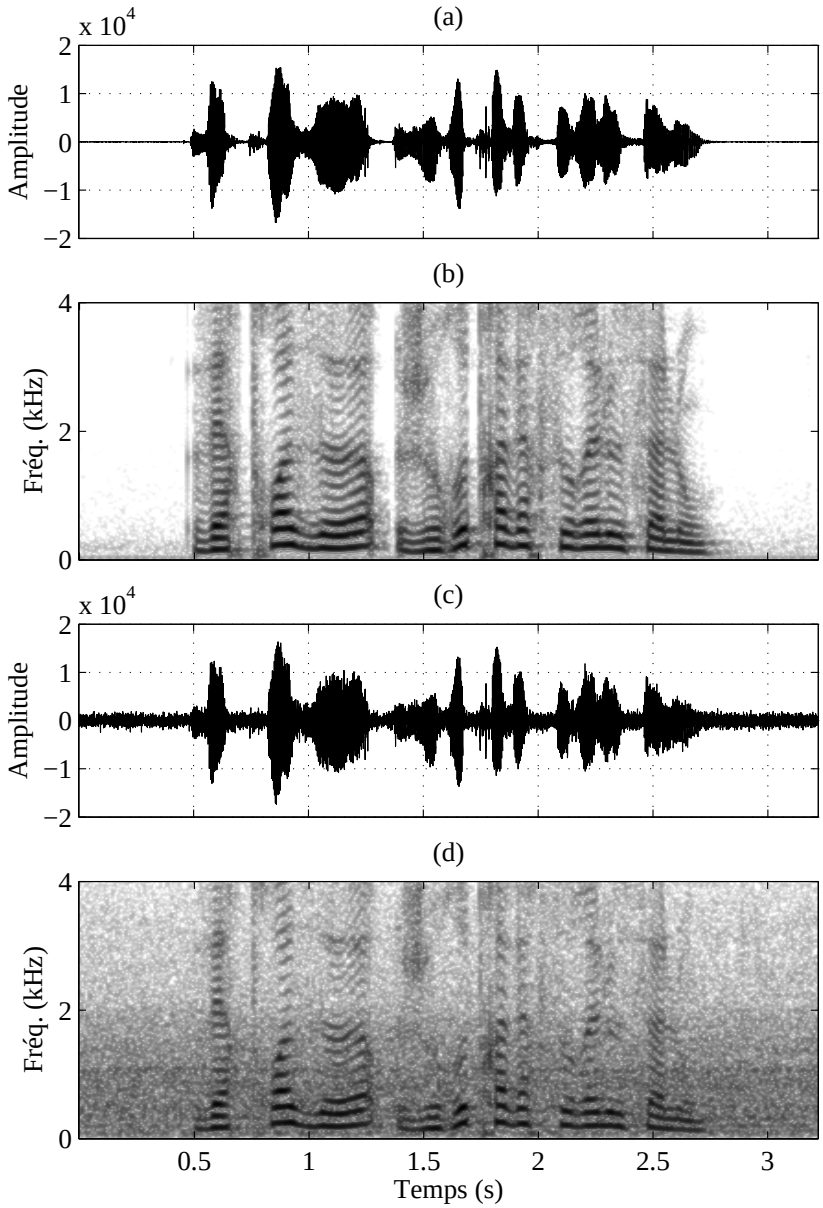

FIG. 3.4 – *(a) Forme d'onde et (b) spectre d'amplitude de la phrase "Vers trois heures je re-traverserai le salon". (c) Forme d'onde et (d) spectre d'amplitude de cette même phrase dégradée par un bruit de voiture avec un RSB global de* 12*dB.*

bruit musical, très net sur la figure 3.5.(a), et a contrario un filtre dépendant seulement du RSB *a priori* entraînera un effet de flou (analogie avec le traitement d'image) lié au lissage de son estimateur (dans le cas présent, les estimateurs du RSB sont ceux proposés dans la partie 3.2.5). Cet effet est peu visible sur le spectrogramme de la figure 3.5.(b) mais est toutefois nettement audible. On peut aussi noter sur cette figure que le niveau de bruit musical est largement réduit par l'utilisation du RSB *a priori*. **Le comportement du RSB est prépondérant sur celui du gain spectral et la qualité du signal restauré est donc essentiellement déterminée par celle des estimateurs de RSB.** En réalité,

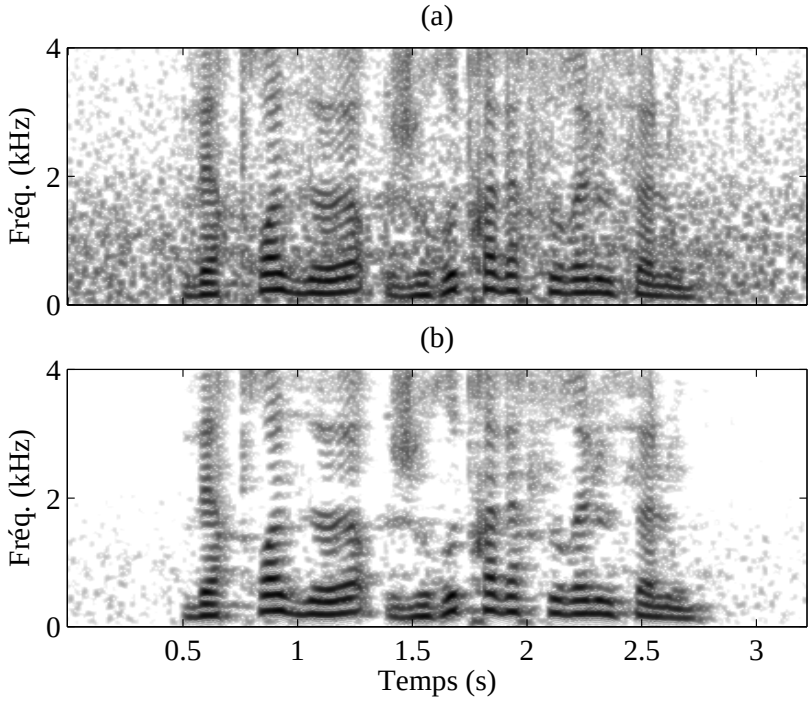

FIG. 3.5 – *Signaux restaurés en utilisant le filtre de Wiener exprimé en fonction (a) du RSB* a posteriori *et (b) du RSB* a priori.

le choix du gain permet "seulement" d'ajuster le compromis entre la distorsion de la parole et la réduction du bruit. Ceci est valable pour tous les filtres s'exprimant exclusivement en fonction de l'un ou l'autre RSB. Il faut considérer séparément les filtres de réduction de bruit où les deux types de RSB interviennent comme c'est le cas par exemple avec l'approche MMSE STSA (*cf.* partie 2.4.2.2). Dans ce cas, le RSB *a priori* impose son comportement qui est corrigé par le RSB *a posteriori*. L'adjonction de l'approche "deux-états" est aussi à mettre à part car elle permet de corriger un filtre globalement dominé par le RSB *a priori* par un facteur de pondération globalement dominé par le RSB *a posteriori* (*cf.* parties 2.4.2.1 et 2.4.2.2).

Les signaux de parole et de bruit étant supposés indépendants (*cf.* partie 2.1), on peut écrire la relation suivante entre les RSB *a posteriori* et *a priori* :

$$\mathrm{E}[RSB_{post}(p,k)] = 1 + RSB_{prio}(p,k). \tag{3.5}$$

Les quantités $RSB_{post}(p,k)$ et $1+RSB_{prio}(p,k)$ ne sont donc théoriquement pas interchangeables car le RSB *a priori* est une mesure à long terme (sur plusieurs trames) alors que le RSB *a posteriori* est une mesure à court terme (valable sur la durée d'une trame). En pratique, cependant, la quantité $RSB_{post}(p,k)$ peut être remplacée dans l'expression du gain par sa version lissée $\mathrm{E}[RSB_{post}(p,k)]$. Ce jeu d'écriture permet alors de créer à partir d'un même gain deux classes de filtres [Scalart 1996a]. En effet, un gain s'exprimant en fonction de l'un des deux RSB peut ainsi toujours être réécrit en fonction de l'autre. On conserve donc les caractéristiques du gain en terme de distorsion tout en créant deux solutions dont les résultats sont très différents. Par exemple, dans la mesure où on s'autorise à exprimer la SSP (*cf.* partie 2.4.1.1) en fonction du RSB *a priori*, les caractéristiques du signal restauré

seront imposées par le RSB *a priori*. Le signal obtenu souffrira donc peu du phénomène de bruit musical pourtant si gênant dans la SSP classique.

3.2.3 Estimateurs idéaux du RSB

Le RSB *a priori* est une quantité estimée à long terme mais qui est utilisée sur une base à court terme. Par exemple, dans le cas du filtre de Wiener, les signaux traités sont supposés stationnaires ce qui explique que le filtre soit exprimé en fonction du RSB *a priori*, quantité lissée. Mais en réalité, le signal de parole est seulement quasi-stationnaire (stationnaire sur la durée d'une trame) et par conséquent le filtre de Wiener est utilisé pour traiter des trames à court terme. Il y a donc une contradiction entre les hypothèses de départ qui justifient le calcul à long terme du RSB *a priori* et l'application du filtre de Wiener à court terme. En réalité, seule l'approche MV intégrant l'aspect "deux-états" (*cf.* partie 2.4.2.1) utilise réellement le RSB *a priori* sur une base à long terme pour corriger l'approche MV classique. Ainsi, le comportement de l'estimateur du RSB *a priori* doit être considéré localement, *i.e.* à court terme, car c'est ainsi que cette quantité est utilisée en pratique. On peut même aller plus loin en arguant que le RSB *a priori*, de la façon dont il est utilisé dans les filtres de réduction de bruit, ne représente en fait qu'un artifice permettant de supprimer le bruit musical par le biais du lissage de son estimateur. D'ailleurs, en pratique, quel que soit l'estimateur utilisé pour le RSB *a priori*, cette quantité est toujours peu lissée de façon à ne pas trop dégrader le signal de parole dont les transitoires sont très importants pour la qualité subjective. Il faut donc trouver le meilleur compromis entre suppression du bruit musical et distorsion de la parole.

Dans la mesure où le signal de parole est quasi-stationnaire et où le filtre est utilisé à court terme, la meilleure estimation du RSB possible ne peut dépendre que des signaux considérés sur la trame traitée. Ceci est aussi valable pour le signal de bruit, bien qu'il puisse généralement être considéré comme stationnaire (*cf.* partie 3.3). Idéalement, la connaissance du signal propre et du bruit permet de calculer ce que l'on appellera le RSB *a posteriori* local et le RSB *a priori* local. Ces quantités sont définies comme suit

$$RSB_{post}^{local}(p,k) = \frac{|X(p,k)|^2}{|B(p,k)|^2}, \tag{3.6}$$

et

$$RSB_{prio}^{local}(p,k) = \frac{|S(p,k)|^2}{|B(p,k)|^2}. \tag{3.7}$$

Ces quantités locales constituent donc les buts respectifs à atteindre ou autrement dit les bornes maximales de qualité pour les estimateurs des RSB *a posteriori* et *a priori*.

3.2.4 Outil d'analyse pour les estimateurs du RSB

Les performances des techniques de réduction de bruit dépendent de la fonction de gain choisie, cependant, elles sont avant tout gouvernées par l'efficacité des estimateurs du RSB (*cf.* partie 3.2.2). L'analyse de ces estimateurs n'est pas aisée, par exemple, l'estimateur decision-directed du RSB *a priori* (*cf.* partie 3.2.5) réalise un barycentre pondéré entre deux quantités ce qui rend son analyse difficile. Pour pallier ce problème, nous proposons d'utiliser une approche inspirée de celle exposée dans [Renevey 2001]. Le principe consiste à analyser les nuages de points définis par le

couple (RSB_{post},RSB_{prio}), le RSB *a priori* étant représenté en fonction du RSB *a posteriori*. Plutôt que d'analyser ces deux types de RSB séparément, nous montrons qu'il est préférable d'analyser le couple (RSB_{post},RSB_{prio}) étant donné les relations qui existent entre ces deux quantités. En effet, le bruit étant supposé additif (*cf.* partie 2.1), l'amplitude du signal bruité peut s'écrire ainsi :

$$|X(p,k)| = \sqrt{|S(p,k)|^2 + |B(p,k)|^2 + 2|S(p,k)||B(p,k)|\cos\alpha(p,k)} \tag{3.8}$$

où $\alpha(p,k)$ représente la différence de phase entre $S(p,k)$ et $B(p,k)$. En remplaçant $|X(p,k)|$ dans l'équation (3.6) par son expression (3.8) et en utilisant (3.7), on peut écrire la relation suivante :

$$RSB_{post}^{local}(p,k) = 1 + RSB_{prio}^{local}(p,k) + 2\sqrt{RSB_{prio}^{local}(p,k)}\cos\alpha(p,k) \tag{3.9}$$

qui exprime le lien entre les expressions locales des RSB *a posteriori* et *a priori*.

Les techniques de réduction de bruit font généralement l'hypothèse que le signal utile et le bruit sont indépendants, ce qui en terme de RSB revient à :

$$\mathrm{E}[RSB_{post}(p,k)] = 1 + RSB_{prio}(p,k). \tag{3.10}$$

Cette relation, valide à long terme, permet de supprimer l'influence du terme de phase qui apparaît dans l'équation (3.9). Le bruit et le signal de parole sont donc supposés toujours s'ajouter en quadrature (*i.e.* $\alpha(p,k) = \frac{\pi}{2}$) ce qui n'est pas le cas à court terme où cette phase joue un rôle important dans l'estimation du RSB *a priori*. **En effet, même en connaissant le module du bruit en plus de celui du signal bruité on ne peut pas en déduire directement celui du signal de parole** car il faudrait connaître la phase $\alpha(p,k)$ (*cf.* partie 3.4). La borne maximale de qualité pour le RSB *a priori* définie par l'équation (3.7) ne peut donc théoriquement pas être atteinte (excepté dans le cas où la phase réelle $\alpha(p,k)$ serait effectivement égale à $\frac{\pi}{2}$). De plus, on peut noter qu'en pratique les RSB *a posteriori* et *a priori* sont estimés en utilisant une estimation à long terme de la DSP du bruit qui ne prend donc pas en compte les évolutions fines du bruit entre les trames successives (*cf.* partie 3.3). Les estimateurs du RSB ne peuvent donc en aucun cas atteindre les bornes maximales de qualité définies par les équations (3.6) et (3.7). La quantité $|X(p,k)|^2$ étant observée, l'amélioration de l'estimateur du RSB *a posteriori* passe par celle de l'estimateur de la DSP du bruit. Par contre, il reste une marge d'amélioration assez importante pour l'estimateur du RSB *a priori*, indépendamment de la DSP du bruit, comme on le verra dans le chapitre suivant.

Dans la suite, les nuages de points seront, par défaut, obtenus pour 100 trames du signal bruité présenté dans la partie 3.2.1 dont 50 de bruit seul et 50 de parole bruitée.

La relation exprimée par l'équation (3.9) est illustrée par la figure 3.6. Le RSB *a priori* y est représenté en fonction du RSB *a posteriori* dans le cas idéal où les modules du signal de parole et du bruit sont connus. Le nuage de points qui en résulte est délimité par deux courbes. La courbe en trait plein correspond à l'équation (3.9) dans le cas où $\alpha(p,k) = 0$ (*i.e.* le signal et le bruit s'ajoutent en phase) et la courbe en pointillé à celui où $\alpha(p,k) = \pi$ (*i.e.* le signal et le bruit s'ajoutent en opposition de phase). Ces deux courbes définissent une zone dans laquelle la distribution des points dépend de la différence de phase $\alpha(p,k)$. Le nuage de points de la figure 3.6 correspond au cas idéal où les estimateurs sont parfaits mais est donc très éloigné de la pratique. Nous allons donc nous rapprocher de la réalité en remplaçant la connaissance du module carré du bruit $|B(p,k)|^2$ dans les équations (3.6)

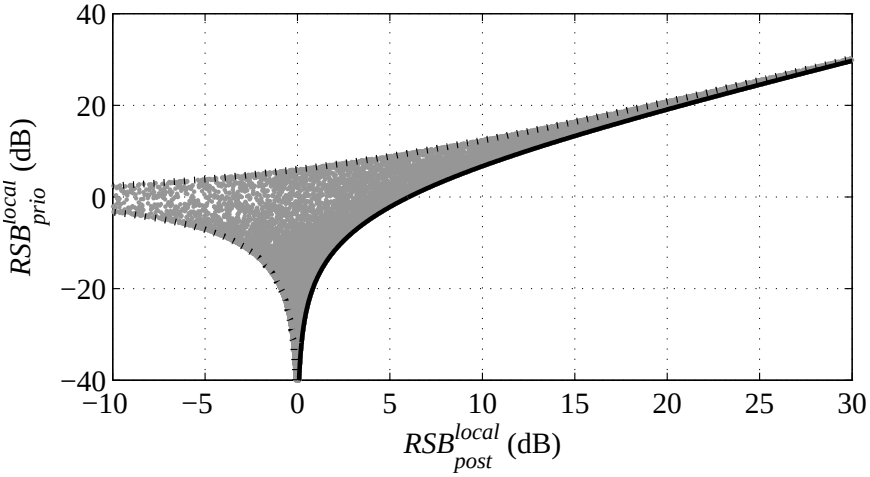

FIG. 3.6 – *RSB_{prio}^{local} en fonction du RSB_{post}^{local} dans le cas où les modules du signal propre et du bruit sont connus. Les deux lignes matérialisent l'équation (3.9) dans le cas où $\alpha(p,k) = 0$ (trait plein) et dans le cas où $\alpha(p,k) = \pi$ (pointillé).*

et (3.7) par l'estimée de sa DSP $\hat{\gamma}_b(k)$. La technique d'estimation retenue ainsi que les paramètres associés sont détaillés dans la partie 6.3. La figure 3.7 représente ce cas et on peut noter que, par rapport à la figure 3.6, le fait d'estimer la DSP du bruit provoque une dispersion assez importante du nuage de points pour les faibles valeurs du RSB. Ceci est cohérent avec le fait que l'estimation du

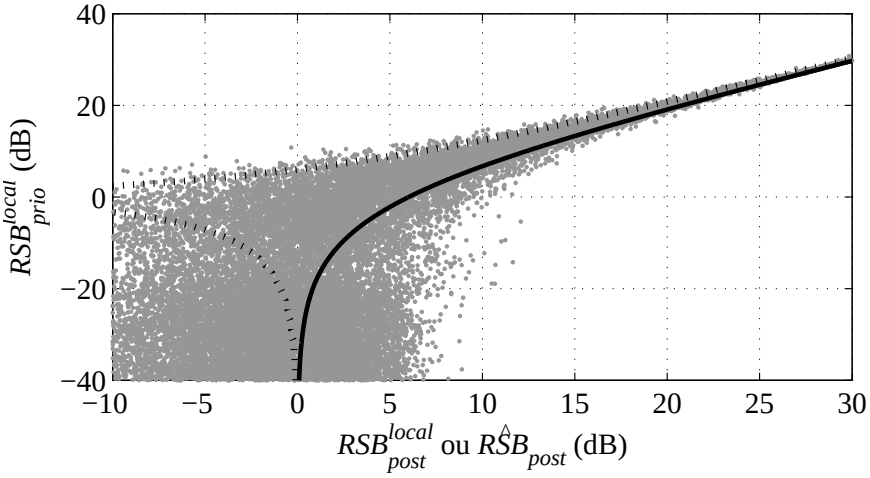

FIG. 3.7 – *RSB_{prio}^{local} en fonction du RSB_{post}^{local} dans le cas où le module du signal propre est connu et la DSP du bruit est estimée. Les deux lignes matérialisent l'équation (3.9) dans le cas où $\alpha(p,k) = 0$ (trait plein) et dans le cas où $\alpha(p,k) = \pi$ (pointillé).*

bruit est d'autant plus délicate que le RSB est faible. Ainsi, commettre une erreur d'estimation du bruit quand l'énergie de la composante de parole est importante n'est pas très dommageable, par contre si la composante de parole est faible cela peut avoir pour effet de la dégrader voire même de la détruire. Évidemment, l'utilisation d'une estimée du bruit à la place de la valeur réelle entraîne une diminution

de la qualité de la parole restaurée. Cette dégradation dépend de la technique d'estimation du bruit ainsi que du RSB global : plus celui-ci est faible, plus les erreurs d'estimations sont importantes. Pour une estimée fixée de la DSP du bruit, le cas qui vient d'être décrit donne donc une borne maximale de qualité réaliste pour les estimateurs de RSB (le module du signal de parole reste supposé connu). Cette borne bien qu'encore une fois irréalisable en pratique va nous permettre de situer les performances des estimateurs existants et de ceux que nous proposons dans les parties 4.3 et 4.4. On peut d'ailleurs noter que, dans tous les cas, l'approche utilisée pour estimer le RSB *a posteriori* restera la même dans la mesure où il n'y a pas d'alternative. La comparaison entre deux nuages de points n'a donc finalement pour but que de faciliter l'analyse des différents estimateurs du RSB *a priori*.

3.2.5 Estimateur decision-directed

3.2.5.1 Principe de l'approche decision-directed

L'estimation de la DSP du bruit $\hat{\gamma}_b(k)$ est un préliminaire indispensable au calcul des RSB *a posteriori* et *a priori*. La partie 2.5 présente différentes approches permettant de calculer cette DSP. Les RSB sont alors obtenus ainsi :

$$\hat{RSB}_{post}(p,k) = \frac{|X(p,k)|^2}{\hat{\gamma}_b(k)}, \tag{3.11}$$

et

$$\hat{RSB}^{DD}_{prio}(p,k) = \beta \frac{|\hat{S}_{DD}(p-1,k)|^2}{\hat{\gamma}_b(k)} + (1-\beta)\max\left(\hat{RSB}_{post}(p,k) - 1, 0\right), \tag{3.12}$$

où $\hat{S}_{DD}(p-1,k)$ est le spectre du signal de parole estimé à la trame précédente. Cet estimateur du RSB *a priori*, nommé decision-directed (DD) ce qui signifie dirigé par la décision, a été proposé dans [Ephraïm 1984] et son comportement est contrôlé par le paramètre β (toujours proche de 1 et typiquement égal à 0,98) [Cappé 1994]. Finalement, de façon générale, le gain spectral est une fonction qui dépend du RSB *a priori* et éventuellement du RSB *a posteriori* :

$$G_{DD}(p,k) = g\left(\hat{RSB}^{DD}_{prio}(p,k), \hat{RSB}_{post}(p,k)\right). \tag{3.13}$$

La fonction de gain $g(.)$ peut être l'une des fonctions décrites dans la partie 2.4. Le spectre du signal de parole restauré est ensuite obtenu par :

$$\hat{S}_{DD}(p,k) = G_{DD}(p,k)X(p,k). \tag{3.14}$$

Dans la suite, par défaut, la fonction de gain retenue correspondra au filtre de Wiener ce qui conduit à :

$$G_{DD}(p,k) = \frac{\hat{RSB}^{DD}_{prio}(p,k)}{1+\hat{RSB}^{DD}_{prio}(p,k)}. \tag{3.15}$$

Le principe de l'approche DD ainsi définie par les équations (3.11) à (3.13) est résumé par le schéma de la figure 3.8.

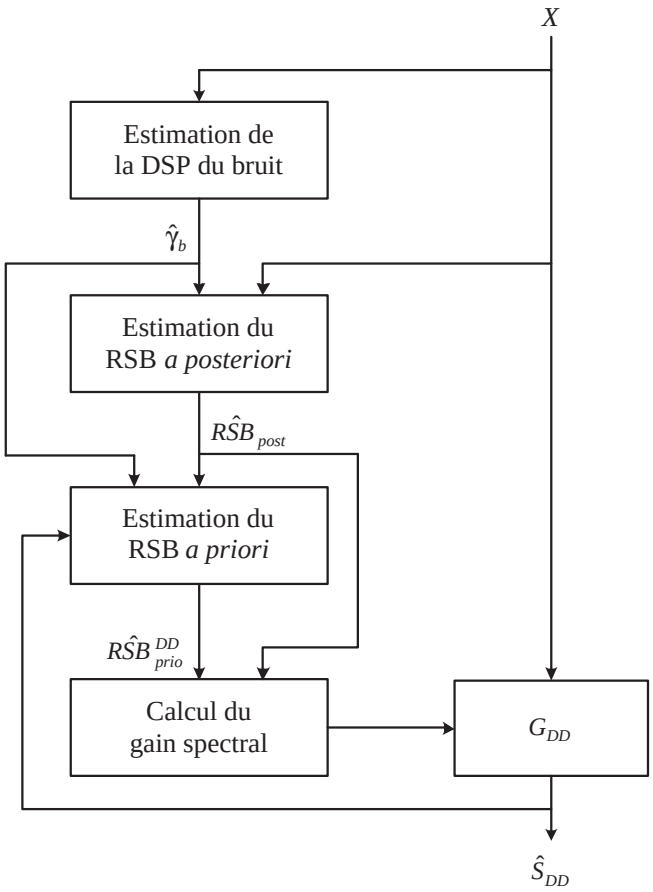

FIG. 3.8 – *Schéma du principe général de l'approche DD.*

3.2.5.2 Analyse de l'algorithme decision-directed

La figure 3.5 présentée dans la partie 3.2.2 permet d'illustrer le comportement de l'approche DD. Comme exposé dans la partie 2.4.1.3, le filtre de Wiener peut s'exprimer en fonction soit du RSB *a priori* (3.12), soit du RSB *a posteriori* (3.11), il s'agit alors du filtre pseudo-Wiener. Ce qui est évident au premier abord c'est que l'utilisation du RSB *a priori* (*cf.* figure 3.5.(b)) permet de limiter de façon conséquente le niveau de bruit musical résultant de l'utilisation du RSB *a posteriori* (*cf.* figure 3.5.(a)). Une analyse de l'estimateur DD, basée sur le RSB *instantané* défini par l'équation (2.7), a été réalisée dans [Cappé 1994] où les deux effets suivants sont mis en avant :

– Quand le RSB *instantané* est beaucoup plus grand que 0dB, l'estimateur $R\hat{S}B^{DD}_{prio}(p,k)$ correspond à une version retardée d'une trame du RSB *instantané*. En effet, lorsque le RSB est important l'atténuation apportée est négligeable et donc on peut écrire que

$$\hat{S}_{DD}(p-1,k) \approx X(p-1,k). \tag{3.16}$$

L'équation (3.12) peut donc se réécrire ainsi :

$$R\hat{S}B^{DD}_{prio}(p,k) \approx \beta R\hat{S}B_{post}(p-1,k) + (1-\beta)\max\left(R\hat{S}B_{post}(p,k) - 1,0\right). \tag{3.17}$$

De plus, comme le paramètre β est généralement très proche de 1 et étant donné que le RSB *instantané* est très supérieur à 1, l'approximation suivante peut être réalisée :

$$\hat{RSB}^{DD}_{prio}(p,k) \approx \hat{RSB}_{inst}(p-1,k). \tag{3.18}$$

– Quand le RSB *instantané* est plus petit ou proche de 0dB, l'estimateur $\hat{RSB}^{DD}_{prio}(p,k)$ correspond à une version retardée et très lissée du RSB *instantané*. Dans ce cas, la variance du RSB *a priori* est très réduite par rapport à celle du RSB *instantané*. Un filtre exprimé (uniquement) en fonction du RSB *a priori* introduira donc peu de bruit musical.

Ce comportement est illustré par la figure 3.9 où l'évolution des RSB *instantané* et *a priori* est représentée au cours des trames pour la bande de fréquence centrée sur 467Hz (qui correspond à la 30ème bande avec les paramètres choisis pour l'analyse). Précisons que l'estimation de la DSP du bruit est réalisée en utilisant l'approche avec DAV décrite dans la partie 2.5.2. Les 20 premières et les 17 dernières trames contiennent uniquement du bruit et les 19 au centre contiennent de la parole bruitée (RSB global de 12dB). On retrouve bien l'effet de lissage du RSB *a priori* pour les faibles RSB ainsi

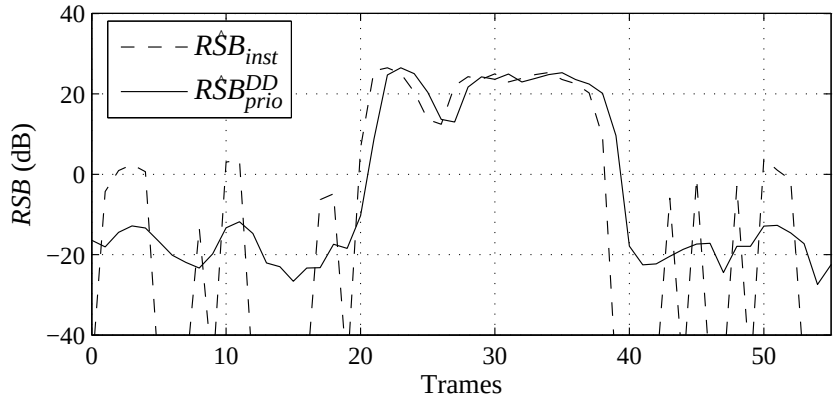

FIG. 3.9 – *Evolution temporelle du RSB pour l'approche DD (pour la bande de fréquence centrée sur* 467*Hz) avec seuillage du filtre de Wiener. En pointillé : RSB* instantané *; en trait plein : RSB* a priori *de l'approche DD.*

que le retard (de la durée d'une trame) introduit par son estimateur lorsque le RSB est important. On peut noter que l'estimateur DD n'introduit pas de lissage pour les composantes de parole (fort RSB) s'écartant ainsi de la définition du RSB *a priori* donnée dans l'équation (2.6). Cette propriété est très importante car elle permet d'éviter les distorsions de la parole. **Cette dualité de comportement en fait un estimateur très intéressant qui est largement utilisé dans les techniques de réduction de bruit. Cependant, le fait que l'estimateur DD introduise un retard est un inconvénient car le RSB estimé est biaisé et ne caractérise pas exactement la trame courante.** On peut même dire qu'il correspond mieux à la trame précédente (idée qui sera développée dans la partie 4.3). Ce biais limite donc les performances du système de réduction de bruit et engendre un effet de réverbération pour les signaux restaurés. De plus, ce biais est particulièrement gênant lors des attaques et des extinctions de parole car c'est durant ces périodes qu'il introduit les erreurs les plus importantes comme on peut le voir sur la figure 3.9. Lorsque le filtre de Wiener est utilisé avec l'approche DD, ce qui est

le cas dans l'exemple de la figure 3.9, le lissage (pour les faibles RSB) est assuré par un seuillage du gain spectral :

$$G_{DD}(p,k) = \max(G_{DD}(p,k), G_{min}) \tag{3.19}$$

où G_{min} représente le seuil. Dans l'exemple de la figure 3.9 ce seuil est de -20dB ($G_{min} = 0{,}1$). De façon équivalente, il est possible de réaliser un seuillage du RSB *a priori*. À l'origine, le filtre

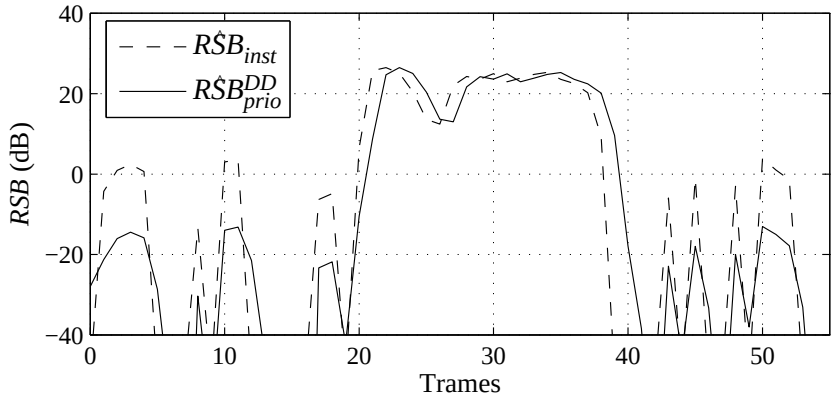

FIG. 3.10 – *Evolution temporelle du RSB pour l'approche DD (pour la bande de fréquence centrée sur* 467*Hz) sans seuillage du filtre de Wiener. En pointillé : RSB* instantané *; en trait plein : RSB* a priori *de l'approche DD.*

MMSE STSA est utilisé avec l'approche DD [Ephraïm 1984] ; le seuillage du RSB *a priori* est alors assuré de par le comportement particulier de ce filtre qui a tendance à remonter le plancher spectral autour des pics fréquentiels isolés de façon à masquer le bruit musical (*cf.* partie 2.4.2.2). Comme le montre la figure 3.10, si le seuillage est supprimé ($G_{min} = 0$) alors l'effet de lissage disparaît, toutefois, les pics localisés du RSB *instantané*, qui sont à l'origine du bruit musical, sont fortement atténués. **L'estimateur DD permet donc à lui seul de limiter le phénomène de bruit musical sans pour autant introduire un plancher spectral.** De ce point de vue, l'application avec le filtre MMSE STSA constitue finalement un cas particulier qui génère un phénomène de seuillage interprétable comme un lissage. Le fait d'utiliser le filtre de Wiener permet de rester plus général et c'est pourquoi il est utilisé comme référence. De plus, l'utilisation du filtre de Wiener permet par un simple seuillage d'obtenir un bruit résiduel qui respecte la coloration du bruit original, contrairement au filtre MMSE STSA. De cette façon il est aussi possible de régler simplement le niveau de bruit résiduel désiré.

Comme support à l'analyse de l'estimateur DD, l'outil d'analyse présenté dans la partie 3.2.4 est mis à profit. Ainsi, les couples de RSB $(RSB_{post}, RSB^{DD}_{prio})$ sont représentés dans la figure 3.11. Les RSB *a posteriori* et *a priori* sont estimés respectivement en utilisant les équations (3.11) et (3.12). L'analyse de ce nuage de points doit être réalisée conjointement à celle de la figure 3.7 qui fait office de référence. On peut d'ailleurs remarquer que dans cette figure le RSB^{local}_{post} correspond au $\hat{RSB}_{post}$ étant donné que la DSP du bruit est estimée. Ainsi, par rapport à cette référence, dans la figure 3.11 une grande partie des RSB *a priori* est sous-estimée (environ 60% dans ce cas) ce qui illustre parfaitement l'effet du biais de l'estimateur DD. Si l'on considère le cas où une composante spectrale de parole apparaît subitement à la trame p et en supposant que le RSB *a priori* est nul à la trame $p-1$ ($\hat{S}_{DD}(p-1,k)$ est par conséquent nul également) alors pour la trame courante on peut

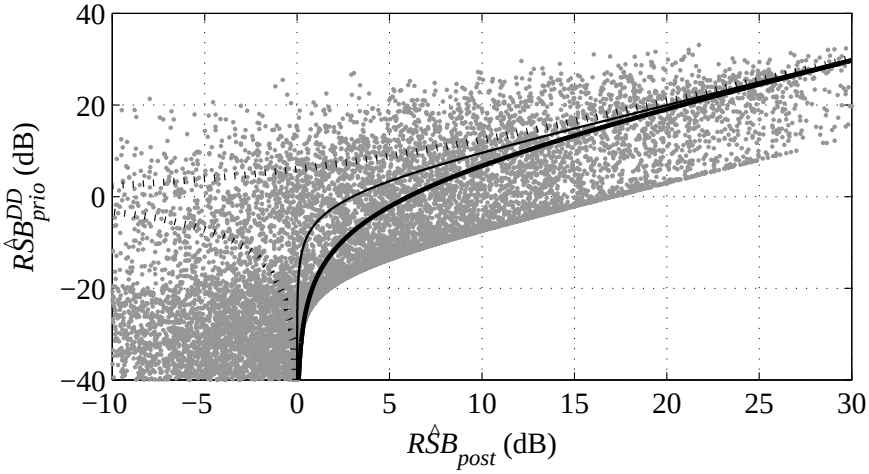

FIG. 3.11 – *$\hat{RSB}^{DD}_{prio}$ en fonction du $\hat{RSB}_{post}$ pour l'approche DD. Les trois lignes matérialisent l'équation (3.9) dans le cas où $\alpha(p,k) = 0$ (trait fort), $\alpha(p,k) = \pi$ (pointillé) et $\alpha(p,k) = \frac{\pi}{2}$ (trait fin).*

écrire :

$$\hat{RSB}^{DD}_{prio}(p,k) = (1-\beta)\max\left(\hat{RSB}_{post}(p,k) - 1,0\right). \tag{3.20}$$

Autrement dit, le RSB *a priori* estimé est une version atténuée par le facteur $(1-\beta)$ du RSB *instantané*. Si l'on choisit la valeur typique de $\beta = 0{,}98$ alors l'atténuation est d'environ 17dB. Nous allons voir que l'effet de la relation (3.20) est clairement visible sur la figure 3.11. En effet, si $\alpha(p,k) = \frac{\pi}{2}$, alors l'équation (3.9) devient

$$RSB^{local}_{prio}(p,k) = RSB^{local}_{post}(p,k) - 1 = RSB^{local}_{inst}(p,k). \tag{3.21}$$

Cette relation est représentée sur la figure 3.11 par le trait fin. L'atténuation par le facteur $(1-\beta)$ dans l'équation (3.20) est clairement matérialisée par une concentration importante de points aux alentours d'une version décalée d'environ -17dB de cette courbe en trait fin (limite inférieure du nuage de points lorsque $0 < \hat{RSB}_{post}(p,k) < 30$dB). Ce décalage correspond au biais maximum qui est introduit lors des attaques de parole ou plus généralement quand le niveau d'une composante spectrale croît rapidement. On peut noter que l'on peut réduire le niveau de bruit musical en augmentant la valeur de β mais en contrepartie le biais de l'estimateur du RSB *a priori* augmente et la parole subit donc davantage de dégradations.

Parallèlement à la sous-estimation du RSB *a priori*, on peut aussi remarquer sur la figure 3.11 que certaines occurrences du RSB *a priori* sont surestimées. Ce cas apparaît quand une composante de parole disparaît subitement, *i.e.* $\max\left(\hat{RSB}_{post}(p,k) - 1,0\right) = 0$, ce qui conduit au résultat suivant :

$$\hat{RSB}^{DD}_{prio}(p,k) = \beta\frac{|\hat{S}_{DD}(p-1,k)|^2}{\hat{\gamma}_b(k)}. \tag{3.22}$$

Le RSB *a priori* obtenu dépend du spectre de parole estimé à la trame précédente alors qu'une valeur nulle serait la meilleure estimation. Ce cas apparaît plus généralement dès que le niveau d'une composante spectrale de parole diminue rapidement. L'effet de réverbération caractéristique de l'approche DD est ainsi dû à la combinaison de la sous-estimation et de la surestimation des RSB *a priori* suivant les fluctuations du niveau des composantes spectrales de parole.

3.2.6 Comparaison des RSB *a posteriori* et *a priori*

L'approche DD fait appel au calcul de deux quantités, le RSB *a posteriori* et le RSB *a priori*. Certaines fonctions de gain ne s'expriment qu'en fonction de l'une ou l'autre de ces deux quantités et d'autres en fonction des deux en même temps (*cf.* partie 2.4). Ces dernières réalisent un compromis entre le comportement de ces deux RSB. Nous allons donc nous intéresser aux différences qui existent entre ces deux types de RSB dans le but de quantifier les défauts et avantages de chacun. Lorsqu'un filtre de réduction de bruit s'exprime uniquement en fonction du RSB *a posteriori*, le signal restauré contient du bruit musical et est donc jugé de mauvaise qualité. C'est pour cette raison que le RSB *a priori* de l'approche DD lui est très souvent préféré. En effet, cet estimateur permet de réduire de façon importante le niveau de bruit musical (sans toutefois le supprimer complètement). **Cependant, si l'on fait abstraction du bruit musical, l'utilisation du RSB *a posteriori* donne de meilleurs résultats que celle du RSB *a priori* pour les composantes de parole.** En effet, comme nous l'avons vu dans la partie précédente, l'estimateur du RSB *a priori* est biaisé ce qui diminue ses performances pendant l'activité vocale et conduit à l'effet de réverbération caractéristique de l'approche DD.

Pour mesurer les performances de ces deux estimateurs de RSB, nous proposons de comparer les valeurs estimées aux valeurs réelles, ces dernières étant calculées en connaissant le bruit et le signal utile. La figure 3.12 est composée de deux nuages de points, chacun est obtenu en représentant le RSB estimé (*cf.* équations (3.11) et (3.12)) en fonction du RSB réel (ou autrement dit local, *cf.* équations (3.6) et (3.7)). Ces nuages de points sont obtenus pour 50 trames d'activité vocale dans le but d'analyser le comportement des estimateurs uniquement par rapport aux composantes de parole. Dans les deux cas, le trait fort représente un estimateur idéal, *i.e.* $\hat{RSB} = RSB^{local}$ qui sert de référence pour évaluer les performances des estimateurs. Il apparaît clairement que le nuage de points correspondant au RSB *a posteriori* (*cf.* figure 3.12.(a)) est moins dispersé autour de l'estimateur idéal que le nuage de points correspondant au RSB *a priori* (*cf.* figure 3.12.(b)). La dispersion observée dans ces deux figures peut être caractérisée par le coefficient de corrélation de chaque nuage de points qui peut se calculer ainsi :

$$\rho = \frac{\mathrm{E}[(\hat{RSB} - \mathrm{E}[\hat{RSB}])(RSB^{local} - \mathrm{E}[RSB^{local}])]}{\sqrt{\mathrm{E}[(\hat{RSB} - \mathrm{E}[\hat{RSB}])^2]\mathrm{E}[(RSB^{local} - \mathrm{E}[RSB^{local}])^2]}}. \tag{3.23}$$

Dans le cas particulier de la figure 3.12, on obtient $\rho_{post} = 0{,}79$ et $\rho_{prio} = 0{,}23$ ce qui est cohérent avec les dispersions constatées dans les figures 3.12.(a) et (b), un plus petit coefficient de corrélation traduisant une plus grande dispersion des points. En généralisant à d'autres conditions de bruit et de RSB, la relation suivante se dégage : $\rho_{prio} \approx \rho_{post} - 0{,}5$.

Dans les figures 3.12.(a) et (b), le trait fin représente la moyenne conditionnelle du RSB estimé et est obtenue ainsi :

$$\mathrm{E}[\hat{RSB} \mid RSB^{local}] = \int \hat{rsb}\ \mathrm{p}(\hat{rsb} \mid RSB^{local})\ \mathrm{d}\hat{rsb} \tag{3.24}$$

où $\mathrm{p}(.)$ représente la densité de probabilité (histogramme obtenu expérimentalement) du RSB estimé sachant le RSB réel. La moyenne conditionnelle du RSB estimé est plus proche de l'estimateur idéal dans le cas du RSB *a posteriori*. En effet, elle est sous-estimée pour les RSB importants alors que celle du RSB *a priori* est largement sous-estimée pour les valeurs de RSB supérieures à -17dB. Paradoxalement, étant donné que la dispersion des points est importante pour le RSB *a priori*, des

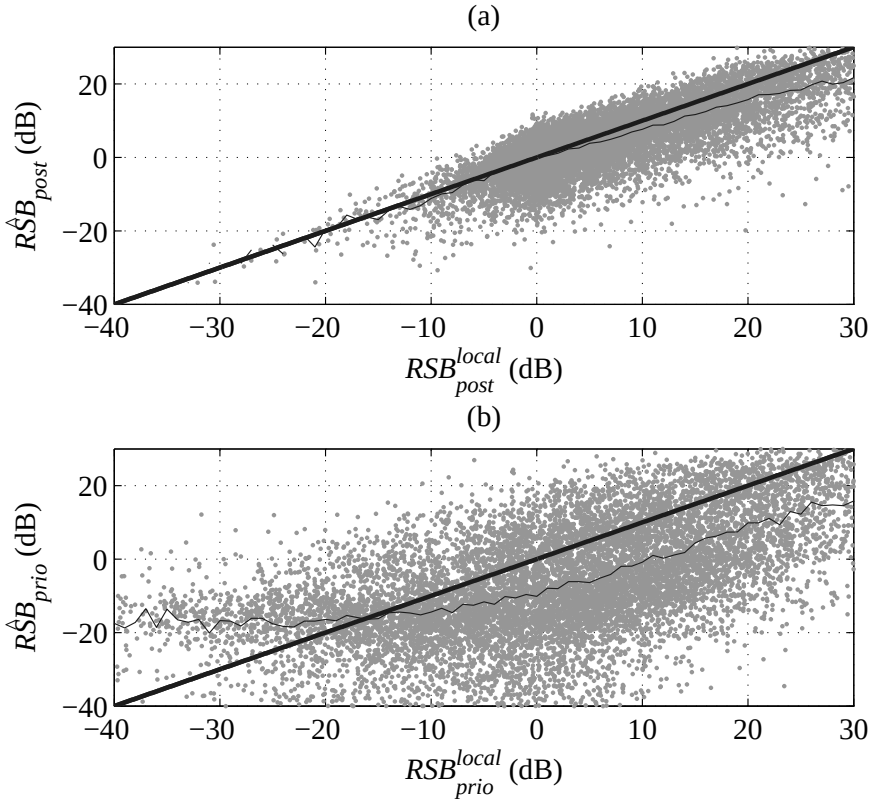

FIG. 3.12 – *RSB estimé représenté en fonction du RSB réel (*i.e. *RSB local) dans le cas du RSB* a posteriori *(a) et du RSB* a priori *(b). Le trait fort représente un estimateur idéal et le trait fin la moyenne du RSB estimé en fonction du RSB réel.*

cas de surestimation se produisent également. De plus, le RSB *a priori* est surestimé pour des valeurs de RSB inférieures à -17dB. **Finalement, ces résultats confirment que l'estimateur du RSB *a posteriori* est meilleur que celui du RSB *a priori* (de l'approche DD) pour les composantes de parole.**

3.2.7 Convergence des estimateurs du RSB

À la lumière de la partie 3.2.3, le but à atteindre pour le comportement d'un estimateur du RSB est de se rapprocher de celui du RSB *a posteriori* local (*cf.* équation (3.6)) ou, ce qui revient sensiblement au même, de celui du RSB *a priori* local (*cf.* équation (3.7)). Dans la mesure où l'estimation de la DSP du bruit est réalisée à long terme (*cf.* partie 3.3) et où le terme de phase $\alpha(p,k)$ apparaissant dans l'équation (3.9) n'est pas pris en compte, en pratique ce but ne peut pas être atteint. L'analyse proposée dans la partie 3.2.6 permet cependant de dégager l'idée qu'**il serait intéressant de converger vers un estimateur unique du RSB combinant les avantages respectifs du RSB *a posteriori* et du RSB *a priori* de l'approche DD.** Les caractéristiques d'un tel estimateur sont les suivantes :

– Pour les composantes de signal de parole, il faut conserver l'estimateur du RSB *a posteriori*

donné par l'équation (3.11) car son estimation est non biaisée. Son utilisation dans un filtre de réduction de bruit permet donc d'éviter l'effet de réverbération caractéristique de l'estimateur DD du RSB *a priori*.

– Pour les composantes de bruit seul, il faut conserver l'estimateur du RSB *a priori* donné par l'équation (3.12) car le lissage qu'il introduit permet de limiter efficacement le bruit musical.

Un estimateur "idéal" (noté par l'exposant *id*) du RSB *a priori*, dans le sens où il se rapproche du RSB *a priori* local, peut donc se résumer ainsi :

$$\hat{RSB}^{id}_{prio}(p,k) = \begin{cases} \hat{RSB}_{post}(p,k) - 1 & \text{pour les composantes de parole bruitée,} \\ \hat{RSB}^{DD}_{prio}(p,k) & \text{pour les composantes de bruit seul.} \end{cases} \tag{3.25}$$

Le RSB *a posteriori* idéal est alors automatiquement défini par

$$\hat{RSB}^{id}_{post}(p,k) = \hat{RSB}^{id}_{prio}(p,k) + 1. \tag{3.26}$$

Ces deux quantités portent bien sûr les mêmes informations et sont interchangeables. En réalité, il s'agit seulement de deux façons d'exprimer un estimateur du RSB unique. On peut noter que pour les composantes de bruit seul, le RSB *a priori* estimé devrait théoriquement être nul. Ceci permettrait alors de supprimer complètement le bruit musical qui reste faiblement audible même en utilisant le RSB *a priori* de l'approche DD. Les parties 4.3 et 4.4 exposent le principe de deux approches permettant peu ou prou d'atteindre ce comportement.

On peut remarquer que, si l'on converge vers un estimateur unique, cela n'a plus aucun sens d'utiliser les techniques qui font intervenir les RSB *a posteriori* et *a priori* car l'intérêt de celles-ci repose justement sur les différences de comportement entre ces deux quantités. Par exemple, le fait d'utiliser la technique MMSE STSA n'a plus de sens car son comportement serait très proche de celui de la SSP (*cf.* partie 2.4.2.2) mais pour un coût beaucoup plus important en terme de calcul. Ce type de compromis n'a d'ailleurs plus lieu d'être puisque celui-ci est réalisé directement par l'estimateur du RSB en fonction de la nature de la composante de signal traitée (parole ou bruit).

3.3 Limitations liées à l'estimation de la DSP du bruit

La partie 2.5 fait apparaître deux classes d'estimateurs de la DSP du bruit, d'une part ceux qui utilisent une DAV et d'autre part ceux qui fonctionnent en continu, *i.e.* sans DAV. Les techniques à base de DAV figent l'estimée de la DSP du bruit pendant l'activité vocale et ont donc de bonnes performances lorsque le bruit est stationnaire, hypothèse de base des techniques de réduction de bruit. Cependant, le bruit est très souvent de nature non-stationnaire même si l'on peut généralement le considérer comme "plus stationnaire" que le signal de parole. Il faut alors être capable de suivre rapidement ses variations pour éviter les distorsions qui découlent d'une mauvaise estimation de la DSP du bruit. D'où l'intérêt des approches sans DAV qui permettent d'actualiser cette estimée aussi pendant l'activité vocale.

La figure 3.13 permet d'illustrer ces deux familles d'approches. La famille des techniques sans DAV sera représentée par l'approche MCRA [Cohen 2002a] présentée dans la partie 2.5.3.5 et qui

offre de bonnes performances. Les paramètres choisis sont les suivants : $\lambda_X = 0{,}8$, $\lambda_P = 0{,}2$, $\lambda_B = 0{,}95$, $\delta = 5$ et $L = 62$. Le taux de recouvrement entre les trames de 32ms est de 75%. Avec ces paramètres, on déduit de l'équation (2.64) que $\tilde{\lambda}_B \geq \lambda_B = 0{,}95$; la constante de temps équivalente sera donc toujours supérieure ou égale à 312ms fournissant ainsi, dans le meilleur des cas, un lissage de la DSP de bruit déjà relativement important. Pour permettre une comparaison équitable avec l'approche MCRA, la constante de temps de l'approche avec DAV [Scalart 1996a] (*cf.* partie 2.5.2), qui est fixe, sera également de 312ms. La DAV est choisie idéale, *i.e.* obtenue à partir du signal propre, dans le but de se déconnecter du problème de l'estimation robuste de la DAV. Ainsi, sur la figure 3.13

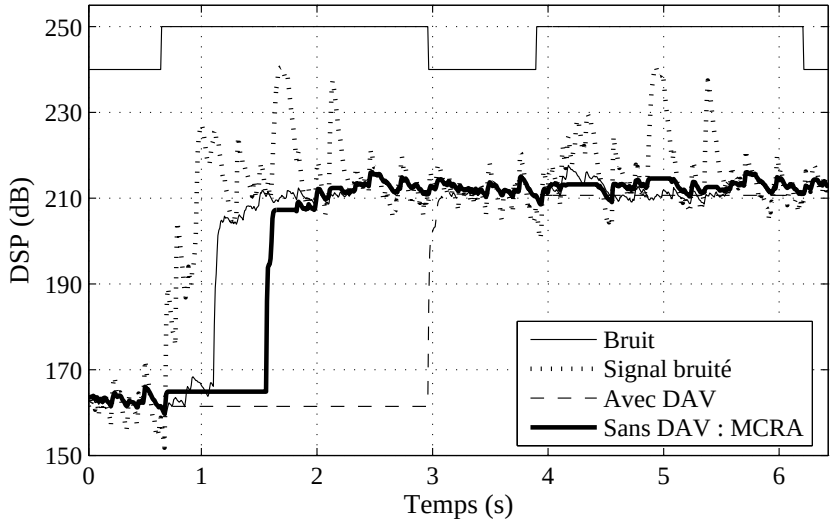

FIG. 3.13 – *DSP lissée du signal bruité (constante de temps de 72ms) en pointillé et du bruit de voiture (constante de temps de 312ms) en trait fin. DSP du bruit estimée par une approche avec DAV (tirets) et sans DAV (trait fort). Le niveau de bruit augmente de 21dB pendant l'activité vocale à l'instant 1,1s. Le RSB global passe ainsi de 21 à 0dB. Ces DSP sont toutes représentées pour la bande de fréquence centrée sur 498Hz. La DAV idéale a été rajoutée en haut de la figure. Le signal de parole est identique pour les deux périodes d'activité vocale (état haut).*

on peut observer que lors d'une brusque augmentation du niveau de bruit pendant l'activité vocale, l'estimateur MCRA permet d'actualiser la DSP du bruit au bout de seulement 500ms alors qu'en utilisant l'approche avec DAV il faut attendre la prochaine période de silence pour la remettre à jour (délai de près de 2s dans cet exemple). Dans ce cas, le bruit résiduel reste très présent tant que la DSP n'a pas été remise à jour ce qui est gênant. Bien sûr, il se peut également que le niveau du bruit chute rapidement pendant l'activité vocale, dans ce cas les techniques sans DAV permettent de diminuer rapidement l'estimée du bruit pour limiter les distorsions de la parole qui ne peuvent être évitées avec une DAV. On peut aussi noter que, pendant l'activité vocale, l'estimateur MCRA permet de suivre les évolutions lentes (la tendance) de la DSP du bruit alors que pour l'approche avec DAV la DSP du bruit reste figée durant toute l'activité vocale.

Cependant, un tel suivi lissé de la DSP du bruit se révèle insuffisant notamment dans un cas critique de bruit de foule (babble) qui n'est rien d'autre qu'un mélange de signaux de parole et qui en l'occurrence ne peut être considéré que comme quasi-stationnaire. En réalité, même un bruit sta-

tionnaire, dans le sens où sa DSP est constante (ou moins restrictif, varie lentement), peut présenter une variabilité à court terme très importante. Ceci est illustré par la figure 3.14.(a) où l'excursion par rapport à la DSP (long terme) est souvent très importante, *i.e.* jusqu'à plus de 30dB. Le signal bruité utilisé est celui présenté dans la partie 3.2.1 dont le bruit peut pourtant être considéré comme stationnaire (bruit de voiture). La figure 3.14.(b) permet de voir l'impact du lissage de la DSP du bruit sur

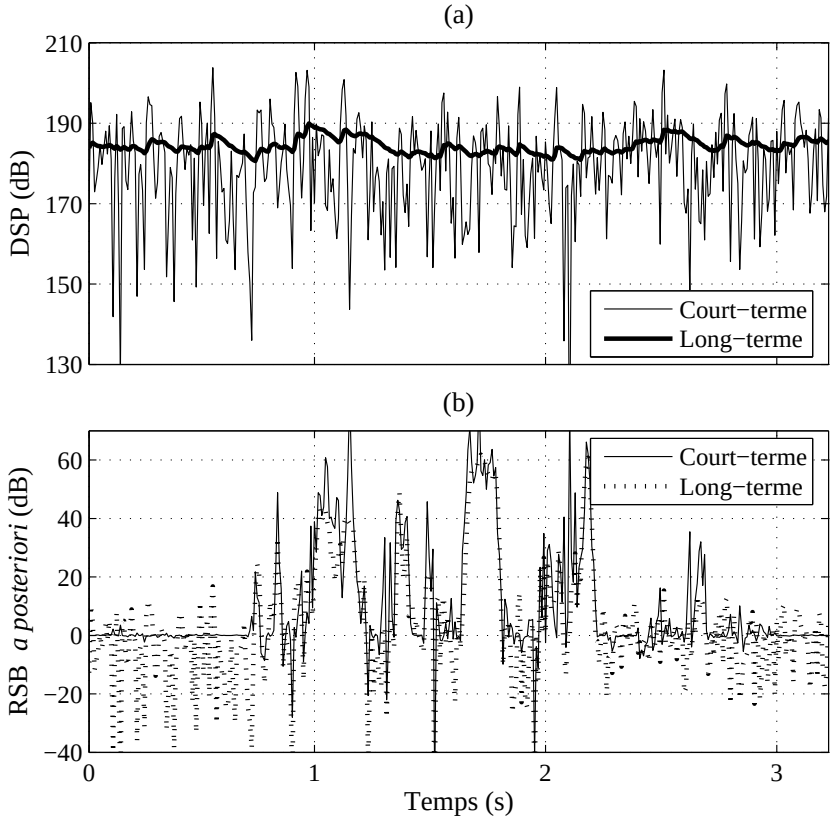

FIG. 3.14 – *(a) Module carré (court terme en trait fin) et DSP (long terme en trait fort) du bruit (constante de temps de 312ms) représentés pour la bande de fréquence centrée sur 498Hz. (b) RSB* a posteriori *calculé à partir de ces deux quantités dans le cas où le bruit est connu. Le RSB est représenté en trait fin pour la quantité court terme et en pointillé pour la quantité long terme. Le RSB global du signal bruité présenté dans la partie 3.2.1 est de 12dB.*

l'estimation du RSB *a posteriori* dans le cas idéal où le bruit est connu (mais lissé ou non). Ce lissage introduit de nombreuses erreurs d'estimation qui se traduisent (après filtrage) par le phénomène de bruit musical pendant l'inactivité vocale et par des distorsions du signal de parole pendant l'activité vocale. Tout ceci va dans le sens de ce qui a été présenté dans la partie 3.2.3. En effet, la borne maximale de qualité pour l'estimation du RSB (*a posteriori* ici) ne peut être atteinte qu'en utilisant des quantités court terme de par la nature non-stationnaire des signaux de parole et de bruit. **Le fait que l'estimée de la DSP du bruit soit lissée est donc une limitation en soi.** Ceci est d'ailleurs illustré dans la partie 3.2.4 par la figure 3.7. Cependant, il est très difficile de se départir de cette limitation car s'il était possible d'estimer avec précision le module carré du bruit alors le problème de la ré-

duction de bruit serait du même coup réglé (exception faite des problèmes liés à la phase exposés dans la partie 3.4). En pratique, la connaissance de la DSP du bruit est indispensable, et le signal de bruit sera donc dans tous les cas considéré comme stationnaire au moins à moyen terme (quelques trames). **En réalité, les techniques de réduction de bruit doivent se contenter d'une estimée que l'on peut qualifier de grossière de la DSP du bruit pour estimer le signal de parole. C'est ce qui leur permet d'être robustes mais introduit en contrepartie une limitation de leur performance.**

On verra dans la partie 4.5 qu'il est toutefois possible de contourner cette limitation. En effet, le fait d'utiliser une DSP lissée (auquel s'ajoute l'impact de la phase, *cf.* partie 3.4) entraîne des dégradations qui se traduisent généralement par la suppression de certaines composantes de signal. Nous démontrerons qu'il est possible d'en restaurer une partie en exploitant la structure harmonique de la parole.

3.4 Rôle de la phase dans la réduction de bruit

Dans les approches par atténuation spectrale à court terme la phase a reçu très peu d'attention. Nous proposons d'en rappeler la raison et d'aller un peu plus loin en montrant que, si la phase n'est effectivement pas importante pour la reconstruction du signal (la synthèse), elle a tout de même une influence non négligeable dans l'estimation des paramètres (notamment le RSB) utilisés pour calculer le gain spectral.

3.4.1 De l'importance de la phase

Les techniques de réduction de bruit par atténuation spectrale à court terme consistent à estimer uniquement le module du signal de parole. La phase du signal bruité est directement utilisée comme estimée de celle du signal de parole restauré. Cela se justifie d'un point de vue physiologique par le fait que l'oreille humaine est peu sensible à une modification raisonnable de la phase. La TF décompose un signal temporel en une somme de sinusoïdes, chacune déterminée par son amplitude, sa fréquence et sa phase. Celle-ci détermine la position temporelle relative des différentes composantes sinusoïdales. On comprend donc que **la phase à court terme a relativement peu d'importance pour l'intelligibilité alors qu'à long terme elle joue un rôle plus important que le module [Oppenheim 1979, Oppenheim 1981].** Ceci est confirmé dans [Wang 1982] où il est montré que, dans les applications de réduction de bruit, le fait d'améliorer l'estimation de la phase n'apporte pas de gain subjectif en qualité, excepté pour les très faibles RSB. Une analyse plus détaillée [Vary 1985] montre que, lorsque la résolution fréquentielle est suffisante, une modification de la phase n'entraîne pas de dégradation dans la mesure où le RSB local des composantes spectrales de parole est supérieur à 6dB.

Cependant, la phase est seulement considérée à la fin du traitement lorsqu'il faut redonner une relation de phase à chaque amplitude spectrale du signal restauré. Pourtant, le déphasage entre les composantes de parole et de bruit a un impact important sur l'estimation du module du signal propre. Pour illustrer ceci, prenons l'approche SSA (*cf.* partie 2.4.1.1) qui a l'avantage d'être aisément interprétable sur un diagramme de Fresnel. Considérons deux cas, dans le premier on suppose que le

bruit est constructif, *i.e.* $|X(p,k)| > |S(p,k)|$, et dans le second le bruit est supposé destructif, *i.e.* $|X(p,k)| < |S(p,k)|$. Dans les deux cas, le module à court terme du bruit $|B(p,k)|$ est supposé parfaitement connu. **Sous cette hypothèse idéale il n'est pourtant pas possible d'obtenir une réduction de bruit parfaite, *i.e.* $|\hat{S}(p,k)| = |S(p,k)|$, sans la connaissance de la différence de phase entre les composantes $S(p,k)$ et $B(p,k)$.**

En effet, dans le premier cas où le bruit est constructif, illustré dans la figure 3.15, bien que le module du bruit soit parfaitement connu, celui du signal de parole $|\hat{S}(p,k)|$ est sous-estimé. Ce

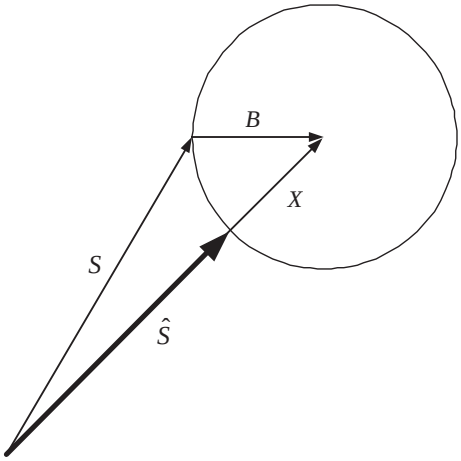

FIG. 3.15 – *Impact de la phase dans le plan de Fresnel pour l'approche SSA. Cas où le bruit est constructif.*

problème est symptomatique du fait que la phase n'est pas prise en compte. En réalité, dans la SSA, le signal de bruit est implicitement supposé s'ajouter en phase avec la parole, seul cas qui n'occasionne donc pas de dégradation. Dès lors que l'on s'écarte de cette hypothèse, des erreurs d'estimation sont commises. Ce phénomène se retrouve dans toutes les autres techniques de réduction de bruit sous des formes plus ou moins faciles à interpréter. Ainsi, par exemple, la SSP suppose implicitement que les signaux sont ajoutés en quadrature de phase. Dans les approches qui s'expriment en fonction du RSB, celui-ci répercute les erreurs liées à la non prise en compte de la phase étant donné que le RSB est une quantité purement énergétique.

Le deuxième cas, illustré par la figure 3.16, est encore plus défavorable dans la mesure où le bruit est maintenant destructif. La phase n'étant pas prise en compte, le module du bruit est soustrait à celui du signal bruité dégradant encore davantage la parole. Pour restaurer la composante fréquentielle du signal utile, il faudrait au contraire appliquer à $X(p,k)$ un gain supérieur à un ou au pire éviter une dégradation supplémentaire avec un gain égal à un. On imagine bien que, plus le RSB global du signal est défavorable, plus ce type d'erreur est fréquent et dommageable. **Ces deux exemples montrent donc que la phase joue un rôle important, bien que complètement négligé, dans l'estimation du module du signal de parole.** En prenant en compte les composantes fréquentielles significatives, *i.e.* $|S| > |B|$, il ressort que le bruit est aussi souvent destructif que constructif en raison de la phase du bruit uniformément distribuée. Comme aucune tendance ne ressort et qu'il n'existe pas d'approche pour estimer l'impact du bruit (constructif ou destructif), il n'est donc pas possible d'adapter la stratégie de

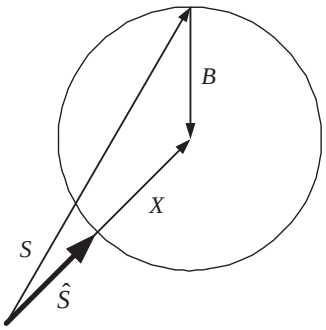

FIG. 3.16 – *Impact de la phase dans le plan de Fresnel pour l'approche SSA. Cas où le bruit est destructif.*

réduction de bruit par rapport à cela. La partie 4.5 met en avant une approche alternative qui permet de restaurer certaines composantes détruites ou dégradées à cause de telles erreurs d'estimation.

3.4.2 Information portée par la phase

Lorsqu'un modèle est nécessaire pour la phase, celle-ci est supposée distribuée uniformément sur $[0,2\pi[$ de façon à simplifier les calculs [Ephraïm 1984, McAulay 1980]. De plus elle est supposée ne pas porter d'information utile. Il est vrai qu'à court terme le module est prépondérant sur la phase [Oppenheim 1979, Oppenheim 1981]. Cependant dans [Hayes 1980] il est démontré théoriquement que, sous des hypothèses peu restrictives (respectées par le signal de parole), il est possible à partir uniquement de la phase d'un signal de reconstruire le signal temporel correspondant à un facteur d'échelle près. La faisabilité de cette approche est illustrée par la figure 3.17 où une trame de parole voisée est reconstruite uniquement à partir de l'information de phase en utilisant l'approche itérative proposée dans [Hayes 1980]. Au facteur d'échelle près, l'allure du signal temporel a été effectivement retrouvée. De la même façon, mais sous des conditions beaucoup plus restrictives (peu réalistes en pratique pour le signal de parole), il est possible de reconstruire le signal temporel à partir du module seul. La phase, même à court terme, contient donc beaucoup d'informations dont une partie est d'ailleurs redondante avec celles du module.

Chaque harmonique d'un signal de parole possède une allure en amplitude et en phase imposée par le lobe principal de la TFCT de la fenêtre d'analyse. Ce qui les différencie est donc leurs phases ainsi que leurs niveaux relatifs. Ceci est visible sur la figure 3.18 où chaque harmonique, dont le module est de forme caractéristique, possède une phase de pente constante. En effet, la dérivée de la phase en fréquence possède des paliers localisés au niveau des harmoniques dont la valeur de $-\pi/2$ correspond à la pente de la phase. Celle-ci est donc fortement liée à la structure du module. Ainsi, connaissant la position des harmoniques (*cf.* partie 5.6) il est tout à fait possible de redonner la bonne allure à la phase (droite de pente constante) lorsque celle-ci a été altérée. Cependant, les essais effectués dans le cadre de cette thèse ont montré que pour bien restaurer la phase il faudrait connaître, en plus de l'allure de la phase, le déphasage relatif entre les harmoniques. Malheureusement, cette information fondamentale, lorsqu'elle est détruite, est irrécupérable avec les approches actuelles.

Rappelons que l'intérêt de restaurer la phase ne se situe pas dans la reconstruction du signal estimé

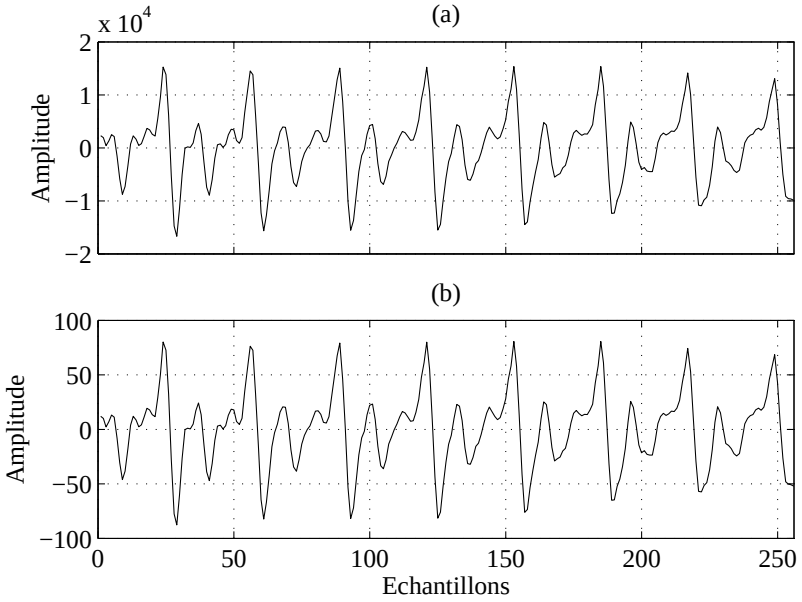

FIG. 3.17 – *(a) Trame de signal voisé de référence. (b) Trame reconstruite uniquement à partir de la phase de la trame de référence.*

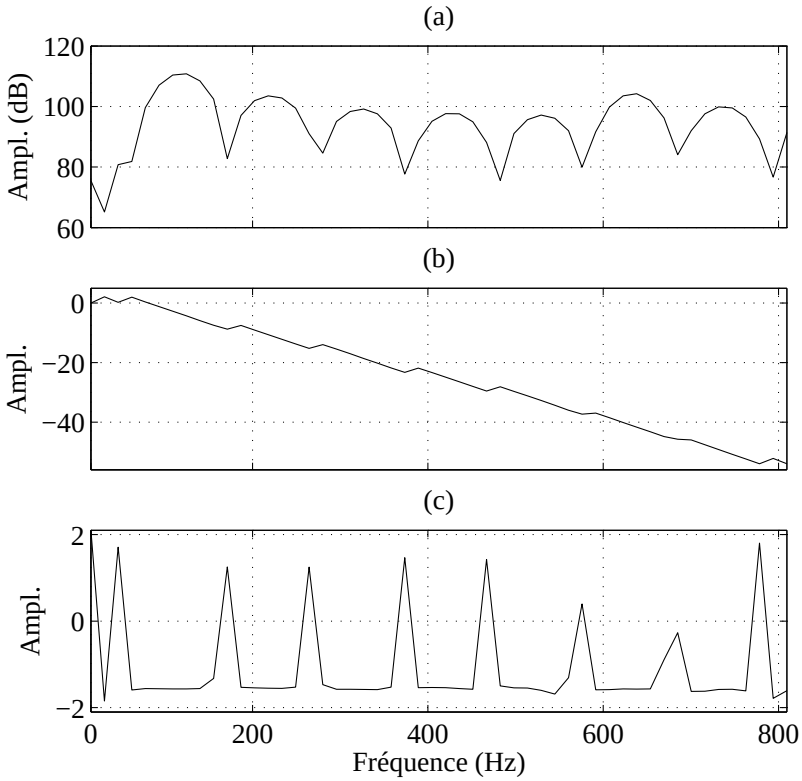

FIG. 3.18 – *(a) Zoom sur les basses fréquences du module d'une trame de signal voisé. (b) Phase déroulée correspondante. (c) Dérivée en fréquence de la phase.*

mais bien dans l'estimation des paramètres comme le RSB. Ainsi, à partir de l'équation (3.9), il est possible de relier la différence de phase $\alpha(p,k)$ entre le signal et le bruit avec les RSB *a posteriori* et *a priori*. Il est donc toujours possible d'exprimer une de ces quantités en fonction des deux autres, par exemple :

$$\cos\alpha(p,k) = \frac{RSB_{post}^{local}(p,k) - 1 - RSB_{prio}^{local}(p,k)}{2\sqrt{RSB_{prio}^{local}(p,k)}}. \tag{3.27}$$

Cependant, une bonne estimation de deux de ces trois quantités est indispensable pour obtenir une estimation fiable de la troisième. En pratique, avec les estimateurs disponibles du RSB *a posteriori* et du RSB *a priori*, l'estimation de la phase ne donne pas de résultat convaincant. Finalement, malgré le fait que la phase puisse contenir beaucoup d'informations, il ressort que celle-ci n'est pas exploitable.

3.5 Conclusion

Quatre freins à la qualité des techniques de réduction de bruit par atténuation spectrale à court terme ont été identifiés. Pour résumer, rappelons que :

- Les modèles statistiques classiquement utilisés pour les signaux de parole et de bruit simplifient grandement l'expression (et le calcul) des estimateurs mais sont très éloignés de la réalité (très net pour la parole).
- Il existe deux estimateurs de RSB couramment utilisés mais qui ont chacun un défaut majeur. Ainsi, le RSB a posteriori n'est pas biaisé pour les composantes de parole mais son utilisation introduit énormément de bruit musical. Son pendant, le RSB a priori permet de limiter cet effet mais, en contrepartie, son estimateur est biaisé pour les composantes de parole, ce qui se traduit par un effet de réverbération.
- L'estimation de la DSP du bruit conditionne la qualité des estimateurs de RSB, or on a vu que les estimateurs existants sont loins d'être parfaits. Même les approches qui permettent de poursuivre les non-stationnarités du bruit sont limitées par le fait même que l'estimation est réalisée à long terme, ce qui par ailleurs est nécessaire pour assurer la robustesse de l'estimateur.
- Dans une moindre mesure, l'impact de la phase doit être souligné car il influence également la qualité des estimateurs de RSB. Son impact est maximal lorsque le niveau du bruit est important.

L'analyse de ces limitations a servi de base à la conception des approches proposées dans le chapitre 4 dont voici les trois axes majeurs :

- De nouveaux estimateurs du signal de parole basés sur des modèles statistiques adaptés aux signaux traités sont proposés et permettent ainsi de limiter la distorsion du signal de parole.
- Trois nouveaux estimateurs du RSB seront également présentés. Nous verrons qu'ils permettent de limiter voire de supprimer les défauts des estimateurs classiques du RSB.
- Finalement, nous verrons qu'il est possible de contourner les limitions inhérentes à l'estimation du bruit et de la phase en tirant parti de la structure voisée du signal de parole de façon à limiter les distorsions du signal restauré.

Références

[Breithaupt 2003] C. Breithaupt, et R. Martin, “MMSE Estimation of Magnitude-Squared DFT Coefficients with Supergaussian Priors,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Vol. 1, pp. 896–899, 2003.

[Cappé 1994] O. Cappé, “Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor,” *IEEE Trans. Speech Audio Processing*, Vol. 2, No. 2, pp. 345–349, Avril 1994.

[Cohen 2002a] I. Cohen, et B. Berdugo, “Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement,” *IEEE Signal Processing Lett.*, Vol. 9, No. 1, pp. 12–15, Janvier 2002.

[Ephraïm 1984] Y. Ephraïm, et D. Malah, “Speech Enhancement Using a Minimum Mean Square Error Short-Time Spectral Amplitude Estimator,” *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-32, No. 6, pp. 1109–1121, Décembre 1984.

[Hayes 1980] M. H. Hayes, J. S Lim, et A. V. Oppenheim, “Signal Reconstruction from Phase or Magnitude,” *Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, pp. 672-680, Décembre 1980.

[Kullback 1958] S. Kullback, “Information Theory and Statistics,” *Dover Publication*, 1958.

[Martin 2002] R. Martin, “Speech Enhancement Using MMSE Short Time Spectral Estimation with Gamma Distributed Speech Priors,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Orlando, États-Unis, Vol. 1, pp. 253–256, Mai 2002.

[McAulay 1980] J. McAulay, et M. Malpass, “Speech Enhancement Using a Soft-Decision Noise Suppression Filter,” *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, No. 2, pp. 137–145, Avril 1980.

[Oppenheim 1979] A. V. Oppenheim, J. S. Lim, G. Kopec, et S. C. Pohlig, “Phase in Speech and Pictures,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Washington, États-Unis, Vol. 4, pp. 632–637, Avril 1979.

[Oppenheim 1981] A. V. Oppenheim, et J. S. Lim, “The Importance of Phase in Signals,” *Proc. IEEE*, Vol. 69, No. 5, pp. 529–541, Mai 1981.

[Renevey 2001] P. Renevey, et A. Drygajlo, “Detection of Reliable Features for Speech Recognition in Noisy Conditions Using a Statistical Criterion,” *Proc. Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis*, Aalborg, Denmark, pp. 71–74, Septembre 2001.

[Scalart 1996a] P. Scalart, et J. Vieira Filho, “Speech Enhancement Based on a Priori Signal to Noise Estimation,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Atlanta, États-Unis, Vol. 2, pp. 629–632, Mai 1996.

[Vary 1985] P. Vary, “Noise Suppression by Spectral Magnitude Estimation - Mechanism and Theoretical Limits,” *Signal Processing*, Vol. 8, pp. 387–400, 1985.

[Wang 1982] D. L. Wang, et J. S. Lim, “The Unimportance of Phase in Speech Enhancement,” *Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-30, No. 4, pp. 679–681, Août 1982.

Chapitre 4

Amélioration des techniques de réduction de bruit

L'identification de certaines limitations majeures des techniques de réduction de bruit (*cf.* chapitre précédent) a permis de dégager des pistes qui ont abouti aux méthodes proposées et analysées dans le présent chapitre. Ainsi, la partie 4.1 montre que l'utilisation de modèles statistiques super-gaussiens, mieux adaptés aux signaux traités que le modèle gaussien, débouche sur de nouveaux estimateurs du signal de parole dont le comportement sera analysé. Les trois parties suivantes, 4.2, 4.3 et 4.4 sont consacrées à l'amélioration de l'estimateur du RSB avec pour but d'atteindre le comportement idéal décrit dans la partie 3.2.7. Cependant, même en atteignant ce comportement, les performances de l'estimateur du RSB restent limitées par celles de l'estimateur de la DSP du bruit (*cf.* partie 3.3) et par l'impact de la phase décrit dans la partie 3.4. Une estimation précise du bruit ou de la phase est particulièrement délicate à mener, ainsi, nous proposons dans la partie 4.5 une méthode alternative qui permet de limiter la dégradation du signal de parole restauré en tirant parti de sa structure harmonique. Les approches présentées dans ce chapitre seront illustrées à partir d'un signal bruité représentatif et des résultats complets (sur un large panel de signaux) seront donnés dans le chapitre 6.

4.1 Modèles statistiques super-gaussiens

De nombreuses techniques de réduction de bruit sont basées sur la connaissance *a priori* des densités de probabilité des signaux de parole et de bruit dans le domaine fréquentiel. Ces hypothèses permettent alors d'expliciter le gain spectral correspondant. Ainsi, l'hypothèse la plus courante suppose que les signaux de parole et de bruit suivent le modèle gaussien. Ce choix s'explique notamment par le fait qu'il introduit beaucoup de simplifications dans le calcul des estimateurs. **Cependant, les densités de probabilités de Laplace et Gamma modélisent mieux le signal de parole que la densité de probabilité gaussienne comme démontré dans la partie 3.1** [Martin 2002, Breithaupt 2003, Guédon 2002]. De même, le modèle de Laplace est mieux adapté aux bruits non-stationnaires en comparaison avec le modèle gaussien. En utilisant de tels modèles super-gaussiens pour l'estimation spectrale du signal de parole, il est donc légitime de s'attendre à une

amélioration par rapport au modèle gaussien classique.

Généralement le module et la phase des signaux sont supposés statistiquement indépendants permettant ainsi de traiter le module seul et de réutiliser la phase bruitée (*cf.* partie 2.3). Pour un modèle gaussien cela implique que les parties réelles et imaginaires sont aussi indépendantes. Ceci n'est plus vrai pour des densités de probabilité super-gaussiennes, cependant, la dépendance étant faible en moyenne, les parties réelles et imaginaires seront supposées indépendantes [Martin 2002]. Le but n'est donc plus d'estimer seulement le module mais d'estimer séparément les composantes réelle et imaginaire du signal de parole. Dans la suite, les grandeurs spectrales seront donc séparées en leurs parties réelle et imaginaire et les indices temporels et fréquentiels seront omis afin d'alléger les notations. Ainsi, par exemple, le signal $Y(p,k)$ sera noté Y ; de plus Y_R et Y_I seront définis par $Y = Y_R + iY_I$ comme étant respectivement les parties réelle et imaginaire du signal Y. On fera aussi l'hypothèse que l'énergie des signaux est équi-répartie entre les parties réelle et imaginaire. Toujours pour des raisons de simplicité d'écriture, dans cette partie, on adopte la notation γ pour désigner la variance (évaluée sur chaque composante spectrale d'analyse) d'un signal. Si γ_y représente la variance du signal Y alors la variance des signaux Y_R et Y_I sera $\gamma_y/2$. La moyenne des processus ne portant aucune information, celle-ci est toujours supposée nulle.

Sous ces hypothèses, et selon la loi choisie, la densité de probabilité de la partie réelle (ou imaginaire) du signal utile et du bruit peut s'écrire :

– loi de Gauss :

$$\mathrm{p}(Y_R) = \frac{1}{\sqrt{\pi\gamma_y}} exp\left(-\frac{Y_R^2}{\gamma_y}\right), \tag{4.1}$$

– loi de Laplace :

$$\mathrm{p}(Y_R) = \frac{1}{\sqrt{\gamma_y}} exp\left(-\frac{2|Y_R|}{\sqrt{\gamma_y}}\right), \tag{4.2}$$

– loi Gamma :

$$\mathrm{p}(Y_R) = \frac{\sqrt[4]{3}}{2\sqrt{\pi}\sqrt[4]{2\gamma_y}} |Y_R|^{-\frac{1}{2}} exp\left(-\frac{\sqrt{3}|Y_R|}{\sqrt{2\gamma_y}}\right). \tag{4.3}$$

Grâce à l'hypothèse d'indépendance des parties réelle et imaginaire, l'estimateur optimal au sens de l'EQMM pour les coefficients complexes S du signal de parole peut être décomposé en deux estimateurs, l'un pour la partie réelle et l'autre pour la partie imaginaire. De cette hypothèse, il découle donc directement que :

$$\hat{S} = \mathrm{E}[S \mid X] = \mathrm{E}[S_R \mid X_R] + i\mathrm{E}[S_I \mid X_I]. \tag{4.4}$$

Désormais nous nous intéresserons uniquement au problème de l'estimation de la partie réelle S_R, les résultats obtenus étant facilement transposables à la partie imaginaire en remplaçant les quantités réelles par leurs équivalents imaginaires. Le point de départ pour exprimer les nouveaux estimateurs sous différents modèles de bruit et de signal sera donc :

$$\hat{S}_R = \mathrm{E}[S_R \mid X_R] = \frac{\int_{-\infty}^{\infty} s_r \mathrm{p}(X_R \mid s_r)\mathrm{p}(s_r)\mathrm{d}s_r}{\int_{-\infty}^{\infty} \mathrm{p}(X_R \mid s_r)\mathrm{p}(s_r)\mathrm{d}s_r}. \tag{4.5}$$

Dans le cas des techniques faisant l'hypothèse de densités de probabilités super-gaussiennes, nous ne donnerons pas les gains équivalents aux estimateurs obtenus. En effet, cela n'apporterait rien en

regard de la complexité des écritures. Toutefois, on peut dans tous les cas obtenir l'expression de ces gains en utilisant l'équation :

$$G = \frac{\mathrm{E}[S_R \mid X_R] + i\mathrm{E}[S_I \mid X_I]}{X}. \tag{4.6}$$

Les quantités γ_b et γ_s sont utilisées pour exprimer les différents estimateurs proposés car il n'est pas intéressant de faire apparaître le RSB dans l'expression des estimateurs. En pratique, l'estimée de la DSP du bruit, $\hat{\gamma}_b$, sera obtenue classiquement en utilisant une des approches évoquées dans la partie 2.5. La quantité γ_s sera quant à elle estimée en utilisant l'approche DD présentée dans la partie 3.2.5 :

$$\hat{\gamma}_s = \hat{\gamma}_b \hat{RSB}^{DD}_{prio}. \tag{4.7}$$

Les quatre nouvelles approches proposées ci-après sont regroupées sous le terme d'approche SG pour super-gaussienne. On retient pour le bruit les modèles suivants :

- loi de Gauss (meilleure pour les bruits stationnaires),
- loi de Laplace (meilleure pour les bruits non-stationnaires),

et pour la parole :

- loi de Gauss qui fait office de référence,
- loi de Laplace qui approche mieux la loi expérimentale,
- loi Gamma qui approche fidèlement la loi expérimentale.

Le formalisme utilisé pour analyser ces techniques est repris de [Martin 2002]. La variance de la partie réelle du signal bruité est supposée fixe (uniquement pour l'analyse) : $\gamma_x = \gamma_s + \gamma_b = 2$ et l'estimée de la partie réelle du signal de parole $\hat{S}_R = \mathrm{E}[S_R \mid X_R]$ est représentée en fonction de X_R ($0 \leq X_R \leq 5$) pour 3 valeurs du RSB *a priori* $\left(\frac{\gamma_s}{\gamma_b}\right)$.

4.1.1 Modèle gaussien pour le bruit

4.1.1.1 Modèle gaussien pour la parole

Ce cas n'est évidemment pas nouveau car quand le bruit et la parole sont supposés gaussiens, l'estimateur qui en découle correspond au filtre de Wiener (*cf.* partie 2.4.1.3). Le signal estimé peut donc s'exprimer ainsi :

$$\hat{S} = \mathrm{E}[S \mid X] = \mathrm{E}[S_R \mid X_R] + i\mathrm{E}[S_I \mid X_I] = \frac{\gamma_s}{\gamma_s + \gamma_b} X. \tag{4.8}$$

Cet estimateur servira de référence de comparaison pour ceux obtenus à partir de densités de probabilité super-gaussiennes. On peut rappeler que cet estimateur introduit un effet de flou (analogie avec le traitement d'image) dans la parole traitée qui est lié au lissage du RSB *a priori*. Quand cette quantité est estimée à partir de l'estimateur decision-directed (ce qui est généralement le cas) cet effet peut être interprété comme de la réverbération.

4.1.1.2 Modèle laplacien pour la parole

Sous ces hypothèses (modèle gaussien pour le bruit et laplacien pour la parole), l'equation (4.5) devient l'estimateur Gauss-Laplace qui s'écrit ainsi :

$$\hat{S}_R = \mathrm{E}[S_R \mid X_R] = \frac{\int_{-\infty}^{\infty} s_r exp\left(-\frac{(X_R - s_r)^2}{\gamma_b}\right) exp\left(-\frac{2|s_r|}{\sqrt{\gamma_s}}\right) \mathrm{d}s_r}{\int_{-\infty}^{\infty} exp\left(-\frac{(X_R - s_r)^2}{\gamma_b}\right) exp\left(-\frac{2|s_r|}{\sqrt{\gamma_s}}\right) \mathrm{d}s_r}. \tag{4.9}$$

En introduisant les notations suivantes :

$$L_{R+} = \sqrt{\frac{\gamma_b}{\gamma_s}} + \frac{X_R}{\sqrt{\gamma_b}} \tag{4.10}$$

et

$$L_{R-} = \sqrt{\frac{\gamma_b}{\gamma_s}} - \frac{X_R}{\sqrt{\gamma_b}} \tag{4.11}$$

et en utilisant les théorèmes 3.462,1 (pour le numérateur) et 3.322,2 (pour le dénominateur) de la table [Gradshteyn 1994] on obtient l'estimateur suivant :

$$\hat{S}_R = \mathrm{E}[S_R \mid X_R] = X_R + \frac{\gamma_b}{\sqrt{\gamma_s}} \frac{exp\left(\frac{2X_R}{\sqrt{\gamma_s}}\right) erfc(L_{R+}) - exp\left(-\frac{2X_R}{\sqrt{\gamma_s}}\right) erfc(L_{R-})}{exp\left(\frac{2X_R}{\sqrt{\gamma_s}}\right) erfc(L_{R+}) + exp\left(-\frac{2X_R}{\sqrt{\gamma_s}}\right) erfc(L_{R-})} \tag{4.12}$$

où $erfc(.)$ est la fonction d'erreur complémentaire définie par :

$$erfc(x) = 1 - \frac{2}{\sqrt{\pi}} \int_0^x exp\left(-\frac{t}{2}\right) dt. \tag{4.13}$$

La figure 4.1 représente les résultats obtenus avec l'estimateur Gauss-Laplace. Lorsque le RSB *a priori* est important, le comportement de l'estimateur Gauss-Laplace est similaire à celui du filtre de Wiener. Par contre, pour les faibles RSB *a priori*, son comportement devient non-linéaire. De par la différence entre les modèles statistiques de la parole et du bruit, quand X_R est supérieur à plusieurs fois l'écart-type $\sqrt{\gamma_x}$, la parole devient plus probable que le bruit et l'estimateur proposé délivre donc une valeur significativement plus grande que le filtre de Wiener. Dans le cas contraire, le bruit est plus probable que la parole et l'estimateur délivre alors une valeur plus faible que le filtre de référence. Ce comportement permet ainsi de limiter les distorsions de la parole traitée tout en supprimant plus efficacement le bruit que le filtre de Wiener.

Pour faciliter l'analyse, il est intéressant d'introduire la notion de RSB *a posteriori* pour la partie réelle de X :

$$RSB_{post}^{X_R} = \frac{|X_R|^2}{\mathrm{E}[|B_R|^2]} = 2X_R^2/\gamma_b. \tag{4.14}$$

Pour chacune des courbes représentées, la valeur γ_b est figée et donc une valeur élevée (>1) de X_R correspond à une valeur élevée de ce RSB *a posteriori*. On sait que pour le filtre MMSE STSA (*cf.* partie 2.4.2.2) le RSB *a posteriori* agit comme une correction du RSB *a priori*. Cependant, cette correction va à l'encontre de ce que le RSB *a posteriori* indique. Ce comportement bien que paradoxal permet de limiter efficacement le bruit musical avec en contrepartie une augmentation de l'effet de

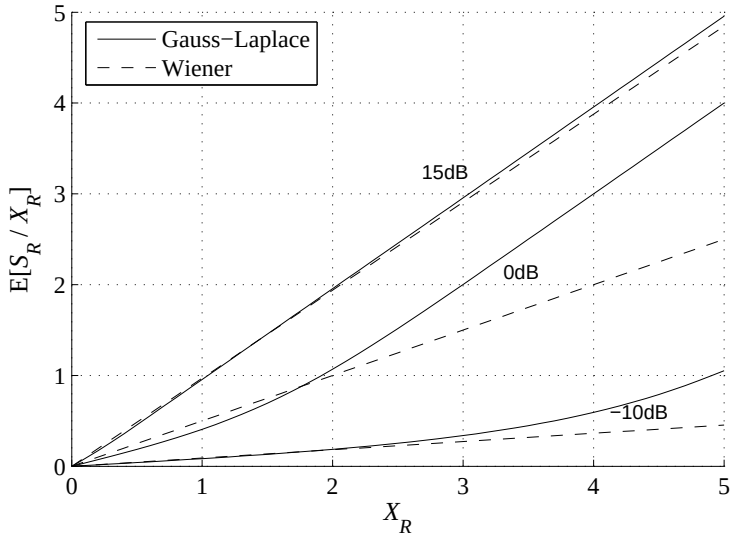

FIG. 4.1 – *Signal estimé $E[S_R \mid X_R]$ dans le cas où le bruit suit un modèle gaussien et la parole un modèle laplacien et cela pour 3 valeurs du RSB* a priori *(trait plein) : 15, 0 et -10 dB avec la contrainte que $\gamma_s + \gamma_b = 2$. La référence est le filtre de Wiener (tirets).*

réverbération. Dans l'approche proposée ici, le RSB *a posteriori* agit aussi comme une correction pour le RSB *a priori* mais celle-ci va dans le sens du RSB *a posteriori*. En effet, lorsque le RSB *a posteriori* est plus grand que le RSB *a priori* alors l'estimateur délivre une valeur plus importante que celle du filtre de Wiener et inversement. **Cette caractéristique permet de corriger les erreurs d'estimation du RSB *a priori* dues au lissage de son estimateur. L'effet de réverbération est donc diminué par rapport à celui obtenu avec le filtre de Wiener.**

4.1.1.3 Modèle Gamma pour la parole

Sous ces nouvelles hypothèses (modèle gaussien pour le bruit et Gamma pour la parole), l'équation (4.5) devient l'estimateur Gauss-Gamma qui s'écrit ainsi :

$$\hat{S}_R = \mathrm{E}[S_R \mid X_R] = \frac{\int_{-\infty}^{\infty} s_r |s_r|^{-\frac{1}{2}} exp\left(-\frac{(X_R - s_r)^2}{\gamma_b}\right) exp\left(-\frac{\sqrt{3}|s_r|}{\sqrt{2\gamma_s}}\right) \mathrm{d}s_r}{\int_{-\infty}^{\infty} |s_r|^{-\frac{1}{2}} exp\left(-\frac{(X_R - s_r)^2}{\gamma_b}\right) exp\left(-\frac{\sqrt{3}|s_r|}{\sqrt{2\gamma_s}}\right) \mathrm{d}s_r}. \tag{4.15}$$

En introduisant les notations suivantes [Martin 2002] :

$$G_{R+} = \frac{\sqrt{3}}{2\sqrt{2}} \sqrt{\frac{\gamma_b}{\gamma_s}} + \frac{X_R}{\sqrt{\gamma_b}} \tag{4.16}$$

et

$$G_{R-} = \frac{\sqrt{3}}{2\sqrt{2}} \sqrt{\frac{\gamma_b}{\gamma_s}} - \frac{X_R}{\sqrt{\gamma_b}} \tag{4.17}$$

et en utilisant les théorèmes 3.462,1 de la table [Gradshteyn 1994] (pour le numérateur et le dénominateur), on obtient l'estimateur suivant [Martin 2002] :

$$\hat{S}_R = \mathrm{E}[S_R \mid X_R] = \sqrt{\gamma_b}\frac{exp\left(\frac{G_{R-}^2}{2}\right)D_{-1.5}\left(\sqrt{2}G_{R-}\right) - exp\left(\frac{G_{R+}^2}{2}\right)D_{-1.5}\left(\sqrt{2}G_{R+}\right)}{exp\left(\frac{G_{R-}^2}{2}\right)D_{-1.5}\left(\sqrt{2}G_{R-}\right) + exp\left(\frac{G_{R+}^2}{2}\right)D_{-1.5}\left(\sqrt{2}G_{R+}\right)} \tag{4.18}$$

où $D_p(.)$ est une fonction de cylindre parabolique définie par le théorème 9.240 de cette même table [Gradshteyn 1994].

La figure 4.2 représente les résultats obtenus avec l'estimateur Gauss-Gamma. Le comportement

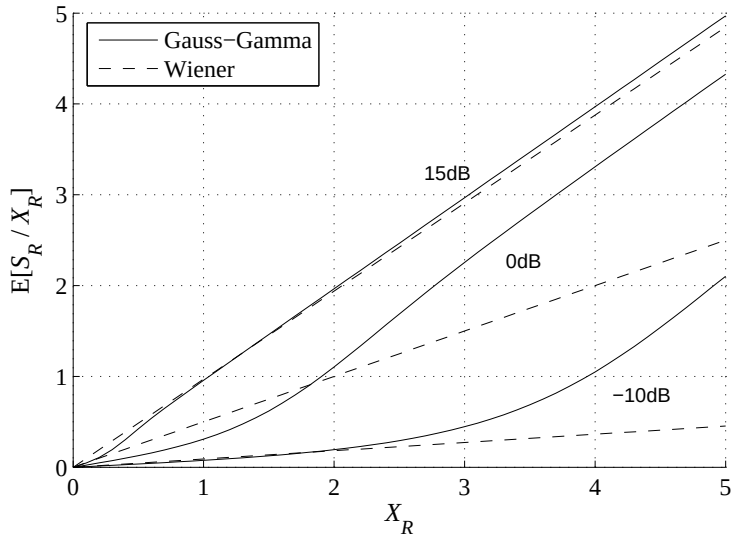

FIG. 4.2 – *Signal estimé $E[S_R \mid X_R]$ dans le cas où le bruit suit un modèle gaussien et la parole un modèle Gamma et cela pour 3 valeurs du RSB* a priori *(trait plein) : 15, 0 et -10 dB avec la contrainte que $\gamma_s + \gamma_b = 2$. La référence est le filtre de Wiener (tirets).*

de cet estimateur est similaire à celui du cas précédent (Gauss-Laplace), mais présente des non-linéarités encore plus prononcées qui s'expliquent par le fait que le modèle Gamma est plus éloigné du modèle gaussien que ne l'est le modèle laplacien. **Les non-linéarités étant plus importantes que pour l'estimateur Gauss-Laplace, la correction apportée par le RSB *a posteriori*, défini par l'équation (4.14), est plus marquée. La distorsion de la parole est donc moindre et l'effet de réverbération est très réduit.**

4.1.2 Modèle laplacien pour le bruit

4.1.2.1 Modèle laplacien pour la parole

Quand le bruit et la parole suivent tous deux un modèle laplacien, l'équation (4.5) devient l'estimateur Laplace-Laplace qui s'écrit ainsi :

$$\hat{S}_R = \mathrm{E}[S_R \mid X_R] = \frac{\int_{-\infty}^{\infty} s_r exp\left(-\frac{|X_R - s_r|}{\sqrt{\gamma_b}}\right) exp\left(-\frac{2|s_r|}{\sqrt{\gamma_s}}\right) \mathrm{d}s_r}{\int_{-\infty}^{\infty} exp\left(-\frac{|X_R - s_r|}{\sqrt{\gamma_b}}\right) exp\left(-\frac{2|s_r|}{\sqrt{\gamma_s}}\right) \mathrm{d}s_r}. \tag{4.19}$$

Si $\sqrt{\gamma_b} \neq \sqrt{\gamma_s}$ alors l'estimateur devient :

$$\hat{S}_R = \mathrm{E}[S_R \mid X_R] = sign(X_R) \frac{\frac{\gamma_s \gamma_b}{\gamma_b - \gamma_s}\left(exp\left(-\frac{2|X_R|}{\sqrt{\gamma_b}}\right) - exp\left(-\frac{2|X_R|}{\sqrt{\gamma_s}}\right)\right) - |X_R| exp\left(-\frac{2|X_R|}{\sqrt{\gamma_s}}\right)\sqrt{\gamma_s}}{exp\left(-\frac{2|X_R|}{\sqrt{\gamma_b}}\right)\sqrt{\gamma_b} - exp\left(-\frac{2|X_R|}{\sqrt{\gamma_s}}\right)\sqrt{\gamma_s}} \tag{4.20}$$

et pour $\sqrt{\gamma_b} = \sqrt{\gamma_s}$ on obtient :

$$\hat{S}_R = \mathrm{E}[S_R \mid X_R] = \frac{X_R}{2}. \tag{4.21}$$

Ce cas particulier rejoint d'ailleurs la solution du filtre de Wiener.

La figure 4.3 représente les résultats obtenus avec l'estimateur Laplace-Laplace. Pour les RSB *a*

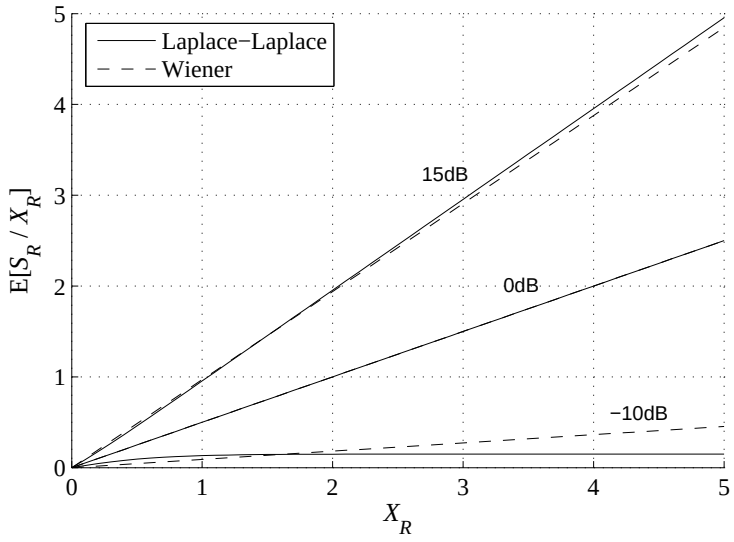

FIG. 4.3 – *Signal estimé $E[S_R \mid X_R]$ dans le cas où le bruit et la parole suivent un modèle laplacien et cela pour 3 valeurs du RSB* a priori *(trait plein) : 15, 0 et -10 dB avec la contrainte que $\gamma_s + \gamma_b = 2$. La référence est le filtre de Wiener (tirets).*

priori supérieurs à 0dB, le comportement de cet estimateur est très proche de celui du filtre de Wiener. **Par contre, pour les faibles RSB *a priori*, cet estimateur a tendance à délivrer une valeur constante, indépendante de la valeur du signal bruité ce qui a pour effet de limiter le bruit musical résiduel.** D'un autre côté, ce comportement augmente les distorsions de la parole en ne permettant pas au RSB *a posteriori* (4.14) de corriger les erreurs du RSB *a priori* dues à son estimation lissée (*cf.* estimateur Gauss-Laplace). L'effet de réverbération reste donc aussi très présent.

4.1.2.2 Modèle Gamma pour la parole

Sous ces hypothèses (modèle laplacien pour le bruit et Gamma pour la parole), l'équation (4.5) devient l'estimateur Laplace-Gamma qui s'écrit ainsi :

$$\hat{S}_R = \mathrm{E}[S_R \mid X_R] = \frac{\int_{-\infty}^{\infty} s_r |s_r|^{-\frac{1}{2}} exp\left(-\frac{2|X_R - s_r|}{\sqrt{\gamma_b}}\right) exp\left(-\frac{\sqrt{3}|s_r|}{\sqrt{2\gamma_s}}\right) \mathrm{d}s_r}{\int_{-\infty}^{\infty} |s_r|^{-\frac{1}{2}} exp\left(-\frac{2|X_R - s_r|}{\sqrt{\gamma_b}}\right) exp\left(-\frac{\sqrt{3}|s_r|}{\sqrt{2\gamma_s}}\right) \mathrm{d}s_r}. \tag{4.22}$$

En introduisant les notations suivantes [Martin 2002] :

$$G_+ = \frac{\sqrt{3}}{2\sqrt{2}} \frac{1}{\sqrt{\gamma_s}} + \frac{1}{\sqrt{\gamma_b}} \tag{4.23}$$

et

$$G_+ = \frac{\sqrt{3}}{2\sqrt{2}} \frac{1}{\sqrt{\gamma_s}} - \frac{1}{\sqrt{\gamma_b}} \tag{4.24}$$

et en utilisant le théorème 3.381 de la table [Gradshteyn 1994], on obtient l'estimateur suivant pour $X_R \geq 0$ [Martin 2002] :

$$\hat{S}_R = \mathrm{E}[S_R \mid X_R] = \frac{\sqrt[4]{3}}{2\sqrt{\pi\gamma_b}\sqrt[4]{2\gamma_s}\mathrm{p}(X_R)} \left[\frac{2}{3} exp\left(-\frac{2X_R}{\sqrt{\gamma_b}}\right) X_R^{3/2} M\left(\frac{3}{2}, \frac{5}{2}; -2G_- X_R\right)\right.$$

$$\left. + exp\left(-\sqrt{\frac{3}{2}}\frac{1}{\sqrt{\gamma_s}} X_R\right) (2G_+)^{-3/2} \Psi\left(-\frac{1}{2}, -\frac{1}{2}; 2G_+ X_R\right) - exp\left(-\frac{2X_R}{\sqrt{\gamma_b}}\right) (2G_+)^{-3/2} \Gamma\left(\frac{3}{2}\right)\right]. \tag{4.25}$$

La quantité $\mathrm{p}(X_R)$ s'écrit :

$$\mathrm{p}(X_R) = \frac{\sqrt[4]{3}}{2\sqrt{\pi\gamma_b}\sqrt[4]{2\gamma_s}} \left[2exp\left(-\frac{2X_R}{\sqrt{\gamma_b}}\right) \sqrt{X_R} M\left(\frac{1}{2}, \frac{3}{2}; -2G_- X_R\right)\right.$$

$$\left. + exp\left(-\sqrt{\frac{3}{2}}\frac{1}{\sqrt{\gamma_s}} X_R\right) \frac{1}{\sqrt{2G+}} \Psi\left(\frac{1}{2}, \frac{1}{2}; 2G_+ X_R\right) + exp\left(-\frac{2X_R}{\sqrt{\gamma_b}}\right) \sqrt{\frac{\pi}{2G_+}}\right] \tag{4.26}$$

où $M(.)$ est la fonction hypergéométrique confluente utilisée dans l'approche MMSE STSA (*cf.* partie 2.4.2.2) et $\Psi(.)$ en est également une définie par le théorème 9.210 de la table [Gradshteyn 1994]. La fonction Gamma Γ est définie dans le théorème 8.310 de cette même table [Gradshteyn 1994]. Pour $X_R < 0$ on utilise la relation $\mathrm{E}[S_R \mid X_R] = -\mathrm{E}[S_R \mid -X_R]$.

La figure 4.4 représente les résultats obtenus avec l'estimateur Laplace-Gamma. Le comportement de cet estimateur est similaire à celui du cas précédent (estimateur Laplace-Laplace) dans le sens où, pour les RSB *a priori* très faibles, la sortie est quasiment constante quelle que soit la valeur de l'entrée, limitant ainsi le bruit musical. Par contre, on peut noter que pour des RSB *a priori* proches de 0dB cet estimateur a alors un comportement proche du cas Gauss-Gamma ce qui lui permet de limiter les distorsions de la parole et d'atténuer sensiblement l'effet de réverbération. Pour des RSB *a priori* importants, nous retrouvons un comportement très proche du filtre de Wiener.

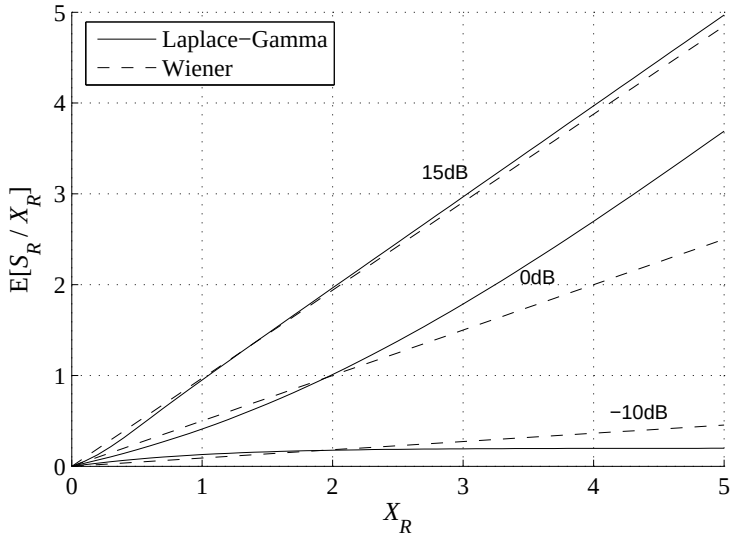

FIG. 4.4 – *Signal estimé $E[S_R \mid X_R]$ dans le cas où le bruit suit un modèle laplacien et la parole un modèle Gamma et cela pour 3 valeurs du RSB* a priori *(trait plein) : 15, 0 et -10 dB avec la contrainte que $\gamma_s + \gamma_b = 2$. La référence est le filtre de Wiener (tirets).*

4.1.3 Complexité des approches SG

L'utilisation de modèles statistiques plus proches de la réalité que le modèle gaussien permet de limiter la distorsion du signal restauré et améliore sensiblement les résultats par rapport à l'approche DD (avec le filtre de Wiener). Ceci est d'ailleurs confirmé par les résultats donnés dans la partie 6.3.2. **Cependant, il faut souligner que les modèles super-gaussiens utilisés augmentent largement la complexité des estimateurs les rendant ainsi rédhibitoires en terme de coût de calcul** qui doit rester raisonnable pour les raisons exposées dans la partie 1.2.2.

4.2 Généralisation de l'approche decision-directed

La partie 3.2.5 décrit et met en avant les défauts et avantages de l'estimateur decision-directed du RSB *a priori*. Cette approche est efficace sur des portions stationnaires du signal de parole comme par exemple une voyelle longue (signal voisé) dont l'amplitude et la fréquence de chaque harmonique restent constantes. En fait, c'est le seul cas où l'estimateur DD n'est pas biaisé, le retard de l'estimateur n'étant pas un problème du fait de la stationnarité du signal. Le biais de cet estimateur est par contre particulièrement gênant dans les zones transitoires du signal de parole. Ces transitoires sont typiquement des attaques ou des extinctions de signal. Cependant, on retrouve aussi de telles transitions pendant la parole continue tout simplement lorsque la fréquence d'une harmonique change d'une trame à l'autre : saut de la fréquence (discrète) k' à la fréquence k.

Admettons qu'une harmonique d'amplitude constante suive le scénario décrit dans la figure 4.5 où entre les trames $p-1$ et p l'harmonique passe de la fréquence k' à la fréquence k. Comme l'estimateur DD considère chaque fréquence indépendamment, le changement de fréquence est tout simplement

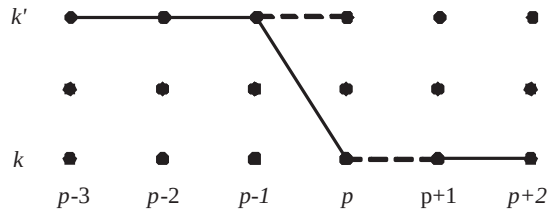

FIG. 4.5 – *Saut de fréquence pour une harmonique dont l'amplitude est constante. Les tirets représentent les zones où l'estimateur DD engendre des erreurs.*

interprété comme une disparition de l'harmonique à la fréquence k' accompagnée de l'apparition d'une autre harmonique à la fréquence k. Ainsi, la quantité $RSB_{prio}(p,k')$ sera surestimée créant un effet de traînage dû au biais de l'estimateur DD. Pour la même raison, la quantité $RSB_{prio}(p,k)$ sera sous-estimée. Ces deux types d'erreurs font partie intégrante de l'effet de réverbération caractéristique de l'approche DD. Si la fréquence de cette harmonique était restée constante, l'estimateur n'aurait pas commis d'erreur. Il faut souligner que ceci n'est vrai que dans le cas où l'amplitude est constante, c'est-à-dire le seul cas où le biais de l'estimateur ne provoque pas de dégradation. **Pour éviter ce type de dégradation, nous proposons un estimateur decision-directed capable de suivre les évolutions fréquentielles des harmoniques du signal de parole.** Cette généralisation permet alors de se ramener dans le cas où la fréquence reste constante. En pratique, le suivi harmonique est assuré en associant à chaque fréquence k de la trame p la fréquence $\tilde{k}$ de la trame $p-1$. Nous obtenons ainsi un estimateur decision-directed généralisé (ou DDG) qui s'exprime comme ceci :

$$\hat{RSB}^{DD}_{prio}(p,k) = \beta\frac{|\hat{S}_{DD}(p-1,\tilde{k})|^2}{\hat{\gamma}_b(\tilde{k})} + (1-\beta)\max\left(\hat{RSB}_{post}(p,k)-1,0\right). \tag{4.27}$$

Le paramètre β joue un rôle identique à celui de l'équation (3.12). Par exemple, si l'on reprend le cas décrit par la figure 4.5, le calcul de la quantité $\hat{RSB}^{DD}_{prio}(p,k)$ fait intervenir des composantes à la fréquence k' dans le premier terme de l'équation (4.27) (trame $p-1$) et à la fréquence k dans le second terme (trame p) rendant ainsi le changement de fréquence de l'harmonique complètement transparent.

Cette nouvelle approche n'est bien entendu applicable que dans la mesure où il est possible d'identifier le changement de fréquence des composantes du signal de parole. Seules les parties voisées sont donc concernées mais cela reste tout de même intéressant dans la mesure où elles représentent en moyenne 80% des sons prononcés. Il faut donc être capable d'identifier la position de chaque harmonique du signal pour chaque trame. Pour ce faire, dans un objectif de validation de la méthode, le suivi harmonique est réalisé par marquage manuel à partir d'un signal de parole propre et peut donc être considéré comme idéal. À chaque composante fréquentielle $S(p,k)$ est donc associé un élément binaire $I(p,k)$ qui indique si cette composante contient une harmonique ($I(p,k)=1$) ou non ($I(p,k)=0$). En pratique, l'influence de la fenêtre d'analyse a pour effet d'étaler les harmoniques sur plusieurs fréquences d'analyse mais seule la fréquence centrale est ici considérée pour représenter l'harmonique toute entière. La figure 4.6.(b) représente le suivi harmonique obtenu pour le signal de parole servant d'exemple représenté par le figure 4.6.(a). Comme les harmoniques s'étalent sur plusieurs bandes de fréquence, il faut donc, à partir de la connaissance de leurs positions, déterminer une

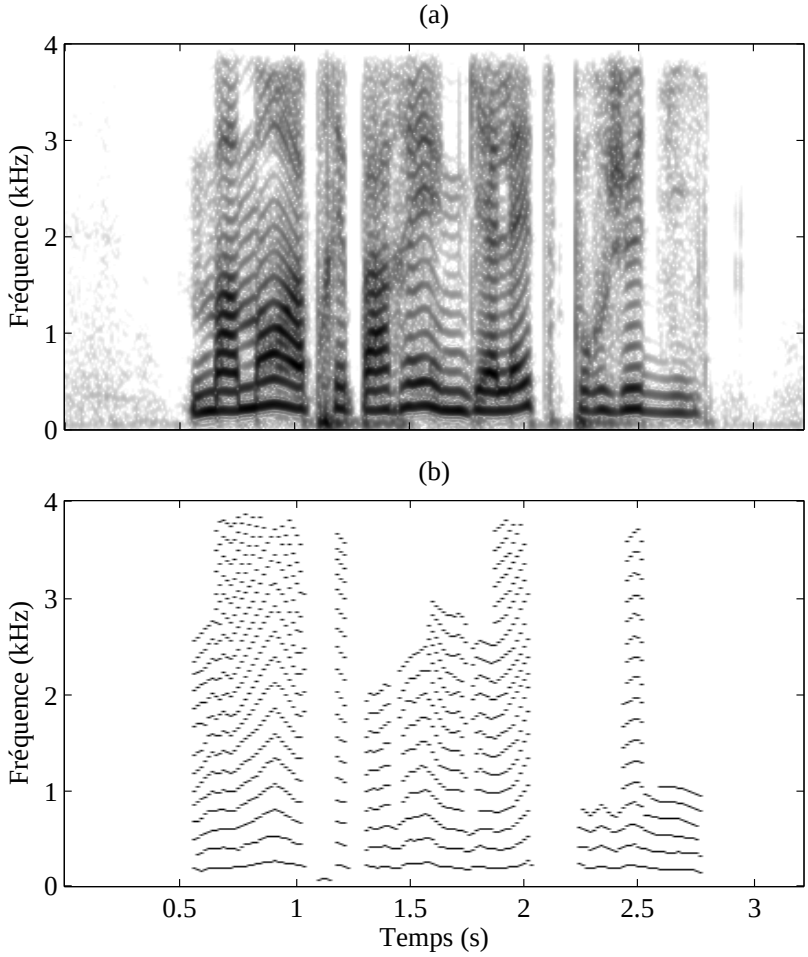

FIG. 4.6 – *(a) Spectrogramme du signal de parole considéré et (b) marquage idéal des harmoniques (suivi harmonique).*

correspondance entre toutes les fréquences de la trame p et celles de la trame $p-1$ comme illustré par la figure 4.7. Ceci est réalisé par interpolation entre chaque paire d'harmoniques de façon à traduire l'évolution fréquentielle de toutes les bandes de fréquence correspondant à chacune des harmoniques du signal de parole. Selon les paramètres choisis pour l'analyse, d'une trame à l'autre, l'amplitude du saut en fréquence d'une harmonique peut être très variable. Les paramètres retenus dans cet exemple sont classiques : $Fe = 8$kHz, trame de 256 points, FFT de 512 points et recouvrement de 50%. Dans ces conditions, le saut fréquentiel est au maximum de 6 bandes de fréquence soit environ 100Hz. Considérons la n$^{\text{ème}}$ harmonique du signal de parole. Soit $I(p-1,k') = 1$ et $I(p,k) = 1$ indiquant qu'entre les trames $p-1$ et p la position de celle-ci est passée de la fréquence k' à la fréquence k. Considérons également la n$+1^{\text{ème}}$ harmonique, son évolution étant résumée par $I(p-1,k'+M) = 1$ et $I(p,k+N) = 1$. Il est alors possible d'interpoler le suivi harmonique entre ces deux harmoniques comme illustré sur la figure 4.7. À la trame p, le RSB *a priori* à la fréquence $k+n$ sera donc calculé, selon l'équation (4.27), à partir du signal de parole estimé à la trame précédente (premier terme de

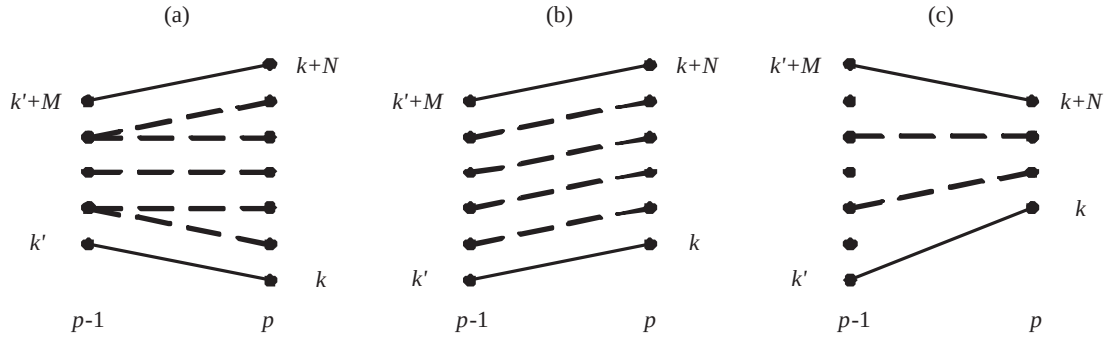

FIG. 4.7 – *Interpolation des chemins entre deux harmoniques. Les traits pleins représentent les harmoniques et les tirets les chemins interpolés. Les cas (a), (b) et (c) illustrent les 3 possibilités : l'écart entre les deux harmoniques augmente, reste constant ou bien diminue.*

l'équation (4.27)) mais à la fréquence :

$$k' + \left[n\frac{M}{N} + \frac{1}{2} \right] \text{ pour } n = 1, \ldots, N-1 \tag{4.28}$$

où l'opérateur [.] permet d'obtenir la partie entière. Le rapport $\frac{M}{N}$ est généralement proche de 1 ce qui témoigne d'une évolution plutôt lente de la fréquence fondamentale. Les harmoniques étant des multiples de cette fréquence, cet effet se retrouve démultiplié ce qui explique que la fréquence d'une harmonique peut évoluer très rapidement. Toutefois, même dans ce cas, les écarts de fréquence successifs entre deux harmoniques évoluent lentement.

La figure 4.8 montre que l'**on obtient des améliorations sensibles dans les zones où la fréquence des harmoniques évolue rapidement. La continuité des harmoniques est en effet mieux respectée qu'avec l'approche DD classique.** Par contre, lorsque la fréquence varie lentement alors

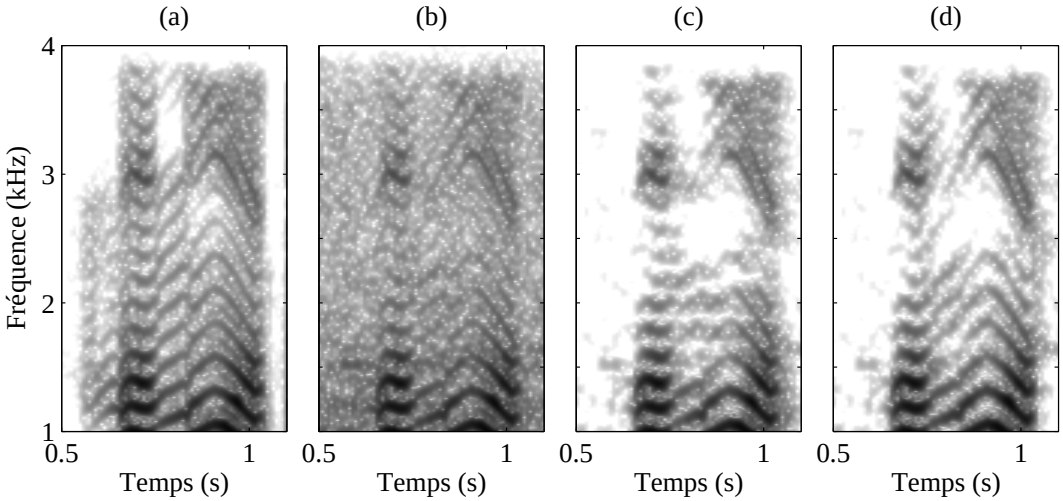

FIG. 4.8 – *Spectrogrammes. (a) Signal de parole propre ; (b) signal de parole bruité ; (c) signal restauré avec l'approche DD ; (d) signal restauré avec l'approche DDG.*

l'étalement fréquentiel des harmoniques (dû à la fenêtre d'analyse) fait que l'estimateur DD bénéficie tout de même d'une partie de la contribution de l'harmonique (de la trame précédente) ce qui limite donc son biais et rend l'impact de l'approche DDG moins important. Il faut aussi souligner que cette approche ne permet pas d'améliorer l'estimateur DD dans le cas des attaques et extinctions de signal ni dans le cas où l'amplitude des harmoniques fluctue. **Ainsi, même si objectivement une amélioration existe, elle reste peu audible car elle n'est pas suffisante pour supprimer l'effet de réverbération qui reste très présent.** Cette approche ne sera donc pas exploitée dans des conditions réelles (suivi estimé à partir d'un signal bruité) qui de toute façon posent le problème de l'obtention d'un suivi précis et fiable des harmoniques dans les zones fortement bruitées.

4.3 Réestimation du RSB *a priori* par une approche en deux passes

4.3.1 Principe de l'approche en deux passes

À la lumière de l'analyse réalisée dans les parties 3.2.5.2 et 3.2.6, et de la conclusion qui en est tirée dans la partie 3.2.3, nous proposons d'estimer le RSB *a priori* par une procédure en deux passes qui a pour but de supprimer le biais de l'estimateur DD tout en conservant ses propriétés de réduction du niveau de bruit musical. Nous avons vu dans la section 3.2.5 que l'approche DD introduit un retard d'une trame dans l'estimation du RSB *a priori*. Ceci est particulièrement net sur la figure 3.9. Le gain spectral calculé pour la trame p à partir de ce RSB biaisé est donc mieux adapté à la trame $p-1$ qu'à la trame courante. À partir de cette analyse, nous proposons de calculer le gain spectral à la trame $p+1$ et de l'appliquer à la trame courante p. Cela nous conduit à proposer un algorithme en deux passes appelé "Two-step noise reduction technique" ou TSNR [Plapous 2004, Plapous à paraître].

Dans la première passe, de façon classique, le gain spectral $G_{DD}(p,k)$ est calculé par l'approche DD comme décrit dans l'équation (3.15). Dans la seconde passe, ce gain est utilisé pour estimer le RSB *a priori* à la trame $p+1$:

$$\hat{RSB}_{prio}^{TSNR}(p,k) = \hat{RSB}_{prio}^{DD}(p+1,k) = \beta' \frac{|G_{DD}(p,k)X(p,k)|^2}{\hat{\gamma}_b(k)} + (1-\beta')\max\left(\hat{RSB}_{post}(p+1,k) - 1,0\right), \tag{4.29}$$

où β' joue le même rôle que β dans l'équation (3.12) mais peut avoir une valeur différente. Il est important de remarquer que le calcul de $\hat{RSB}_{post}(p+1,k)$ requiert la connaissance de la trame future $X(p+1,k)$, introduisant ainsi un retard algorithmique supplémentaire souvent incompatible avec les applications visées (contraintes de temps réel par exemple, *cf.* partie 1.2.2). Cette limitation peut être contournée en utilisant un cas dégénéré de l'équation (4.29) où $\beta' = 1$ ce qui conduit donc à un estimateur qui ne dépend plus de la trame future :

$$\hat{RSB}_{prio}^{TSNR}(p,k) = \frac{|G_{DD}(p,k)X(p,k)|^2}{\hat{\gamma}_b(k)} = \frac{|\hat{S}_{DD}(p,k)|^2}{\hat{\gamma}_b(k)} \tag{4.30}$$

où $\hat{S}_{DD}(p,k)$ est le spectre du signal restauré en utilisant l'approche DD. Bien entendu, un tel choix n'est valide que pour la deuxième passe dont le but est de raffiner l'estimation de la première ($\beta = 0{,}98$ pour la première passe). Il a d'ailleurs été vérifié que cela permet de supprimer plus de bruit musical

qu'en choisissant $\beta' = 0{,}98$. Finalement, de façon générale, le gain spectral est obtenu ainsi :

$$G_{TSNR}(p,k) = h\left(R\hat{S}B_{prio}^{TSNR}(p,k), R\hat{S}B_{post}(p,k)\right). \tag{4.31}$$

La fonction de gain $h(.)$ peut être différente de la fonction $g(.)$ utilisée dans l'équation (3.13). Le spectre du signal de parole restauré est ensuite obtenu par :

$$\hat{S}_{TSNR}(p,k) = G_{TSNR}(p,k)X(p,k). \tag{4.32}$$

Dans la suite, par défaut, la fonction de gain retenue sera le filtre de Wiener :

$$G_{TSNR}(p,k) = \frac{R\hat{S}B_{prio}^{TSNR}(p,k)}{1 + R\hat{S}B_{prio}^{TSNR}(p,k)}. \tag{4.33}$$

Le principe de cet algorithme en deux passes défini par les équations (3.11), (3.12), (3.15), (4.30) et (4.31) est résumé par le schéma de la figure 4.9.

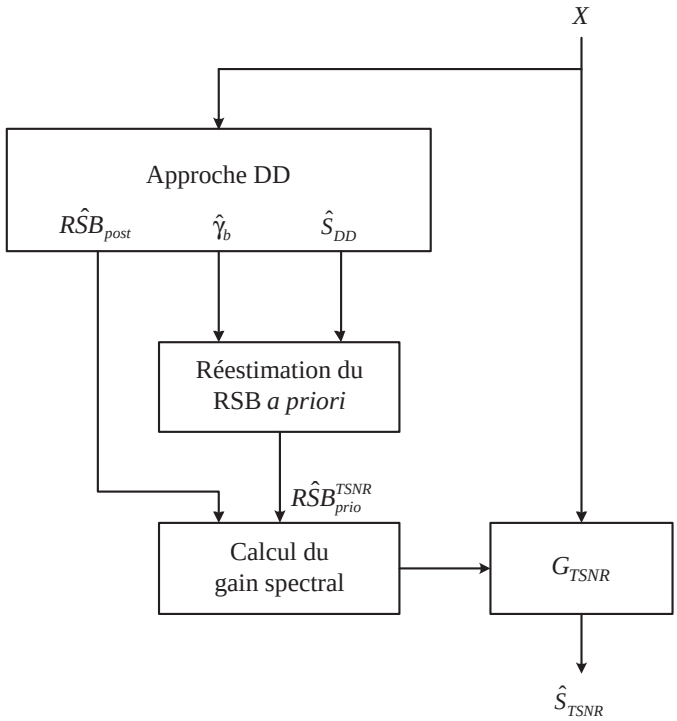

FIG. 4.9 – *Schéma du principe général de l'approche TSNR.*

4.3.2 Analyse théorique de l'approche TSNR

Le signal bruité décrit dans la section 3.2.1 a été traité par les approches DD et TSNR. Les variations temporelles des différents RSB mis en jeu sont représentées sur la figure 4.10. Les conditions sont identiques à celles de la figure 3.9 à savoir que les RSB sont représentés pour la bande de fréquence centrée sur $467Hz$ ($30^{\text{ème}}$ bande de fréquence). Les 20 premières et les 17 dernières trames

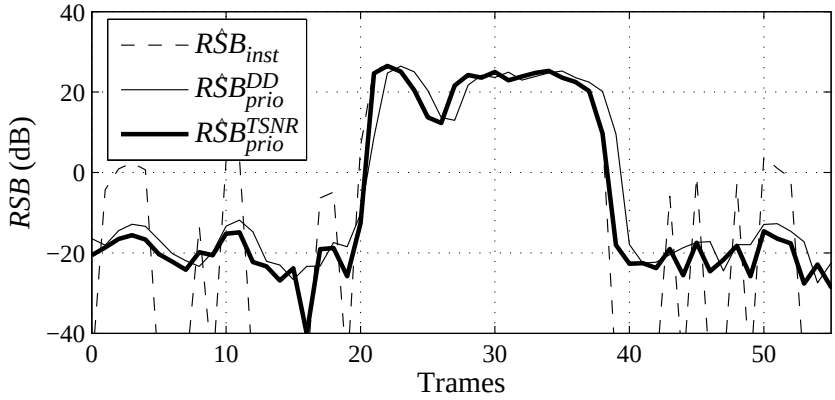

FIG. 4.10 – *Evolution temporelle du RSB pour les approches DD et TSNR (pour la fréquence* 467*Hz). En pointillé : RSB* instantané *; en trait fin : RSB* a priori *de l'approche DD ; en trait fort : RSB* a priori *de l'approche TSNR.*

contiennent uniquement du bruit et les 19 au centre contiennent de la parole bruitée (RSB global de 12dB). Le comportement de l'estimateur du RSB *a priori* de l'approche TSNR est illustré par la figure 4.10 et peut se résumer de la façon suivante :

- Quand le RSB *instantané* est beaucoup plus grand que 0dB, le $\hat{RSB}^{TSNR}_{prio}(p,k)$ suit parfaitement et sans retard le RSB *instantané* contrairement à $\hat{RSB}^{DD}_{prio}(p,k)$. De plus ce suivi est aussi assuré lorsque le $\hat{RSB}_{inst}(p,k)$ croît ou décroît rapidement (attaque ou extinction de parole). Ce point est important car c'est durant ces périodes que les erreurs d'estimation du $\hat{RSB}^{DD}_{prio}(p,k)$ sont les plus importantes.
- Quand le RSB *instantané* est plus petit ou proche de 0dB, le $\hat{RSB}^{TSNR}_{prio}(p,k)$ atteint (généralement) des valeurs sensiblement inférieures à celles du $\hat{RSB}^{DD}_{prio}(p,k)$. De plus on peut remarquer que la deuxième passe supprime le retard introduit par l'approche DD même pour les faibles RSB tout en conservant l'effet de lissage désiré. Ce comportement est cohérent avec le fait que $\beta' = 1$ dans la seconde passe (4.30) qui est aussi un estimateur decision-directed, le fait d'augmenter β' réduisant le niveau de bruit musical.

L'approche TSNR permet donc de supprimer le retard de l'approche DD tout en garantissant un niveau de bruit musical inférieur. Cette technique permet donc de bien préserver les phases transitoires de parole (attaques et extinctions) et de supprimer l'effet de réverbération de l'approche DD dû au biais de son estimateur du RSB *a priori*. On peut préciser qu'en pratique cet effet de réverbération peut être réduit en augmentant le recouvrement entre les trames analysées (ce qui augmente du même coup la charge de calcul) mais il ne peut pas être supprimé alors que l'approche TSNR le permet avec un recouvrement classique de 50%.

L'outil d'analyse présenté dans la partie 3.2.4 est utilisé comme support à l'analyse de l'approche TSNR. Les couples $(\hat{RSB}_{post}, \hat{RSB}^{TSNR}_{prio})$ sont donc représentés dans la figure 4.11 où les RSB *a posteriori* et *a priori* sont estimés en utilisant respectivement les équations (3.11) et (4.30). Pour quantifier l'apport de l'approche TSNR par rapport à l'approche DD, ce nuage de points est analysé en regard de celui obtenu dans la figure 3.11. Deux zones de la figure 4.11 contiennent une importante densité

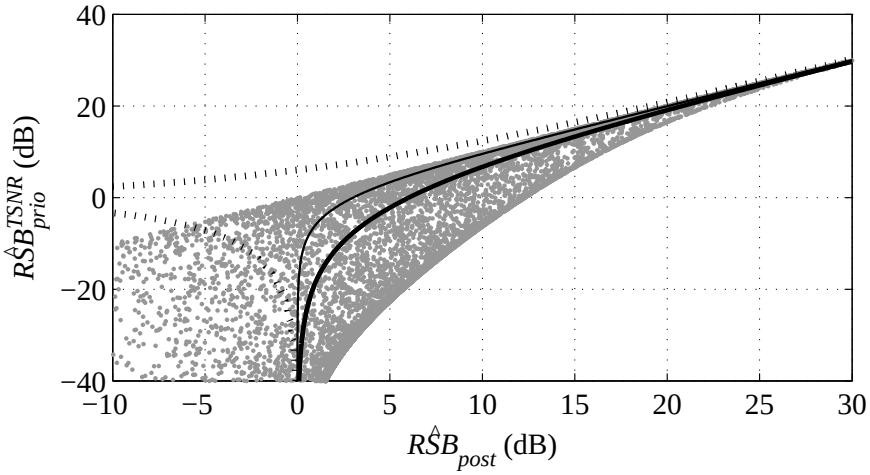

FIG. 4.11 – $\hat{RSB}^{TSNR}_{prio}$ *en fonction du* $\hat{RSB}_{post}$ *dans le cas de l'approche TSNR. Les trois lignes matérialisent l'équation (3.9) dans le cas où* $\alpha(p,k) = 0$ *(trait fort),* $\alpha(p,k) = \pi$ *(pointillé) et* $\alpha(p,k) = \frac{\pi}{2}$ *(trait fin).*

de points, elles correspondent en fait à deux comportements asymptotiques de l'approche TSNR.

La limite basse du nuage de points correspond au cas où la parole est absente à la trame $p-1$, *i.e.* $\hat{S}(p-1,k) = 0$. Ainsi, à la trame p l'estimateur DD du RSB *a priori* (équation (3.11)) conduit à :

$$\hat{RSB}^{DD}_{prio}(p,k) = (1-\beta)\max\left(\hat{RSB}_{post}(p,k) - 1,0\right). \tag{4.34}$$

Si un signal de parole apparaît à la trame p, le RSB *instantané* correspondant sera atténué d'environ 17dB provoquant ainsi une dégradation de l'attaque de ce signal. La seconde passe de l'approche TSNR permet de raffiner cette estimation. Ainsi, en utilisant les équations (3.15) et (3.11), on obtient à partir de (4.30) l'estimation suivante :

$$\hat{RSB}^{TSNR}_{prio}(p,k) = \left(\frac{(1-\beta)\max\left(\hat{RSB}_{post}(p,k) - 1,0\right)}{1+(1-\beta)\max\left(\hat{RSB}_{post}(p,k) - 1,0\right)}\right)^2 \hat{RSB}_{post}(p,k). \tag{4.35}$$

En cherchant le point de croisement entre les courbes définies par les équations (4.34) et (4.35), on peut montrer que si l'on a :

$$\hat{RSB}_{post}(p,k) > \frac{1}{2\beta}\left(1+2\beta+\sqrt{\frac{1+3\beta}{1-\beta}}\right) \tag{4.36}$$

alors l'approche TSNR délivre un RSB supérieur à celui de l'approche DD. Dans le cas classique où $\beta = 0{,}98$ ce seuil est égal à environ 9,4dB. Donc, si une composante de parole apparaît subitement à la trame p, augmentant ainsi le RSB *a posteriori*, l'approche TSNR permettra de compenser partiellement l'atténuation du signal de parole provoquée par l'approche DD. D'ailleurs, plus le RSB *a posteriori* est grand et plus le biais de l'approche DD est réduit comme illustré dans la figure 4.11. Dans ce cas, le RSB *a priori* estimé tend vers le RSB *a posteriori* (équivalent au RSB *instantané* pour les RSB importants) ce qui permet de supprimer le retard introduit par l'approche DD. Dans le cas

contraire, si la parole est aussi absente à la trame p, laissant le RSB *a posteriori* à un faible niveau, alors le RSB *a priori* estimé devient plus faible que dans le cas de l'approche DD ce qui permet de réduire encore plus le niveau de bruit musical. Il faut noter que, si pour des RSB importants le biais est complètement supprimé, celui-ci subsiste pour des RSB plus faibles (aux alentours du seuil de 9,4dB) mais qui correspondent tout de même à des composantes de parole.

La limite haute du nuage de points de la figure 4.11 correspond au cas où le RSB *a priori* est élevé (qu'il soit surestimé par l'approche DD ou non) à la trame $p-1$. Dans ce cas, à partir de l'équation (3.12), l'approximation suivante peut être réalisée [Cappé 1994] :

$$\hat{RSB}_{prio}^{DD}(p,k) \approx \beta \hat{RSB}_{inst}(p-1,k). \tag{4.37}$$

Ainsi, le gain spectral obtenu par l'approche DD (première passe de l'approche TSNR) devient :

$$G_{DD}(p,k) \approx \frac{\beta \hat{RSB}_{inst}(p-1,k)}{1+\beta \hat{RSB}_{inst}(p-1,k)}. \tag{4.38}$$

Or, si l'on considère que $\hat{RSB}_{inst}(p-1,k) \gg 1$ et que β est proche de 1, alors l'équation (4.38) se réduit à $G_{DD}(p,k) \approx 1$. En introduisant cette approximation dans l'équation (4.30), cela conduit à :

$$\hat{RSB}_{prio}^{TSNR}(p,k) \approx \hat{RSB}_{post}(p,k) \tag{4.39}$$

ce qui explique que la limite haute du nuage de points soit une droite. Pour les RSB *a posteriori* importants, l'équation (4.39) peut donc être approchée ainsi :

$$\hat{RSB}_{prio}^{TSNR}(p,k) \approx \hat{RSB}_{inst}(p,k) \tag{4.40}$$

ce qui correspond au comportement désiré qui est détaillé dans la partie 3.2.3. Le raffinement apporté par l'approche TSNR supprime donc les cas de surestimation du RSB *a priori* introduits par l'approche DD.

Pour conclure, la sous-estimation du RSB *a priori* due au retard de l'approche DD est supprimée pour les RSB importants, cependant elle est conservée pour les plus faibles RSB ce qui permet de limiter le niveau de bruit musical résiduel. La surestimation du RSB *a priori*, quant à elle, est aussi supprimée. L'approche TSNR permet ainsi de préserver les transitoires de la parole (attaques et extinctions de parole) et supprime l'effet de réverbération caractéristique de l'approche DD.

4.3.3 Illustration du comportement de l'approche TSNR

Le principe et une analyse de l'approche TSNR ont été proposés dans les deux parties précédentes 4.3.1 et 4.3.2. La figure 4.12 permet d'illustrer son comportement à partir du signal bruité décrit dans la partie 3.2.1. Le signal de la figure 4.12.(b) est obtenu en utilisant l'approche DD et constitue donc la référence. En comparant les figures (b) et (c), on remarque tout d'abord que le bruit musical résiduel de l'approche DD a été davantage réduit par la seconde passe de l'approche TSNR. La suppression de l'effet de réverbération s'avère difficile à faire ressortir sur un spectrogramme. Toutefois, le zoom de la figure 4.13, où le contraste a été forcé, permet de voir que le traînage de l'approche DD est limité par l'approche TSNR. On peut souligner que bien que cette amélioration soit difficile à faire ressortir sur un spectrogramme, elle est malgré tout bien perceptible.

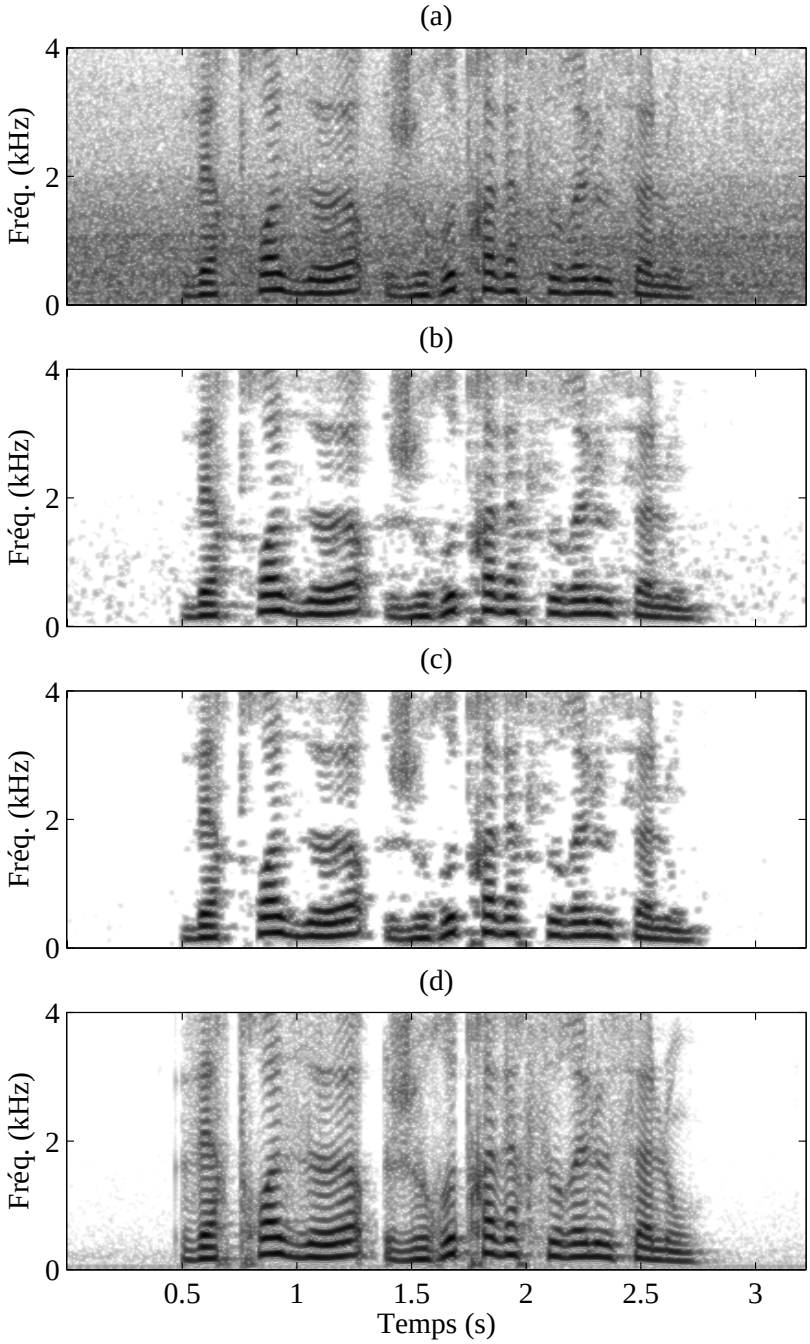

FIG. 4.12 – *Spectres d'amplitude pour (a) le signal de parole bruité, (b) le signal de parole restauré par l'approche DD, (c) le signal de parole restauré par l'approche TSNR et (d) le signal de parole propre.*

4.3.4 Autres approches proposées dans la littérature

On peut trouver dans la littérature des approches dont le but est le même que celui de l'approche TSNR, à savoir supprimer le biais de l'estimateur DD. Ainsi, l'approche proposée dans

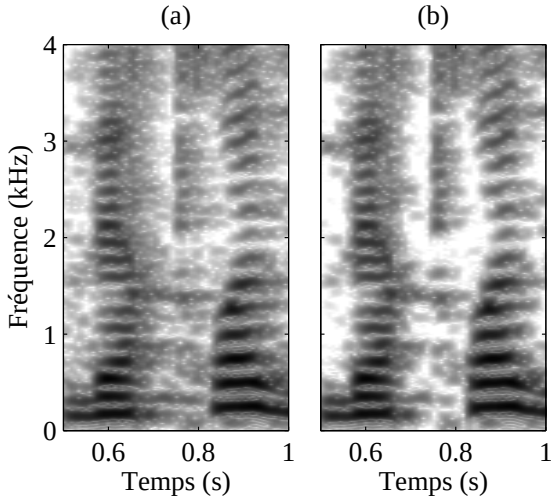

FIG. 4.13 – *Zoom sur une partie du signal de la figure 4.12. (a) Approche DD et (b) approche TSNR.*

[Beaugeant 1999a] consiste à rendre le coefficient β de l'estimateur DD adaptatif de façon à s'adapter au type de signal. Ainsi, en présence de parole β tend vers une valeur faible pour limiter la distorsion de la parole et en présence de bruit seul, β tend vers 1 de façon à supprimer le bruit musical. Le problème vient de la technique utilisée pour calculer le paramètre β. Deux approches sont proposées, l'une basée sur le RSB et l'autre sur le taux de variation spectrale [Quatieri 1997]. Mais dans le premier cas, on note une remontée audible du niveau de bruit musical pendant la parole et, dans le second, ce phénomène a lieu durant les phases d'extinction du signal de parole. Une autre approche proposée dans [Cohen 2004] consiste à modifier l'estimateur DD de façon à prendre en compte la corrélation temporelle entre les trames successives. Deux approches sont proposées, l'une "causale" et l'autre "non causale". Cette dernière permet d'atteindre des performances équivalentes à l'approche TSNR mais avec un retard supplémentaire important dans la mesure où la connaissance des trois trames futures est nécessaire.

4.4 Sélection des composantes fiables du RSB *a posteriori*

La technique proposée ci-après s'inscrit dans une démarche concurrente de l'approche TSNR, le but étant le même c'est-à-dire de supprimer le biais de l'estimateur de l'approche DD. L'approche TSNR permet de limiter ce biais mais ne le supprime pas complètement pour les RSB relativement faibles (*cf.* partie 4.3.2). Pour atteindre le comportement idéal d'un estimateur du RSB, défini dans la partie 3.2.3, nous proposons donc de sélectionner et d'utiliser uniquement les composantes fiables du RSB *a posteriori*. Par fiables on entend qui génèrent des composantes de parole (après filtrage) et, par extension, les composantes non fiables sont celles qui génèrent du bruit musical.

4.4.1 Principe et analyse

Dans la partie 3.2.6, on a vu que l'estimateur du RSB *a posteriori* est plus intéressant pour estimer les composantes de parole que l'estimateur du RSB *a priori* de l'approche DD. **Une stratégie judicieuse est donc de déterminer quand il est possible d'utiliser directement le RSB *a posteriori* et quand cette utilisation risque d'engendrer du bruit musical.** Pour réussir à sélectionner les composantes fiables du RSB *a posteriori*, nous proposons de séparer le nuage de points défini par le couple $(\hat{RSB}_{post}, \hat{RSB}^{DD}_{prio})$ en utilisant deux seuils. À partir d'un seuil η appliqué au RSB *a priori*, il est possible d'en déduire un seuil δ pour le RSB *a posteriori* en utilisant l'équation (3.9). La relation entre ces deux seuils dépend donc du paramètre de phase $\alpha(p,k)$. Ainsi, comme illustré sur la figure 4.14, le nuage de points va être divisé en quatre quadrants. Nous proposons de choisir $\alpha(p,k) = \pi$ car, pour un η fixé, cette valeur conduit au plus petit seuil possible pour δ ce qui permet donc de préserver les composantes du RSB *a posteriori* correspondant à de la parole quelle que soit la différence de phase entre la parole et le bruit. Ce choix est naturel étant donné qu'il est impossible d'estimer cette différence de phase $\alpha(p,k)$ (*cf.* partie 3.4), cependant, toute autre valeur peut être choisie. Il est possible d'obtenir un signal de sortie en utilisant les composantes du RSB *a posteriori*

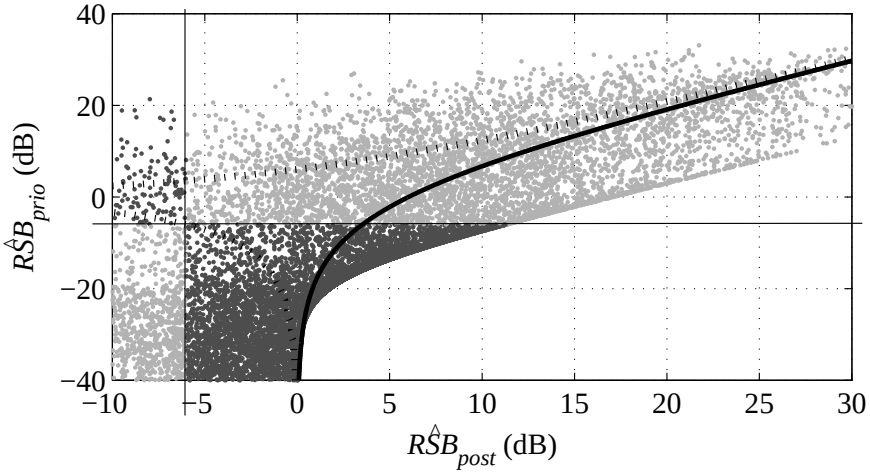

FIG. 4.14 – *Division du nuage de points défini par le couple $(\hat{RSB}_{post}, \hat{RSB}^{DD}_{prio})$ en quatre quadrants (deux en points foncés et les deux autres en points clairs) en utilisant deux seuils portant sur le $\hat{RSB}_{post}$ et le $\hat{RSB}^{DD}_{prio}$. Les deux courbes matérialisent l'équation (3.9) dans le cas où $\alpha(p,k) = 0$ (trait fort) et $\alpha(p,k) = \pi$ (pointillé).*

contenues dans chacun des quadrants. Dans le cas présenté, un filtre pseudo-Wiener (*cf.* partie 2.4.1.3) a été utilisé pour obtenir les signaux de sortie correspondant à chaque quadrant :

$$G_W(p,k) = 1 - \frac{1}{RSB_{post}(p,k)}. \tag{4.41}$$

Des écoutes informelles ont ensuite permis de régler les seuils η et δ de façon à ce qu'une bonne classification soit réalisée. La valeur ainsi obtenue pour η est de -6dB ce qui correspond pour δ à une valeur d'environ -6dB également. Ce choix est illustré par la figure 4.14. Ainsi, les points foncés du quadrant droit génèrent du bruit musical d'un niveau important et les points clairs et foncés des deux

quadrants gauche génèrent des composantes de signal très faibles et inaudibles qui sont donc inutiles. Finalement, les points clairs du quadrant droit peuvent être classifiés comme générant uniquement des composantes de parole, *i.e.* sans bruit musical. **Une classification efficace peut donc être obtenue en exploitant les comportements complémentaires des estimateurs du RSB *a posteriori* et *a priori*.**

En se basant sur cette classification, il est possible de réestimer le RSB *a posteriori* en sélectionnant seulement ses composantes fiables et en les utilisant pour calculer le gain spectral. Cette approche appelée "Reliable Features Selection Noise Reduction" ou RFSNR [Plapous 2005b] se résume en quatre étapes :

1 : Les RSB *a posteriori* et *a priori* sont respectivement calculés à partir des équations (3.11) et (3.12) de l'estimateur DD.

2 : Le RSB *a posteriori* est alors réestimé en utilisant l'équation suivante :

$$\hat{RSB}_{post}^{seuil}(p,k) = \begin{cases} \hat{RSB}_{post}(p,k) & \text{si } \hat{RSB}_{post}(p,k) \geq \delta \\ & \text{et } \hat{RSB}_{prio}^{DD}(p,k) \geq \eta, \\ 1 & \text{sinon,} \end{cases} \tag{4.42}$$

où l'exposant *seuil* le différencie du RSB *a posteriori* classique de l'équation (3.11).

3 : Le RSB réestimé, $\hat{RSB}_{post}^{seuil}(p,k)$, est utilisé directement pour calculer le gain spectral. On peut bien sûr utiliser un gain faisant intervenir le RSB *a priori* (*cf.* l'approche MMSE STSA présentée dans la partie 2.4.2.2 par exemple) mais l'idéal est justement de n'utiliser que le RSB *a posteriori* réestimé pour ne pas subir l'influence du biais de l'estimateur DD. Ainsi, sans perte de généralité, le filtre pseudo-Wiener 2.4.1.3 est choisi :

$$G_{RFSNR}(p,k) = 1 - \frac{1}{\hat{RSB}_{post}^{seuil}(p,k)}. \tag{4.43}$$

Ce gain est ensuite appliqué au spectre de parole bruitée :

$$\hat{S}_{RFSNR}(p,k) = G_{RFSNR}(p,k)X(p,k). \tag{4.44}$$

4 : Un autre gain spectral, $G_{DD}(p,k)$, est calculé à partir des RSB *a posteriori* et *a priori* calculés à l'étape 1 et est utilisé pour obtenir le signal $\hat{S}_{DD}(p,k)$ requis à l'étape 1 pour les calculs de la trame suivante. C'est en fait ce qui est réalisé classiquement dans l'approche DD (*cf.* partie 3.2.5.1).

La figure 4.15 représente le schéma de principe de cette approche.

On peut noter que si l'on choisit de conserver les deux quadrants de droite sur la figure 4.14 lors de la réestimation du RSB *a posteriori*, cela correspondrait au cas où un seuillage est appliqué seulement au RSB *a posteriori* :

$$\hat{RSB}_{post}(p,k) \geq \delta. \tag{4.45}$$

On retrouve d'ailleurs une telle condition dans l'équation (2.18) de la partie 2.4.1.2 consacrée à la soustraction spectrale généralisée :

$$\left(\hat{RSB}_{post}^{\gamma}(p,k) - \alpha\right)^{\frac{1}{\gamma}} \geq \beta \tag{4.46}$$

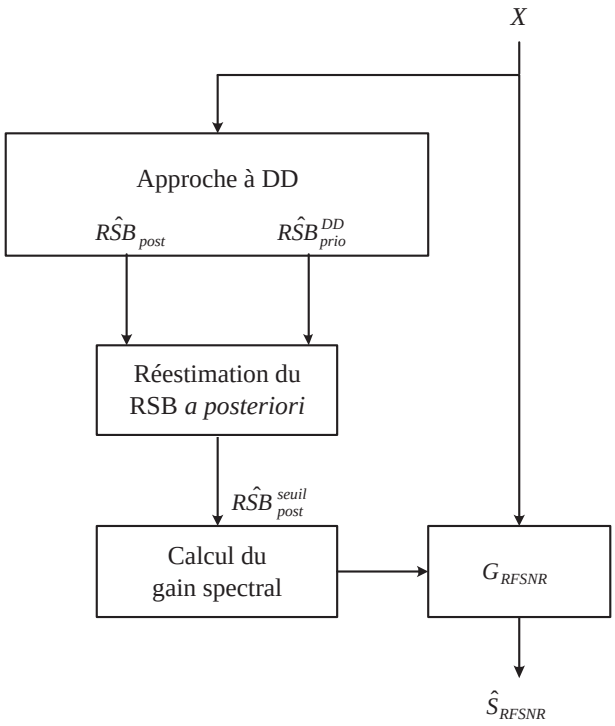

FIG. 4.15 – *Schéma du principe général de l'approche RFSNR.*

dans le cas particulier où les paramètres sont les suivants : $\alpha = \delta$, $\beta = 0$ (à ne pas confondre avec le paramètre de l'approche DD) et $\gamma = 1$. Ainsi, pour supprimer complètement le bruit musical (points foncés du quadrant droit) il faudrait utiliser un seuil d'une valeur d'environ 10dB mais alors toutes les composantes de parole correspondant aux points clairs compris entre les valeurs de -6 à 10dB (sur l'axe des abscisses) seraient également supprimées, engendrant ainsi beaucoup de dégradations. **Finalement, le fait d'utiliser deux seuils dans l'équation (4.42) permet de contourner ce problème et de préserver les composantes correspondant à de la parole tout en supprimant celles qui génèrent du bruit musical.**

4.4.2 Illustration du comportement de l'approche RFSNR

Dans cet exemple, le gain spectral choisi pour les approches DD et RFSNR est respectivement le filtre de Wiener et pseudo-Wiener (*cf.* partie 2.4.1.3). Le signal bruité, dont le spectre d'amplitude est représenté par la figure 4.16.(a), est identique à celui présenté dans la partie 3.2.1 (bruit de voiture avec un RSB de 12dB). Le signal de la figure 4.16.(b), qui est exempt de bruit musical, est généré en utilisant seulement les points clairs du quadrant droit de la figure 4.14 ce qui confirme l'efficacité de la réestimation du RSB *a posteriori* proposée dans l'équation (4.42). De plus, les composantes de parole sont restaurées à partir des composantes fiables du RSB *a posteriori* et ne souffrent donc pas du biais de l'estimateur DD. Ainsi, l'effet de réverbération caractéristique de cette approche est supprimé. Le signal de la figure 4.16.(c) est généré avec les points des 3 autres quadrants de la figure 4.14 ; cet important niveau de bruit musical serait celui obtenu en utilisant directement le RSB

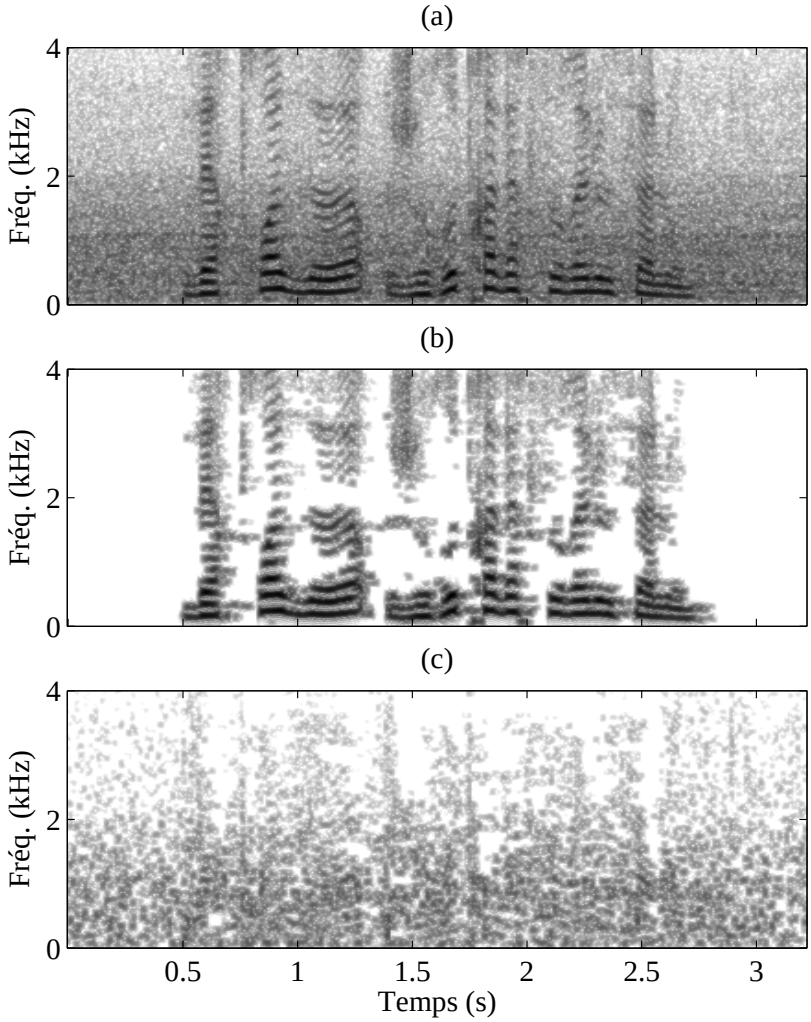

FIG. 4.16 – *Spectres d'amplitude pour (a) le signal de parole bruité, (b) le signal de parole restauré par l'approche RFSNR et (c) le bruit musical supprimé en utilisant cette approche.*

a posteriori de l'équation (3.11) pour restaurer le signal de parole. Toutefois, on peut préciser qu'à l'écoute les différences avec l'approche TSNR sont peu perceptibles (ce point sera détaillé dans la partie 6.3.5 consacrée aux résultats). Ceci est dû au fait que l'amélioration de l'approche RFSNR porte essentiellement sur les plus faibles composantes de parole.

4.5 Régénération des harmoniques de la parole

Dans les approches TSNR ou RFSNR, présentées dans les parties 4.3 et 4.4, l'estimateur du RSB a été amélioré permettant ainsi de limiter les distorsions du signal de parole restauré. Cependant, ces approches engendrent tout de même certaines dégradations qui sont essentiellement dues aux erreurs d'estimations de la DSP du bruit et à l'impact de la phase (*cf.* parties 3.3 et 3.4). Étant donné qu'en moyenne 80% des sons prononcés sont voisés, ces dégradations correspondent très souvent à

des distorsions harmoniques, c'est-à-dire que certaines harmoniques (généralement celles qui sont le plus affectées par le bruit) sont détruites ou fortement dégradées. **Nous proposons donc de tirer avantage de la structure harmonique des composantes voisées de la parole pour éviter ce type de distorsion. Pour ce faire, le signal distordu est traité (dans le domaine temporel) par une fonction non-linéaire capable de régénérer les harmoniques manquantes.** Le signal artificiel ainsi obtenu est alors mis à profit pour réestimer le RSB *a priori* qui sera utilisé pour calculer un gain spectral à même de préserver les harmoniques détruites par la majorité des approches de la littérature, y compris les approches TSNR et RFSNR. Cette approche originale est appelée HRNR pour "Harmonic Regeneration Noise Reduction" [Plapous 2005a, Plapous à paraître].

4.5.1 Principe de la régénération d'harmonicité

Considérons un signal restauré $\hat{s}(t)$ obtenu par une technique de réduction de bruit qualifiée de classique (dans le sens où elle introduit de la distorsion harmonique). Dans ce cas, une solution simple et efficace pour restaurer les harmoniques dégradées de ce signal de parole consiste à lui appliquer une fonction non-linéaire NL (*e.g.* valeur absolue, minimum ou maximum par rapport à un seuil, *etc.*) dans le domaine temporel. Soit $s_{harmo}(t)$ le signal artificiel ainsi restauré par application de cette fonction non-linéaire :

$$s_{harmo}(t) = NL(\hat{s}(t)). \tag{4.47}$$

Les harmoniques restaurées du signal $s_{harmo}(t)$ sont créées aux mêmes positions que celles du signal de parole propre. Cette propriété très intéressante est assurée par le fait même qu'une non-linéarité dans le domaine temporel est utilisée pour régénérer les harmoniques (ce point sera détaillé dans la partie 4.5.2). La figure 4.17 permet d'illustrer le comportement typique de la non-linéarité et du même coup son intérêt. Un signal bruité (celui présenté dans la partie 3.2.1) est traité par la technique TSNR. En comparant les figures 4.17.(a) et (b), il est évident que l'approche TSNR détruit certaines harmoniques et que d'autres sont sévèrement dégradées. La figure 4.17.(c) représente le module du signal artificiel obtenu en utilisant l'équation (4.47) où la non-linéarité choisie est le redressement mono-alternance (*i.e.* le maximum par rapport à 0). Cette non-linéarité appliquée au signal $\hat{s}(t)$ a permis de restaurer les harmoniques détruites ou dégradées aux mêmes positions que celles du signal de parole. **Cependant, les amplitudes des harmoniques de ce signal artificiel sont biaisées et le module du signal $s_{harmo}(t)$ ne peut donc pas être utilisé directement comme estimation du module de la parole propre. Par contre, ce signal possède des informations très importantes et nous proposons de les exploiter en raffinant l'estimation du RSB *a priori* selon l'expression suivante** :

$$\hat{RSB}_{prio}^{HRNR}(p,k) = \frac{\rho(p,k)|\hat{S}(p,k)|^2 + (1-\rho(p,k))|S_{harmo}(p,k)|^2}{\hat{\gamma}_b(k)}. \tag{4.48}$$

Le paramètre de mixage $\rho(p,k)$ ($0 < \rho(p,k) < 1$) est utilisé pour réaliser un compromis (dépendant du temps et de la fréquence) entre la contribution du signal distordu $\hat{S}(p,k)$ et celle du signal artificiel $S_{harmo}(p,k)$ dans l'estimation du RSB *a priori* selon le comportement suivant :

– quand l'estimation de $\hat{S}(p,k)$ obtenue par l'approche classique est fiable, alors la régénération d'harmonicité est inutile et $\rho(p,k)$ doit être proche de 1,

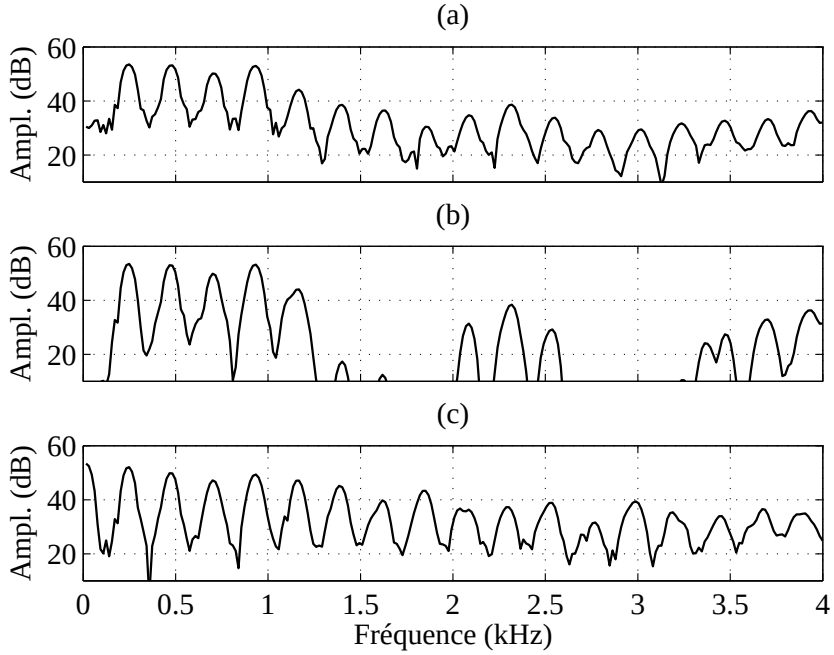

FIG. 4.17 – *Effet de la non-linéarité sur une trame voisée. (a) Module de la trame de parole propre ; (b) module de la trame de parole restaurée par l'approche TSNR ; (c) module de la trame du signal artificiel obtenu après la régénération des harmoniques.*

– quand l'estimation de $\hat{S}(p,k)$ n'est pas fiable, alors la régénération d'harmonicité est nécessaire pour corriger l'estimation du RSB *a priori* et $\rho(p,k)$ doit être proche de 0 (ou de toute autre constante pouvant dépendre de la fonction non-linéaire choisie).

Pour obtenir simplement ce comportement, nous proposons de choisir $\rho(p,k) = G(p,k)$, le gain $G(p,k)$ étant préalablement obtenu par la technique de réduction de bruit classique. Le paramètre $\rho(p,k)$ peut aussi être choisi constant afin d'obtenir un compromis entre les deux quantités $|\hat{S}(p,k)|^2$ et $|S_{harmo}(p,k)|^2$. De façon générale, le nouveau gain spectral $G_{HRNR}(p,k)$ à même de préserver les harmoniques du signal de parole s'exprime ainsi :

$$G_{HRNR}(p,k) = v\left(R\hat{S}B_{prio}^{HRNR}(p,k), R\hat{S}B_{post}(p,k)\right). \tag{4.49}$$

La fonction $v(.)$ peut être une des fonctions de gain décrites dans la partie 2.4. Sans perte de généralité, dans la suite le choix se porte sur un filtre de Wiener, et donc :

$$G_{HRNR}(p,k) = \frac{R\hat{S}B_{prio}^{HRNR}(p,k)}{1 + R\hat{S}B_{prio}^{HRNR}(p,k)}. \tag{4.50}$$

Finalement, le spectre de parole restauré est obtenu ainsi :

$$\hat{S}_{HRNR}(p,k) = G_{HRNR}(p,k)X(p,k). \tag{4.51}$$

Le schéma de principe de cette approche, définie par les équations (4.47) à (4.49), qui permet de préserver les harmoniques détruites par les approches classiques et donc de limiter les distorsions est représenté par la figure 4.18.

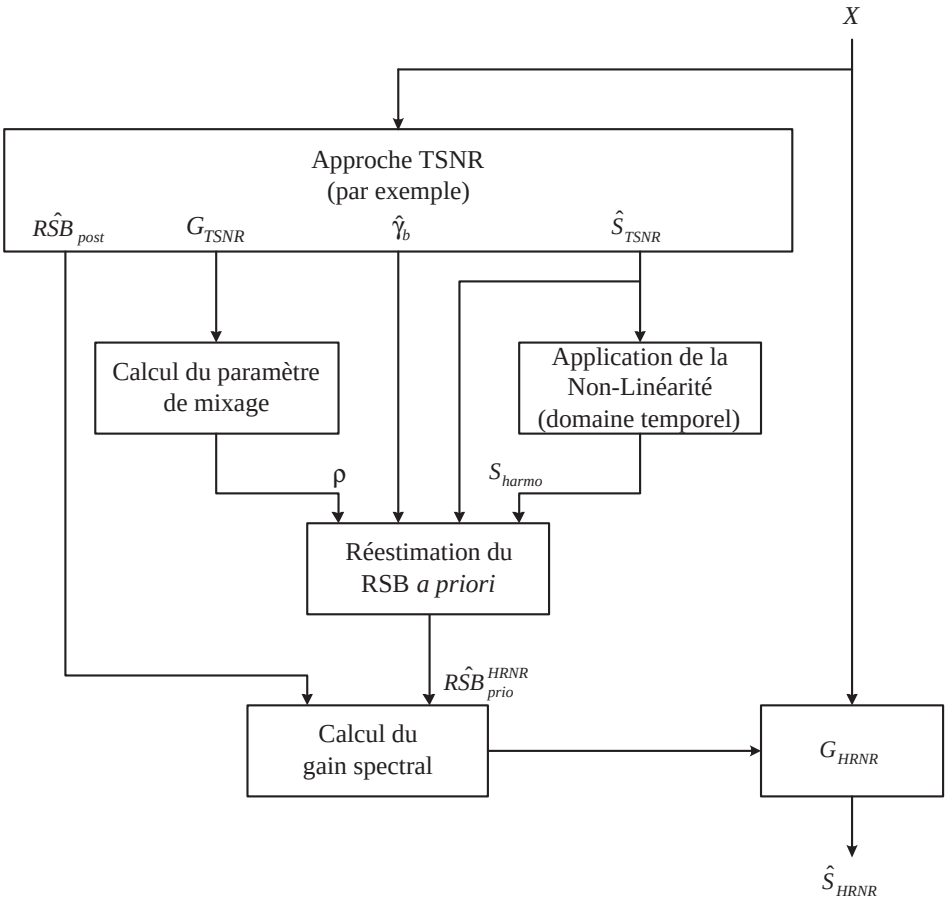

FIG. 4.18 – *Schéma du principe général de l'approche HRNR.*

4.5.2 Analyse théorique de la régénération d'harmonicité

Pour analyser le phénomène de régénération d'harmonicité nous allons nous focaliser, sans perte de généralité, sur une fonction non-linéaire en particulier à savoir le redressement mono-alternance (ou maximum par rapport à zéro). En remplaçant la fonction générique *NL* dans l'équation (4.47) par cette fonction non-linéaire, on obtient :

$$s_{harmo}(t) = \max(\hat{s}(t),0) = \hat{s}(t)p(\hat{s}(t)) \tag{4.52}$$

où la fonction $p(.)$, non-linéaire elle aussi, est définie par

$$p(u) = \begin{cases} 1 & \text{si } u > 0 \\ 0 & \text{si } u < 0. \end{cases} \tag{4.53}$$

La figure 4.19 représente une trame de parole voisée du signal $\hat{s}(t)$ ainsi que le signal $p(\hat{s}(t))$ correspondant qui a été mis à l'échelle pour une question de lisibilité. Ce signal $p(\hat{s}(t))$ correspond à la répétition d'un motif élémentaire avec une période T identique à celle du signal de parole voisé $\hat{s}(t)$. En supposant que le signal de parole est stationnaire (hypothèse non réaliste mais qui permet

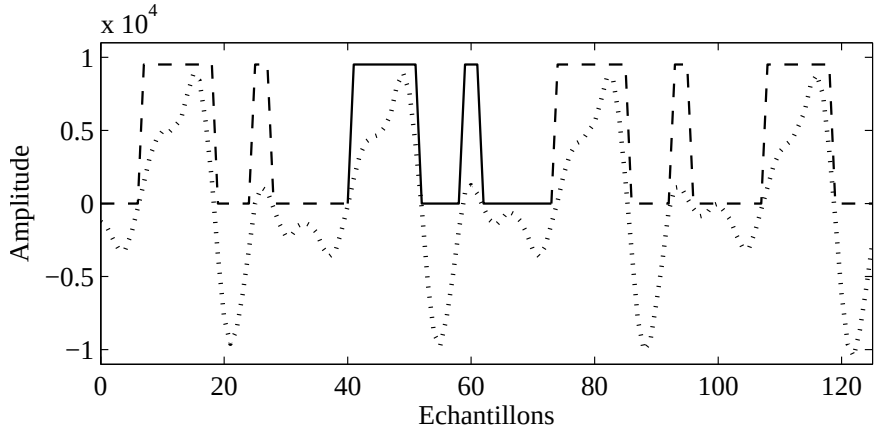

FIG. 4.19 – *Trame de parole voisée $\hat{s}(t)$ (pointillé) et signal (mis à l'échelle) $p(\hat{s}(t))$ correspondant (tirets). Motif élémentaire répété (trait plein).*

d'obtenir des relations exploitables aussi dans le cas quasi-stationnaire), la TF du signal périodique $p(\hat{s}(t))$ est une version échantillonnée (avec un pas de $\frac{1}{T}$) de la TF du motif élémentaire répété :

$$TF(p(\hat{s}(t))) = \frac{1}{T}\sum_{m=-\infty}^{\infty} R\left(\frac{m}{T}\right)\delta\left(f-\frac{m}{T}\right) \tag{4.54}$$

où δ représente la distribution de Dirac, f le domaine des fréquences continues et $R(\frac{m}{T})$ la TF du motif élémentaire prise à la fréquence $\frac{m}{T}$. Étant donné que le pas d'échantillonnage est égal à la fréquence fondamentale du motif élémentaire, l'échantillonnage coïncide parfaitement avec la position des harmoniques de la TF de ce motif. Finalement, en utilisant l'équation (4.52), la TF du signal $s_{harmo}(t)$ s'écrit :

$$TF(s_{harmo}(t)) = TF(\hat{s}(t)) * \frac{e^{-j\theta}}{T}\sum_{m=-\infty}^{\infty} R\left(\frac{m}{T}\right)\delta\left(f-\frac{m}{T}\right) \tag{4.55}$$

où $*$ représente l'opérateur de convolution et θ est la phase à l'origine. La figure 4.20 permet d'illustrer l'équation (4.55) dans le cas réaliste où le signal de parole est supposé quasi-stationnaire. Les trois figures de la colonne de gauche (a), (b) et (c) représentent respectivement une trame des signaux $\hat{s}(t)$, $p(\hat{s}(t))$ et $s_{harmo}(t)$ qui interviennent dans l'équation (4.52). Les trois figures en vis-à-vis dans la colonne de droite (d), (e) et (f) représentent leurs équivalents dans le domaine fréquentiel. À court terme et en temps discret, le peigne harmonique idéal de l'équation (4.54) subit l'influence de la fenêtre d'analyse et est aussi sujet au repliement fréquentiel, la bande fréquentielle du signal $p(\hat{s}(t))$ étant en principe infinie. Le spectre harmonique résultant, représenté par la figure 4.20.(e), n'est donc pas à proprement parler un peigne harmonique. Dans le domaine temporel, le signal restauré, $s_{harmo}(t)$ (c), résulte de la multiplication du signal de parole dégradé $\hat{s}(t)$ (a) et du signal non-linéaire $p(\hat{s}(t))$ (b). Dans le domaine fréquentiel, le spectre du signal restauré (f) résulte donc de la convolution du spectre du signal dégradé (d) et du spectre harmonique (e). Ces deux spectres ayant la même fréquence fondamentale (celle du signal voisé $\hat{s}(t)$), des harmoniques vont donc être régénérées par le biais de la convolution là où celles-ci étaient supprimées ou dégradées (comparaison entre les figures (d) et (f)). On comprend bien que les amplitudes des harmoniques régénérées par ce procédé ne peuvent pas être contrôlées, le peigne harmonique (e) dépendant du signal de parole dégradé (d). **Les avantages de**

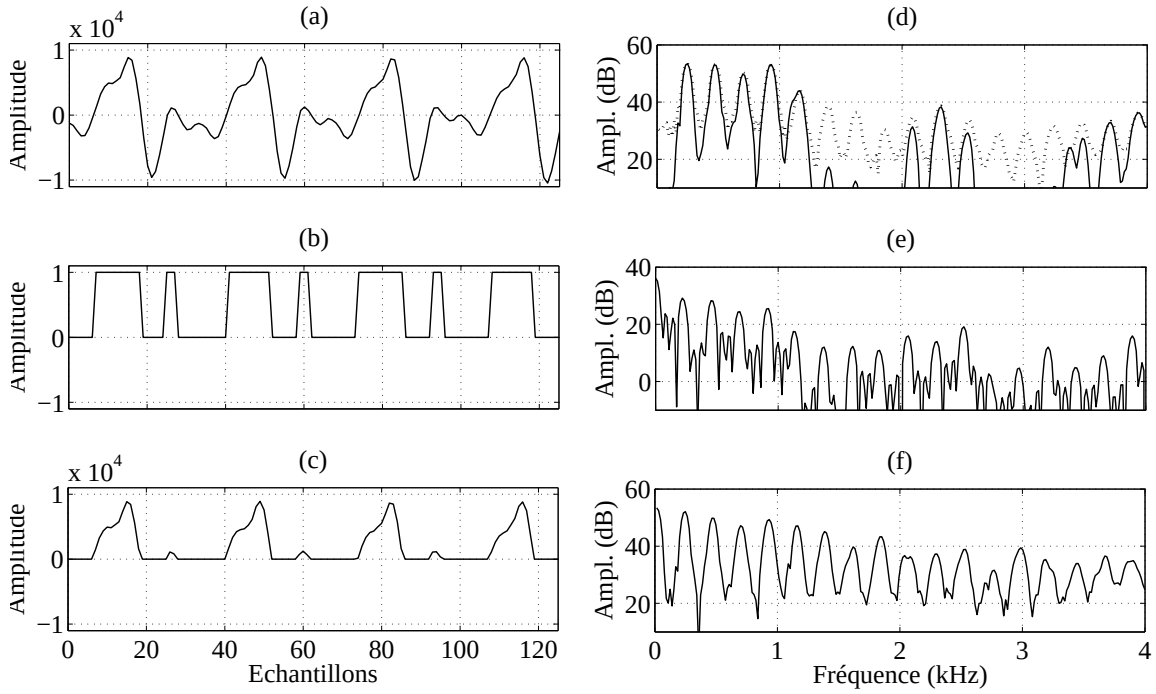

FIG. 4.20 – *Dans la colonne de gauche : une trame de parole voisée $\hat{s}(t)$ (a), le signal $p(\hat{s}(t))$ correspondant (b) et le signal $s_{harmo}(t)$ résultant de la non-linéarité (c). Dans la colonne de droite : leurs équivalents dans le domaine fréquentiel (d), (e) et (f). Le module du signal propre a été rajouté (en pointillé) sur la figure (d) pour identifier les zones où les harmoniques ont été détruites. L'échelle de la figure (e) est différente de celle des figures (d) et (f).*

cette méthode sont sa simplicité et la robustesse avec lesquelles les harmoniques sont régénérées aux positions désirées. De plus, on peut préciser que, étant donné que l'enveloppe du spectre $TF(p(\hat{s}(t)))$ a tendance à décroître quand $|m|$ augmente (dans l'équation (4.54)), une harmonique n'est ainsi régénérée qu'en utilisant (majoritairement) les informations des quelques harmoniques voisines. Ceci assure donc une certaine cohérence dans la façon dont les harmoniques sont régénérées. Le revers de cette propriété est que si trop d'harmoniques sont manquantes alors la régénération d'harmonicité perd de son efficacité étant donné qu'il n'y a plus assez d'information pour assurer la reconstruction.

Le cas qui vient d'être traité concernait la régénération d'harmonicité d'une trame de parole complètement voisée. Cependant, en moyenne 20% des sons prononcés ne le sont pas et il est donc important de connaître le comportement du processus de régénération d'harmonicité lorsque la trame de parole est non voisée ou seulement partiellement voisée. Dans un premier temps, considérons un cas hybride où la partie basse du spectre de parole est voisée et la partie haute non voisée. La TF du signal $p(\hat{s}(t))$ exprimée par l'équation (4.54) sera donc, comme dans le cas complètement voisé, un peigne harmonique dont la fréquence fondamentale est imposée par la partie basse et voisée du spectre du signal $\hat{s}(t)$. Tout se passe donc comme dans le cas complètement voisé et étant donné que l'enveloppe du peigne harmonique décroît avec la fréquence, chaque bande de fréquence régénérée

(4.55) sera obtenue à partir des composantes de son voisinage (dans le spectre du signal $\hat{s}(t)$). Ainsi, les parties non voisées seront reconstruites majoritairement à partir des composantes non voisées du spectre du signal $\hat{s}(t)$. De même, les harmoniques de la partie basse du spectre ne sont utilisées que dans la reconstruction des parties voisées du spectre. Considérons maintenant le cas où tout le spectre de parole est non voisé. Un signal non voisé s'apparentant à un bruit, le signal $p(\hat{s}(t))$ ne présente pas de périodicité et son spectre ne sera donc pas un peigne harmonique. Cependant, la convolution dans l'équation (4.55) entre ce spectre et le spectre non voisé de parole générera automatiquement un spectre non voisé. **Ces deux comportements assurent que les parties non voisées du spectre de parole ne sont pas dégradées par le processus de régénération d'harmonicité dans le sens où des harmoniques ne seront pas régénérées à la place de composantes non voisées.**

4.5.3 Illustration du comportement de l'approche HRNR

Le principe et une analyse de l'approche HRNR ont été proposés dans les deux parties précédentes 4.5.1 et 4.5.2. Nous proposons donc d'illustrer son comportement à partir d'un cas représentatif de signal bruité (*cf.* signal décrit dans la partie 3.2.1). Dans cet exemple, l'approche TSNR est choisie pour obtenir le signal qui sert de base à l'approche HRNR (et dont les harmoniques sont dégradées). En comparant les cas (b), (c) et (d) de la figure 4.21, il apparaît que de nombreuses harmoniques sont préservées en utilisant l'approche HRNR alors qu'elles sont complètement supprimées par l'approche TSNR. Cet exemple confirme donc qu'il est possible de se servir du caractère voisé de certaines composantes de la parole pour éviter les distorsions harmoniques inhérentes aux techniques de réduction de bruit classiques, que le bruit soit mal estimé ou que les harmoniques soient complètement détruites par celui-ci (bruit ajouté en opposition de phase par exemple). Des résultats plus complets et notamment un test subjectif formel seront donnés dans le chapitre 6 permettant ainsi de quantifier l'intérêt de l'approche HRNR.

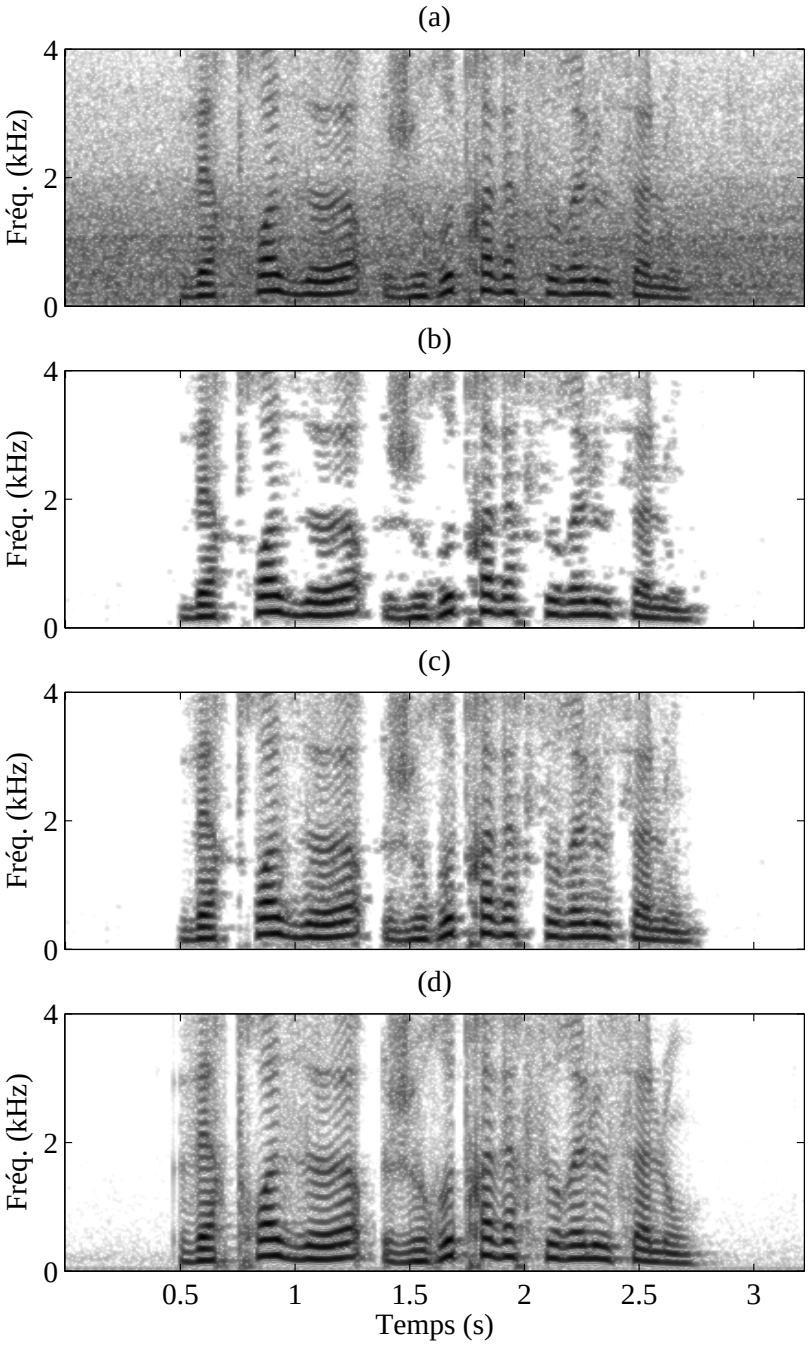

FIG. 4.21 – *Spectres d'amplitudes d'un signal de parole pour les différents cas suivants : (a) Signal de parole dégradé par du bruit de voiture avec un RSB de* 12*dB. (b) Signal de parole restauré par l'approche TSNR. (c) Signal de parole restauré par l'approche HRNR. (d) Signal de parole propre.*

4.6 Conclusion

Les nouvelles techniques de réduction de bruit proposées dans ce chapitre ont été analysées et leur comportement a également été illustré. Toutes ces approches sont de près ou de loin reliées ou basées sur l'approche DD qui constitue donc une référence. On a vu que les approches SG permettent, par l'utilisation de modèles adaptés aux signaux traités, de limiter de façon intéressante les distorsions du signal restauré. En particulier, le bruit musical est réduit et l'effet de réverbération est très limité.

Trois nouveaux estimateurs du RSB ont également été proposés dont deux permettent de supprimer (du point de vue subjectif) les défauts de l'approche DD. L'approche DDG est une généralisation de l'estimateur DD dans le sens où l'estimateur proposé permet de suivre les évolutions fréquentielles des composantes harmoniques de parole. Cependant, cette approche ne permet pas d'améliorer suffisamment l'approche DD pour s'en démarquer subjectivement. Par contre, les approches TSNR et RFSNR permettent de limiter (voire supprimer pour le RFSNR) le biais de l'estimateur decision-directed permettant ainsi de supprimer l'effet de réverbération de l'approche DD et de réduire davantage le niveau de bruit musical. Bien que basées sur deux concepts différents, ces techniques apportent des résultats subjectifs équivalents.

L'approche HRNR permet de dépasser les limitations liées aux problèmes d'estimation de la DSP du bruit et de la phase qui engendrent une distorsion harmonique. En effet, il est possible de régénérer les harmoniques détruites par des approches classiques (y compris TSNR et RFSNR) en utilisant un traitement non-linéaire du signal distordu. Cette approche permet donc de limiter de façon significative la distorsion du signal de parole.

Afin de compléter l'analyse de ces techniques et de quantifier précisément leurs performances, le chapitre 6 livre les résultats obtenus à partir d'un corpus complet de signaux bruités où de nombreuses conditions sont représentées. Le chapitre suivant est quant à lui consacré à la mise en œuvre des techniques de réduction de bruit qui doit être réalisée avec soin de façon à éviter que certains artefacts ne viennent dégrader la qualité du signal restauré.

Références

[Beaugeant 1999a] C. Beaugeant, et P. Scalart, "Noise Reduction Using Perceptual Spectral Change," *Eurospeech*, Budapest, Hongrie, Septembre 1999.

[Breithaupt 2003] C. Breithaupt, et R. Martin, "MMSE Estimation of Magnitude-Squared DFT Coefficients with Supergaussian Priors," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Vol. 1, pp. 896–899, 2003.

[Cappé 1994] O. Cappé, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor," *IEEE Trans. Speech Audio Processing*, Vol. 2, No. 2, pp. 345–349, Avril 1994.

[Cohen 2004] I. Cohen, "On the Decision-Directed Estimation Approach of Ephraim and Malah," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Montréal, Canada, Vol. 1, pp. 293–296, Mai 2004.

[Gradshteyn 1994] I. S. Gradshteyn, et I. M. Ryzhik, "Table of Integrals, Series, and Products," *Academic press, 5ème édition*, 1994.

[Guédon 2002] L. Guédon, "Mise en œuvre de Nouvelles Hypothèses dans les Algorithmes de Réduction de Bruit par Atténuation Spectrale," *Document interne FT R&D*, Août 2002.

[Martin 2002] R. Martin, "Speech Enhancement Using MMSE Short Time Spectral Estimation with Gamma Distributed Speech Priors," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Orlando, États-Unis, Vol. 1, pp. 253–256, Mai 2002.

[Plapous 2004] C. Plapous, C. Marro, L. Mauuary, et P. Scalart, "Two-Step Noise Reduction Technique," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Montréal, Canada, Vol. 1, pp. 289–292, Mai 2004.

[Plapous 2005a] C. Plapous, C. Marro, et P. Scalart, "Speech Enhancement using Harmonic Regeneration," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Philadelphie, États-Unis, Vol. 1, pp. 157–160, Mars 2005.

[Plapous 2005b] C. Plapous, C. Marro, et P. Scalart, "Reliable a Posteriori Signal-to-Noise Ratio Features Selection," *IEEE Work. Appl. of Signal Processing to Audio and Acoust.*, New Paltz, États-Unis, Octobre 2005.

[Plapous à paraître] C. Plapous, C. Marro, et P. Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement," *IEEE Trans. Speech Audio Processing*, accepté et à paraître.

[Quatieri 1997] T. F. Quatieri, et R. A. Baxter, "Noise Reduction Based on Spectral Change," *IEEE Work. Appl. of Signal Processing to Audio and Acoust.*, Mohonk, États-Unis, Octobre 1997.

Chapitre 5

Mise en œuvre

Nous avons choisi de consacrer un chapitre à la mise en œuvre car la qualité finale du signal restauré en est fortement dépendante. En effet, à quoi bon avoir de très bons estimateurs du signal de parole si le résultat final est entaché d'artefacts ("clics", rugosité, effet tuyau, *etc.*). Les "clics" sont généralement liés à une mauvaise réalisation de la synthèse du signal. Ainsi, la partie 5.1 détaille les différentes options qui existent pour la synthèse et les différentes contraintes qu'elles impliquent. Selon la méthode de synthèse retenue, un degré de liberté supplémentaire apparaît dans le choix de la fenêtre d'analyse. La partie 5.2 met en avant l'avantage apporté par le choix non conventionnel d'une fenêtre dissymétrique. Il est généralement nécessaire de masquer les distorsions du signal de parole en conservant une partie du bruit original. Ceci se justifie également d'un point de vue usage dans la mesure où l'ambiance sonore bruitée doit être réduite mais tout de même transmise car elle fait partie intégrante de la communication. La partie 5.3 expose une technique plus judicieuse qu'un classique seuillage du gain spectral. Un autre point important concerne les nombreux artefacts audibles qui peuvent être engendrés à cause d'une mauvaise gestion de la résolution fréquentielle du gain spectral. La partie 5.4 explique comment les supprimer ; néanmoins, cela se fait au prix d'un autre type de dégradation que l'approche psychoacoustique permet de supprimer efficacement (*cf.* partie 5.5). Finalement, la partie 5.6 met en avant une mise en œuvre particulière qui permet à la fois de limiter une grande partie des artefacts du signal restauré tout en autorisant une réduction de bruit importante.

5.1 Traitement par blocs

5.1.1 Principe

Le traitement par blocs se justifie par l'utilisation de la FFT que ce soit pour des raisons de complexité (*e.g.* filtrage rapide) ou que l'estimation de paramètres dans le domaine fréquentiel représente un passage obligé (*e.g.* réduction de bruit). Dans le cas qui nous intéresse, à savoir la réduction de bruit, le traitement par blocs est rendu nécessaire par le fait que l'analyse du signal et le calcul du filtre de réduction de bruit se font dans le domaine fréquentiel. De plus ce traitement par blocs ou trames est adapté au caractère quasi-stationnaire de la parole. Cette partie est inspirée du travail effectué dans [Guérin 2005].

Le traitement par blocs consiste à traiter les échantillons au rythme trame et se traduit généralement par un filtrage (convolution). Le filtre peut être invariant dans le temps et dans ce cas le traitement par blocs se justifie théoriquement. Par contre, dans le cadre de la réduction de bruit le filtre est amené à varier d'une trame à l'autre. Bien que dans ce cas la justification théorique ne tienne plus, en pratique le traitement par blocs peut être utilisé moyennant certaines précautions pour éviter les artefacts liés à la variabilité du filtre. Le filtrage peut être réalisé soit dans le domaine temporel par convolution soit dans le domaine fréquentiel par multiplication point à point. La figure 5.1 illustre le

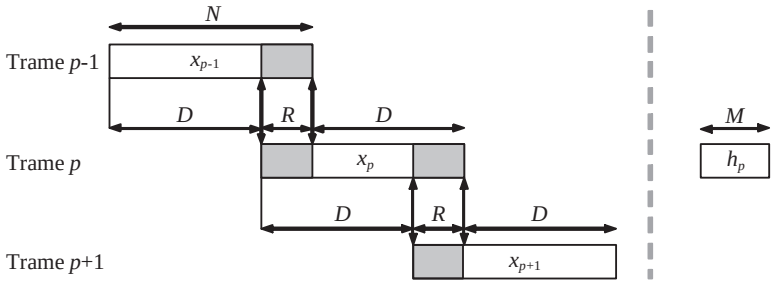

FIG. 5.1 – *Illustration du traitement par blocs et notations associées.*

principe du traitement par blocs et synthétise les notations utilisées : chaque trame x_p du signal d'entrée est constituée de N échantillons $x_p(n)$ avec $n \in [0,N-1]$ et p représente l'indice de la trame. Le support temporel de chaque trame x_p est $[0,N-1]$ selon la notation à référence temporelle glissante [Cappé 1993]. Le recouvrement entre les trames est noté R et donc $D = N - R$ représente le décalage à chaque nouvelle trame. La taille du filtre h_p associé à la trame x_p est notée M et ses coefficients sont notés $h_p(m)$ avec $m \in [0,M-1]$.

Dès l'instant où les trames se recouvrent ($R > 0$), l'utilisation de fenêtres d'analyse et de synthèse est indispensable. Les différents paramètres (choix des fenêtres et taux de recouvrement) doivent satisfaire à la contrainte de reconstruction parfaite, *i.e.* si aucun traitement n'est appliqué alors le signal de sortie reconstruit doit être identique au signal d'entrée.

5.1.2 Équivalence entre filtrage dans le domaine fréquentiel et temporel

Le filtrage dans le domaine fréquentiel correspond dans le domaine temporel à une convolution circulaire et non pas à une convolution linéaire classique. La convolution linéaire d'un vecteur x_p de taille N (taille du bloc) par un filtre h_p de taille M produit un vecteur y_p de taille $N+M-1$. Du fait du traitement par blocs, les $M-1$ premiers et derniers échantillons sont filtrés en incluant des zéros en début et fin du bloc. On n'obtient donc que $N-M+1$ échantillons "valides". Ceci est représenté sur la figure 5.2 dans le cadre "Filtrage temporel". Les parties triangulaires représentent les échantillons $y_p(n)$ filtrés avec un nombre restreint d'échantillons $x_p(n)$ ou de façon équivalente avec un filtre tronqué. Cette convolution peut être réalisée dans le domaine fréquentiel par multiplication point à point suivie d'une IFFT (*cf.* cadre "Filtrage fréquentiel"). Lors de cette opération apparaît le phénomène de repliement temporel (directement relié au fait que la convolution est alors circulaire) qui affecte les $M-1$ premiers échantillons. En effet, tout se passe comme si les $M-1$ derniers échantillons du

résultat de la convolution linéaire (triangle grisé) étaient repliés sur les $M-1$ premiers échantillons. C'est ce qui explique que le filtrage dans le domaine fréquentiel ne délivre que N échantillons alors que la convolution linéaire en délivre $N+M-1$.

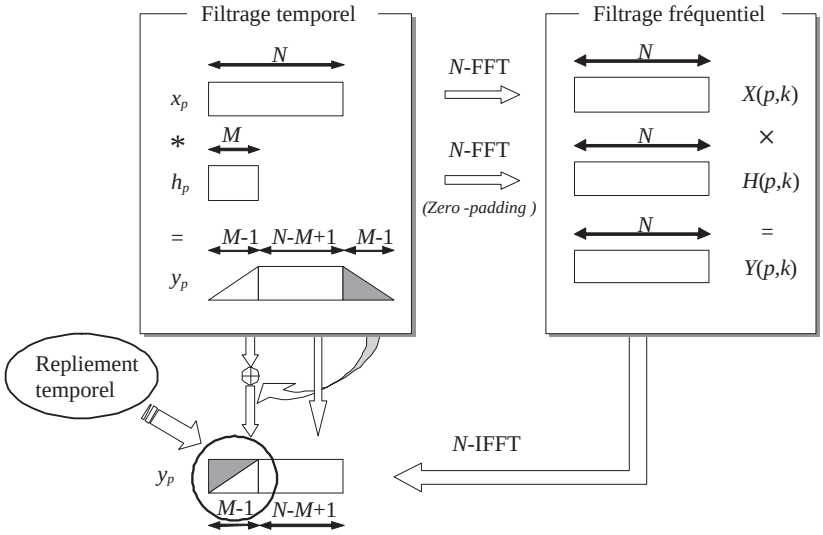

FIG. 5.2 – *Convolutions linéaire et circulaire.*

5.1.3 L'"overlap and save" ou OLS

5.1.3.1 Principe

La méthode OLS [Crochiere 1983, Kunt 1984] consiste à ne conserver que les échantillons valides à l'issue de l'opération de filtrage. Les autres échantillons sont considérés non valides car soit ils sont obtenus à partir d'un filtre tronqué (filtrage temporel) soit ils subissent le phénomène de repliement temporel (filtrage fréquentiel). Dans le cas de la Figure 5.2, on ne conservera donc que le bloc valide de $N-M+1$ échantillons filtrés.

Historiquement, l'OLS s'implémente dans le domaine fréquentiel (application de convolution rapide par FFT). L'utilisation de l'OLS pour la réduction de bruit nécessite cependant d'élargir le cadre classique. En effet, dans l'application qui nous intéresse, chaque trame est fenêtrée et les trames se recouvrent (avec un taux minimum de 50%). Donc étant donné que l'OLS ne conserve que les échantillons valides à l'issue du filtrage, il est impossible de satisfaire à la contrainte de reconstruction parfaite par sommation des trames filtrées, cette sommation impliquant des échantillons repliés et donc non valides. Par contre il est possible de contourner ce problème en filtrant directement les échantillons non fenêtrés. Nous allons donc détailler les implémentations fréquentielle et temporelle qui découlent de ce cadre particulier.

5.1.3.2 Implémentation dans le domaine fréquentiel

De fait, à l'analyse la trame de signal est pondérée par une fenêtre (*e.g.* fenêtre de Hanning) ce qui implique que pour effectuer la reconstruction selon le principe de l'OLS, il faut appliquer le filtre à la FFT du signal non fenêtré. Cette mise en œuvre exige donc le calcul d'une FFT supplémentaire. Comme le montre la figure 5.3, dans le domaine fréquentiel la technique OLS nécessite un recouvrement minimal de $M-1$ échantillons, soit $R \geq M-1$, afin de garantir une reconstruction parfaite. Les échantillons grisés subissent l'effet du repliement temporel et donc dans le cas où $R < M-1$, les échantillons en gris foncé ne peuvent être fournis sans distorsion. D'un point de vue pratique, si un recouvrement R faible est imposé pour une raison de coût de calcul (par exemple) il est possible de limiter le support temporel du filtre afin de respecter la condition $R \geq M-1$ (*cf.* partie 5.4). Toutefois, il est possible de supprimer cette contrainte au prix du calcul d'une FFT sur $D+M-1$ échantillons au lieu de $D+R=N$ (les échantillons supplémentaires sont récupérés de la trame précédente). En effet, dans ce cas le repliement n'affectera pas les D échantillons à fournir. Le retard introduit par l'implémentation fréquentielle de l'OLS est seulement de D échantillons et correspond uniquement à l'acquisition des nouveaux échantillons. On ne prend pas ici en compte le retard lié au filtre ni le temps de calcul nécessaire au filtrage, ces retards supplémentaires sont détaillés dans la partie 5.1.5.

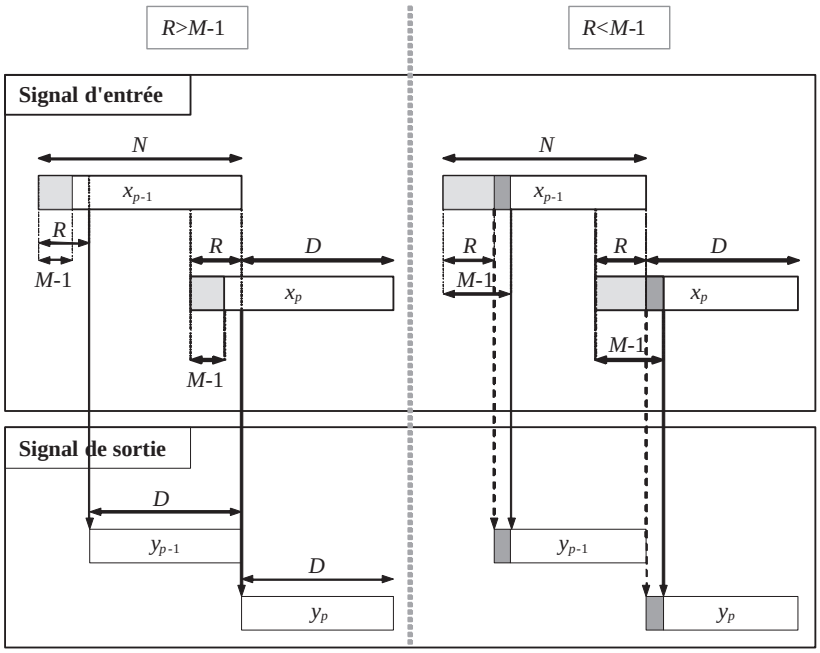

FIG. 5.3 – *Principe de l'OLS dans le domaine fréquentiel.*

5.1.3.3 Implémentation dans le domaine temporel

Il est aussi possible d'implémenter la méthode OLS dans le domaine temporel, l'approche est alors sensiblement différente. Dans le domaine temporel, la méthode OLS consiste simplement à filtrer les D nouveaux échantillons de la trame p par le filtre associé à cette même trame. Le filtre est

calculé à partir de la trame fenêtrée mais il est appliqué à des échantillons non fenêtrés comme dans le cas fréquentiel. Le principe de l'OLS temporel est illustré par la figure 5.4. Pour que les échantillons filtrés soient valides, le buffer utilisé pour le filtrage doit contenir les D nouveaux échantillons précédés des $M-1$ échantillons nécessaires au filtrage des $M-1$ premiers échantillons (le buffer est donc de taille $D+M-1$). Dans le cas où $R < M-1$, les échantillons peuvent être prélevés de la trame $p-1$, la contrainte $R \geq M-1$ de l'implémentation fréquentielle n'existe donc pas en temporel. La méthode OLS ne nécessite pas en elle-même de recouvrement (rendu cependant nécessaire par la réduction de bruit, point que nous verrons plus loin) et, dans ce cas limite, la taille du buffer utilisé pour le filtrage est de $N+M-1$. Le retard introduit par l'implémentation temporelle de l'OLS est identique au cas fréquentiel et est donc de D échantillons.

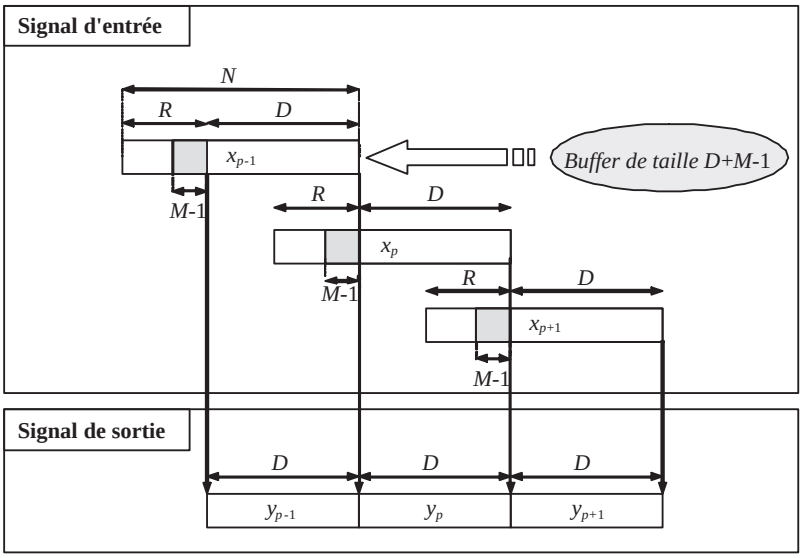

FIG. 5.4 – *Principe de l'OLS dans le domaine temporel.*

5.1.3.4 Implémentation en sous-trames pour la réduction des "clics"

Si le filtre à appliquer est invariant dans le temps alors la méthode OLS donne des résultats exacts, que ce soit dans le domaine temporel ou fréquentiel. Cependant, lorsque le filtre varie d'une trame à l'autre (*e.g.* réduction de bruit), cette technique peut produire des artefacts audibles sous la forme de "clics" qui apparaissent au rythme trame et qui sont dus à de fortes variations entre les filtres successifs (généralement durant l'activité vocale). **Pour éviter ce type d'artefact, la transition entre les filtres des trames successives doit être rendue moins abrupte.** Une façon d'assurer une transition souple entre deux filtres successifs consiste à calculer un nouveau filtre par barycentre des filtres des trames p et $p-1$, et de faire varier la pondération au cours des échantillons comme illustré dans la figure 5.5. Le coefficient de pondération $\alpha(n)$ avec $n \in [R,N-1]$ (selon la référence temporelle glissante définie par la trame x_p) doit être croissant et doit respecter les conditions initiale et finale suivantes :

$$\begin{cases} \alpha(R) = 0 \\ \alpha(N-1) = 1. \end{cases} \tag{5.1}$$

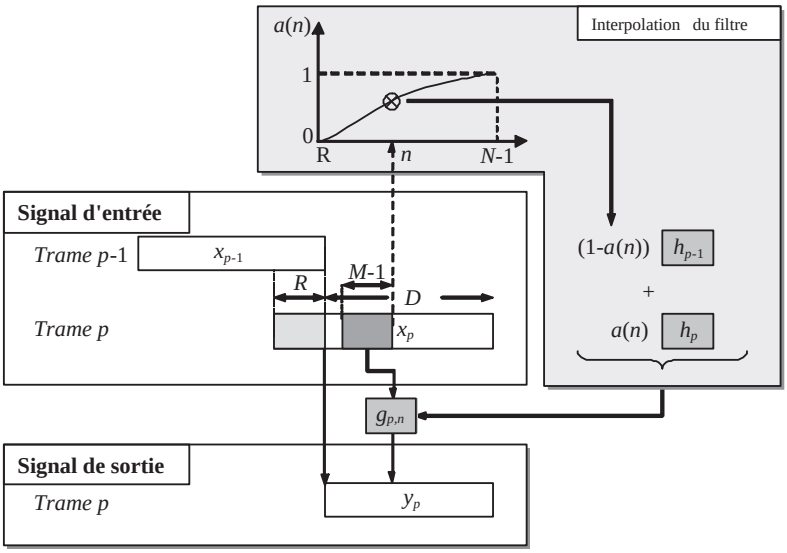

FIG. 5.5 – *Réduction des "clics" par interpolation de filtres.*

L'implémentation dans le domaine temporel s'appuie sur la Figure 5.5. Pour chaque nouvel échantillon $x_p(n)$ à filtrer, avec $n \in [R,N-1]$, un nouveau filtre est calculé par la formule :

$$g_{p,n}(m) = \alpha(n)h_p(m) + (1-\alpha(n))h_{p-1}(m), \tag{5.2}$$

où $g_{p,n}$ signifie que le filtre g est associé à l'échantillon n de la trame p et $m \in [0,M-1]$ désigne l'indice relatif aux coefficients du filtre. L'échantillon filtré $y_p(n)$ est alors obtenu par convolution :

$$y_p(n) = \sum_{m=0}^{M-1} x_p(n-m)g_{p,n}(m). \tag{5.3}$$

Pour la loi définissant l'évolution du coefficient de pondération $\alpha(n)$ de multiples choix sont envisageables, il est toutefois possible de différencier deux catégories de lois comme illustré sur la figure 5.6 :

- Loi continue : implémentation par échantillons. Dans ce cas, $\alpha(n)$ varie à chaque nouvel échantillon. Cette loi peut être linéaire (C^0), voire suivre la forme d'une fenêtre de Hanning (C^1), par exemple. Dans ce dernier cas, la continuité de la dérivée améliore l'interpolation et réduit les discontinuités entre les filtres interpolés. Néanmoins, la différence reste inaudible.
- Loi discontinue : implémentation par sous-trames. Afin de réduire la complexité ajoutée par l'interpolation du filtre à chaque échantillon, $\alpha(n)$ peut ne varier qu'un nombre restreint de fois ce qui revient à faire une implémentation par morceaux ou sous-trames.

Ces approches par interpolation de filtre pour supprimer les "clics" liés à la mise en œuvre peuvent avoir un impact sur la qualité du signal. En effet, comme le montre la figure 5.5 et l'équation (5.2),

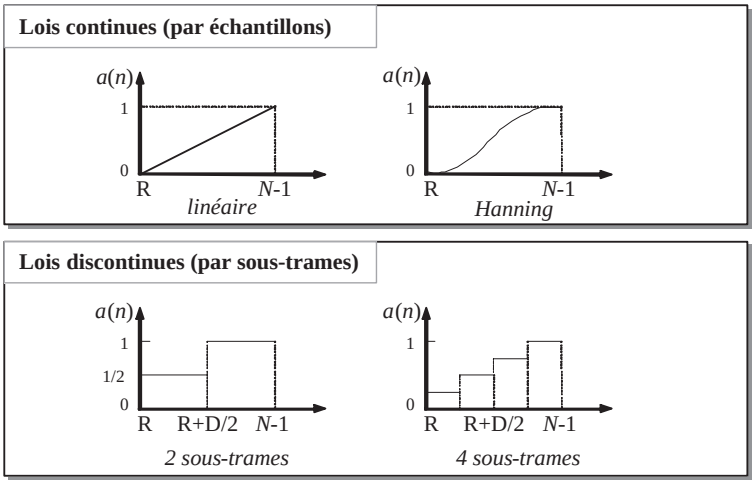

FIG. 5.6 – *Lois d'interpolation du filtre par échantillons ou par sous-trames.*

le filtre $g_{p,n}$ utilisé pour filtrer les échantillons de la trame p contient des informations statistiques de la trame précédente $p-1$ par le biais du filtre h_{p-1}. Il en résulte un retard au niveau de l'estimation du filtre, qui peut parfois se manifester par l'apparition d'une certaine rugosité dans la voix. Afin de limiter cet effet, on peut restreindre l'interpolation du filtre aux $M-1$ premiers échantillons soit $\alpha(n) = 1$ pour $n \in [R+M-1, N-1]$ (dans la mesure où $M-1 < D$ évidemment). L'audibilité de ce phénomène est d'ailleurs fortement liée à la profondeur de réduction de bruit souhaitée et il est en général masqué par le bruit résiduel tant que le niveau de bruit supprimé reste inférieur à environ 12dB.

5.1.4 L'"overlap and add" ou OLA

5.1.4.1 Principe

La technique OLA [Crochiere 1983, Kunt 1984] consiste à conserver les échantillons filtrés partiellement (filtre tronqué) de la trame x_p, puis à les combiner avec ceux issus des trames précédentes ou suivantes (selon que l'on considère le début ou la fin de la trame) afin d'obtenir des échantillons filtrés avec un filtre complet. Cette technique ne souffre pas des problèmes de "clics" que l'on rencontre avec la technique OLS car la transition entre les filtres des trames successives est assurée du fait de l'interpolation implicite d'une trame à l'autre.

Comme pour la technique OLS, l'implémentation peut se faire dans le domaine fréquentiel ou dans le domaine temporel. Au même titre que l'OLS, à l'origine l'OLA s'implémente dans le domaine fréquentiel (application de convolution rapide par FFT). Cependant son utilisation pour la réduction de bruit nécessite d'élargir le cadre classique. Comme pour la technique OLS, nous allons détailler les implémentations fréquentielle et temporelle qui découlent de cette application.

5.1.4.2 Implémentation dans le domaine fréquentiel

Durant la phase d'analyse, chaque trame x_p est pondérée par une fenêtre (*e.g.* fenêtre de Hanning). Afin d'éviter le problème de la convolution circulaire évoqué dans la partie 5.1.2, il est nécessaire d'allonger le support temporel de la trame par des zéros (zero-padding) afin d'assurer que les échantillons qui se replient soient uniquement des zéros [Kunt 1984, Crochiere 1983, Marro 1996]. Par conséquent, pour supprimer tout repliement, la taille M du filtre doit être inférieure au nombre de zéros rajoutés. Cette opération augmente du même coup la taille de la FFT.

Pour assurer la contrainte de reconstruction parfaite (puisque les trames se recouvrent), le signal rehaussé doit être obtenu par sommation sur plusieurs trames. Le nombre de trames dépend du taux de recouvrement, 50% ou 75% classiquement, et de la taille du filtre. Comme le montre la figure 5.7 (recouvrement de 50%), cette reconstruction introduit, pour les besoins de la sommation (sommation des parties des trames $p-1$, p et $p+1$ dans la zone grisée), **un retard incompressible de la durée d'une trame soit N échantillons.**

5.1.4.3 Implémentation dans le domaine temporel

La figure 5.8 représente l'implémentation dans le domaine temporel sans recouvrement de trames. Il s'agit là simplement d'illustrer le principe car le cadre de la réduction de bruit impose un recouvrement. Les parties triangulaires représentent les échantillons filtrés avec un filtre tronqué. Dans les zones de recouvrement des signaux filtrés, tout se passe comme si les coefficients manquants du filtre de la trame $p-1$ étaient progressivement remplacés par les coefficients du filtre de la trame p (*cf.* zone de sommation grisée).

L'implémentation dans le domaine temporel permet plus de souplesse que dans le domaine fréquentiel et deux implémentations aux caractéristiques distinctes sont possibles.

- L'implémentation peut être identique à celle du domaine fréquentiel à cela près que le filtrage est réalisé dans le domaine temporel. Les caractéristiques en terme de retard et de qualité sont donc les mêmes. Cependant, l'intérêt de réaliser la même approche dans le domaine temporel est limité.
- Il est également possible de réaliser une version possédant un plus faible retard et moins complexe de surcroît. Dans ce cas, on contourne la contrainte de reconstruction parfaite en se ramenant au cas de l'OLA sans recouvrement (*cf.* figure 5.8). Le buffer utilisé pour le filtrage contient alors uniquement les D nouveaux échantillons, non fenêtrés, complétés par $M-1$ zéros de part et d'autre pour réaliser l'OLA. Bien entendu cette réalisation impose une contrainte sur la longueur du filtre : $M \leq D$. Le retard de cette implémentation est de D échantillons. Cependant, il faut préciser que **des artefacts audibles peuvent apparaître sous la forme de grésillements, notamment pour des signaux dont le RSB global est fort** (supérieur à 18dB). Ceci peut s'expliquer par l'utilisation implicite d'un filtre hybride reprenant une partie des coefficients du filtre de la trame $p-1$ et une partie de ceux du filtre de la trame p et qui n'a donc pas réellement de sens physique. Cet artefact n'existe pas lorsque l'OLA est réalisé de façon classique (reconstruction parfaite avec recouvrement de 50 ou 75%) car la fenêtre d'analyse pondère l'impact du filtre hybride.

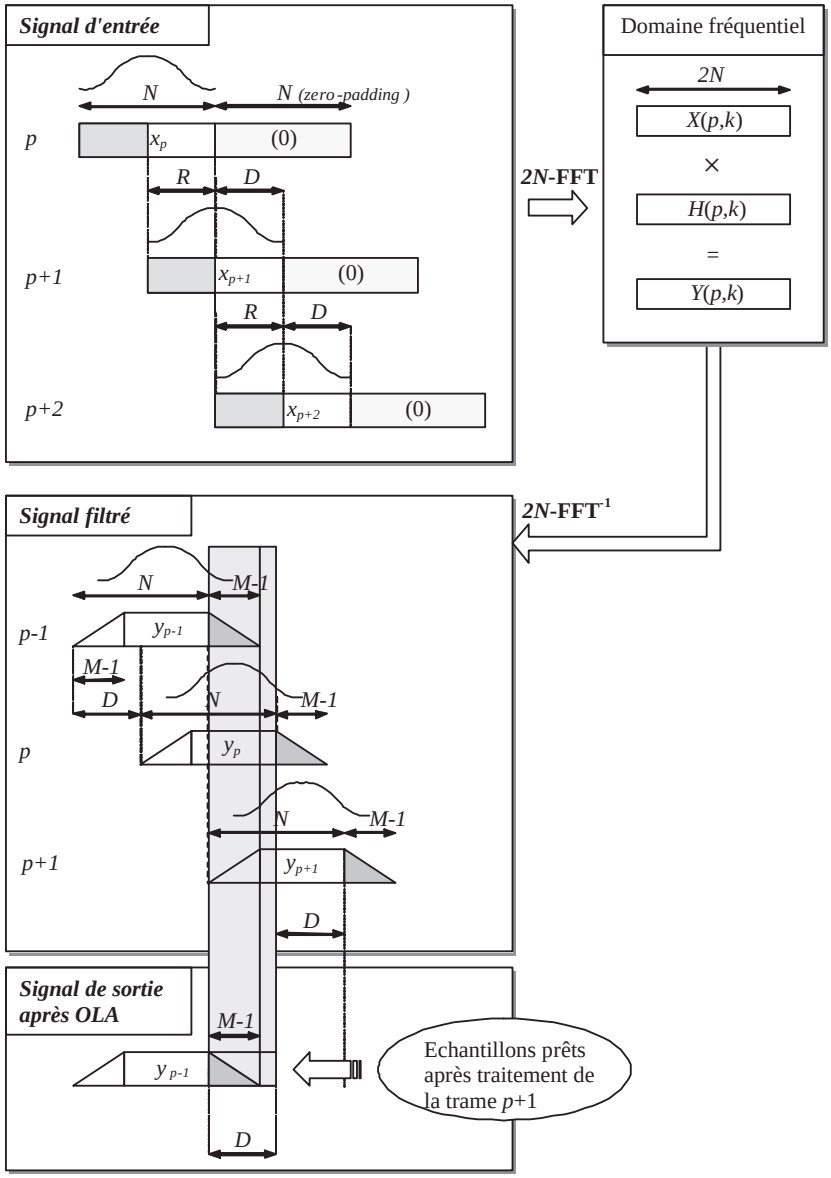

FIG. 5.7 – *Principe de l'OLA dans le domaine fréquentiel avec recouvrement des trames de 50%.*

5.1.5 Retards et caractéristiques des différentes approches et implémentations

Le tableau 5.1 synthétise les retards algorithmiques introduits par les approches OLS et OLA (selon l'implémentation de ces techniques) ainsi que les caractéristiques liées à la qualité des signaux filtrés (pour des cas classiques, *i.e.* taux de recouvrement de 50 ou 75%). Les retards (en échantillons) sont donnés sous forme d'une somme de 3 termes. Le premier correspond au retard introduit par la méthode utilisée de par son implémentation. Le second terme (toujours D) correspond au fait que les calculs sont réalisés pendant l'acquisition de la trame suivante. Finalement le terme $(M-1)/2$ correspond au retard introduit par le filtre de taille M (supposé impair) à phase linéaire (retard fixe pour toutes les fréquences). Le phénomène de rugosité de la parole traitée étant inaudible quand le

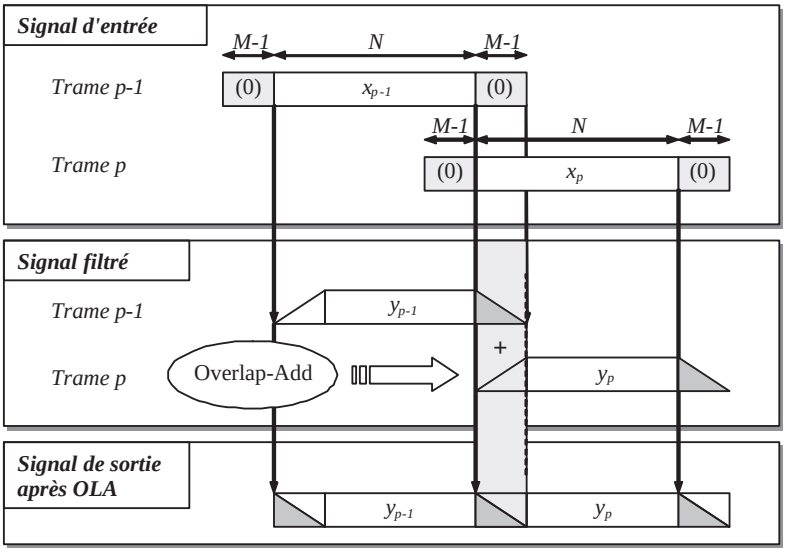

FIG. 5.8 – *Principe de l'OLA dans le domaine temporel sans recouvrement des trames ($R = 0$).*

TAB. 5.1 – *Retards et qualité des différentes approches et implémentations.*

Méthode	Retard	Qualité
OLS fréquentiel	$D+D+(M-1)/2$	clics
OLS temporel	$D+D+(M-1)/2$	clics
OLS en sous-trames	$D+D+(M-1)/2$	rugosité
OLA fréquentiel	$N+D+(M-1)/2$	pas d'artefact
OLA temporel	$N+D+(M-1)/2$	pas d'artefact
OLA temporel (faible retard)	$D+D+(M-1)/2$	grésillements

niveau de bruit supprimé reste inférieur à environ 12dB, **la méthode OLS en sous-trames est celle qui permet d'atteindre la meilleure qualité tout en introduisant le plus faible retard.** Rappelons que ces caractéristiques ne sont valables que dans le cas de la réduction de bruit (filtre variant dans le temps) et que les questions de qualité ne se posent pas dans le cas du filtrage classique.

5.2 Choix de la fenêtre d'analyse

Du point de vue de la mise en œuvre, fenêtre d'analyse et de synthèse sont interchangeables [Cappé 1993], cependant, en pratique la fenêtre d'analyse est très souvent choisie douce (idéalement à bords nuls) et celle de synthèse rectangulaire. En effet, le spectre de la trame analysée est convolué par la réponse fréquentielle de la fenêtre d'analyse qui possède un lobe principal plus large que celui d'une fenêtre rectangulaire (mais avec des lobes latéraux beaucoup plus faibles). L'atténuation spectrale qui est fonction du spectre estimé possède donc naturellement un profil relativement lissé dans la mesure où on ne peut pas trouver un canal complètement atténué jouxtant un canal peu ou pas atténué. Ce

choix permet donc de limiter le phénomène de repliement lié à un filtre trop agressif.

Comme il a été montré dans la partie 5.1, l'implémentation dans le domaine temporel des techniques de synthèse OLS et OLA ne nécessite pas de satisfaire à la contrainte de reconstruction parfaite. Dans cette optique, il devient donc possible de choisir n'importe quelle fenêtre d'analyse indépendamment du recouvrement et de la fenêtre de synthèse. En particulier, il est possible d'utiliser une fenêtre dissymétrique permettant de donner plus de poids à la fin de la trame, *i.e.* aux échantillons récents comme illustré par la figure 5.9.(a). Une telle fenêtre $w(n)$ peut être réalisée à partir de deux

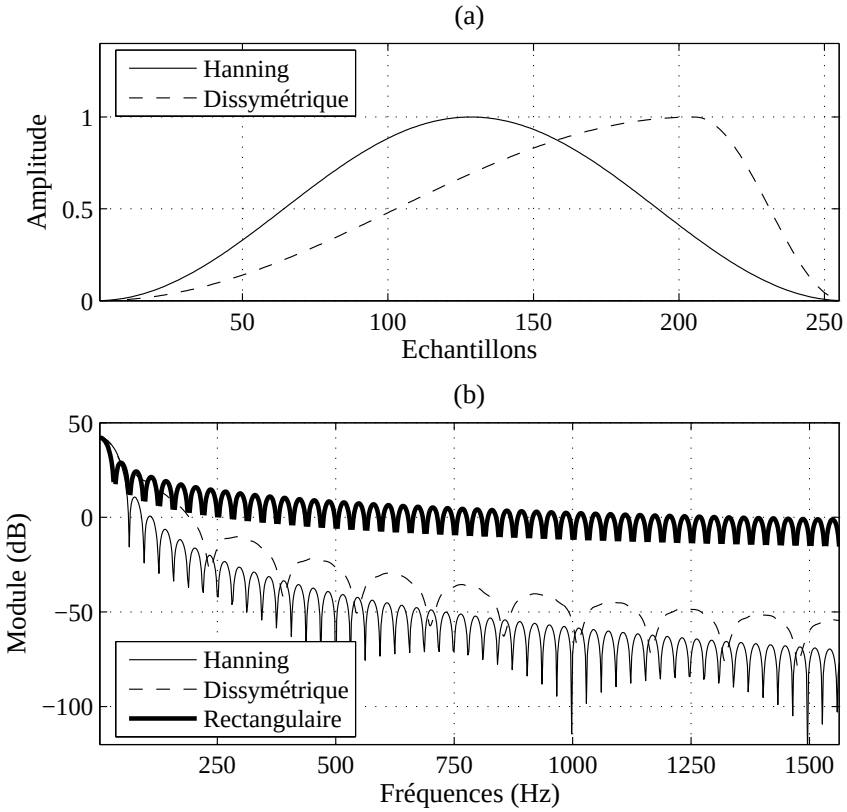

FIG. 5.9 – *(a) Allure temporelle d'une fenêtre symétrique de Hanning (trait plein) et d'une fenêtre dissymétrique à* 80% *(tirets). (b) Réponses fréquentielles (zoom sur les basses fréquences) équivalentes avec en plus la réponse de la fenêtre rectangulaire (trait fort) pour comparaison.*

demi-fenêtres de Hanning de tailles différentes :

$$w(n) = \begin{cases} 0.5 - 0.5cos\left(\frac{\pi n}{L_1}\right) \text{ pour } n \in [0,L_1-1] \\ 0.5 + 0.5cos\left(\frac{\pi(n-L_1+1)}{L_2}\right) \text{ pour } n \in [L_1,L_1+L_2-1]. \end{cases} \tag{5.4}$$

Cette fenêtre peut être caractérisée par son taux de dissymétrie : $\tau = L_1/(L_1+L_2)$, ainsi, si $\tau < 50\%$ le passé de la trame sera privilégié sinon ce sera le futur. Le cas particulier où $\tau = 50\%$ correspond à la fenêtre de Hanning de taille L_1+L_2. L'allure d'une fenêtre dissymétrique à 80% est donnée par la figure 5.9.(a) ; une fenêtre de Hanning classique y est également représentée.

L'avantage d'une fenêtre privilégiant les échantillons les plus récents est lié à un phénomène

récurrent dans les techniques de réduction de bruit. En effet, on a vu dans la partie 3.2.5 que les techniques basées sur l'utilisation du RSB *a priori*, calculé en utilisant l'approche decision-directed, subissent un retard dans l'estimation du gain spectral qui se traduit par un effet de réverbération. **Ainsi, des écoutes informelles ont permis de montrer que le fait de privilégier les échantillons récents de la trame analysée réduit sensiblement cet effet de réverbération (il n'est cependant pas supprimé)**. Il faut tout de même noter qu'une fenêtre de ce type entraîne une remontée des lobes secondaires, visible sur la figure 5.9.(b), par rapport à une fenêtre classique comme une fenêtre de Hanning mais qui reste tout de même raisonnable, *i.e.* le niveau des lobes secondaires reste inférieur à celui des lobes de la fenêtre rectangulaire. On peut remarquer que cette technique ne permet pas d'atteindre les performances des approches TSNR et RFSNR (*cf.* parties 4.3 et 4.4) qui lui seront préférées pour supprimer l'effet de réverbération introduit par l'approche DD.

5.3 Limitation des distorsions par seuillage du gain

Les techniques de réduction de bruit pour les applications de communication doivent répondre à certaines exigences de qualité. De façon très générale, le signal restauré doit être plus agréable à l'écoute que le signal bruité, ce qui, il faut en convenir, est très subjectif. Comme on l'a déjà vu, notamment dans les parties 4.3 et 4.4, si l'on cherche à supprimer la totalité du bruit alors certaines composantes de parole, en particulier de nombreuses harmoniques, sont aussi supprimées. Ainsi le signal restauré peut paraître moins riche que le signal bruité car le bruit comble l'impression de manque, l'oreille humaine étant capable de reconstituer les harmoniques manquantes à partir du timbre [Cappé 1993]. Le fait de conserver une certaine quantité de bruit résiduel permet donc de masquer ce manque d'harmoniques. Cela permet aussi de masquer partiellement certains artefacts comme l'effet tuyau ou le bruit musical discutés dans la partie 5.4. Notons qu'il est de toute façon nécessaire de conserver l'ambiance sonore (atténuée) car elle fait partie de la communication au même titre que le signal de parole.

Une approche classique et qui peut être appliquée à toutes les techniques d'atténuation spectrale consiste à appliquer un seuil au gain spectral. Ce seuil, $0 \leq G_{min} < 1$, est déterminé de façon à obtenir un compromis entre la qualité du signal désiré et la quantité de bruit que l'on désire supprimer. Le gain à appliquer au spectre bruité s'exprime donc ainsi :

$$G_{seuil}(p,k) = \max(G(p,k), G_{min}). \tag{5.5}$$

Cette opération est non-linéaire ce qui se traduit par le fait que le niveau de bruit supprimé n'est pas uniforme. En effet, pendant les périodes de bruit seul, où généralement $G(p,k) < G_{min}$, une partie du bruit est réinjectée par l'intermédiaire de G_{min} ce qui n'est pas le cas pendant les périodes d'activité vocale pour les composantes de parole (où généralement $G(p,k) > G_{min}$).

Dans le cadre de l'obtention du filtre de Wiener, il est possible de diminuer la distorsion globale (parole et bruit) en partant de la nouvelle hypothèse que le signal désiré n'est plus $S(p,k)$ mais $S(p,k) + G_{min}B(p,k)$. Ainsi, l'erreur à minimiser au sens de l'EQMM sera :

$$e(p,k) = \mathrm{E}[(S(p,k) + G_{min}B(p,k) - G(p,k)X(p,k))^2]. \tag{5.6}$$

Sous les hypothèses posées dans la partie 2.1 (signaux supposés stationnaires ; bruit additif indépendant du signal de parole) cette expression devient :

$$e(p,k) = (1-G(p,k))^2 \mathrm{E}[|S(p,k)|^2] + (G_{min} - G(p,k))^2 \mathrm{E}[|B(p,k)|^2] \tag{5.7}$$

et son minimum est atteint pour :

$$G(p,k) = \frac{\mathrm{E}[|S(p,k)|^2] + G_{min}\mathrm{E}[|B(p,k)|^2]}{\mathrm{E}[|S(p,k)|^2] + \mathrm{E}[|B(p,k)|^2]}. \tag{5.8}$$

Ce gain spectral peut être réécrit de façon à faire apparaître l'expression du filtre de Wiener classique $G_W(p,k)$ (2.20) :

$$\begin{aligned} G(p,k) &= \frac{\mathrm{E}[|S(p,k)|^2]}{\mathrm{E}[|S(p,k)|^2] + \mathrm{E}[|B(p,k)|^2]} + G_{min}\left(1 - \frac{\mathrm{E}[|S(p,k)|^2]}{\mathrm{E}[|S(p,k)|^2] + \mathrm{E}[|B(p,k)|^2]}\right) \\ &= G_W(p,k) + G_{min}(1 - G_W(p,k)). \end{aligned} \tag{5.9}$$

Le deuxième terme correspond à la réinjection d'une partie du bruit de façon à ce que le niveau de bruit supprimé soit le même pour les composantes de bruit seul et de parole. Le principe de cette approche obtenue dans le cadre du filtre de Wiener peut être repris avec n'importe quelle autre fonction de gain, il suffit alors de remplacer dans l'équation (5.9) le terme $G_W(p,k)$ par le gain voulu. Comme le montre la figure 5.10, si le gain obtenu par seuillage classique est déjà proche de 0dB alors l'impact du seuillage uniforme ne sera pas sensible. Par contre, l'apport de cette approche se ressent dans les

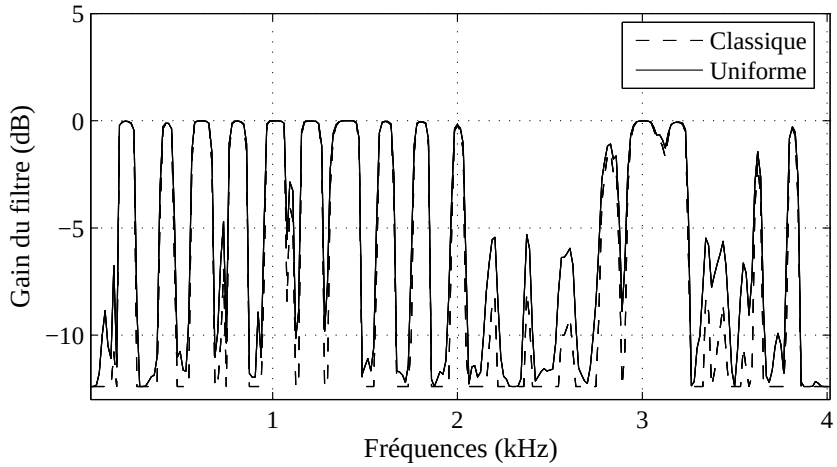

FIG. 5.10 – *Gain obtenu pour une trame voisée avec un seuillage classique (pointillé) et un seuillage uniforme (trait fin).*

zones où le gain est relativement faible (parole peu énergétique ou signal très bruité). **Le gain obtenu en utilisant le seuillage uniforme est alors sensiblement supérieur au gain classique permettant ainsi de mieux préserver la richesse du signal de parole.**

5.4 Contrôle de l'agressivité du filtre de réduction de bruit

5.4.1 Pourquoi limiter l'agressivité du filtre ?

La partie 5.1 montre que, selon la méthode de synthèse choisie, la taille du filtre de réduction de bruit doit respecter certaines contraintes. Ainsi, l'implémentation de la méthode OLS dans le domaine fréquentiel impose que la taille du filtre soit inférieure à celle du recouvrement. L'implémentation de la méthode OLA dans le domaine fréquentiel impose quant à elle que la taille du filtre soit inférieure à la taille du zero-padding effectué [Kunt 1984, Crochiere 1983, Marro 1996]. Les implémentations temporelles permettent plus de souplesse et le filtre peut alors être aussi long que la trame de signal à filtrer. Hormis ces contraintes d'implémentation, il peut être nécessaire de limiter la taille du filtre afin de réduire la complexité (convolution moins coûteuse) ou pour diminuer le retard introduit par l'algorithme de réduction de bruit (terme $(M-1)/2$ dans le tableau 5.1).

Cependant le choix de la taille du filtre influe aussi grandement sur la qualité du signal de parole restauré. Très souvent le signal restauré est entaché de bruit musical que même l'approche decision-directed (*cf.* partie 3.2.5) ne permet pas de supprimer complètement, en particulier lorsque le bruit est de nature non-stationnaire. Du point de vue réponse fréquentielle du filtre, le bruit musical est caractérisé par des valeurs de gain importantes localisées de façon aléatoire en fréquence et qui ne dépassent pas statistiquement la durée d'une trame. **Il est donc possible de supprimer ce bruit musical en empêchant le gain de fluctuer rapidement, c'est-à-dire en lissant sa réponse fréquentielle ou ce qui est équivalent en limitant son support temporel**. En effet, la longueur de la réponse impulsionnelle est directement reliée à son "agressivité" dans le domaine fréquentiel, *i.e.* sa capacité à varier rapidement d'une fréquence à l'autre. De plus, **le fait de limiter l'agressivité du filtre permet de diminuer l'effet "tuyau" (nasalisation) caractéristique d'un filtre trop agressif** qui a tendance à supprimer tout le signal présent entre les harmoniques (car il est généralement noyé dans le bruit). Enfin, cela permet aussi de diminuer la variabilité du filtre d'une trame à l'autre ce qui limite du même coup l'importance des artefacts du traitement par blocs (clics et grésillement, *cf.* partie 5.1 et tableau 5.1). Cette solution n'est toutefois pas sans effet sur la qualité de la parole. En effet, plus le filtre est choisi court et plus la parole restaurée semble étouffée (perte de présence). La raison de cette dégradation sera expliquée dans la partie 5.4.3. **Il s'agira donc selon la qualité désirée de trouver le bon compromis entre dégradation de la parole, artefacts et niveau de bruit musical.**

5.4.2 Comment limiter l'agressivité du filtre ?

L'atténuation spectrale à court terme est une approche qui consiste à appliquer un gain au spectre du signal bruité dans le but d'estimer celui du signal de parole propre. En réalité, seul le module est restauré (la phase bruitée est réutilisée telle quelle), par conséquent le gain appliqué est réel. Dans la suite, les exemples permettant d'illustrer le discours sont obtenus à partir d'un signal de parole perturbé par un bruit de voiture avec un RSB global de 12dB (*cf.* partie 3.2.1).

À partir du gain réel spécifié dans le domaine fréquentiel, il est possible de limiter son agressivité en contraignant la taille de son support temporel [Marro 1996]. Tout d'abord, il faut obtenir la réponse

impulsionnelle $h(p,n)$ qui correspond au gain spectral $H(p,k)$:

$$h(p,n) = h_p(n) = IFFT(H(p,k)). \tag{5.10}$$

Cette réponse impulsionnelle est non causale car sa phase (linéaire) est nulle pour toutes les fréquences (le gain étant réel). La figure 5.11.(a) en représente un exemple obtenu pour une trame voisée de voix féminine. Il faut dans un premier temps rendre causale cette réponse impulsionnelle :

$$\begin{cases} h_{causal}(p,n) = h(p,n+L/2) \text{ pour } n \in [0,L/2-1] \\ h_{causal}(p,n) = h(p,n-L/2) \text{ pour } n \in [L/2,L-1] \end{cases} \tag{5.11}$$

où L représente la taille de la FFT qui est supposée paire. Idéalement, L est une puissance de 2 car la FFT (ainsi que la IFFT) est numériquement optimale lorsqu'elle s'applique sur des trames de longueur en puissance de 2. La figure 5.11.(b) représente la réponse impulsionnelle de la figure 5.11.(a) une fois

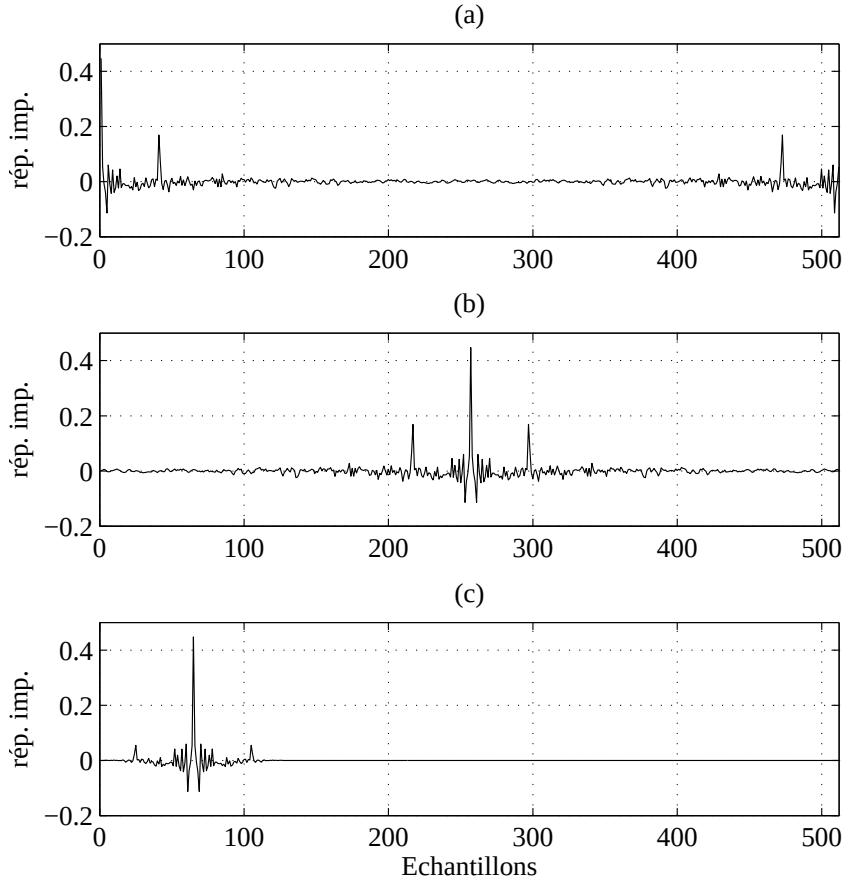

FIG. 5.11 – *Réponse impulsionnelle (a) du filtre non contraint et non causal, (b) du filtre non contraint mais rendu causal et (c) du filtre causal contraint à* 129 *échantillons.*

rendue causale : $h_{causal}(p,n)$. Dans cet exemple $L = 512$ et donc le retard introduit par ce filtre serait de 255 échantillons. Ensuite la taille M (impair) du filtre est choisie en fonction de l'implémentation et de la qualité de parole désirée :

$$h_{tronc}(p,n) = h_{causal}(p,n+L/2-(M+1)/2) \text{ pour } n \in [0,M-1]. \tag{5.12}$$

Finalement, pour éviter les effets de bord, cette réponse impulsionnelle est pondérée par une fenêtre à bords nuls (de préférence) w_{filt}, *e.g.* fenêtre de Hanning, de longueur M :

$$h_w(p,n) = w_{filt}(n) h_{causal}(p,n) \text{ pour } n \in [0,M-1]. \tag{5.13}$$

Ce fenêtrage peut être interprété comme un filtrage passe bas. La figure 5.11.(c) représente la réponse impulsionnelle $h_w(p,n)$ causale et dont la taille est contrainte à $M = 129$ échantillons. Le retard introduit par ce filtre contraint est donc seulement de 64 échantillons. **Limiter la taille du filtre permet donc d'améliorer la qualité du signal restauré (diminution voire suppression du bruit musical et des artefacts) tout en diminuant le retard du filtre mais aux prix d'un phénomène d'étouffement de la parole.**

5.4.3 Impact de la contrainte temporelle sur la réponse fréquentielle

Dans les trois exemples qui vont suivre, la méthode OLA est choisie pour la synthèse (implémentation fréquentielle), avec un zero-padding d'ordre 2 ($L = 2N$, N zéros sont ajoutés pour compléter la trame de taille N). La taille maximale du filtre (au delà du repliement temporel apparaît) est donc de $M = 255$ échantillons.

La limitation de l'agressivité du filtre permet entre autres de réduire l'effet de bruit musical entachant très souvent le signal restauré. Cet effet est particulièrement marqué en utilisant la technique de la soustraction spectrale en puissance (*cf.* partie 2.4.1.1) qui a donc été retenue pour illustrer, dans la figure 5.12, l'intérêt de la contrainte temporelle du filtre sur une trame de bruit seul. On peut observer que, naturellement, plus le support temporel du filtre est court, plus sa réponse fréquentielle est lissée. En période de bruit seul, le gain fréquentiel devrait atténuer de façon importante toutes les

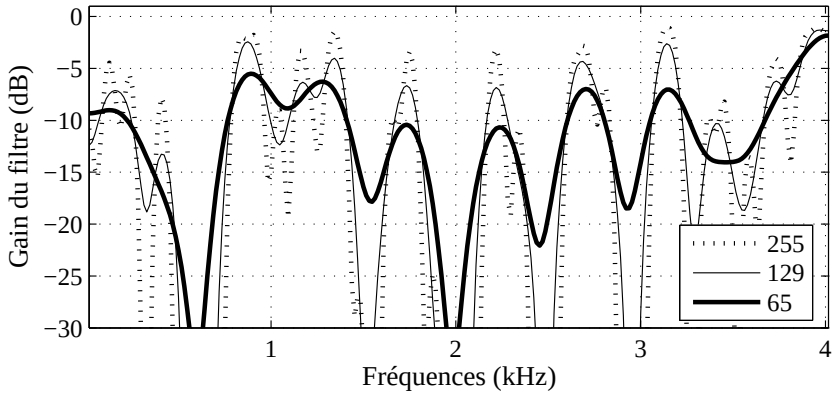

FIG. 5.12 – *Effet de la limitation de la taille du filtre sur sa réponse en fréquence pour une trame de bruit de voiture. Réponses fréquentielles du filtre contraint à* 255 *coefficients (pointillé),* 129 *(trait fin) et* 65 *(trait fort).*

fréquences. Cependant, de par les erreurs d'estimation du bruit et des différents RSB utilisés dans le calcul du gain, de nombreuses fréquences réparties aléatoirement sur le spectre et au cours du temps sont seulement faiblement atténuées créant par conséquent le phénomène très gênant de bruit musical. Ainsi, sur la figure 5.12, cet effet est très marqué pour le filtre qui possède la taille maximale de 255

coefficients. La limitation de l'agressivité du gain fréquentiel, notamment pour $M = 65$ coefficients, permet d'étaler en fréquence les occurrences isolées de gain élevé. Cette atténuation supplémentaire du gain permet ainsi de limiter fortement le phénomène de bruit musical voire de le supprimer complètement.

Dans les deux exemples qui suivent, le filtre de réduction de bruit choisi sera le filtre de Wiener (*cf.* partie 2.4.1.3). La figure 5.13 permet d'illustrer l'impact de la longueur du filtre sur sa réponse fréquentielle pour une trame voisée de voix féminine. Sur la réponse impulsionnelle de la figure 5.13.(a),

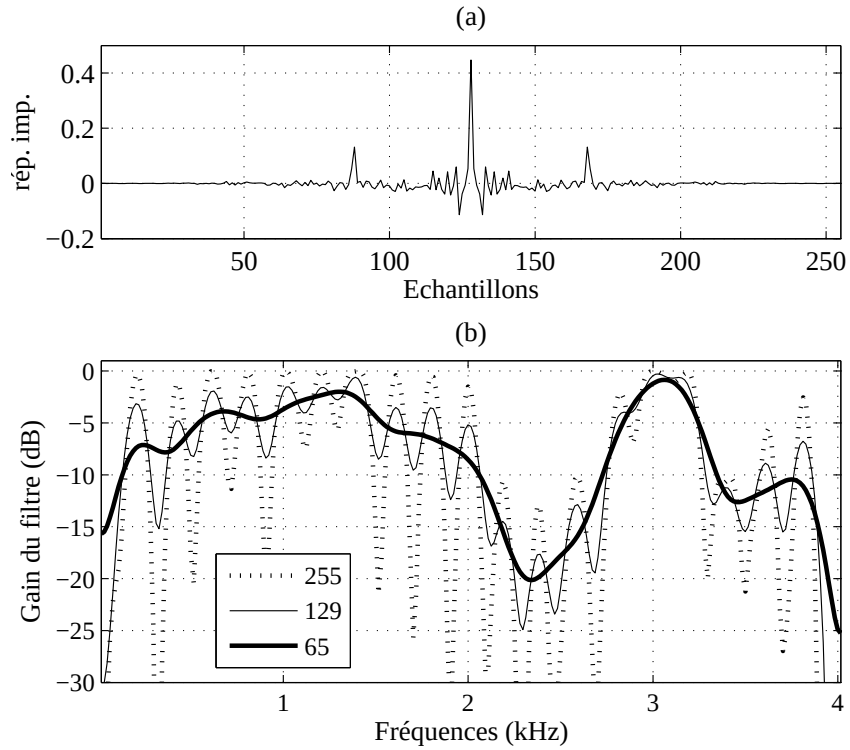

FIG. 5.13 – *Effet de la limitation de la taille du filtre sur la réponse en fréquence pour une voix féminine. (a) Réponse impulsionnelle limitée à* 255 *coefficients. (b) Réponses fréquentielles du filtre contraint à* 255 *coefficients (pointillé),* 129 *(trait fin) et* 65 *(trait fort).*

il existe 2 pics secondaires bien marqués qui sont directement reliés à l'harmonicité importante du filtre fréquentiel visible dans la figure 5.13.(b). Dans le cas où $M = 65$, la réponse fréquentielle est tellement lissée qu'elle ne peut plus suivre la structure harmonique de la parole. La réduction de bruit se fait alors plus au niveau des formants que des harmoniques. La perte de résolution fréquentielle (voulue) s'accompagne d'une atténuation plus importante que celle désirée pour les harmoniques du signal. Plus le filtre est court et plus l'atténuation apportée aux harmoniques est importante. Par exemple, on observe dans le cas $M = 65$ une atténuation de l'ordre de 8dB dans les basses fréquences. **Cette atténuation des harmoniques est présente dans tout le spectre et entraîne un effet d'étouffement de la parole.**

On peut par ailleurs observer sur la figure 5.14, où le filtre est cette fois ci obtenu pour une trame voisée de voix masculine, que cette atténuation est beaucoup moins présente dans ce cas où les har-

moniques sont plus proches les unes des autres et que par conséquent, le lissage ne s'accompagne pas d'une atténuation aussi importante que pour des voix féminines. On peut noter que les pics secondaires (sur la figure 5.14.(a)) sont ici très faibles (échantillon 50 pour celui de gauche) témoignant du fait que la résolution fréquentielle du gain ($M = 255$ sur la figure 5.14.(b)) est trop faible pour suivre parfaitement l'harmonicité de cette voix. L'atténuation est donc fonction de la fréquence fondamentale et par conséquent les voix dont la fréquence fondamentale est faible souffrent peu de la sensation d'étouffement précédemment décrite.

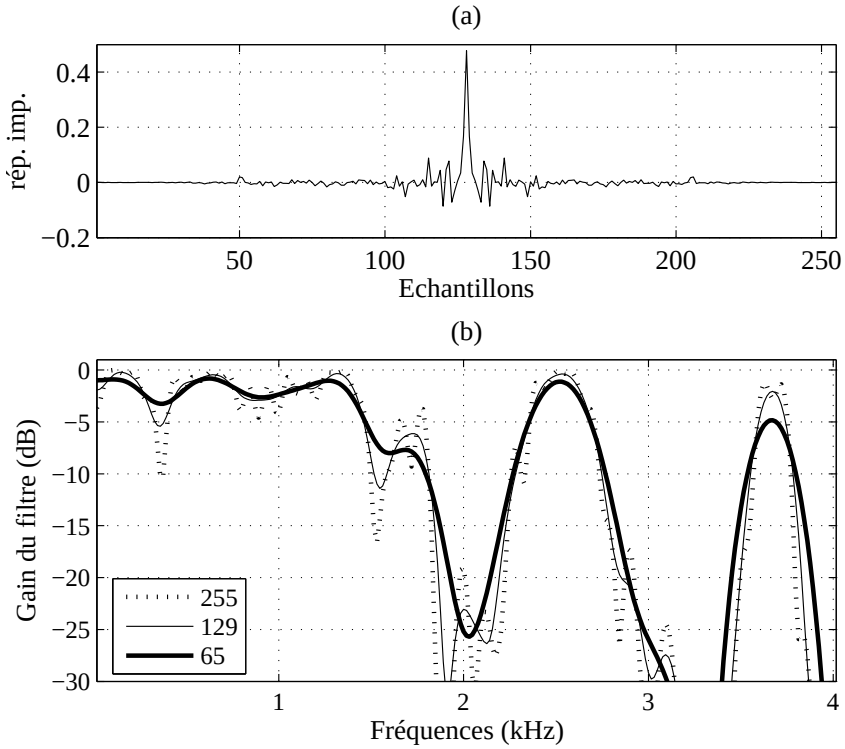

FIG. 5.14 – *Effet de la limitation de la taille du filtre sur la réponse en fréquence pour une voix masculine. (a) Réponse impulsionnelle limitée à* 255 *coefficients. (b) Réponses fréquentielles du filtre contraint à* 255 *coefficients (pointillé),* 129 *(trait fin) et* 65 *(trait fort).*

Le fait d'appliquer un seuil minimal au gain (*cf.* partie 5.3), qui peut entre autres servir à masquer les défauts de la réduction de bruit, avant de contraindre la taille du filtre limite sensiblement ce phénomène. Toutefois, c'est l'utilisation de l'approche psychoacoustique qui permet de résoudre ce problème (*cf.* partie 2.4.3) tout en conservant le caractère lissé du filtre qui permet de supprimer le bruit musical résiduel et les artefacts.

5.5 Apport de l'approche psychoacoustique

La partie 2.4.3 met en avant différentes façons d'exploiter le masquage psychoacoustique. Celle proposée dans [Lin 2002] apparaît très prometteuse car elle permet de supprimer uniquement la partie audible du bruit réduisant ainsi la distorsion du signal de parole. **Lin a construit sa solution à partir**

du filtre de Wiener mais il est en fait possible d'utiliser son approche quel que soit le filtre utilisé. Ceci est rendu possible en modifiant dans l'expression des RSB *a posteriori* et *a priori* la quantité de bruit à supprimer qui n'est plus $\mathrm{E}[|B(p,k)|^2]$ mais seulement sa partie audible (notée *aud*) $\mathrm{E}[|B^{aud}(p,k)|^2]$ exprimée par l'équation (2.55). Ainsi, les expressions des deux RSB qui interviennent dans le calcul des filtres de réduction de bruit peuvent être transformées ainsi :

$$RSB^{aud}_{post}(p,k) = \frac{|X(p,k)|^2}{\mathrm{E}[|B^{aud}(p,k)|^2]} = \frac{|X(p,k)|^2}{\max(\mathrm{E}[|B(p,k)|^2] - T(p,k),0)}, \tag{5.14}$$

$$RSB^{aud}_{prio}(p,k) = \frac{\mathrm{E}[|S(p,k)|^2]}{\mathrm{E}[|B^{aud}(p,k)|^2]} = \frac{\mathrm{E}[|S(p,k)|^2]}{\max(\mathrm{E}[|B(p,k)|^2] - T(p,k),0)}. \tag{5.15}$$

où rappelons que $T(p,k)$ représente le seuil de masquage. Ces deux quantités n'interviennent que dans le calcul final du filtre et les estimateurs classiques du RSB *a posteriori* et *a priori* (*cf.* équations (3.11) et (3.12) de la partie 3.2.5) doivent tout de même être calculés car ils permettent de mettre à jour la DSP $\mathrm{E}[|S(p,k)|^2]$ utilisée dans l'expression (5.15). En pratique, la quantité $RSB^{aud}_{prio}(p,k)$ peut donc être estimée selon :

$$\hat{RSB}^{aud}_{prio}(p,k) = \frac{\hat{RSB}^{DD}_{prio}(p,k)\hat{\gamma}_b}{\max(\hat{\gamma}_b - T(p,k),0)}. \tag{5.16}$$

Il est aussi possible de tirer parti des approches TSNR ou HRNR (*cf.* parties 4.3 et 4.5), par exemple, en remplaçant dans l'équation ci-dessus la quantité $\hat{RSB}^{DD}_{prio}(p,k)$ par la quantité $\hat{RSB}^{TSNR}_{prio}(p,k)$ ou encore $\hat{RSB}^{HRNR}_{prio}(p,k)$. Outre les avantages inhérents à l'approche psychoacoustique (*cf.* partie 2.4.3), l'utilisation du masquage fréquentiel permet aussi de résoudre le problème dû à la limitation de l'agressivité du filtre de réduction de bruit (lié au choix de mise en œuvre, *cf.* partie 5.4). En effet, on a vu que l'on peut être amené à choisir un filtre peu agressif en fréquence pour supprimer les défauts audibles du signal restauré comme le bruit musical ou l'effet tuyau. Ce type de contrainte permet d'améliorer nettement la qualité du signal restauré mais en contrepartie la parole subit une perte de niveau qui est d'ailleurs d'autant plus marquée que la fréquence fondamentale de la parole est importante (voix aiguës).

Les techniques classiques (comprendre qui n'utilisent pas le masquage psychoacoustique) ont tendance à supprimer la totalité du bruit entre les harmoniques. Par conséquent, la limitation de l'agressivité du filtre, équivalente à un lissage fréquentiel, entraîne une forte atténuation du gain à appliquer aux harmoniques (*cf.* partie 5.4). D'où l'intérêt de l'approche psychoacoustique qui permet d'obtenir un gain toujours supérieur ou égal à celui du filtre classique et généralement largement supérieur entre les harmoniques où le bruit est masqué par le signal de parole comme illustré par la figure 5.15. La trame de parole considérée ici est une trame de signal voisé (voix féminine) extraite d'un signal perturbé par un bruit de voiture avec un RSB global de 12dB. Pour la trame représentée, le seuil de masquage "idéal" obtenu à partir du signal de parole propre correspond peu ou prou à une version décalée de 2 à 5dB du seuil obtenu en pratique à partir d'une première restauration du signal bruité. Cela n'est pas très gênant dans la mesure ou la courbe de masquage reste sous-estimée par rapport à la courbe idéale. Le cas où elle est surestimée peut aussi se produire, il est alors possible que le bruit ne soit plus complètement masqué. Pour éviter ce problème on peut choisir de sous-estimer le seuil de masquage de quelques dB. On peut remarquer que le seuil de masquage est aussi élevé entre les harmoniques qu'au niveau des harmoniques elles-mêmes, indiquant qu'une quantité importante de bruit peut être masquée entre les harmoniques ce qui se traduira par une augmentation du gain.

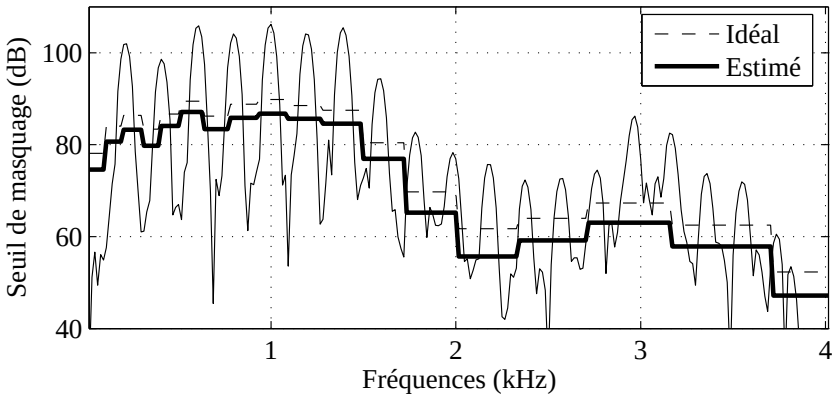

FIG. 5.15 – *Spectre de puissance (trait fin) d'une trame de parole propre (voix féminine) et les courbes de masquage obtenues à partir d'une première réduction de bruit (trait fort) et à partir de la trame de parole propre (tirets).*

Ce phénomène est illustré par la figure 5.16 qui permet de visualiser l'apport de l'approche psychoacoustique dans le cadre de la limitation de l'agressivité du filtre de réduction de bruit. Le gain fréquentiel (filtre de Wiener) est calculé pour la même trame que celle utilisée dans la figure 5.15. La méthode OLA fréquentielle (*cf.* partie 5.1) est choisie pour la synthèse du signal traité, ainsi la taille maximale du filtre est de 255 coefficients pour une FFT de 512 points sur une fenêtre de 256 échantillons. Le lissage du filtre de Wiener classique (65 coefficients) introduit de larges atténuations

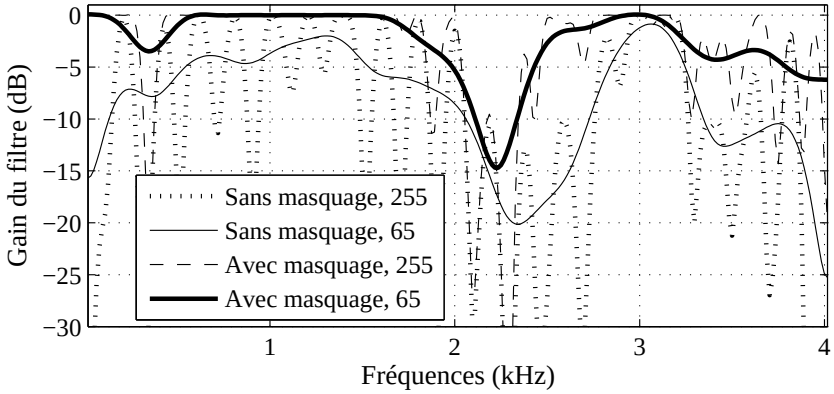

FIG. 5.16 – *Apport de l'approche psychoacoustique dans le cadre de la limitation de l'agressivité du filtre de réduction de bruit. Filtre de Wiener classique contraint à* 255 *coefficients (pointillé) et sa version lissée à* 65 *coefficients (trait fin). Filtre de Wiener intégrant l'approche psychoacoustique contraint à* 255 *coefficients (tirets) et sa version lissée à* 65 *coefficients (trait fort).*

au niveau des harmoniques provoquant une sensation d'étouffement de la parole ce qui n'est pas le cas avec l'approche psychoacoustique. En effet, le gain contraint à 255 coefficients est nettement plus élevé au niveau des harmoniques et surtout entre celles-ci qu'avec l'approche classique. Le gain atteint même parfois 0dB (bruit complètement masqué) alors que le gain classique peut être de -10dB. En utilisant l'approche psychoacoustique le gain fréquentiel est donc naturellement lissé et quand il est

contraint à 65 coefficients l'atténuation provoquée reste négligeable. Le signal traité ne souffre donc pas de l'effet d'étouffement précédemment décrit. Bien que le gain obtenu soit naturellement peu agressif pendant l'activité vocale, le lissage du filtre reste tout de même indispensable pour supprimer complètement le bruit musical (surtout pendant l'inactivité vocale). Notons que ces observations ont été corroborées par des écoutes informelles et que des résultats objectifs seront donnés dans la partie 6.3.8.

Il est donc possible d'obtenir un signal rehaussé exempt de défauts tels que le bruit musical ou l'étouffement de la parole en intégrant l'approche psychoacoustique à une technique classique. Il faut cependant préciser que le niveau de réduction de bruit (contrôlé par seuillage du gain, *cf.* partie 5.3) que l'on peut atteindre est limité à environ 12dB. Dans le cas contraire, des remontées de bruit deviennent audibles pendant les périodes d'activité vocale. Cet effet est tout simplement dû au fait que le seuil de masquage est estimé et qu'il n'est donc pas parfait. Cette approche ne permet donc pas de supprimer autant de bruit qu'on le désire mais c'est aussi le cas pour les approches classiques (pour d'autres raisons : artefacts, bruit musical, effet tuyau).

5.6 Traitement distinct des composantes voisées et non voisées de la parole

L'intérêt de l'approche HRNR est de pouvoir restaurer les harmoniques du signal de parole qui sont détruites par des approches classiques (*cf.* partie 4.5). Cette approche limite donc les distorsions de la parole et par conséquent, du point de vue de la mise en œuvre, il est possible de supprimer davantage de bruit qu'en utilisant une approche classique. Toutefois, le fait de conserver un filtre agressif entraîne des artefacts qui deviennent audibles lorsque le niveau de bruit résiduel est faible. La partie 5.4 met en avant que, pour limiter ce type de distorsion, il est judicieux d'opter pour un filtre peu agressif. Cependant, l'intérêt est pour le moins limité de régénérer les harmoniques pour ensuite les atténuer suite à l'utilisation d'un filtre lissé. L'introduction de la psychoacoustique dans l'approche HRNR pourrait résoudre ce problème mais cela a plutôt tendance à limiter son intérêt dans le sens où il n'est alors pas possible de supprimer plus de 12dB de bruit sans occasionner une remontée audible de bruit dans le signal restauré (*cf.* partie 5.5). **Pour résumer, il faudrait pouvoir traiter le signal de parole avec un filtre court (lissé) excepté pour ses harmoniques où il est nécessaire de conserver un filtre long (agressif).** Ceci nécessite donc d'extraire les composantes harmoniques et de les traiter séparément et différemment du reste du signal bruité. Cette approche qui limite ainsi la distorsion du signal de parole et les artefacts sera nommée VNV du fait du traitement distinct de ses composantes voisées et non voisées.

5.6.1 Principe de l'approche VNV

Considérons pour l'instant l'extraction de la partie harmonique du signal de parole comme un problème annexe et supposons que la partie voisée, et bruitée, $X_V(p,k)$ du signal $X(p,k)$ est disponible. Le principe de l'approche VNV consiste à restaurer le signal de parole en utilisant une approche classique (TSNR ou RFSNR par exemple, *cf.* parties 4.3 et 4.4), *i.e.* sans régénération des harmoniques,

avec un filtre peu agressif et de remplacer toutes les composantes harmoniques du signal restauré par celles obtenues avec l'approche HRNR en utilisant bien sûr un filtre agressif. Il suffit donc d'appliquer au signal $|X_V(p,k)|$ le gain spectral obtenu par l'approche HRNR pour obtenir une estimation fiable de la partie voisée $|\hat{S}_V(p,k)|$ du signal de parole propre. Notons que nous travaillons uniquement avec les spectres d'amplitude des différents signaux. Ensuite le spectre d'amplitude du signal de parole propre est estimé ainsi :

$$|\hat{S}_{VNV}(p,k)| = \max(|\hat{S}_{liss}(p,k)|,|\hat{S}_V(p,k)|), \tag{5.17}$$

où $\hat{S}_{liss}(p,k)$ est le signal obtenu par l'approche classique avec un filtre lissé. Ainsi, les harmoniques dégradées par ce lissage sont tout simplement remplacées par les harmoniques obtenues par l'approche HRNR. Non seulement ces harmoniques ne sont pas atténuées mais en plus certaines harmoniques qui sont dégradées par les techniques classiques sont restaurées. L'estimée du spectre du signal de parole est finalement obtenue en appliquant la phase bruitée au signal $|\hat{S}_{VNV}(p,k)|$:

$$\hat{S}_{VNV}(p,k) = |\hat{S}_{VNV}(p,k)|e^{i\phi_X(p,k)}. \tag{5.18}$$

Le schéma de principe de l'approche VNV décrite ci-dessus est représenté par la figure 5.17.

La difficulté de cette approche consiste à extraire uniquement et de façon robuste les composantes harmoniques de la parole. Pour chacune des trames il faut déterminer la position de chaque harmonique. De nombreuses techniques permettent de déterminer la fréquence fondamentale du signal de parole [McAulay 1990] dont certaines de façon plus ou moins robuste au bruit [Prasanna 2004]. Par la suite il est donc possible de déterminer celles des harmoniques qui en sont des multiples. Toutefois nous proposons plutôt d'estimer directement le peigne harmonique correspondant à chaque trame en utilisant une approche robuste et relativement peu coûteuse.

5.6.2 Extraction des composantes harmoniques du signal de parole

En premier lieu, il faut déterminer le peigne harmonique de chaque trame voisée qui servira ensuite à extraire les composantes harmoniques de la parole. La fréquence fondamentale est supposée comprise entre 80 et 500Hz. Les paramètres retenus dans cet exemple sont classiques : $Fe = 8$kHz, trame de 256 points, FFT de 512 points et recouvrement de 50%. Pour donner un ordre d'idée, avec ces paramètres, les dents du peigne sont construites comme des multiples de la fréquence fondamentale F_0 dont la valeur peut varier de la $5^{\text{ème}}$ à la $32^{\text{ème}}$ bande de fréquence avec un pas de $\frac{1}{6}$ jugé suffisant pour obtenir une bonne précision. Il existe donc 163 peignes harmoniques possibles.

Nous proposons de chercher le meilleur candidat à partir du signal $S_{harmo}(p,k)$ qui est utilisé comme intermédiaire de calcul dans l'approche HRNR (*cf.* partie 4.5). En effet, ce signal artificiel possède un fort caractère harmonique ce qui va donc faciliter la recherche du peigne correspondant. Nous avons vérifié que cela permet effectivement de limiter les erreurs de détection. Il est toutefois difficile de donner des statistiques précises car cela nécessiterait de marquer la position des harmoniques d'un nombre conséquent de signaux de parole. Le peigne harmonique noté $Y(p,k)$ est défini par $Y(p,k) = 1$ où les harmoniques sont présentes et $Y(p,k) = 0$ ailleurs. Pour déterminer le meilleur candidat, il faut minimiser une distance entre le spectre d'amplitude $|S_{harmo}(p,k)|$ noté $Z(p,k)$ et l'ensemble des peignes disponibles. Cette distance notée $d(Z,Y)$ est bâtie à partir de la norme L_2 :

$$d(Z,Y) = \|Z - gY\| \tag{5.19}$$

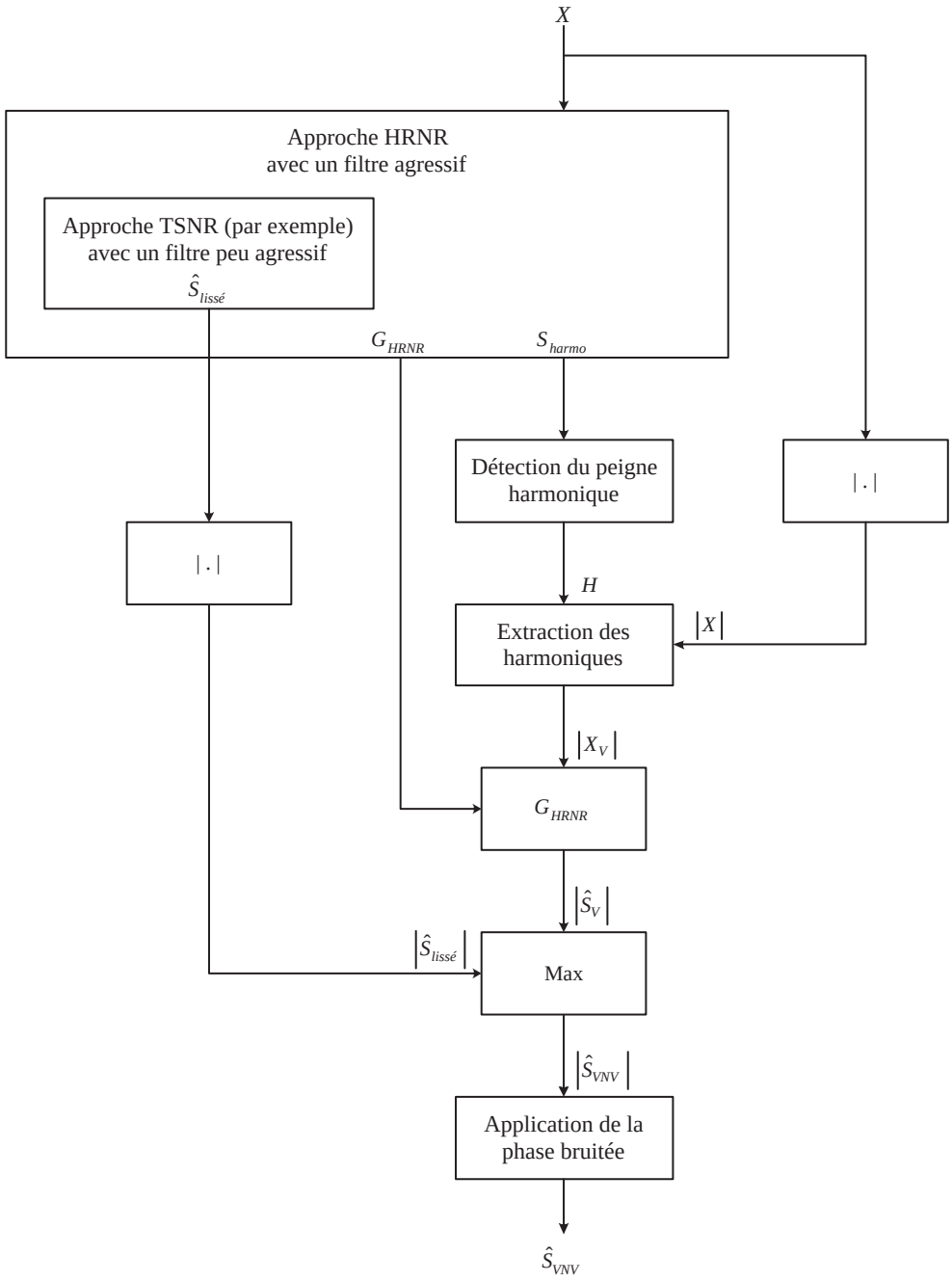

FIG. 5.17 – *Schéma du principe général de l'approche VNV.*

où g est un gain qui permet de normaliser chaque peigne tel que $\|gY\| = 1$ de façon à ce qu'ils aient tous le même poids. Minimiser l'équation (5.19) revient à maximiser le produit scalaire suivant :

$$\frac{< Z, gY >}{\|Z\|\|gY\|} = \frac{< Z, gY >}{\|Z\|}. \tag{5.20}$$

Pour une trame fixée, la quantité $\|Z\|$ est une constante et donc maximiser l'expression ci-dessus revient à maximiser la quantité $< Z, gY >$. Comme Y est un peigne harmonique composé uniquement

de 0 et de 1, on peut écrire que :

$$g = \frac{1}{\|Y\|} = \frac{1}{\sqrt{N_{harmo}}} \tag{5.21}$$

où N_{harmo}, le nombre d'harmoniques composant ce peigne, s'écrit :

$$N_{harmo} = \left[\frac{Fe}{2F_0} + \frac{1}{2} \right]. \tag{5.22}$$

Finalement, le calcul de la quantité $< Z,gY >$ est aisé et peu coûteux car il s'agit tout simplement de la moyenne des N_{harmo} composantes du spectre d'amplitude $|S_{harmo}(p,k)|$ qui vérifient $Y(p,k) = 1$. Le peigne harmonique Y retenu à l'issue de ce processus de décision, *i.e.* celui qui maximise $< Z,gY >$, est alors convolué (dans le domaine fréquentiel) par la TFD de la fenêtre d'analyse (Hanning par exemple) puis normalisé de façon à ce que l'amplitude de la bande de fréquence centrale de chaque harmonique soit égale à 1. Le spectre harmonique ainsi obtenu est noté $H(p,k)$. Ceci permet d'imposer le comportement de la fenêtre d'analyse au peigne harmonique comme c'est le cas pour le signal bruité $X(p,k)$. Finalement, en multipliant le spectre d'amplitude du signal bruité $|X(p,k)|$ par $H(p,k)$, il est possible d'extraire la partie harmonique de ce spectre d'amplitude, toujours bruitée, qui est ensuite restaurée en utilisant l'approche HRNR :

$$|\hat{S}_V(p,k)| = G_{HRNR}(p,k)(H(p,k)|X(p,k)|). \tag{5.23}$$

Cette opération d'extraction est illustrée par la figure 5.18.(b) qui montre que les composantes harmoniques sont extraites avec succès du signal bruité représenté par la figure 5.18.(a) (décrit dans la partie 3.2.1). Cette approche est robuste aux deux types d'erreurs qui peuvent être commises dans l'estima-

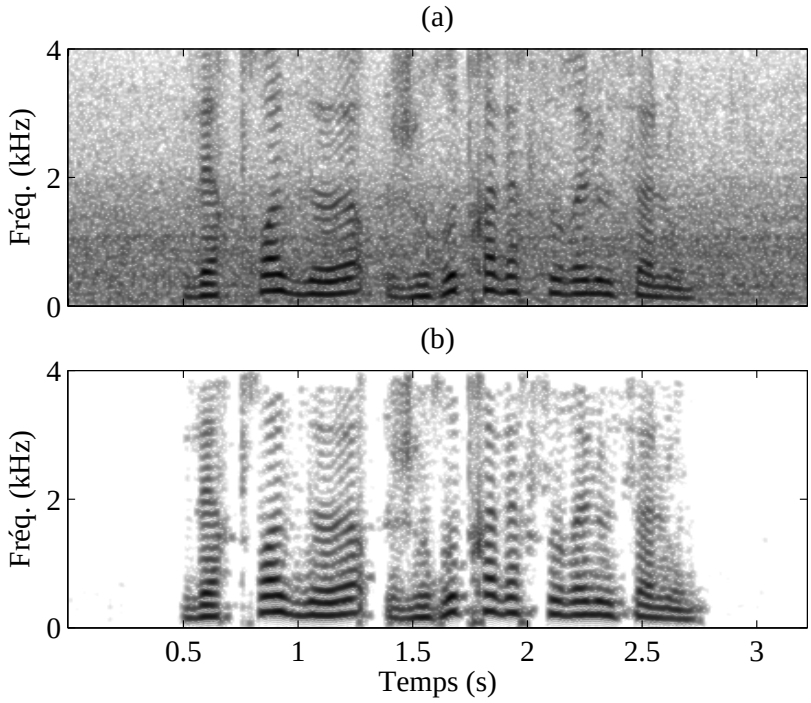

FIG. 5.18 – *Spectrogrammes (a) du signal bruité et (b) des composantes harmoniques restaurées.*

tion du peigne harmonique, à savoir une sous-estimation (généralement $F_0/2$) ou une surestimation (généralement $2F_0$) de la fréquence fondamentale. En effet, dans le cas d'une sous-estimation du type $F_0/2$, le filtre $G_{HRNR}(p,k)$ qui apparaît dans l'équation (5.23) permet de corriger les harmoniques surnuméraires en les atténuant ce qui supprime implicitement le problème. Dans le cas d'une surestimation du type $2F_0$, une harmonique sur deux sera purement et simplement supprimée du signal $\hat{S}_V(p,k)$. Cependant, le spectre d'amplitude final est donné par l'équation (5.17) et donc ses harmoniques ne subiront pas plus de distorsion que celle apportée par l'approche classique (avec un filtre peu agressif). De toute façon ce type d'erreur reste rare étant donné que la recherche des harmoniques est effectuée à partir du signal $S_{harmo}(p,k)$ qui possède un caractère harmonique très marqué.

5.6.3 Illustration de l'approche VNV

Finalement, cette approche VNV permet de recourir à un filtre peu agressif pour les composantes non voisées (ainsi que pour le bruit seul) tandis qu'un comportement agressif est conservé uniquement pour les harmoniques. De cette façon, les dégradations de ces deux types de composantes sont limitées ce qui n'est pas possible avec une approche classique (par exemple TSNR ou HRNR) où il faut toujours faire un compromis. Les résultats de l'approche VNV sont globalement proches de ceux de l'approche HRNR, cependant, on note une amélioration pour les composantes non voisées due au lissage du filtre ainsi que dans certains cas pour les composantes voisées. En effet, l'approche HRNR a tendance à faire remonter le niveau de bruit entre les harmoniques de faibles niveaux. Cet effet est dû à la non-linéarité utilisée pour régénérer les harmoniques et est généralement inaudible surtout lorsqu'un seuillage est appliqué au filtre (*cf.* partie 5.3). Toutefois ce phénomène peut devenir audible lorsque le seuil est choisi très faible comme dans l'exemple de la figure 5.19 (pas de seuil) où ce bruit se traduit par une rugosité de la parole. L'approche VNV permet de suppri-

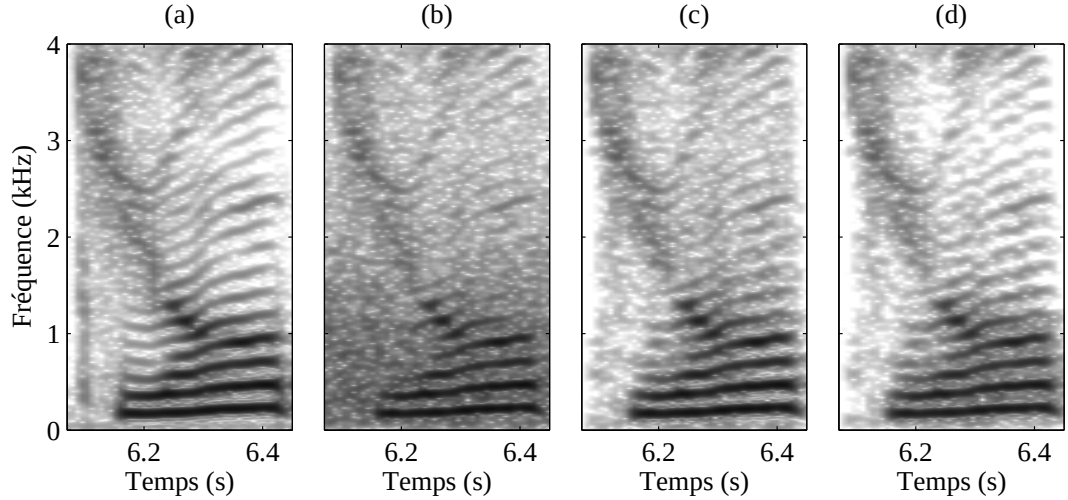

FIG. 5.19 – *Spectrogrammes. (a) Signal de parole propre ; (b) signal de parole bruité ; (c) signal restauré avec l'approche HRNR ; (d) signal restauré avec l'approche VNV.*

mer cet effet puisque entre les harmoniques c'est un filtre classique (TSNR lissé par exemple) qui est utilisé. Cette amélioration est nettement visible en comparant les figures 5.19.(c) et 5.19.(d). Un test

avec des auditeurs experts a montré que l'amélioration apportée par l'approche VNV est sensible et que le signal restauré paraît globalement plus clair.

5.7 Conclusion

Ce chapitre permet de faire une mise au point sur certains aspects de la mise en œuvre des techniques de réduction de bruit susceptibles de provoquer des artefacts ou des distorsions gênantes si celle-ci est mal maîtrisée. On peut noter en particulier qu'une bonne connaissance des techniques de synthèse et de leurs contraintes permet d'éviter certains artefacts gênants ("clics"). D'ailleurs, si leur implémentation est réalisée dans le domaine temporel alors il est possible de choisir librement la fenêtre d'analyse. Si celle-ci est choisie dissymétrique (le poids est mis sur la fin de la trame) alors cela permet de limiter l'effet de réverbération de l'approche DD. Un point important concerne la limitation de l'agressivité du gain spectral qui permet d'obtenir un signal restauré agréable à l'écoute. Toutefois, ceci peut dans certains cas engendrer un phénomène d'étouffement de la parole qu'il est alors possible de supprimer en utilisant l'approche psychoacoustique dans un cadre différent de ce qui est fait classiquement. L'approche VNV joue également sur l'agressivité du gain spectral pour encore améliorer le résultat de l'approche HRNR en réalisant un traitement distinct pour les composantes voisées et non voisées du signal de parole. Un autre point important est la gestion du bruit résiduel qui doit autant que possible avoir la même nature et la même coloration que le bruit original de façon à obtenir un résultat naturel. Cela peut se faire par seuillage du gain spectral dont nous proposons une solution quelque peu différente de celle qui est classiquement réalisée. Les approches proposées dans le chapitre précédent permettent de limiter la distorsion du signal restauré mais il reste difficile de la supprimer complètement. Pour masquer ces distorsions il faut donc réinjecter un niveau de bruit suffisant qui dépend des approches utilisées et surtout de la distorsion qu'elles génèrent comme illustré dans les parties 6.3.6 et 6.3.7 du chapitre suivant.

Références

[Cappé 1993] O. Cappé, “Techniques de Réduction de Bruit pour la Restauration d’Enregistrements Musicaux,” *Thèse de l’École Nationale Supérieure des Télécommunications*, Paris, Septembre 1993.

[Crochiere 1983] R. E. Crochiere, et L. R. Rabiner, “Multirate Digital Signal Processing,” *Prentice-Hall, Première édition*, 1983.

[Guérin 2005] A. Guérin, “Traitement par Blocs : Reconstruction par Méthode Overlap-Save (OLS) et Overlap-Add (OLA),” *Document interne FT R&D*, 2005.

[Kunt 1984] M. Kunt, “Traité d’Électricité - Traitement Numérique des Signaux,” *Presses Polytechniques et universitaires Romandes, Artech House, Troisième édition*, 1998.

[Lin 2002] L. Lin, W. H. Holmes, et E. Ambikairajah, “Speech Denoising Using Perceptual Modification of Wiener Filtering,” *IEEE Electronics Lett.*, Vol. 38, No. 23, pp. 1486–1487, Novembre 2002.

[Marro 1996] C. Marro, “Traitement de Déréverbération et de Débruitage pour le Signal de Parole dans des Contextes de Communication Interactive,” *Thèse de l’Université de Rennes 1*, 1996.

[McAulay 1990] R. J. McAulay, et T. F. Quatieri, “Pitch Estimation and Voicing Detection Based on a Sinusoidal Speech Model,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Albuquerque, États-Unis, Vol. 1, pp. 249–252, Avril 1990.

[Prasanna 2004] S. R. Mahadeva Prasanna, et B. Yegnanarayana, “Extraction of Pitch in Adverse Conditions,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Montréal, Canada, Vol. 1, pp. 109–112, Mai 2004.

Chapitre 6

Évaluation des approches étudiées

Plusieurs techniques de réduction de bruit ont été proposées et analysées dans le chapitre 4. Dans le présent chapitre nous proposons d'évaluer l'amélioration apportée par ces approches à partir d'un corpus représentatif de nombreuses conditions de bruit (*cf.* partie 6.1). Les approches psychoacoustique et VNV qui relèvent plutôt de la mise en œuvre et qui sont présentées dans le chapitre 5 seront également évaluées. Les résultats objectifs seront donnés à partir de deux critères. D'une part, le RSB segmental permettra de quantifier la réduction effective de bruit et d'autre part, la mesure de la distance cepstrale permettra de quantifier la distorsion générée par les techniques de réduction de bruit. Lorsque cela a été jugé nécessaire, des tests subjectifs ont également été réalisés pour affiner les résultats objectifs. La partie 6.2 est consacrée à la présentation des mesures objectives et des tests subjectifs utilisés. Les résultats sont quant à eux présentés et analysés dans la partie 6.3. Il existe une certaine hiérarchie dans les performances de ces approches, comme dans leur conception d'ailleurs, ce qui explique que nous ayons choisi d'analyser les différentes approches en les comparant par paires.

6.1 Description du corpus utilisé

Avant de décrire les mesures objectives et les tests subjectifs destinés à mesurer les performances de la réduction de bruit, nous proposons de décrire le corpus de signaux utilisé. Ce corpus est constitué comme ceci :

- 4 locuteurs : 2 femmes et 2 hommes,
- 9 doubles phrases (séquences) de 8s par locuteur,
- 4 types de bruits : Bureau, Voiture, Rue et Foule,
- 5 valeurs de RSB : 0, 6, 12, 18 et 24dB.

Une condition correspond à un type de bruit et à un RSB fixés, il existe donc 20 conditions différentes. Les résultats objectifs seront donc moyennés sur les 36 doubles phrases disponibles par condition. Ce

corpus (720 doubles phrases) représente une durée totale de 72mn. Les signaux sont échantillonnés à 8kHz et les séquences de parole propre sont normalisées à -22dB par rapport à la valeur de saturation pour assurer un niveau d'écoute constant lors des tests subjectifs. Cette normalisation est réalisée en utilisant l'outil speech voltmeter (SV56) suivant la recommandation ITU-T P.56 [P56 1996]. Les signaux de parole sont des phrases phonétiquement équilibrées enregistrées au calme et les bruits sont enregistrés séparément en condition réelle. Les signaux bruités qui constituent le corpus sont obtenus par addition des signaux de parole et de bruit pondérés de façon à obtenir le RSB souhaité. En pratique, l'outil SV56 [P56 1996] est également utilisé pour fixer le RSB.

Les 4 bruits choisis sont représentatifs des cas susceptibles d'être rencontrés. Pour chacun, leur spectre d'amplitude et leur DSP (la tendance) sont représentés dans la figure 6.1. Les bruits Bureau (ventilateurs) et Voiture (roulement) sont stationnaires et respectent en cela une des hypothèses majeures des techniques de réduction de bruit. Les bruits Rue (voiture, piétons, parole) et surtout Foule (mélange de signaux de parole) sont beaucoup plus critiques dans la mesure où ils ne sont pas stationnaires.

6.2 Outils pour l'analyse objective et méthodes de test subjectif

L'évaluation des performances des techniques de réduction de bruit représente une véritable problématique. En effet, ce sont des êtres humains qui sont amenés à juger de la qualité des traitements, jugement qui est donc par essence subjectif. L'idéal serait donc de pouvoir effectuer à volonté des tests subjectifs formels sur des panels relativement importants de personnes afin d'obtenir un jugement fiable. Malheureusement ceci n'est pas possible pour des raisons évidentes de coût. Il est par contre possible d'utiliser des mesures objectives qui tentent de se rapprocher au mieux de l'évaluation subjective, l'idéal étant de confirmer ces résultats objectifs par des tests subjectifs formels ou bien encore informels, réalisés avec un nombre relativement restreint d'auditeurs experts en écoute.

6.2.1 Critères objectifs de qualité

Une mesure couramment utilisée est le gain sur le RSB qui traduit la puissance de bruit effectivement éliminé en prenant en compte le niveau de bruit résiduel et la dégradation de la parole. En termes de distorsion, la mesure de la distance cepstrale (DC) est très usitée car elle est assez bien corrélée avec les mesures subjectives. De nombreuses autres distances existent mais sont plus difficiles à interpréter [Le Bouquin 1991, Akbari Azirani 1995a]. On peut noter que les mesures objectives présentées ici ne sont applicables que dans le cas où les signaux bruités sont simulés car elles nécessitent la connaissance du signal utile $s(n)$ en plus du signal restauré $\hat{s}(n)$. On a vu dans la partie 5.1.5 que l'opération de filtrage introduit un retard, le signal restauré $\hat{s}(n)$ doit donc bien sûr être au préalable synchronisé avec le signal de parole $s(n)$.

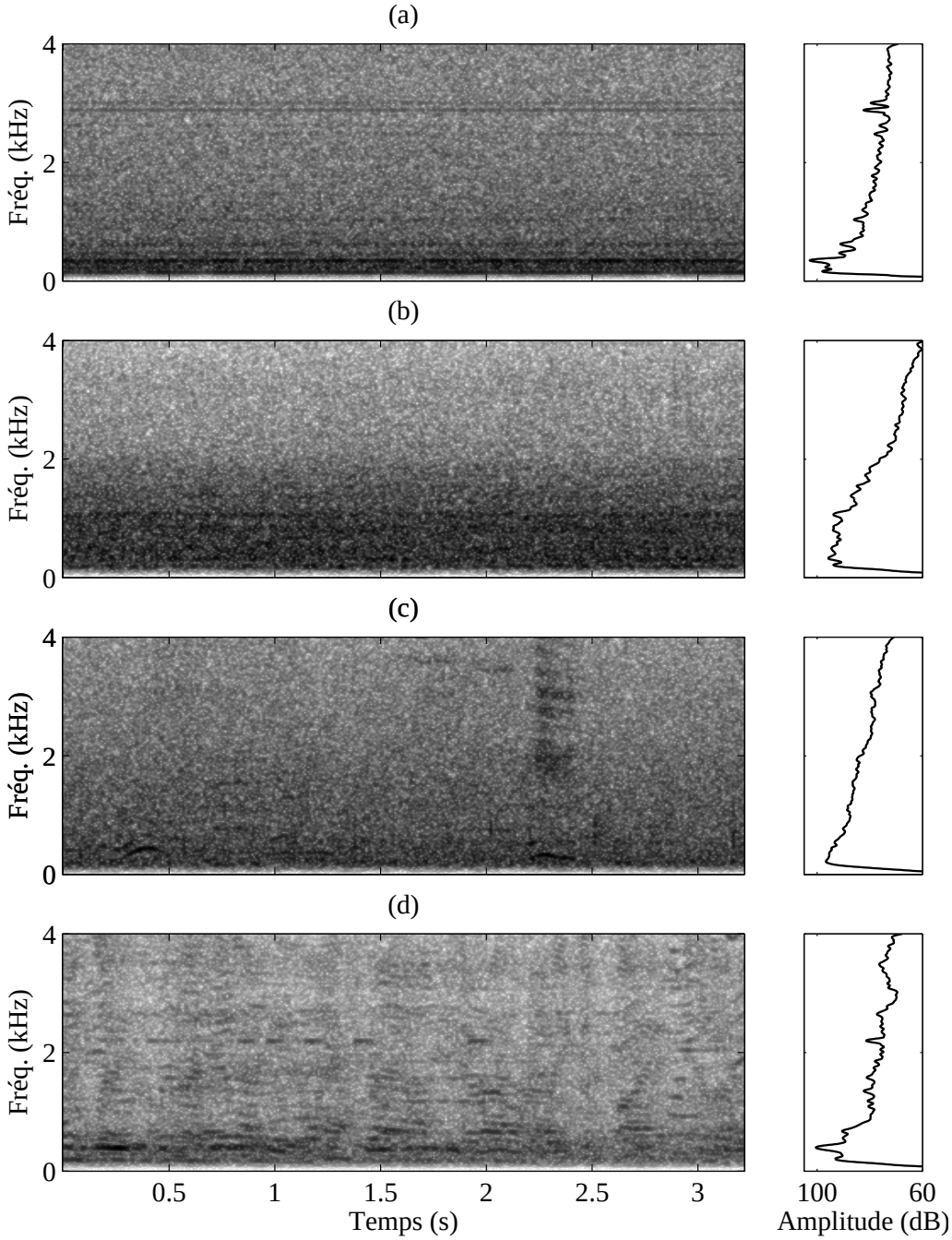

FIG. 6.1 – *Spectres d'amplitudes et DSP des bruits utilisés. (a) Bureau, (b) Voiture, (c) Rue et (d) Foule.*

6.2.1.1 Gain sur le RSB segmental

Le RSB est une mesure à long terme qui se prête peu au signal de parole qui est par essence non-stationnaire. Pour l'évaluation des techniques de réduction de bruit, on lui préfère le RSB segmental [Quackenbush 1988, Faucon 1993] qui est calculé sur des trames d'une durée de quelques dizaines de

ms. Le RSB segmental permet donc de suivre l'évolution du RSB au cours des trames. Afin d'obtenir une valeur représentative et non biaisée de la valeur moyennée du RSB segmental, notée RSB_{seg}, le calcul ne prend pas en compte les trames de silence :

$$RSB_{seg} = \frac{1}{P} \sum_{p \in \mathcal{P}} 10\log_{10} \left(\frac{\sum_{i=0}^{N-1} s^2(Np+i)}{\sum_{i=0}^{N-1} (\hat{s}(Np+i) - s(Np+i))^2} \right) \tag{6.1}$$

où $\mathcal{P}$ représente l'ensemble des indices des trames contenant de la parole et P leur nombre ; N est la taille de la trame comme défini dans la partie 5.1. Dans le cas où on prendrait en compte les trames de silence dans ce calcul et en supposant que l'on supprime complètement le bruit pendant ces périodes (facile à réaliser avec une DAV robuste), cela aurait pour effet d'augmenter artificiellement la valeur du RSB_{seg} du signal de sortie mais aussi de la rendre inexploitable.

Le gain sur le RSB est alors défini à partir du RSB segmental comme la différence entre le RSB en sortie et en entrée :

$$G_{RSB} = \frac{1}{P} \sum_{p \in \mathcal{P}} \left[10\log_{10} \left(\frac{\sum_{i=0}^{N-1} s^2(Np+i)}{\sum_{i=0}^{N-1} (\hat{s}(Np+i) - s(Np+i))^2} \right) - 10\log_{10} \left(\frac{\sum_{i=0}^{N-1} s^2(Np+i)}{\sum_{i=0}^{N-1} b^2(Np+i)} \right) \right]. \tag{6.2}$$

6.2.1.2 Mesure de la distance cepstrale (DC)

La mesure DC est une mesure de similarité entre les cepstres réels de deux signaux qui correspond à une version tronquée de la distance spectrale logarithmique (norme L_2) de ces spectres [Kubichek 1991, Le Bouquin 1991]. L'utilisation de cette distance se justifie par le fait que l'oreille humaine est plus sensible aux variations du logarithme du spectre qu'à celles du spectre lui même. Elle est calculée pour chaque trame p à partir des coefficients cepstraux $c_i(p)$ et $c_i^{'}(p)$ correspondant respectivement aux signaux s et $s^{'}$ à comparer [Le Bouquin 1991, Faucon 1993, Kubichek 1991] :

$$DC(p) = 2 \sum_{i=1}^{M} (c_i(p) - c_i^{'}(p))^2. \tag{6.3}$$

On peut noter que le terme $(c_0(p) - c_0^{'}(p))^2$ qui correspond à une amplification n'est pas pris en compte dans le calcul de la mesure DC. Les variations fines sont d'autant mieux prises en compte que le nombre M de coefficients utilisés est élevé. Ces coefficients cepstraux peuvent être calculés à partir des deux définitions du cepstre : le cepstre Fourier et le cepstre paramétrique. Les modèles étant différents, les cepstres résultants le seront aussi.

– 1. Cepstre Fourier

La suite, notée $c_n(p)$, des coefficients cepstraux $c_i(p)$ obtenus pour la trame $s_p(n)$ sont calculés comme ceci :

$$c_n(p) = IFFT(\ln|FFT(s_p(n))|). \tag{6.4}$$

– 2. Cepstre paramétrique

Dans ce cas, les coefficients cepstraux sont obtenus par l'analyse autorégressive LPC (linear predictive coding) du signal considéré, et ce à l'aide des relations suivantes [Boite 1987] :

$$c_i(p) = -a_i(p) - \sum_{k=1}^{i-1} \left(1 - \frac{k}{i}\right) c_{i-k}(p) a_k(p) \text{ avec } i > 0 \tag{6.5}$$

$$c_0(p) = ln(\sigma^2) \tag{6.6}$$

où σ^2 est la puissance du signal sur la trame considérée et les a_k sont les coefficients de l'analyse LPC. Pour obtenir un résultat précis, M est généralement choisi égal à deux fois l'ordre de l'analyse LPC [Le Bouquin 1991].

De façon à caractériser chaque signal restauré par une seule valeur, la mesure DC peut être moyennée sur l'ensemble des trames d'activité vocale. Les périodes de silence ne sont pas prises en compte de façon à ne pas biaiser le résultat. C'est cette valeur moyenne (la DC est obtenue à partir du cepstre Fourier) qui est utilisée dans la partie 6.3. On peut noter qu'une différence de 0,05 entre deux mesures DC est considérée comme significative (dans notre cas, compte tenu des caractéristiques du corpus utilisé).

6.2.2 Critères subjectifs de qualité

Les mesures objectives sont largement utilisées pour juger de la qualité des traitements. Cependant, elles ne permettent pas de prendre en compte certains effets comme le bruit musical, le naturel du bruit ou la réverbération (due au lissage des estimateurs) qui sont pénalisants dans les tests subjectifs. L'idéal est donc dans la mesure du possible de confirmer les résultats objectifs par des tests subjectifs.

Les tests subjectifs consistent à faire écouter des séquences de signaux restaurés à un panel d'auditeurs dans des conditions bien définies. Ces auditeurs peuvent être choisis "naïfs", c'est-à-dire non habitués aux dégradations susceptibles d'affecter les signaux, ou au contraire "experts", c'est-à-dire habitués au traitement de la parole ou plus généralement du son. Il existe différents types de tests subjectifs, parmi les plus courants, on peut citer les suivants, issus de la recommandation UIT-T P.800 [P800 1996] :

– Le test ACR (Absolute Category Rating) consiste à faire écouter aux auditeurs chaque séquence restaurée sans référence. Ils donnent ensuite une note sanctionnant la qualité de cette séquence et par conséquent le traitement qui l'a généré.
– Le test DCR (Degradation Category Rating) consiste à faire écouter aux auditeurs la séquence à évaluer après un signal de référence. Ils doivent ensuite noter la dégradation de cette séquence par rapport à la référence.
– Le test CCR (Comparison Category Rating) consiste à faire écouter aux auditeurs des paires A-B de séquences. Ils doivent ensuite noter la séquence B par rapport à la séquence A.

Dans le cadre des travaux menés, le test CCR a été retenu car il est reconnu comme le plus discriminant pour évaluer les systèmes de réduction de bruit et qu'il permet de comparer facilement deux traitements. Les consignes pour ce type de test sont de "prendre en compte la distorsion de la parole et le niveau de réduction de bruit" pour donner une note selon l'échelle suivante :

3 Bien meilleure
2 Meilleure
1 Légèrement meilleure
0 A peu près équivalente
−1 Un peu moins bonne
−2 Moins bonne
−3 Beaucoup moins bonne

Les notes d'opinions sont données en score CMOS (pour Comparative Mean Opinion Score). En utilisant une échelle de ce type, les auditeurs formulent implicitement deux jugements avec une seule réponse : "Quel est la séquence de meilleure qualité ?" et "Quelle est la différence de qualité entre les deux séquences ?". Ceci permet donc à l'issue du test de savoir quel est le traitement préféré et d'appréhender la distance de qualité qui les sépare.

Il est possible en s'éloignant quelque peu de la recommandation UIT-T P.800 [P800 1996] de dilater ou de contracter cette échelle selon les besoins. Ainsi, des tests informels ont été réalisés avec le concours des personnes expertes en écoute du laboratoire. Ce sont des tests dits "ABX"qui consistent à donner sa préférence à la séquence A ou B ou à aucune des deux (X ou égal) si l'auditeur ne peut pas les différencier. Cela permet de bien discriminer deux traitements mais ne permet pas d'appréhender la distance entre ceux-ci (note de préférence et non d'opinion). Ce test peut être assimilé à un test CCR dont l'échelle a été réduite au minimum. Il est aussi possible de réaliser des test dits "AB" où le choix des auditeurs est "forcé" dans la mesure où on interdit l'égalité.

On peut noter que, dans ce type de tests (CCR, AB et ABX), les paires A-B de séquences sont présentées dans un ordre aléatoire et autant de fois dans l'ordre A-B que dans l'ordre B-A pour compenser le fait que les auditeurs, si les signaux A et B sont proches, ont tendance à préférer le dernier entendu.

Il faut souligner que, lors des tests subjectifs, seul un nombre restreint de séquences et de conditions sont sélectionnées afin d'éviter que ceux-ci soient rédhibitoires.

6.3 Analyse des résultats

Cette partie est consacrée à l'analyse des résultats d'un certain nombre d'approches proposées et décrites dans les chapitres 4 et 5. Les approches considérées sont les suivantes :

- Approche DD : c'est notre point de départ et la référence initiale de qualité utilisée pour comparer les nouvelles techniques proposées. Cette approche permet de limiter le niveau de bruit musical mais occasionne une distorsion des composantes de parole.
- Approche SG : de nouveaux estimateurs du signal de parole sont obtenus en utilisant des modèles statistiques super-gaussiens plus proches de la réalité que le modèle gaussien classique.

- Approche DDG : cette approche constitue une généralisation de l'estimateur decision-directed qui lui permet de poursuivre les évolutions fréquentielles des composantes de parole.
- Approche TSNR : il s'agit d'une technique en deux passes. La seconde permet de raffiner l'estimation du RSB *a priori* obtenue dans la première qui correspond à l'estimateur DD.
- Approche RFSNR : le but de cette approche est également de raffiner l'estimation du RSB ce qui est réalisé en sélectionnant et en utilisant uniquement les composantes fiables du RSB *a posteriori*. Ces composantes sont obtenues par seuillage des RSB *a posteriori* et *a priori*.
- Approche HRNR : les erreurs d'estimation de la DSP du bruit et l'impact de la phase génèrent de la distorsion harmonique. L'approche proposée ici tire parti du caractère voisé de la parole pour restaurer les harmoniques normalement détruites par des approches classiques.
- Approche VNV : cette technique permet de traiter de façon distincte les composantes voisées et non voisées de la parole. Les composantes voisées sont restaurées en utilisant un filtre agressif alors que les autres le sont en utilisant un filtre lissé. Ceci permet de limiter la présence d'artefacts dans le signal restauré sans dégrader les harmoniques (à cause d'un filtre lissé).
- Approche psychoacoustique : cette approche est à considérer à part, comme une solution permettant d'utiliser un filtre peu agressif (ce qui limite certains artefacts) sans pour autant occasionner un phénomène d'étouffement de la parole.

Il existe des dépendances de conception entre ces approches qui sont matérialisées par une flèche sur le diagramme de la figure 6.2. Chaque nouveau palier vertical indique une amélioration significative en performance (d'un point de vue subjectif). Sur un même palier les différentes approches peuvent donc être considérées comme équivalentes. Chacune des techniques représentées dans la figure 6.2 peut donc être considérée comme l'amélioration d'une autre technique. Les résultats seront par conséquent toujours donnés sous forme comparative entre une approche de référence et sa version améliorée. On peut par ailleurs noter que toutes les approches proposées découlent de près ou de loin de l'approche DD ce qui confirme l'intérêt majeur de cette technique.

Les paramètres choisis pour chacune des approches proposées seront détaillés dans leurs parties respectives. Toutefois, bon nombre de paramètres sont communs à toutes les techniques et il est donc intéressant de les regrouper ici.

Ainsi, le signal bruité est découpé en trames de 256 échantillons (32ms) fenêtrées avec une fenêtre de Hanning. Ces trames se recouvrent avec un taux de 50% compatible avec cette fenêtre d'analyse. Un zero-padding de 256 points est utilisé pour créer des trames de 512 échantillons qui seront analysées dans le domaine fréquentiel par FFT. Ce choix est lié d'une part à la volonté d'améliorer la résolution fréquentielle et d'autre part au fait que l'approche OLA fréquentielle est retenue pour la synthèse (*cf.* partie 5.1) rendant ce zero-padding nécessaire. Avec ces paramètres, la taille maximale du filtre de réduction de bruit, pour éviter le repliement, est donc de 255 coefficients. Cette taille de filtre (filtre agressif) sera retenue pour toutes les approches exceptée l'approche VNV qui fait aussi intervenir un filtre lissé.

La DSP du bruit sera calculée en utilisant l'estimateur récursif présenté dans la partie 2.5.2 avec le paramètre $\lambda_B = 0.8825$ qui assure une constante de temps de 128ms. Cette technique requiert une DAV qui sera toujours supposée idéale de façon à se déconnecter du problème de son estimation. Celle-ci est également nécéssaire afin d'exploiter les critères objectifs de qualité uniquement pendant

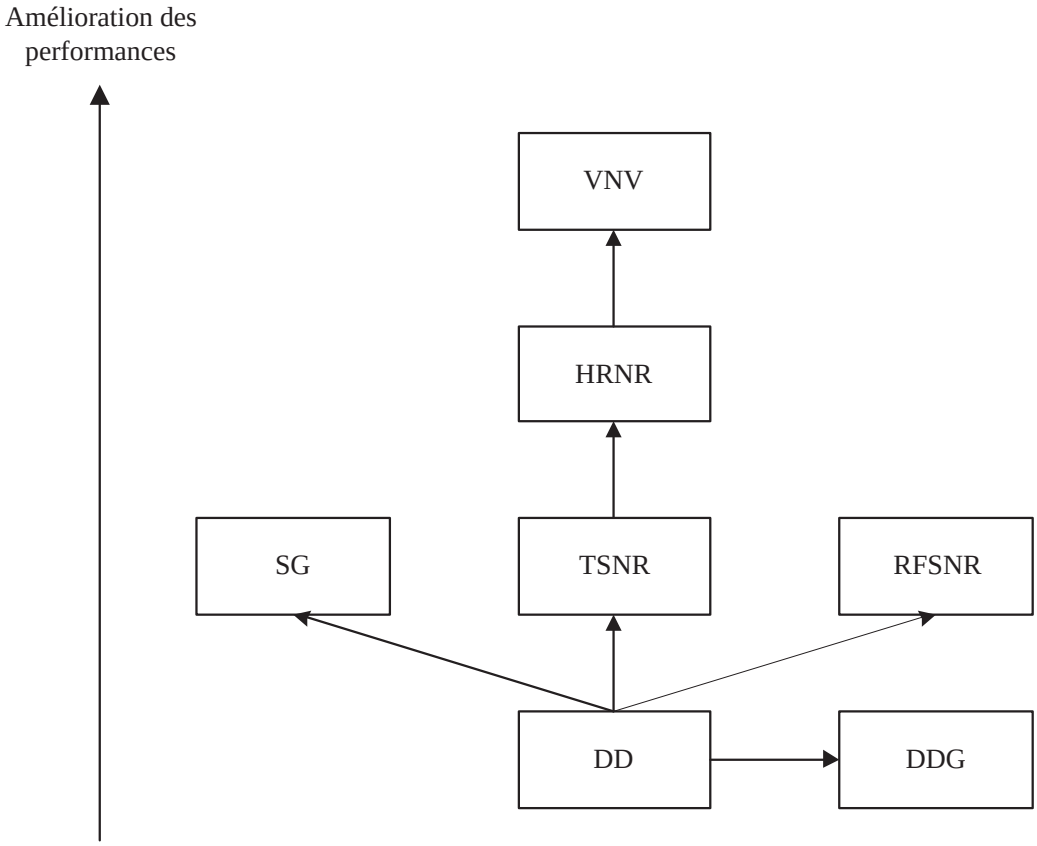

FIG. 6.2 – *Inter-dépendance de conception entre les différentes approches et indication de leur niveau de performance.*

les périodes d'activité vocale. En pratique, n'importe quel estimateur de la DSP du bruit présenté dans la partie 2.5 peut être utilisé et en particulier ceux qui ne nécéssitent pas de DAV car ils sont plus efficaces lorsque le bruit est de type non stationnaire. Les résultats donnés dans la suite permettent donc d'illustrer le comportement des approches proposées pour un estimateur donné de la DSP du bruit.

Pour toutes les approches présentées la fonction de gain retenue est toujours le filtre de Wiener (ou pseudo-Wiener dans le cas de l'approche RFSNR, *cf.* partie 2.4.1.3). Ceci n'est bien entendu pas valable pour l'approche SG qui ne fait pas intervenir de fonction de gain. Dans le cadre des mesures objectives, le gain spectral n'est jamais seuillé (*cf.* partie 5.3) de façon à mesurer les performances réelles des techniques proposées. Il en va autrement pour les tests subjectifs où un seuillage est nécessaire pour assurer la qualité du signal restauré. Les seuils utilisés seront donc précisés dans les parties concernées.

6.3.1 Approche decision-directed (DD)

L'approche DD (*cf.* partie 3.2.5) constitue le point de départ de toutes les approches proposées. Elles sont en effet toutes plus ou moins dépendantes de cette technique qui possède une efficacité

indéniable pour limiter le niveau de bruit musical résiduel. Les approches proposées cherchent donc essentiellement à conserver cet avantage tout en améliorant la qualité du signal de parole restauré, notamment en supprimant le phénomène de réverbération introduit par l'approche DD. La quantité β qui apparaît dans l'équation (3.12) est le paramètre clé de cette approche et est très classiquement choisi égal à 0,98.

6.3.2 Approches super-gaussiennes (SG)

L'approche SG (*cf.* partie 4.1) est un peu particulière dans la mesure où le but n'est pas d'améliorer l'estimation du RSB comme dans la majorité des approches proposées. En effet, partis du constat qu'il existe des modèles mieux adaptés que la loi de Gauss pour les signaux de parole et de bruit, nous avons exprimé des estimateurs de la partie réelle et imaginaire du signal de parole en tenant compte de ces nouveaux modèles. L'estimateur DD intervient dans l'approche SG pour estimer la DSP du signal propre (4.7) nécessaire dans le calcul des estimateurs proposés. Là encore on choisit $\beta = 0{,}98$ dans l'équation (3.12). Sous l'hypothèse Gauss-Gauss (bruit-parole), cet estimateur se réduit tout simplement à l'expression du filtre de Wiener ce qui nous ramène donc à l'approche DD qui sert ici de référence pour les quatre nouvelles techniques nommées : Gauss-Laplace, Gauss-Gamma, Laplace-Laplace et Laplace-Gamma. Le tableau 6.1 regroupe l'ensemble des résultats pour ces approches en terme de RSB segmental. L'approche Gauss-Gamma est supérieure aux autres, y compris à l'approche

TAB. 6.1 – *Tableau comparatif, en terme de RSB segmental, des approches DD (Gauss-Gauss) et SG.*

Bruit	RSB (dB)	RSB segmental pour les approches				
		Gauss Gauss	Gauss Laplace	Gauss Gamma	Laplace Laplace	Laplace Gamma
Bureau	24	24,5	**25,2**	**25,2**	24,5	24,9
	18	19,0	19,9	**20,0**	18,9	19,5
	12	13,6	14,6	**14,8**	13,2	13,9
	6	8,4	9,4	**9,6**	7,8	8,6
	0	4,0	4,7	**5,0**	3,3	4,3
Voiture	24	23,3	24,3	**24,4**	23,2	23,8
	18	18,1	19,1	**19,3**	17,9	18,5
	12	13,1	14,1	**14,4**	12,8	13,5
	6	8,5	9,4	**9,6**	8,0	8,7
	0	4,4	5,1	**5,3**	3,8	4,6
Rue	24	23,3	24,3	**24,4**	23,2	23,8
	18	17,7	18,8	**19,0**	17,5	18,2
	12	12,5	13,5	**13,7**	12,1	12,8
	6	7,5	8,4	**8,7**	6,9	7,6
	0	3,1	3,8	**3,9**	2,5	3,4
Foule	24	24,1	**25,2**	**25,2**	24,1	24,6
	18	18,2	19,3	**19,4**	17,9	18,6
	12	12,5	13,5	**13,7**	12,1	12,8
	6	7,4	8,2	**8,3**	7,0	7,6
	0	3,2	3,6	**3,8**	2,9	3,5

Gauss-Gauss (ou DD). Cela confirme l'analyse réalisée dans la partie 4.1.1.3. En effet, en utilisant cette approche, le biais (effet de réverbération) de l'approche DD est fortement réduit ce qui se traduit

directement par une augmentation du RSB segmental. On peut remarquer que l'approche Gauss-Laplace donne aussi des résultats intéressants alors que les approches Laplace-Laplace et Laplace-Gamma restent en deçà (même si elles restent supérieures à l'approche Gauss-Gauss). On peut donc en déduire que les bonnes performances de l'approche Gauss-Gamma s'expliquent par le fait que la loi Gamma est beaucoup mieux adaptée au signal de parole que les deux autres (Gauss et Laplace) et que la loi de Laplace pour le bruit a finalement tendance à pénaliser les performances en terme de RSB segmental.

Il en va autrement en terme de distorsion, mesurée par la distance cepstrale, dont les résultats sont regroupés dans le tableau 6.2. En effet, on remarque que pour les bruits stationnaires (Bureau et Voi-

TAB. 6.2 – *Tableau comparatif, en terme de distance cepstrale, des approches DD (Gauss-Gauss) et SG.*

Bruit	RSB (dB)	Distance cepstrale pour les approches				
		Gauss Gauss	Gauss Laplace	Gauss Gamma	Laplace Laplace	Laplace Gamma
Bureau	24	0,22	0,21	**0,19**	0,25	0,20
	18	0,36	0,34	**0,31**	0,39	0,33
	12	0,50	0,48	**0,45**	0,52	0,47
	6	0,65	0,63	**0,61**	0,65	**0,61**
	0	0,83	0,82	0,82	0,81	**0,77**
Voiture	24	0,30	0,28	**0,24**	0,34	0,26
	18	0,42	0,40	**0,35**	0,46	0,38
	12	0,53	0,51	**0,49**	0,56	0,52
	6	0,64	0,62	**0,60**	0,66	0,62
	0	0,81	0,80	0,77	0,81	**0,74**
Rue	24	0,39	0,38	**0,33**	0,42	0,35
	18	0,58	0,57	**0,52**	0,60	0,53
	12	0,79	0,78	0,74	0,78	**0,72**
	6	1,04	1,04	1,01	1,00	**0,96**
	0	1,37	1,35	1,36	1,31	**1,25**
Foule	24	0,33	0,32	**0,27**	0,35	0,28
	18	0,54	0,52	0,47	0,56	**0,46**
	12	0,81	0,78	0,74	0,81	**0,72**
	6	1,12	1,09	1,08	1,11	**1,04**
	0	1,51	1,46	1,48	1,48	**1,41**

ture) l'approche Gauss-Gamma donne les meilleurs résultats (sauf dans le cas 0dB). Ceci s'explique par le fait que la loi de Gauss est mieux adaptée que la loi de Laplace aux bruits stationnaires comme démontré dans la partie 3.1. Pour les bruits non-stationnaires (Rue et surtout Foule), la loi de Laplace est la mieux adaptée et en effet, c'est alors l'estimateur Laplace-Gamma qui obtient les meilleurs résultats (excepté pour les forts RSB). La supériorité de la loi Gamma pour modéliser la parole n'est donc pas démentie par cette mesure de distorsion qui permet par contre d'obtenir un résultat plus fin et de montrer ainsi que bien modéliser le bruit se révèle également important. Ces résultats sont donc cohérents avec l'analyse de la partie 3.1 et confirment que l'utilisation de meilleurs modèles pour les signaux de parole et de bruit permet d'apporter une amélioration aux signaux restaurés.

Aucun test subjectif formel n'a été réalisé dans la mesure où des écoutes expertes ont montré que l'approche TSNR donne de meilleurs résultats pour un coût en calcul nettement moindre (*cf.* résultats

de la partie 6.3.4). Le point faible des approches utilisant des lois super-gaussiennes est en effet leur complexité rédhibitoire (*cf.* partie 1.2.2).

6.3.3 Approche decision-directed généralisée (DDG)

Le principe de l'approche DDG (*cf.* partie 4.2) consiste à rendre l'estimateur decision-directed capable de suivre les évolutions fréquentielles des composantes voisées de parole. Cependant, l'amélioration apportée par cette technique est insuffisante pour se démarquer subjectivement de l'approche DD. De plus, la mise en œuvre dans des conditions réelles se révèle irréaliste (*cf.* partie 4.2), par conséquent, aucun résultat ne sera donné pour cette approche. On précisera tout de même que là encore le paramètre β contrôlant l'estimateur DDG est égal à 0,98 dans l'équation (4.27).

6.3.4 Approche "two-step noise reduction" (TSNR)

Le but de l'approche TSNR (*cf.* partie 4.3) est de supprimer le biais de l'approche DD tout en conservant sa capacité à limiter le niveau de bruit musical résiduel. La première passe est donc un estimateur DD pour lequel on choisit tout à fait classiquement $\beta = 0{,}98$ dans l'équation (3.12). La seconde passe qui est aussi un estimateur DD permet alors de réestimer le RSB *a priori* mais dans le cas particulier où $\beta' = 1$ dans l'équation (4.29) ce qui conduit donc à l'expression (4.30).

L'approche TSNR donne systématiquement de meilleurs résultats que l'approche DD en terme de RSB segmental (*cf.* tableau 6.3). Cela s'explique par sa capacité à supprimer efficacement le biais

TAB. 6.3 – *Tableau comparatif, en terme de RSB segmental, des approches DD et TSNR.*

Bruit	RSB (dB)	RSB segmental pour les approches	
		DD	TSNR
Bureau	24	24,4	**25,0**
	18	19,0	**19,8**
	12	13,6	**14,5**
	6	8,4	**9,2**
	0	3,9	**4,4**
Voiture	24	23,3	**24,1**
	18	18,1	**18,9**
	12	13,1	**14,0**
	6	8,5	**9,3**
	0	4,4	**5,0**
Rue	24	23,2	**24,0**
	18	17,7	**18,6**
	12	12,4	**13,3**
	6	7,4	**8,2**
	0	3,0	**3,4**
Foule	24	24,1	**24,8**
	18	18,1	**19,0**
	12	12,5	**13,3**
	6	7,4	**7,9**
	0	3,2	**3,4**

de l'approche DD. Le filtre de réduction de bruit préserve donc mieux le signal de parole, en particulier lors des attaques et extinctions, ce qui se traduit naturellement par une augmentation du RSB segmental.

Cependant, en terme de distorsion l'approche DD donne systématiquement les meilleurs résultats (*cf.* tableau 6.4). Ceci peut paraître paradoxal dans la mesure où, on l'a vu dans la partie 4.3, l'ap-

TAB. 6.4 – *Tableau comparatif, en terme de distance cepstrale, des approches DD et TSNR.*

Bruit	RSB (dB)	Distance cepstrale pour les approches	
		DD	TSNR
Bureau	24	**0,22**	0,28
	18	**0,35**	0,43
	12	**0,49**	0,58
	6	**0,63**	0,73
	0	**0,78**	0,88
Voiture	24	**0,29**	0,34
	18	**0,40**	0,46
	12	**0,51**	0,60
	6	**0,62**	0,75
	0	**0,75**	0,89
Rue	24	**0,37**	0,42
	18	**0,53**	0,58
	12	**0,68**	0,74
	6	**0,84**	0,90
	0	**1,01**	1,04
Foule	24	**0,31**	0,34
	18	**0,48**	0,52
	12	**0,67**	0,70
	6	**0,86**	0,89
	0	**1,06**	1,09

proche TSNR préserve mieux le signal de parole. Cela s'explique tout simplement par le fait que la mesure DC n'est pas sensible à la différence d'énergie globale de deux trames (*cf.* partie 6.2.1.2). Ainsi, une trame qui serait globalement sous-estimée de 17dB (sous-estimation maximale occasionnée par l'approche DD) n'entraîne pas pour autant une augmentation de la mesure DC. De plus, la surestimation occasionnée par l'approche DD (pour les extinctions par exemple) réduit naturellement la dégradation de la parole en limitant le niveau de réduction de bruit. Par contre, cette mesure ne permet manifestement pas de quantifier le gain subjectif apporté par la suppression de l'effet de réverbération de l'approche DD. Des écoutes informelles ont été jugées suffisantes pour valider l'amélioration apportée par l'approche TSNR, nous ne disposons donc pas de résultats de test subjectif pour cette approche.

6.3.5 Approche “reliable features selection noise reduction” (RFSNR)

L’analyse de l’approche TSNR (*cf.* partie 4.3.2) montre qu’une partie du biais de l’approche DD subsiste pour les composantes de parole dont le RSB *a posteriori* est faible. Un tel comportement est nécessaire pour obtenir une bonne suppression du bruit musical résiduel. Le but de l’approche RFSNR (*cf.* partie 4.4) est de supprimer complètement le biais de l’approche DD tout en assurant une suppression efficace du bruit musical. Bien que basée sur l’approche DD (avec $\beta = 0{,}98$ dans l’équation (3.12)), cette nouvelle technique sera donc comparée avec l’approche TSNR. Les deux seuils permettant de sélectionner les composantes valides du RSB *a posteriori* sont les suivants : $\eta = -6$dB et $\delta = -6$dB.

En terme de RSB segmental (*cf.* tableau 6.5), l’approche RFSNR donne de meilleurs résultats que l’approche TSNR, excepté pour les forts RSB. Cependant, dans tous les cas, les différences sont faibles. Cela s’explique par le fait que cette nouvelle approche apporte une amélioration (suppression

TAB. 6.5 – *Tableau comparatif, en terme de RSB segmental, des approches TSNR et RFSNR.*

Bruit	RSB (dB)	RSB segmental pour les approches	
		TSNR	RFSNR
Bureau	24	**25,0**	24,7
	18	**19,8**	19,7
	12	14,5	**14,6**
	6	9,2	**9,5**
	0	4,4	**4,7**
Voiture	24	**24,1**	24,0
	18	**18,9**	**18,9**
	12	14,0	**14,1**
	6	9,3	**9,7**
	0	5,0	**5,2**
Rue	24	**24,0**	23,9
	18	**18,6**	**18,6**
	12	13,3	**13,5**
	6	8,2	**8,4**
	0	3,4	**3,7**
Foule	24	**24,8**	24,7
	18	19,0	**19,1**
	12	13,3	**13,5**
	6	7,9	**8,1**
	0	3,4	**3,5**

du biais de l’approche DD) uniquement pour les faibles composantes de signal ou en tout cas celles qui présentent un RSB relativement faible. L’amélioration en terme de RSB segmental reste donc peu importante.

En terme de distorsion, la différence entre les approches TSNR et RFSNR est également faible (*cf.* tableau 6.6). Malgré tout, on peut observer une certaine tendance : l’approche RFSNR semble générer moins de dégradations pour les RSB importants et plus de dégradations pour les RSB faibles. Cependant, les différences entre ces deux approches ne sont pas suffisamment significatives pour être prises en compte. D’un point de vue subjectif, même s’il existe une différence entre ces deux approches

TAB. 6.6 – *Tableau comparatif, en terme de distance cepstrale, des approches TSNR et RFSNR.*

Bruit	RSB (dB)	Distance cepstrale pour les approches	
		TSNR	RFSNR
Bureau	24	0,28	**0,25**
	18	0,43	**0,41**
	12	**0,58**	**0,58**
	6	**0,73**	0,76
	0	**0,88**	0,91
Voiture	24	0,34	**0,31**
	18	0,46	**0,44**
	12	0,60	**0,59**
	6	**0,75**	0,76
	0	**0,89**	0,93
Rue	24	0,42	**0,40**
	18	0,58	**0,57**
	12	**0,74**	0,76
	6	**0,90**	0,93
	0	**1,04**	1,07
Foule	24	0,34	**0,33**
	18	0,52	**0,50**
	12	**0,70**	**0,70**
	6	**0,89**	0,90
	0	**1,09**	1,11

celle-ci reste très faible. Aucun test subjectif formel ne sera donc réalisé car on peut considérer les approches TSNR et RFSNR comme globalement très proches.

6.3.6 Approche “harmonic regeneration noise reduction” (HRNR)

La partie 3.3 met en avant que les techniques d’estimation du bruit dont on dispose ne sont pas assez performantes constituant ainsi une limite à la performance des techniques de réduction de bruit. À cela s’ajoute le fait que la phase ne peut pas être prise en compte dans l’estimation des quantités impliquées dans le calcul du signal restauré (*cf.* partie 3.4). La distorsion générée à cause de ces deux limitations se traduit majoritairement par une distorsion des harmoniques du signal de parole étant donné qu’en moyenne 80% des sons prononcés sont voisés. L’approche HRNR (*cf.* partie 4.5) permet d’aller plus loin que les techniques conventionnelles en restaurant les harmoniques que des techniques classiques détruisent. L’approche TSNR, dont les paramètres sont précisés dans la partie 6.3.4, est choisie comme base de l’approche HRNR (*cf.* partie 4.5). Ces deux techniques seront donc comparées de façon à quantifier l’amélioration apportée par l’approche HRNR. La fonction non-linéaire retenue pour l’équation (4.47) est le redressement mono-alternance ou autrement dit $\max(.,0)$. Le paramètre de mixage intervenant dans l’équation (4.48) est quant à lui choisi égal au gain de l’approche TSNR : $\rho(p,k) = G_{TSNR}(p,k)$.

L’approche HRNR apporte systématiquement une amélioration en terme de RSB segmental (*cf.* tableau 6.7). Ceci s’explique par le fait que des harmoniques sont régénérées de manière efficace sans affecter les composantes déjà restaurées par l’approche TSNR. En effet, le paramètre de mixage

TAB. 6.7 – *Tableau comparatif, en terme de RSB segmental, des approches TSNR et HRNR.*

Bruit	RSB (dB)	RSB segmental pour les approches	
		TSNR	HRNR
Bureau	24	25,0	**25,1**
	18	19,8	**20,2**
	12	14,5	**15,0**
	6	9,2	**9,8**
	0	4,4	**4,7**
Voiture	24	24,1	**24,7**
	18	18,9	**19,7**
	12	14,0	**14,8**
	6	9,3	**9,9**
	0	5,0	**5,3**
Rue	24	24,0	**24,6**
	18	18,6	**19,4**
	12	13,3	**14,2**
	6	8,2	**8,8**
	0	3,4	**3,7**
Foule	24	24,8	**25,7**
	18	19,0	**20,0**
	12	13,3	**14,2**
	6	7,9	**8,5**
	0	3,4	**3,7**

$\rho(p,k)$ (*cf.* partie 4.5.1) autorise la régénération d'harmonicité seulement dans les zones où les harmoniques ont potentiellement été détruites.

Ce comportement est confirmé par la mesure de distorsion (*cf.* tableau 6.8) qui montre que celle-ci est significativement réduite par l'approche HRNR dans toutes les conditions. À 0dB toutefois, l'efficacité de cette approche se trouve réduite car le signal fourni par l'approche TSNR est alors trop dégradé pour permettre une régénération efficace des harmoniques (*cf.* partie 4.5.2).

Un test subjectif formel a été mis en place dans le but de valider ces résultats objectifs et de quantifier l'amélioration apportée par l'approche HRNR. Il s'agit d'un test CCR (*cf.* partie 6.2.1) qui a été réalisé avec 24 auditeurs dont 16 naïfs et 8 experts. Les écoutes ont été faites sur combiné téléphonique ; les signaux ont donc été préalablement filtrés SRI (système de référence intermédiaire) suivant la recommandation ITU-T P.48 [P48 1988] pour simuler une communication téléphonique sur réseau filaire RTC. Ces tests ont été réalisés suivant la recommandation UIT-T P.800 [P800 1996], cependant, une échelle allant de -5 à 5 (au lieu de -3 à 3) a été choisie de façon a donner plus de précision (de dynamique) aux notes des auditeurs. À partir du corpus disponible, décrit dans la partie 6.1, il a fallu sélectionner un nombre limité de conditions de façon à obtenir un test de durée raisonnable (45mn environ). Les bruits Voiture, Rue et Foule ont été retenus ainsi que les RSB 12, 18 et 24dB, les deux autres niveaux (0 et 6dB) ayant été jugés trop critiques. Finalement, pour chacune de ces 9 conditions, 4 séquences ont été sélectionnées de façon à les représenter de manière équilibrée.

Rappelons que l'approche HRNR permet de limiter la distorsion du signal de parole. En pratique il est donc possible de supprimer plus de bruit qu'avec l'approche TSNR à distorsion égale (*cf.* partie 4.5). Dans ce test, le niveau de réduction de bruit de l'approche TSNR a donc été limité à 12dB

TAB. 6.8 – *Tableau comparatif, en terme de distance cepstrale, des approches TSNR et HRNR.*

Bruit	RSB (dB)	Distance cepstrale pour les approches	
		TSNR	HRNR
Bureau	24	0,28	**0,19**
	18	0,43	**0,33**
	12	0,58	**0,46**
	6	0,73	**0,64**
	0	0,88	**0,83**
Voiture	24	0,34	**0,22**
	18	0,46	**0,32**
	12	0,60	**0,44**
	6	0,75	**0,62**
	0	0,89	**0,85**
Rue	24	0,42	**0,31**
	18	0,58	**0,44**
	12	0,74	**0,61**
	6	0,90	**0,81**
	0	1,04	**1,00**
Foule	24	0,34	**0,25**
	18	0,52	**0,40**
	12	0,70	**0,58**
	6	0,89	**0,79**
	0	1,09	**1,03**

pour toutes les conditions. Cela permet d'obtenir un signal restauré avec un niveau acceptable de distorsion. Pour l'approche HRNR, le niveau de réduction de bruit a été limité respectivement à 17, 19 et 21dB pour les RSB 12, 18 et 24dB. Les auditeurs devaient donner une note globale tenant compte d'une part de la distorsion de la parole et d'autre part de la réduction de bruit. Les résultats obtenus sont représentés sur la figure 6.3. Une note CMOS positive indique que l'approche HRNR est préférée à l'approche TSNR. L'approche HRNR est systématiquement préférée avec une note

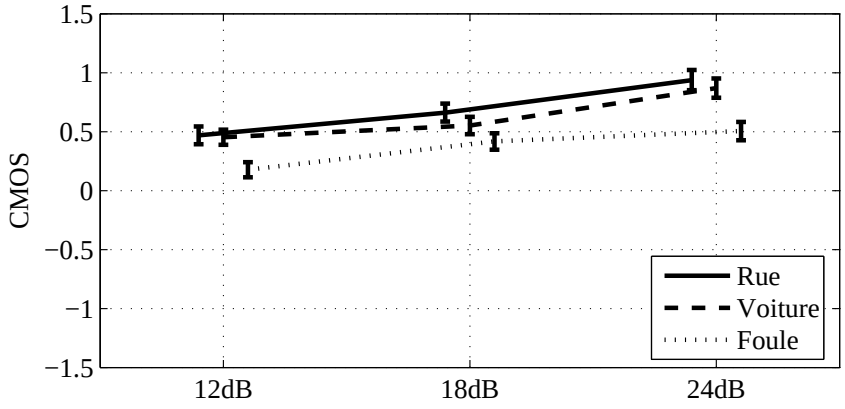

FIG. 6.3 – *Résultats du test CCR formel permettant de comparer les approches TSNR et HRNR. La note CMOS ainsi que les intervalles de confiance sont représentés pour 3 niveaux de RSB (12, 18 et 24dB) et 3 types de bruit (Rue, Voiture et Foule).*

CMOS significative ce qui confirme les résultats objectifs. On peut remarquer que l'amélioration est

moindre pour le Foule ce qui n'est pas étonnant étant donné la nature hautement non-stationnaire de ce bruit (*cf.* partie 3.3). Malgré cela, il existe un comportement commun à tous les types de bruit : les performances de l'approche HRNR augmentent avec le RSB ce qui confirme l'analyse de la partie 4.5.2. En effet, plus le RSB augmente, plus l'approche HRNR permet de régénérer efficacement les harmoniques détruites et dégradées par l'approche TSNR.

6.3.7 Approche voisé-non voisé (VNV)

Il est possible de supprimer certaines distorsions et artefacts du signal restauré en ayant recours à un filtre lissé (*cf.* partie 5.4). Cependant, le prix à payer est l'introduction d'un phénomène d'étouffement de la parole lié à l'atténuation de certaines harmoniques (*cf.* figure 5.13). L'approche VNV (*cf.* partie 5.6) permet de pallier ce problème en conservant le comportement de l'approche HRNR avec un filtre agressif pour les composantes harmoniques et le comportement de l'approche TSNR avec un filtre lissé pour les autres composantes. Le signal $\hat{S}_{liss}(p,k)$ qui apparaît dans l'équation (5.17) est obtenu en utilisant l'approche TSNR telle que décrite dans la partie 6.3.4 mais avec un filtre lissé de seulement 65 coefficients au lieu de 255. L'approche HRNR est quant à elle utilisée pour l'extraction et l'estimation de la partie voisée $\hat{S}_V(p,k)$ de la parole propre (*cf.* équation (5.17)) avec les paramètres donnés dans la partie 6.3.4.

On note une diminution assez conséquente du RSB segmental dans toutes les conditions (*cf.* tableau 6.9). Ce phénomène est lié à l'emploi du filtre lissé pour certaines composantes et notamment

TAB. 6.9 – *Tableau comparatif, en terme de RSB segmental, des approches HRNR et VNV.*

Bruit	RSB (dB)	RSB segmental pour les approches	
		HRNR	VNV
Bureau	24	**25,1**	23,6
	18	**20,2**	18,0
	12	**15,0**	12,9
	6	**9,8**	8,2
	0	**4,7**	3,9
Voiture	24	**24,7**	23,9
	18	**19,7**	18,4
	12	**14,8**	13,4
	6	**9,9**	8,8
	0	**5,3**	4,7
Rue	24	**24,6**	23,6
	18	**19,4**	17,9
	12	**14,2**	12,6
	6	**8,8**	7,7
	0	**3,7**	3,3
Foule	24	**25,7**	24,6
	18	**20,0**	18,5
	12	**14,2**	12,8
	6	**8,5**	7,7
	0	**3,7**	3,4

celles qui se trouvent entre les harmoniques. En effet, comme le montre la figure 5.13, le lissage du

filtre entraîne une augmentation conséquente du gain entre les harmoniques ce qui occasionne donc une remontée du niveau de bruit. Celui-ci reste inaudible car masqué par le signal de parole mais cela se ressent tout de même sur le RSB segmental.

On note aussi une augmentation systématique de la mesure de distorsion (*cf.* tableau 6.10). Cette augmentation reste malgré tout trop faible pour être significative.

TAB. 6.10 – *Tableau comparatif, en terme de distance cepstrale, des approches HRNR et VNV.*

Bruit	RSB (dB)	Distance cepstrale pour les approches	
		HRNR	VNV
Bureau	24	**0,19**	0,20
	18	**0,33**	0,34
	12	**0,46**	0,50
	6	**0,64**	0,68
	0	**0,83**	0,87
Voiture	24	**0,22**	0,24
	18	**0,32**	0,34
	12	**0,44**	0,46
	6	**0,62**	0,63
	0	**0,85**	0,84
Rue	24	**0,31**	0,32
	18	**0,44**	0,46
	12	**0,61**	0,63
	6	**0,81**	0,83
	0	**1,00**	1,02
Foule	24	**0,25**	0,26
	18	**0,40**	0,42
	12	**0,58**	0,61
	6	**0,79**	0,81
	0	**1,03**	**1,03**

Les résultats objectifs ne sont pas en faveur de l'approche VNV, un test subjectif informel a donc été réalisé pour trancher entre les approches HRNR et VNV. Il s'agit d'un test ABX (*cf.* partie 6.2.1) qui a été réalisé avec 10 auditeurs experts ce qui donne du poids et de la fiabilité aux résultats de ces tests informels. Les écoutes ont été faites au casque. À partir du corpus disponible, décrit dans la partie 6.1, un nombre restreint de conditions ont été sélectionnées de façon à obtenir un test court (15mn environ). Les bruits Voiture et Foule ont été retenus ainsi que les RSB 12 et 18dB. Finalement, 3 séquences pour chacune de ces 4 conditions ont été sélectionnées. Les niveaux de réduction de bruit ont été limités de la même façon pour les deux approches à savoir respectivement 17 et 19dB pour les RSB de 12 et 18dB (*cf.* description du test formel de la partie 6.3.6). Les résultats obtenus sont présentés dans le tableau 6.11. La préférence va toujours en faveur de l'approche VNV et l'opinion des auditeurs recueillie à l'issue du test confirme que grâce à cette technique le signal de parole paraît globalement plus clair (*cf.* partie 5.6.3).

TAB. 6.11 – *Résultats du test AB permettant de comparer les approches HRNR et VNV. La préférence pour l'approche VNV est indiquée en pourcentage pour 2 niveaux de RSB (12 et 18dB) et 2 types de bruit (Voiture et Foule).*

	Type de bruit	
Niveau (en dB)	Voiture	Foule
12	69%	60%
18	56%	58%

6.3.8 Introduction de la psychoacoustique dans l'approche TSNR

L'approche psychoacoustique qui est classiquement utilisée pour limiter le phénomène de bruit musical ainsi que la distorsion de la parole est utilisée ici dans un objectif différent. En effet, en couplant l'approche psychoacoustique avec l'utilisation d'un filtre lissé (*cf.* partie 5.5) il est possible de limiter les artefacts liés à un filtre trop agressif tout en évitant l'effet d'étouffement de la parole dû au lissage du filtre (*cf.* partie 5.5). L'approche TSNR est choisie pour estimer le seuil de masquage puis pour calculer la quantité $\hat{RSB}^{aud}_{prio}(p,k)$ utilisée dans l'équation (5.16) qui permet finalement d'obtenir le gain spectral. L'agressivité de celui-ci est alors réduite en contraignant la taille du filtre à 65 coefficients seulement.

TAB. 6.12 – *Tableau comparatif, en terme de RSB segmental, de l'approche TSNR avec et sans psychoacoustique.*

Bruit	RSB (dB)	RSB segmental pour les approches	
		TSNR sans psychoacoustique	TSNR avec psychoacoustique
Bureau	24	**25,0**	24,5
	18	**19,8**	18,8
	12	**14,5**	13,5
	6	**9,2**	9,0
	0	4,4	**4,6**
Voiture	24	24,1	**24,5**
	18	18,9	**19,1**
	12	**14,0**	13,9
	6	**9,3**	9,2
	0	**5,0**	**5,0**
Rue	24	24,0	**24,5**
	18	18,6	**18,8**
	12	13,3	**13,4**
	6	**8,2**	**8,2**
	0	3,4	**3,5**
Foule	24	24,8	**26,5**
	18	19,0	**20,4**
	12	13,3	**14,3**
	6	7,9	**8,3**
	0	3,4	**3,5**

Pour l'analyse objective, nous avons choisi de conserver un filtre agressif de façon à mesurer

l'impact réel de l'approche psychoacoustique. Dans le cas des bruits stationnaires (Voiture et surtout Bureau), l'avantage va plutôt à l'approche sans psychoacoustique (*cf.* tableau 6.12). Ceci est dû au fait que la partie masquée du bruit n'est pas supprimée. En effet, le RSB segmental étant une mesure énergétique il ne peut pas tenir compte du fait qu'en réalité ce bruit est masqué. De plus, pour ce type de bruit la distorsion du signal de parole reste relativement faible et l'amélioration apportée aux composantes de parole ne compense donc pas complètement la remontée du niveau de bruit résiduel. Par contre, on remarque que pour les bruits non-stationnaires (Rue et Foule), l'approche psychoacoustique donne systématiquement les meilleurs résultats (*cf.* tableau 6.12). Ceci s'explique par le fait que ces types de bruit occasionnent beaucoup de distorsions qui sont supprimées en utilisant le masquage psychoacoustique. Dans ce cas, il apparaît que l'amélioration apportée aux composantes de parole est plus importante que la remontée du niveau de bruit résiduel.

En terme de distorsion, la tendance qui ressort est une amélioration pour les forts RSB et une dégradation pour les faibles RSB (*cf.* tableau 6.13). Toutefois les différences sont trop minimes pour être

TAB. 6.13 – *Tableau comparatif, en terme de distance cepstrale, de l'approche TSNR avec et sans psychoacoustique.*

Bruit	RSB (dB)	Distance cepstrale pour les approches	
		TSNR sans psychoacoustique	TSNR avec psychoacoustique
Bureau	24	0,28	**0,25**
	18	0,43	**0,41**
	12	**0,58**	0,59
	6	**0,73**	0,76
	0	**0,88**	0,90
Voiture	24	0,34	**0,30**
	18	0,46	**0,42**
	12	0,60	**0,58**
	6	**0,75**	0,76
	0	**0,89**	0,92
Rue	24	0,42	**0,39**
	18	0,58	**0,56**
	12	**0,74**	0,75
	6	**0,90**	0,92
	0	**1,04**	1,07
Foule	24	0,34	**0,30**
	18	0,52	**0,48**
	12	0,70	**0,68**
	6	**0,89**	**0,89**
	0	**1,09**	1,10

significatives. En pratique, on s'éloigne du cadre de cette analyse objective car l'intérêt de l'approche TSNR avec psychoacoustique est de supprimer l'effet d'étouffement constaté à cause du lissage du filtre de réduction de bruit. Des écoutes informelles ont permis de confirmer que ce but est effectivement atteint avec une amélioration nette pour les voix féminines particulièrement touchées par ce phénomène d'étouffement (*cf.* partie 5.5).

6.3.9 Compilation des résultats

Nous proposons avec les tableaux 6.14 et 6.15 une compilation des résultats obtenus par les approches DD, TSNR, RFSNR, HRNR et VNV afin d'avoir une vision globale de ces différentes approches. On peut juste noter qu'en terme de distorsion toutes les techniques de réduction de bruit

TAB. 6.14 – *Tableau récapitulatif, en terme de RSB segmental, des approches DD, TSNR, RFSNR, HRNR et VNV.*

Bruit	RSB (dB)	RSB segmental pour les approches				
		DD	TSNR	RFSNR	HRNR	VNV
Bureau	24	24,4	25,0	24,7	**25,1**	23,6
	18	19,0	19,8	19,7	**20,2**	18,0
	12	13,6	14,5	14,6	**15,0**	12,9
	6	8,4	9,2	9,5	**9,8**	8,2
	0	3,9	4,4	**4,7**	**4,7**	3,9
Voiture	24	23,3	24,1	24,0	**24,7**	23,9
	18	18,1	18,9	18,9	**19,7**	18,4
	12	13,1	14,0	14,1	**14,8**	13,4
	6	8,5	9,3	9,7	**9,9**	8,8
	0	4,4	5,0	5,2	**5,3**	4,7
Rue	24	23,2	24,0	23,9	**24,6**	23,6
	18	17,7	18,6	18,6	**19,4**	17,9
	12	12,4	13,3	13,5	**14,2**	12,6
	6	7,4	8,2	8,4	**8,8**	7,7
	0	3,0	3,4	**3,7**	**3,7**	3,3
Foule	24	24,1	24,8	24,7	**25,7**	24,6
	18	18,1	19,0	19,1	**20,0**	18,5
	12	12,5	13,3	13,5	**14,2**	12,8
	6	7,4	7,9	8,1	**8,5**	7,7
	0	3,2	3,4	3,5	**3,7**	3,4

donnent de meilleurs résultats lorsque le bruit est stationnaire. Ceci est d'ailleurs évident dans la mesure où la stationnarité du bruit facilite grandement l'estimation de la DSP du bruit qui rappelons-le est réalisée à long terme.

TAB. 6.15 – *Tableau récapitulatif, en terme de distance cepstrale, des approches DD, TSNR, RFSNR, HRNR et VNV.*

Bruit	RSB (dB)	Distance cepstrale pour les approches				
		DD	TSNR	RFSNR	HRNR	VNV
Bureau	24	0,22	0,28	0,25	**0,19**	0,20
	18	0,35	0,43	0,41	**0,33**	0,34
	12	0,49	0,58	0,58	**0,46**	0,50
	6	**0,63**	0,73	0,76	0,64	0,68
	0	**0,78**	0,88	0,91	0,83	0,87
Voiture	24	0,29	0,34	0,31	**0,22**	0,24
	18	0,40	0,46	0,44	**0,32**	0,34
	12	0,51	0,60	0,59	**0,44**	0,46
	6	**0,62**	0,75	0,76	**0,62**	0,63
	0	**0,75**	0,89	0,93	0,85	0,84
Rue	24	0,37	0,42	0,40	**0,31**	0,32
	18	0,53	0,58	0,57	**0,44**	0,46
	12	0,68	0,74	0,76	**0,61**	0,63
	6	0,84	0,90	0,93	**0,81**	0,83
	0	1,01	1,04	1,07	**1,00**	1,02
Foule	24	0,31	0,34	0,33	**0,25**	0,26
	18	0,48	0,52	0,50	**0,40**	0,42
	12	0,67	0,70	0,70	**0,58**	0,61
	6	0,86	0,89	0,90	**0,79**	0,81
	0	1,06	1,09	1,11	**1,03**	1,03

6.4 Conclusion

Toutes les approches proposées et analysées dans ce chapitre sont de près ou de loin reliées ou basées sur l'approche DD qui constitue la référence initiale de qualité. On a vu que les approches SG (surtout avec les hypothèses Gauss-Gamma) permettent de limiter de façon intéressante les distorsions du signal restauré. Ces approches ont cependant un défaut majeur. En effet, l'utilisation de modèles super-gaussiens complexifie de façon spectaculaire les estimateurs ce qui rend leur utilisation rédhibitoire (en terme de coût de calcul).

Trois nouveaux estimateurs du RSB ont été proposés dont deux permettent d'obtenir de meilleurs résultats que l'approche SG pour une complexité seulement légèrement supérieure à l'approche DD de référence. L'approche DDG généralise l'estimateur DD, cependant cela ne permet pas de l'améliorer suffisamment pour s'en démarquer subjectivement. Par contre, les approches TSNR et RFSNR permettent de limiter (voire supprimer pour le RFSNR) les défauts de l'estimateur decision-directed permettant ainsi de supprimer l'effet de réverbération et de réduire davantage le niveau de bruit musical. Bien que basées sur deux concepts différents, ces techniques apportent des résultats subjectifs équivalents. L'approche psychoacoustique peut être associée à ces approches car elle permet de limiter l'effet d'étouffement de la parole lorsqu'un filtre peu agressif est choisi pour la mise en œuvre.

L'approche TSNR est retenue de par sa simplicité, sa faible complexité et bien sûr son efficacité pour servir de base à l'approche HRNR qui permet de dépasser les limitations liées aux problèmes d'estimation de la DSP du bruit et de la phase. Ceux-ci se traduisent essentiellement par la distorsion

des harmoniques du signal de parole. L'approche HRNR permet donc de régénérer les harmoniques détruites par des approches classiques en utilisant un traitement non-linéaire du signal distordu. En pratique, la limitation de la distorsion harmonique du signal de parole permet de supprimer plus de bruit qu'avec une technique classique. L'approche VNV est une évolution de l'approche HRNR qui tire parti de la mise en œuvre. En effet, nous avons proposé une approche permettant de séparer les composantes voisées et non voisées et par suite de les traiter de façon distincte. Un filtre agressif est conservé pour les composantes harmoniques et un filtre lissé est utilisé pour les autres composantes. Les artefacts et le bruit musical sont donc fortement réduits sans dégrader la structure voisée de la parole.

Références

[Akbari Azirani 1995a] A. Akbari Azirani, “Rehaussement de la Parole en Ambiance Bruitée. Application aux Télécommunications Mains-Libres,” *Thèse de l’Université de Rennes 1*, 1995.

[Boite 1987] R. Boite, et M. Kunt, “Traitement de la Parole,” *Presses Polytechniques Romandes, Complément au Traité d’électricité, Première édition*, 1987.

[Faucon 1993] G. Faucon, R. Le Bouquin, et A. Akbari Azirani, “Mesures Objectives de la Réduction de Bruit,” *GRETSI*, Juan-les-Pins, France, Septembre 1993.

[Kubichek 1991] R. F. Kubichek, “Standards and Technology Issues in Objective Voice Quality Assessment,” *Digital Signal Processing*, Vol. 1, pp. 38–44, 1991.

[Le Bouquin 1991] R. Le Bouquin, “Traitements pour la Réduction du Bruit sur de la Parole. Application aux Communications Radio-Mobiles,” *Thèse de l’Université de Rennes 1*, 1991.

[P48 1988] ITU-T Recommendation P.48, “Spécification d’un Système de Référence Intermédiaire,” 1988.

[P56 1996] ITU-T Recommandation P.56, “Telephone Transmission Quality - Objective Measuring Apparatus,” Mars 1996.

[P800 1996] ITU-T Recommendation P.800, “Methods for Subjective Determination of Transmission Quality,” Août 1996.

[Quackenbush 1988] S. R. Quackenbush, T. P. Barnwell, et M. A. Clements, “Objective Measures of Speech Quality,” *Prentice Hall*, 1988.

Bilan et perspectives

L'avènement des télécommunications mobiles a renforcé la nécessité d'améliorer la prise de son notamment lorsqu'elle est dégradée par une perturbation qualifiée de bruit. De nombreuses solutions permettent d'atténuer ce bruit. Les travaux menés au cours de cette thèse se restreignent aux techniques monovoies, sans doute le cas le plus courant mais aussi le plus critique. De façon générale, les techniques de réduction de bruit sont soumises à un compromis : plus on réduit le niveau de bruit et plus le signal restauré souffre de distorsions. Ainsi, l'objectif de cette thèse était de limiter ces distorsions pour pouvoir supprimer plus de bruit que ne le permettent les approches classiques. Le domaine de la réduction de bruit est très actif depuis près de 30 ans mais il n'a pas connu de révolution depuis l'approche proposée par Ephraïm et Malah en 1984 [Ephraïm 1984]. Cette approche dite decision-directed (DD) permet de réduire de façon impressionnante le phénomène de bruit musical résiduel, souvent plus désagréable que le bruit original.

C'est dans ce cadre que nous avons proposé des approches permettant d'améliorer le rendu du signal restauré par rapport aux approches de référence. L'approche DD est largement utilisée et reconnue comme très efficace, cependant, elle souffre tout de même d'un défaut : celui de générer une distorsion du signal de parole interprétable comme un phénomène de réverbération. Nous avons proposé deux axes principaux qui permettent de supprimer ce défaut :

– Le premier consiste à introduire des modèles super-gaussiens adaptés aux signaux traités pour exprimer les estimateurs du signal de parole. Ceci permet de limiter de façon intéressante les distorsions du signal restauré. En particulier, le bruit musical est réduit et l'effet de réverbération se retrouve limité.

– Le second consiste à améliorer l'estimateur du rapport signal à bruit (RSB) qui est à la fois la cause du bruit musical (RSB *a posteriori*) et de l'effet de réverbération (RSB *a priori*). Trois nouveaux estimateurs du RSB ont été proposés dont deux permettent de supprimer les défauts de l'approche DD. Ainsi, l'estimateur DD généralisé permet de suivre les évolutions fréquentielles des composantes harmoniques de parole de façon à limiter le biais de l'estimateur DD. Cependant, cette approche ne permet pas d'améliorer suffisamment l'approche DD pour s'en démarquer subjectivement. Par contre, les approches TSNR ("two-step noise reduction") et RFSNR ("reliable features selection noise reduction") permettent de limiter (voire supprimer dans le cas du RFSNR) le biais de l'estimateur DD permettant ainsi de supprimer l'effet de réverbération de l'approche DD et de réduire davantage le niveau de bruit musical. Bien que basées sur deux concepts différents, ces techniques apportent des résultats subjectifs équivalents. L'approche TSNR opère en deux passes, la seconde permet de raffiner l'estimation du RSB fournie

par la première. L'approche RFSNR est par contre basée sur la détermination des composantes fiables du RSB *a posteriori*. L'utilisation de ces composantes non biaisées permet de supprimer complètement le défaut de l'estimateur DD.

Le défaut de l'estimateur de RSB étant supprimé, nous avons constaté que cette amélioration reste insuffisante pour supprimer toutes les distorsions du signal restauré. En effet, de nombreuses dégradations sont causées par les erreurs d'estimation de la DSP du bruit et par l'impact de la phase. Ces erreurs se manifestent généralement par la distorsion des composantes harmoniques de la parole. Nous avons donc proposé l'approche HRNR ("harmonic regeneration noise reduction") qui tire justement parti de la structure harmonique de la parole pour limiter ce type de distorsion. En pratique, il s'avère possible de régénérer les harmoniques détruites par des approches classiques (y compris TSNR et RFSNR) en utilisant un traitement non-linéaire du signal distordu. Cette approche permet de limiter de façon significative la distorsion du signal de parole.

Par delà les techniques de réduction de bruit, un soin particulier a été apporté à leur mise en œuvre. La littérature est quasi-inexistante dans ce domaine et nous avons proposé différentes techniques permettant d'améliorer la qualité du signal restauré. En effet, il faut prendre des précautions afin d'éviter certains artefacts gênants. En particulier, une bonne maîtrise de l'étape de synthèse permet de supprimer les "clics" qui apparaissent à cause de la variabilité du filtre de réduction de bruit. Un autre point important concerne l'agressivité du gain spectral car on ne peut pas se permettre de conserver un filtre trop agressif qui génère beaucoup de distorsion et d'artefacts (nasalisation) et favorise l'apparition du bruit musical. Par conséquent, il est très intéressant de limiter l'agressivité du filtre de façon à limiter voire supprimer ces dégradations. Il est également indispensable de bien gérer le bruit résiduel qui permet entre autres de masquer les distorsions de la parole utile. D'ailleurs, le bruit réinjecté doit autant que possible avoir la même nature et la même coloration que le bruit original de façon à obtenir un résultat naturel.

Les approches proposées dans le présent mémoire ont permis d'améliorer significativement les performances des techniques de réduction de bruit. Cependant, il existe encore une marge de progression notamment lorsque le RSB est très défavorable. En effet, il a été établi que les erreurs d'estimation de la DSP du bruit et l'impact de la phase provoquent une distorsion harmonique du signal restauré. L'approche HRNR permet de limiter significativement ce type de distorsion en régénérant les harmoniques dégradées, mais lorsque le RSB est très défavorable, les distorsions sont telles qu'il est impossible de les régénérer efficacement. Une solution plus directe à ce problème consisterait à améliorer l'estimation de la DSP du bruit qui souffre de nombreuses limitations. On peut d'ailleurs remarquer que beaucoup de travaux portent sur ce sujet depuis ces dernières années et que le problème de la réduction de bruit a donc tendance à se reporter sur celui de l'estimation robuste du bruit. Le besoin dans ce domaine est d'obtenir une estimation à court terme capable de suivre parfaitement les évolutions du bruit trame à trame. Dans un domaine proche, il serait également intéressant de pouvoir estimer la différence de phase entre les composantes de signal et de bruit. Cela contribuerait à améliorer l'estimation du RSB notamment lorsque celui-ci est très défavorable. Cependant, les études portant sur la phase ont montré que l'estimation de cette quantité reste un problème très délicat.

Références

Règles d'atténuation spectrale à court-terme

[Akbari Azirani 1995a] A. Akbari Azirani, "Rehaussement de la Parole en Ambiance Bruitée. Application aux Télécommunications Mains-Libres," *Thèse de l'Université de Rennes 1*, 1995.

[Arslan 1995] L. Arslan, A. McCree, et V. Viswanathan, "New Methods for Adaptive Noise Suppression," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Detroit, États-Unis, Vol. 1, pp. 812–815, Mai 1995.

[Beaugeant 1999a] C. Beaugeant, et P. Scalart, "Noise Reduction Using Perceptual Spectral Change," *Eurospeech*, Budapest, Hongrie, Septembre 1999.

[Beaugeant 1999b] C. Beaugeant, "Réduction de Bruit et Contrôle d'Écho pour les Applications Radiomobiles," *Thèse de l'Université de Rennes 1*, 1999.

[Berouti 1979] M. Berouti, R. Schwartz, et J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Washington, États-Unis, pp. 208–211, Avril 1979.

[Boll 1979] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-27, No. 2, pp. 113–120, Avril 1979.

[Bouquin 1996] R. Le Bouquin, "Enhancement of Noisy Speech Signals: Application to Mobile Radio Communications," *Speech Commun.*, Vol. 18, pp. 3–19, 1996.

[Breithaupt 2003] C. Breithaupt, et R. Martin, "MMSE Estimation of Magnitude-Squared DFT Coefficients with Supergaussian Priors," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Vol. 1, pp. 896–899, 2003.

[Cappé 1993] O. Cappé, "Techniques de Réduction de Bruit pour la Restauration d'Enregistrements Musicaux," *Thèse de l'École Nationale Supérieure des Télécommunications*, Paris, Septembre 1993.

[Cappé 1994] O. Cappé, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor," *IEEE Trans. Speech Audio Processing*, Vol. 2, No. 2, pp. 345–349, Avril 1994.

[Cappé 1995] O. Cappé, et J. Laroche, "Evaluation of Short-Time Spectral Attenuation Techniques for the Restoration of Musical Recordings," *IEEE Trans. Speech Audio Processing*, Vol. 3, No.1, Janvier 1995.

[Chen 2005] B. Chen, et P. C. Loizou, "Speech Enhancement using a MMSE Short Time Spectral Amplitude Estimator with Laplacian Speech Modeling," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Philadelphie, États-Unis, Vol. 1, pp. 1097–1100, Mars 2005.

[Cho 2001] Y. D. Cho, K. Al-Naimi, et A. Kondoz, "Mixed Decision-Based Noise Adaptation for Speech Enhancement," *IEE Electronics Lett.*, Vol. 37, Issue 8, pp. 540–542, Avril 2001.

[Choi 2005] M.-S. Choi, et H.-G. Kang, "An Improved Estimation of a Priori Speech Absence Probability for Speech Enhancement: In Perspective of Speech Perception," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Philadelphie, États-Unis, Vol. 1, pp. 1117–1120, Mars 2005.

[Cohen 2001b] I. Cohen, "On Speech Enhancement Under Signal Presence Uncertainty," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Salt Lake City, États-Unis, Vol. 1, pp. 661–664, Mai 2001.

[Cohen 2001c] I. Cohen, et B. Berdugo, "Speech Enhancement for Non-Stationary Noise Environments," *Elsevier Signal Processing*, No. 81, pp. 2403–2418, 2001.

[Cohen 2002b] I. Cohen, "Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator," *IEEE Signal Processing Lett.*, Vol. 9, Issue 4, pp. 113–116, Avril 2002.

[Cohen 2004] I. Cohen, "On the Decision-Directed Estimation Approach of Ephraim and Malah," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Montréal, Canada, Vol. 1, pp. 293–296, Mai 2004.

[Ephraïm 1984] Y. Ephraïm, et D. Malah, "Speech Enhancement Using a Minimum Mean Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-32, No. 6, pp. 1109–1121, Décembre 1984.

[Ephraïm 1985] Y. Ephraïm, et D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-33, No.3, pp. 443–445, Avril 1985.

[Ephraïm 2004] Y. Ephraïm, et I. Cohen, "Recent Advancements in Speech Enhancement," *The Electrical Engineering Handbook*, CRC Press, à paraître.

[Etter 1994] W. Etter, et G. S. Moschytz, "Noise Reduction by Noise-Adaptive Spectral Magnitude Expansion," *J. Audio Eng. Soc.*, Vol. 42, No. 5, pp. 341–349, Mai 1994.

[Fan 2004] N. Fan, "Low Distortion Speech Denoising using an Adaptive Parametric Wiener Filter," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Montréal, Canada, Vol. 1, pp. 309–312, Mai 2004.

[Fillon 2004] T. Fillon, "Traitements Numériques du Signal Acoustique pour une Aide aux Malentendants," *Thèse de l'ENST*, 2004.

[Fingscheidt 2005] T. Fingscheidt, C. Beaugeant, et S. Suhadi, "Overcoming The Statistical Independance Assumption w.r.t Frequency in Speech Enhancement," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Philadelphie, États-Unis, Vol. 1, pp. 1081–1084, Mars 2005.

[Furuta 2002] S. Furuta, et S. Takahashi, "A Noise Suppresser for the AMR Speech Codec and Evaluation Test Results Based on 3GPP Specifications," *Workshop on Speech Coding*, Tsukuba, Japon, pp. 159–161, Octobre 2002.

[Gazor 2004] S. Gazor, "Employing Laplacian-Gaussian Densities for Speech Enhancement," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Montréal, Canada, Vol. 1, pp. 297–300, Mai 2004.

[Gemello 2004] R. Gemello, F. Mana, et R. De Mori, "A Modified Ephraim-Malah Noise Suppression Rule for Automatic Speech Recognition," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Montréal, Canada, Vol. 1, pp. 957–960, Mai 2004.

[Godsill 1998] S. J. Godsill, P. J. W. Rayner, et O. Cappé, "Digital Audio Restoration," *Appl. of Digital Signal Processing to Audio and Acoust.*, Kluwer Academic Publishers, pp. 133–193, 1993.

[Goh 1998] Z. Goh, K.-C. Tan, et B. T. G. Tan, "Postprocessing Method for Suppressing Musical Noise Generated by Spectral Subtraction," *IEEE Trans. Speech Audio Processing*, Vol. 6, No. 3, pp. 287–292, Mai 1998.

[Guédon 2002] L. Guédon, "Mise en œuvre de Nouvelles Hypothèses dans les Algorithmes de Réduction de Bruit par Atténuation Spectrale," *Document interne FT R&D*, Août 2002.

[Guérin 2002] A. Guérin, "Rehaussement de la Parole pour les Communications mains-libres. Réduction de Bruit et Annulation d'Écho Non Linéaire," *Thèse de l'Université de Rennes 1*, 2002.

[Guérin 2005] A. Guérin, "Traitement par Blocs : Reconstruction par Méthode Overlap-Save (OLS) et Overlap-Add (OLA)," *Document interne FT R&D*, 2005.

[Gustafsson 2001] H. Gustafsson, S. E. Nordholm, et I. Claesson, "Spectral Subtraction Using Reduced Delay Convolution and Adaptive Averaging," *IEEE Trans. Speech Audio Processing*, Vol. 9, No. 8, pp. 799–807, Novembre 2001.

[Hasan 2004] M. K. Hasan, S. Salahuddin, et M. R. Khan, "A Modified a Priori SNR for Speech Enhancement Using Spectral Subtraction Rules," *IEEE Signal Processing Lett.*, Vol. 11, Issue 4, pp. 450–453, Avril 2004.

[Hu 2001] Y. Hu, M. Bhatnagar, et P. C. Loizou, "A Cross-Correlation Technique for Enhancing Speech Corrupted with Correlated Noise," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Salt Lake City, États-Unis, Vol. 1, pp. 673–676, Mai 2001.

[Kim 2001a] H.-G. Kim, K. Obermayer, et D. Ruwisch, "Real-Time Noise Cancelling with Spectral Diffusion," *Document technique Cortologic AG*, Berlin, 2001.

[Kim 2001b] H.-G. Kim, K. Obermayer, M. Bode, et D. Ruwisch, "Efficient Speech Enhancement by Diffusive Gain Factors (DGF)," *Proc. Eurospeech*, Scandinavie, Vol. 1, pp. 1867–1870, 2001.

[Kim 2003] H.-G. Kim, M. Schwab, N. Moreau, et T. Sikora, "Speech Enhancement of Noisy Speech Using Log-Spectral Amplitude Estimator and Harmonic Tunnelling," *Intl. Work. Acoust. Echo and Noise Control*, Kyoto, Japon, Septembre 2003.

[Le Bouquin 1991] R. Le Bouquin, "Traitements pour la Réduction du Bruit sur de la Parole. Application aux Communications Radio-Mobiles," *Thèse de l'Université de Rennes 1*, 1991.

[Lim 1979] J. S. Lim, et A. V. Oppenheim, "Enhancement and Bandwith Compression of Noisy Speech," *Proc. IEEE*, Vol. 67, No. 12, pp. 1586–1604, Décembre 1979.

[Malah 1999] D. Malah, R. V. Cox, et A. J. Accardi, "Tracking Speech-Presence Uncertainty to Improve Speech Enhancement in Non-Stationary Noise Environments," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Phoenix, États-Unis, pp. 789–792, Mars 1999.

[Martin 2002] R. Martin, “Speech Enhancement Using MMSE Short Time Spectral Estimation with Gamma Distributed Speech Priors,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Orlando, États-Unis, Vol. 1, pp. 253–256, Mai 2002.

[McAulay 1980] J. McAulay, et M. Malpass, “Speech Enhancement Using a Soft-Decision Noise Suppression Filter,” *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, No. 2, pp. 137–145, Avril 1980.

[Ogata 2001] S. Ogata, et T. Shimamura, “Reinforced Spectral Subtraction Method to Enhance Speech Signal,” *IEEE Region 10 Intl. Conf. on Electrical and Electronic Technology*, Vol. 1, pp. 242–245, Août 2001.

[Okazaki 2004] M. Okazaki, T. Kunimoto, et T. Kobayashi, “Multi-Stage Subtraction for Enhancement of Audio Signals,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Montréal, Canada, Vol. 2, pp. 805–808, Mai 2004.

[Plapous 2004] C. Plapous, C. Marro, L. Mauuary, et P. Scalart, “Two-Step Noise Reduction Technique,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Montréal, Canada, Vol. 1, pp. 289–292, Mai 2004.

[Plapous 2005a] C. Plapous, C. Marro, et P. Scalart, “Speech Enhancement using Harmonic Regeneration,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Philadelphie, États-Unis, Vol. 1, pp. 157–160, Mars 2005.

[Plapous 2005b] C. Plapous, C. Marro, et P. Scalart, “Reliable a Posteriori Signal-to-Noise Ratio Features Selection,” *IEEE Work. on Appl. of Signal Processing to Audio and Acoust.*, New Paltz, États-Unis, Octobre 2005.

[Plapous à paraître] C. Plapous, C. Marro, et P. Scalart, “Improved Signal-to-Noise Ratio Estimation for Speech Enhancement,” *IEEE Trans. Speech Audio Processing*, à paraître.

[Porter 1984] J. E. Porter, et S. F. Boll, “Optimal Estimators for Spectral Restoration of Noisy Speech,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, San Diego, États-Unis, Vol. 2, pp. pp. 18A2.1–2.4, Mars 1984.

[Puder 1999] H. Puder, “Single Channel Noise Reduction using Time-Frequency Dependent Voice Activity Detection,” *Intl. Work. Acoust. Echo and Noise Control*, Pocono Manor, États-Unis, pp. 68–71, Septembre 1999.

[Quatieri 1997] T. F. Quatieri, et R. A. Baxter, “Noise Reduction Based on Spectral Change,” *IEEE Work. Appl. of Signal Processing to Audio and Acoust.*, Mohonk, États-Unis, Octobre 1997.

[Scalart 1996a] P. Scalart, et J. Vieira Filho, “Speech Enhancement Based on a Priori Signal to Noise Estimation,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Atlanta, États-Unis, Vol. 2, pp. 629–632, Mai 1996.

[Scalart 1996b] P. Scalart, et A. Benamar, “A System for Speech Enhancement in the Context of Hands-Free Radiotelephony with Combined Noise Reduction and Acoustic Echo Cancellation,” *Speech Commun.*, Vol. 20, pp. 203–214, 1996.

[Schmidt 2004] G. Schmidt, “Single-Channel Noise Suppression Based on Spectral Weighting - An Overview,” *EURASIP News Lett.*, Vol. 15, No. 1, pp. 9–24, Mars 2004.

[Sim 1998] B. L. Sim, Y. C. Tong, J. S. Chang, et C. T. Tan, “A Parametric Formulation of the Generalized Spectral Subtraction Method,” *IEEE Trans. Speech Audio Processing*, Vol. 6, No. 4, pp. 328–337, Juillet 1998.

[Sovka 1996] P. Sovka, P. Pollak, et J. Kybic, "Extended Spectral Subtraction," *European Signal Processing Conf.*, Trieste, Italie, pp. 963-966, Septembre 1996.

[Virette 2002a] D. Virette, P. Scalart, et C. Lamblin, "Analysis of Background Noise Reduction Techniques for Robust Speech Coding," *Document interne FT R&D*, 2002.

[Virette 2002b] D. Virette, "Étude d'Algorithmes de Codage de la Parole Destinés aux Environnements Bruités," *Document interne FT R&D*, 2002.

[Wang 1993] F. M. Wang, P. Kabal, R. P. Ramachandran, et D. O'Shaughnessy, "Frequency Domain Adaptive Postfiltering for Enhancement of Noisy Speech," *Speech Commun.*, Vol. 12, pp. 41–56, 1993.

[Wei 2000] W. Wei, et C. Yanpu, "Speech Enhancement by Spectral Component Selection," *IEEE Intl. Conf. Signal Processing*, Vol. 2, No. 5, pp. 674–678, 2000.

[Xie 1996] F. Xie, et D. Van Compernolle, "Speech Enhancement by Spectral Magnitude Estimation - A Unifying Approach," *Speech Commun.*, Vol. 19, pp. 89–104, 1996.

[Yang 1993] J. Yang, "Frequency Domain Noise Suppression Approaches in Mobile Telephone Systems," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Minneapolis, États-Unis, Vol. 2, pp. 363–366, Avril 1993.

Estimation de la DSP du bruit

[Cohen 2001a] I. Cohen, et B. Berdugo, "Spectral Enhancement by Tracking Speech Presence Probability in Subbands," *Hands-Free Speech Commun.*, Kyoto, Japan, pp. 95–98, Avril 2001.

[Cohen 2002a] I. Cohen, et B. Berdugo, "Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement," *IEEE Signal Processing Lett.*, Vol. 9, No. 1, pp. 12–15, Janvier 2002.

[Cohen 2003] I. Cohen, "Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging," *IEEE Trans. Speech Audio Processing*, Vol. 11, No. 5, pp. 466–475, Septembre 2003.

[Doblinger 1995] G. Doblinger, "Computationally Efficient Speech Enhancement by Spectral Minima Tracking in Subbands," *Eurospeech*, Madrid, Espagne, Vol. 2, pp. 1513–1516, Septembre 1995.

[Evans 2002] N. W. D. Evans, J. S. Mason, et B. Fauve, "Efficient Real-Time Noise Estimation without Explicit Speech, non-Speech Detection: an Assessment on the AURORA Corpus," *Intl. Conf. Digital Signal Processing*, Vol. 2, pp. 985–988, Juillet 2002.

[Hirsch 1995] H. G. Hirsch, et C. Ehrlicher, "Noise Estimation Techniques for Robust Speech Recognition," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Detroit, États-Unis, Vol. 1, pp. 153–156, Mai 1995.

[Kato 2003] M. Kato, M. Serizawa, N. Toki, U. Mori, Y. Morishita, et K. Hayashi, "Noise Suppression with High Speech Quality Based on Weighted Noise Estimation for 3G Handsets," *NEC Res. & Develop. Special issue on Device and Systems for Mobile Communications*, Vol. 44, No. 4, pp. 340–348, Octobre 2003.

[Lin 2003] L. Lin, W. H. Holmes, et E. Ambikairajah, “Adaptive Noise Estimation Algorithm for Speech Enhancement,” *Electronics Lett.*, Vol. 39, Issue 9, pp. 754–755, Mai 2003.

[Lin 2005] Z. Lin, et R. Goubran, “Instant Noise Estimation using Fourier Transform of AMDF and Variable Start Minima Search,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Philadelphie, États-Unis, Vol. 1, pp. 161–164, Mars 2005.

[Martin 1993] R. Martin, “An Efficient Algorithm to Estimate the Instantaneous SNR of Speech Signals,” *Eurospeech*, Berlin, Allemagne, pp. 1093–1096, Septembre 1993.

[Martin 1994] R. Martin, “Spectral Subtraction Based on Minimum Statistics,” *Eusipco*, Edinburgh, Royaume-Uni, pp. 1182–1185, Septembre 1994.

[Martin 2001] R. Martin, “Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics,” *IEEE Trans. Speech Audio Processing*, Vol. 9, No. 5, pp. 504–512, Juillet 2001.

[Paliwal 1988] K. K. Paliwal, “Estimation of Noise Variance From the Noisy AR Signal and Its Application in Speech Enhancement,” *Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-36, No. 2, pp. 292–294, Février 1988.

[Rangachari 2004] S. Rangachari, P. C. Loizou, et Y. Hu, “A Noise Estimation Algorithm with Rapid Adaptation for Highly Non-Stationary Environments,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Montréal, Canada, Vol. 1, pp. 305–308, Mai 2004.

[Sugiyama 2003] A. Sugiyama, M. Kato, et M. Serizawa, “Test Results of NEC’s New Low Complexity AMR-NS Solution Based on TS 26.077,” *TSG-SA4#28 Meeting*, Erlangen, Allemagne, Septembre 2003.

[Takeda 2005] K. Takeda, T. H. Dat, H. Fujimura, et F. Itakura, “SNR and Local Noise Power Estimations Based on Gaussian Mixture Modeling on The Log-Power Domain,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Philadelphie, États-Unis, Vol. 4, pp. 881–884, Mars 2005.

À propos de la phase

[Almeida 1981] L. B. Almeida, et J. M. Tribolet, “A Model for Short-Time Phase Prediction of Speech,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Atlanta, États-Unis, Vol. 6, pp. 213–216, Avril 1981.

[Alsteris 2004] L. D. Alsteris, et K. K. Paliwal, “Importance of Window Shape for Phase-Only Reconstruction of Speech,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Montréal, Canada, Vol. 1, pp. 573–576, Mai 2004.

[Cox 1980] R. C. Cox, et D. M. Robinson, “Some Notes on Phase in Speech Signals,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Denver, États-Unis, Vol. 5, pp. 150–153, Avril 1980.

[Espy 1982] C. Espy, et J. S. Lim, “Effects of Noise on Signal Reconstruction from Fourier Transform Phase,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Paris, France, pp. 1833–1836, Mai 1982.

[Hayes 1980] M. H. Hayes, J. S Lim, et A. V. Oppenheim, “Signal Reconstruction from Phase or Magnitude,” *Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, pp. 672-680, Décembre 1980.

[Kang 2002] H.-G. Kang, et H. K. Kim, "A phase Generation Method for Speech Reconstruction From Spectral Envelope and Pitch Intervals," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Orlando, États-Unis, Vol. 1, pp. 429–432, Mai 2002.

[Kim 2001] D.-S. Kim, "On the Perceptually Irrelevant Phase Information in Sinusoidal Representation of Speech," *IEEE Trans. Speech Audio Processing*, Vol. 9, No. 8, pp. 900–905, Novembre 2001.

[Marsh 1988] K. A. Marsh, et J. M. Richardson, "Probabilistic Algorithm for Phase Retrieval," *Optical Society of America*, Vol. 5, No. 7, pp. 993–998, Juillet 1988.

[Mcaulay 1985] R. J. McAulay, et T. F. Quatieri, "Mid-Rate Coding Based on a Sinusoidal Representation of Speech," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Tampa, États-Unis, Vol.10, pp. 945–948, Mars 1985.

[Mcaulay 1986] R. J. McAulay, et T. F. Quatieri, "Phase Modelling and its Application to Sinusoidal Transform Coding," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Tokyo, Japon, Vol.11, pp. 1713–1716, Avril 1986.

[Oppenheim 1979] A. V. Oppenheim, J. S. Lim, G. Kopec, et S. C. Pohlig, "Phase in Speech and Pictures," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Washington, États-Unis, Vol. 4, pp. 632–637, Avril 1979.

[Oppenheim 1981] A. V. Oppenheim, et J. S. Lim, "The Importance of Phase in Signals," *Proc. IEEE*, Vol. 69, No. 5, pp. 529–541, Mai 1981.

[Pobloth 1999] H. Pobloth, et W. B. Kleijn, "On Phase Perception in Speech," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Phoenix, États-Unis, Vol.1, pp. 29–32, Mars 1999.

[Seok 1999] J.-W. Seok, et K.-S. Bae, "Reduction of Musical Noise in Spectral Subtraction Method Using Subframe Phase Randomisation," *IEEE Electronics Lett.*, Vol. 35, No. 2, pp. 123–125, Janvier 1999.

[Thomson 1988] D. L. Thomson, "Parametric Models of the Magnitude/Phase Spectrum for Harmonic Speech Coding," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, New York, États-Unis, Vol.1, pp. 378–381, Avril 1988.

[Vary 1985] P. Vary, "Noise Suppression by Spectral Magnitude Estimation - Mechanism and Theoretical Limits," *Signal Processing*, Vol. 8, pp. 387–400, 1985.

[Wang 1982] D. L. Wang, et J. S. Lim, "The Unimportance of Phase in Speech Enhancement," *Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-30, No. 4, pp. 679–681, Août 1982.

Approches basées sur la psychoacoustique

[Akbari Azirani 1995b] A. Akbari Azirani, R. Le Bouquin Jeannes, et G. Faucon, "Optimizing Speech Enhancement by Exploiting Masking Properties of the Human Ear," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Detroit, États-Unis, Vol. 1, pp. 800–803, Mai 1995.

[Gustafsson 1998] S. Gustafsson, P. Jax, et P. Vary, "A Novel Psychoacoustically Motivated Audio Enhancement Algorithm Preserving Background Noise Characteristics," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Seattle, États-Unis, pp. 397–400, Mai 1998.

[ISO MPEG 1992] ISO/IEC, JTC1/SC29/WG11 MPEG “Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media at Up to about 1.5Mbit.s - Part 3 : Audio,” IS11172-3, 1992.

[Johnston 1988] J. D. Johnston, “Transform Coding of Audio Signals Using Perceptual Noise Criteria,” *IEEE J. on Select. Areas Commun.*, Vol. 6, No. 2, pp. 314–323, Février 1988.

[Lin 2002] L. Lin, W. H. Holmes, et E. Ambikairajah, “Speech Denoising Using Perceptual Modification of Wiener Filtering,” *IEEE Electronics Lett.*, Vol. 38, No. 23, pp. 1486–1487, Novembre 2002.

[Painter 2000] T. Painter, et A. Spanias, “Perceptual Coding of Digital Audio,” *IEEE proc.*, Vol. 88, No. 4, Avril 2000.

[Schroeder 1979] M. R. Schroeder, B. S. Atal, et J. L. Hall, “Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear,” *J. Acoustical Society of America*, 66(6), pp. 1647–1652, Décembre 1979.

[Tsoukalas 1993] D. Tsoukalas, M. Paraskevas, et J. Mourjopoulos, “Speech Enhancement Using Psychoacoustic Criteria,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Minneapolis, États-Unis, Vol. 2, pp. 359–362, Avril 1993.

[Tsoukalas 1997] D. E. Tsoukalas, J. Mourjopoulos, et G. Kokkinakis, “ Audio Noise Cancellation Using a Subjective Signal Representation,” *Intl. Conf. Digital Signal Processing*, Vol. 2, pp. 613–616, juillet 1997.

[Virag 1996] N. Virag, “Speech Enhancement Based on Masking Properties of the Human Auditory System,” *Thèse de l’École Polytechnique Fédérale de Lausanne*, 1996.

[Virag 1999] N. Virag, “Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System,” *IEEE Trans. Speech Audio Processing*, Vol. 7, No. 2, pp. 126–137, Mars 1999.

[You 2004] C. H. You, S. N. Hoh, et S. Rahardja, “An MMSE Speech Enhancement Approach Incorporating Masking Properties,” *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Montréal, Canada, Vol. 1, pp. 725–728, Mai 2004.

[Zwicker 1981] E. Zwicker, et R. Feldtkeller, “Psychoacoustique - L’Oreille Récepteur d’Information,” *Masson et CNET-ENST*, 1981.

Approches basées sur des alternatives à la FFT

DCT

[Hasan 2002] M. K. Hasan, M. S. A. Zilany, et M. R. Khan, “DCT Speech Enhancement with Hard and Soft Thresholding Criteria,” *Electronics Lett.*, Vol. 38, No. 13, pp. 669–670, Juin 2002.

[Salahuddin 2002] S. Salahuddin, S. Z. Al Islam, M. D. Hasan, et M. R. Khan, “Soft Thresholding for DCT Speech Enhancement,” *Electronics Lett.*, Vol. 38, No. 24, pp. 1605–1607, Novembre 2002.

[Soon 1998] I. Y. Soon, S. N. Koh, et C. K. Yeo, “Noisy Speech Enhancement Using Discrete Cosine Transform,” *Speech Commun. Lett.*, No. 24, pp. 249–257, 1998.

[Soon 2000] I. Y. Soon, et S. N. Koh, "Low Distortion Speech Enhancement," *IEE Proc. Vision, Image and Signal Processing*, Vol. 147, Issue 3, pp. 247–253, Juin 2000.

Ondelettes

[Abry 1997] P. Abry, "Ondelettes et Turbulence. Multirésolutions, Algorithmes de Décomposition, Invariance d'Échelles," *Diderot Éditeur*, Paris, 1997.

[Agbinya 1996] J. I. Agbinya, "Discrete Wavelet Transform Techniques in Speech Processing," *IEEE Region 10 Intl. Conf. on Electrical and Electronic Technology*, Perth, Australie, Vol. 2, pp. 514–519, Novembre 1996.

[Chang 2002] S. Chang, Y. Kwon, S.-I. Yang, et I.-J. Kim, "Speech Enhancement for Non-Stationary Noise Environment by Adaptive Wavelet Packet," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Orlando, États-Unis, Vol. 1, pp. 561–564, Mai 2002.

[Cohen 1992] A. Cohen, "Ondelettes et Traitement Numérique du Signal," *Masson, Recherches en Mathématiques Appliquées*, Octobre 1992.

[Donoho 1994] D. L. Donoho, et I. M. Johnstone, "Threshold Selection for Wavelet Shrinkage of Noisy Data," *IEEE Intl. Conf. On Engineering in Medicine and Biology Society*, Vol. 1, pp. A24-A25, Novembre 1994.

[Gade 1997] S. Gade, et K. Gram-Hansen, "The Analysis of Nonstationary Signals," *Sound and Vibration*, pp. 40–46, Janvier 1997.

[Ma 2005] N. Ma, M. Bouchard, et R. A. Goubran, "A Wavelet Kalman Filter with Perceptual Masking for Speech Enhancement in Colored Noise," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Philadelphie, États-Unis, Vol. 1, pp. 149–152, Mars 2005.

[Seok 1997] J.-W. Seok, et K.-S. Bae, "Speech Enhancement with Reduction of Noise Components in the Wavelet Domain," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Munich, Allemagne, Vol. 2, pp. 1323–1326, Avril 1997.

Approches basées sur du filtrage adaptatif

[Deng 2005] J. Deng, M. Bouchard, et T. Yeap, "Speech Enhancement using a Switching Kalman Filter with a Perceptual Post-Filter," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Philadelphie, États-Unis, Vol. 1, pp. 1121–1124, Mars 2005.

[Gabrea 2002] M. Gabrea, "Speech Signal Recovery in Colored Noise using an Adaptative Kalman Filtering," *IEEE Canadian Conf. on Electrical and Computer Engineering*, Vol. 2, pp. 12–15, Mai 2002.

[Kuo 1995] S. M. Kuo, et M. J. Ji, "Development and Analysis of an Adaptive Noise Equalizer," *IEEE Trans. Speech Audio Processing*, Vol. 3, No. 3, pp. 217–222, Mai 1995.

[Oppenheim 1994] A. V. Oppenheim, E. Weinstein, K. C. Zangi, M. Feder, et D. Gauger, "Single Sensor Active Noise Cancellation," *IEEE Trans. Speech Audio Processing*, Vol. 2, No. 2, pp. 285–290, Avril 1994.

[Sambur 1978] M. R. Sambur, "Adaptive Noise Cancelling for Speech Signals," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-26, No. 5, pp. 419–423, Octobre 1978.

[Widrow 1975] B. Widrow *et al.*, "Adaptive Noise Cancelling: Principles and Applications," *IEEE Proc.*, Vol. 63, No.12, Décembre 1975.

Approches basées sur des modèles

HMM

[Ephraïm 1989] Y. Ephraïm, D. Malah, et B.-H. Juang, "On the Application of Hidden Markov Models for Enhancing Noisy Speech," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 37, No. 12, pp. 1846–1856, Décembre 1989.

[Ephraïm 1990] Y. Ephraïm, "A Minimum Mean Square Error Approach for Speech Enhancement," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Albuquerque, États-Unis, Vol. 2, pp. 829–832, Avril 1990.

[Ephraïm 1991] Y. Ephraïm, "On Minimum Mean Square Error Speech Enhancement," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Toronto, Canada, Vol. 2, pp. 997–1000, Mai 1991.

[Ephraïm 1992] Y. Ephraïm, "A Bayesian Estimation Approach for Speech Enhancement using Hidden Markov Models," *IEEE Trans. Speech Audio Processing*, Vol. 40, Issue 4, pp. 725–735, Avril 1992.

GMM

[Ding 2005] G.-H. Ding, X. Wang, Y. Cao, F. Ding, et Y. Tang, "Speech Enhancement Based on Speech Spectral Complex Gaussian Mixture Model," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Philadelphie, États-Unis, Vol. 1, pp. 165–168, Mars 2005.

[Hall 2003] H. Hall, "A Gaussian Mixture Model Spectral Representation for Speech Recognition," *Thèse de l'Université de Cambridge*, Juillet 2003.

[Master 2000] A. S. Master, "Speech Spectrum Modelling from Multiple Sources," *Master de l'Université de Cambridge*, Août 2000.

Approches basées sur les sous-espaces

[Ephraïm 1993] Y. Ephraïm, et H. L. Van Trees, "A Signal Subspace Approach for Speech Enhancement," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Minneapolis, États-Unis, Vol. 2, pp. 355–358, Avril 1993.

[Ephraïm 1995a] Y. Ephraïm, et H. L. Van Trees, "A Spectrally-Based Signal Subspace Approach for Speech Enhancement," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Detroit, États-Unis, Vol. 1, pp. 804–807, Mai 1995.

[Ephraïm 1995b] Y. Ephraïm, et H. L. Van Trees, "A Signal Subspace Approach for Speech Enhancement," *IEEE Trans. Speech Audio Processing*, Vol. 3, No. 4, pp. 251–266, Juillet 1995.

[Hegger 2001] R. Hegger, H. Kantz, et L. Matassini, "Noise Reduction for Human Speech Signals by Local Projections in Embedding Spaces," *IEEE IEEE Trans. Circuits Syst. II*, Vol. 48, No. 12, Décembre 2001.

[Johnson 2003] M. T. Johnson, A. C. Lindgren, R. J. Povinelli, et X. Yuan, "Performance of Nonlinear Speech Enhancement Using Phase Space Reconstruction," *Intl. Work. Acoust. Echo and Noise Control*, Kyoto, Japon, Septembre 2003.

[You 2005] C. H. You, S. N. Koh, et S. Rahardja, "Signal Subspace Speech Enhancement for Audible Noise Reduction," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Philadelphie, États-Unis, Vol. 1, pp. 145–148, Mars 2005.

Approches basées sur les réseaux de neurones

[Fah 2000] L. B. Fah, A. Hussain, et S. A. Samad, "Speech Enhancement by Noise Cancellation Using Neural Network," *IEEE Region 10 Intl. Conf. on Electrical and Electronic Technology*, Kuala Lumpur, Malaisie, Vol. 1, pp. 39–42, Septembre 2000.

[Knecht 1994] W. G. Knecht, "Nonlinear Noise Filtering and Beamforming Using the Perceptron and Its Volterra Approximation," *IEEE Trans. Speech Audio Processing*, Vol. 2, No. 1, pp. 55–62, Janvier 1994.

[Wan 1998] E. A. Wan, et A. T. Nelson, "Handbook of Neural Networks for Speech Processing," *Ed. S. Katagiri, Artech House, Première édition*, 1998.

Divers outils d'analyse

[Boashash 1987] B. Boashash, et P. J. Black, "An Efficient Real-Time Implementation of the Wigner-Ville Distribution," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-35, No. 11, pp. 1611–1618, Novembre 1987.

[Chari 1995] V. R. Chari, et C. Y. Espy-Wilson, "Adaptive Enhancement of Fourier Spectra," *IEEE Trans. Speech Audio Processing*, Vol. 3, No. 1, pp. 35–39, Janvier 1995.

[Dat 2005] T. H. Dat, K. Takeda, et F. Itakura, "Generalized Gamma Modeling of Speech and Its Online Estimation for Speech Enhancement," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Philadelphie, États-Unis, Vol. 4, pp. 181–184, Mars 2005.

[Deller 1994] J. R. Deller, "Tom, Dick, and Mary Discover the DFT," *IEEE Signal Processing Magazine*, Vol. 11, No.2, pp. 36–50, Avril 1994.

[Dix 1994] P. J. Dix, et G. Bloothooft, "A Breakpoint Analysis Procedure Based on Temporal Decomposition," *IEEE Trans. Speech Audio Processing*, Vol. 2, No. 1, pp. 9–17, Janvier 1994.

[Faucon 1993] G. Faucon, R. Le Bouquin, et A. Akbari Azirani, "Mesures Objectives de la Réduction de Bruit," *GRETSI*, Juan-les-Pins, France, Septembre 1993.

[Hendriks 2005] R. C. Hendriks, R. Heusdens, et J. Jensen, "Adaptive Time Segmentation of Noisy Speech for Improved Speech Enhancement," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Philadelphie, États-Unis, Vol. 1, pp. 153–156, Mars 2005.

[Hlawatsch 1992] F. Hlawatsch, et G. F. Boudreaux-Bartels, "Linear and Quadratic Time-Frequency Signal Representations," *IEEE Signal Processing Magazine*, Vol. 9, No. 2, pp. 21–67, Avril 1992.

[Kaiser 1990] J. F. Kaiser, "On a Simple Algorithm to Calculate the 'Energy' of a Signal," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Albuquerque, États-Unis, Vol. 1, pp. 381–384, Avril 1990.

[Kameoka 2004] H. Kameoka, T. Nishimoto, et S. Sagayama, "Separation of Harmonic Structures based on Tied Gaussian Mixture Model and Information Criterion for Concurrent Sounds," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Montréal, Canada, Vol. 4, pp. 297–300, Mai 2004.

[Kraniauskas 1994] P. Kraniauskas, "A Plain Man's Guide to the FFT," *IEEE Signal Processing Magazine*, Vol. 11, No.2, pp. 24–35, Avril 1994.

[Kubichek 1991] R. F. Kubichek, "Standards and Technology Issues in Objective Voice Quality Assessment," *Digital Signal Processing*, Vol. 1, pp. 38–44, 1991.

[Kullback 1958] S. Kullback, "Information Theory and Statistics," *Dover Publication*, 1958.

[Marchand 1998] M. Desainte-Catherine, et S. Marchand, "High Precision Fourier Analysis of Sounds using Signal Derivatives," *Université de Bordeaux*, Mai 1998.

[P56 1996] ITU-T Recommandation P.56, "Telephone Transmission Quality - Objective Measuring Apparatus," Mars 1996.

[Peyrin 1986] F. Peyrin, et R. Prost, "A Unified Definition for the Discrete-Time, Discrete-Frequency, and Discrete-TimeFrequency Wigner Distributions," *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-34, No. 4, pp. 858–866, Août 1986.

[Plante 1998] F. Plante, G. Meyer, et W. A. Ainsworth, "Improvement of Speech Spectrogram Accuracy by the Method of Reassignment," *IEEE Trans. Speech Audio Processing*, Vol. 6, No. 3, pp. 282–287, Mai 1998.

[Quackenbush 1988] S. R. Quackenbush, T. P. Barnwell, et M. A. Clements, "Objective Measures of Speech Quality," *Prentice Hall*, 1988.

[Renevey 2001] P. Renevey, et A. Drygajlo, "Detection of Reliable Features for Speech Recognition in Noisy Conditions Using a Statistical Criterion," *Proc. Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis*, Aalborg, Denmark, pp. 71–74, Septembre 2001.

[Sekhar 2004] S. Chandra Sekhar, et T. V. Sreenivas, "Novel Approach to AM-FM Decomposition with Applications to Speech and Music Analysis," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Montréal, Canada, Vol. 2, pp. 753–756, Mai 2004.

[Smits 1995] R. Smits, et B. Yegnanarayana, "Determination of Instants of Significant Excitation in Speech Using Group Delay Function," *IEEE Trans. Speech Audio Processing*, Vol. 3, No. 5, pp. 325–333, Septembre 1995.

[Yannis 1996] S. Yannis, "Decomposition of Speech Signals Into a Periodic and Non-Periodic Part Based on Sinusoidal Models," *IEEE Intl. Conf. Electronics, Circuits, and Systems*, Vol. 1, pp. 514–517, Octobre 1996.

Divers

[Beritelli 2002] F. Beritelli, S. Casale, G. Ruggeri, et S. Serrano, "Performance Evaluation and Comparison of G.729/AMR/Fuzzy Voice Activity Detectors," *IEEE Signal Processing Lett.*, Vol. 9,

No. 3, Mars 2002.

[Boite 1987] R. Boite, et M. Kunt, "Traitement de la Parole," *Presses Polytechniques Romandes, Complément au Traité d'électricité, Première édition*, 1987.

[Bouteille 2002] F. Bouteille, "Traitement de la Parole dans les Ponts de Conférence à Accès Hétérogènes Synchrones (RNIS, RTC) et Asynchrones (IP)," *Thèse de l'Université de Rennes 1*, 2002.

[Crochiere 1983] R. E. Crochiere, et L. R. Rabiner, "Multirate Digital Signal Processing," *Prentice-Hall, Première édition*, 1983.

[Deller 2000] J. R. Deller Jr., J. H. L. Hansen, et J. G. Proakis, "Discrete-Time Processing of Speech Signal," *IEEE Press, Première publication en 1993*, 2000.

[Gradshteyn 1994] I. S. Gradshteyn, et I. M. Ryzhik, "Table of Integrals, Series, and Products," *Academic press, 5ème édition*, 1994.

[Kunt 1984] M. Kunt, "Traité d'Électricité - Traitement Numérique des Signaux," *Presses Polytechniques et universitaires Romandes, Artech House, Troisième édition*, 1998.

[Marro 1996] C. Marro, "Traitement de Déréverbération et de Débruitage pour le Signal de Parole dans des Contextes de Communication Interactive," *Thèse de l'Université de Rennes 1*, 1996.

[McAulay 1990] R. J. McAulay, et T. F. Quatieri, "Pitch Estimation and Voicing Detection Based on a Sinusoidal Speech Model," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Albuquerque, États-Unis, Vol. 1, pp. 249–252, Avril 1990.

[P48 1988] ITU-T Recommendation P.48, "Spécification d'un Système de Référence Intermédiaire," 1988.

[P800 1996] ITU-T Recommendation P.800, "Methods for Subjective Determination of Transmission Quality," Août 1996.

[Prasanna 2004] S. R. Mahadeva Prasanna, et B. Yegnanarayana, "Extraction of Pitch in Adverse Conditions," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Montréal, Canada, Vol. 1, pp. 109–112, Mai 2004.

[Ramirez 2004] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, et A. Rubio, "Voice Activity Detection with Noise Reduction and Long-Term Spectral Divergence Estimation," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Montréal, Canada, Vol. 2, pp. 1093–1096, Mai 2004.

[Taddei 2004] H. Taddei, C. Beaugeant, et M. de Meuleneire, "Noise Reduction on Speech Codec Parameters," *IEEE Intl. Conf. Acoust., Speech, Signal Processing*, Montréal, Canada, Vol. 1, pp. 497–500, Mai 2004.

[Van Gerven 1997] S. Van Gerven, et F. Xie, "A Comparative Study of Speech Detection Methods," *Eurospeech*, Vol. 3, pp. 1095–1098, Grèce, Septembre 1997.

Liste des figures

Liste des tableaux

Traitements pour la réduction de bruit. Application à la communication parlée.

Avec l'avènement des télécommunications mobiles grand public, le besoin d'améliorer la prise de son, notamment en réduisant la gêne due au bruit, s'est fait de plus en plus présent. Les techniques de réduction du bruit sont soumises à un compromis entre le niveau effectif de réduction et la distorsion qui affecte le signal de parole. Au vu des performances actuelles, il est souhaitable de supprimer plus de bruit tout en conservant un niveau de dégradation acceptable du signal restauré, ceci en particulier lorsque le niveau de bruit est important. Les techniques qui ont suscité le plus d'intérêt au cours de ces 30 dernières années sont les approches par atténuation spectrale à court terme qui consistent à modifier une transformée à court terme du signal bruité en utilisant une règle de suppression. L'essor de cette famille de techniques s'explique essentiellement par le fait qu'elles permettent de respecter les contraintes de temps réel et de complexité inhérentes aux applications de communication parlée.

La première partie de ce document est consacrée à l'analyse des techniques majeures de réduction du bruit par atténuation spectrale à court terme. Ce sera notamment l'occasion d'identifier les limitations, points de blocage et autres défauts de ces méthodes ainsi que de montrer qu'il existe une marge de progression intéressante en terme de qualité par rapport à ces différents points clés. La seconde partie est essentiellement consacrée à la description et l'analyse de solutions originales proposées en réponse aux limitations identifiées dans la première partie. Un soin particulier a également été apporté à la mise en œuvre qui fait partie intégrante des techniques de réduction de bruit et qui conditionne la qualité du signal restauré.

L'analyse des limitations des techniques de réduction du bruit actuelles a permis de dégager plusieurs approches originales permettant de résoudre tout ou partie des problèmes identifiés. Ainsi, l'introduction de nouveaux modèles statistiques, adaptés aux signaux de parole et de bruit, pour déterminer l'expression d'une règle de suppression permet d'obtenir des résultats sensiblement meilleurs qu'en utilisant le modèle gaussien classique. Un problème d'ordre plus général concerne les défauts des estimateurs du rapport signal à bruit, paramètre fondamental qui conditionne les performances des techniques de réduction de bruit. La suppression de ces défauts conduit effectivement à une limitation des distorsions de la parole. Cependant, le signal restauré souffre toujours de certaines dégradations dues notamment aux erreurs d'estimation du bruit et à l'impact de la phase. En effet, l'estimation du bruit, qui constitue une étape clé des techniques de réduction de bruit, souffre de nombreuses limitations surtout lorsque le bruit n'a pas un caractère stationnaire. Dans une moindre mesure, la phase, qui est souvent négligée, a aussi une influence importante dans l'estimation du signal de parole, en particulier lorsque le niveau de bruit est élevé. Une approche originale qui tire parti de la structure voisée du signal de parole pour limiter les distorsions harmoniques engendrées par les techniques classiques est proposée et permet de dépasser les limites de performances des techniques classiques.

Outre ces nouvelles approches, leur mise en œuvre conditionne également la qualité finale du signal restauré. Plusieurs points sensibles sont donc soulevés et des solutions sont données qui permettent d'éviter de nombreux artefacts ("clics", nasalisation, bruit musical) désagréables. Les approches proposées sont évaluées en utilisant des critères objectifs dont les résultats sont au besoin validés par des tests subjectifs. Les résultats obtenus montrent des améliorations significatives par rapport aux performances des techniques de référence.