République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Ecole Nationale Polytechnique
Département d'Electronique
Laboratoire Signal & Communications

Thèse de Doctorat en Electronique

Présentée par :

BOUBAKIR Chabane

Magister en Electronique de l'ENP

Intitulée :

Contribution aux méthodes de rehaussement de la parole pour le codage à bas débit

Soutenue publiquement le **20 / 03 / 2014** devant le jury composé de :

Président :	Mme. GUERTI M'hania	Professeur	ENP
Rapporteur :	M. BERKANI Daoud	Professeur	ENP
Examinateurs :	Mme. HAMAMI Latifa	Professeur	ENP
	M. AMROUCHE Abderrahmane	Professeur	USTHB
	M. TEFFAHI Hocine	Professeur	USTHB
	M. GUESSOUM Abderrezak	Professeur	Univ. Blida

ENP 2014

REMERCIEMENTS

Ces travaux ont été réalisés au sein du Laboratoire Signal et Communications (LSC) de l'Ecole Nationale Polytechnique d'Alger (ENP), et au Laboratoire d'Etudes et de Modélisation en Electrotechnique (LAMEL) de l'université de JIJEL.

Je tiens à remercier Monsieur D. Berkani, Professeur à l'ENP, pour avoir accepté d'être rapporteur de mes travaux, pour sa rigueur scientifique et ses qualités humaines. Ses conseils et ses critiques ont grandement contribués à la réalisation de ce document.

J'exprime ma reconnaissance à Madame M. Guerti, Professeur à l'ENP, pour l'honneur qu'elle m'a fait en acceptant de présider le jury de cette thèse.

Je tiens également à exprimer mes sincères remerciements et ma profonde reconnaissance à Madame L. Hamami, Professeur à l'ENP, pour m'avoir encouragé à finaliser cette thèse, pour ses précieux conseils, sa gentillesse et pour l'intérêt qu'elle a voulu porter à ce travail en acceptant de l'examiner.

Je remercie Monsieur A. Amrouche, Professeur à l'Université des Sciences et de la Technologie Houari Boumediene USTHB, pour l'honneur qu'il m'a fait en participant à l'évaluation de ce travail.

J'exprime ma plus profonde gratitude à Monsieur H. TEFFAHI, Professeur à l'Université des Sciences et de la Technologie Houari Boumediene USTHB, de m'avoir honoré en acceptant d'être examinateur.

Je remercie vivement Monsieur A. Guessoum, Professeur à l'Université de Blida, d'avoir accepté de juger ce travail en tant qu'examinateur et d'avoir contribué à l'amélioration de ce document.

Je tiens à remercier également le professeur Francis Grenez de l'Université Libre de Bruxelles (ULB), pour m'avoir accueilli dans son équipe signaux et communications durant les années 2004 et 2005.

Je suis aussi reconnaissant des aides ponctuelles et efficaces du Professeur Rainer Martin de l'IKA de Ruhr University of Bochum, et de m'avoir accueilli dans son laboratoire durant l'année 2007.

Merci aussi à tous mes collègues et étudiants de l'université de JIJEL qui se reconnaîtront ici. Je leur exprime ma profonde sympathie et leur souhaite beaucoup de succès dans leur carrière.

Je remercie également mes parents, mes frères et sœurs, et ma petite famille pour m'avoir soutenu et encouragé pendant toutes ces années.

TABLE DES MATIÈRES

Liste des figures

Liste des tableaux

	Abréviations
ACR	**A**bsolute **C**ategory **R**ating.
AR	**A**uto-**R**egréssif.
ARMA	**A**uto-**R**égressif à **M**oyenne **A**justée.
CAN	**C**onvertisseur **A**nalogique/**N**umérique.
CELP	**C**ode **E**xcited **L**inear **P**rediction.
CNA	**C**onvertisseur **N**umérique / **A**nalogique.
DAM	**D**iagnostic **A**cceptability **M**easure.
DAP	**D**écodage **A**coustico-**P**honétique.
DD	**D**ecision **D**irected (decision dirigée).
DFT	**D**iscret **F**ourier **T**ransform (transformée de Fourier discrète).
DoD	**D**epartment **o**f **D**efense.
DRT	**D**iagnostic **R**hyme **T**est.
DSP	**D**ensité **S**pectrale de **P**uissance.
EQM	**E**rreur **Q**uadratique **M**oyenne.
EM	**E**xpectation **M**aximization.
FFT	**F**ast **F**ourier **T**ransform (transformée de Fourier rapide).
FS MELP	**F**ederal **S**tandard **MELP.**
IFFT	**I**nverse **F**ast **F**ourier **T**ransform (transformée de Fourier rapide inverse).
IS	**I**takura-**S**aito measure.
LSF	**L**ine **S**pectral **F**requencies.
LLR	**L**og-**L**ikelihood-**R**atio.
LP	**L**inear **P**redictif (prediction linéaire).
LPC	**L**inear **P**redictif **C**oding.
LRT	**L**ikelihood **R**atio **T**est.
LSA	**L**og **S**pectral **A**mplitude.
LSP	**L**ine **S**pectral **P**airs.
MA	**M**oyenne **A**justée.
MAP	**M**aximum **a P**osteriori.
MBE	**M**ulti**B**and **E**xcitation.
MCRA	**M**inima **C**ontrolled **R**ecursive **A**veraging.
MELP	**M**ixed **E**xcitation **L**inear **P**rediction.
MELPe	**e**nhanced-**M**ELP.
ML	**M**aximum **L**ikelihood.
MMSE	**M**inimum **M**ean **S**quare **E**rror.
MM-LSA	**M**ultiplicatively-**M**odified **LSA.**
MOS	**M**ean **O**pinion **S**core.
MRT	**M**odified **R**hyme **T**est.

MS	**M**inimum **S**tatistics.
MSE	**M**ean **S**quare **E**rror.
MSVQ	**M**ulti **S**tage **V**ectors **Q**uantizer.
OTAN	**O**rganisation du **T**raité de l'**A**tlantique **N**ord.
PESQ	**P**erceptual **E**valuation of **S**peech **Q**uality.
RIF	**R**éponse **I**mpulsionnelle **Fi**nie.
RII	**R**éponse **I**mpulsionnelle **I**nfinie.
RSB	**R**apport **S**ignal sur **B**ruit.
SNR	**S**ignal to **N**oise **R**atio.
SPU	**S**ignal **P**resence **U**ncertainty.
STC	**S**inusoidal **T**ransform **C**oder.
STSA	**S**hort **T**ime **S**pectral **A**mplitude.
TDT	**T**ucker **D**avis **T**echnologies.
TFCT	**T**ransformée de **F**ourier à **C**ourt-**T**erme.
TFD	**T**ransformée de **F**ourier **D**iscrète.
UIT	**U**nion **I**nternationale des **T**élécommunications.
UKF	**U**nscented **K**alman **F**ilter.
V/NV	**V**oisé / **N**on **V**oisé.
VAD	**V**oice **A**ctivity **D**etector (détection d'activité vocale).
VQ	**V**ectors **Q**uantizer.
WI	**W**aveform **I**nterpolation.
WSS	**W**eighted **S**pectral **S**lope.

INTRODUCTION GENERALE

Le traitement du signal de la parole a fait l'objet de recherches intenses depuis plusieurs années [1]. En particulier, le codage de la parole pour la transmission numérique, a connu des progrès remarquables et les applications de téléphonies numériques et mobiles se multipliant.

Le codage du signal de parole est la discipline qui cherche à représenter la voix humaine sous un format numérique compact, mais efficace pour la transmission et/ou le stockage. L'objectif est d'utiliser le minimum de bits possible tout en préservant la qualité perceptuelle du signal à sa destination [2][3].

Dans les systèmes de téléphonie filaire classiques, la parole est numérisée à 64 kbit/s. De nombreux algorithmes [4] ont été proposés pour diminuer ce débit tout en essayant de conserver une qualité subjective donnée fonction des exigences de l'application à laquelle le codeur est destiné. On distingue en général trois plages de débits :
• Les hauts débits, supérieurs à 16 kbit/s, correspondant à des algorithmes de codage de la forme d'onde non spécifiques à la parole.
• Les débits moyens, de 4 kbit/s à 16 kbit/s, correspondant à des techniques de codage hybrides utilisant des méthodes de codage de la forme d'onde et prenant en compte certaines propriétés de la parole ou de la perception auditive. Le principal représentant de cette classe est le codage CELP [5].
• Les bas et très bas débits, de quelques dizaines de bits par seconde à 4 kbit/s, correspondant aux vocodeurs (VOice CODER) spécifiques au codage de la parole.

Le procédé de codage de la parole mis en œuvre à bas débit est généralement celui du vocodeur qui se base sur un modèle complètement paramétrique, parmi ces vocodeurs on peut citer le codeur LPC10 et le codeur MELP. Dans cette thèse, nous utiliserons seulement les codeurs à bas débit fonctionnant avec des débits allant de 2.4 kbit/s jusqu'à 4.8 kbit/s. Ainsi, le codeur MELP à 2.4 kbit/s sera exploité dans nos simulations.

La téléphonie cellulaire, les transmissions confidentielles entre organismes gouvernementaux ou pour les applications militaires, le courrier vocal et ainsi de suite, ne sont que quelques exemples des nombreuses applications des codeurs à bas débit. A cause des

perturbations et des bruits apparaissant inévitablement dans ces applications, le signal reconstruit ne pouvait être une réplique exacte du signal propre.

Les performances des codeurs de la parole sont limitées dans les environnements bruités et surtout dans le cas des codeurs à bas débit, comme le cas du MELP à 2.4 kbit/s. Cette limitation a donné lieu à l'idée d'améliorer la qualité et l'intelligibilité de la parole bruitée d'entrée par un système de rehaussement avant son codage [6]-[10].

Le problème du rehaussement de la parole étant un problème très complexe qui a été largement étudié. Dans ce cadre, plusieurs méthodes ont été développées, comme les différentes variantes de la soustraction spectrale : d'amplitude [11], de puissance [12], non linéaire [13], avec masquage perceptuel [14], paramétrique [15] et multi-bandes [16]. En plus, du filtrage de Wiener [17][18],… etc. Ces méthodes parviennent à réduire de manière très efficace le niveau du bruit de fond. Cependant, l'inconvénient de ces méthodes est l'introduction d'un bruit appelé « bruit musical », ce bruit est très gênant du point de vue perceptif.

D'autres méthodes, ont été développées, à partir de la représentation des systèmes dans l'espace d'état par des équations différentielles matricielles du premier ordre. Ces techniques sont basées sur le fait qu'un processus aléatoire peut être modélisé comme étant la sortie d'un système linéaire gouverné par un bruit blanc, c'est le filtre de Kalman [19]. La technique du filtrage de Kalman est largement utilisée aujourd'hui dans la réduction du bruit. Les inconvénients de ces méthodes sont la complexité et les difficultés d'une implémentation en temps réel.

A cet effet, des méthodes Bayesiennes de réduction du bruit plus sophistiquées ont été mises au point pour parvenir à une réduction du bruit efficace tout en limitant la présence du bruit musical. Plusieurs règles de suppression ont été construites à partir de l'hypothèse que les coefficients de la TFD des signaux source et bruit sont des variables aléatoires Gaussiennes et indépendantes [17][20]-[22]. D'autres plus efficaces et plus récentes se basent sur des modèles super-Gaussiens [23]-[34].

Les algorithmes monovoie de rehaussement de la parole seront étudiés et implémentés dans cette thèse. Des améliorations et des adaptations de ces algorithmes seront aussi l'objet de ce travail de recherche.

La majorité des blocs de réduction de bruit des codeurs de la parole à bas débit utilisent les méthodes Bayesiennes, ces méthodes assurent une qualité acceptable de la parole rehaussée, un son intelligible, un bruit résiduel faible et un temps de traitement faible par rapport aux autres méthodes. Ces approches Bayesiennes seront implémentées et testées avec les deux modèles Gaussien et super-Gaussien des coefficients de la transformée de Fourier des signaux de la parole et du bruit. Par la suite, les algorithmes les plus efficaces de chaque catégorie seront associés avec le codeur MELP.

Le bloc de pré-traitement qu'on présente dans cette thèse tente, le mieux possible, de réaliser un codage efficace à bas débit du signal de parole bruité de bande téléphonique (300 Hz – 3400 Hz), donc une fréquence d'échantillonnage de 8 kHz. Notre contribution a consisté à valider et perfectionner les différentes composantes du système de rehaussement en vue d'une meilleure qualité et robustesse [35]-[39].

Plusieurs algorithmes de débruitage on été testé. La performance du système est évaluée à l'aide des mesures subjectives et objectives. Les résultats obtenus démontrent l'efficacité de

l'utilisation d'un système de rehaussement basé sur un modèle super-Gaussien comme un bloc de pré-traitement pour le codage de la parole à bas débit. Certains résultats sont également obtenus dans divers milieux hostiles (comme le bruit babble, le bruit de voiture et dans une rue, …etc.) sous des conditions acoustiques variables (à savoir, différentes valeurs du SNR).

Les simulations ainsi que tous les tests effectués sur les différents algorithmes traités ont été effectués sur un PC Pentium 4, Intel ® Core (TM) 2, 1.86 GHz, 1 Go de RAM, avec un logiciel MATLAB version 7.1.0.246 (R14) service pack3. Ce logiciel permet de simplifier la mise au point des algorithmes de rehaussement sans contraintes de temps et de mémoire, et il est devenu un outil indispensable et simple pour le test des algorithmes de traitement du signal avant l'implémentation en temps réel.

Cette thèse présente une contribution aux méthodes de rehaussement de la parole pour les codeurs bas débit dans un milieu hostile et contient cinq chapitres.

Le premier chapitre présentera les notions élémentaires et les termes relatifs à la description de la parole et du bruit, les corpus les plus utilisés, les propriétés statistiques des coefficients de la transformée de Fourier discrète qui seront la base des méthodes Bayesiennes de rehaussement, et enfin les mesures objectives qui seront utilisées tout au long de ce travail afin de comparer entre les diverses méthodes implémentées.

Le deuxième chapitre sera consacré à la définition du rehaussement et les principales méthodes de rehaussement de la parole. Nous présenterons la méthode de Berouti et le filtre de Wiener qui sont les deux méthodes de base des méthodes STSA basées sur la soustraction spectrale et sur les méthodes Bayesiennes respectivement. Suivi, des notions générales sur l'estimation de la densité spectrale du bruit et du rapport signal sur bruit a priori.

Dans le troisième chapitre, nous présenterons l'utilisation du filtre de Kalman au rehaussement de la parole. Des généralités sur le filtre de Kalman et ses paramètres seront introduites au début, suivi par une modélisation dans le cas d'un bruit blanc et des bruits colorés, et les résultats d'implémentation dans chaque cas.

Au cours du quatrième chapitre, nous étudieront les approches Bayesiennes de rehaussement de la parole. Les estimateurs basés sur le modèle Gaussien et ceux basés sur un modèle super-Gaussien seront détaillés, ainsi qu'une comparaison entre les résultats de ces estimateurs.

Dans le chapitre cinq, nous appliquerons l'opération de prétraitement au codeur MELP à 2.4 kbit/s dans un environnement bruité. On commencera par une étude détaillée du codeur MELP à 2.4 kbit/s, suivi par sa variante MELPe et enfin une comparaison entre ces deux variantes et la variante proposée basée sur des approches super-Gaussien.

Enfin, la conclusion générale donnera les interprétations appropriées aux différents résultats et une vue sur les perspectives du travail.

Les annexes contiendront les différents algorithmes utilisés dans ce travail, ainsi que les schémas blocs du codeur MELP.

CHAPITRE 1

GENERALITES SUR LES SIGNAUX ET NOTIONS D'ESTIMATION

1.1 Introduction

La parole est la faculté de communiquer la pensée par un système de sons articulés, c'est le moyen de communication privilégié entre les humains qui sont les seuls êtres vivants à utiliser un tel système structuré. Le fort développement actuel des télécommunications redonne à la parole un aspect privilégié.

La présence d'un bruit superposé au signal utile dégrade la qualité et l'intelligibilité de la parole. La connaissance du système vocal et de ses propriétés ainsi qu'une bonne estimation du bruit permettent l'amélioration de la qualité de la parole. Les propriétés du système auditif humain peuvent également être exploitées pour améliorer la qualité perceptuelle du signal codé dans un milieu propre ou hostile.

Ce chapitre comporte des généralités sur la parole et le bruit. Nous présenterons tout d'abord le principe de production de la parole, les caractéristiques du signal de la parole, les techniques d'analyse les plus utilisées, les sources et les types de bruit et quelques corpus de parole utilisés. Par la suite, une étude détaillée des propriétés statistiques des coefficients de la Transformée de Fourier Discrète (TFD) sera exposée. Ces propriétés sont très utiles pour les méthodes Bayesiennes de rehaussement de la parole. Enfin, les mesures de qualité utilisées tout au long de ce document pour la comparaison entre les différentes approches de rehaussement de la parole sont décrites.

1.2 Production de la parole

Le signal vocal est engendré par l'appareil phonatoire avant d'être émis puis détecté par un auditeur. A la perception, l'oreille analyse ce signal et transmet au cerveau les informations nécessaires à son interprétation. La figure (1.1) présente le system vocal qui se compose [1] :

- D'une partie subglottique : l'ensemble diaphragme, les poumons et la trachée-artère.
- D'une partie glottique : le larynx et les cordes vocales.
- D'une partie supraglottique (conduit vocal) : le pharynx et les cavités buccales et nasales.

La parole est généralement produite par expiration de l'air à travers la glotte et le conduit vocal, l'air venant des poumons est modulé par vibration des cordes vocales et par déformation (élargissement ou rétrécissement) du conduit. Ce système phonatoire permet le processus de production de la parole qui comporte trois étapes principales :

- ❖ La génération d'une énergie ventilatoire qui mettra en mouvement oscillatoire des cordes vocales.
- ❖ Les vibrations des cordes vocales qui donneront naissance à des sons voisés.
- ❖ L'articulation qui sera réalisée dans les cavités supraglottique.

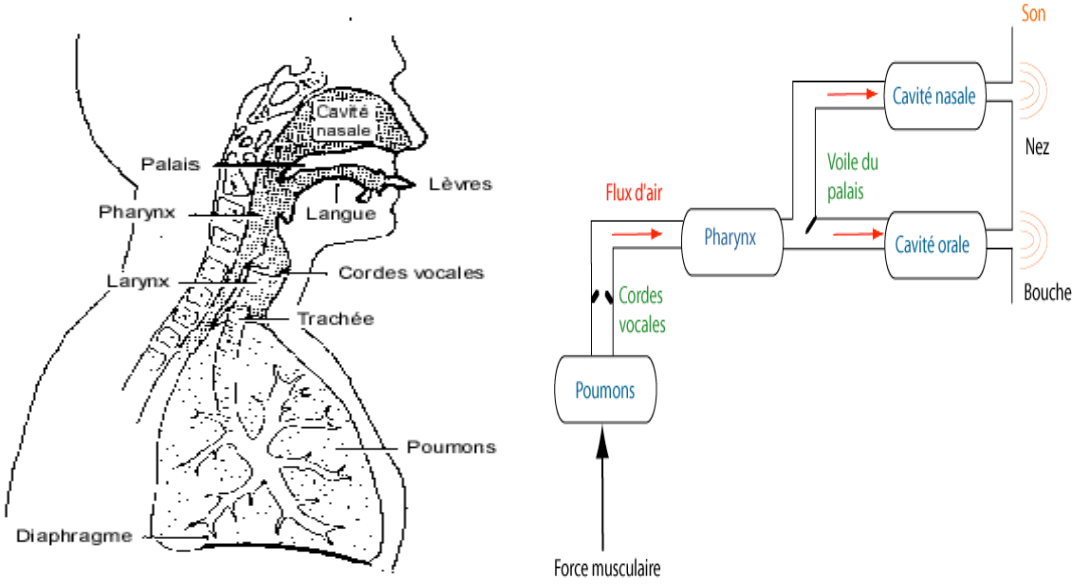


Figure 1.1 : Appareil phonatoire humain**.**

1.3 Eléments de traitement du signal vocal

Le traitement du signal vocal s'inscrit dans une succession de procédures, que ce soit pour le codage, la reconnaissance automatique, le rehaussement ou pour la synthèse de la parole.

Le traitement est aussi utilisé pour réduire la redondance du signal vocal, où on extrait des paramètres pertinents pour le rehaussement et le codage. Ces paramètres obtenus durant l'étape d'analyse du signal bruité, seront la base de toutes les méthodes de rehaussement. Suivant le degré de contamination de ces paramètres par le milieu bruité, des critères de correction seront utilisés. Les paramètres obtenus après correction seront exploités durant l'étape de synthèse pour synthétiser une parole rehaussée.

La majorité des algorithmes de réduction du bruit, d'estimation du niveau de bruit et de classification bruit/parole, qui seront exposés en détails dans les prochains chapitres, sont basés sur les propriétés temporelles et spectrales du signal de parole. Ainsi, quelques mesures ou

certains paramètres doivent être disponibles à chaque trame d'analyse. Ces quantités seront exploitées par le module de rehaussement et peuvent être par exemple : le pitch, l'énergie, le taux de passage par zéro, la fonction d'autocorélation, la classification voisée/non voisée, le spectre d'amplitude ou de puissance du signal, les différents coefficients qui modélisent le signal de parole, les mesures statistiques appliquées à ce signal,…etc. Les éléments de traitement du signal vocal les plus utilisés dans cette thèse seront exposés dans les sections suivantes.

1.3.1 Quasi stationnarité

Le signal de parole peut être considéré comme un processus non stationnaire, c'est à dire que ses propriétés statistiques changent au cours du temps. La non stationnarité résulte des changements au cours du temps de la source ainsi que la forme et les dimensions du conduit vocal. Cependant, l'observation du signal de la parole indique qu'il n'évolue pas ou peu sur des durées de (5 ms à 30 ms). Donc, on peut le considérer comme localement stationnaire durant ce temps (la durée du temps où les variations du conduit vocal sont presque négligeables) [1].

1.3.2 Fenêtrage

Le but du fenêtrage est de découper le signal de parole en petites tranches, où il peut être considéré localement comme quasi-stationnaire. En outre, et pour profiter de l'évolution lente du signal vocal, le fenêtrage permet le traitement en temps réel et facilite aussi l'analyse des signaux [1][40].

Il existe plusieurs types de fenêtres d'analyse, comme la fenêtre rectangulaire, triangulaire, Blackman, Kaiser, la fenêtre de Hanning et la fenêtre de Hamming,…etc. Les deux dernières sont les plus convenables à la parole, car elles entraînent un minimum de distorsion spectral du signal de la parole, par rapport aux autres fenêtres. La fenêtre de Hamming de longueur *N* est donnée par :

$$w(n)=\begin{cases}0.54-0.46\cos\left(\dfrac{2\pi n}{N}\right), & 0\le n\le N-1\\ 0, & \text{ailleurs}\end{cases} \tag{1.1}$$

Et celle de Hanning par :

$$w(n)=\begin{cases}0.5+0.5\cos\left(\dfrac{2\pi n}{N}\right), & 0\le n\le N-1\\ 0, & \textit{ailleurs}\end{cases} \tag{1.2}$$

1.3.3 Sons voisés ou non-voisés

Une décomposition simplifiée d'un signal de parole fait ressortir deux types de sons, voisés ou non-voisés :

Les sons voisés, qui résultent de l'excitation du conduit vocal par des impulsions périodiques, avec une fréquence de récurrence appelée fréquence fondamentale (**PITCH**) qui peut varie de 80 Hz à 200 Hz pour les hommes, de 150 Hz à 450 Hz pour les femmes et de 200 Hz à 600 Hz pour les enfants. Cette différence de fréquence est due à la longueur et la masse des cordes vocales [1].

Les sons non voisés, qui sont obtenus par resserrement du conduit vocal, et sont habituellement d'énergie inférieure aux sons voisés. Les cordes vocales sont écartées et n'entrent pas en vibration et l'excitation dans ce cas est modélisée par un bruit blanc. Ces sons sont considérés comme ayant des caractéristiques proches de celles du bruit.

1.3.4 Énergie du signal

Elle est représentée par l'intensité du son qui est liée à la pression de l'air en amant du larynx. L'amplitude du signal de la parole varie au cours du temps selon le type du son, et son énergie dans une trame de taille N est donnée par :

$$E = \sum_{n=0}^{N-1} s^2(n) \tag{1.3}$$

1.3.5 Spectre

L'enveloppe spectrale ou spectre représente l'intensité de la voix selon la fréquence, elle est généralement obtenue par une analyse de Fourier à court terme. La quasi stationnarité du signal de parole permet de mettre en œuvre des méthodes efficaces d'analyse et de modélisation, utilisées pour le traitement à court terme du signal vocal sur des fenêtres de durée généralement comprise entre 20 ms et 30 ms appelées trames, avec un recouvrement entre ces fenêtres qui assure la continuité temporelle des caractéristiques de l'analyse.

La transformée de Fourier à court terme (TFCT) d'un signal échantillonné est par définition la transformée du signal pondéré.

$$S(k) = \sum_{n=0}^{N-1} s(n) \cdot w(n) \cdot \exp(-j2\pi nk / N), \qquad 0 \le k \le N-1 \tag{1.4}$$

Où : N : Le nombre de points prélevés, $S(k)$: spectre complexe, $s(n)$: segment analysé**,** $w(n)$: fenêtre de pondération.

Le spectre de puissance (appelé aussi densité spectrale de puissance de la transformée de Fourier) est donné par :

$$\left|S(k)\right|^2, \quad 0 \le k \le \frac{N}{2} \tag{1.5}$$

1.3.6 Spectrogramme

Il est souvent intéressant de représenter l'évolution temporelle du spectre à court terme d'un signal, sous la forme d'un spectrogramme.

Un spectrogramme est la transcription des données bidimensionnelles (amplitude/temps) en un motif tridimensionnel (énergie/fréquence/temps). Il est usuellement représenté avec le temps en abscisse, la fréquence en ordonnée, alors que les énergies sont notées en couleurs ou en niveaux de gris. Un spectrogramme d'un signal de parole pondéré par une fenêtre est donné par :

$$S_x(t, f) = \left| \int_{t-T}^{t+T} w(t-\tau).x(\tau).e^{-j2\pi f\tau}.d\tau \right|^2 \tag{1.6}$$

Dans laquelle $x(\tau)$ est le signal, T est la demi longueur de fenêtre temporelle, $w(\tau)$ est une fenêtre de pondération.

On parle de spectrogramme à large bande ou à bande étroite selon la durée de la fenêtre de pondération. Les spectrogrammes à bande large sont obtenus avec des fenêtres de pondération de faible durée (typiquement 10 ms); ils mettent en évidence l'enveloppe spectrale du signal, et

permettent par conséquent de visualiser l'évolution temporelle des formants. Les périodes voisées y apparaissent sous la forme de bandes verticales plus sombres. Les spectrogrammes à bande étroite sont moins utilisés (durée de fenêtres de 30 ms). Ils mettent plutôt la structure fine du spectre en évidence : les harmoniques du signal dans les zones voisées y apparaissent sous la forme de bandes horizontales (figure 1.2).

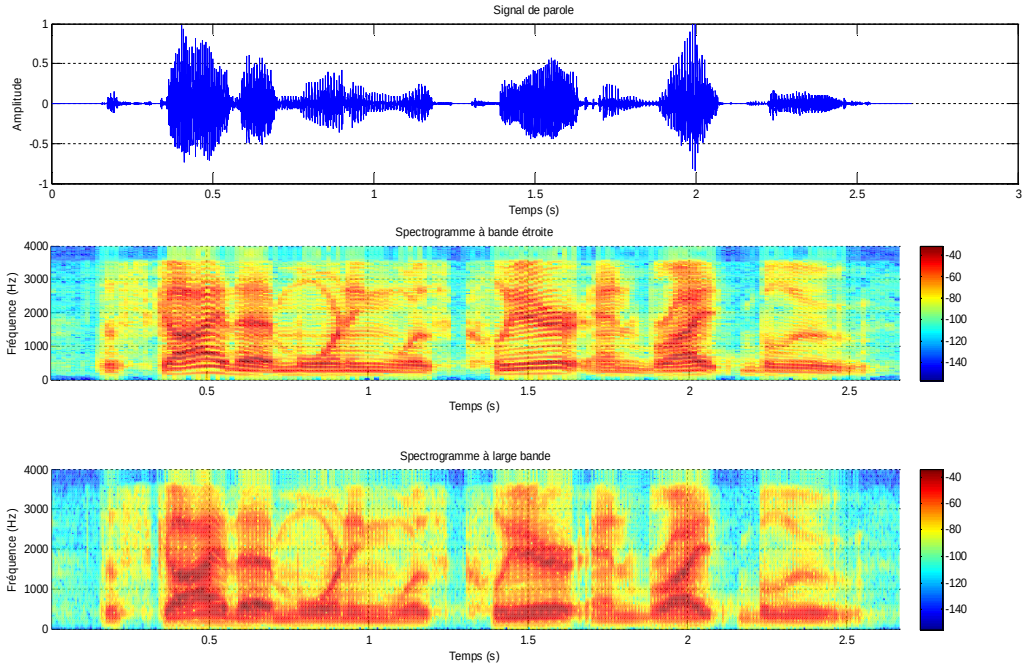


Figure 1.2 : Spectrogrammes à large bande, à bande étroite, et évolution temporelle de la phrase Anglaise « The sky that morning was clear and bright blue ».

1.3.7 Bande passante du signal

La bande passante téléphonique du signal de parole s'étend de 300 Hz à 3400 Hz, soit une plage de 3100 Hz. Pour la plupart des codeurs, la bande passante du signal de parole est limitée à un maximum de 4 kHz, bien qu'en pratique, elle soit plutôt limitée à 3600 Hz. La fréquence d'échantillonnage est donc de 8 kHz.

Pour le codage large bande avec une bande de 7 kHz, le signal a une meilleure qualité que celui limité à la bande téléphonique. En effet, les fréquences basses (50 Hz à 300 Hz) contribuent à améliorer le caractère naturel du signal de parole, ainsi que la reconnaissance du locuteur. De plus, on remarque une amélioration de l'intelligibilité et de la netteté due à l'extension de la bande passante de 3400 Hz à 7000 Hz. Les codeurs larges bandes sont utilisés essentiellement dans des applications spécialisées telles que la vidéoconférence et la vidéophonie où l'interaction avec le réseau téléphonique est rare.

1.4 Audition – perception

La parole est un vecteur de transmission d'information d'une grande complexité. En tant que récepteur de ce vecteur, l'appareil auditif de l'être humain se caractérise par une grande finesse d'analyse de cette complexité et par une grande robustesse à l'environnement. Pour cette raison, de nombreux systèmes de traitement de la parole tentent de reproduire les fonctionnalités de cet appareil [14].

Les ondes sonores sont recueillies par l'appareil auditif, ce qui provoque les sensations auditives. Ces ondes de pression sont analysées dans l'oreille interne qui envoie au cerveau l'influx nerveux qui en résulte; le phénomène physique induit ainsi un phénomène psychique grâce à un mécanisme physiologique complexe.

L'appareil auditif comprend l'oreille externe, l'oreille moyenne, et l'oreille interne (figure 1.3). Captées par l'oreille externe, les vibrations acoustiques sont transmises par l'oreille moyenne au milieu liquidien de la cochlée, organe de l'audition de l'oreille interne. Au sein de la cochlée, les vibrations provoquent la mise en mouvement des liquides et des différentes membranes qui la constituent. Ces mouvements provoquent à leur tour l'inclinaison des stéréocils des cellules ciliées déclenchant ainsi l'activation des fibres nerveuses. Ces dernières transmettent alors un message électrique vers le cortex cérébral.

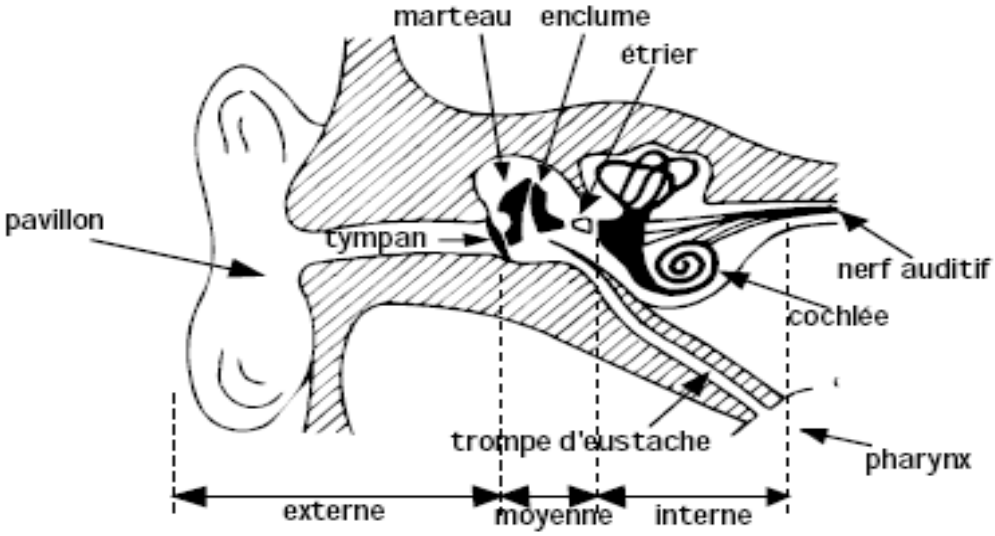


Figure 1.3 : Schéma du système auditif.

L'oreille ne répond pas également à toutes les fréquences. La figure (1.4) présente le champ auditif humain, délimité par la courbe du seuil de l'audition et celle du seuil de la douleur. Sa limite supérieure en fréquence (16000 Hz, variable selon les individus) fixe la fréquence d'échantillonnage maximale utile pour un signal auditif (32000 Hz) [41][42].

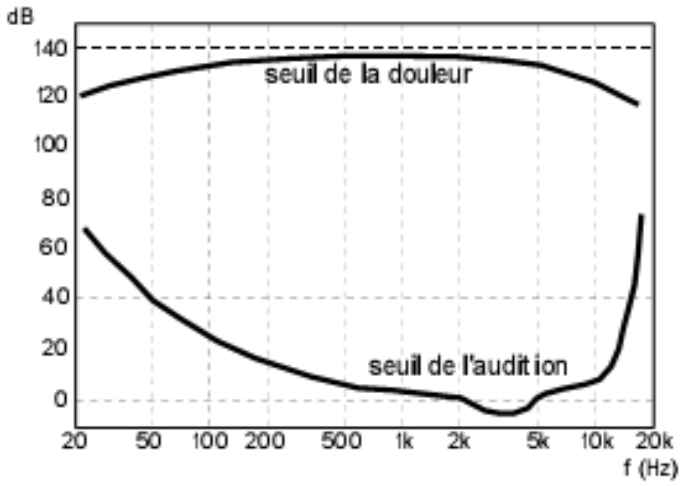


Figure 1.4 : Courbe du champ auditif humain.

1.5 Généralités sur le bruit

1.5.1 Définition du bruit

On appelle bruit, toute sensation auditive désagréable et gênante. Le bruit est un ensemble de sons produits par une ou plusieurs sources, qui provoquent des vibrations de l'air et se propagent en faisant vibrer le tympan de notre oreille. Donc, le bruit peut être défini comme étant tous les phénomènes qui empêchent la transmission d'un message d'une source à sa destination ou tout ce qui détériore la qualité et l'intelligibilité du message transmis.

Au sens le plus large, tout signal indésirable qui affecte à un degré ou à un autre l'intégrité et l'intelligibilité d'un signal utile peut être considéré comme du bruit.

1.5.2 Origine et physiologie du bruit

Le bruit qui affecte les performances d'un système quelconque peut provenir de deux sources différentes : une source interne, où les bruits sont générés à l'intérieur du système (bruit thermique, bruits de fond etc...), et une source externe, où les bruits sont générés à l'extérieur du système et influencent ce dernier. Ils peuvent être naturels (perturbations atmosphériques, solaire, cosmique) ou artificiels (d'origine industrielle comme les bruits causés par les machines, véhicules, etc...).

Les différents paramètres qui caractérisent la physiologie du bruit sont :

- **La fréquence** : la fréquence est le critère qui va permettre d'identifier les différents types de bruit. Notre oreille peut écouter des bandes sonores allant jusqu'à 20 kHz. Il est intéressant de savoir que nous n'allons pas entendre certains types de bruit. Soit parce qu'il s'agit de très basses fréquences, donc de sons très graves, soit parce qu'il s'agit de sons plus aigus.
- **L'intensité** : pour l'homme, un son d'une intensité supérieure à 65 décibels est perçu comme un son désagréable. A partir de 85 décibels environ, nous avons un seuil de gêne. Il existe toute une graduation de 80 à 120 décibels, voire au-delà, qui conduit à des lésions irrémédiables dans le mécanisme de l'oreille.
- **La durée** : un bruit constant, bien qu'il ne soit pas agressif pour une personne, devient, compte tenu de la durée d'exposition, très nocif.
- **La répétition** : il existe certains bruits qui n'ont pas besoin d'être à forte puissance pour être gênants, c'est le cas de certains événements sonores répétitifs selon une fréquence précise qui peuvent entrer en résonance avec notre émotivité.

1.5.3 Différents types de bruit

Plusieurs méthodes de classification des bruits peuvent être opérées (il n'existe pas de classification typique). Ces classifications dépendent le plus souvent de l'application et de l'objectif poursuivis. Ainsi selon l'importance accordée à tel ou tel paramètre du signal, une classification peut apparaître comme étant plus intéressante qu'une autre [43]. On distingue :

1.5.3.1 Bruits additifs, convolutionnels et physiologiques

Les différents bruits pouvant influés sur un message peuvent être divisés en deux grandes catégories : les bruits additifs et les bruits convolutionnels. La distinction entre les deux peut être faite par le nombre d'agents agresseurs extérieurs à la transmission du message. Les bruits additifs sont causés par des agents extérieurs au trinôme source-voie-destinataire alors que les bruits convolutionnels sont causés par la moindre qualité de la voie de communication, celle-ci ayant alors un rôle ambigu, du point de vue du message, de médium et d'agresseur.

a- Les bruits additifs

Les bruits additifs sont dus à la multiplicité des systèmes de communication dans un même environnement. Plusieurs émetteurs et plusieurs récepteurs pouvant être confinés dans un même espace, les messages de tous les émetteurs peuvent donc se trouver en concurrence sur une même voie, sans que les récepteurs ne possèdent un mécanisme infaillible pour isoler le message qui leur est destiné. L'émetteur et le récepteur peuvent aussi se trouver en présence d'un ou de plusieurs équipements générant un bruit de fond d'amplitude variable.
Les bruits additifs peuvent être subdivisés en trois groupes en fonction des lieux où ils peuvent être rencontrés :

● **Bruits des systèmes industriels :** Ils correspondent aux bruits émis par des machines possédant une faible isolation phonique. Ils peuvent être très intenses et sont, par nature, non stationnaires. Ils sont très souvent des bruits rythmiques, ou périodiques, correspondant à la répétition d'une tâche de nature productive. L'automatisation totale des sites de production n'étant pas encore atteinte, il faut également considérer les bruits produits directement ou indirectement par l'homme. Au titre de ceux-ci peut se retrouver le bruit de parole qui est le fait immédiat de l'homme. Il est également possible de classer, dans ce type de bruits produits par les systèmes industriels, le bruit des outils de travail des ouvriers présents sur un site, tel que le bruit du petit matériel électrique.

● **Bruits des moyens de transport :** ils correspondent aux bruits qui peuvent être observés dans divers véhicules tels que les voitures, les trains ou les avions. Ils se caractérisent généralement par une très forte stationnarité qui correspond à la vitesse de fonctionnement des organes moteurs. Le bruit observé est ainsi constitué d'un ou de plusieurs harmoniques et ne comporte que de micro-fluctuations. Ces remarques générales doivent cependant être nuancées par l'observation du bruit de certains moyens de transport tels que le train ou le bateau. Dans le cas du train, un bruit non-stationnaire et rythmique est présent tout au long du déplacement. Dans le cas du bateau, le moteur peut fonctionner de manière très lente, surtout sur de grosses unités, et produire lui aussi un bruit rythmique. Ces bruits peuvent également varier en fonction des conditions de déplacement.

● **Bruits des milieux administratifs et urbains :** ce sont les bruits présents dans les bureaux, les domiciles ou dans les concentrations urbaines. Ces bruits peuvent être très variés (climatisation ou bruit de parole) mais sont peu intenses et sont toujours momentanés au contraire des bruits de moyens de transport où l'auditeur est un passager.

b- Les bruits convolutionnels

Les bruits convolutionnels (ou multiplicatifs) sont dus à la distorsion induite par la voie de communication. Ils résultent de la mauvaise qualité d'un ou de plusieurs éléments de support du message ou, tout simplement, de son étroitesse en bande passante.
Les sociétés modernes utilisent de plus en plus des moyens de communication à longue distance tels que le téléphone, les moyens radiophoniques et radiotéléphoniques. Ces moyens de communication à longue distance ont été élaborés à partir d'un compromis coût/efficacité. La parole, lorsqu'elle est transmise par un tel moyen, est forcément dégradée tout en gardant une grande intelligibilité.

c- Les bruits physiologiques :

Ce sont des bruits spécifiques à l'être humain lors de la phase de production de la parole. L'homme essaie, lui, de s'adapter aux conditions sonores rencontrées en modifiant sa méthode de production de la parole, dans certaines situations la parole est modifiée sans que l'homme ne modifie sa façon de parler de manière volontaire.

1.5.3.2 Classification basée sur la densité de probabilité

Ne pouvant disposer ni d'une formulation analytique et ni évidemment d'une représentation mathématique déterministe du processus aléatoire, nous le caractérisons alors par ses propriétés statistiques. La théorie des probabilités est alors utilisée pour la description statistique des bruits et autres signaux aléatoires. On peut alors distinguer au moins trois types de bruits :

Le bruit blanc : est une réalisation d'un processus aléatoire dans lequel la densité spectrale de puissance est la même pour toutes les fréquences. En synthèse et traitement du son, on ne considère que les fréquences comprises entre 20 Hz et 20 kHz puisque l'oreille humaine n'est sensible qu'à cette bande de fréquences. L'impression obtenue est celle d'un souffle. Le bruit blanc doit son appellation à l'analogie avec la lumière blanche. Cette lumière blanche est due à la présence de photons de toutes les valeurs d'énergie, et est donc composée de toutes les couleurs. De même, un bruit blanc possède un spectre de fréquence continu et d'amplitude moyenne constante indépendante des fréquences. Le bruit blanc reste un modèle théorique et un tel bruit ne peut pas exister naturellement car il aurait sinon une puissance infinie.

Le bruit Gaussien : se dit d'un processus qui possède une densité de probabilité Gaussienne. L'importance des processus Gaussiens résulte du fait qu'ils sont le modèle asymptotique d'un grand nombre de phénomènes naturels. D'autre part, ils sont entièrement définis par leurs valeurs moyennes et covariances. L'autre propriété fondamentale réside dans le fait qu'un bruit Gaussien lorsqu'il transite par un système linéaire reste toujours Gaussien.

Bruit coloré : En physique, en acoustique et en traitement du signal, bien que le bruit soit, par nature, un signal aléatoire, il peut posséder certaines caractéristiques statistiques. La distribution de la densité spectrale est l'une de ses caractéristiques. En se basant sur ce critère, par analogie avec la lumière, on parle de bruit coloré. La notion de couleur du bruit est couramment utilisée, mais ne représente pas de propriété physique particulière. Beaucoup de couleurs s'appliquent à des signaux ayant des composantes à toutes fréquences, avec une densité spectrale par unité de fréquence proportionnelle à $1/f^{\beta}$. Par exemple, le bruit blanc est plat avec $\beta = 0$, alors que $\beta = 2$ pour le bruit brun.

1.5.3.3 Classification basée sur les propriétés du bruit [14]

Un bruit peut être caractérisé par différentes propriétés. La connaissance de ces propriétés permet d'adopter une stratégie de rehaussement plus adéquate. Le tableau (1.1) présente les propriétés essentielles caractérisant un bruit :

Propriétés	Types
Structure	Continu / Impulsionnel / Périodique
Type d'interaction	Additif / Convolutif
Comportement dans le domaine temporel	Stationnaire / Non stationnaire
Bande de fréquence	Large bande / Bande étroite
Dépendance du signal	Corrélé / Non corrélé
Propriétés statistiques	Dépendant / Indépendant
Propriétés spatiales	Cohérent / Incohérent

Tableau 1.1 : Classification du bruit basée sur plusieurs propriétés.

1.5.4 Bruitage de la parole dans un système de communication

Dans un système de communication ou de reconnaissance de la parole, constitué d'un émetteur, récepteur et canal de transmission, le signal de la parole peut être affecté par du bruit à n'importe quelle étape de la chaîne de communication comme suit [44] :

- **A la source de la parole (émetteur)**

Quand la source elle-même est dans un environnement bruité, le bruit qui caractérise cet environnement s'ajoute au signal de la parole prononcé par un locuteur qui se trouve dans cet environnement; la nature du bruit dépend du milieu : bruit peut être stationnaire (par exemple le bruit d'un ventilateur) ou non stationnaire (par exemple bruit du clavier). De plus, les équipements d'enregistrement et de transmission de la parole utilisent des microphones pour capter le signal de la parole, en faisant une transformation d'une énergie acoustique en une énergie électrique. Un mauvais choix de ces capteurs acoustiques résulte à d'autres bruits qui dégradent de plus la parole à la source même. De plus, la distance du locuteur et sa position par rapport au microphone peuvent varier avec le temps, ce qui résulte en une distorsion de l'amplitude du signal.

- **Durant la transmission**

Le signal de parole est généralement transmis à un récepteur lointain à travers des canaux de transmission : quand plusieurs canaux sont présents dans le même moyen de la transmission, ils peuvent influencer l'un sur l'autre, comme dans le cas de la téléphonie, la visiophonie, etc… Durant cette transmission, plusieurs bruits additionnels affectent le signal de parole, à cause du comportement non idéal du canal. D'autres bruits peuvent également s'ajouter durant la conversion analogique/numérique (CAN) des données, avant la transmission et durant la reproduction de la parole à la fin de la chaîne de communication, lors de la conversion numérique/analogique (CNA).

- **Le bruit à la réception**

Parfois, bien que la source de la parole puisse être dans un environnement silencieux, le récepteur du signal parole peut être dans un environnement fortement bruité. Ici également, la fatigue d'écoute surgit pendant que la qualité de la parole reçue est nettement dégradée. Par conséquent, le rehaussement de la parole est nécessaire dans ce cas aussi.

1.5.4.1 Rapport signal sur bruit

Le rapport signal sur bruit (RSB), en anglais : signal to noise ratio (SNR), comme son nom l'indique, fournit le rapport entre la puissance moyenne du signal et celle du bruit, c'est le critère le plus couramment utilisé pour désigner la qualité d'une transmission d'information par rapport aux parasites. Il est défini – en décibel (dB) – par :

$$SNR_{dB} = 10\log_{10}\left[\frac{\sum_{n=-\infty}^{\infty} s^2(n)}{\sum_{n=-\infty}^{\infty} d^2(n)}\right] \qquad (1.7)$$

Où $s(n)$ est le signal propre et $d(n)$ le bruit.

1.5.4.2 Quelques corpus de bruits

Afin de tester les algorithmes de rehaussement de la parole bruitée, on est amené à bruiter le signal de parole par des bruits de référence comme ceux fournis par le corpus NOISEX [45], ou d'utiliser une base de données de parole bruitée de référence comme le corpus Noizeus [46].

a- Le corpus Noise-Rom-0

Le premier corpus développé, le corpus Noise-Rom-0, comprend 24 bruits. Ce corpus a été développé en 1988 par l'Institut TNO pour l'étude de la perception, à Soesterberg aux Pays-Bas, avec l'appui du groupe de recherche et d'étude *Speech Processing* (RSG 10) de l'OTAN (Organisation du Traité de l'Atlantique Nord). Ces 24 bruits sont d'origine très diverses comme le montre la liste complète ci-dessous, liste donnée dans l'ordre croissant des numéros de bruit.
1- bruit généré par une sinusoïde ayant une fréquence de 1000 Hz.
2- bruit rose.
3- bruit blanc.
4- bruit blanc atténué de 6 décibels par octave de 250 Hz.
5- bruit blanc atténué de 12 décibels par octave de 250 Hz.
6- bruit de parole, ayant les propriétés de masquage d'un environnement bruité par de la parole.
7- bruit du M 109 à 30 km/h.
8- bruit de Buccaneer à 190 nœuds et 1000 pieds.
9- bruit du Leopard 2 à 70 km/h.
10- bruit de véhicule de transport à roues à 50-60 km/h.
11- bruit de Buccaneer à 450 nœuds et 300 pieds.
12- bruit de l'hélicoptère Lynx sur plate-forme.
13- bruit du Leopard 1 à 70 km/h.
14- bruit dans une salle de commandement d'un contre-torpilleur.
15- bruit dans une salle des machines d'un contre-torpilleur.
16- bruit de mitrailleuse.
17- bruit de canal radio hautes-fréquences.
18- signal de test du bateau STITEL.
19- bruit de parole (*voice babble, canteen, 100 people*), bruit réel.
20- bruit d'un chasseur F16 biplace à 500 nœuds et 300-600 pieds en place copilote.
21- bruit d'une usine de production de voitures : bruits de soudures électriques lors de l'assemblage du bas de caisse (*car floor production*).
22- bruit d'une usine de production de voitures : bruits du hall d'assemblage.
23- bruit de voiture en déplacement : Volvo 340 à 120 km/h en 4ème vitesse sur une route goudronnée.
24- bruit de voiture en déplacement : Volvo 340 à 50 km/h en 3ème vitesse sur une route pavée.

Tous ces bruits peuvent être regroupés en trois catégories différentes en fonction de leur univers de rattachement : bruits de référence propres au domaine du décodage acoustico-phonétique (DAP), bruits d'origine industrielle (ou plutôt civile...) et bruits d'origine militaire.

Les bruits propres au domaine de DAP regroupent les bruits synthétiques et les bruits de parole. Les bruits synthétiques sont, généralement, générés à partir de fonctions sinusoïdales ou Gaussiennes et possèdent des spectres très stables. Dans cette première sous-catégories se trouvent les bruits 1, 2, 3, 4 et 5 du corpus Noise-Rom-0. Les bruits de parole, la deuxième sous-catégorie de ce groupe, sont les bruits 6 et 19 du corpus.

La deuxième catégorie regroupe les bruits d'origine industrielle et, plus généralement, civile. Elle est composée de deux bruits de véhicules en déplacement (numéros 23 et 24) et de deux bruits d'atelier de fabrication (bruit 21 et 22). Alors que les deux premiers sont stationnaires, les deux derniers ne le sont pas.

La dernière catégorie regroupe tous les bruits d'origine militaire, que ces bruits soient stationnaires ou non. Ces bruits ont été enregistrés à bord de véhicules des trois "armes" (terre, air et mer) des armées de l'OTAN. Les bruits de canal radio hautes-fréquences (bruit 17) et de rafales de mitrailleuse (bruit 16) sont les seules exceptions à cette règle. L'expérimentateur

dispose ainsi de bruits de véhicules terrestres (bruits 7, 9, 10 et 13), de bruits enregistrés à bord d'unités navales (bruits 14, 15 et 18) et de bruits originaires du monde aéronautique (bruits 8, 11, 12, et 20).

Tous ces bruits sont enregistrés à une fréquence d'échantillonnage de 20 kHz et fournis de manières isolées. Aucun signal de parole bruité n'est fourni en complément alors qu'un bruitage effectué a priori aurait permis de disposer d'une base de comparaison encore mieux définie. Pour palier à cet inconvénient, un autre corpus de bruits, comprenant également des signaux de parole bruités selon un certain nombre de rapports signal sur bruit, a été défini.

b- Le corpus Noisex-92 [45]

Le corpus Noisex-92 a été conjointement mis au point, en 1992, à partir du corpus Noise-Rom-0 par l'Institut TNO pour l'étude de la perception et par l'équipe de recherche sur la parole de la *Defense Research Agency* Anglaise. Seuls certains bruits ont été sélectionnés par rapport à l'ensemble de ceux disponibles dans le premier corpus. De plus, ces bruits ont été rééchantillonnés de manière à être compatibles avec des signaux de parole préalablement enregistrés à une fréquence de 16 kHz.

c- La base de données Noizeus [46]

Les signaux qui vont être exploités tout au long de ce document sont extraits de la base de donnée « Noizeus ». L'utilisation d'un tel corpus est pour faciliter la comparaison des algorithmes de rehaussement de la parole et de mesurer leurs performances en terme de distorsion. Cette base de données est devenue une base standard pour l'évaluation des algorithmes de rehaussement. Elle contient 30 phrases phonétiquement équilibrées, ces phrases sont bruitées artificiellement par huit bruits réels à différent rapport signal sur bruit (0 dB, 5 dB, 10 dB et 15 dB).

Les bruits ont été sélectionnés de la base de données AURORA comme : le bruit de parole, le bruit de voiture, train, restaurant, exhibition, rue, aéroport et le bruit dans une station de train. Les trente phrases de la base de données ont été produites par trois locuteurs et trois locutrices (5 phrases / locuteur) et enregistrées dans une salle acoustiquement isolée, avec une fréquence d'échantillonnage de 25 kHz et reéchantillonnées à 8 kHz.

Comme cette base ne contient pas de bruit blanc, une routine Matlab a été élaborée où nous avons ajoutés du bruit blanc aux trente fichiers de paroles propres aux différents rapports signal sur bruit.

1.5.4.3 Densités spectrales de puissance et spectrogrammes de quelques bruits

Nous avons sélectionné une phrase et quelques bruits du corpus Noizeus dont nous disposons, pour illustrer ses caractéristiques temporelles et fréquentielles. En général, le signal vocal se caractérise par une pente de 6 dB/octave, due à l'influence de la source d'excitation et au rayonnement des lèvres. Une pente de 6 dB/octave veut dire que les hautes fréquences ont une énergie plus faible que celle des basses fréquences, ceci est mieux illustré dans la représentation du spectre de puissance de la parole propre à la figure suivante.
Les autres courbes illustrent les densités spectrales de puissance (DSP) d'un bruit blanc et de quelques bruits colorés.

Contrairement au bruit blanc Gaussien, qui a un spectre de puissance plat, le spectre des bruits réels est différent. Par conséquent, le signal du bruit n'affecte pas uniformément le signal de la parole sur le tout spectre, quelques fréquences sont beaucoup plus affectées par rapport aux autres suivant le type de bruit.

On trouve en comparant les DSP des différents bruits, que les caractéristiques spectrales du bruit de parole (babble noise) ressemblent à celles du signal de la parole, son énergie principale est concentrée dans les basses fréquences (les fréquences significatives de la parole

elle-même). Par conséquent, elles sont affectées plus que les hautes fréquences. Donc l'utilisation de ce bruit complique d'avantage le système de débruitage par rapport au bruit blanc. Les autres bruits (voiture, aéroport, rue) permettent des degrés de contamination différente entre celle d'un bruit blanc et celle d'un bruit babble.

En plus du type de bruit choisi, le degré de contamination est lié au RSB utilisé, La figure (1.6) montre l'influence du bruit blanc ajouté à un signal de parole propre pour des RSB différents (RSB = 5 dB et 0 dB). L'évolution temporelle du signal de parole et le spectrogramme montrent l'augmentation de l'influence du bruit à mesure que le RSB diminue, jusqu'à la disparition du signal de parole (pour des RSB très faibles).

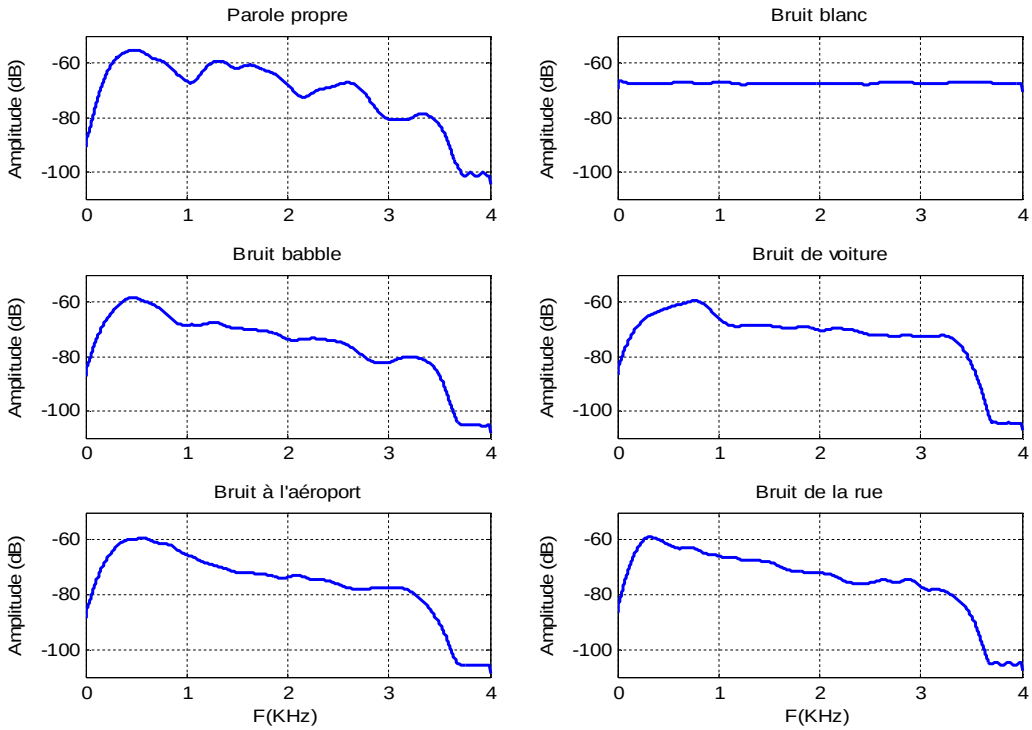


Figure 1.5 : Densités spectrales de puissance.

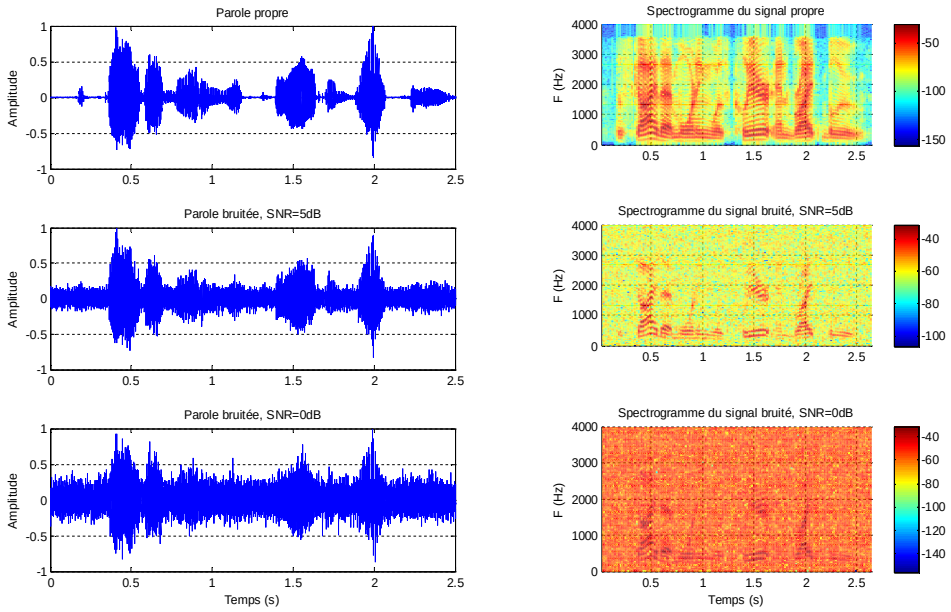


Figure 1.6 : Représentations temporelles et spectrogrammes d'une phrase.

1.6 Propriétés statistiques des coefficients de la TFD

Généralement, on traite et on analyse le signal de la parole dans le domaine spectrale en utilisant une analyse spectrale à court terme et la transformée de Fourier discrète (TFD). Il est donc intéressant d'étudier les propriétés statistiques des coefficients de la TFD. Considérant un signal $y(n)$ pondéré par une fenêtre $w(n)$ de taille N, le calcul de la transformée de Fourier de ce segment fenêtré est donné par :

$$Y_k = \sum_{n=0}^{N-1} w(n) \cdot y(n) \cdot e^{\frac{-j2\pi kn}{N}} \tag{1.8}$$

Où k est l'indice fréquentiel, $k \in \{0,\ 1,\ \ldots,\ N-1\}$.

Donc, les coefficients de Fourier complexes Y_k constituent un processus aléatoire [10].

Y_k peut être présenté comme suit :

$$Y_k = \mathrm{Re}\{Y_k\} + j\,\mathrm{Im}\{Y_k\} \tag{1.9}$$

Ou en fonction de l'amplitude et la phase :

$$Y_k = R_k e^{j\theta_k} \tag{1.10}$$

1.6.1 Propriétés statistiques asymptotiques [10]

Pour étudier les propriétés asymptotiques des coefficients de la TFD nous supposons que : La taille N est suffisamment grande, $N \to \infty$, et La taille N est plus grande que la portée de la corrélation de $y(n)$.

Cette dernière condition, exclut les signaux périodiques de cette discussion.

Si le signal d'entrée est aléatoire, on peut conclure à partir du théorème centrale limite que pour $k \notin \{0, N/2\}$, les deux parties réelle et imaginaire des coefficients de la TFD (Y_k) peuvent être modélisées comme étant des variables aléatoires Gaussiennes, mutuellement indépendantes de moyennes nulles et de variance : $0.5\sigma_{Y_k}^2 = 0.5E\left\{ |Y_k|^2 \right\}$, i.e.,

$$p_{\mathrm{Re}\{Y_k\}}(u) = \frac{1}{\sqrt{\pi}\sigma_{Y_k}} \exp\left(-\frac{u^2}{\sigma_{Y_k}^2} \right) \tag{1.11}$$

$$p_{\mathrm{Im}\{Y_k\}}(v) = \frac{1}{\sqrt{\pi}\sigma_{Y_k}} \exp\left(-\frac{v^2}{\sigma_{Y_k}^2} \right) \tag{1.12}$$

Pour un signal d'entrée $y(n)$ de valeurs réelles et $k \in \{0, N/2\}$, la partie imaginaire de Y_k est nulle et la partie réelle suit une loi normale de variance $\sigma_{Y_k}^2 = \mathrm{E}\left\{ |Y_k|^2 \right\}$.

Donc, pour $k \notin \{0, N/2\}$ la distribution conjointe des deux parties réelle et imaginaire est donnée par :

$$p_{\mathrm{Re}\{Y_k\},\mathrm{Im}\{Y_k\}}(u,v)=\frac{1}{\pi\sigma_{Y_k}^2}\exp\left(-\frac{u^2+v^2}{\sigma_{Y_k}^2}\right) \tag{1.13}$$

Ou avec $z=u+jv$ par

$$p_{Y_k}(z)=\frac{1}{\pi\sigma_{Y_k}^2}\exp\left(-\frac{|z|^2}{\sigma_{Y_k}^2}\right) \tag{1.14}$$

Par passage vers les coordonnées polaire, $Y_k=R_k e^{j\theta_k}$, on obtient une distribution de Rayleigh pour l'amplitude [10] :

$$p_{R_k}(u)=\begin{cases}\dfrac{2u}{\sigma_{Y_k}^2}\exp\left(-\dfrac{u^2}{\sigma_{Y_k}^2}\right), & u\geq 0\\ 0, & u<0\end{cases} \tag{1.15}$$

Et sa phase θ_k suit une loi uniforme entre 0 et 2π :

$$p_{\theta_k}(u)=\begin{cases}\dfrac{1}{2\pi}, & 0\leq u\leq 2\pi\\ 0\ , & \textit{ailleurs}\end{cases} \tag{1.16}$$

Comme pour un modèle Gaussien, l'amplitude et la phase sont statistiquement indépendantes, la densité conjointe est le produit des deux densités :

$$p_{R_k,\theta_k}(u,v)=p_{R_k}(u)\cdot p_{\theta_k}(v)=\begin{cases}\dfrac{u}{\pi\sigma_{Y_k}^2}\exp\left(-\dfrac{u^2}{\sigma_{Y_k}^2}\right), & u\geq 0 \ \text{ et } \ 0\leq v\leq 2\pi\\ 0, & \textit{ailleurs}\end{cases} \tag{1.17}$$

En outre, l'amplitude au carrée de chaque composante fréquentielle $|Y_k|^2=R_k^2$ est une variable aléatoire exponentielle, sa densité de probabilité est donnée par :

$$p_{R_k^2}(u)=\begin{cases}\dfrac{1}{\sigma_{Y_k}^2}\exp\left(-\dfrac{u}{\sigma_{Y_k}^2}\right), & u\geq 0\\ 0, & u<0\end{cases} \tag{1.18}$$

1.6.2 Modèle signal plus bruit

Dans le cas du rehaussement de la parole, le signal observé est : $y(n)=x(n)+d(n)$, qui est la somme d'un signal désiré $x(n)$ et un signal de bruit $d(n)$ statistiquement indépendant de $x(n)$. Par conséquent, cela conduit à un modèle de bruit additif même dans le domaine fréquentiel, $Y_k=X_k+D_k$.

Le calcul de la densité conditionnelle des coefficients Y_k observés sachant les coefficients souhaités $X_k=A_k e^{j\alpha_k}=A_k\left[\cos(\alpha_k)+j\sin(\alpha_k)\right]$ sous l'hypothèse de la distribution Gaussienne sera présenté dans la section suivante.

Comme le signal désiré et le bruit sont additifs et statistiquement indépendants, les densités conditionnelles des deux parties réelle et imaginaire sont données par :

$$p_{\mathrm{Re}\{Y_k\}|\mathrm{Re}\{X_k\}}\left(u\left|\mathrm{Re}\{X_k\}\right.\right)=\frac{1}{\sigma_{D_k}\sqrt{\pi}}\exp\left(-\frac{\left(u-A_k\cos(\alpha_k)\right)^2}{\sigma_{D_k}^2}\right) \tag{1.19}$$

$$p_{\mathrm{Im}\{Y_k\}|\mathrm{Im}\{X_k\}}\left(v\left|\mathrm{Im}\{X_k\}\right.\right)=\frac{1}{\sigma_{D_k}\sqrt{\pi}}\exp\left(-\frac{\left(v-A_k\sin(\alpha_k)\right)^2}{\sigma_{D_k}^2}\right) \tag{1.20}$$

Avec $z=u+jv$, la densité de probabilité conjointe est donnée par :

$$p_{\mathrm{Re}\{Y_k\},\mathrm{Im}\{Y_k\}|X_k}\left(u,v\left|X_k\right.\right)=\frac{1}{\pi\sigma_{D_k}^2}\exp\left(-\frac{\left|z-A_k\exp(j\alpha_k)\right|^2}{\sigma_{D_k}^2}\right)=\frac{1}{\pi\sigma_{D_k}^2}\exp\left(-\frac{|z|^2+A_k^2-2A_k\,\mathrm{Re}\{\exp(-j\alpha_k)z\}}{\sigma_{D_k}^2}\right) \tag{1.21}$$

Comme la rotation dans le plan complexe ne change pas l'amplitude,

$$|z|^2=\left|z\exp(-j\alpha_k)\right|^2=\mathrm{Re}\left\{z\exp(-j\alpha_k)\right\}^2+\mathrm{Im}\left\{z\exp(-j\alpha_k)\right\}^2,$$

La densité de probabilité conjointe peut être aussi écrite ainsi :

$$p_{\mathrm{Re}\{Y_k\},\mathrm{Im}\{Y_k\}|X_k}\left(u,v\left|X_k\right.\right)=\frac{1}{\pi\sigma_{D_k}^2}\exp\left(-\frac{\left(\mathrm{Re}\{\exp(-j\alpha_k)z\}-A_k\right)^2+\mathrm{Im}\{\exp(-j\alpha_k)z\}^2}{\sigma_{D_k}^2}\right) \tag{1.22}$$

Ce qui donne pour l'amplitude conditionnelle [17].

$$p_{R_k|X_k}\left(u\left|X_k\right.\right)=\begin{cases}\dfrac{2u}{\sigma_{D_k}^2}\cdot\exp\left(-\dfrac{u^2+A_k^2}{\sigma_{D_k}^2}\right)\cdot I_0\left(\dfrac{2A_ku}{\sigma_{D_k}^2}\right), & u\geq 0\\ 0, & u<0\end{cases} \tag{1.23}$$

Où $I_0(.)$ est la fonction de Bessel modifiée de première espèce. Quand la parole est absente$(A_k=0)$, l'amplitude suit la distribution de Rayleigh comme décrit avant.

1.6.3 Propriétés statistiques des coefficients de la TFD pour des trames de courtes durées

Contrairement à la section précédente, on considère que la taille N de la TFD est courte, telle qu'il est utilisé dans les communications mobiles et les autres applications. Donc, les hypothèses statistiques asymptotiques ne sont pas bien remplies, spécialement pour les sons voisés qui présentent une corrélation élevée [23].

Donc, pour des trames de taille < 100 ms, les densités de probabilité de Laplace, Gamma et Gamma généralisée modélisent beaucoup mieux les parties réelles et imaginaires des coefficients de la TFD que le modèle Gaussien.

1.7 Estimation optimale

Généralement, on essaye d'estimer une variable aléatoire x quand les observations $\boldsymbol{y}=(y_1, \ldots, y_N)^T$ d'un autre vecteur aléatoire sont données. La valeur estimée résultante est : $\hat{x}=f(\boldsymbol{y})$.

Dans le cas du traitement de la parole, plusieurs critères d'optimisation ont été utilisés [10]. L'estimateur du maximum de vraisemblance (ML : Maximum Likelihood) donne la valeur de x telle que la densité de probabilité conjointe conditionnelle des variables observées est maximale, c'est-à-dire :

$$\hat{x}=\arg \max_x p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|x) \tag{1.24}$$

L'estimateur du maximum a posteriori (MAP) est défini par :

$$\hat{x}=\arg \max_x p_{x|\mathbf{y}}(x|\mathbf{y})=\arg \max_x \frac{p_{\mathbf{y}|x}(\mathbf{y}|x)\,p_x(x)}{p_{\mathbf{y}}(\mathbf{y})} \tag{1.25}$$

Où maintenant la distribution a priori $p_x(x)$ de la variable inconnue x est utilisée. Quand $p_x(x)$ est uniformément distribuée, l'estimateur MAP donne la même estimation que l'estimateur (ML).

D'une manière générale, on minimise la moyenne statistique d'une fonction de coût $C(x,\hat{x})$.

$$\mathrm{E}_{x\mathbf{y}}\{C(x,\hat{x})\}=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} C(u, f(\mathbf{y}))p_{x\mathbf{y}}(u, y_1,\cdots y_N)\,du dy_1\cdots dy_N \tag{1.26}$$

Le plus important de ces estimateurs est l'estimateur basé sur la minimisation de l'erreur quadratique moyenne (MMSE estimator : Minimum Mean Square Error estimator), où $C(x,\hat{x})=(x-\hat{x})^2$, qui sera détaillé dans la section suivante.

1.7.1 Estimateur basé sur la minimisation de l'erreur quadratique moyenne

La solution optimale de $\hat{x}$ au sens du MMSE est l'espérance conditionnelle de x sachant le vecteur d'observation $\mathbf{y}=(y_1,\cdots,y_N)^T$

$$\hat{x}=\mathrm{E}_x\{x|\mathbf{y}\}=\int_{-\infty}^{\infty} u\cdot p_{x|\mathbf{y}}(u\mid y_1,\cdots y_N)\,du \tag{1.27}$$

L'erreur quadratique moyenne de cet estimateur est :

$$\begin{aligned}\mathrm{E}_{x\mathbf{y}}\{(x-\hat{x})^2\}&=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(u-\hat{x})^2 p_{x\mathbf{y}}(u, y_1,\cdots y_N)\,du\cdot dy_1\cdots dy_N\\&=\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}(u-\hat{x})^2\, p_{x|\mathbf{y}}(u|y_1,\cdots,y_N)\cdot p_{\mathbf{y}}(y_1,\cdots,y_N)\,du\cdot dy_1\cdots dy_N\end{aligned} \tag{1.28}$$

Comme la densité de probabilité est non négative, il suffit de minimiser l'intégrale $\int_{-\infty}^{\infty}(u-\hat{x})^2\, p_{x|\mathbf{y}}(u|y_1,\cdots,y_N)\,du$ pour chaque vecteur des observations $\mathbf{y}$ donné.

Le résultat de l'estimateur est obtenu en annulant la dérivée première par rapport à $\hat{x}$ [10].

$$\hat{x} = \int_{-\infty}^{\infty} u \cdot p_{x|y}\left(u|y_1,\cdots,y_N\right) du = \mathrm{E}_x\left\{x|\mathbf{y}\right\} \tag{1.29}$$

En général, $\hat{x}$ est une fonction non linéaire des valeurs observées du vecteur $\mathbf{y}$.

1.7.2 Estimateur linéaire optimal

Nous simplifions la procédure d'estimation, en considérant que l'estimation $\hat{x}$ est une combinaison linéaire des données observées,

$$\hat{x} = \mathbf{h}^T \cdot \mathbf{y} \tag{1.30}$$

Où $\mathbf{h}^T = \left(h_1,\cdots,h_N\right)$ est le vecteur des pondérations constantes. Le développement de $\mathrm{E}\left\{\left(x-\hat{x}\right)^2\right\}$ conduit à :

$$\mathrm{E}\left\{\left(x-\hat{x}\right)^2\right\} = \mathrm{E}\left\{x^2\right\} - 2\mathrm{E}\left\{x\mathbf{y}^T\right\}\mathbf{h} + \mathbf{h}^T\mathrm{E}\left\{\mathbf{y}\mathbf{y}^T\right\}\mathbf{h} \tag{1.31}$$

La minimisation de l'erreur quadratique moyenne conduit à :

$$\mathbf{h} = \mathbf{R}_{yy}^{-1} \cdot \mathbf{r}_{xy} \tag{1.32}$$

Où $\mathbf{r}_{xy}$ est définie comme suit :

$$\mathbf{r}_{xy} = \left(\mathrm{E}\left\{xy_1\right\},\cdots,\mathrm{E}\left\{xy_N\right\}\right)^T \tag{1.33}$$

Donc, pour un vecteur d'observation $\mathbf{y}$, on calcule la valeur estimée par :

$$\hat{x} = \mathbf{y}^T \cdot \mathbf{R}_{yy}^{-1} \cdot \mathbf{r}_{xy} \tag{1.34}$$

Contrairement à la solution générale qui est non linéaire, les densités de probabilité du signal et du bruit ne sont pas utilisées dans la solution linéaire. Le vecteur des pondérations $\mathbf{h} = \mathbf{R}_{yy}^{-1} \cdot \mathbf{r}_{xy}$ est une fonction des statistiques d'ordre 2 mais pas une fonction du vecteur $\mathbf{y}$ lui-même.

1.7.3 Cas Gaussien

Dans cette section, l'observation $y = x + d$ est constituée par l'addition de deux signaux x et d, considérés comme Gaussiens. Ces deux signaux sont non corrélés et de moyennes nulles, la densité de probabilité de y est donnée par :

$$p_y\left(v\right) = \frac{1}{\sqrt{2\pi\left(\sigma_x^2 + \sigma_d^2\right)}} \exp\left(-\frac{v^2}{2\left(\sigma_x^2 + \sigma_d^2\right)}\right) \tag{1.35}$$

Et comme la convolution de deux Gaussiennes est Gaussienne, la densité conditionnelle $p_{y|x}\left(v|u\right) = p_d\left(v-u\right)$ et la densité de x peuvent être écrites ainsi :

$$p_d(v-u)=\frac{1}{\sqrt{2\pi}\sigma_d}\exp\left(-\frac{(v-u)^2}{2\sigma_d^2}\right) \tag{1.36}$$

Et

$$p_x(u)=\frac{1}{\sqrt{2\pi}\sigma_x}\exp\left(-\frac{u^2}{2\sigma_x^2}\right) \tag{1.37}$$

respectivement. L'estimateur non linéaire au sens du MMSE est donc donné par :

$$\hat{x}=\int_{-\infty}^{\infty}u\cdot p_{x|y}(u|v)\,du=\frac{1}{p_y(v)}\int_{-\infty}^{\infty}u\cdot p_{y|x}(v|u)\,p_x(u)\,du \tag{1.38}$$

$$=K\int_{-\infty}^{\infty}u\cdot\exp\left(-u^2\frac{\sigma_x^2+\sigma_d^2}{2\sigma_x^2\sigma_d^2}+\frac{vu}{\sigma_d^2}\right)du \tag{1.39}$$

Où :

$$K=\frac{1}{p_y(v)}\cdot\frac{\exp\left(-v^2/2\sigma_d^2\right)}{2\pi\sqrt{\sigma_d^2\sigma_x^2}}$$

Avec :

$$\int_{-\infty}^{\infty}x\exp\left(-px^2+2qx\right)dx=\frac{q}{p}\sqrt{\frac{\pi}{p}}\cdot\exp\left(\frac{q^2}{p}\right),\qquad \mathrm{Re}\{p\}>0 \tag{1.40}$$

Enfin, on obtient l'estimateur en fonction de la variable aléatoire observée y comme suit :

$$\hat{x}=\frac{\sigma_x^2}{\sigma_x^2+\sigma_d^2}y=\frac{\xi}{1+\xi}y \tag{1.41}$$

Où ξ : est le rapport signal sur bruit a priori défini par :

$$\xi=\frac{\sigma_x^2}{\sigma_d^2} \tag{1.42}$$

Donc, l'estimateur non linéaire au sens du MMSE est identique à l'estimateur linéaire dans le cas Gaussien.

1.7.4 Estimation et détection conjointe

Dans une application pratique, le signal désiré x n'est pas toujours présent dans le signal bruité observé y. L'estimateur optimal doit être adapté à cette incertitude de la présence du signal désiré et doit fournir une estimation optimale indépendamment de la présence ou l'absence de la parole.

En général, on suppose qu'il y a deux versions du signal désiré x_0 et x_1, qui sont présentes dans le signal observé avec deux probabilités a priori $P\left(H^{(0)}\right)$ et $P\left(H^{(1)}\right)=1-P\left(H^{(0)}\right)$ respectivement, avec les deux hypothèses :

$H^{(0)}$: la présence de $x_0\,(x=x_0)$; $H^{(1)}$: la présence de $x_1\,(x=x_1)$

Les deux versions du signal x sont traitées comme des variables aléatoires avec différentes possibilités des fonctions de densité de probabilité. Avec ces suppositions, la fonction de densité de probabilité de x peut-être écrite comme suit :

$$p_x(u) = p_{x|H^{(0)}}\left(u\middle|H^{(0)}\right)p\left(H^{(0)}\right) + p_{x|H^{(1)}}\left(u\middle|H^{(1)}\right)p\left(H^{(1)}\right) \tag{1.43}$$

Comme précédemment, on utilise la fonction du coût $C(x,\hat{x}) = (x-\hat{x})^2$, où $\hat{x}$ est en général une fonction de la variable observée $y = x + d$. On minimise le coût total de Bayes.

$$\tilde{J} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C\left(u,\hat{x}(v)\right)p_{xy}(u,v)\,du\,dv$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\hat{x}(v)-u\right)^2\left(p_{xy|H^{(0)}}\left(u,v\middle|H^{(0)}\right)p\left(H^{(0)}\right) + p_{xy|H^{(1)}}\left(u,v\middle|H^{(1)}\right)p\left(H^{(1)}\right)\right)du\,dv \tag{1.44}$$

L'annulation de la dérivée première par rapport à $\hat{x}$ de l'intégrale intérieure donne :

$$\int_{-\infty}^{\infty}\left(\hat{x}(v)-u\right)\left(p_{xy|H^{(0)}}\left(u,v\middle|H^{(0)}\right)p\left(H^{(0)}\right) + p_{xy|H^{(1)}}\left(u,v\middle|H^{(1)}\right)p\left(H^{(1)}\right)\right)du = 0 \tag{1.45}$$

Et la substitution de :

$$p_{xy|H^{(0)}}\left(u,v\middle|H^{(0)}\right)p\left(H^{(0)}\right) = p_{y|H^{(0)}}\left(v\middle|H^{(0)}\right)p_{x|y,H^{(0)}}\left(u\middle|v,H^{(0)}\right)p\left(H^{(0)}\right) \tag{1.46}$$

$$p_{xy|H^{(1)}}\left(u,v\middle|H^{(1)}\right)p\left(H^{(1)}\right) = p_{y|H^{(1)}}\left(v\middle|H^{(1)}\right)p_{x|y,H^{(1)}}\left(u\middle|v,H^{(1)}\right)p\left(H^{(1)}\right) \tag{1.47}$$

Conduit à :

$$\hat{x}(v)\left[p_{y|H^{(0)}}\left(v\middle|H^{(0)}\right)p\left(H^{(0)}\right) + p_{y|H^{(1)}}\left(v\middle|H^{(1)}\right)P\left(H^{(1)}\right)\right]$$

$$= p_{y|H^{(0)}}\left(v\middle|H^{(0)}\right)p\left(H^{(0)}\right)\int_{-\infty}^{\infty} up_{x|y,H^{(0)}}\left(u\middle|v,H^{(0)}\right)du + p_{y|H^{(1)}}\left(v\middle|H^{(1)}\right)p\left(H^{(1)}\right)\int_{-\infty}^{\infty} up_{x|y,H^{(1)}}\left(u\middle|v,H^{(1)}\right)du \tag{1.48}$$

On introduit le rapport de vraisemblance généralisé :

$$\Lambda(v) = \frac{p_{y|H^{(1)}}\left(v\middle|H^{(1)}\right)p\left(H^{(1)}\right)}{p_{y|H^{(0)}}\left(v\middle|H^{(0)}\right)p\left(H^{(0)}\right)} \tag{1.49}$$

On obtient la solution [47] :

$$\hat{x}(v) = E_x\left\{x\middle|v,H^{(0)}\right\}\frac{1}{1+\Lambda(v)} + E_x\left\{x\middle|v,H^{(1)}\right\}\frac{\Lambda(v)}{1+\Lambda(v)} \tag{1.50}$$

Le problème d'estimation et de détection conjointe au sens de l'EQM conduit à une combinaison linéaire des deux estimateurs au sens de l'EQM pour les deux hypothèses $H^{(0)}$ et $H^{(1)}$.

1.8 Mesures de qualité

Afin de comparer entre les différentes méthodes de rehaussement de la parole qui seront développées dans les prochains chapitres, on utilise des tests d'écoute, des représentations temporelles et fréquentielles, et en plus on doit disposer de mesures fiables et variées.

La qualité vocale est un phénomène multidimensionnel. En effet, elle peut être évaluée selon différents critères de qualité. Les deux principaux critères sont la sonie et l'intelligibilité (i.e. le niveau et la compréhensibilité du signal de parole, respectivement), qui permettent à l'auditeur d'entendre et de comprendre le message du locuteur.

De façon générale, la qualité dépend de la personne qui la juge. La qualité vocale est donc une notion complexe à définir du fait de sa forte subjectivité, elle peut être aussi évaluée objectivement par une série de calcul appliquée aux signaux sonores testés [48].

1.8.1 Evaluation subjective de la qualité vocale

Le jugement de la qualité vocale est avant tout subjectif. La meilleure façon d'évaluer la qualité vocale est donc de faire appel à des utilisateurs et de les interroger, sous forme de sondages ou de tests en laboratoire. Ce sont les méthodes d'évaluation subjective de la qualité.

Lors d'un test subjectif, on demande à des participants de tester un système de rehaussement dans différentes conditions et de noter sur une échelle de qualité, la qualité vocale de ce système. Les notes des participants pour une condition de test donnée sont moyennées pour obtenir une note moyenne d'opinion dénommée « note MOS » (Mean Opinion Score). Le fait de moyenner les notes individuelles permet de diminuer l'effet subjectif sur l'évaluation de la qualité vocale.
La notation s'effectue selon l'une des méthodes définies par l'Union Internationale des Télécommunications (UIT). La plus utilisée est la méthode d'évaluation par catégories absolues (Absolute Category Rating, ACR) avec les catégories : 5 = Excellente, 4 = Bonne, 3 = Passable, 2 = Médiocre, 1 = Mauvaise.

D'autres mesures subjectives orientées vers l'intelligibilité des signaux de mauvaise qualité peuvent être utilisées comme le DRT (Diagnostic Rhyme Test) et le MRT (Modified Rhyme Test), différents mots sont présentés et on doit choisir le plus proche au son écouté.

Les tests subjectifs sont indispensables pour l'évaluation de la qualité vocale, puisqu'ils représentent le jugement humain de la qualité vocale. Cependant, les tests subjectifs nécessitent de mobiliser beaucoup de moyens (temps, personnes et argent). Ces tests sont difficiles à mettre en œuvre et par conséquent ils ne seront pas implémentés dans le cadre de ce travail.

Les méthodes objectives se présentent comme une alternative aux méthodes subjectives et permettent d'automatiser l'évaluation de la qualité vocale. Néanmoins, elles doivent présenter une forte corrélation avec les résultats des tests subjectifs, qui représentent le jugement des utilisateurs [49].

1.8.2 Evaluation objective de la qualité vocale

L'évaluation objective de la qualité vocale s'est tout d'abord effectuée avec des outils simples de traitement du signal tels que le rapport signal sur bruit (SNR), le rapport signal sur bruit segmental (SNR_{seg}), l'erreur quadratique moyenne (EQM) pour les mesures temporelles, la distance cepstrale, et la distance spectrale pour les mesures fréquentielles.

Les mesures objectives les plus communément utilisées sont [49][50] :

➢ **Mesure de distorsion d'Itakura-Saito (IS)**

Pour une trame de la parole propre avec un vecteur des coefficients de la prédiction linéaire, $\vec{a}_\Phi$, et un vecteur des coefficients de la parole rehaussée, $\vec{a}_d$, la mesure de distorsion d'Itakura-Saito est donnée par :

$$d_{IS}\left(\vec{a}_d,\vec{a}_\Phi\right)=\left[\frac{\sigma_\Phi^2}{\sigma_d^2}\right]\cdot\left[\frac{\vec{a}_d R_\Phi \vec{a}_d^T}{\vec{a}_\Phi R_\Phi \vec{a}_\Phi^T}\right]+\log\left(\frac{\sigma_d^2}{\sigma_\Phi^2}\right)-1 \qquad (1.51)$$

Où : σ_d^2 et σ_Φ^2 représentent les gains de la prédiction linéaire pour le signal rehaussé et le signal propre respectivement, et R_Φ la fonction d'autocorrélation.

➢ **Mesure LLR (Log-Likelihood Ratio)**

La mesure du rapport de la log-vraisemblance (LLR) est presque identique à la mesure d'Itakura-Saito (IS), elle s'intéresse seulement à la différence entre la forme spectrale des deux modèles LPC des signaux propre et rehaussé, elle est donnée par :

$$d_{LLR}\left(\vec{a}_d,\vec{a}_\Phi\right)=\log\left(\frac{\vec{a}_d R_\Phi \vec{a}_d^T}{\vec{a}_\Phi R_\Phi \vec{a}_\Phi^T}\right) \qquad (1.52)$$

➢ **Mesure du rapport signal sur bruit segmental**

Comme le rapport signal sur bruit global a une faible corrélation avec les tests subjectifs de la qualité, le rapport signal sur bruit segmental est largement utilisé comme mesure objective de la qualité. Il présente une forte corrélation avec les tests subjectifs et il est obtenu en moyennant les rapports signal sur bruit obtenus dans chaque trame d'analyse.

$$d_{AvgSegSNR}=\frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Nm}^{Nm+N-1}s_\Phi^2(n)}{\sum_{n=Nm}^{Nm+n-1}\left[s_d(n)-s_\Phi(n)\right]^2} \qquad (1.53)$$

Où chaque segment "m" est de longueur N. Les trames qui ont un SNR qui dépasse 35 dB peuvent affecter la corrélation du SNR segmental avec les tests subjectifs, généralement une limite supérieure à 35 dB est utilisée. Egalement, durant les périodes de silence, les valeurs du SNR peuvent atteindre des valeurs négatives très faibles, ces trames affectent le degré d'exactitude de la qualité perceptuelle mesurée objectivement dans ce cas. Alors, une limite inférieure est à considérer (on a utilisé -10 dB, mais la gamme de [-20dB, 0dB] peut être utilisée).

➢ **Mesure WSS (Weighted Spectral Slope)**

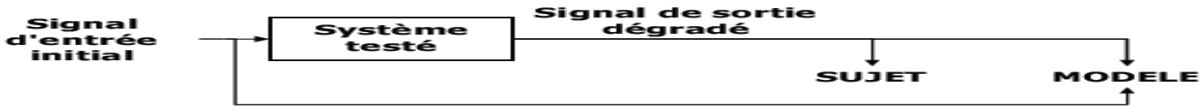
La mesure WSS ou la pente spectrale pondérée est basée sur le modèle auditif. 36 filtres qui se chevauchent sont utilisés pour l'estimation du spectre court-terme lissé de la parole. Cette mesure calcule une différence pondérée entre les pentes spectrales dans chaque bande. La valeur de chaque pondération indique si la bande est près d'un pic spectral ou une vallée, et si le pic est le plus grand dans le spectre. Pour chaque trame la valeur du WSS en décibels est :

$$d_{WSS}(j)=K_{spl}\left(K-\hat{K}\right)+\sum_{\omega=1}^{36}w_a(\omega)\left(S(\omega)-\hat{S}(\omega)\right)^2 \qquad (1.54)$$

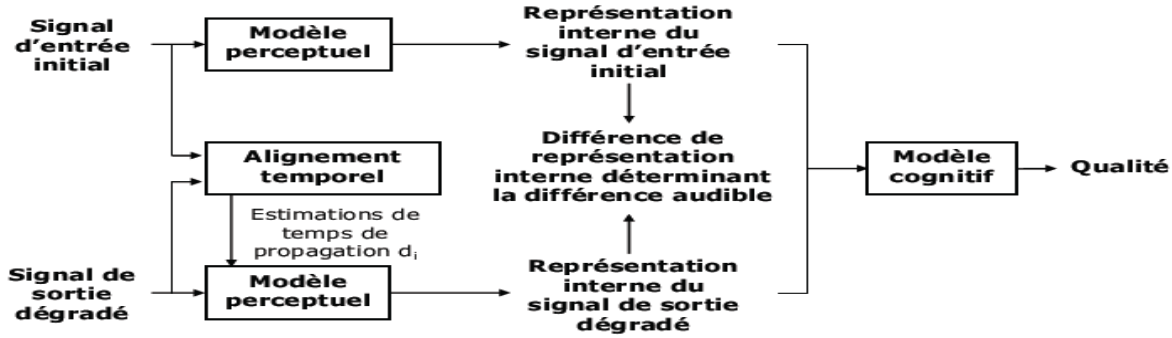
Où : K et $\hat{K}$ représentent les niveaux de pression totale du son original et rehaussé, et K_{spl} est un paramètre qui peut être varié afin d'augmenter la performance totale.

- **Mesure PESQ (Perceptual Evaluation of Speech Quality)**

L'évaluation perceptuelle de qualité de la parole est un indice défini par l'Union Internationale des Télécommunications (ITU) pour évaluer la qualité perçue de la voix (ITU-T P.862/P.862.1) [51]. Il a une corrélation élevée avec les tests subjectifs. Son principe de fonctionnement est présenté dans la figure (1.7). Lors d'un test subjectif d'écoute, le sujet juge le signal de sortie dégradé par le système à tester, sans avoir accès au signal de référence correspondant à la phrase prononcée. La comparaison s'effectue avec les conditions de référence (i.e. sans dégradation) présentées durant le test, ainsi qu'avec la référence propre à chaque sujet. Le modèle PESQ a accès pour chaque condition testée aux signaux de référence et dégradé, qu'il transforme et compare pour obtenir une note de qualité d'écoute. Cette note PESQ est appliquée à une échelle de type MOS, sous forme d'un scalaire compris entre 1 et 4.5 [51][52].



(a) Le sujet juge le signal dégradé, en comparaison avec sa référence de qualité interne



(b) Le modèle PESQ transforme le signal dégradé et le signal de référence correspondant en représentations internes, puis utilise leur différence pour calculer une note de qualité d'écoute

Figure 1.7 : Principe de fonctionnement du modèle PESQ [51].

1.9 Conclusion

Dans ce chapitre, nous avons abordé des généralités sur le signal utile dans notre travail, qui est le signal de la parole et sur le signal indésirable qui est le bruit. Ce signal biologique est un signal réel, continue, d'énergie finie et localement stationnaire sur de courte durée. Sa structure est complexe et variable en fonction du temps, ce qui exige un traitement temporel ou fréquentiel plus adapté et une modélisation adéquate. Dans un système de communication ou de reconnaissance de la parole, le signal utile est affecté par du bruit du milieu hostile où se trouve le locuteur. Nous nous intéresserons aux bruits additifs, blancs ou colorés, stationnaires ou non stationnaires de bande de fréquence large ou étroite, car c'est les types de bruits rencontrés dans le codage à bas débit dans les environnements bruités. Pour nos simulations, les fichiers de parole bruités par ces bruits à des rapports signal sur bruit différents sont extraits de la base de données Noizeus.

Nous avons aussi présenté les propriétés statistiques des coefficients de la transformée de Fourier discrète, qui seront la base des approches Bayesiennes de rehaussement de la parole dans les chapitres suivants. Enfin, dans toutes les simulations et les résultats qui seront présentés tout au long de cette thèse, en plus des tests d'écoute durant les manipulations et les représentations temporelles, fréquentielles et les spectrogrammes, nous utiliserons des mesures objectives de la parole rehaussée comme le LLR, WSS, SNRseg et le PESQ, pour comparer entre les différents algorithmes implémentés.

CHAPITRE 2

PRINCIPE DU REHAUSSEMENT DE LA PAROLE

2.1 Introduction

Le débruitage de la parole en vue de l'amélioration de l'intelligibilité et de la qualité est un domaine de recherche très actif et présent dans de nombreux champs d'applications. Les méthodes classiques sont largement utilisées et se basent sur la soustraction spectrale. Ces méthodes parviennent efficacement à réduire le bruit additif. En contre partie, elles produisent un bruit résiduel gênant à la perception humaine connu sous le nom de bruit musical.

Dans le cadre de ce chapitre, l'intérêt du débruitage et la classification des systèmes de rehaussement seront présentés au début, suivi par une description détaillée des systèmes de rehaussement monovoie. Nous exposerons aussi la méthode de Berouti et le filtre de Wiener qui sont les deux méthodes de base des méthodes STSA basées sur la soustraction spectrale et sur les méthodes Bayesiennes respectivement. Enfin, des notions générales sur l'estimation de la densité spectrale du bruit et du rapport signal sur bruit a priori seront introduites.

2.2 Intérêt du débruitage

Le rehaussement signifie l'amélioration de la valeur ou de la qualité de quelque chose. Une fois appliqué à la parole, ceci signifie simplement l'amélioration de l'intelligibilité et/ou de la qualité d'un signal dégradé en utilisant des outils de débruitage des signaux.

Les algorithmes de réduction du bruit tentent d'améliorer deux caractéristiques propres au signal de parole bruité : la première, qualitative, est une mesure subjective qui montre jusqu'à quel niveau le son du signal débruité plaît aux auditeurs. L'amélioration de la qualité sonore n'est pas seulement du plaisir esthétique, mais de la facilité de compréhension afin de minimiser la fatigue pour l'oreille humaine. La deuxième caractéristique, intelligibilité, est une mesure objective qui montre la quantité d'information que nous sommes capables d'extraire du signal de parole bruité sans se soucier de sa qualité.

Les techniques de rehaussement de la parole sont généralement exploitées dans les domaines suivants :

- Reconnaissance de la parole.
- Médical, comme pour les prothèses auditives des malentendants.
- Téléphonie mobile.
- Téléphonie main libre.
- Les systèmes de prétraitement dans le codage de la parole dans un milieu hostile.

2.3 Classification des systèmes de rehaussement

Un système de rehaussement de la parole est généralement basé sur un certain nombre de suppositions et contraintes liées à l'application et à l'environnement. Dans la littérature les systèmes de rehaussement peuvent être classés selon le tableau (2.1) [14][44].

Domaines	**Stratégies possibles**
Nombre de canaux d'entrée	un / deux / plusieurs
Domaine de traitement	temps / fréquence
Type d'algorithme	adaptatif / non adaptatif

Tableau 2.1 : Classification des techniques de rehaussement.

Typiquement, la classification la plus utilisée est celle qui divise ces techniques en fonction du nombre des canaux d'entrée en : techniques monovoie (un seul microphone) et techniques multivoies (plusieurs microphones).

2.3.1 Techniques monovoie

S'appliquent aux situations dans lesquelles seulement un microphone d'acquisition est disponible pour la parole et le bruit (figure 2.1). Ces systèmes sont simples à mettre en œuvre et comparativement moins cher que les systèmes multivoies.

Le problème major de ces méthodes est la non disponibilité d'une référence pour le bruit. De plus, les performances sont limitées dans le cas des bruits non stationnaires ou dans le cas d'un RSB très faible [44].

Dans cette étude et pour remplir les exigences décrites dans le cahier des charges, nous avons choisi de se limiter au cas monovoie car c'est le contexte le plus courant en traitement de la parole, pour le codage ou la reconnaissance de la parole dans un milieu bruité.

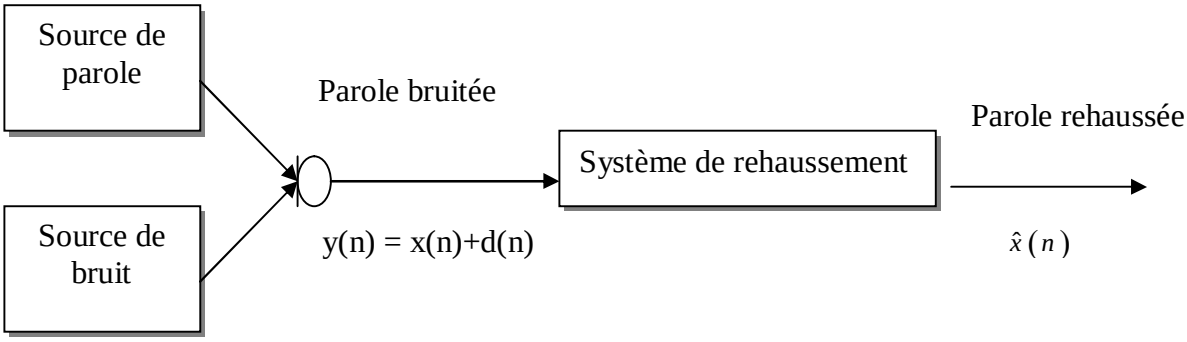


Figure 2.1 : Système monovoie de réduction du bruit.

2.3.2 Techniques multivoies

Ces méthodes à plusieurs microphones ont l'avantage de l'existence de plusieurs signaux à l'entrée du système de rehaussement, en plus d'un signal de référence pour le bruit. Bien que ces systèmes soient plus complexes que les systèmes monovoie, plusieurs limitations de ces derniers peuvent être surmontées en particulier dans le cas des bruits non stationnaires.

Les techniques multivoies sont largement utilisées dans l'annulation adaptative du bruit, mais rarement utilisées dans les blocs de réduction de bruit des codeurs de la parole [53].

2.4 Systèmes monovoie

Un système de réduction du bruit monovoie est constitué de deux composantes principales, le bloc d'estimation du spectre de bruit et celui de l'estimation du spectre de la parole, en plus des autres composantes : analyse, synthèse et modification comme illustré dans la figure (2.2).

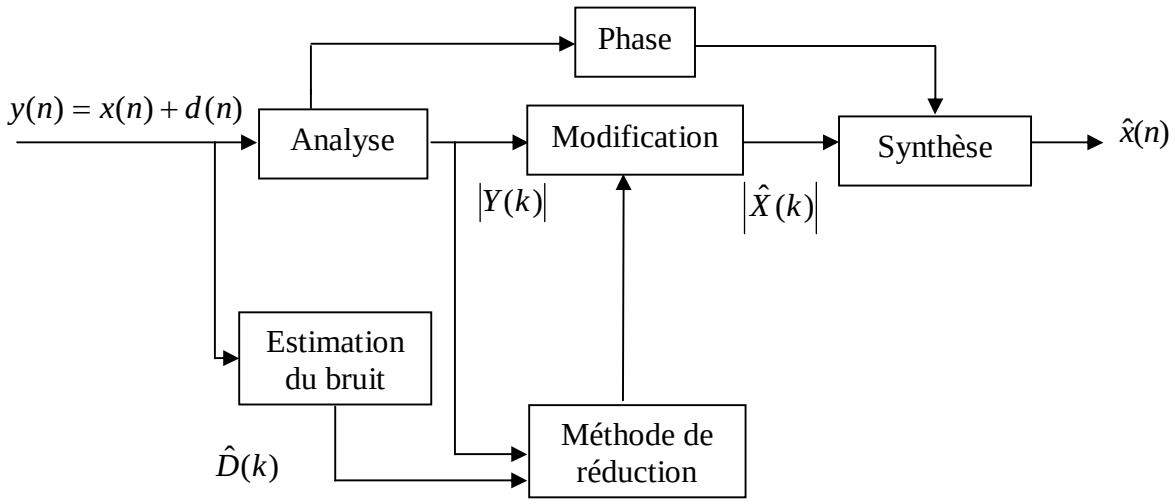


Figure 2.2 : Schéma général d'un système de réduction de bruit monovoie**.**

Chaque technique de rehaussement monovoie exploite une ou plusieurs propriétés du signal de parole pour séparer le bruit. Il n'existe aucune classification standard pour les méthodes disponibles, la plupart d'entre elles relèveront d'une des catégories présentées ci-dessous :

- Les méthodes basées sur la périodicité de la parole,
- Les méthodes basées sur un modèle de la parole,

- Les méthodes basées sur l'estimation de l'amplitude spectrale à court terme (STSA : Short Time Spectral Amplitude),
- Les méthodes basées sur les critères perceptuelles.

Les méthodes basées sur la périodicité de la parole se servent du fait, que la forme d'onde d'un son voisé présente une forme quasi-périodique avec une période qui corresponde à la fréquence fondamentale [54][55]. L'inconvénient principal de ces techniques est dans le fait qu'elles ne peuvent rehausser que les segments voisés de la parole. Ces méthodes ne sont pas applicables aux segments de parole non voisés, les transitions et les fricatifs. Si une méthode de traitement séparée est utilisée pour rehausser les segments non voisés, une détection voisé / non voisé efficace est indispensable. Les performances de cette détection se dégradent avec la présence du bruit, ainsi que l'estimation exacte de la période du pitch.

Les méthodes basées sur la modélisation de la parole sont utilisées dans le cas où on n'a aucune connaissance des propriétés statistiques de la parole ou du bruit. A la place de l'observation bruitée, les modèles de production de parole sont utilisés comme le modèle ARMA, AR ou MA. Pour estimer les paramètres du modèle de la parole, les trois règles d'estimation : le maximum de vraisemblance (ML : Maximum Likelihood), le maximum a posteriori (MAP) et l'estimation basée sur l'erreur quadratique moyenne minimale (MMSE) sont largement utilisées [56]. L'estimateur ML est utilisé souvent pour les paramètres déterministes. Les règles d'estimation MAP et MMSE sont généralement utilisées pour les paramètres qui peuvent être considérés comme des variables aléatoires avec des densités de probabilité a priori connues.

Une approche de cette catégorie a été proposée dans [57], où un modèle AR variant dans le temps est utilisé pour le signal de parole. Tous les deux, le modèle et le signal sont estimés à partir du signal bruité en utilisant l'approche d'estimation du maximum a posteriori. De nombreuses variantes basées sur cette approche ont été proposées par la suite dans [58]-[63]. Ces méthodes seront l'objet du troisième chapitre.

Les méthodes basées sur l'estimation de l'amplitude spectrale à court –terme, appelées aussi méthodes STSA ou subtractive - type algorithmes sont les plus utilisées dans les blocs de réduction du bruit des codeurs de la parole. Plusieurs techniques STSA existent, les plus couramment employées sont des généralisations des techniques comme le filtrage de Wiener à court-terme et les méthodes dites de soustraction spectrale. La soustraction spectrale est la méthode de base des techniques soustractives de réduction de bruit, qui sont des améliorations de cette approche. Les autres méthodes de débruitage sont basées sur des approches Bayesiennes où le filtre de Wiener est la méthode de base. Le quatrième chapitre sera consacré à l'étude et l'implémentation de ces techniques.

Les méthodes basées sur les critères perceptuelles exploitent les propriétés de masquage du système auditif humain [41][42]. La perception d'un signal audio ou de la parole est le résultat de plusieurs effets physiologiques et psychologiques, qui ne sont pas encore pleinement comprises. Encouragés par le succès des modèles perceptuels dans les applications du codage audio [64], les chercheurs ont tenté de résoudre le problème de rehaussement de la parole d'une manière similaire. Ces méthodes vont de l'intégration des effets de masquage auditif dans les règles de suppression du bruit à des systèmes de réduction de bruit mises en œuvre entièrement dans le domaine perceptuel. La plupart de ces techniques visent à surmonter le compromis classique entre la réduction du bruit et la distorsion de la parole rehaussée [14], où le bruit audible est masqué au lieu d'être supprimé davantage, réduisant ainsi les chances d'une distorsion supplémentaire de la parole.

2.5 Méthodes STSA basées sur la soustraction spectrale

Plusieurs méthodes de débruitage de la parole basées sur la soustraction spectrale ont été présentées dans la littérature, comme la soustraction spectrale d'amplitude [11], la soustraction spectrale de puissance [12], la soustraction spectrale non linéaire [13], la soustraction spectrale paramétrique [15] et la soustraction spectrale multi-bandes [16]. Seulement une vue générale du principe de la soustraction spectrale de puissance sera présentée dans la section suivante.

2.5.1 Définition

C'est une méthode largement implémentée dans les systèmes de codage et de reconnaissance de la parole dans un milieu hostile. Elle permet de réduire l'influence du bruit avant l'étape de paramétrisation. La soustraction spectrale propose de calculer une estimée du bruit sur des portions du signal ne contenant pas de parole [65][11][12]. Sous l'hypothèse que le bruit soit stationnaire, l'estimée du bruit est soustraite du spectre de puissance du signal bruité. L'efficacité de cette méthode dépend de la qualité de l'estimation de la densité spectrale de puissance du bruit. Elle est devenue une référence aussi bien pour le rehaussement de la parole que pour la restauration d'enregistrements anciens.

2.5.2 Principe de base de la soustraction spectrale

Afin d'aborder le principe de fonctionnement général de la soustraction spectrale à court-terme, considérons un signal d'observation bruité, y, composé d'un signal de parole, x, corrompu par un bruit additif, d. Pour chaque indice temporel, n, le signal d'observation bruité, est donné par :

$$y(n) = x(n) + d(n) \tag{2.1}$$

Dans le domaine spectral, le spectre de puissance du signal bruité $\left(|Y(k)|^2\right)$ est égal à la somme du spectre de puissance de la parole propre $\left(|X(k)|^2\right)$ et celui du bruit $\left(|D(k)|^2\right)$.

$$|\mathrm{Y}(k)|^2 = |X(k)|^2 + |D(k)|^2 \tag{2.2}$$

L'estimation directe du bruit est impossible d'où vient l'importance de l'utilisation de l'opérateur d'espérance E[.], qui donne une approximation de l'estimation du bruit $\mathrm{E}\left[|D(k)|^2\right]$, il est aussi noté $\left|\hat{D}(k)\right|^2$.

L'estimation du signal propre $\left|\hat{X}(k)\right|^2$ est liée à l'estimation du spectre du bruit par la relation suivante :

$$\left|\hat{X}(k)\right|^2 = |\mathrm{Y}(k)|^2 - \left|\hat{D}(k)\right|^2 \tag{2.3}$$

L'idée de base de cette méthode consiste à calculer le spectre de puissance de chaque fenêtre du signal bruité et de lui soustraire une estimation du spectre de puissance du bruit.

Une estimation initiale du spectre du bruit se fait pendant les premières trames de silence, avant qu'un locuteur ne commence à parler, ces périodes de l'inactivité vocale nous permettent d'avoir une estimation initiale du bruit dans cet environnement. Par la suite, les techniques d'estimation du bruit permettent de calculer une mise à jour du spectre de bruit.

La phase initiale est conservée pendant le traitement du signal [66]. La surestimation du bruit provoque le problème d'apparition des valeurs négatives pour l'estimation du spectre de puissance du signal propre. Deux rectifications ont été utilisées :

- La rectification complète de la trame (full wave rectification), par la conversion des valeurs négatives en valeurs positives.
- La rectification demi-trame (half wave rectification), par la mise à zéro des valeurs négatives.

Les meilleurs résultats sont assurés par la rectification demi trame, comme suit [65]:

$$\left|\hat{X}(k)\right|^2 = \begin{cases} \left|Y(k)\right|^2 - \left|\hat{D}(k)\right|^2 & \text{si} \quad \left|\mathrm{Y(k)}\right|^2 > \left|\hat{D}(k)\right|^2 \\ 0 & \text{ailleurs} \end{cases} \tag{2.4}$$

Durant l'application de cette méthode, on doit prendre en considération un nouveau bruit apparaissant connu sous le nom de bruit musical. Afin de diminuer l'influence de ce bruit et avoir une bonne qualité du signal à la sortie, Berouti et al [12] ont proposé des modifications à l'algorithme de base.

2.5.3 Soustraction spectrale de Berouti et al

Berouti et al [12] ont révolutionné la méthode de la soustraction spectrale de puissance en apportant des modifications à l'algorithme de base. Elle est devenue plus efficace et diffère des autres méthodes par la soustraction d'une estimation du spectre du bruit pondérée par un facteur α , et la limitation imposée aux composantes spectrales du signal d'aller au-dessous d'une certaine limite. Elle est donnée comme suit :

$$\left|\hat{X}(k)\right|^2 = \begin{cases} \left|Y(k)\right|^2 - \alpha\left|\hat{D}(k)\right|^2 & \text{si} \quad \left|\hat{\mathrm{X}}(\mathrm{k})\right|^2 > \beta\left|\hat{D}(k)\right|^2 \\ \beta\left|\hat{D}(k)\right|^2 & \text{ailleurs} \end{cases} \tag{2.5}$$

Où : α est un facteur de soustraction (surestimation), (α >1). Et β : un paramètre de lissage spectral, (0< β <<1). La méthode de Berouti et al est représentée par le diagramme suivant:

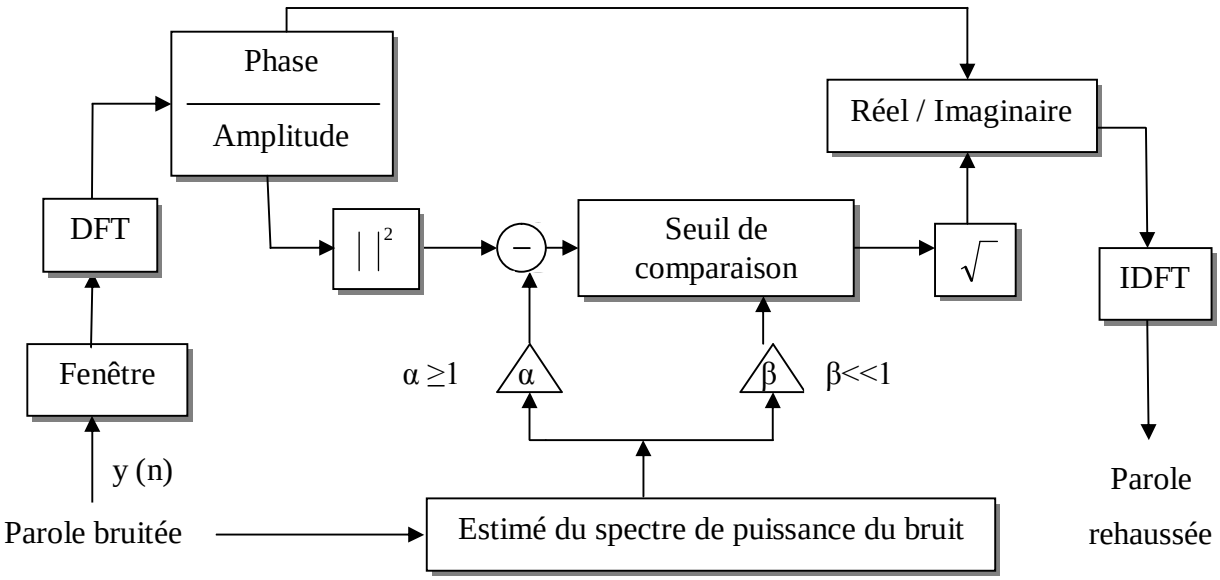


Figure 2.3 : Diagramme de la soustraction spectrale proposée par Berouti et al [12]**.**

2.5.4 Influence des paramètres

Le signal de la parole rehaussée, qui résulte de l'application de cette méthode est affecté par deux types de bruits : bruit à large bande, connu sous le nom de bruit résiduel et bruit à bande étroite, connu sous le nom de bruit musical. Ces deux bruits apparaissent sous forme de pics et de vallées dans le spectre du signal rehaussé. Ils ont une distribution aléatoire et changent aléatoirement en fréquence et en amplitude d'une trame à l'autre.

La réduction des pics spectraux du bruit résiduel est effectuée par le facteur de soustraction qui prend toujours une valeur supérieure à l'unité ($\alpha > 1$). Si une valeur élevée de α est prise, la réduction du bruit à large bande est assurée, mais cela provoque une distorsion du signal de la parole. La valeur du facteur de soustraction α est fonction du rapport signal sur bruit segmental estimé. Pour chaque trame, la valeur de α est donnée par [12] :

$$\alpha = \begin{cases} 5 & \text{SNRseg} < \text{-5dB} \\ \alpha_0 - (SNRseg/s) & \text{-5dB} < \text{SNRseg} < \text{20dB} \\ 1 & \text{SNRseg} > \text{20dB} \end{cases} \qquad (2.6)$$

Avec :

- α_0 : la valeur de α pour un SNRseg = 0. En pratique, elle varie entre 3 et 6.
- 1/s : la pente de la droite dans la figure (2.4)
- SNRseg : le rapport signal sur bruit segmental estimé.

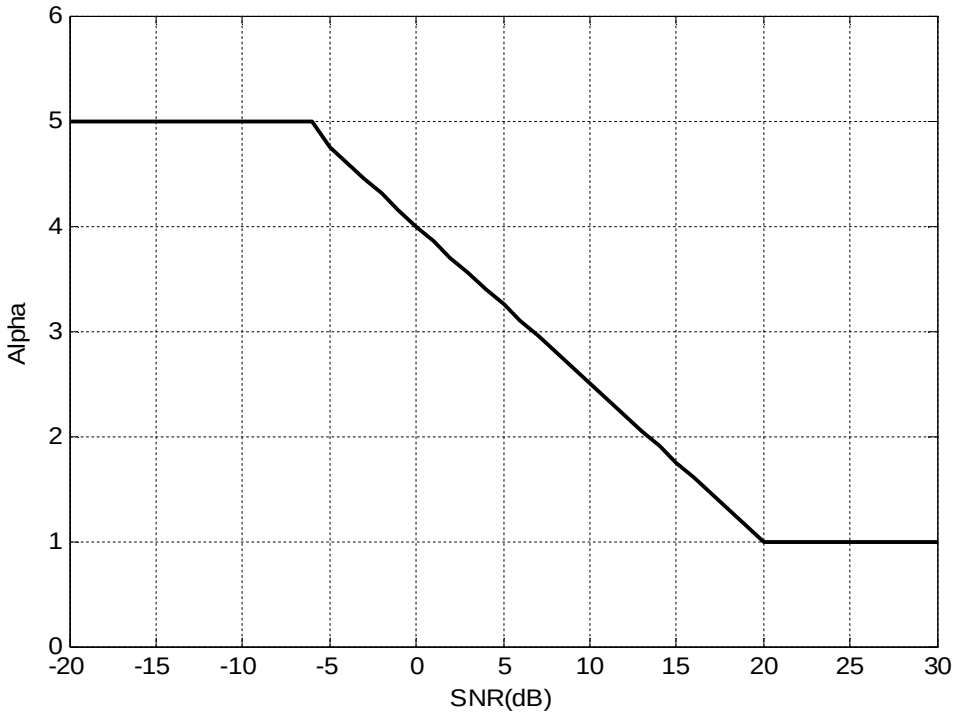


Figure 2.4 : Valeurs de α en fonction du SNRseg**.**

Outre la réduction des pics, il y a le problème du remplissage des vallées d'où la réduction du bruit musical. Cela est assuré par le facteur de lissage β qui prend des valeurs dans l'intervalle $0 < \beta << 1$.

- $\beta > 0$: les pics du bruit résiduel sont masqués par les composantes spectrales voisines.

- $\beta << 1$: le bruit à large bande est plus bas par rapport à celui obtenu dans le cas où $\beta = 0$.

Donc le choix de β a une importance majeure pour le spectre de puissance du signal propre estimé $\left|\hat{X}(k)\right|^2$, Il est constaté aussi que :

- Si β est faible : le bruit résiduel sera réduit, mais le bruit musical sera audible.
- Si β est grande : le bruit musical n'est pas audible mais le bruit résiduel reste présent.

L'inconvénient majeur de cette technique est l'apparition d'un bruit de fond résiduel ayant un caractère musical. Ce caractère est dû à l'apparition des pics, appelés aussi tonales, dans le spectre du signal débruité. Toutefois, malgré cet inconvénient du bruit musical, la méthode de soustraction spectrale reste performante en termes d'atténuation du bruit.

2.5.5 Phénomène du bruit musical

Ce phénomène est caractéristique des méthodes d'atténuation spectrale à court-terme. Du fait du caractère tonal particulier de cet artéfact, il est désigné par le terme de bruit musical. Son spectre à court-terme correspond approximativement à une distribution aléatoire des pics spectraux. L'origine du bruit musical est la variance des estimateurs locaux de la densité spectrale des signaux. En effet, comme le spectre à court-terme du bruit fluctue autour des valeurs moyennes, son amplitude atteint à certains instants et pour certains indices fréquentiels des valeurs largement supérieures à la moyenne.

Considérons le cas où le signal est fortement bruité dans une certaine zone du spectre où le rapport signal à bruit est par conséquent, relativement bas. Sporadiquement, du fait de la variance du bruit, l'amplitude spectrale du signal atteint des valeurs largement supérieures au niveau moyen estimé du bruit conduisant donc à une surestimation locale et instantanée du rapport signal sur bruit, la bande de fréquences correspondante est alors traitée comme du signal utile et est relativement moins atténuée que les composantes fréquentielles voisines. De manière sporadique, des pics fréquentiels isolés se dégagent donc du spectre atténué, engendrant ainsi le phénomène du bruit musical**.** La figure (2.5) montre le phénomène du bruit musical dans le spectre du signal de parole rehaussé. La nature tonale du bruit musical est très gênante du point de vue auditif. En effet, un tel bruit pourrait être considéré comme un signal utile par certains processus et être amplifié au détriment des composantes issues du signal source. Il est donc rigoureusement nécessaire de prévenir l'apparition du bruit musical.

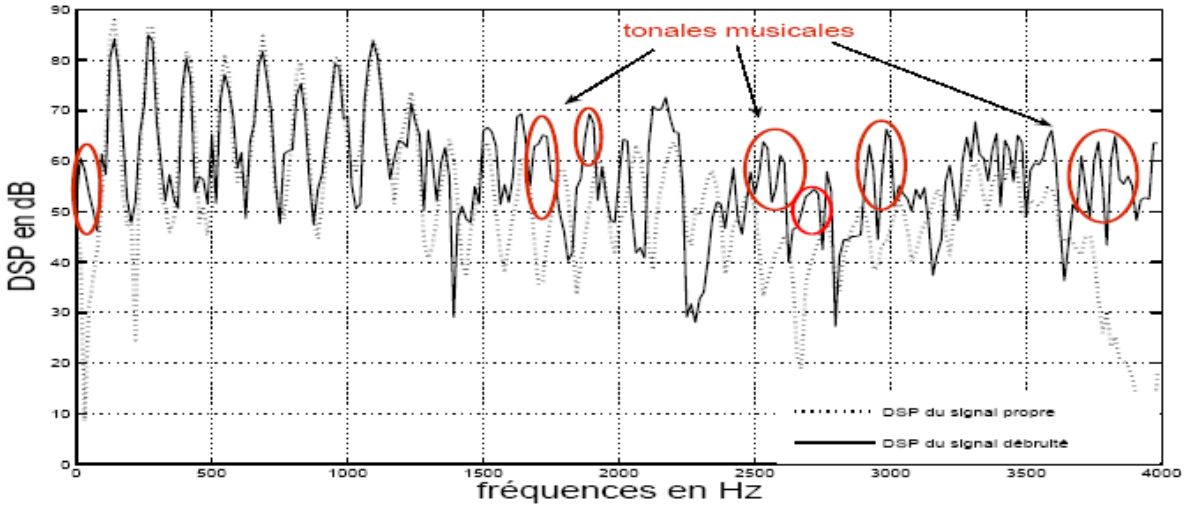


Figure 2.5 : DSP du signal propre et du signal débruité**.**

2.5.6 Limitation des méthodes basées sur la soustraction spectrale

Les méthodes de réduction de bruit basées sur la soustraction spectrale parviennent à réduire de manière très efficace le niveau du bruit de fond, la réduction du bruit résiduel, mais l'inconvénient principal de ces méthodes réside dans le bruit musical. Ce bruit est très gênant du point de vue perceptif.

Ainsi, d'autres méthodes ont été proposées afin de mieux réduire ce bruit musical tout en conservant la simplicité d'implémentation et l'utilisation en temps réel de ces méthodes. Les approches Bayesiennes sont les plus utilisées pour combattre le phénomène du bruit musical dans les codeurs de la parole et les systèmes de reconnaissance de la parole dans un milieu hostile. Ces méthodes seront largement étudiées et appliquées dans le cadre de notre travail de recherche.

2.6 Méthodes STSA basées sur les méthodes Bayesiennes

Les méthodes Bayesiennes ont été introduites par Ephraim et Malah [20][21]. Elles sont basées sur l'hypothèse que les coefficients de la TFCT des signaux de la parole et du bruit sont des variables aléatoires Gaussiennes et indépendantes (comme décrit dans le premier chapitre). Plusieurs variantes ont été présentées par la suite dans la littérature en se basant sur d'autres modèles probabilistes, qui modélisent mieux les coefficients de la TFCT sur de courtes durées. Les densités de Laplace, gamma, gamma généralisée sont les plus adaptées [23]-[34].

Comme dans le cas des méthodes de la soustraction spectrale où la méthode de la soustraction spectrale de puissance est la méthode de référence, les approches Bayesiennes sont comparées par rapport au filtre de Wiener qui est la méthode de référence dans ce cas.

2.6.1 Filtrage de Wiener

Le filtre de Wiener est conçu pour minimiser l'erreur quadratique moyenne entre la sortie et une sortie désirée [17][18]. Il est dit optimal au sens du critère de l'erreur quadratique moyenne et nous verrons que dans ce cas, que la réponse du filtre est liée à la fonction d'autocorrélation du signal d'entrée et à l'intercorrélation entre les signaux d'entrée et de sortie désirée. Le filtrage de Wiener est adéquat pour les situations dans lesquelles le signal et le bruit sont stationnaires, comme dans notre application**.**

- **Problème du filtrage de Wiener**

On souhaite, à partir d'un message $y(n)$, contenant un signal utile (signal désiré) $x(n)$ et un bruit qui sont deux processus aléatoires stationnaires, à déterminer le meilleur filtre (optimal).

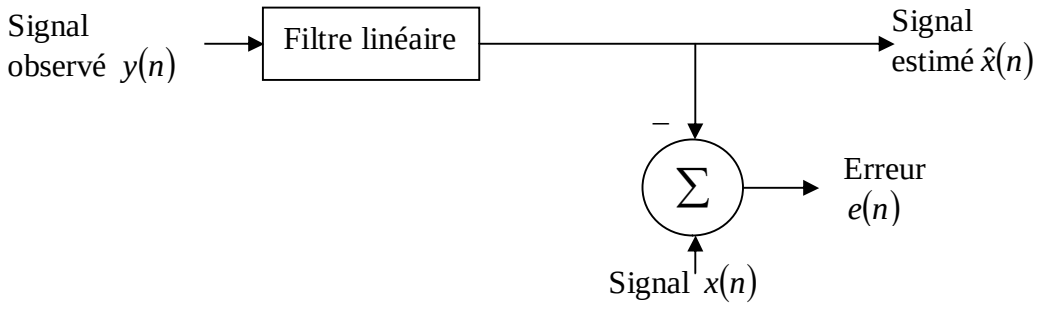


Figure 2.6 : Schéma général du filtrage de Wiener [67].

Le problème qui se pose est comment retrouver $x(n)$ à partir de $y(n)$. Une solution consiste à filtrer $y(n)$ de telle sorte que la sortie $\hat{x}(n)$ soit la plus proche possible de $x(n)$. On peut mesurer la qualité de l'estimation par $e(n)$ définie par :

$$e(n) = x(n) - \hat{x}(n) \tag{2.7}$$

Evidemment, plus $e(n)$ sera faible, plus l'estimation sera bonne. On cherche donc un filtre qui minimisera l'erreur. Il est pratique de chercher à minimiser $e^2(n)$ car c'est une fonction quadratique facilement dérivable. Par ailleurs, étant donné que les signaux intéressants sont aléatoires, la fonction qui sera minimisée est l'erreur quadratique moyenne (EQM en anglais MSE : Mean Square Error) définie par :

$$J = E(e^2(n)) \tag{2.8}$$

Donc, le filtre optimal de Wiener correspond à un filtre qui minimisera l'EQM.

Appelons H, le filtre que nous recherchons et N la longueur de sa réponse impulsionnelle donnée avec une notation matricielle par :

$$\mathbf{h} = \begin{bmatrix} h_0 & h_1 & \dots & h_{N-1} \end{bmatrix}^T \tag{2.9}$$

Le signal estimé $\hat{x}(n)$ peut alors s'écrire :

$$\hat{x}(n) = \sum_{i=0}^{N-1} h_i y(n-i) \tag{2.10}$$

Où encore en introduisant la notation matricielle pour

$$\hat{x}(n) = h^T \mathrm{y}(n) \Leftrightarrow \hat{x}(n) = \mathrm{y}^T(n) h \tag{2.11}$$

Avec :

$$\mathbf{y}(n) = \begin{bmatrix} y(n) & y(n-1) & \dots & y(n-(N-1)) \end{bmatrix}^T \tag{2.12}$$

En faisant l'hypothèse que les signaux $\mathbf{x}(n)$ et $\mathbf{y}(n)$ sont stationnaires, et si on introduit les équations (2.7) et (2.11) dans l'équation (2.8), on arrive à la fonction suivante :

$$\begin{aligned} J = E\left[\left(x(n) - \mathbf{h}^T \mathbf{y}(n) \right)^2 \right] &= E\left[x^2(n) - 2\mathbf{h}^T \mathbf{y}(n) x(n) + \mathbf{h}^T \mathbf{y}(n) \mathbf{y}^T(n) \mathbf{h} \right] \\ &= E\left[x^2(n) \right] - 2\mathbf{h}^T \mathbf{r}_{yx} + \mathbf{h}^T \mathbf{R}_{yy} \mathbf{h} \end{aligned} \tag{2.13}$$

Où $\mathbf{R}_{yy}$ est la fonction d'autocorrélation de y définie par :

$$\mathbf{R}_{yy} = E\left[\mathbf{y}(n) \mathbf{y}^T(n) \right] \tag{2.14}$$

Et où $\mathbf{r}_{yx}$ est la fonction d'intercorrélation des signaux x et y définie par :

$$\mathbf{r}_{yx} = E\left[\mathbf{y}(n) x(n) \right] = E\left[\begin{pmatrix} y(n) & \mathrm{y}(n-1) & \dots & \mathrm{y}(n-M+1) \end{pmatrix} x(n) \right] \tag{2.15}$$

L'équation (2.13) montre que la fonction J dépend de la réponse impulsionnelle h. Pour on obtenir le minimum, il suffit de chercher les conditions d'annulation de la dérivée de la fonction J par rapport à la réponse impulsionnelle du filtre.

La dérivée de la fonction J par rapport à la réponse impulsionnelle est donnée par :

$$\frac{\partial J}{\partial \mathbf{h}} = -2\mathbf{r}_{yx} + 2\mathbf{h}^T\mathbf{R}_{yy} \tag{2.16}$$

Le vecteur optimum $\mathbf{h}^*$ est celui qui annule le gradient du critère:

$$\frac{\partial J}{\partial \mathbf{h}} = 0 \Leftrightarrow \frac{\partial J}{\partial \mathbf{h}} = -2\mathbf{r}_{yx} + 2\mathbf{h}^T\mathbf{R}_{yy} = 0 \Rightarrow \mathbf{h}^* = \mathbf{R}_{yy}^{-1}\mathbf{r}_{yx} \tag{2.17}$$

2.6.2 Filtrage de Wiener et la réduction du bruit [67]

On dispose d'un message $y(n) = x(n) + d(n)$, où le signal et le bruit sont deux processus aléatoires stationnaires.

$$\mathbf{R}_{yy} = E\left[\mathbf{y}\mathbf{y}^T\right] = E\left[(x+d)(x+d)^T\right] = E\left[xx^T\right] + E\left[dd^T\right] + E\left[xd^T\right] + E\left[dx^T\right] \tag{2.18}$$

Comme le signal et le bruit sont indépendants, on aura donc :

$$\mathbf{R}_{yy} = \mathbf{R}_{xx} + \mathbf{R}_{dd} \tag{2.19}$$

On remplace l'équation (2.19) dans l'équation (2.17) on trouve :

$$\mathbf{h}^* = \left(\mathbf{R}_{xx} + \mathbf{R}_{dd}\right)^{-1}\mathbf{r}_{yx} \tag{2.20}$$

- **Dans le domaine fréquentielle**

On a :

$$\hat{x}(n) = h(n) * y(n) \xrightarrow{TF} \hat{X}(k) = H(k)Y(k) \tag{2.21}$$

Donc :

$$E(k) = X(k) - \hat{X}(k) = X(k) - H(k)Y(k) \tag{2.22}$$

$$E\left[\left|E(k)\right|^2\right] = E\left\{\left[X(k) - H(k)Y(k)\right]^*\left[X(k) - H(k)Y(k)\right]\right\}$$

$$= E\left[\left|X(k)\right|^2\right] - H(k)E\left[X^*(k)Y(k)\right] - H^*(k)E\left[Y^*(k)X(k)\right] + \left|H(k)\right|^2 E\left[\left|Y(k)\right|^2\right] \tag{2.23}$$

$$J_2 = E\left[\left|E(k)\right|^2\right] = E\left[\left|X(k)\right|^2\right] - H(k)P_{yx}(k) - H^*(k)P_{xy}(k) + \left|H(k)\right|^2 P_{yy}(k) \tag{2.24}$$

Où : $P_{xy}(k)$ est la densité interspectrale de puissance de x et y, et $P_{yy}(k)$ est la densité spectrale de puissance de y.

$$\frac{\partial J_2}{\partial H(k)} = H^*(k)P_{yy}(k) - P_{yx}(k) = \left[H(k)P_{yy}(k) - P_{xy}(k)\right]^* = 0 \tag{2.25}$$

$$H(k) = \frac{P_{xy}(k)}{P_{yy}(k)} \tag{2.26}$$

$$Y(k) = X(k) + D(k) \tag{2.27}$$

$$P_{xy}(k) = E\left[X(k)\left\{X(k) + D(k)\right\}^*\right] = P_{xx}(k) \tag{2.28}$$

$$P_{yy}(k) = E\left[\left\{X(k) + D(k)\right\}\left\{X(k) + D(k)\right\}^*\right] = P_{xx}(k) + P_{dd}(k) \tag{2.29}$$

Avec : $P_{xx}(k)$ est la densité spectrale de puissance de x, et $P_{dd}(k)$ est la densité spectrale de puissance de d. En remplaçant les équations (2.28) et (2.29) dans (2.26), on trouve :

$$H(k) = \frac{P_{xx}(k)}{P_{xx}(k) + P_{dd}(k)} \tag{2.30}$$

Comme le rapport signal sur bruit a priori (SNR a priori) ξ_k est défini par :

$$\xi_k = \frac{P_{xx}(k)}{P_{dd}(k)} \tag{2.31}$$

Si on introduit l'équation (2.31) dans l'équation (2.30), on arrive à la fonction de transfert du filtre suivante :

$$H(k) = \frac{\xi_k}{\xi_k + 1} \tag{2.32}$$

Donc, la fonction du gain de Wiener est :

$$G_w(\xi_k, \gamma_k) = \frac{\xi_k}{\xi_k + 1} = H(k) \tag{2.33}$$

On remarque que lorsque $\xi_k \to 0$, $H(k) \approx 0$ et lorsque $\xi_k \to \infty$, $H(k) \approx 1$. Et il est indépendant du rapport signal sur bruit a posteriori (γ_k).

- **Mise en œuvre du filtre**

La mise en œuvre du filtre nécessite que :

L'on segmente le signal $y(n)$ en blocs à traiter successivement ;

L'on connaisse ou que l'on estime les densités spectrales de puissance du bruit et du signal.

2.7 Estimation de la densité spectrale du bruit

L'estimation du spectre de puissance du bruit est un des éléments critiques de toute opération de rehaussement de la parole. Ainsi, l'efficacité des méthodes de rehaussement repose essentiellement sur une estimation préalable correcte du niveau de bruit, cela permet d'avoir une information à la fois sur le niveau du bruit et sur son contenu spectral. Le problème majeur qui se pose est la quasi-inexistence des données sur la densité spectrale de puissance réelle du bruit, d'où la nécessité de considérer une hypothèse de départ pour les différentes méthodes d'estimations du bruit qui consiste en la stationnarité ou la quasi-stationnarité de ce dernier. Plus explicitement, cela revient à dire que les statistiques du bruit de fond varient lentement comparées à celles du signal source.

2.7.1 Estimation initiale du spectre de bruit

Dans les différents contextes applicatifs de réduction du bruit, le signal source n'est, en général, pas présent dans toutes les trames temporelles. Dans le cadre du traitement de la parole, notamment, le signal est régulièrement ponctué de silences au sein desquels seul le bruit persiste. On peut exploiter les périodes de silences au début de l'enregistrement. Sous l'hypothèse de quasi-stationnarité du bruit, il est alors commode d'utiliser ces périodes de silence pour estimer la densité spectrale de puissance du bruit. Le spectre moyen de quelques dizaines de trames qui se trouvent au début du fichier de la parole nous donne une estimation initiale du spectre du bruit.

Cette estimation doit être mise à jour en continue par les algorithmes d'estimation continue du bruit ou seulement durant les périodes de non activité de la parole par les méthodes d'estimation discontinues du bruit basées sur la détection d'activité vocale.

2.7.2 Détection d'activité vocale

La détection d'activité vocale (VAD : Voice Activity Detection) nous permet de faire la distinction entre les segments d'un signal de parole qui comportent de la voix humaine (période d'activité) et les segments du même signal qui n'en ont pas (période de non activité). C'est une partie importante de tout système de communication. Par exemple, on trouve un détecteur d'activité vocale dans presque tous les codeurs de la parole, dans les systèmes de reconnaissance de la parole dans un milieu bruité et dans les systèmes de rehaussement.

Tous les VAD s'appuient sur le même principe de fonctionnement. Ils découpent un signal de parole en trames de 10 à 30 ms, selon l'algorithme, afin de rendre le signal de chaque trame approximativement invariant dans le temps. Ensuite, ces algorithmes prennent différents types de mesures afin de déterminer si le segment sous observation est actif, c'est à dire contenant de la voix humaine, ou non. Les mesures les plus courantes sont [68][69]:

1. Niveau d'énergie du signal

Une comparaison de la puissance de la trame avec un certain seuil de comparaison, adaptatif ou non, peut révéler la présence de parole ou non.

2. Passage par zéro

Le nombre de fois dans une trame où le signal de la parole change de signe (+/-), peut indiquer la présence de parole.

3. Forme spectrale

La distribution d'énergie dans diverses gammes de fréquences. Une certaine forme spectrale peut indiquer la présence de parole.

4. Coefficients d'autocorrélation

Ces coefficients, utilisés dans la prédiction linéaire de la parole (LPC), peuvent indiquer la présence de la voix humaine, voisée ou non.

Après avoir évalué ces critères, ces algorithmes les comparent à un certain seuil, afin de déterminer si le segment sous observation est actif ou non. Ces seuils sont habituellement dynamiques et la densité spectrale estimée du bruit est mise à jour pendant les périodes de non activité selon l'équation suivante :

$$\lambda_d(k,l) = (1-\alpha) R_k^2(l) + \alpha \lambda_d(k,l-1) \qquad (2.34)$$

Où $\lambda_d(k,l)$ est la densité spectrale de puissance estimée du bruit, α un facteur de lissage et $R_k^2(l)$ est la densité spectrale de puissance du signal bruité de la trame d'indice l.

Un signal vocal comporte jusqu'à 60% de silence ou de bruit de fond. Pour réduire la quantité d'informations à transmettre, il est connu de discriminer les portions de signal vocal qui contiennent réellement des signaux utiles et les portions qui ne contiennent que du silence ou du bruit ; et de les coder respectivement selon deux algorithmes différents, chaque portion qui ne contient que du silence ou du bruit étant codée avec très peu d'informations représentant les caractéristiques du bruit ambiant. Généralement, les codeurs comportent un dispositif de détection d'activité vocale qui réalise cette discrimination.

Le décodeur chargé de décoder le signal vocal codé doit utiliser alternativement deux algorithmes de décodage correspondant respectivement aux portions de signal codées comme de la voix et aux portions de signal codées comme du silence ou bruit de fond. Le passage d'un algorithme à l'autre est synchronisé par les informations codant les périodes de silence ou bruit.

Les codeurs connus qui implémentent des VAD comme la norme ITU-T G.729A, annexe B, par exemple, ne sont plus capables de faire la distinction entre le signal utile et le bruit lorsque le niveau de bruit est supérieur à 8000 échelons de l'échelle de quantification définie par cette norme. Il en résulte de nombreuses transitions inutiles du signal de détection d'activité vocale, et donc la perte de portions du signal utile.

Les différentes variantes de la VAD utilisées par les codeurs de la parole sont très efficaces dans les milieux sans bruit, mais une fois le niveau du bruit ambiant devient important et variable les performances de ces détecteurs se détériorent. Une fois le codeur est précédé par un bloc de réduction de bruit la précision de ces détecteurs s'améliore d'avantage. De plus, des algorithmes de la VAD ont été utilisés dans les blocs de réduction de bruit pour une meilleure estimation de la densité spectrale de puissance du bruit, mais ses performances restent limitées par rapport à l'estimation continue [8][9][35].

2.7.3 Estimation continue

L'estimation continue du niveau du bruit suppose que le signal de parole apparaît de manière transitoire au milieu du bruit de fond. En exploitant cette hypothèse, on peut considérer approximativement que toute hausse instantanée du niveau au-dessus de la valeur moyenne estimée du bruit témoigne de la présence du signal. Ainsi, il est alors possible d'estimer le niveau du bruit à partir d'un lissage récursif du premier ordre comparable à celui de l'équation (2.34), en employant des valeurs différentes du paramètre de lissage α suivant que la trame correspond à une attaque ou non.

$$\alpha = \begin{cases} \alpha_a & si \quad R_k^2(l) > \lambda_d(k) \\ \alpha_d & si \quad R_k^2(l) < \lambda_d(k) \end{cases} \tag{2.35}$$

Pour réaliser l'estimation souhaitée, il convient d'employer une constante de temps longue pour les périodes d'attaque et plus courte pour les périodes de descente. Le temps d'attaque (temps de réponse à -3 dB), T_{att}, permet de définir la valeur du paramètre α_a tel que $\alpha_a = 0.5^{T_e/T_{att}}$; où T_e est la période d'échantillonnage [70].

La plupart des algorithmes conçus pour l'estimation continue du bruit peuvent être largement classifiés en deux catégories. La première catégorie est basée sur la mise à jour de l'estimation du bruit en suivant les régions de silence de la parole et l'autre catégorie est basée sur la mise à jour de l'estimation du bruit en utilisant l'histogramme du spectre de puissance de la parole bruitée.

Plusieurs approches ont été proposées dans la littérature [71]-[76], les plus citées sont :

- La méthode de la moyenne pondérée (weighted averaging technique) proposée par Hirsch [72].

- L'approche d'estimation du bruit basée sur le contrôle des minima par une moyenne récursive (Minima Controlled Recursive Averaging : MCRA) proposée dans [74] par Cohen et Berdugo. L'estimation du bruit est mise à jour en faisant la moyenne des valeurs des spectres de puissance précédents. Cette estimation est contrôlée par des facteurs de lissage dans le domaine temporel et fréquentiel. Ces facteurs de lissage sont obtenus en se basant sur la probabilité de présence du signal dans chaque composante fréquentielle séparément. Cette probabilité est calculée en utilisant le rapport entre le spectre de puissance du signal bruité et son minimum local durant une fenêtre de durée fixe. Ce rapport est comparé avec un seuil, où un rapport faible indique l'absence de la parole.

- Dans [76], S. Rangachari et al ont présenté deux nouveaux algorithmes d'estimation du bruit, basés sur la méthode du MCRA. Ces deux algorithmes sont appelés méthode du MCRA_2. Dans le premier algorithme, l'estimation du bruit est mise à jour dans chaque trame basée sur la VAD. Si la parole est absente dans une trame, le bruit estimé sera mis à jour avec un facteur de lissage constant. La décision de présence de la parole est basée sur le rapport entre le spectre de la parole bruitée et son minimum local. Le deuxième algorithme est une amélioration du premier dans les deux aspects suivants : la mise à jour du bruit estimé est faite sans l'utilisation explicite de VAD, et l'exploitation de la corrélation entre les composantes fréquentielles des spectres des trames adjacentes pour l'estimation de la probabilité de la présence de la parole.

- La méthode de base de la majorité des algorithmes d'estimation du bruit est sans doute celle proposée par Martin en 1994 [71], sa version améliorée en 2001 [73] et par la suite adaptée au codeur MELP [9]. Elle est appelée «MS : Minimum Statistics method », basée sur le minimum des statistiques et un facteur de lissage optimal de la densité spectrale de puissance de la parole bruitée. Cette méthode est très efficace et devenue un standard dans l'estimation continue de la densité spectrale de puissance du bruit. L'annexe A sera consacrée au principe et aux détails d'implémentation de cette technique.

2.8 Estimation du SNR a priori

L'implémentation du filtre de Wiener et de la majorité des méthodes Bayesiènnes nécessite le calcul du rapport signal sur bruit a priori ξ_k. Cette quantité doit être estimée pour chaque trame. Deux approches sont considérées ici. Dans la première, un estimateur du maximum de vraisemblance (Maximum Likelihood : ML) de la variance des composantes spectrales de la parole est utilisé. La deuxième approche est basée sur une méthode d'estimation dite décision dirigée (Decision-Directed), les deux approches nécessitent la connaissance de la densité spectrale du bruit.

2.8.1 Approche d'estimation du maximum de vraisemblance [20]

L'approche d'estimation du maximum de vraisemblance est généralement la plus utilisée pour l'estimation d'un paramètre inconnu, de densité de probabilité donnée, quand aucune information a priori sur ce paramètre n'est disponible. Par exemple, l'estimation de la densité spectrale de puissance du signal rehaussé dans notre cas (la variance $\hat{\lambda}_x(k)$).

L'estimateur du ML de $\hat{\lambda}_x(k)$ durant la trame d'analyse (l) est basée sur M observations consécutives $Y_k(l) = \{y_k(l), y_k(l-1), \ldots, y_k(l-M+1)\}$, qui sont considérés comme statistiquement indépendants. Cette supposition est raisonnable quand l'analyse est faite sur des fenêtres non chevauchées. Cependant, dans le système utilisé ici, le recouvrement est fait. Néanmoins, on continue cette supposition puisque la dépendance statistique est difficile à être modélisée et manipulée. On considère également que les variances de la k$^{\text{ieme}}$ composante spectrale du signal et du bruit $\lambda_x(k)$ et $\lambda_d(k)$ respectivement, sont deux paramètres qui varient lentement, de sorte qu'ils puissent être considérés comme constants durant les M observations précédentes. Finalement, nous supposons que la variance de la k$^{\text{ième}}$ composante spectrale du bruit est connue.
L'estimateur du ML $\hat{\lambda}_x(k)$ de $\lambda_x(k)$, est l'argument non négatif qui maximise la densité de probabilité conjointe conditionnelle de $Y_k(l)$ sachant $\lambda_x(k)$ et $\lambda_d(k)$. En se basant sur le modèle statistique Gaussien et l'indépendance statistique des composantes spectrales, cette densité de probabilité est donnée par :

$$p\left(Y_k(l)\middle|\lambda_x(k), \lambda_d(k)\right) = \prod_{m=0}^{M-1} \frac{1}{\pi\left(\lambda_x(k)+\lambda_d(k)\right)} \cdot \exp\left(-\frac{R_k^2(l-m)}{\lambda_x(k)+\lambda_d(k)}\right) \qquad (2.36)$$

Où $R_k(m) = |Y_k(m)|$. La valeur $\hat{\lambda}_x(k)$ est facilement obtenue à partir de (2.36), et égale :

$$\hat{\lambda}_x(k) = \begin{cases} \dfrac{1}{M}\sum_{m=0}^{M-1} R_k^2(l-m) - \lambda_d(k), & \textit{si non négatif} \\ 0\,, & \textit{ailleurs} \end{cases} \qquad (2.37)$$

Cet estimateur produit l'estimateur suivant pour le SNR a priori ξ_k :

$$\hat{\xi}_k = \begin{cases} \dfrac{1}{M}\sum_{m=0}^{M-1} \gamma_k(l-m) - 1, & \textit{si non négatif} \\ 0\,, & \textit{ailleurs} \end{cases} \qquad (2.38)$$

Où : $\gamma_k(m) = |Y_k(m)|^2 / \lambda_d(k)$ est le SNR a posteriori dans la m[ième] fenêtre d'analyse.

Dans la pratique, la moyenne courante requise en (2.38) est remplacée par une moyenne récursive avec une constante de temps comparable à la période de corrélation de γ_k. C'est-à-dire, l'estimateur de ξ_k dans la l[ième] fenêtre d'analyse est obtenu par :

$$\overline{\gamma}_k(l) = \alpha \overline{\gamma}_k(l-1) + (1-\alpha)\frac{\gamma_k(l)}{\beta}; \qquad 0 \le \alpha \le 1 \;;\; \beta \ge 1 \tag{2.39}$$

$$\hat{\xi}_K(l) = \begin{cases} \overline{\gamma}_k(l) - 1, & \overline{\gamma}_k(l) - 1 \ge 0 \\ 0, & \textit{ailleurs} \end{cases} \tag{2.40}$$

β est un facteur de correction et les valeurs de α et β sont déterminées par des tests d'écoute. D'après [20], des valeurs de $\alpha = 0.725$ et $\beta = 2$ assurent une qualité acceptable soit avec l'estimateur MMSE-STSA ou avec l'estimateur de Wiener.

2.8.2 Approche d'estimation de décision dirigée (decision-directed)

La méthode de décision dirigée est la plus utilisée pour l'estimation du SNR a priori dans chaque composante spectrale. Cet estimateur s'avère très utile quand il est combiné avec les estimateurs au sens du MMSE ou l'estimateur d'amplitude de Wiener [20].

Soient $\xi_k(l)$, $A_k(l)$, $\lambda_d(k,l)$ et $\gamma_k(l)$ le SNR a priori, l'amplitude, la variance du bruit et le SNR a posteriori, respectivement, de la k[ième] composante spectrale correspondante dans la trame d'analyse d'indice (l).

Le calcul de l'estimateur du SNR a priori est basé ici sur la définition de $\xi_k(l)$ et sa relation au SNR a posteriori $\gamma_k(l)$, comme donné dans les deux équations suivantes :

$$\xi_k(l) = \frac{\mathrm{E}\left[A_k^2(l)\right]}{\lambda_d(k,l)} \tag{2.41}$$

$$\xi_k(l) = \mathrm{E}\left[\gamma_k(l) - 1\right] \tag{2.42}$$

Utilisant (2.41) et (2.42), on peut écrire :

$$\xi_k(l) = \mathrm{E}\left\{\frac{1}{2}\frac{A_k^2(l)}{\lambda_d(k,l)} + \frac{1}{2}\left[\gamma_k(l) - 1\right]\right\} \tag{2.43}$$

L'estimateur $\hat{\xi}_k(l)$ de $\xi_k(l)$ est déduit de l'équation (2.43), et est donné par :

$$\hat{\xi}_k(l) = \alpha \frac{\hat{A}_k^2(l-1)}{\lambda_d(k,l-1)} + (1-\alpha)\, p\left[\gamma_k(l) - 1\right], \qquad 0 \le \alpha < 1 \tag{2.44}$$

Où $\hat{A}_k(l-1)$ est l'amplitude estimée de la k[ième] composante spectrale du signal dans la trame d'analyse d'indice $(l-1)$, et $p[.]$ est un opérateur défini par :

$$p[x]=\begin{cases} x\,, & si\ x\geq 0 \\ 0\,, & ailleurs \end{cases} \tag{2.45}$$

$p[.]$ est utilisé pour assurer que l'estimateur $\hat{\xi}_k(l)$ est toujours positif si $\gamma_k(l)-1$ est négatif.

L'estimateur $\hat{\xi}_k(l)$ de $\xi_k(l)$ est un estimateur de type « décision dirigée », parce que la mise à jour de $\hat{\xi}_k(l)$ est basée sur l'amplitude estimée de la trame précédente.

En utilisant $\hat{A}_k(l)=G\left[\hat{\xi}_k(l),\gamma_k(l)\right]\cdot R_k(l)$, où $G[.,.]$ est une fonction de gain qui résulte de l'estimateur au sens du MMSE ou de l'estimateur d'amplitude de Wiener. L'équation (2.44) peut être écrite ainsi :

$$\hat{\xi}_k(l)=\alpha\cdot G^2\left[\hat{\xi}_k(l-1),\gamma_k(l-1)\right]\cdot\gamma_k(l-1)+(1-\alpha.)\cdot P\left[\gamma_k(l)-1\right] \tag{2.46}$$

Plusieurs conditions initiales ont été examinées par simulation dans les travaux de recherches [20]-[22]. L'utilisation de $\hat{\xi}_k(0)=\alpha+(1-\alpha)P\left[\gamma_k(0)-1\right]$ est appropriée, puisqu'elle minimise les effets des transitions initiales dans la parole rehaussée.

2.9 Conclusion

Au vu de ce chapitre, le principe du filtre de Wiener qui est la méthode de base des approches Bayesiennes a été présenté. Toutes les autres approches qui seront détaillées dans le chapitre suivant seront comparées avec le filtre de Wiener. En plus, la méthode d'estimation de décision dirigée sera utilisée dans toutes les approches, pour l'estimation du SNR a priori par rapport à l'estimateur ML, à cause de son efficacité, comme largement citées dans les travaux de recherche.

CHAPITRE 3

APPLICATION DU FILTRE DE KALMAN AU REHAUSSEMENT DE LA PAROLE

3.1 Introduction

Plusieurs méthodes, ont été développées, à partir de la représentation des systèmes dans l'espace d'état par des équations différentielles matricielles du premier ordre. Ces techniques sont basées sur le fait qu'un processus aléatoire peut être modélisé comme étant la sortie d'un système linéaire gouverné par un bruit blanc, c'est le filtre de Kalman [19].

L'objectif du filtre de Kalman est d'obtenir une estimation récursive optimale du vecteur d'état, qui est basée sur une mesure disponible jusqu'à l'instant courant et compte tenu de l'estimation à l'instant précédent.

La technique du filtrage de Kalman qui sera présentée dans ce chapitre, est largement utilisée aujourd'hui dans plusieurs domaines, surtout en traitement de la parole et en réduction de bruit. Elle a prouvé sa supériorité non seulement par rapport aux méthodes classiques mais également par rapport au filtre de Wiener.

3.2 Généralités sur le filtre de Kalman

3.2.1 Définition du filtre de Kalman

Le filtre de Kalman doit son nom à *Rudolf Kalman* bien que *Thorvald Nicolai Thiele* [77] et *Peter Swerling* aient développé un algorithme similaire avant lui. La paternité du filtre fait l'objet d'une petite controverse dans la communauté scientifique. Le filtre a été décrit dans diverses publications par Swerling (1958), Kalman (1960) [19] et Kalman-Bucy (1961) [78]. En 1960, R.E. Kalman a édité son article célèbre décrivant une solution récursive au problème de filtrage linéaire de données discrètes [19]. Il a introduit un filtre, à partir de la représentation des systèmes dans l'espace d'état par des équations différentielles matricielles du premier ordre [67][79].

Stanley Schmidt est reconnu comme ayant réalisé la première implémentation du filtre. C'était lors d'une visite de Rudolf Kalman au NASA Ames Research Center qu'il vit le potentiel de son filtre pour l'estimation de la trajectoire pour le programme Apollo. Ceci conduisit à l'utilisation du filtre dans l'ordinateur de navigation.

Depuis cette époque, Une grande variété de filtres de Kalman a été développée à partir de la formulation originale dite filtre de Kalman simple. Schmidt développa le filtre de Kalman étendu, Bierman, Thornton et bien d'autres développèrent toute gamme de filtres racine carré.

Le filtre de Kalman est un estimateur récursif, cela signifie que pour estimer l'état courant, seulement l'état précédent et les mesures actuelles sont nécessaires. L'historique des observations et des estimations n'est ainsi pas requis. Statistiquement, cet estimateur est optimal au sens du critère de l'erreur quadratique moyenne. Dans la pratique, ce filtre de Kalman est une des plus grandes découvertes dans l'histoire de la théorie d'estimation statistique.

Il est utilisé dans plusieurs domaines technologiques (radar, vision électronique, télécommunication,…), en météorologie, finance et en navigation.

Ces dernières années, plusieurs systèmes de pré-traitement des codeurs de la parole utilisent le filtre de Kalman et ses variantes pour la réduction de bruit. Ses performances vis-à-vis les méthodes STSA de réduction de bruit [58]-[63], nous ont poussés à l'étudier dans le cadre de ce travail [37]-[39].

3.2.2 Calcul du filtre de Kalman

Le filtre de Kalman sert à estimer l'état $x \in \Re^n$ d'un système qui est soumis à une équation différentielle stochastique linéaire avec une mesure $z \in \Re^m$. Ce qui donne à chaque pas k les équations d'état et de mesure suivantes [79][80] :

$$x_k = Ax_{k-1} + Bu_k + w_{k-1} \text{ (Équation d'état)} \quad (3.1)$$

$$z_k = Hx_k + v_k \text{ (Équation de mesure ou d'observation)} \quad (3.2)$$

Avec :

- **A** $(n \times n)$ est la matrice de prédiction qui relie l'état x_{k-1} à l'état x_k.
- **B** $(n \times l)$ relie l'état x_k à un éventuel signal de contrôle (ou consigne) $u \in \Re^l$. C'est le vecteur de consigne.
- **H** $(m \times n)$ est l'équation de mesure qui relie l'état x_k à la mesure z_k.
- u_k est une consigne appliquée en entrée au système.

Les variables aléatoires w_k et v_k représentent respectivement le bruit du système et de la mesure. Ils sont supposés indépendants l'un par rapport à l'autre. De plus, ce sont des

bruits blancs dont, par définition, la distribution suit une loi normale de moyenne nulle et de matrice de covariance non nulle, ce qui nous donne, pour les densités de probabilité des variables w et v notées $\mathrm{P}(w)$ et $\mathrm{P}(v)$:

$$\mathrm{P}(w) \sim N(0, Q) \tag{3.3}$$
$$\mathrm{P}(v) \sim N(0, R) \tag{3.4}$$

Donc :

$$\begin{cases} E[w_k] = 0. \\ E[v_k] = 0. \\ E[w_k v_i^T] = 0, \; pour 1 \le i \le k. \\ E[w_k w_i] = Q, \; pour 1 \le i \le k. \\ E[v_k v_i] = R, \; pour 1 \le i \le k. \end{cases} \tag{3.5}$$

Avec :

- Q et R sont des matrices de covariance associées aux bruits sur le processus à estimer et aux bruits sur la mesure.

Les matrices A, B, H**,** Q et R changent en fonction de l'évolution du processus.

On définit :

- $\hat{x}_k^- \in \Re^n$ l'état estimé a priori pour l'étape k (ou estimateur a priori), la connaissance du système étant donnée avant l'étape k.
- $\hat{x}_k \in \Re^n$ l'état estimé a posteriori (ou estimateur a posteriori) de x_k, connaissant la mesure z_k.

Le vecteur d'état x_k contient l'information pertinente du système à tout moment et décrit l'évolution du système en fonction du temps (ou d'une étape) k et des données de contrôle u_k.

Soit une estimation a priori et a posteriori de l'erreur de mesure :

$$e_k^- \equiv x_k - \hat{x}_k^- \tag{3.6}$$
$$e_k \equiv x_k - \hat{x}_k \tag{3.7}$$

Alors les covariances de l'erreur estimée a priori et a posteriori sont :

$$P_k^- = E[e_k^- e_k^{-T}] \tag{3.8}$$
$$P_k = E[e_k e_k^T] \tag{3.9}$$

Notre objectif est d'obtenir une estimation récursive optimale du vecteur d'état, qui est basée sur une mesure disponible jusqu'à l'instant $k+1$ et compte tenu de l'estimation à l'instant k.

On a :

$$\hat{x}_k = k_k \hat{x}_k^- + K_k z_k \tag{3.10}$$

Les matrices k_k et K_k jusqu'ici sont inconnues. Nous cherchons leurs valeurs telles que la nouvelle estimation $\hat{x}_k$ satisfera les deux conditions suivantes :

$$E([x_k - \hat{x}_k] z_i^T) = 0, \; i = 1, 2, \; \ldots, k-1 \tag{3.11}$$
$$E([x_k - \hat{x}_k] z_k^T) = 0 \tag{3.12}$$

On remplace les formules de x_k et $\hat{x}_k$ dans l'équation (3.11), on obtient donc la relation suivante :

$$E[(Ax_{k-1} + Bu_k + w_{k-1} - k_k\hat{x}_k^- - K_k z_k)z_i^T] = 0,\ i = 1,2,\ \ldots\ ,k-1 \qquad (3.13)$$

Et avec $z_k = Hx_k + v_k$, l'équation (3.13) devient :

$$E[(Ax_{k-1} + Bu_k + w_{k-1} - k_k\hat{x}_k^- - K_k Hx_k - K_k v_k)z_i^T] = 0,\ i = 1,2,\ \ldots\ ,k-1 \qquad (3.14)$$

Puis on remplace aussi x_k par sa formule, on obtient donc :

$$E[(Ax_{k-1} + Bu_k + w_{k-1} - k_k\hat{x}_k^- - K_k HAx_{k-1} - K_k HBu_k - K_k Hw_{k-1} - K_k v_k)z_i^T] = 0 \qquad (3.15)$$

Comme :

$$\begin{cases} E[w_k z_i^T] = 0. \\ E[v_k z_i^T] = 0. \end{cases}$$

Alors l'équation précédente sera réduite ainsi :

$$\begin{gathered} AE[x_{k-1}z_i^T] + BE[u_k z_i^T] - k_k E[\hat{x}_k^- z_i^T] - K_k HAE[x_{k-1}z_i^T] - K_k HBE[u_k z_i^T] = 0 \\ E([x_k - K_k Hx_k - k_k x_k] - k_k(\hat{x}_k^- - x_k))z_i^T = 0 \\ [I - k_k - K_k H]E(x_k z_i^T) = 0 \end{gathered} \qquad (3.16)$$

L'équation (3.16) peut être satisfaite pour n'importe quelle donné x_k si :

$$k_k = I - K_k H \qquad (3.17)$$

Si on remplace cette formule de k_k dans l'équation (3.10) on obtient les équations suivantes :

$$\begin{gathered} \hat{x}_k = (I - K_k H)\hat{x}_k^- + K_k z_k \\ \hat{x}_k = \hat{x}_k^- + K_k[z_k - H\hat{x}_k^-] \end{gathered} \qquad (3.18)$$

Où :

- K_K est le gain de Kalman.
- La partie $(z_k - H\hat{x}_k^-)$ de l'équation (3.18) représente la différence entre la mesure z_k et la prédiction $H\hat{x}_k^-$ (appelée aussi l'innovation ou le résiduel). Lorsque le résiduel est nul, alors la mesure est identique à la prédiction obtenue à l'itération précédente [79].

- **Calcul du gain de Kalman**

Le gain de Kalman K_K est un facteur qui permet de donner plus ou moins de poids à la prédiction antérieure pour l'estimation de l'état, le calcul du gain est effectué de telle sorte qu'il minimise la covariance de l'erreur a posteriori P_k. Ce dernier contient l'espérance de l'erreur au carré et c'est ce qui fait qu'un filtre de Kalman est une solution optimale du problème de minimisation des moindres carrés (« minimum least square »).
On définit l'erreur sur z_k par :

$$\tilde{z}_k \cong z_k - \hat{z}_k^-$$
$$\tilde{z}_k \cong z_k - H\ \hat{x}_k^- \tag{3.19}$$

Le paramètre $\hat{x}_k$ dépend linéairement de x_k, et aussi de z_k. Par conséquent, de l'équation (3.12) on peut écrire que :

$$E([x_k - \hat{x}_k]z_k^{-T}) = 0 \tag{3.20}$$

Par la soustraction de (3.20) et (3.12) on trouve :

$$E([x_k - \hat{x}_k]\tilde{z}_k^T) = 0 \tag{3.21}$$

Remplaçant x_k, $\hat{x}_k$ et $\tilde{z}_k$ de (3.1), (3.18), (3.19) respectivement par leurs formules dans l'équation (3.21) on obtient l'équation suivante :

$$E[Ax_{k-1} + Bu_k + w_{k-1} - \hat{x}_k^- - K_k(z_k - H\ \hat{x}_k^-)][z_k - H\ \hat{x}_k^-]^T = 0 \tag{3.22}$$

Et avec $z_k = Hx_k + v_k$, l'équation (3.22) peut être réécrite comme suit :

$$E[Ax_{k-1} + Bu_k - \hat{x}_k^- + K_k H\hat{x}_k^- - K_k Hx_k - K_k v_k][Hx_k + v_k - H\hat{x}_k^-]^T = 0$$
$$E[(\underbrace{x_k - \hat{x}_k^-}_{e_k^-}) - K_k H\ \underbrace{(x_k - \hat{x}_k^-)}_{e_k^-} - K_k v_k][H\ \underbrace{(x_k - \hat{x}_k^-)}_{e_k^-} + v_k]^T = 0$$
$$E[[I - K_k H]\underbrace{(x_k - \hat{x}_k^-)}_{e_k^-} - K_k v_k][H\ \underbrace{(x_k - \hat{x}_k^-)}_{e_k^-} + v_k]^T = 0$$
$$[I - K_k H]E\underbrace{(x_k - \hat{x}_k^-)}_{e_k^-}\underbrace{(x_k - \hat{x}_k^-)^T}_{e_k^{-T}} H^T - K_k E(v_k v_k^T) = 0$$
$$[I - K_k H]P_k^- H^T - K_k R = 0$$

Donc, le gain de Kalman K_k peut être exprimé ainsi :

$$K_k = P_k^- H^T[HP_k^- H^T + R]^{-1} = \frac{P_k^- H^T}{H\ P_k^- H^T + R} \tag{3.23}$$

Où R est la covariance du bruit de mesure.

En analysant la forme de ce gain, les conclusions suivantes sont faites:

- $\lim_{R\to 0} K_k = H^{-1}$, donc plus la covariance du bruit de mesure diminue, plus le gain augmente, jusqu'à H^{-1}. Ainsi, moins l'incertitude sur la mesure est grande, plus la mesure z_k sera prise en compte au détriment de la prédiction. En effet, si K_k est remplacé avec $R = 0$ dans l'équation (3.18), $\hat{x}_k = H^{-1}z_k$ est obtenu où les termes référant à l'état de l'étape précédente ($\hat{x}_k^-$) ont étés éliminés.
- $\lim_{P^-\to 0} K_k = 0$, évidemment, si la covariance de l'erreur estimée a priori est nulle, cela signifie que l'état actuel correspond à l'état prédit à l'étape précédente, donc $e_k^- = 0$ et $x_k = \hat{x}_k^-$ (3.6) et qu'aucun correctif n'est apporté par le gain.

En résumé, lorsque la covariance du bruit de mesure diminue, le filtre fait d'autant plus confiance à la mesure z_k, alors que lorsque la covariance de l'erreur estimée a priori diminue, c'est la prédiction $\hat{x}_k^-$ qui obtient la faveur du filtre.

- **Calcul de la matrice de covariance a posteriori P_k**

Par la soustraction de (3.1) et (3.18) on trouve :

$$x_k - \hat{x}_k = x_k - \hat{x}_k^- - K_k H x_k - K_k v_k + K_k H \hat{x}_k^-$$
$$e_k = e_k^- - K_k H e_k^- - K_k v_k = (I - K_k H) e_k^- - K_k v_k \qquad (3.24)$$

On a :

$$P_k = E(e_k e_k^T) = E[((I - K_k H)e_k^- - K_k v_v)((I - K_k H)e_k^- - K_k v_k)^T]$$
$$P_k = E[(I - K_k H)e_k^- e_k^{-T}(I - K_k H)^T + K_k v_k v_k^T K_k^T]$$
$$P_k = (I - K_k H)P_k^-(I - K_k H)^T + K_k R_k K_k^T]$$
$$P_k = P_k^- - K_k H P_k^- - P_k^- H^T K_k^T + K_k H P_k^- H^T K_k^T + K_k R_k K_k^T$$
$$P_k = (I - K_k H)P_k^- - P_k^- H^T K_k^T + \underbrace{K_k (H P_k^- H^T + R_k)}_{P_k^- H^T} K_k^T$$

Donc la matrice de covariance a posteriori P_k peut être exprimée comme suit :

$$P_k = (I - K_k H)P_k^- \qquad (3.25)$$

- **Calcul de la matrice de covariance a priori P_k^-**

On a :

$$\hat{x}_k^- = A\hat{x}_{k-1} + Bu_k \qquad (3.26)$$

Par la soustraction de (3.1) et (3.26) on trouve :

$$x_k - \hat{x}_k^- = x_k - A\hat{x}_{k-1} - Bu_k$$
$$e_k^- = Ax_{k-1} + Bu_k + w_{k-1} - A\hat{x}_{k-1} - Bu_k$$
$$e_k^- = Ae_{k-1} + w_{k-1}$$

On a :

$$P_k^- = E(e_k^- e_k^{-T})$$
$$P_k^- = E[(Ae_{k-1} + w_{k-1})(Ae_{k-1} + w_{k-1})^T]$$
$$P_k^- = AE[e_{k-1}e_{k-1}^T]A^T + E[w_{k-1}w_{k-1}^T]$$

Donc, la matrice de covariance a priori P_k^- peut être exprimée ainsi :

$$P_k^- = AP_{k-1}A^T + Q_{k-1} \qquad (3.27)$$

Où : I est la matrice identité, P_k^- est la covariance de l'erreur estimée a priori pour l'étape courante et P_k (3.25) est la covariance de l'erreur a posteriori, utilisée pour calculer la covariance de l'erreur estimée a priori pour l'étape suivante (3.27). La covariance de l'erreur Q_{k-1} est additionnée à la covariance de l'erreur a priori multipliée par la matrice A et A^T respectivement, afin de projeter une prédiction sur la prochaine covariance de l'erreur a priori.

3.2.3 Algorithme du filtre de Kalman discret

Le filtre de Kalman est un estimateur récursif cela signifié que pour estimer l'état courant, seuls l'état précédent et les mesures actuelles sont nécessaires. L'historique des observations et des estimations n'est ainsi pas requis. Le filtre de Kalman estime un processus en utilisant une forme du contrôle avec retour d'état : Le filtre estime l'état du processus à un instant donné et après il obtient une rétroaction sous forme des observations bruitées. Par conséquent, le filtre de Kalman a deux phases distinctes : prédiction (les équations de la mise à jour en temps) et mise à jour (les équations de la mise à jour en mesure).

La phase de prédiction utilise l'état estimé à l'instant précédent pour produire une estimation de l'état courant. Dans l'étape de mise à jour, les observations à l'instant courant sont utilisées pour corriger l'état prédit dans le but d'obtenir une estimation plus précise. Les équations de mise à jour en temps sont considérées comme des équations du prédicteur, alors que les équations de mise à jour en mesure peuvent être considérées comme des équations du correcteur. Enfin, l'algorithme du filtre de Kalman discret ressemble à un algorithme de prédiction, correction pour la résolution des problèmes numériques comme illustré par la figure suivante [80].

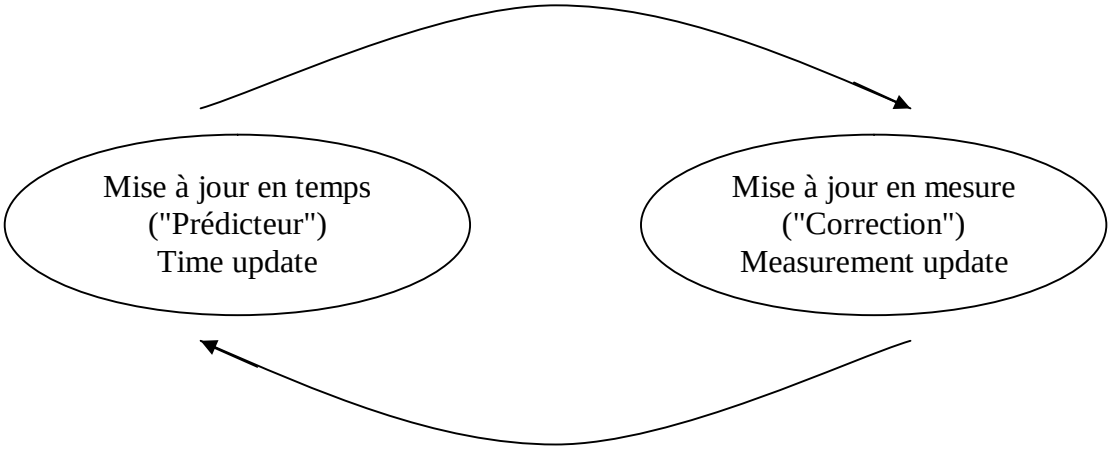


Figure 3.1 : Fonctionnement du filtre de Kalman discret.

Les équations spécifiques pour la mise à jour en temps (prédiction) sont :

$\hat{x}_k = A\hat{x}_{k-1} + Bu_k$ (état prédit)	(3.28)
$P_k^- = AP_{k-1}A^T + Q$ (mise à jour de la matrice de covariance de l'état)	(3.29)

Tableau 3.1 : Équations de mise à jour en temps.

Les équations spécifiques de la mise à jour en mesure (correction) sont :

$K_k = P_k^- H^T (HP_k^- H^T + R)^{-1}$ (gain de Kalman optimal)	(3.30)
$\hat{x}_k = \hat{x}_k^- + K_k(z_k - H\hat{x}_k^-)$ (mise à jour de l'état)	(3.31)
$P_k = (I - K_k H)P_k^-$ (mise à jour de la covariance)	(3.32)

Tableau 3.2 : Équations de mise à jour en mesure.

La figure 3.2 montrée ci-dessous, offre une image beaucoup plus complète et plus claire du fonctionnement du filtre de Kalman discret.

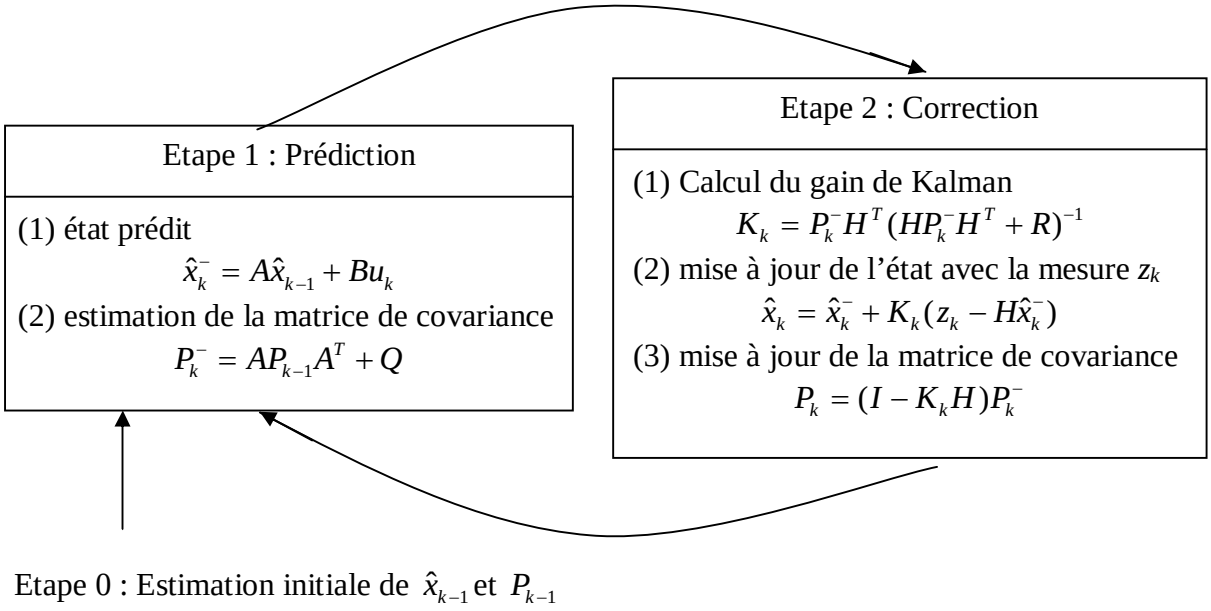


Figure 3.2 : Image complète du fonctionnement du filtre de Kalman discret.

L'étape 0 : constitue l'initialisation des paramètres du filtre avant de commencer à faire de la prédiction. Cette étape ne s'exécute qu'une seule fois.

L'étape 1 : se charge d'estimer le futur état du système, connaissant l'état précédent.

L'étape 2 : estime l'état réel du système à partir du vecteur de mesures.

On passe de l'étape 1 à l'étape 2 et inversement aussi longtemps que l'on a besoin d'estimer l'état du système.

3.2.4 Paramètres et accord du filtre

La covariance de bruit de mesure R est généralement calculée avant l'opération de filtrage de Kalman. En prenant les premières trames, une estimation initiale de cette covariance est pratiquement réalisable. En général, chaque locuteur prend un instant de silence de $0 \approx 0.2s$ avant le commencement du dialogue réel, cette période initiale nous informe sur les caractéristiques du bruit ambiant. Pendant l'opération du filtrage, la valeur de R est mise à jour continuellement ou pendant la non activité vocale. Par contre, la détermination de la covariance du bruit de processus Q est une tâche très délicate. Ceci est due ou fait que parfois le processus à estimer n'est pas observable directement. Un modèle AR simple et une analyse LPC peuvent être utilisés par exemple où Q est le gain de cette prédiction. En plus, dans le cas où Q et R sont constantes, les deux estimés P_k (la matrice de covariance d'erreur) et K_k (le gain de Kalman) stabiliserons rapidement et demeureront constantes.

Plusieurs approches de rehaussement de la parole basées sur le filtrage de Kalman ont été présentées dans la littérature. Ces approches se diffèrent l'une de l'autre par la modélisation choisie et l'algorithme utilisé pour l'estimation des paramètres du modèle.

3.3 Modélisation dans le cas d'un bruit blanc

La parole bruitée est donnée sous la forme :

$$y(n) = s(n) + v(n) \tag{3.33}$$

Où : $s(n)$ est la parole propre, $v(n)$ le bruit additif et $y(n)$ la parole bruitée observée. L'estimation de $s(n)$ au sens de l'erreur quadratique moyenne minimale est donnée par :

$$\hat{s}(n) = E\left[s(n) / y(0), y(1), \ldots, y(n)\right] \tag{3.34}$$

On considère que le signal de la parole $s(n)$ est modélisé par un processus autorégressif (AR) d'ordre p :

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + u(n) \tag{3.35}$$

$u(n)$ est le bruit du processus. Les séquences de bruits $u(n)$ et $v(n)$ sont considérées comme des bruits indépendants, blancs, Gaussiens, de moyennes nulles et de matrices de covariance respectives Q et R. Les p coefficients a_i sont les coefficients de la prédiction linéaire (LP).

La formulation de ce problème dans l'espace d'état est [58][60] :

$$\begin{cases} x(n) = Fx(n-1) + Gu(n) \\ y(n) = Hx(n) + v(n) \end{cases} \tag{3.36}$$

Où : $x(n)$ est le vecteur d'état $(p \times 1)$ constitué par les p dernières valeurs du signal $s(n)$

$$x(n) = \left[s(n-p+1), \ldots, s(n)\right]^T \tag{3.37}$$

G : le vecteur d'entrée $(p \times 1)$,

$$G = \left[0, \ldots, 0, 1\right]^T \tag{3.38}$$

F : la matrice de transition $(p \times p)$ (par la suite elle sera notée F_s),

$$F = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 1 \\ a_p & a_{p-1} & a_{p-2} & \cdots & a_1 \end{bmatrix} \tag{3.39}$$

H : est le vecteur d'observation $(1 \times p)$,

$$H = \left[0, \ldots, 0, 1\right] \tag{3.40}$$

Les équations d'estimation et de mise à jour du filtre de Kalman sont [58][62] :

Etape d'initialisation :

$$\hat{x}(0/0) = E\{x(0)\} \tag{3.41}$$

$$P(0/0) = E\left\{\left[x(0) - \hat{x}(0/0)\right]\left[x(0) - \hat{x}(0/0)\right]^T\right\} \tag{3.42}$$

Etape de prédiction :

$$\hat{x}(n/n-1) = F(n)\hat{x}(n-1/n-1) \tag{3.43}$$

$$P(n/n-1) = F(n)P(n-1/n-1)F(n)^T + Q(n) \tag{3.44}$$

Etape de correction :

$$\tilde{z}(n) = y(n) - H\hat{x}(n/n-1) \tag{3.45}$$

$$S(n) = HP(n/n-1)H^T + R(n) \tag{3.46}$$

$$K(n) = P(n/n-1)H^T S^{-1}(n) \tag{3.47}$$

$$\hat{x}(n/n) = \hat{x}(n/n-1) + K(n)\tilde{z}(n) \tag{3.48}$$

$$P(n/n) = \left(I - K(n)H\right)P(n/n-1) \tag{3.49}$$

3.4 Modélisation dans le cas d'un bruit coloré

Dans le cas d'un signal de la parole bruité par un bruit coloré [59][61][63], le signal de la parole $s(n)$ et le bruit additif $v(n)$ peuvent être modéliser par deux modèles AR d'ordre p et q respectivement :

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + u(n) \tag{3.50}$$

$$v(n) = \sum_{j=1}^{q} b_j v(n-j) + w(n) \tag{3.51}$$

$$y(n) = s(n) + v(n) \tag{3.52}$$

Où : $s(n)$ est le $n^{ème}$ échantillon du signal de la parole, $v(n)$ est le $n^{ème}$ échantillon de bruit additif, $y(n)$ est le $n^{ème}$ échantillon d'observation, a_i est le $i^{ème}$ paramètre AR du modèle de la parole et b_j est le $j^{ème}$ paramètre AR du modèle de bruit.

Ce système peut être représenté dans l'espace d'état par le modèle composé suivant :

$$\begin{cases} x(n) = Fx(n-1) + Gu(n) \\ y(n) = Hx(n) \end{cases} \tag{3.53}$$

avec $x(n)$ est le vecteur d'état $(p+q)\times 1$,

$$x(n)=\left[s(n-p+1),\ldots,s(n),v(n-q+1),\ldots,v(n)\right]^T \tag{3.54}$$

G : est le vecteur d'entrée $(p+q)\times 1$,

$$G=\left[0,\ldots,0,1,0,\ldots,0,1\right]^T \tag{3.55}$$

F : est la matrice de transition $(p+q)\times(p+q)$,

$$F=\begin{bmatrix} F_s & 0 \\ 0 & F_v \end{bmatrix} \tag{3.56}$$

$$F_s=\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 1 \\ a_p & a_{p-1} & a_{p-2} & \cdots & a_1 \end{bmatrix} \tag{3.57}$$

$$F_v=\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 1 \\ b_q & b_{q-1} & b_{q-2} & \cdots & b_1 \end{bmatrix} \tag{3.58}$$

H : est le vecteur d'observation $1\times(p+q)$,

$$H=\left[0,\ldots,0,1,0,\ldots,0,1\right] \tag{3.59}$$

3.5 Estimation des paramètres

L'application du filtre de Kalman au rehaussement de la parole s'effectue en deux étapes séparées :

1) Estimation des coefficients AR $\{a_1,a_2,\ldots,a_p\}$ et les matrices de covariances des bruits Q et R pour chaque segment où la parole est stationnaire, en plus des coefficients AR $\{b_1,b_2,\ldots,b_q\}$ dans le cas d'un bruit coloré.

2) L'application de l'algorithme du filtre de Kalman en utilisant les valeurs des paramètres estimées.

Enfin, L'échantillon de la parole estimé à chaque fois $\hat{s}(n)$, est le $p^{ème}$ composant du vecteur d'état estimé $\hat{x}(n/n)$.

Le problème concernant l'estimation des paramètres doit être examiné attentivement, la robustesse de la détermination de ces paramètres influe sur la qualité de la parole rehaussée par le filtre [57]. Pour cela on a le choix entre deux méthodes pour estimer ces paramètres ; soit directement à partir de la parole propre pour comparer entre les variantes des filtres (si l'accès à la source est possible) ou bien de la parole bruitée (si l'accès à la source n'est pas possible). Cette dernière est la méthode la plus utilisée dans le rehaussement de la parole, malgré sa mauvaise qualité par rapport à la première, car dans la réalité la source propre de la parole est inconnue.

Une autre technique plus robuste, consiste à estimer et corriger ces valeurs d'une manière récursive tout en rehaussant la parole (l'algorithme EM) [81]-[83]. La section 3.6 suivante sera consacrée à cet algorithme.

La figure (3.3) illustre un exemple d'un signal de parole propre contaminé par un bruit blanc additif, avec un SNR = 5 dB. Ce signal est filtré par un filtre de Kalman d'ordre $p=10$ (modélisation de la parole seulement), un signal de parole rehaussé à la sortie est reconstruit. Une fenêtre d'analyse de 5 ms est utilisée, les coefficients AR du filtre sont mises à jour à chaque nouvelle trame.

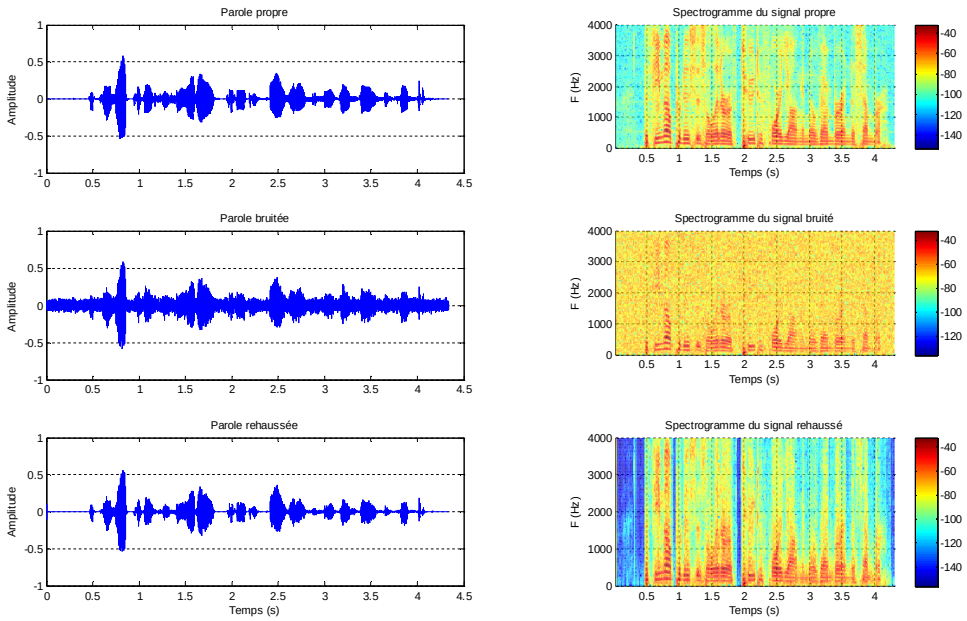


Figure 3.3 : Formes d'ondes et spectrogrammes, cas d'un bruit blanc.

La figure (3.4) montre un autre exemple dans le cas d'un signal de parole bruité par un bruit coloré, le signal rehaussé est filtré par un filtre de Kalman d'ordre $p=10$ pour la modélisation de la parole et $q=8$ pour la modélisation du bruit coloré. Ce bruit coloré $v(n)$ est obtenu par l'excitation d'un filtre AR d'ordre 8 par un bruit blanc $w(n)$ comme suit [61] :

$$\begin{aligned} v(n) = &-0.0851v(n-1)+0.19126v(n-2)+0.0458.v(n-3)+0.0229v(n-4)+ \\ &0.1097v(n-5)+0.1553.v(n-6)-0.132v(n-7)-0.76v(n-8)+w(n) \end{aligned} \quad (3.60)$$

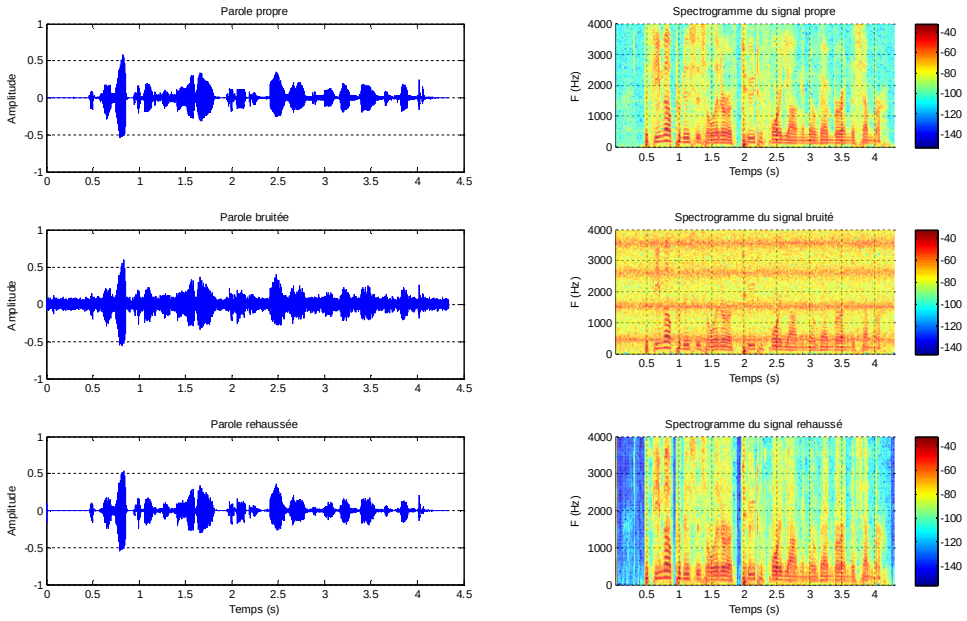


Figure 3.4 : Formes d'ondes et spectrogrammes, cas d'un bruit coloré.

3.6 Résultats de simulation [37][38]

L'algorithme du filtre de Kalman a été appliqué a des signaux prisent de la base de données « Noizeus ». L'utilisation d'un tel corpus (Noizeus) est pour faciliter la comparaison des variantes du filtre de Kalman en se basant sur plusieurs phrases phonétiquement équilibrées, sous divers bruits [46] et à des rapports signal sur bruit différents. Ainsi, pour chaque méthode, pour un type de bruit bien défini et pour un rapport signal sur bruit désiré, toutes les trente phrases de la parole propre et les trente phrases de la parole bruitée correspondantes seront utilisées pour chaque mesure objective, où la moyenne est appliquée à la fin. Le SNR global est calculé selon la norme ITU P.56 [84].

3.6.1 Sans modèle de bruit

Dans le cas du filtre de Kalman simple (modélisation dans le cas d'un bruit blanc), une routine Matlab a été mise au point, où un filtre de Kalman simple a été appliqué sur tous les 30 fichiers de la base de données choisie. Dans cette première approche, seulement le signal de parole est modélisé par un modèle AR d'ordre $p=10$. Ces 10 coefficients AR sont obtenus à chaque trame d'analyse de durée 20 ms, par une analyse LPC appliquée directement sur le signal propre.

Le bruit additif est considéré comme étant un bruit blanc, même dans le cas des bruits réels de la base de données.

Les résultats de test des mesures (LLR, SNRseg, WSS, PESQ) dans ce cas sont présentés dans le tableau (3.3). Les deux premières lignes D(blanc) et R(blanc) du tableau donnent les résultats des mesures du signal dégradé par un bruit blanc D(blanc), et celui du signal rehaussé dans ce cas R(blanc). Les lignes suivantes correspondent respectivement aux résultats dans le cas d'un bruit de voiture, bruit de train, bruit babble, bruit dans un restaurant, bruit dans une station de train, bruit à l'aéroport, bruit d'exhibition et bruit dans une rue. Chaque valeur du tableau correspond à une moyenne effectuée sur 30 phrases différentes de la base de données présentant les mêmes caractéristiques (même type de bruit, même niveau de bruit).

	0 dB				5 dB			
	LLR	**SNR**	**WSS**	**PESQ**	**LLR**	**SNR**	**WSS**	**PESQ**
D(blanc)	1.80	-5.08	82.69	1.539	1.54	-2.33	69.97	1.79
R(blanc)	0.70	3.17	68.79	2.345	0.58	4.86	57.47	2.625
D(voiture)	1.01	-4.96	66.94	1.634	0.79	-2.17	54.08	1.891
R(voiture)	0.88	1.96	67.24	2.073	0.75	3.62	54.39	2.350
D(train)	1.18	-4.50	60.25	1.60	0.99	-1.69	48.12	1.859
R(train)	0.74	2.45	61.49	2.20	0.61	4.36	48.39	2.515
D(babble)	0.89	-4.63	70.35	1.705	0.71	-1.78	56.02	2.006
R(babble)	1.03	1.58	69.57	2.068	0.86	3.34	55.45	2.352
D(restaurant)	0.84	-4.19	66.42	1.754	0.68	-1.39	53.60	2.001
R(restaurant)	1.01	1.86	65.96	2.093	0.84	3.62	52.64	2.354
D(station de train)	0.94	-4.71	69.05	1.665	0.73	-1.89	54.67	1.958
R(station de train)	0.98	1.86	67.94	2.062	0.84	3.47	54.19	2.366
D(aéroport)	0.86	-4.41	71.52	1.726	0.69	-1.67	56.05	2.021
R(aéroport)	1.06	1.85	69.09	2.051	0.89	3.52	54.73	2.338
D(exhibition)	1.20	-4.67	63.52	1.585	0.94	-1.84	51.93	1.882
R(exhibition)	0.84	2.61	62.75	2.194	0.71	4.34	50.72	2.463
D(rue)	0.99	-4.26	63.34	1.563	0.80	-1.58	50.04	1.904
R(rue)	0.90	2.54	62.07	2.194	0.73	4.06	49.77	2.448

Tableau 3.3 : Résultats de test pour la modélisation dans le cas d'un bruit blanc.

D'après les résultats du tableau (3.3), on observe une amélioration en termes de mesures (LLR, SNRseg, WSS, PESQ) pour les deux valeurs de rapport signal sur bruit 0 dB et 5 dB. Les tests d'écoute et les formes d'ondes obtenues pour chaque phrase confirment qu'il y a une amélioration de la qualité de la parole rehaussée, un niveau de bruit résiduel faible et une intelligibilité acceptable.

On remarque aussi, que les mesures obtenues dans le cas du bruit blanc sont nettement supérieures aux mesures obtenues dans le cas des bruits réels, car les calculs de base du filtre de Kalman sont basés sur l'hypothèse que le bruit additif est un bruit blanc. Pour les bruits colorés réels, une modélisation du bruit en plus que celle de la parole et plus que nécessaire.

3.6.2 Avec modèle de bruit

Les résultats des tests (LLR, SNRseg, WSS, PESQ) sont présentés dans les tableaux suivants, pour le cas d'un bruit de voiture (Tableau 3.4), bruit de train (Tableau 3.5), bruit dans une station de train (Tableau 3.6), bruit de rue (Tableau 3.7), bruit à l'aéroport (Tableau 3.8), bruit au restaurant (Tableau 3.9), bruit d'exhibition (Tableau 3.10) et bruit babble (Tableau 3.11) successivement. Dans ce cas, tous ces résultats sont obtenues avec une modélisation du signal de la parole propre d'ordre $p = 10$, une durée de trame d'analyse de 20 ms et une modélisation du bruit réel avec un ordre q variable : 2, 4, 6, 8, 10.

	0dB				5dB			
	LLR	**SNR**	**WSS**	**PESQ**	**LLR**	**SNR**	**WSS**	**PESQ**
Deg	1.01	-4.96	66.94	1.634	0.79	-2.17	54.08	1.891
2	0.61	2.26	66.60	2.130	0.51	3.88	53.80	2.398
4	0.56	2.35	65.03	2.162	0.47	3.99	52.71	2.435
6	0.55	2.34	64.39	2.154	0.46	3.99	52.06	2.431
8	0.53	2.30	64.01	2.149	0.45	3.96	51.59	2.430
10	0.53	2.28	64.36	2.146	0.44	3.93	51.78	2.425

Tableau 3.4 : Résultats de test pour un bruit de voiture.

	0 dB				5 dB			
	LLR	**SNR**	**WSS**	**PESQ**	**LLR**	**SNR**	**WSS**	**PESQ**
Deg	1.18	-4.50	60.25	1.605	0.99	-1.69	48.12	1.859
2	0.62	2.14	61.92	2.194	0.52	4.34	48.64	2.502
4	0.54	2.60	61.15	2.229	0.45	4.48	48.32	2.534
6	0.54	2.66	59.73	2.209	0.46	4.53	47.29	2.517
8	0.53	2.67	58.85	2.205	0.45	4.54	46.45	2.517
10	0.53	2.67	58.70	2.201	0.44	4.53	46.24	2.514

Tableau 3.5 : Résultats de test pour un bruit de train.

	0dB				5dB			
	LLR	**SNR**	**WSS**	**PESQ**	**LLR**	**SNR**	**WSS**	**PESQ**
Deg	0.94	-4.71	69.05	1.665	0.73	-1.89	54.67	1.958
2	0.56	2.34	65.90	2.166	0.46	3.98	53.01	2.437
4	0.54	2.35	65.06	2.174	0.44	4.01	52.23	2.453
6	0.54	2.35	64.78	2.170	0.44	4.01	51.74	2.443
8	0.53	2.34	64.03	2.169	0.43	3.99	51.17	2.438
10	0.52	2.32	63.84	2.165	0.43	3.97	51.11	2.437

Tableau 3.6 : Résultats de test pour un bruit dans une station de train.

	0dB				5dB			
	LLR	**SNR**	**WSS**	**PESQ**	**LLR**	**SNR**	**WSS**	**PESQ**
Deg	0.99	-4.26	63.34	1.563	0.80	-1.58	50.04	1.904
2	0.57	2.81	60.96	2.220	0.46	4.36	49.19	2.483
4	0.54	2.86	59.54	2.242	0.45	4.38	48.61	2.493
6	0.54	2.85	59.63	2.239	0.45	4.36	48.47	2.485
8	0.52	2.84	58.83	2.237	0.44	4.34	47.73	2.482
10	0.51	2.83	58.56	2.235	0.44	4.32	47.72	2.478

Tableau 3.7 : Résultats de test pour un bruit de rue.

Les deux mesures qui ont plus de corrélation avec les tests subjectifs et les tests d'écoute sont le PESQ et le SNRseg. A partir des tableaux (3.4 - 3.7) on observe que les bruits (de voiture, train, station et rue) peuvent être bien modélisés par un modèle AR d'ordre q = 4. Pour cette valeur les meilleurs résultats sont obtenus et le spectre peut être bien modélisé par un processus autorégressif d'ordre q = 4.

	0dB				5dB			
	LLR	**SNR**	**WSS**	**PESQ**	**LLR**	**SNR**	**WSS**	**PESQ**
Deg	0.86	-4.41	71.52	1.726	0.69	-1.67	56.05	2.021
2	0.57	2.58	65.25	2.177	0.46	4.19	52.49	2.457
4	0.52	2.63	63.84	2.196	0.44	4.26	51.35	2.473
6	0.52	2.64	63.24	2.195	0.43	4.27	50.62	2.472
8	0.51	2.63	62.63	2.196	0.42	4.26	50.09	2.469
10	0.49	2.63	62.14	2.197	0.41	4 .25	49.89	2.471

Tableau 3.8 : Résultats de test pour le bruit à l'aéroport.

A partir du tableau 3.8, on observe qu'un ordre de q = 6 assure les meilleurs résultats dans le cas du bruit à l'aéroport.

Pour le bruit au restaurant, le bruit de l'exhibition et le bruit babble (les tableaux 3.9 – 3.11), un ordre plus élevé pour la modélisation par rapport aux autres bruits est justifié par le fait que ce bruit à la forme d'un signal de parole. Un ordre q = 8 donne les meilleurs résultats.

Le cas où q = 1 correspond aux résultats obtenus dans la section précédente (sans modèle de bruit).

	0dB				5dB			
	LLR	**SNR**	**WSS**	**PESQ**	**LLR**	**SNR**	**WSS**	**PESQ**
Deg	0.84	-4.19	66.42	1.754	0.68	-1.39	53.60	2 .001
2	0.56	2.55	63.62	2.197	0.47	4.31	51.05	2.471
4	0.54	2.63	62.31	2.222	0.45	4.38	50.11	2.488
6	0.53	2.69	60.72	2.229	0.44	4.44	48.89	2.494
8	0.52	2.70	59.83	2.238	0.43	4.45	48.25	2.496
10	0.51	2.69	59.51	2.235	0.43	4.44	48.04	2.491

Tableau 3.9 : Résultats de test pour le bruit au restaurant.

	0dB				5dB			
	LLR	**SNR**	**WSS**	**PESQ**	**LLR**	**SNR**	**WSS**	**PESQ**
Deg	1.20	-4.67	63.52	1.585	0.94	-1.84	51.93	1.882
2	0.60	2.65	62.05	2.208	0.51	4.41	50.19	2.479
4	0.56	2.75	60.68	2.245	0.48	4.52	49.05	2.512
6	0.56	2.82	59.67	2.252	0.47	4.59	48.14	2.522
8	0.54	2.87	58.04	2.262	0.46	4.62	46.63	2.530
10	0.53	2.88	57.73	2.262	0.45	4.62	46.45	2.528

Tableau 3.10 : Résultats de test pour le bruit de l'exhibition.

	0dB				5dB			
	LLR	**SNR**	**WSS**	**PESQ**	**LLR**	**SNR**	**WSS**	**PESQ**
Deg	0.89	-4.63	70.35	1.705	0.72	-1.78	56.02	2.006
2	0.58	2.17	67.16	2.150	0.47	3.90	54.15	2.432
4	0.57	2.24	65.77	2.164	0.45	3.98	52.91	2.449
6	0.56	2.30	64.58	2.165	0.44	4.04	51.66	2.457
8	0.54	2.31	63.43	2.166	0.44	4.04	50.98	2.460
10	0.53	2.29	63.47	2.164	0.43	4.02	50.79	2.455

Tableau 3.11 : Résultats de test pour un bruit babble.

Les remarques précédentes concernant le rehaussement de la parole avec modélisation des bruits réels sont valables quelque soit l'origine des paramètres AR de la parole (soit directement à partir de la parole propre pour comparer entre les variantes des filtres, ou bien de la parole bruitée.

Les figures (3.5) et (3.6) illustrent les formes d'ondes de la parole propre, bruitée et rehaussée et leurs spectrogrammes de la 25ème phrase de la base de données (A good book informs of what we ought to know), bruitée par un bruit babble à 0 dB et rehaussée par une modélisation d'ordre q = 2 et q=8 respectivement.

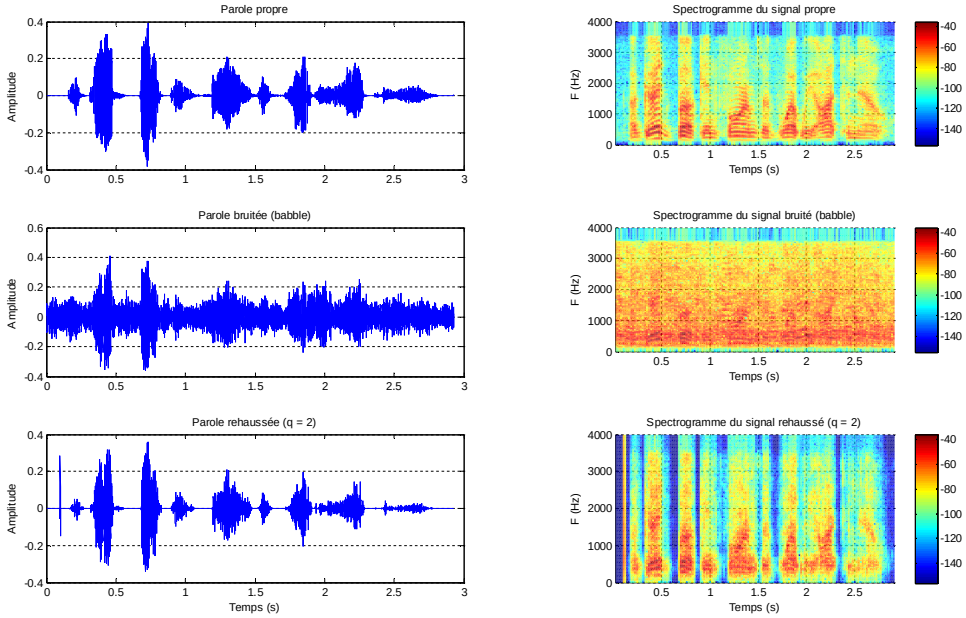


Figure 3.5 : Formes d'ondes et spectrogrammes, cas d'un bruit babble à 0 dB et une modélisation d'ordre q = 2.

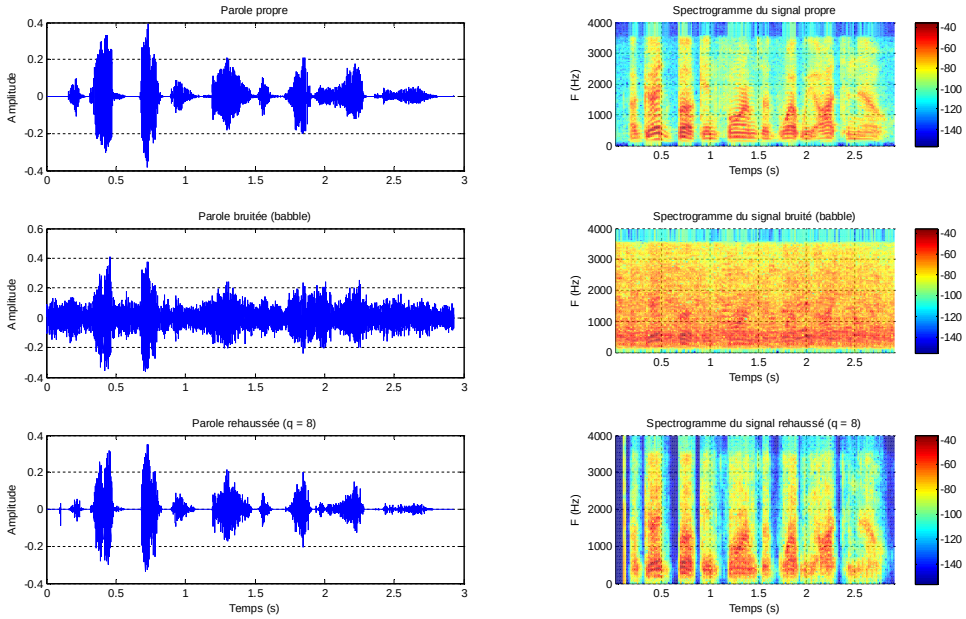


Figure 3.6 : Formes d'ondes et spectrogrammes, cas d'un bruit babble à 0 dB et une modélisation d'ordre q = 8.

3.7 Algorithme EM

3.7.1 Principe

L'algorithme Espérance - Maximisation (en Anglais Expectation-Maximization algorithm), souvent abrégé EM, proposé par Dempster et al. (1977), garantit la convergence vers un estimateur du maximum de vraisemblance (ML) de tous les paramètres inconnus, ou à la limite vers un maximum local de la fonction de vraisemblance, où chaque itération augmente la probabilité d'une meilleure estimation des paramètres [81]-[83].

On utilise souvent Espérance-maximisation pour la classification de données, en apprentissage machine, la suppression de bruit ou en reconnaissance de formes. Espérance-maximisation alterne des étapes d'évaluation de l'espérance (E), où l'on calcule l'espérance de la vraisemblance en tenant compte des dernières variables observées, et une étape de maximisation (M), où l'on estime le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape E. On utilise ensuite les paramètres trouvés en M comme point de départ d'une nouvelle phase d'évaluation de l'espérance, et l'on itère ainsi.

L'algorithme EM est utilisé pour estimer les paramètres d'un système dynamique linéaire, connu aussi sous le nom de modèle espace-état linéaire Gaussien**.** Ce système peut être décrit par les deux équations suivantes [82] :

$$\begin{cases} x_t = \Phi x_{t-1} + w_t, & t = 1, 2, \cdots, N \\ y_t = h x_t + v_t, & t = 1, 2, \cdots, N \end{cases} \tag{3.61}$$

Où la sortie y_t est une fonction linéaire de l'état x_t, et l'état actuel dépend linéairement de l'état précédent. Les deux bruits du processus et de mesure, w_t et v_t, sont des variables aléatoires distribuées selon la loi normale, de moyennes nulles et de matrices de covariance Q et R, respectivement. Seule la sortie du système est observée, l'état et les paramètres des bruits doivent être estimés.

Au lieu de considérer l'état comme une valeur déterministe affectée par un bruit aléatoire, nous combinons la variable d'état et le bruit du processus en une seule variable aléatoire Gaussienne, nous formons une combinaison similaire pour la sortie. En se basant sur l'équation (3.61), nous pouvons écrire les densités conditionnelles de l'état et de la mesure comme suit,

$$P\left(y_t / x_t\right) = \exp\left\{-\frac{1}{2}\left[y_t - h x_t\right]^T R^{-1}\left[y_t - h x_t\right]\right\}\left(2\pi\right)^{-p/2}\left|R\right|^{-1/2} \tag{3.62}$$

$$P\left(x_t / x_{t-1}\right) = \exp\left\{-\frac{1}{2}\left[x_t - \Phi x_{t-1}\right]^T Q^{-1}\left[x_t - \Phi x_{t-1}\right]\right\}\left(2\pi\right)^{-k/2}\left|Q\right|^{-1/2} \tag{3.63}$$

Une séquence de N vecteurs de sortie $\left(y_1, y_2, \cdots, y_N\right)$ est désignée par $\left\{y\right\}$, et la séquence de N vecteurs d'état $\left(x_1, x_2, \cdots, x_N\right)$ est désignée par $\left\{x\right\}$.

Par la propriété de Markov implicite dans ce modèle,

$$P\left(\left\{x\right\}, \left\{y\right\}\right) = P\left(x_0\right)\prod_{t=1}^{N} P\left(x_t / x_{t-1}\right)\prod_{t=1}^{N} P\left(y_t / x_t\right) \tag{3.64}$$

Supposant que l'état initial suit une loi normale :

$$P(x_0) = \exp\left\{-\frac{1}{2}[x_0-\mu]^T \Sigma^{-1}[x_0-\mu]\right\}(2\pi)^{-k/2}|\Sigma|^{-1/2} \tag{3.65}$$

Par conséquent, le log de la probabilité conjointe est la somme suivante :

$$\begin{aligned}\log P(\{x\},\{y\}) = & -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_0-\mu)^T \Sigma^{-1}(x_0-\mu) - \frac{N(p+k)}{2}\log(2\pi) \\ & -\frac{N}{2}\log|Q| - \frac{1}{2}\sum_{t=1}^{N}(x_t-\Phi x_{t-1})^T Q^{-1}(x_t-\Phi x_{t-1}) \\ & -\frac{N}{2}\log|R| - \frac{1}{2}\sum_{t=1}^{N}(y_t-hx_t)^T R^{-1}(y_t-hx_t)\end{aligned} \tag{3.66}$$

Où $\log P(\{x\},\{y\})$ doit être maximisé en fonction des paramètres $\mu, \Sigma, \Phi, Q,$ et R. Comme la fonction log-vraisemblance donnée ci-dessus dépend de la série des données non observées $x_t,\ t = 0,1,\ldots,N,$ l'algorithme EM peut être appliqué conditionnellement en fonction des données observées $(y_1, y_2, \cdots, y_N)$. Donc, les paramètres estimés à l'étape (r+1) seront les valeurs de μ, Σ, Φ, Q, R qui maximisent :

$$G(\mu,\Sigma,\Phi,Q,R) = E_r\left(\log P(\{x\},\{y\}) / y_1,\ldots,y_N\right) \tag{3.67}$$

Où E_r désigne la moyenne conditionnelle relative à une densité contenant les valeurs itérées à l'itération (r) $\mu(r), \Sigma(r), \Phi(r), Q(r)$ et $R(r)$.

Une procédure itérative définie comme une séquence de telles étapes peut donner des vraisemblances non décroissantes.

L'espérance conditionnelle conduit à :

$$\begin{aligned}G(\mu,\Sigma,\Phi,Q,R) = & -\frac{1}{2}\log|\Sigma| - \frac{1}{2}tr\left\{\Sigma^{-1}\left(P_0^N + (x_0^N-\mu)(x_0^N-\mu)^T\right)\right\} \\ & -\frac{N(p+k)}{2}\log 2\pi \\ & -\frac{N}{2}\log|Q| - \frac{1}{2}tr\left\{Q^{-1}\left(C - B\Phi^T - \Phi B^T + \Phi A\Phi^T\right)\right\} \\ & -\frac{N}{2}\log|R| - \frac{1}{2}tr\left\{R^{-1}\sum_{t=1}^{N}\left[(y_t-hx_t^N)(y_t-hx_t^N)^T + hP_t^N h^T\right]\right\}\end{aligned} \tag{3.68}$$

Où tr désigne la trace et A, B et C seront présentés dans la section suivante.

Les termes du filtre de Kalman $x_t^N, P_t^N,$ et $P_{t,t-1}^N$ sont calculés avec les valeurs des paramètres $\mu(r), \Phi(r), Q(r), R(r)$ utilisant les récursivités avant et arrière du filtre de Kalman à l'étape E.

1- The E step:

À partir des estimations initiales suivantes $\mu(0), \Phi(0), Q(0)$, et $R(0)$, calculer $x_t^N = E(x_t / y_1, \ldots, y_N), P_t^N = \mathrm{cov}(x_t / y_1, \ldots, y_N)$ avec les récursivités avant du filtre de Kalman suivantes :

Pour $t = 1, \ldots, N$

$$x_t^{t-1} = \Phi_t x_{t-1}^{t-1} \tag{3.69}$$

$$p_t^{t-1} = \Phi_t P_{t-1}^{t-1} \Phi_t^T + Q_t \tag{3.70}$$

$$K_t = P_t^{t-1} h^T \left(h P_t^{t-1} h^T + R_t \right)^{-1} \tag{3.71}$$

$$x_t^t = x_t^{t-1} + K_t \left(y_t - h x_t^{t-1} \right) \tag{3.72}$$

$$P_t^t = P_t^{t-1} - K_t h P_t^{t-1} \tag{3.73}$$

Où on prend $x_0^0 = \mu$ et $P_0^0 = \sum$.

Afin de calculer les valeurs lissées de x_t^N et P_t^N, on effectue une série des récursivités arrières pour $t = N, N-1, \cdots, 1$ sur les équations :

$$J_{t-1} = P_{t-1}^{t-1} \Phi_t^T \left(P_t^{t-1} \right)^{-1} \tag{3.74}$$

$$x_{t-1}^N = x_{t-1}^{t-1} + J_{t-1} \left(x_t^N - \Phi_t x_{t-1}^{t-1} \right) \tag{3.75}$$

$$P_{t-1}^N = P_{t-1}^{t-1} + J_{t-1} \left(P_t^N - P_t^{t-1} \right) J_{t-1}^T \tag{3.76}$$

Nous avons également besoin de la covariance $P_{t,t-1}^N = \mathrm{cov}(x_t, x_{t-1} / y_1, y_2, \cdots, y_N)$ qui peut être obtenu grâce à la récursivité arrière (backward recursions).

Pour $t = N, N-1, \cdots, 2$

$$P_{t-1,t-2}^N = P_{t-1}^{t-1} J_{t-2}^T + J_{t-1} \left(p_{t,t-1}^N - \Phi_t P_{t-1}^{t-1} \right) J_{t-2}^T \tag{3.77}$$

Qui est initialisée par : $P_{N,N-1}^N = (I - K_N h) \Phi_N P_{N-1}^{N-1}$.

2- Estimer $\mu_0(1) = x_0^N$ et fixer la valeur de Σ à un niveau raisonnable. Calculer $\Phi(1), Q(1),$ et $R(1)$ respectivement avec les équations suivantes :

$$\Phi(r+1) = BA^{-1} \tag{3.78}$$

$$Q(r+1) = \frac{1}{N} \left(C - BA^{-1} B^T \right) \tag{3.79}$$

Et

$$R(r+1) = \frac{1}{N} \sum_{t=1}^{N} \left[\left(y_t - h x_t^N \right) \left(y_t - h x_t^N \right)^T + h P_t^N h^T \right] \tag{3.80}$$

$$A = \sum_{t=1}^{N} \left(P_{t-1}^N + x_{t-1}^N x_{t-1}^{N\,T} \right) \tag{3.81}$$

$$B = \sum_{t=1}^{N} \left(P_{t,t-1}^N + x_t^N x_{t-1}^{N\,T} \right) \tag{3.82}$$

$$C = \sum_{t=1}^{N} \left(P_t^N + x_t^N x_t^{NT} \right) \tag{3.83}$$

3- Répété les deux étapes 1 et 2 précédentes jusqu'à ce que les estimations et la fonction log-vraisemblance sont stables. Généralement, on utilise la forme simplifiée proposée par Gupta et Mehra en 1974 suivante :

$$\log L = -\frac{1}{2}\sum_{t=1}^{N}\log\left|hP_t^{t-1}h^T + R_t\right| - \frac{1}{2}\sum_{t=1}^{N}\left(y_t - hx_t^{t-1}\right)^T\left(hP_t^{t-1}h^T + R_t\right)^{-1}\left(y_t - hx_t^{t-1}\right) \quad (3.84)$$

3.7.2 Application de l'algorithme EM au rehaussement de la parole

Les méthodes de rehaussement de la parole basées sur le filtre de Kalman nécessitent l'estimation des paramètres AR du signal de la parole et ceux de bruits, pour former la matrice de transition F à chaque trame ou la stationnarité est vérifiée. En plus, les matrices de covariances Q et R des deux bruits d'observation et de mesure doivent être disponibles aussi.
Pour tester la robustesse du filtre de Kalman vis-à-vis la modélisation et le rehaussement dans le cas d'un bruit blanc ou un bruit coloré, on utilise généralement une analyse LPC sur le fichier de la parole propre. Par conséquent, la matrice F et la covariance Q sont très proches des valeurs réelles et la valeur de R est obtenue par les méthodes d'estimation du bruit. Cependant, dans les situations réelles comme dans les communications mobiles, on a une seule information à l'entrée de notre système qui est l'observation bruitée (y). A partir de cette information et en appliquant l'algorithme EM, on peut estimer tous ces paramètres tout en rehaussant le signal.

3.7.3 Résultats de l'algorithme EM

L'application du filtre de Kalman au rehaussement de la parole avec l'utilisation de l'algorithme EM, pour l'estimation des différents paramètres directement à partir de l'observation bruitée a été implémentée sous Matlab. Dans ce cas, une modélisation AR d'ordre p = 10 est utilisée. Ainsi, à chaque trame d'analyse de durée 10 ms, la matrice de transition (F) et les matrices de covariances Q et R sont mises à jour durant l'étape de maximisation de l'algorithme EM. Ces paramètres seront utilisés pour affiner l'estimation du vecteur d'état dans l'étape d'espérance de l'algorithme EM. Ces deux étapes sont itérées plusieurs fois jusqu'à ce que la convergence soit atteinte ou le nombre maximal d'itération est atteint.

Pour toutes les simulations effectuées dans cette section, l'estimation initiale de la matrice de transition (F) utilise les 10 coefficients de l'analyse LPC appliquée sur la première trame de l'observation bruitée, les deux matrices de covariances Q et R sont initialisées en se basant sur la variance de l'erreur de prédiction de la même analyse LPC appliquée à la première trame. Le vecteur d'état x est initialisé par un vecteur aléatoire Gaussien de dimension $(p \times 1)$ et sa matrice de covariance est initialisée par une matrice aléatoire Gaussienne de dimension $(p \times p)$.

Afin d'étudier l'influence du nombre d'itérations sur les performances de l'algorithme itératif EM, on a fait tourner l'algorithme pour plusieurs itérations ($k = 1$, 2, 3 et 4) et pour chaque cas on va identifier l'itération maximale la plus adaptée. De plus, un autre paramètre essentiel dans ce cas, qui peut influencer les résultats obtenus est la trame d'analyse. Cette taille est liée à la stationnarité du signal de la parole où nous avons choisi deux taille différentes 10 ms (N = 80 échantillons) et 20 ms (N = 160 échantillons).

Les résultats des tests des mesures (LLR, SNRseg, WSS, PESQ) sont présentés dans le tableau (3.12) pour la trame d'analyse de 10 ms et dans le tableau (3.13) pour le cas de 20 ms. Ces tests sont effectués en utilisant la $25^{ème}$ phrase de la base de données (A good book informs of what we ought to know) bruitée par un bruit blanc avec deux rapports signal sur bruit de 0 dB et 5 dB.

	0dB				5dB			
	LLR	**SNRseg**	**WSS**	**PESQ**	**LLR**	**SNRseg**	**WSS**	**PESQ**
Dégradé	0.8114	-2.3028	33.5856	0.8576	0.6845	-0.8600	31.5813	0.9476
k=1	0.5763	-0.0680	**31.4260**	1.0142	0.4717	0.9299	**28.9386**	1.1743
k=2	**0.5754**	**0.1780**	32.9666	**1.0454**	**0.4701**	**1.2760**	29.1371	**1.2064**
k=3	0.6356	**0.1886**	36.4669	0.9974	0.4855	**1.3258**	30.0620	1.1998
k=4	0.6623	0.0653	40.1783	0.9513	0.5083	1.2708	31.5128	1.1745

Tableau 3.12 : Résultats de tests avec l'algorithme EM, N = 80 (10 ms).

	0dB				5dB			
	LLR	**SNRseg**	**WSS**	**PESQ**	**LLR**	**SNRseg**	**WSS**	**PESQ**
Dégradé	0.8114	-2.3028	33.5856	0.8576	0.6845	-0.8600	31.5813	0.9476
k=1	0.5865	-0.3807	**31.2644**	0.9479	0.4894	0.5910	29.4031	1.1270
k=2	0.5760	-0.0012	31.5850	1.0217	0.4716	1.0541	**29.0400**	1.1967
K=3	**0.5713**	0.1290	32.4546	1.0426	**0.4699**	1.2393	29.3258	1.2211
k=4	0.5747	**0.1920**	33.8280	**1.0469**	0.4754	**1.3254**	29.7961	**1.2282**

Tableau 3.13 : Résultats de tests avec l'algorithme EM, N = 160 (20 ms).

D'après les tests d'écoute, les formes d'ondes et les résultats des tests objectifs des deux tableaux (3.12) et (3.13), on peut conclure que l'utilisation du filtre de Kalman avec l'algorithme EM permet d'avoir une qualité acceptable de la parole rehaussée.

De plus, les résultats obtenus dans le cas du SNR = 5 dB sont meilleurs par rapport à ceux du SNR = 0 dB, car le niveau du bruit dans le premier cas est faible par rapport au deuxième cas.

En se basant sur les deux mesures objectives (SNRseg et PESQ), les résultats des deux tableaux considérés nous permet de conclure que : plus le nombre d'itérations augmente, les résultats s'améliorent d'avantage, une taille réduite permet d'avoir de bon résultats à l'itération k = 2 mais avec un temps de calcul élevé et pour une taille de 20 ms, un nombre d'itérations k = 4 est suffisant dans ce cas avec un temps de calcul moins par rapport au précédent.

La figure (3.7) illustre un exemple d'un signal de parole bruitée par un bruit blanc additif avec SNR = 5 dB, les formes d'ondes obtenues après chaque itération et ses spectrogrammes dans le cas N = 80 échantillons (10 ms). La figure (3.8) montre les mêmes résultats dans le cas d'un SNR = 0 dB.

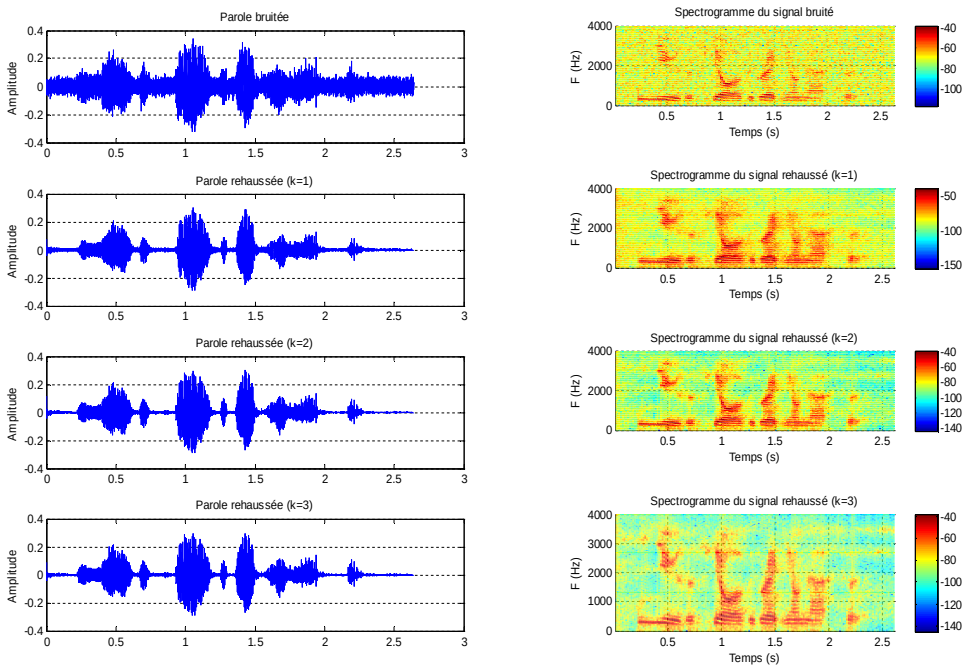


Figure 3.7 : Formes d'ondes et spectrogrammes de la parole bruitée et rehaussée (k=1,2 et 3), cas d'un bruit blanc et un SNR = 5 dB.

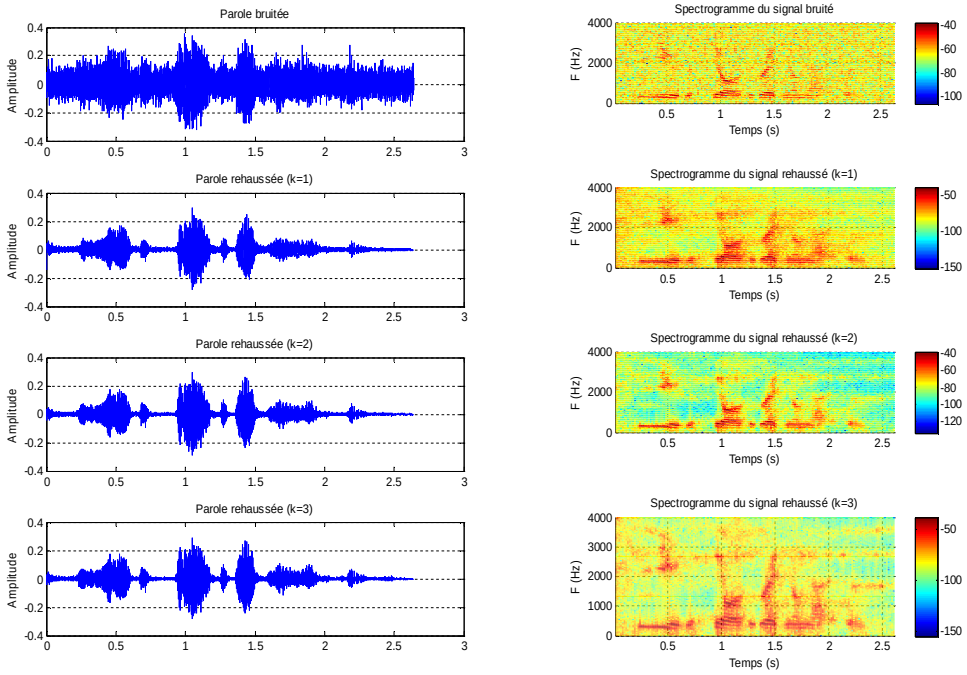


Figure 3.8 : Formes d'ondes et spectrogrammes de la parole bruitée et rehaussée (k=1,2 et 3), cas d'un bruit blanc et un SNR = 0 dB.

3.8 Conclusion

Ce chapitre nous a permet d'introduire le filtre de Kalman conventionnel, ses équations, ses étapes de l'initialisation en passant par l'étape de propagation et enfin l'étape de la correction. Par la suite, l'application du filtre de Kalman au rehaussement de la parole a été présentée dans le cas d'un bruit blanc et les bruits réels colorés.

Les mesures objectives calculées, les tests d'écoute et les formes d'ondes présentés dans ce chapitre, confirment que le filtre de Kalman est un outil efficace pour le rehaussement de la parole même avec des corpus de parole, sous différents types de bruits réels et à des rapports signal sur bruit différents.

Les mesures obtenues dans le cas du bruit blanc sont nettement supérieures aux mesures obtenues dans le cas des bruits réels, car les calculs de base du filtre de Kalman sont basés sur l'hypothèse que le bruit additif est un bruit blanc.

En outre, pour un bruit coloré réel, en plus de la modélisation de la parole, une modélisation du bruit avec un modèle AR d'ordre q = 4 (bruit de voiture, bruit de train, bruit dans une station de train, bruit dans une rue), q = 6 (bruit à l'aéroport) et q = 8 (bruit au restaurant, bruit d'exhibition, bruit babble) est nécessaire pour assurer des meilleures performances.

Quelques résultats de l'application de l'algorithme EM au rehaussement de la parole sont présentés dans le cas d'un bruit blanc additif, pour la détermination des paramètres AR à partir du signal bruité. Les résultats obtenus dans le cas du SNR = 5 dB sont meilleurs par rapport à ceux du SNR = 0 dB, car le niveau du bruit dans le premier cas est faible par rapport au deuxième cas. Par ailleurs, plus le nombre d'itérations augmente, les résultats s'améliorent d'avantage, une taille de 10 ms permet d'avoir de bon résultats à l'itération k = 2 mais avec un temps de calcul élevé et pour une taille de 20 ms, un nombre d'itérations k = 4 est suffisant dans ce cas avec un temps de calcul moins par rapport au précédent.

D'autres travaux sur des versions non linéaires du filtre de Kalman ont été réalisés mais ne sont pas présentés dans ce chapitre. Dans [39], l'estimation des trajectoires des paramètres LSF par le filtre de Kalman non parfumé (Unscented Kalman Filter) [85]-[88] a été étudiée, où nous avons présenté les résultats obtenus.

CHAPITRE 4

APPROCHES BAYESIENNES DE REHAUSSEMENT DE LA PAROLE

4.1 Introduction

Dans ce chapitre nous fournirons les détails des approches Bayesiennes de rehaussement de la parole qui seront intégrées aux blocs de réduction de bruit des codeurs à bas débit. Ces méthodes assurent une réduction du bruit musical meilleure que les méthodes basées sur la soustraction spectrale, un temps de calcul réduit et une implémentation en temps réel plus simple que les méthodes basées sur le filtre de Kalman et ses variantes. On distingue deux classes : les estimateurs basés sur des modèles Gaussiens où les coefficients de la DFT du signal de la parole ont été supposés comme des variables aléatoires Gaussiennes et indépendantes [20][21], et les estimateurs basés sur des modèles super-Gaussiens où les histogrammes mesurés des coefficients de la DFT du signal de la parole et de l'amplitude de ces coefficients ont montré que les coefficients de la DFT peuvent être mieux modélisés en utilisant des densités de probabilité super-Gaussiennes [27][28].

4.2 Estimateurs basés sur des modèles Gaussiens

Ephraim et Malah ont introduit différentes règles de suppression que nous allons présentées dans ce chapitre. Ces règles ont été construites à partir de l'hypothèse que les transformées de Fourier à court terme des signaux sources et bruit sont des variables aléatoires Gaussiennes et indépendantes, comme présenté dans le premier chapitre.
L'estimateur de l'amplitude spectrale est conçu pour minimiser l'erreur quadratique moyenne (EQM) sur le spectre d'amplitude. Cet estimateur est couramment désigné comme l'estimateur d'amplitude spectrale à court terme au sens de l'erreur quadratique moyenne, (MMSE-STSA : Minimum Mean Square Error – Short Time Spectral Amplitude) [20].
Par la suite Ephraim et Malah ont élaboré d'autres règles de suppression à partir des mêmes hypothèses. L'estimateur présenté minimise l'EQM du logarithme du spectre (MMSE-log STSA : Minimum Mean Square Error – log Spectral Amplitude estimator) [21]. De plus, Malah et al. ont proposé un estimateur correspondant à une modification du MMSE-LSA appelé Multiplicatively-modified LSA [89]. L'ensemble de ces règles conduit à un rehaussement du signal de parole très satisfaisant aussi bien du point de vue de la réduction du bruit que vis-à-vis du bruit musical et seront détaillées dans les sections suivantes.

4.2.1 Estimateur d'amplitude spectrale MMSE-STSA

4.2.1.1 Principe de l'estimateur MMSE-STSA

Considérons un signal d'observation, y, composé d'un signal de parole x, corrompu par un bruit additif d, pour chaque indice temporel discret n, le signal d'observation bruité y, est donné par :

$$y(n) = x(n) + d(n) \tag{4.1}$$

Soient $X_k = A_k\ e^{j\alpha_k}$, D_k et $Y_k = R_k\ e^{j\theta_k}$ représentent les spectres à court terme du signal propre $x(n)$, du bruit $d(n)$ et du signal bruité $y(n)$ respectivement.

Le but de l'algorithme MMSE-STSA est d'estimer le module $\hat{A}_k$ de chaque coefficient complexe de Fourier du signal de parole, à partir du signal bruité. Sous l'hypothèse que la parole et le bruit sont Gaussiennes et ses composantes spectrales sont statistiquement indépendantes, l'estimateur d'amplitude spectrale au sens du MMSE de A_k peut être dérivé de Y_k comme suit :

$$\hat{A}_k = E\left\{A_k \left| Y_0,\ Y_1,\ ...\right.\right\} \tag{4.2}$$

D'après la théorie d'estimation (chapitre 1), l'estimateur au sens du MMSE est toujours la moyenne de la densité a posteriori (la moyenne conditionnelle). Donc, $\hat{A}_k$ est la moyenne conditionnelle de A_k sachant l'observation Y_k, et elle est écrite comme suit :

$$\hat{A}_k = \mathrm{E}\left\{\mathrm{A}_k \left| Y_k \right.\right\} = \frac{\int_0^{\infty}\int_0^{2\pi} a_k\ p\left(Y_k \left| a_k, \alpha_k\right.\right)\ p\left(a_k, \alpha_k\right)\ d\alpha_k da_k}{\int_0^{\infty}\int_0^{2\pi} p\left(Y_k \left| a_k, \alpha_k\right.\right)\ p\left(a_k, \alpha_k\right)\ d\alpha_k\ da_k} \tag{4.3}$$

Où $E\{\ .\ \}$ est l'opérateur d'espérance et $p(\ .\ )$ est la densité de probabilité.

Sous le modèle statistique considéré, $p\left(Y_k \left| a_k, \alpha_k\right.\right)$ et $p\left(a_k, \alpha_k\right)$ sont données par :

$$p\left(Y_k \left| a_k, \alpha_k \right.\right) = \frac{1}{\pi\lambda_d(k)} \exp\left\{ -\frac{1}{\lambda_d(k)} \left| Y_k - a_k e^{j\alpha_k} \right|^2 \right\} \tag{4.4}$$

$$p\left(a_k, \alpha_k\right) = \frac{a_k}{\pi\lambda_x(k)} \exp\left\{ -\frac{a_k^2}{\lambda_x(k)} \right\} \tag{4.5}$$

Où $\lambda_x(k) = E\left\{ \left|X_k\right|^2 \right\}$ et $\lambda_d(k) = E\left\{ \left|D_k\right|^2 \right\}$ désignent les variances de X_k et D_k respectivement.
On introduit (4.4) et (4.5) dans (4.3), on obtient [20] :

$$\hat{A}_k = \Gamma(1.5)\frac{\sqrt{\nu_k}}{\gamma_k} \cdot M(-0.5;\ 1;\ -\nu_k) \cdot R_k = \Gamma(1.5)\frac{\sqrt{\nu_k}}{\gamma_k} \exp\left(-\frac{\nu_k}{2}\right) \cdot \left[(1+\nu_k) \cdot I_0\left(\frac{\nu_k}{2}\right) + \nu_k \cdot I_1\left(\frac{\nu_k}{2}\right) \right] \cdot R_k \tag{4.6}$$

Où $\Gamma(.)$ est la fonction gamma, avec $\Gamma(1.5) = \frac{\sqrt{\pi}}{2}$; $M(a\,;c\,;x)$ est la fonction hypergéométrique confluente ; $I_0(.)$ et $I_1(.)$ sont les fonctions de Bessel modifiées d'ordre 0 et du premier ordre respectivement, définies par :

$$I_n(z) = \frac{1}{2\pi} \int_0^{2\pi} \cos(\beta n) \cdot \exp(z\cos\beta)\, d\beta \tag{4.7}$$

Dans l'équation (4.6) ν_k est défini par :

$$\nu_k = \frac{\xi_k}{1+\xi_k} \cdot \gamma_k \tag{4.8}$$

Où l'on rappelle que le SNR a priori (ξ_k) et le SNR a posteriori (γ_k) sont donnés par :

$$\xi_k = \frac{\lambda_x(k)}{\lambda_d(k)} \tag{4.9}$$

$$\gamma_k = \frac{R_k^2}{\lambda_d(k)} \tag{4.10}$$

Il est à noter que γ_k est calculé directement, par contre ξ_k est estimé par l'approche d'estimation de décision dirigée (section **2.8.2**).

4.2.1.2 Comportement de l'estimateur MMSE-STSA aux SNR élevés

Il est intéressant d'étudier le comportement asymptotique de l'estimation $\hat{A}_k$ dans le cas où le rapport signal sur bruit est élevé, c'est-à-dire quand $\xi_k >> 1$.
Considérons la distribution exponentielle du ν_k :

$$p(\nu_k) = \frac{1}{\xi_k} \exp\left(\frac{-\nu_k}{\xi_k}\right) \tag{4.11}$$

Il est simple de remarquer que quand $\xi_k >> 1$, la valeur de $\nu_k >> 1$ avec une probabilité élevée. Donc, pour étudier $\hat{A}_k$ dans le cas où $\xi_k >> 1$, la fonction hypergéométrique confluente est approximée comme suit :

$$M\left(-0.5\ ;\ 1\ ;\ -\nu_k\right)=\frac{\sqrt{\nu_k}}{\Gamma(1.5)}\ ,\qquad \nu_k >>1 \tag{4.12}$$

On obtient :

$$\hat{A}_k=\frac{\xi_k}{1+\xi_k}\cdot R_k=A_k^w\ ,\qquad \xi_k >>1 \tag{4.13}$$

Cet estimateur est équivalent à l'estimateur d'amplitude de Wiener qui ne dépend que du SNR a priori. L'estimation de l'amplitude spectrale du signal rehaussé $\hat{X}_k$ est alors simplement déduite à partir de Y_k selon la formule :

$$\hat{X}_k=\frac{\xi_k}{1+\xi_k}\cdot Y_k=X_k^w\ ,\quad \xi_k >>1 \tag{4.14}$$

4.2.1.3 Fonction du gain

Il est plus pratique de considérer que l'estimateur d'amplitude $\hat{A}_k$ dans (4.6) est obtenu à partir de R_k, par une fonction multiplicative non linéaire du gain, définie par :

$$G_{MMSE}\left(\xi_k,\gamma_k\right)=\frac{\hat{A}_k}{R_k} \tag{4.15}$$

A partir de l'équation (4.6), on observe que cette fonction du gain ne dépend que du SNR a priori et a posteriori ξ_k et γ_k, respectivement.

La figure (4.1) présente l'évolution du gain spectral G_{MMSE} résultant des équations (4.6) et (4.13) en fonction du SNR instantané (γ_k-1).

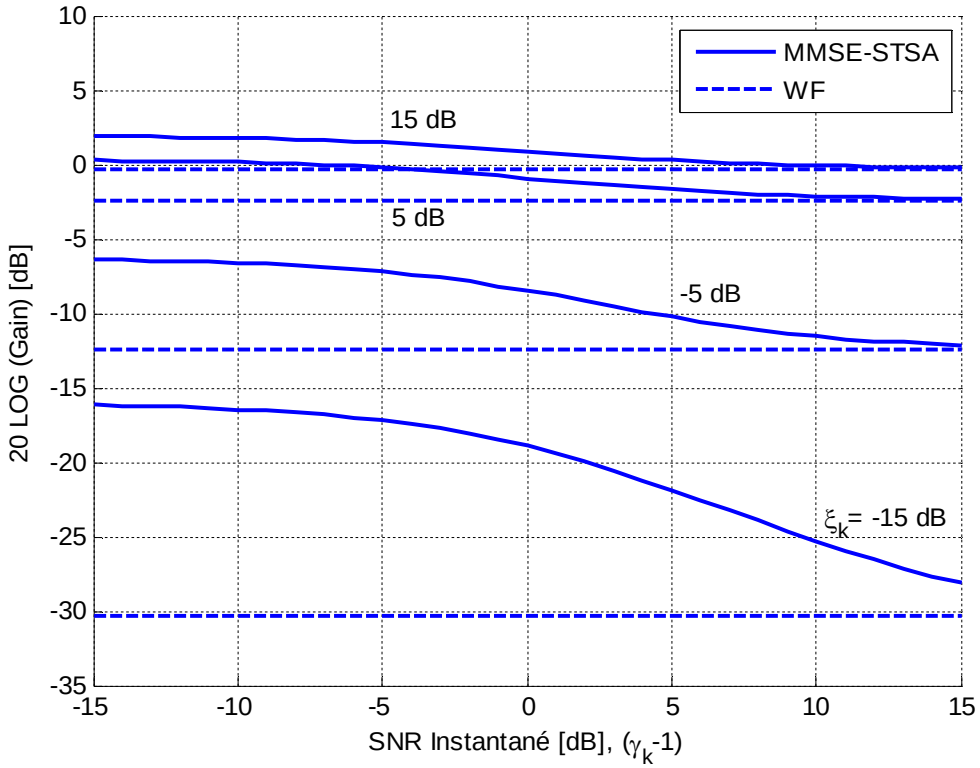


Figure 4.1 : Courbes du gain du MMSE STSA et du filtre de Wiener.

Quand le SNR a priori est constant, les courbes du gain dans la figure (4.1) montrent une augmentation du gain quand le SNR instantané décroît.

Cette figure illustre de plus, plusieurs courbes du gain correspondant à la fonction du gain de Wiener donnée par :

$$G_w\left(\xi_k,\gamma_k\right)=\frac{\xi_k}{1+\xi_k} \tag{4.16}$$

Qui est indépendante de γ_k .

La convergence de l'estimateur MMSE-STSA vers l'estimateur de Wiener à des SNRs élevés est clairement démontrée dans la figure (4.1).

4.2.2 Estimateur d'amplitude sous l'incertitude de la présence du signal

L'estimateur au sens du MMSE, qui tient compte de l'incertitude de la présence du signal dans les observations bruitées, a été développé par Middleton et Esposito (voir détection et estimation conjointe du chapitre 1).
Comme la parole dans l'observation bruitée est soit présente ou absente, deux hypothèses peuvent être considérées :

L'hypothèse nulle, H_0^k : Parole absente : $\left|R_k\right|=\left|D_k\right|$
L'hypothèse alternative, H_1^k : Parole présente : $\left|R_k\right|=\left|A_k+D_k\right|$

En se basant sur ces deux hypothèses, l'équation (4.3), peut être réécrite sous une forme plus explicite comme suit :

$$\hat{A}_k=E\left\{A_k\left|Y_k,H_1^k\right.\right\}P\left(H_1^k\left|Y_k\right.\right)+E\left\{A_k\left|Y_k,H_0^k\right.\right\}P\left(H_0^k\left|Y_k\right.\right) \tag{4.17}$$

Où $P\left(H_i^k\left|Y_k\right.\right),\left(\text{pour } i=0,1\right)$, est la probabilité que la parole est dans l'état H_i^k, pour la composante spectrale k, sachant l'observation bruitée Y_k.
Comme $E\left\{A_k\left|Y_k,H_0^k\right.\right\}$ est nulle, donc, l'équation (4.17) peut être simplifiée comme suit :

$$\hat{A}_k=E\left\{A_k\left|Y_k,H_1^k\right.\right\}P\left(H_1^k\left|Y_k\right.\right) \tag{4.18}$$

Notant que $E\left\{A_k\left|Y_k,H_1^k\right.\right\}$ remplace $E\left\{A_k\left|Y_k\right.\right\}$, comme dans l'équation (4.3) et $P\left(H_1^k\left|Y_k\right.\right)$ définit la modification multiplicative de l'estimateur optimal sous l'hypothèse de la présence de la parole. Exploitant la règle de Bayes, on obtient :

$$P\left(H_1^k\left|Y_k\right.\right)=\frac{\Lambda\left(Y_k,q_k\right)}{1+\Lambda\left(Y_k,q_k\right)}=G_M\left(k\right) \tag{4.19}$$

L'estimateur devient :

$$\hat{A}_k=G_M\left(k\right)E\left\{A_k\left|Y_k,H_k^1\right.\right\} \tag{4.20}$$

Où: $\Lambda\left(Y_k,q_k\right)$ est le rapport de vraisemblance généralisé défini par :

$$\Lambda\left(Y_k, q_k\right) = \mu_k \frac{P\left(Y_k \left| H_k^1\right.\right)}{P\left(Y_k \left| H_k^0\right.\right)} \tag{4.21}$$

Avec $\mu_k = \frac{1-q_k}{q_k}$, et q_k est la probabilité de l'absence du signal dans la k[ième] composante spectrale. Par conséquent, afin de déterminer le nouvel estimateur d'amplitude de l'équation (4.20), on a besoin de calculer le rapport $\Lambda(k)$ seulement. Ceci peut être obtenu facilement en utilisant le model statistique Gaussien considéré pour les composantes spectrales, ou d'une manière équivalente, en utilisant les équations (4.4) et (4.5), alors :

$$\Lambda\left(Y_k, q_k\right) = \mu_k \frac{\exp\left(\nu_k\right)}{1+\xi_k} \tag{4.22}$$

Où: ξ_k est maintenant défini par:

$$\xi_k = \frac{E\left\{A_k^2 \left| H_k^1\right.\right\}}{\lambda_d(k)} \tag{4.23}$$

Cette définition est identique à celle dans (4.9) où le signal est implicitement considéré sûrement présent dans les composantes spectrales bruitées.

Il est plus commode d'exprimer $\Lambda\left(Y_k, q_k\right)$ et l'estimateur d'amplitude résultant en fonction de $\eta_k = E\left\{A_k^2\right\} / \lambda_d(k)$ qui est plus facile à estimer que ξ_k. η_k est lié à ξ_k par :

$$\eta_k = \frac{E\left\{A_k^2\right\}}{\lambda_d(k)} = \left(1-q_k\right)\frac{E\left\{A_k^2 \left| H_k^1\right.\right\}}{\lambda_d(k)} = \left(1-q_k\right)\xi_k \tag{4.24}$$

Ainsi, en considérant $\Lambda\left(Y_k, q_k\right)$ dans l'équation (4.22) comme $\Lambda\left(\xi_k, \gamma_k, q_k\right)$, et utilisant $E\left\{A_k \left| y_k, H_k^1\right.\right\} = G_{MMSE}\left(\xi_k, \gamma_k\right) R_k$, où $G_{MMSE}\left(\xi_k, \gamma_k\right)$ est la fonction du gain définie par (4.6) et (4.15), l'estimateur d'amplitude de l'équation (4.20) peut être écrit comme suit :

$$\begin{aligned}\hat{A}_k &= \frac{\Lambda\left(\xi_k, \gamma_k, q_k\right)}{1+\Lambda\left(\xi_k, \gamma_k, q_k\right)} G_{MMSE}\left(\xi_k, \gamma_k\right) R_k \Big|_{\xi_k = \eta_k / \left(1-q_k\right)} \\ &= G_{MM}\left(\xi_k, \gamma_k, q_k\right) G_{MMSE}\left(\xi_k, \gamma_k\right) R_k = G_{MMSE}^D\left(\xi_k, \gamma_k, q_k\right) R_k\end{aligned} \tag{4.25}$$

Si $q_k = 0$ alors $\frac{\Lambda\left(\xi_k, \gamma_k, q_k\right)}{1+\Lambda\left(\xi_k, \gamma_k, q_k\right)} = 1$, et aussi $\eta_k = \xi_k$. Dans ce cas $G_{MMSE}^D\left(\xi_k, \gamma_k, q_k\right)$ est identique à $G_{MMSE}\left(\xi_k, \gamma_k\right)$. Ainsi, l'estimateur d'amplitude de l'équation (4.6) peut être considéré comme un cas particulier de l'estimateur d'amplitude dans (4.25). Les courbes de gain résultants du G_{MMSE}^D dans l'équation (4.25) sont présentés dans la figure (4.2) pour $q_k = 0.2$.

Il est intéressant de comparer ces courbes du gain de G_{MMSE}^D à celles du G_{MMSE}. Particulièrement dans le cas où le SNR a priori est élevé (5dB et 15dB dans la figure), on remarque une diminution du gain G_{MMSE}^D ($q_k > 0$) lorsque le SNR a posteriori γ_k diminue, par contre dans le cas

du $G_{MMSE}(q_k = 0)$ on a une augmentation du gain dans ces conditions. Ceci est probablement le résultat du fait de favoriser l'hypothèse d'absence du signal par l'estimateur d'amplitude de l'équation (4.25) dans de telles situations.

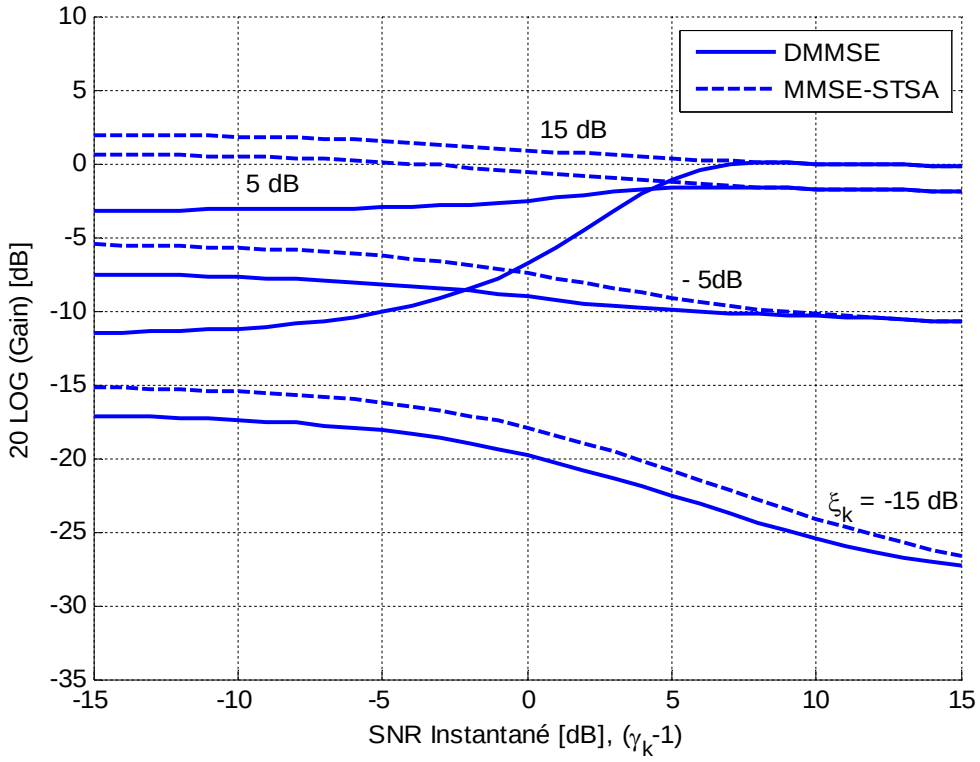


Figure 4.2 : Courbes du G_{MMSE} et G^{D}_{MMSE} en fonction du SNR instantané.

4.2.3 Estimateur d'amplitude MMSE-log STSA

Une année après la proposition de l'estimateur MMSE-STSA, un autre estimateur basé sur le logarithme du spectre pour le rehaussement de la parole a été proposé par les mêmes auteurs [21]. La large utilisation du logarithme du spectre dans les mesures de distorsion, est la motivation principale de l'étude d'un estimateur d'amplitude basé sur la minimisation de l'erreur quadratique moyenne du logarithme du spectre (MMSE-log-STSA).

Avec les définitions données dans l'équation (4.1) et ses transformées de Fourier respectives, l'estimateur $\hat{A}_k$ du logarithme de A_k, est donné par :

$$\hat{A}_k = E\left\{ \left(\log A_k - \log \hat{A}_k\right)^2 \middle| Y_k \right\} \tag{4.26}$$

Par conséquent, l'estimateur est :

$$\hat{A}_k = \exp\left\{ E\left[\ln A_k \middle| Y_k \right] \right\}, \quad 0 \le k \le N-1 \tag{4.27}$$

Et qui est indépendant de la base choisie du logarithme. L'évaluation de $E\left[\ln A_k \middle| Y_k\right]$, peut être simplifiée en utilisant la fonction génératrice des moments $\Phi_{\ln A_k | Y_k}(\mu_k)$.

On pose : $Z_k = \ln A_k$.

$$\Phi_{Z_k|Y_k}(\mu_k) = E\left\{\exp(j\mu_k Z_k)\middle|Y_k\right\} = \mathrm{E}\left\{\mathrm{A}_\mathrm{k}^{j\mu_k}\middle|Y_k\right\} = \frac{\int_0^\infty\int_0^{2\pi} a_k^\mu\ p\left(Y_k\middle|a_k,\alpha_k\right)\ p\left(a_k,\alpha_k\right)\ d\alpha_k da_k}{\int_0^\infty\int_0^{2\pi} p\left(Y_k\middle|a_k,\alpha_k\right)\ p\left(a_k,\alpha_k\right)\ d\alpha_k\ da_k} \quad (4.28)$$

Sous l'hypothèse du modèle Gaussien des coefficients de la DFT, les fonctions des densités de probabilité, $p\left(Y_k\middle|a_k,\alpha_k\right)$ et $p\left(a_k,\alpha_k\right)$ sont données par :

$$p\left(Y_k\middle|a_k,\alpha_k\right) = \frac{1}{\pi\lambda_d(k)}\exp\left\{-\frac{1}{\lambda_d(k)}\left|Y_k - a_k e^{\mathrm{j}\alpha_k}\right|^2\right\} \quad (4.29)$$

$$p\left(a_k,\alpha_k\right) = \frac{a_k}{\pi\lambda_x(k)}\exp\left\{-\frac{a_k^2}{\lambda_x(k)}\right\} \quad (4.30)$$

D'après les équations (4.29) et (4.30), et en utilisant la représentation intégrale de la fonction de Bessel modifiée d'ordre 0, on peut réécrire la formule (4.28) comme suit :

$$\Phi_{Z_k|Y_k}(\mu) = \frac{\int_0^\infty a_k^{\mu+1}\exp\left(-a_k^2/\lambda_k\right)\cdot I_0\left(2a_k\sqrt{\nu_k/\lambda_k}\right)da_k}{\int_0^\infty a_k\exp\left(-a_k^2/\lambda_k\right)I_0\left(2a_k\sqrt{\nu_k/\lambda_k}\right)da_k} \quad (4.31)$$

Où λ_k satisfait la relation suivante :

$$\frac{1}{\lambda_k} = \frac{1}{\lambda_x(k)} + \frac{1}{\lambda_d(k)} \quad (4.32)$$

Et ν_k est définie par : $\nu_k = \frac{\xi_k}{1+\xi_k}\cdot\gamma_k$; $\xi_k = \frac{\lambda_x(k)}{\lambda_d(k)}$; $\gamma_k = \frac{R_k^2}{\lambda_d(k)}$

ξ_k et γ_k sont les SNR a priori et a posteriori. Le calcul des intégrales de l'équation (4.31) donne :

$$\Phi_{Z_k|Y_k}(\mu) = \lambda_k^{\mu/2}\Gamma(\mu/2+1)\cdot M(-\mu/2\ ;\ 1\ ;\ -\nu_k) \quad (4.33)$$

La dérivée première de $\Phi_{Z_k|Y_k}(\mu)$ évaluée à $\mu = 0$ génère le moment du premier ordre $E\left\{Z_k\middle|Y_k\right\}$.

$$E\left\{Z_k\middle|Y_k\right\} = E\left\{\ln A_k\middle|Y_k\right\} = \frac{d}{d\mu}\left[\Phi_{\ln A_k|Y_k}(\mu)\right]\Big|_{\mu=0} \quad (4.34)$$

L'estimateur de l'amplitude spectrale est donné par :

$$\hat{A}_k = \frac{\xi_k}{1+\xi_k}\cdot\exp\left\{\frac{1}{2}\int_{\nu_k}^\infty\frac{e^{-t}}{t}dt\right\}\cdot R_k \quad (4.35)$$

Le rapport $\frac{\hat{A}_k}{R_k}$ désigne la fonction du gain de l'estimateur MMSE-logSTSA donné par :

$$G_{LSA}(\xi_k,\gamma_k)=\frac{\hat{A}_k}{R_k}=\frac{\xi_k}{1+\xi_k}\cdot\exp\left\{\frac{1}{2}\int_{\nu_k}^{\infty}\frac{e^{-t}}{t}dt\right\} \tag{4.36}$$

Les variations de ce gain en fonction du SNR instantané (γ_k-1) sont présentées dans la figure (4.3). Cette figure contient aussi les courbes du gain de l'estimateur MMSE STSA dérivé dans la section précédente. L'explication du comportement de ces courbes reste la même pour les courbes de l'estimateur MMSE-log STSA. A noter que la nouvelle fonction du gain (l'équation 4.36), donne toujours un gain inférieur à celui du MMSE STSA, cela peut être démontré en utilisant l'inégalité de Jensen.

$$\hat{A}_k=\exp\left\{E\left[\ln A_k\,|Y_k\right]\right\}\le\exp\left\{\ln E\left[A_k\,|Y_k\right]\right\}=E\left[A_k\,|Y_k\right] \tag{4.37}$$

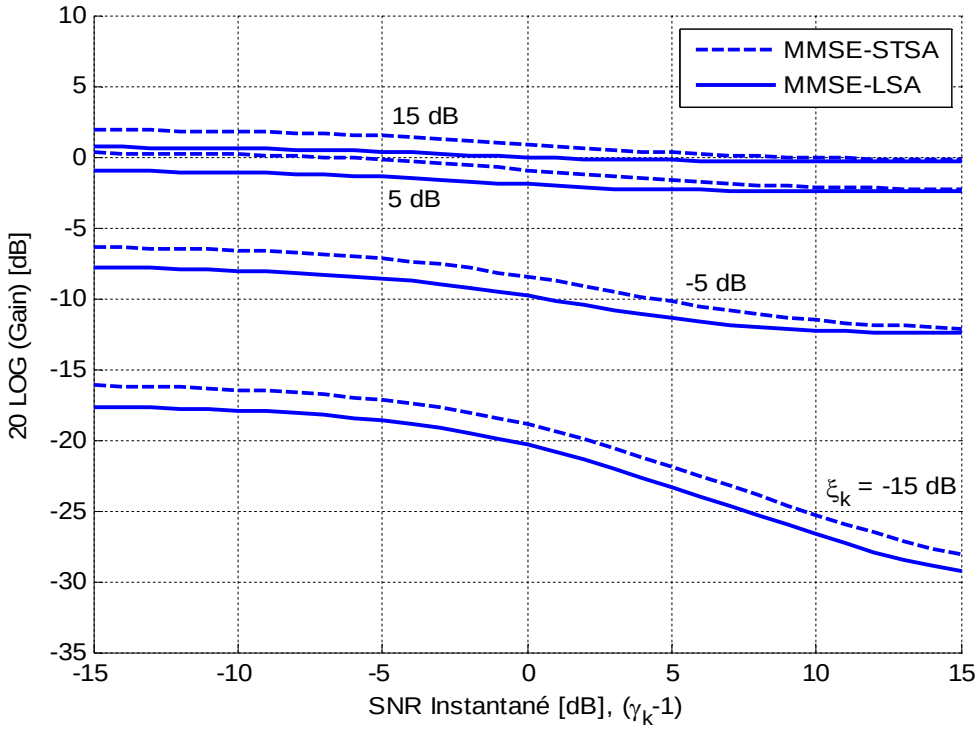


Figure 4.3 : Courbes du gain du MMSE STSA et du MMSE log STSA.

4.2.4 Estimateur MMSE-log STSA sous l'incertitude de la présence du signal

En suivant les mêmes développements pour la dérivation de l'estimateur d'amplitude sous l'incertitude de la présence du signal (section 4.2.2). D'une manière générale, notant C_k une fonction de l'amplitude spectrale à court terme (exemple : A_k, $\log A_k$, A_k^2). L'estimateur MMSE de C_k est donné par :

$$\tilde{C}_k=E\left\{C_k\,|Y_k,H_1^k\right\}P\left(H_1^k\,|Y_k\right)+E\left\{C_k\,|Y_k,H_0^k\right\}P\left(H_0^k\,|Y_k\right) \tag{4.38}$$

Quand la parole est absente $E\left\{C_k\,|Y_k,H_0^k\right\}=0$, l'équation (4.38) peut être simplifiée comme suit :

$$\tilde{C}_k=E\left\{C_k\,|Y_k,H_1^k\right\}P\left(H_1^k\,|Y_k\right) \tag{4.39}$$

Où : $P\left(H_1^k \left| Y_k \right.\right)$ est la modification « soft decision » de l'estimateur optimal sous l'hypothèse que la parole est présente, donnée par :

$$P\left(H_1^k \left| Y_k \right.\right) = \frac{\Lambda(k)}{1+\Lambda(k)} = G_M(k) \tag{4.40}$$

- **Estimateur MM-LSA**

Dans ce cas $C_k = \log A_k$ et l'estimateur d'amplitude a la forme :

$$\tilde{A}_{LSA} = \exp\left[E\left\{ \log A_k \left| Y_k, H_1^k \right.\right\} G_M(k) \right] = \left[G_{LSA}(k) R_k \right]^{G_M(k)} \tag{4.41}$$

Où $G_M(k)$ est la modification du gain définie dans l'équation (4.40). Comme la modification $G_M(k)$ dans l'équation (4.41) n'est pas multiplicative, donc pas d'améliorations significatives par rapport à l'utilisation seule de $G_{LSA}(k)$, dans [89] un estimateur de type LSA avec une modification de type multiplicatif a été proposé et désigné par MM-LSA (Multiplicatively-Modified LSA). Cet estimateur est présenté comme suit :

$$\tilde{A}_{MM} = G_M(k) G_{LSA}(k) R_k = G_{MM}(k) R_k \tag{4.42}$$

4.2.5 Estimation de la probabilité a priori

En plus de la connaissance du rapport signal sur bruit a posteriori γ_k et l'application de la méthode de "décision dirigée" pour l'estimation du rapport signal sur bruit a priori ξ_k, pour calculer la modification multiplicative $G_{MM}(k)$, une estimation fiable de la probabilité a priori de l'absence de la parole q_k (pour toutes les fréquences dans la trame bruitée en cours) est essentielle. Malah, Cox et Accardi ont proposé en 1999, un autre ensemble d'hypothèses binaires pour estimer q_k [8] :

Hypothèse nulle, H_0 : Parole présente dans la k$^{\text{ieme}}$ fréquence : $\eta_k \geq \eta_{\min}$.

Hypothèse alternative, H_1 : Parole absente dans la k$^{\text{ieme}}$ fréquence : $\eta_k \leq \eta_{\min}$.

Où $\eta_k = (1-q_k)\xi_k$, est le SNR a priori.

Sous les hypothèses précédentes, la fonction de densité de probabilité de γ_k est donnée par :

$$p(\gamma_k) = \frac{1}{1+\eta_k} \exp\left(-\frac{\gamma_k}{1+\eta_k} \right), \gamma_k \geq 0 \tag{4.43}$$

On observe que la fonction de densité de probabilité de γ_k est fonction de η_k, cela signifie que la connaissance de γ_k est suffisante pour estimer la probabilité a priori q_k, et le problème de décision précédent est équivalent à :

$$\gamma_k \underset{H_A}{\overset{H_0}{\lessgtr}} \gamma_{th} \tag{4.44}$$

Où γ_{th} est un seuil.

Basant sur l'équation (4.44) une décision binaire est attribuée à I_k ($I_k = 1$ si H_0 est rejetée et $I_k = 0$ si H_0 est acceptée) et utilisée pour calculer q_k :

$$q_k = \alpha_q \cdot q_{k-1} + \left(1 - \alpha_q\right) I_k \tag{4.45}$$

Où : $\gamma_{th} = 0.8$ et $\alpha_q = 0.98$ sont utilisés.

La figure (4.4) illustre les effets de la modification multiplicative $G_{MM}(k)$ sur l'estimateur LSA, où on remarque une diminution du gain quand q_k augmente. C'est intuitif, car l'estimateur atténue beaucoup plus le bruit quand la probabilité de l'absence de la parole est élevée pour des valeurs faibles du SNR instantané.

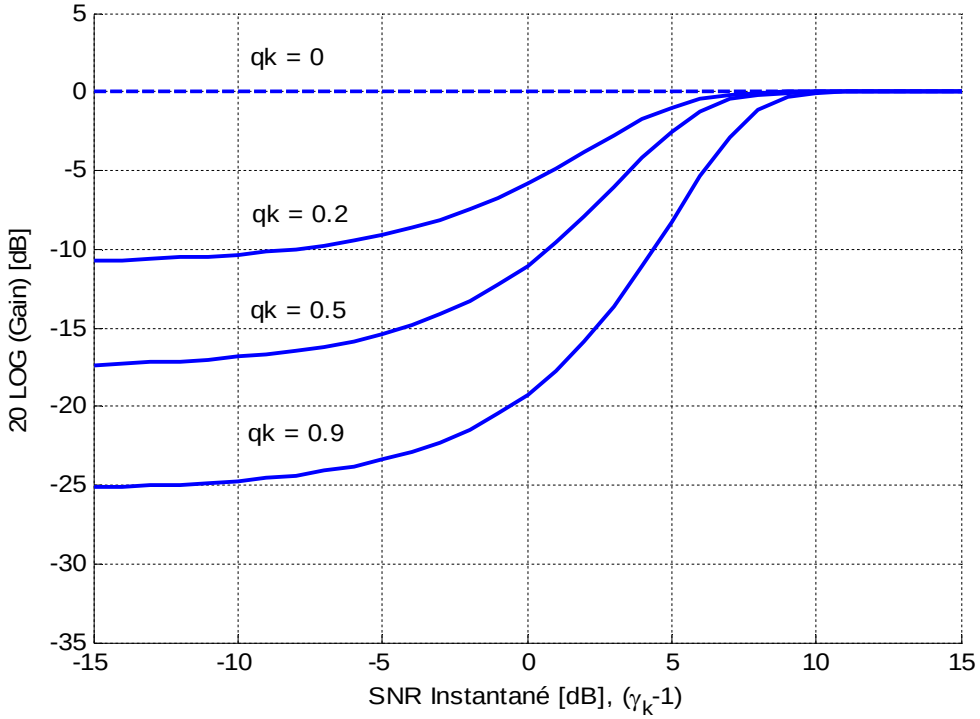


Figure 4.4 : Effet de q_k sur le gain pour ξ_k donné.

4.2.6 Résultats

4.2.6.1 Conditions d'expérimentations

Nous appliquons maintenant les différentes méthodes étudiées précédemment à des signaux prisent de la base de donnée « Noizeus », puis nous présentons quelques résultats objectifs. Ainsi, pour chaque méthode, pour un type de bruit bien défini et pour un rapport signal sur bruit désiré, toutes les trente phrases de la parole propre et les trente phrases de la parole bruitée correspondantes seront utilisées pour chaque mesure objective, où la moyenne est appliquée à la fin. Pour les différents tests, les paramètres suivants sont utilisés :

- Une fréquence d'échantillonnage de 8 kHz.

- Une fenêtre de Hanning de durée 20 ms (160 échantillons) avec un chevauchement de 50 % est utilisée durant l'étape d'analyse, et une fenêtre rectangulaire pour la synthèse avec la méthode d'addition et chevauchement (overlap and add).
- Une taille de 256 échantillons pour la DFT est utilisée pour le calcul des densités spectrales.
- Une méthode simple d'estimation continue de la densité spectrale de puissance du bruit est utilisée.
- L'approche de décision dirigée a été utilisée pour l'estimation du SNR a priori dans chaque trame, où un facteur de pondération $\alpha = 0.98$ et une valeur minimale de -25 dB du SNR a priori, permettent d'avoir un bon compromis entre la réduction du bruit et la distorsion du signal rehaussé.
- Une limite inférieure pour le gain ($G_{min} = 0.05$) a été imposée dans chaque trame, assure un niveau de bruit uniforme dans les trames de bruit seul du signal rehaussé et une qualité agréable.
- Une probabilité a priori d'absence de la parole de valeur qk = 0.3 est utilisé dans les algorithmes où la modification soft est introduite dans la formule total du gain.

4.2.6.2 Evaluation des performances

Dans cette section, nous évaluons objectivement les performances des différentes méthodes étudiées précédemment. Les résultats de test des mesures (LLR, SNR_{seg}, PESQ) sont présentés dans les tableaux suivants pour les signaux bruités par un bruit blanc et des bruits réels (bruit de voiture, bruit de train et le bruit de parole (babble)). Chaque valeur du tableau correspond à une moyenne effectuée sur 30 phrases différentes de la base de données.

	0 dB			**5 dB**		
	LLR	**SNR_{seg}**	**PESQ**	**LLR**	**SNR_{seg}**	**PESQ**
Dégradé	1.8024	-5.0828	1.5402	1.5450	-2.3266	1.7995
Wiener	1.2393	1.4570	2.2550	1.0395	3.8230	2.6078
STSA	0.9386	0.9815	2.3953	0.7538	3.2508	2.6892
STSA (SPU=1)	0.9998	2.2353	2.4966	0.8262	4.5509	2.8390
LSA	0.9884	1.5257	2.4521	0.8105	3.7964	2.7782
MM-LSA	1.0357	2.3536	2.4735	0.8721	4.6679	2.8231

Tableau 4.1 : Evaluation objective de la qualité pour un bruit blanc.

	0 dB			**5 dB**		
	LLR	**SNR_{seg}**	**PESQ**	**LLR**	**SNR_{seg}**	**PESQ**
Dégradé	1.0140	-4.9597	1.6337	0.7954	-2.1730	1.8913
Wiener	0.7376	0.9071	2.2546	0.5920	3.1959	2.6023
STSA	0.5297	0.3974	2.4025	0.4143	2.5825	2.6910
STSA (SPU=1)	0.5643	1.6765	2.4897	0.4435	3.9320	2.8235
LSA	0.5562	0.9654	2.4498	0.4451	3.1565	2.7616
MM-LSA	0.6026	1.7855	2.4619	0.4830	4.0381	2.8019

Tableau 4.2 : Evaluation objective de la qualité pour un bruit de voiture.

	0 dB			5 dB		
	LLR	**SNR_{seg}**	**PESQ**	**LLR**	**SNR_{seg}**	**PESQ**
Dégradé	1.1789	-4.5038	1.6048	0.9881	-1.6914	1.8594
Wiener	0.8211	1.2920	2.2208	0.6893	3.6683	2.5849
STSA	0.5792	0.7692	2.3970	0.4700	3.0595	2.6827
STSA (SPU=1)	0.6306	2.0574	2.4761	0.5252	4.4198	2.8231
LSA	0.6156	1.3240	2.4289	0.5126	3.6195	2.7517
MM-LSA	0.6733	2.1517	2.4470	0.5727	4.5262	2.8028

Tableau 4.3 : Evaluation objective de la qualité pour un bruit de train.

	0 dB			5 dB		
	LLR	**SNR_{seg}**	**PESQ**	**LLR**	**SNR_{seg}**	**PESQ**
Dégradé	0.8950	-4.6320	1.7054	0.7153	-1.7833	2.0061
Wiener	0.7055	0.9062	2.3336	0.5564	3.2833	2.6963
STSA	0.4986	0.4120	2.4781	0.3891	2.6672	2.7621
STSA (SPU=1)	0.5356	1.6935	2.5603	0.4161	4.0185	2.9113
LSA	0.5312	0.9879	2.5221	0.4208	3.2483	2.8366
MM-LSA	0.5728	1.7923	2.5287	0.4527	4.1254	2.8937

Tableau 4.4 : Evaluation objective de la qualité pour un bruit de parole.

4.2.6.3 Interprétations

De façon générale, pour les différentes méthodes, les types de bruit et le rapport signal sur bruit utilisé, on remarque une amélioration acceptable de la qualité en terme des mesures (LLR, SNRseg, PESQ) comparativement aux celles du signal dégradé. De plus, en termes de SNR globale, les résultats sont meilleurs pour les hauts SNR où le niveau du bruit ajouté au signal est faible.

L'utilisation de l'algorithme de Wiener avec l'approche de décision dirigée où $\alpha = 0.98$, donne un signal rehaussé avec plus de distorsion comparativement à celui obtenu par la méthode du MMSE-STSA dans les mêmes conditions. Par contre, le niveau du bruit résiduel dans l'estimation de Wiener est plus faible que celui du MMSE-STSA. Ces remarques sont confirmées par les tests d'écoute et le tableau 4.1 des mesures objectives de la qualité pour un bruit blanc sous un rapport signal sur bruit de 0 dB et 5 dB. En diminuant la valeur de α, la distorsion dans le signal rehaussé est réduite, mais le niveau du bruit résiduel augmente. La valeur de $\alpha = 0.98$ est un bon compromis pour toutes les méthodes de rehaussement.

La méthode du MMSE-STSA qui prend en compte la présence incertaine de la parole dans le signal observé, désignée par STSA (SPU=1) dans les tableaux, produit une qualité meilleure de la parole rehaussée dans ce cas, que celle obtenue avec l'estimateur STSA (SPU=0).

Spécialement avec une probabilité qk = 0.3, on obtient une réduction plus du bruit résiduel, avec des distorsions additionnelles négligeables dans le signal de la parole rehaussée.

On remarque de plus, que les deux variantes du STSA et l'estimateur de Wiener, sont presque équivalentes dans les SNR élevés. De l'autre coté, l'estimateur STSA produit de bons résultats à des SNR faibles, comparativement à l'estimateur de Wiener surtout la version avec SPU=1. L'estimateur MMSE-LSA est meilleur que celui du MMSE-STSA, parce que le niveau du bruit résiduel produit dans ce cas est plus faible, mais la qualité de la parole rehaussée est proche de celle obtenue avec le STSA (SPU=1).

Les meilleurs résultats sont obtenus sans doute avec la version **MM-LSA**. La mesure objective la plus importante et qui a plus de corrélation avec les tests subjectifs est le SNR segmental, ses valeurs dans les tableaux précédents sont les meilleures pour tous les types de bruits et pour les deux valeurs du SNR global. Les valeurs de la mesure PESQ et du SNR segmental dans ces mêmes tableaux confirment que la variante **STSA (SPU=1)** est la méthode qui assure des résultats plus proches mais moins par rapport à ceux du **MM-LSA,** mais meilleures que ceux des variantes de Wiener et **STSA (SPU=0).** Enfin, la mesure LLR est plus faible, donc meilleure dans le cas du **STSA (SPU=0)** dans tous les tableaux.

En plus des conclusions tirées des mesures objectives précédentes, les confirmations des tests d'écoutes durant nos simulations des différentes variantes des algorithmes et sous divers conditions d'expérimentations. Les représentations temporelles des formes d'ondes des signaux propres, bruités et rehaussés peuvent être utilisées pour démontrer la robustesse vis-à-vis de la réduction du bruit. Les figures suivantes illustrent les formes d'ondes des signaux propres, bruités et rehaussés par le filtre de Wiener, la méthode du MMSE-STSA avec SPU = 1 et celle du MMSE-LSA avec SPU =1 respectivement, et les spectrogrammes correspondants de la première phrase de la base de données sous un SNR = 0 dB et 5 dB. Où trois types de bruit on été sélectionnés (bruit blanc, bruit de voiture et le bruit babble).

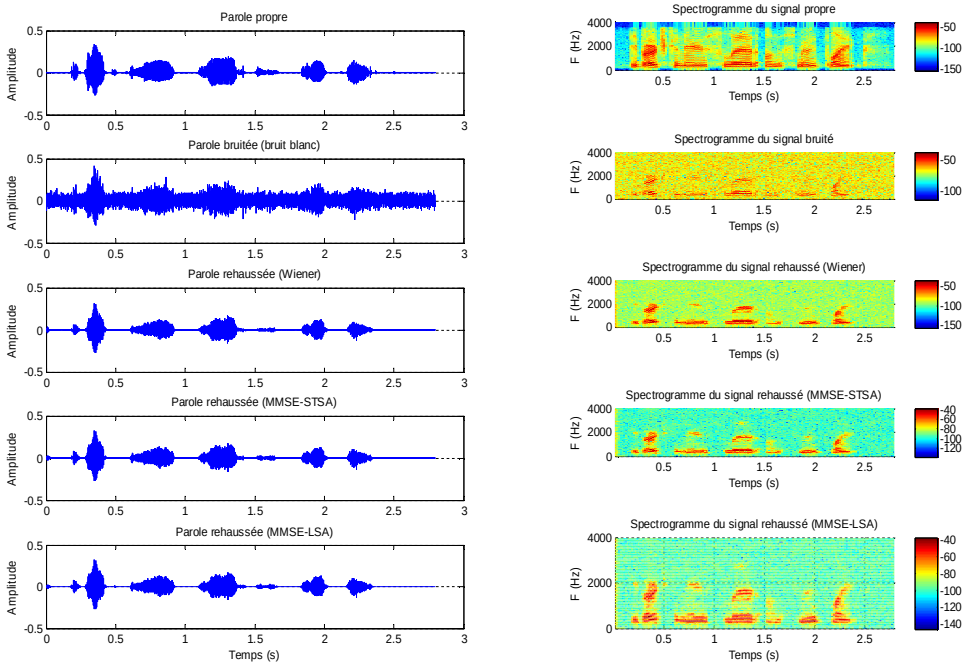


Figure 4.5.a) : Formes d'ondes et spectrogrammes, cas d'un bruit blanc et SNR = 0 dB.

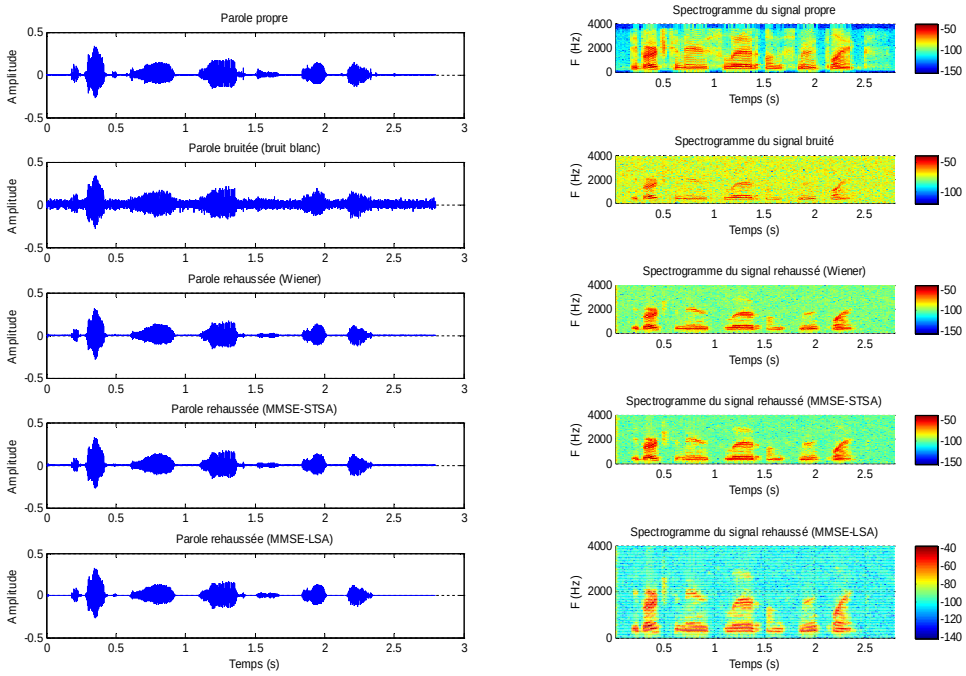


Figure 4.5.b) : Formes d'ondes et spectrogrammes, cas d'un bruit blanc et SNR = 5 dB.

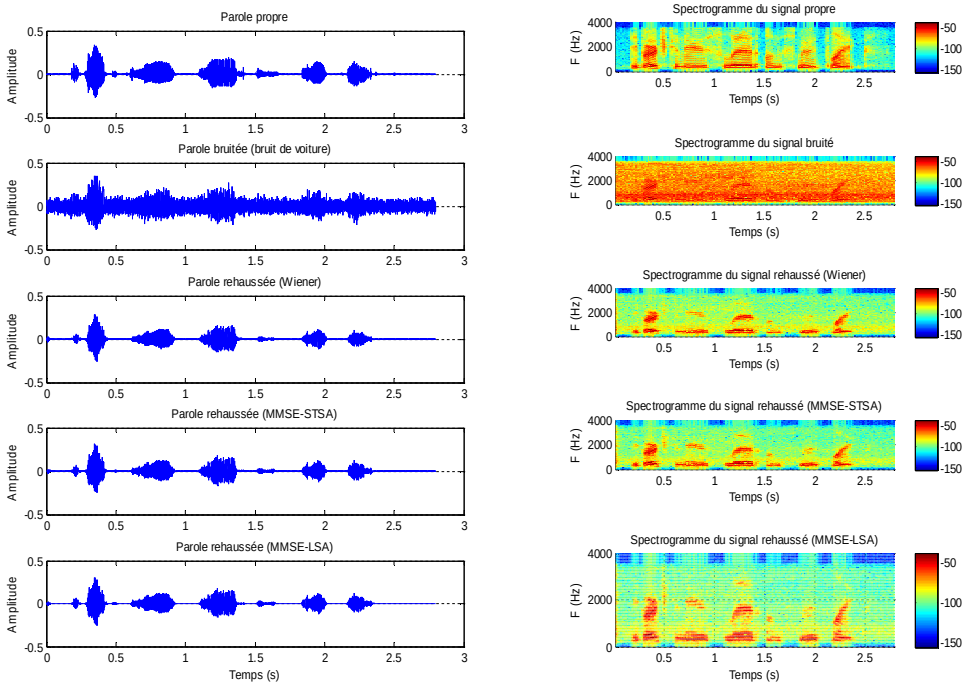


Figure 4.6.a) : Formes d'ondes et spectrogrammes, cas d'un bruit de voiture et SNR = 0 dB.

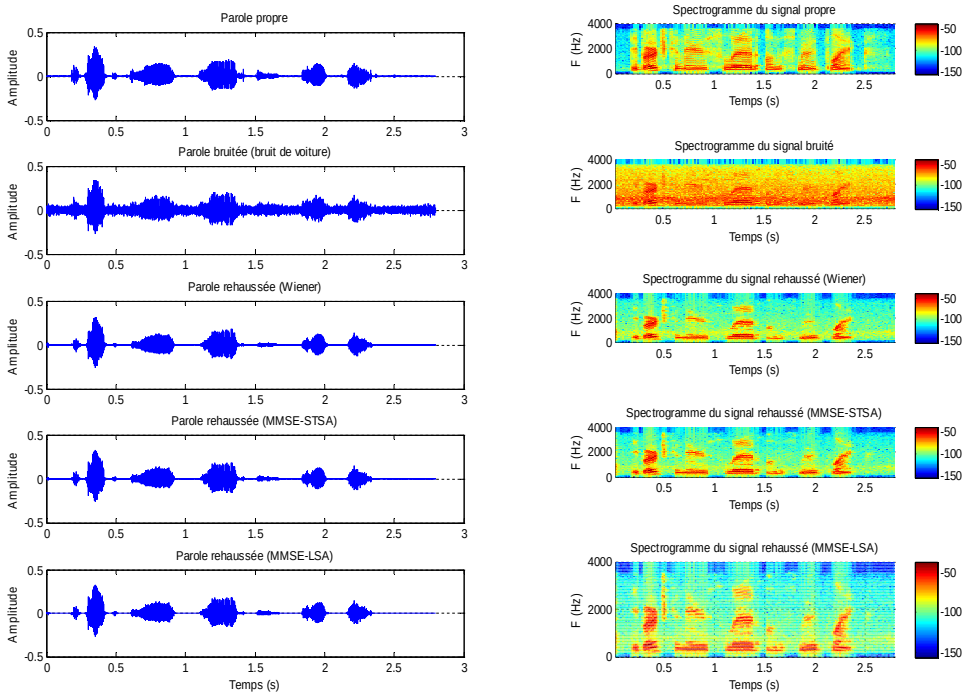


Figure 4.6.b) : Formes d'ondes et spectrogrammes, cas d'un bruit de voiture et SNR = 5 dB.

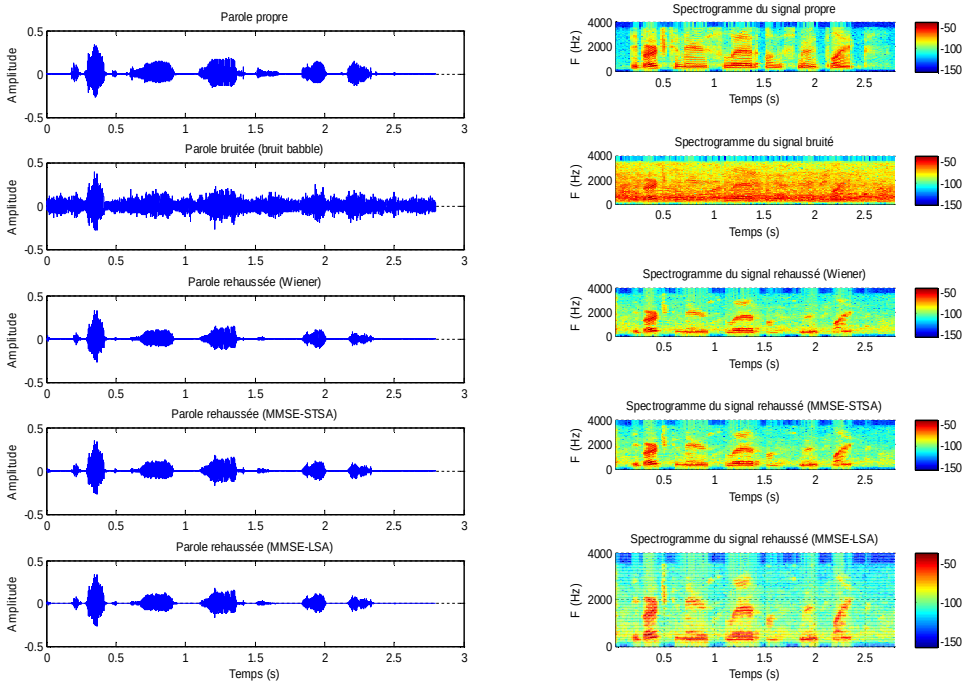


Figure 4.7.a) : Formes d'ondes et spectrogrammes, cas d'un bruit babble et SNR = 0 dB.

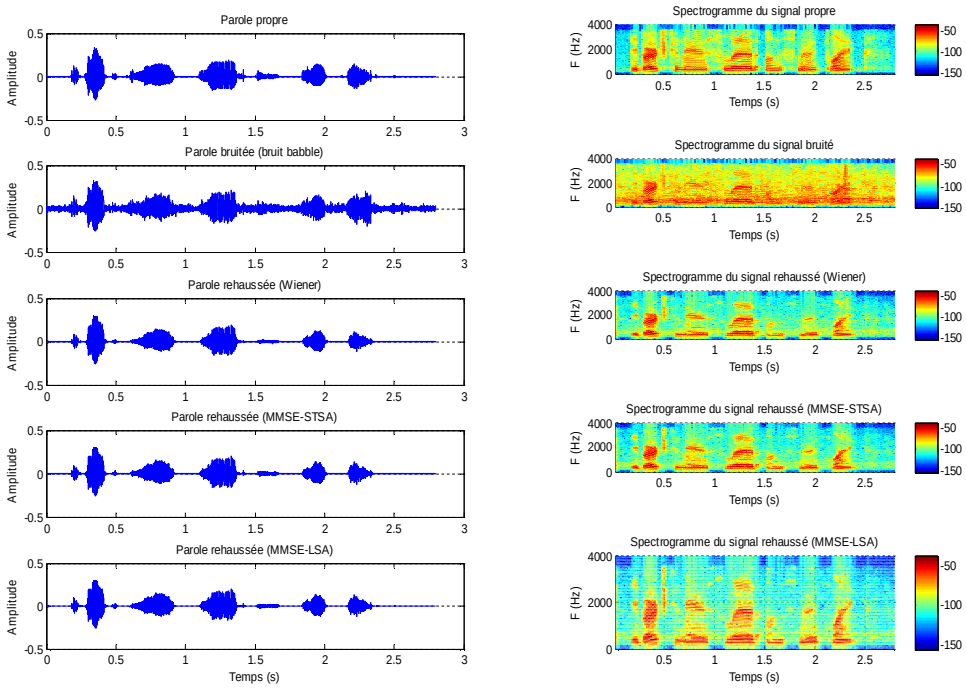


Figure 4.7.b) : Formes d'ondes et spectrogrammes, cas d'un bruit babble et SNR = 5 dB.

4.3 Estimateurs basés sur des modèles super-Gaussiens

4.3.1 Modélisation des amplitudes de la TFD du signal de parole

Les approches Bayesiennes (Wiener, MMSE-STSA, LSA, …) sont dérivées sous l'hypothèse que les coefficients de Fourier complexes de la parole et du bruit sont Gaussiens. Par conséquent, l'amplitude spectrale est modélisée par une distribution de Rayleigh. Cette supposition est vraie dans le cas asymptotique des larges trames où la durée de corrélation du signal est plus courte que la taille de la trame.

Dans le cas des trames courtes, qui est la situation la plus rencontrée, les coefficients de la TFD de la parole sont mieux modélisés par des distributions super-Gaussiennes comme : Gamma, Laplace et Gamma généralisée. La distribution de Laplace est donnée par [24] :

$$f_A(a) = \frac{1}{\sigma}\exp\left(-\frac{2a}{\sigma}\right), \quad a \geq 0 \tag{4.46}$$

Et la distribution Gamma généralisée est donnée par [27] :

$$f_A(a) = \frac{\gamma\beta^{\nu}}{\Gamma(\nu)} a^{\gamma\cdot\nu-1}\exp\left(-\beta a^{\gamma}\right), \qquad a \geq 0, \beta > 0, \gamma > 0, \nu > 0, \tag{4.47}$$

Où la variable aléatoire A représente l'amplitude spectrale et σ représente la variance du signal. La figure (4.8) présente deux exemples des densités de probabilité pour $\gamma = 1 \text{ et } \gamma = 2$, respectivement où les densités ont été normalisées à une variance unité.

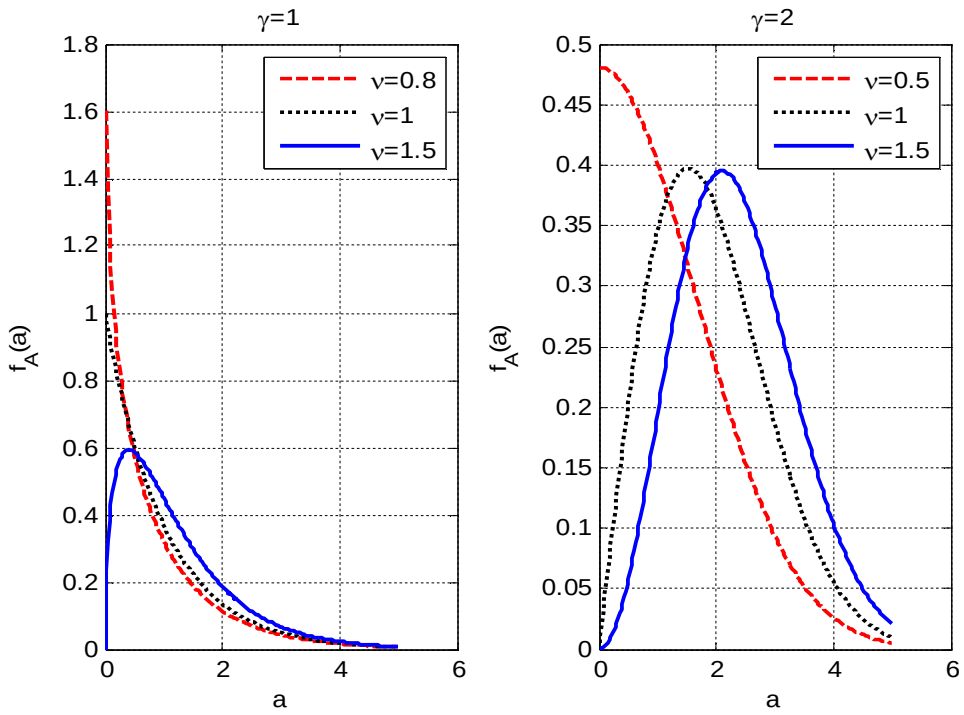


Figure 4.8 : Densités de probabilités $f_A(a)$ pour $\gamma = 1$ et $\gamma = 2$.

Comme mentionné dans le premier chapitre, les estimateurs des coefficients de la TFD s'appuyer sur un certain nombre d'hypothèse. La figure (4.9.a) présente l'histogramme de la partie réelle des coefficients de la TFD et les modèles théoriques de Gauss, Laplace et Gamma. Les coefficients de la TFD expérimentales calculés sont obtenus à partir d'un corpus de parole prononcé par trois locuteurs et trois locutrices [23]. L'histogramme complet (figure 4.9.a), ainsi que la version élargie (figure 4.9.b) montrent que la distribution de Laplace (discontinu) et Gamma (continu) modélisent bien les données expérimentales de la TFD que la distribution Gaussienne (en pointillés).

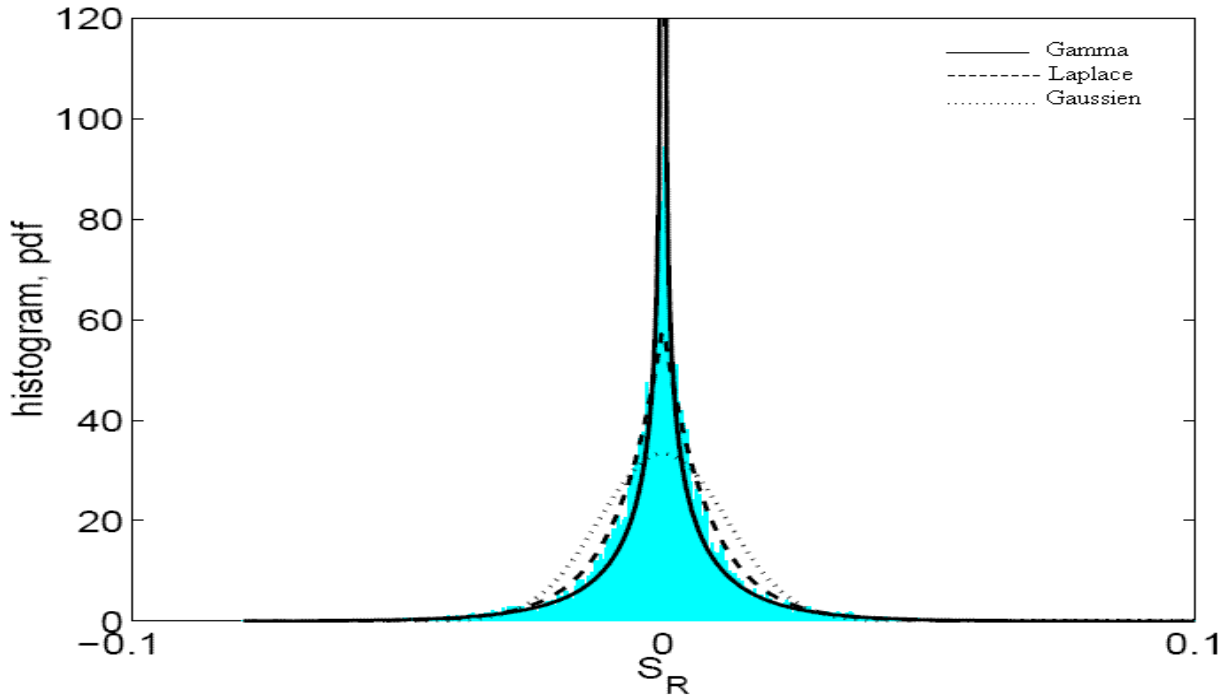


Figure 4.9.a : Histogramme des coefficients de la TFD de la parole [23].

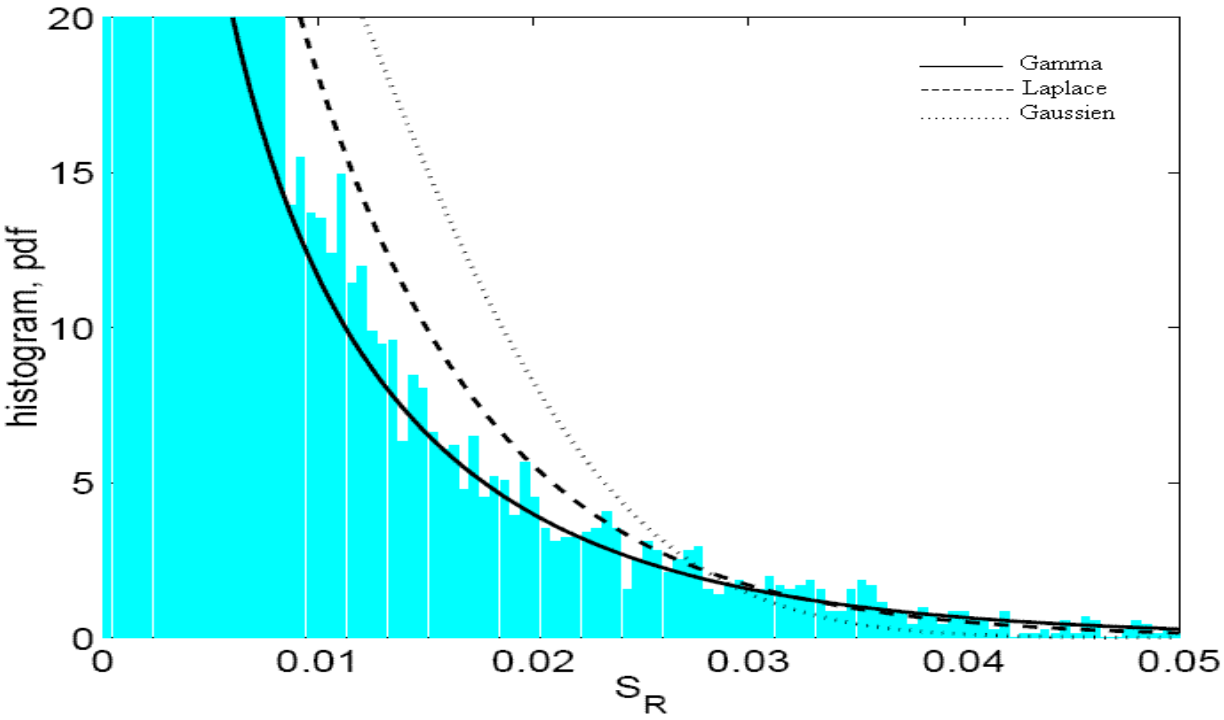


Figure 4.9.b : Histogramme des coefficients de la TFD de la parole (élargi) [23].

Récemment plusieurs estimateurs basés sur des modèles super-Gaussiens ont été proposés, les estimateurs basés sur la distribution Gamma généralisée sont les plus récents et les plus performants d'après les travaux de recherches actuels. Ces estimateurs seront développés par la suite dans ce chapitre et comparés avec les estimateurs basés sur des modèles Gaussiens.

4.3.2 Estimateur MMSE des amplitudes de la TFD du signal de parole

On a vu que l'estimateur au sens du MMSE est la moyenne de l'amplitude du signal de la parole A_k sachant l'amplitude du signal bruité R_k , $\hat{A}_k = E\{A_k | R_k\}$.

En utilisant la formule de Bayes, l'estimateur au sens du MMSE $\hat{A}_k$ est donné par :

$$\hat{A}_k = E\{A_k | R_k\} = \frac{\int_0^\infty a f_{R_k|A_k}(R_k|a) f_{A_k}(a) da}{\int_0^\infty f_{R_k|A_k}(R_k|a) f_A(a) da} \tag{4.48}$$

De plus, puisque le bruit est supposé Gaussien, $f_{R_k|A_k}(R_k|a)$ peut être écrite comme suit [17] :

$$f_{R_k|A_k}(R_k|a) = \frac{2R_k}{\lambda_d(k)} \exp\left(-\frac{R_k^2 + a^2}{\lambda_d(k)}\right) I_0\left(\frac{2aR_k}{\lambda_d(k)}\right) \tag{4.49}$$

Avec $\lambda_d(k) = E\{|D^2|\}$ est la variance du bruit et $I_0(.)$ est la fonction de Bessel modifiée de première espèce d'ordre 0.

- **Les estimateurs du premier et du deuxième ordre**

On s'intéresse à deux valeurs de γ de la distribution Gamma généralisée : $\gamma = 1$ et $\gamma = 2$. Les estimateurs d'amplitude MMSE pour la distribution de ces deux classes ont été dérivés

dans [28]. Pour le cas $\gamma = 2$ la dérivation est exacte et démontrée dans [27]. Dans le cas $\gamma = 1$ c'est impossible de faire une dérivation analytique exacte de l'estimateur d'amplitude au sens du MMSE, ce qui rend à l'application de deux approximations analytiques.

4.3.3 Estimateur d'amplitude pour $\gamma = 2$

On introduit l'équation (4.47) avec $\gamma = 2$ et l'équation (4.49) dans (4.48), on obtient :

$$\hat{A}_k^{(2)} = \frac{\int_0^\infty a^{2v} \exp\left(-\frac{a^2}{\lambda_d(k)} - \beta a^2\right) I_0\left(\frac{2aR_k}{\lambda_d(k)}\right) da}{\int_0^\infty a^{2v-1} \exp\left(-\frac{a^2}{\lambda_d(k)} - \beta a^2\right) I_0\left(\frac{2aR_k}{\lambda_d(k)}\right) da} \qquad (4.50)$$

Où l'indice (2) indique que $\gamma = 2$.

Le calcul des intégrales de l'équation (4.50) pour $v>0$, après la substitution de la relation entre β et la variance $\lambda_x(k), \beta = v / \lambda_x(k)$ (annexe B), donne [29] :

$$\hat{A}_k^{(2)} = \frac{\Gamma(v+1/2)\sqrt{Q}\, M(v+0.5;1;Q)}{\Gamma(v)\gamma_k M(v;1;Q)} R_k \qquad (4.51)$$

Où : $Q = \gamma_k \xi_k / (v + \xi_k)$ et $M(a;c;x)$ est la fonction hypergéométrique confluente.

➢ Dans le cas où $v = 1$, l'équation (4.47) devient une distribution de Rayleigh, par conséquent l'estimateur d'amplitude au sens du MMSE est identique à l'estimateur MMSE-STSA.

La figure suivante illustre les courbes du gain de l'estimateur d'amplitude au sens du MMSE sous la distribution Gamma généralisée avec $\gamma = 2$, pour les deux valeurs du SNR a priori -5 dB et 5 dB et les différentes valeurs de v (0.5 ,1 , 1.5).

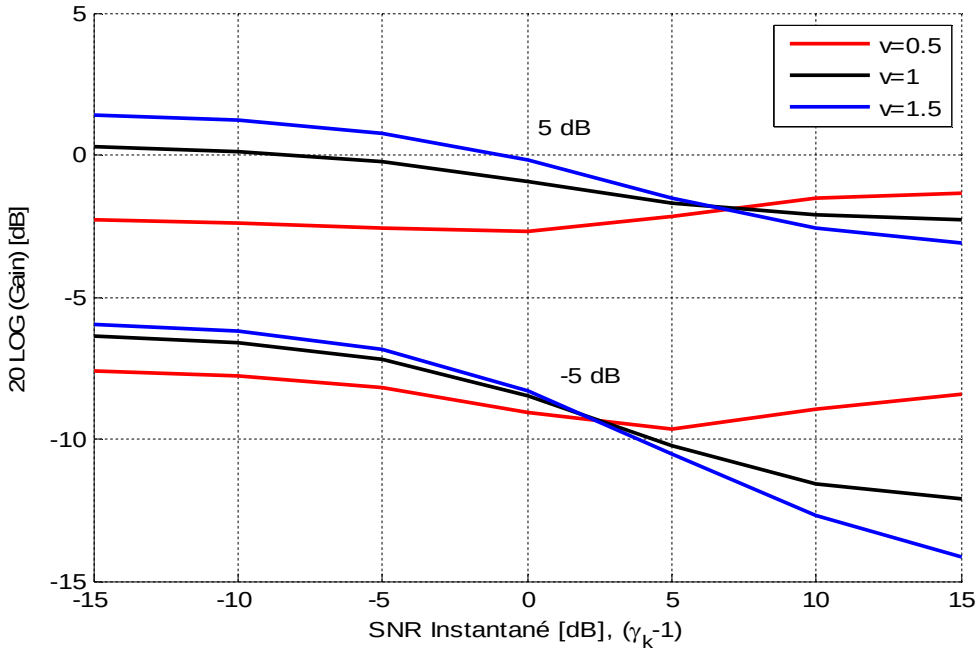


Figure 4.10 : Courbes du gain pour $\gamma = 2$ et $\xi_k = -5$ dB et +5 dB.

D'après les courbes on remarque une diminution du gain lorsque le SNR instantané augmente dans le cas où $v>1$ et le SNR a priori constant, par contre pour $v<1$ on a une augmentation du gain pour les valeurs élevées du SNR instantané et une diminution du gain pour les faibles valeurs du SNR instantané.
Pour $v=1$ les courbes du gain sont identiques à celles de l'estimateur d'amplitude MMSE-STSA basé sur un modèle Gaussien.

4.3.4 Estimateur d'amplitude pour $\gamma=1$

On introduit l'équation (4.47) pour $\gamma=1$ et l'équation (4.49) dans (4.48), on obtient :

$$\hat{A}^{(1)}=\frac{\int_0^\infty a^v \exp\left(-\frac{a^2}{\lambda_d(k)}-\beta a\right) I_0\left(\frac{2aR_k}{\lambda_d(k)}\right) da}{\int_0^\infty a^{v-1} \exp\left(-\frac{a^2}{\lambda_d(k)}-\beta a\right) I_0\left(\frac{2aR_k}{\lambda_d(k)}\right) da} \tag{4.52}$$

Avec $\beta=\sqrt{v(v+1)/\lambda_x(k)}$ (annexe B).
Les solutions analytiques exactes de ces intégrales sont inconnues, mais l'introduction de deux approximations des fonctions de Bessel nous permet de résoudre les intégrales analytiquement.

Par un changement de variables $x=\frac{2aR_k}{\lambda_d(k)}$ et l'introduction de la relation $\beta=\sqrt{v(v+1)/\lambda_x(k)}$, l'équation (4.52) devient :

$$\hat{A}^{(1)}=\frac{\lambda_d(k)}{2R_k}\frac{\int_0^\infty x^v \exp\left(-\frac{x^2}{4\gamma_k}-\frac{\mu x}{2\sqrt{\gamma_k\xi_k}}\right) I_0(x)\,dx}{\int_0^\infty x^{v-1} \exp\left(-\frac{x^2}{4\gamma_k}-\frac{\mu x}{2\sqrt{\gamma_k\xi_k}}\right) I_0(x)\,dx} \tag{4. 53}$$

Avec $\mu=\sqrt{v(v+1)}$.

La fonction $x^v \exp\left(-\frac{x^2}{4\gamma_k}-\frac{\mu x}{2\sqrt{\gamma_k\xi_k}}\right)$ atteint son maximum pour des valeurs faibles de x, lorsque l'exponentiel décroît rapidement et x^v augmente lentement. Dans ce cas, il est particulièrement important d'approximer bien la fonction de Bessel pour les petits arguments. Cela se produit lorsque ξ_k ou $\sqrt{\gamma_k\xi_k}$ est petit et v est petit. On note que le paramètre γ_k est le plus dominant dans la fonction par rapport au paramètre ξ_k. Dans d'autre cas, où le SNR est élevé, la fonction de Bessel devrait être approximée correctement pour des grands arguments.

a)Approximation de la fonction de Bessel pour les petits arguments

Pour les petits arguments de la fonction de Bessel I_0, l'approximation se fait par l'utilisation du développement de Taylor au voisinage de $x=0$. Après *K* termes, la série de Taylor de I_0 est donnée par [90, Eq.9.6.10] :

$$I_0(x;K)=\sum_{k=0}^{K-1}\left(\frac{x}{2}\right)^{2k}\frac{1}{(k!)^2} \tag{4. 54}$$

Dans la figure (4.11) nous montrons I_0 et plusieurs approximations de la série de Taylor pour $K=1,\ 2$ et 3. Nous constatons que pour des petits arguments, I_0 est très bien approximée seulement avec quelques termes. Plus l'ordre du développement de Taylor est élevé, l'approximation devient meilleure pour les arguments les plus grands.

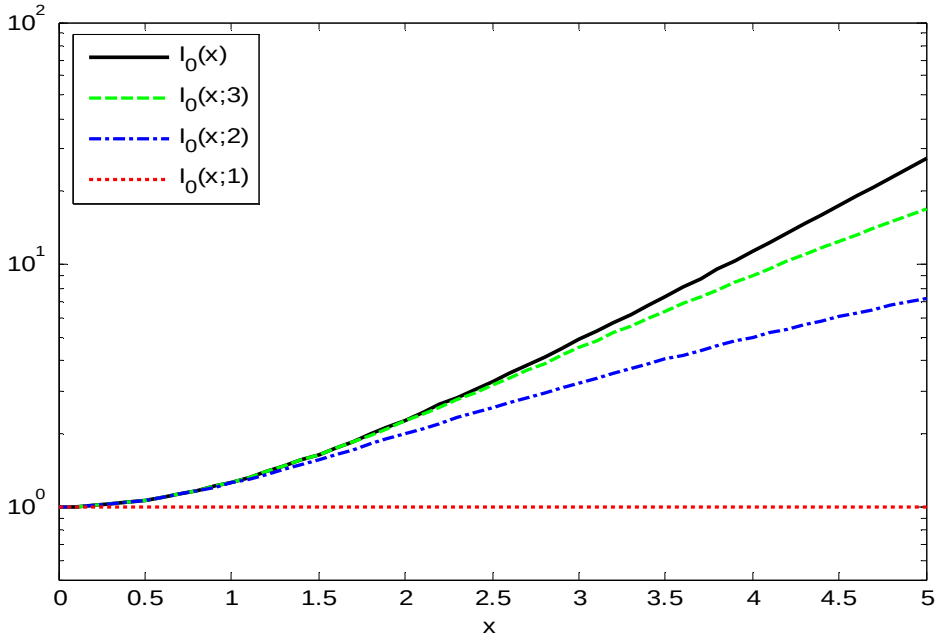


Figure 4.11 : I_0 et ses approximations par le développement en série de Taylor.

Par conséquent, l'introduction de l'équation (4.54) dans (4.53) donne :

$$\hat{A}^{(1)}_{<<,K} = \frac{\lambda_d(k)}{2R_k} \frac{\int_0^\infty x^{v} \exp\left(-\frac{x^2}{4\gamma_k} - \frac{\mu x}{2\sqrt{\gamma_k \xi_k}}\right) \sum_{k=0}^{K-1} \left(\frac{1}{k!}\right)^2 \left(\frac{x}{2}\right)^{2k} dx}{\int_0^\infty x^{v-1} \exp\left(-\frac{x^2}{4\gamma_k} - \frac{\mu x}{2\sqrt{\gamma_k \xi_k}}\right) \sum_{k=0}^{K-1} \left(\frac{1}{k!}\right)^2 \left(\frac{x}{2}\right)^{2k} dx} \tag{4.55}$$

Pour $v>0$, le calcul des intégrales conduit à [91, Eq.3.462.1] :

$$\hat{A}^{1}_{<<,K} = \frac{1}{\sqrt{2\gamma_k}} \frac{\sum_{k=0}^{K-1} \left(\frac{1}{k!}\right)^2 \left(\frac{\gamma_k}{2}\right)^k \Gamma(v+2k+1) D_{-(v+2k+1)}(T)}{\sum_{k=0}^{K-1} \left(\frac{1}{k!}\right)^2 \left(\frac{\gamma_k}{2}\right)^k \Gamma(v+2k) D_{-(v+2k)}(T)} R_k \tag{4.56}$$

Où $T=\sqrt{(v+1)v/2\xi_k}$ et D_v est une fonction parabolique cylindrique d'ordre v.

Pour $K\to\infty$, $\hat{A}^{(1)}_{<<,K} \to \hat{A}^{(1)}$ puisque le développement limité de la série de Taylor (l'équation 4.54) converge quelque soit x et l'ordre d'intégration et de sommation de l'équation (4.55) peut être changé à $K\to\infty$ selon le théorème de Fubini [92].

b) Approximation de la fonction de Bessel pour les grands arguments

Il est nécessaire d'introduire une grande valeur de *K* pour approximer exactement la fonction de Bessel I_0, qui peut avoir comme conséquence des problèmes numériques et des calculs prohibitifs, en raison de $\left(\frac{1}{k!}\right)^2$, les fonctions $\Gamma(x)$ et les fonctions paraboliques

cylindriques. Dans ce cas l'approximation la plus précise à des SNRs élevés qu'on applique est [90, Eq.9.7.1] :

$$I_0(x) \approx \frac{1}{\sqrt{2\pi x}} \exp(x) \tag{4.57}$$

La figure (4.12) illustre I_0 et ses approximations pour les grands arguments sur une échelle logarithmique. Introduisant cette approximation dans l'équation (4.53) et en utilisant [91, Eq.3.462.1], on trouve pour $\nu > 0.5$:

$$\hat{A}_{>>}^{(1)} = (\nu - 1/2)\sqrt{\frac{1}{2\gamma_k}} \frac{D_{-(\nu+1/2)}(p)}{D_{-(\nu-1/2)}(p)} R_k \tag{4.58}$$

Avec $p = \frac{\sqrt{\nu(\nu+1)}}{\sqrt{2\xi_k}} - \sqrt{2\gamma_k}$.

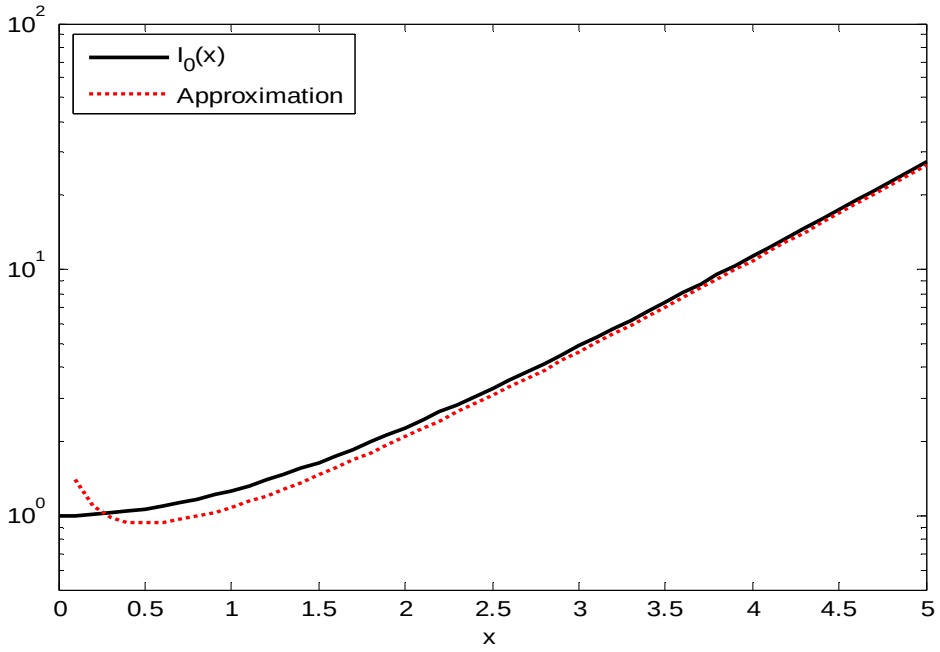


Figure 4.12 : I_0 et ses approximations pour les grands arguments.

4.3.5 Tests et résultats

Afin d'évaluer objectivement les performances de l'estimateur MMSE sous la distribution Gamma généralisée, nous commençons par le cas $\gamma = 1$. Généralement, la version de l'estimateur obtenue avec des approximations de la fonction de Bessel par des petits arguments $\left(\hat{A}_{<<,K}^{(1)}\right)$, est plus précise à des SNRs faibles. L'autre version ($\hat{A}_{>>}^{(1)}$: équation 4.58), est plus précise à des SNRs élevés et des grandes valeurs de ν. Dans [29], il a été démontré que $\hat{A}_{<<,K}^{(1)}$ est toujours inférieur à $\hat{A}^{(1)}$ pour toutes les valeurs de K. De plus, $\hat{A}_{>>}^{(1)}$ peut être inférieure et parfois supérieure à $\hat{A}^{(1)}$, suivant les valeurs des paramètres. Il s'avère que l'estimateur combiné $\hat{A}_{C,K}^{(1)} = \max\left(\hat{A}_{>>}^{(1)}, \hat{A}_{<<,K}^{(1)}\right)$, obtenu à partir d'une décision binaire simple conduit à une approximation meilleure de $\hat{A}^{(1)}$, pour toutes les valeurs de γ_k, ξ_k et ν. La figure (4.13) montre

les courbes des gains en fonction du SNR a posteriori en dB, dans le cas de $\nu = 0.6$ et plusieurs valeurs du SNR a priori (-5 dB, +5 dB et +15 dB). Dans chaque partie du graphe, les gains $G_{<<,5}^{(1)} = \hat{A}_{<<,5}^{(1)} / R$, $G_{>>}^{(1)} = \hat{A}_{>>}^{(1)} / R$ et $G_{MMSE}^{(1)} = \hat{A}_{MMSE}^{(1)} / R$ sont présentés, où $\hat{A}_{MMSE}^{(1)}$ est obtenu par intégration numérique de l'équation (4.53). Nous observons que le max entre $G_{<<,5}^{(1)}$ et $G_{>>}^{(1)}$ conduit à un gain proche de $G_{MMSE}^{(1)}$.

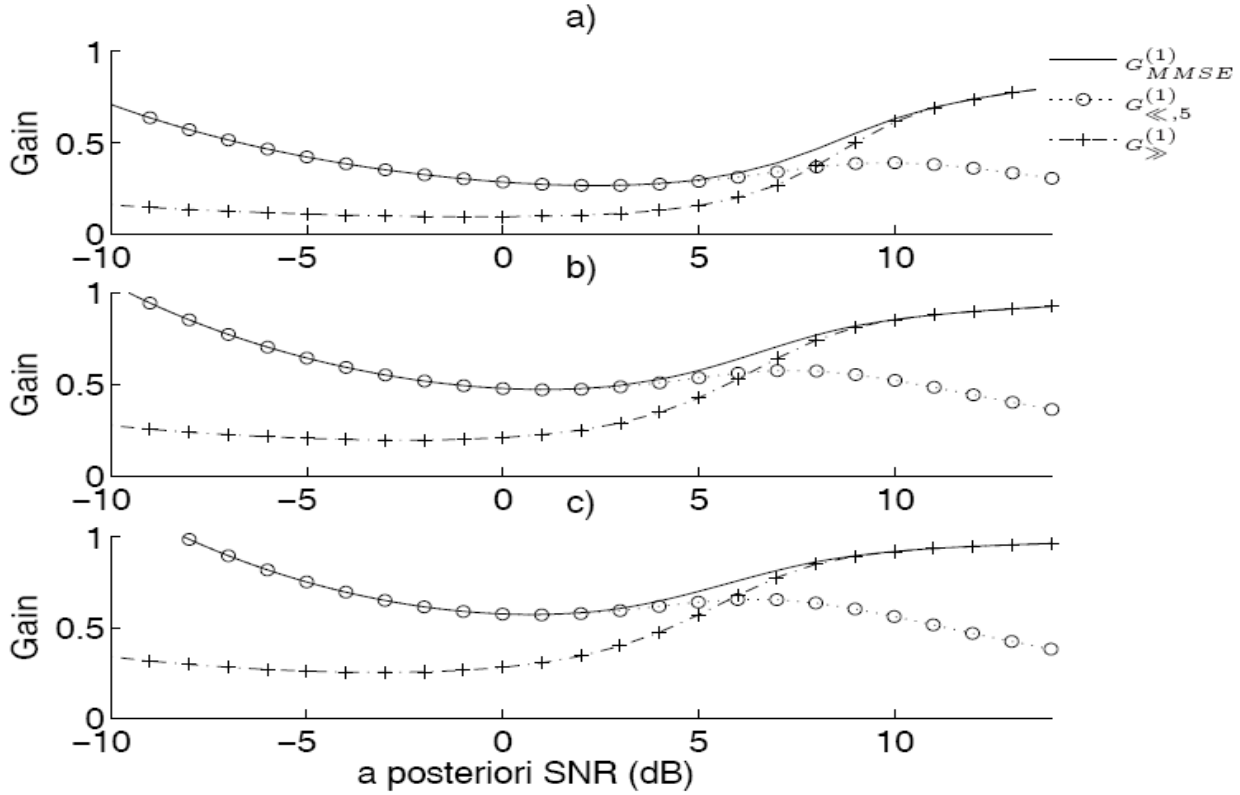


Figure 4.13 : Comparaison entre les gains $G_{<<,5}^{(1)}$, $G_{>>}^{(1)}$ et $G_{MMSE}^{(1)}$ pour $\nu = 0.6$ et a) $\xi = -5$ dB, b) $\xi = +5$ dB, et c) $\xi = +15$ dB [29].

L'implémentation des différents estimateurs sous Matlab nécessite l'évaluation des fonctions confluentes hypergéométriques et les fonctions cylindriques paraboliques. Des procédures sous Matlab disponibles à partir de [93] ont été utilisées. L'estimateur d'amplitude pour $\gamma = 2$ peut être évalué en temps réel avec un temps de calcul raisonnable. L'estimateur combiné pour $\gamma = 1$ est plus complexe, à cause de l'évaluation des deux estimateurs et surtout les sommes dans l'équation (4.56), des valeurs de K = 5 et K = 20 ont été utilisées dans les simulations qui assurent en plus des résultats acceptables obtenus, un temps de calcul raisonnable.

Nous avons effectué une comparaison entre les variations du SNRseg et celles du PESQ en fonction de *v*, en utilisant les trois variantes des gains pour $\gamma = 1$, Les résultats de la comparaison sont présentés dans la figure (4.14). On remarque que l'estimateur combiné $\hat{A}_{C,K}^{(1)} = \max\left(\hat{A}_{>>}^{(1)}, \hat{A}_{<<,K}^{(1)}\right)$ assure des performances meilleures par rapport à l'utilisation seule de l'estimateur $\hat{A}_{>>}^{(1)}$.

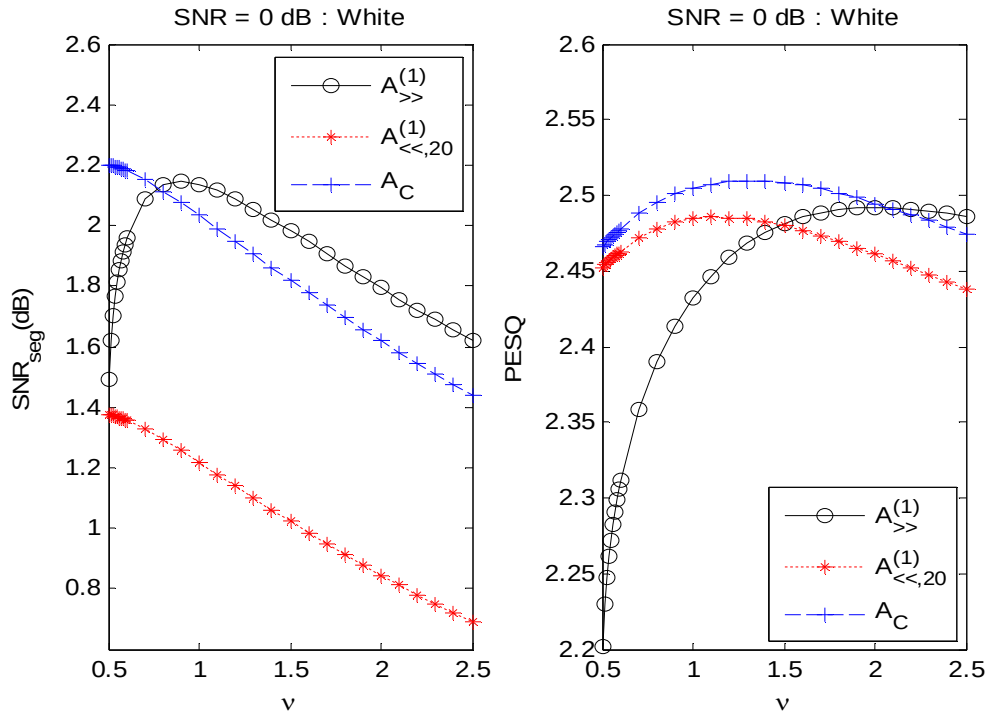


Figure 4.14 : SNRseg et PESQ en fonction de *ν*, cas d'un bruit blanc avec SNR = 0 dB de l'estimateur $\gamma = 1$.

Afin de comparer entre les performances de l'estimateur pour $\gamma = 2$ désigné par $\hat{A}^{(2)}$ et l'estimateur $\hat{A}_{C,K}^{(1)}$, nous présenterons les variations du SNRseg et du PESQ en fonction de *ν* pour $\hat{A}_{C,K}^{(1)}$ et $\hat{A}^{(2)}$, dans le cas d'un bruit blanc et d'un bruit babble avec des SNR = 0 dB et 5 dB (Figures 4.15, 4.16, 4.17, 4.18) :

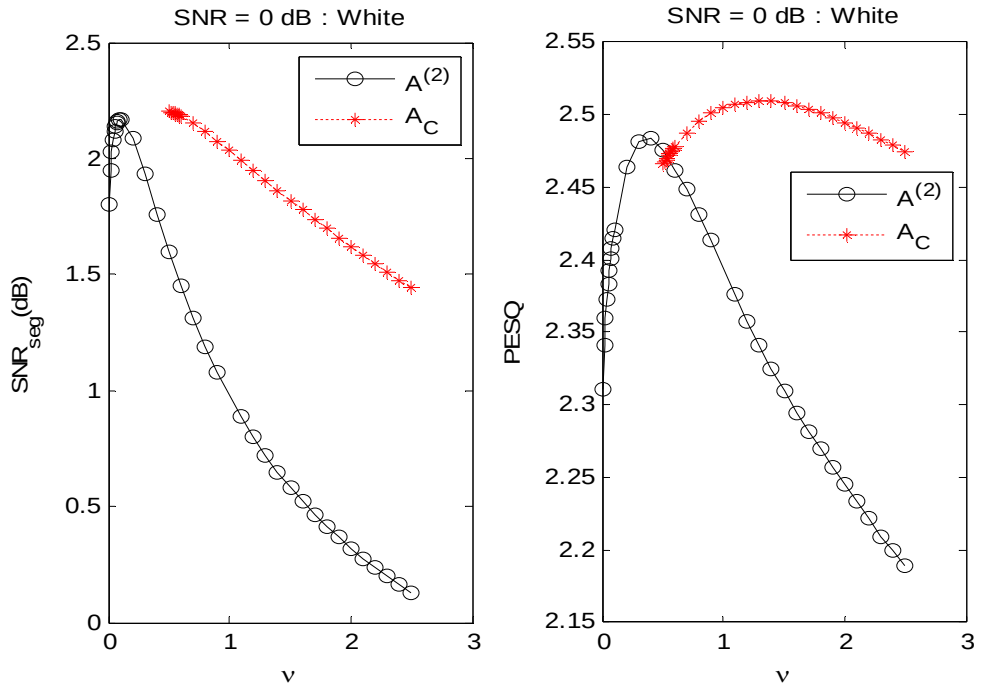


Figure 4.15 : Représentation des performances (SNRseg, PESQ en fonction de *ν*) de l'estimateur $\hat{A}_{C,K}^{(1)}$ et $\hat{A}^{(2)}$ pour un bruit blanc avec SNR = 0 dB.

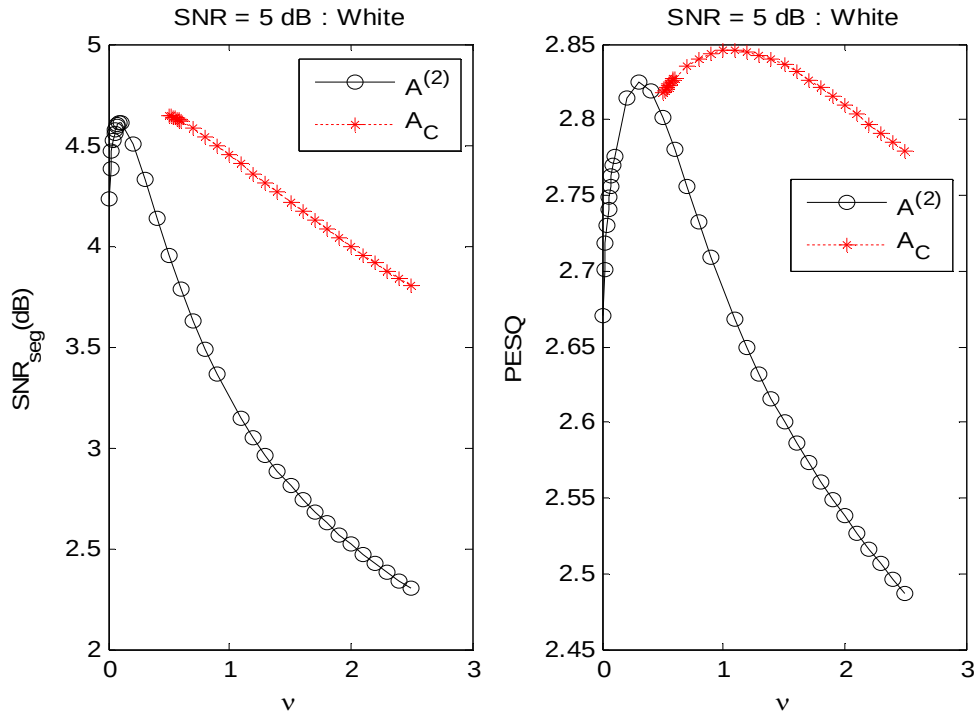


Figure 4.16 : Représentation des performances (SNRseg, PESQ en fonction de *v*) de l'estimateur $\hat{A}_{C,K}^{(1)}$ et $\hat{A}^{(2)}$ pour un bruit blanc avec SNR = 5 dB.

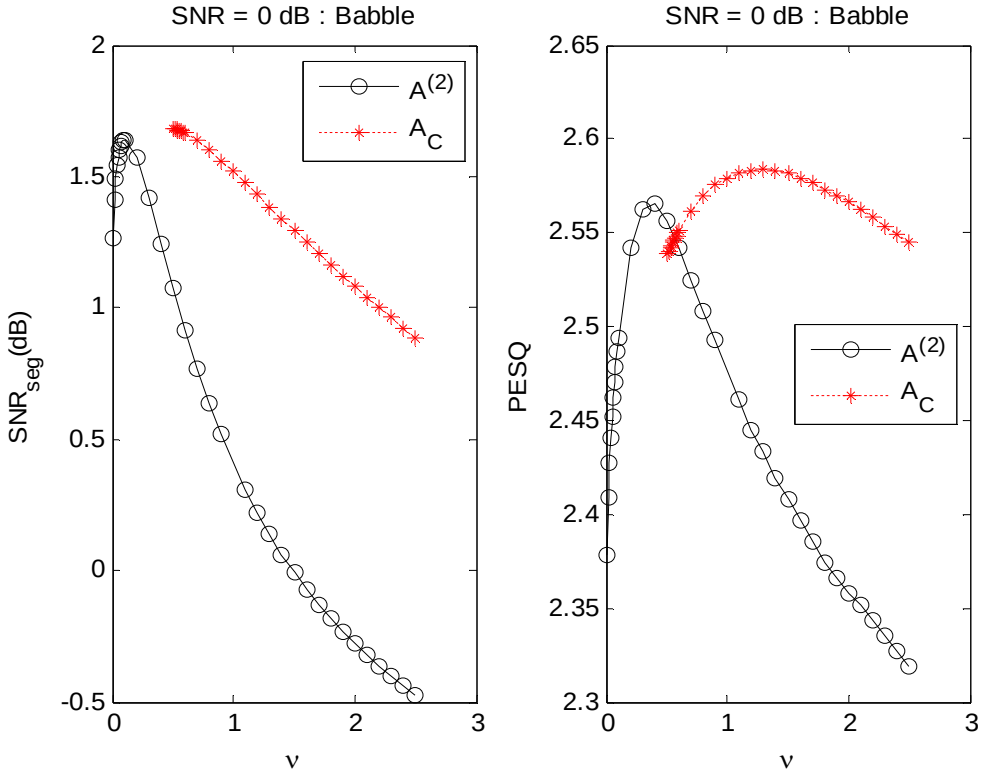


Figure 4.17 : Représentation des performances (SNRseg, PESQ en fonction de *v*) de l'estimateur $\hat{A}_{C,K}^{(1)}$ et $\hat{A}^{(2)}$ pour un bruit babble avec SNR = 0 dB.

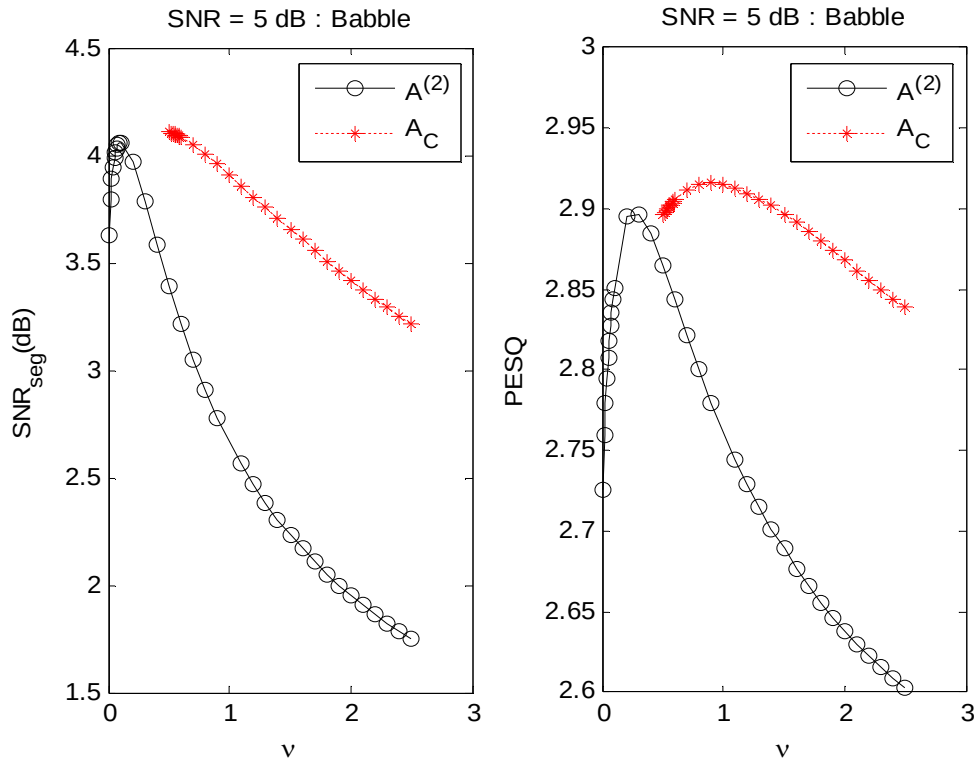


Figure 4.18 : Représentation des performances (SNRseg, PESQ en fonction de *v*) de l'estimateur $\hat{A}_{C,K}^{(1)}$ et $\hat{A}^{(2)}$ pour un bruit babble avec SNR = 5 dB.

On observe que l'estimateur $\hat{A}_{C,K}^{(1)}$ assure de meilleures performances par rapport à $\hat{A}^{(2)}$, surtout en terme du PESQ. De plus, l'estimateur $\hat{A}_{C,K}^{(1)}$ est relativement insensible à *v* est présente le maximum des performance à *v* =1.6, par contre $\hat{A}^{(2)}$ est beaucoup plus sensible à *v* est assure le maximum des performances à *v* =0.1.

Les valeurs maximales du SNRseg et du PESQ obtenues avec les deux estimateurs $\hat{A}_{C,K}^{(1)}$ (avec *v* =1.6) et $\hat{A}^{(2)}$ (avec *v* =0.1) sont approximativement identiques. Pour des applications en temps réels, l'estimateur $\hat{A}^{(2)}$ est plus adapté comparativement au temps d'exécution nécessaire à l'estimateur $\hat{A}_{C,K}^{(1)}$.

Le changement de type de bruit et du niveau du SNR a une influence négligeable sur la valeur de *v* pour les deux estimateurs.

Une fois, les valeurs de *v* qui donnent des meilleurs résultats sont obtenues, l'estimateur $\hat{A}_{C,K}^{(1)}$ et $\hat{A}^{(2)}$ seront comparés à la version MM-LSA, qui est l'estimateur le plus efficace des approches Bayesiennes basées sur un modèle Gaussien. Les résultats sont présentés dans le tableau (4.5).

D'après les résultats du tableau (4.5), on remarque que les estimateurs basés sur des modèles super-Gaussiens ($\hat{A}_{C,K}^{(1)}$ et $\hat{A}^{(2)}$) présentent les meilleures mesures objectives comparativement à l'estimateur MM-LSA pour tous les bruits à 0 dB et 5 dB et un niveau de bruit musical plus faible.

Bruit	Méthode	5 dB SNR			0 dB SNR		
		LLR	**SegSNR**	**PESQ**	**LLR**	**SegSNR**	**PESQ**
Blanc	**Bruité**	1.545	-2.327	1.799	1.802	-5.081	1.539
	MM-LSA	0.945	4.204	2.706	1.142	1.822	2.351
	$\hat{A}_{C,K}^{(1)}$	**0.763**	4.172	**2.831**	**0.953**	1.777	**2.507**
	$\hat{A}^{(2)}$	0.891	**4.609**	2.775	1.087	**2.169**	2.421
Babble	**Bruité**	0.715	-1.783	2.006	0.895	-4.632	1.705
	MM-LSA	0.478	3.661	2.791	0.617	1.288	2.441
	$\hat{A}_{C,K}^{(1)}$	**0.367**	3.605	**2.891**	**0.484**	1.251	**2.579**
	$\hat{A}^{(2)}$	0.437	**4.059**	2.851	0.577	**1.639**	2.494
Voiture	**Bruité**	0.795	-2.173	1.891	1.014	-4.959	1.634
	MM-LSA	0.511	3.589	2.703	0.652	1.273	2.365
	$\hat{A}_{C,K}^{(1)}$	**0.394**	3.519	**2.819**	**0.516**	1.208	**2.507**
	$\hat{A}^{(2)}$	0.469	**3.977**	2.763	0.612	**1.609**	2.426

Tableau 4.5 : Evaluation des performances des estimateurs MM-LSA, $\hat{A}_{C,K}^{(1)}$ et $\hat{A}^{(2)}$.

4.4 Conclusion

Dans ce chapitre, nous avons considéré les techniques basées sur la TFD pour le rehaussement de la parole mono-voie. La première partie a été consacrée aux approches basées sur le modèle Gaussien où nous avons montré par simulation que l'estimateur MM-LSA est le plus performant parmi ces approches, en utilisant un corpus de parole avec différents types de bruits et à des SNRs variés. Dans la deuxième partie de ce chapitre, les nouveaux estimateurs au sens du MMSE basés sur les modèles super-Gaussiens ont été étudiés, présentés et adaptés à notre situation. Les résultats obtenus lors des différentes simulations, montrent bien que les deux variantes $\hat{A}_{C,K}^{(1)}$ et $\hat{A}^{(2)}$ présentent une bonne réduction du bruit comparativement aux estimateurs basés sur des modèles Gaussiens. De plus, l'estimateur $\hat{A}^{(2)}$ semble bien adapté aux codeurs de la parole dans un milieu bruité pour ses performances et son temps d'exécution réduit.

CHAPITRE 5

PRETRAITEMENT AU CODEUR MELP

5.1 Introduction

Les performances des codeurs de la parole sont limitées dans les environnements bruités et surtout dans le cas des codeurs à bas débit, comme le cas du MELP à 2.4 Kbit/s. Cette limitation a donné lieu à l'idée d'opérer un rehaussement de la parole bruitée avant son codage. La combinaison de l'algorithme de réduction de bruit développé à AT&T Labs Research [6][7][89], et le codeur MELP a donné naissance au codeur MELPe. L'algorithme de réduction est une version améliorée et adaptée de l'algorithme MM-LSA, déjà détaillé dans le chapitre précédent. En plus, nous avons combiné le meilleur estimateur des méthodes Bayesiennes basées sur la distribution Gamma généralisée avec le codeur MELP, le codeur obtenu est désigné par MELPeg.

Ce chapitre sera consacré à une étude détaillée du codeur MELP comme présenté par le standard fédéral, une description des différents blocs du codeur MELPe et MELPeg, suivi par une évaluation objective de la qualité de ces codeurs.

5.2 Codage de la parole à bas débit

5.2.1 Généralités sur les codeurs de la parole à bas débit

Depuis les années 80, un progrès considérable a été réalisé dans le domaine du codage de la parole numérique, dans la bande téléphonique. Le développement des codeurs de la parole de haute qualité, fonctionnant à un débit faible, a été motivé par le marché croissant des systèmes digitaux de télécommunication et d'enregistrement, où les applications les plus importantes sont les systèmes de radiocommunications avec les mobiles, les systèmes de communications par satellite, les systèmes de communications pour les multimédia, la téléphonie par internet et la visiophonie. Ce progrès a été rendu possible grâce aux nouveaux processeurs rapides de traitement du signal, grâce à la meilleure compréhension des processus de production et de perception du signal de parole, et finalement grâce au développement d'algorithmes efficaces de codage.

Tout système de codage de la parole réalise un compromis parmi plusieurs contraintes. Idéalement, on voudrait un système capable de représenter le signal de parole avec un débit très faible, produisant un signal synthétisé d'une qualité transparente (c'est-à-dire : le signal décodé indiscernable du signal original) et ceci même en présence de différentes formes de bruit de fond lors de la prise de son. Ce codeur utiliserait en plus un algorithme de faible complexité et de faible demande en mémoire. Pour les applications de télécommunications, il faudrait aussi maintenir le délai du codage très court et assurer une parfaite robustesse contre les erreurs de transmission.

La réalisation des codeurs de haute qualité fonctionnant à bas débit exige un traitement par blocs qui sont codés comme une unité. Ces blocs sont appelés des trames et leur longueur varie habituellement entre 80 et 240 échantillons pour la fréquence d'échantillonnage de 8 kHz. L'accumulation des échantillons nécessaires pour le traitement par trame augmente en général le délai de codage, la complexité de l'algorithme et aussi la demande en mémoire. La plupart des codeurs de parole modernes réalisent une modélisation paramétrique du signal sous la forme d'un signal d'excitation passant au travers d'un filtre, en exploitant d'une certaine manière les propriétés de la perception humaine. Le filtre, appelé filtre de synthèse, est généralement modélisé par la prédiction linéaire (LP). Le plus souvent, il s'agit d'un filtre autorégressif pur.

5.2.2 Classification des codeurs de la parole à bas débit

On sépare habituellement les codeurs de parole en deux classes : les codeurs de forme d'onde et les codeurs paramétriques. On peut définir les codeurs de forme d'onde comme des codeurs dans lesquels le signal synthétisé converge vers le signal original quand le débit augmente et les codeurs paramétriques lorsque le signal synthétisé ne converge pas vers le signal original [4].

Les codeurs de forme d'onde s'efforcent de reconstruire le signal de parole en minimisant un critère de différence entre le signal original et le signal synthétisé. Pour le codage à débit réduit, les seuls codeurs de forme d'onde qui ont « survécu » à la baisse de débit sont les codeurs de type CELP (Code Excited Linear Rediction) [5]. Les codeurs de type CELP dominent le codage de la parole à bas débit au dessus d'environ 5 kb/s et en général ils atteignent une meilleure performance que les codeurs paramétriques pour d'autres signaux que la parole (musique, bruit, ...). Comme le débit descend en bas de 5 kb/s, le nombre de bits commence à être insuffisant pour décrire la forme d'onde du signal de parole et ce sont les codeurs paramétriques qui deviennent plus efficaces.

Les codeurs paramétriques utilisent le modèle de production et les caractéristiques de la perception humaine pour décrire le signal de parole par un ensemble de paramètres. Cet ensemble ne

permet pas de reconstruire la forme d'onde mais il permet de synthétiser un signal perceptiellement similaire au signal d'origine. Ainsi, en augmentant le débit, le signal synthétisé ne converge pas vers la forme du signal original et sa qualité est limitée par la précision du modèle. Comme ces codeurs reposent fortement sur le modèle de production de la parole leur performance est d'habitude très faible pour d'autres signaux. Les codeurs paramétriques peuvent être classés en trois groupes : les vocodeurs, les codeurs sinusoïdaux et les codeurs par interpolation de forme d'onde.

Les vocodeurs les plus utilisés sont sans doute ceux pour lesquels le conduit vocal est modélisé par un filtre autorégressif obtenu par la prédiction linéaire. Les vocodeurs basés sur la prédiction linéaire diffèrent principalement dans la manière de construire le signal d'excitation. Le plus simple des codeurs à prédiction linéaire (LPC) est le codeur pour lequel l'excitation est générée par un train d'impulsions espacées de périodes de pitch pour les sons voisés et par un bruit aléatoire pour les sons non-voisés [3]. Cette décision binaire pour modéliser le signal d'excitation devient trop simple quand il s'agit du codage des transitions de voisement ou du codage de la parole faiblement voisée. Le modèle de l'excitation mixte a été proposé par Makhoul [94]. Dans ce modèle, le signal d'excitation est composé d'un train d'impulsions dans les fréquences basses et d'un bruit dans les fréquences hautes. Ce modèle a été élaboré par McCree et Bamwell dans le codeur MELP (Mixed Excitation Linear Prediction) où le mélange de la composante harmonique et de la composante de bruit se fait à l'aide de deux filtres FIR variables dans le temps [95]. Une version de ce codeur, utilisée pour les tests sera détaillée dans ce chapitre.

Les codeurs siousoïdaux synthétisent la parole par une somme de sinusoïdes dont les amplitudes décrivent le spectre à court terme du signal de parole. Pour les bas débits, uniquement les amplitudes sont quantifiées et les fréquences des sinusoïdes sont habituellement harmoniques. Pour les débits plus élevés, l'information concernant les phases et les nuances des sinusoïdes peut être transmise. Ce modèle convient particulièrement au codage de la parole voisée où le signal synthétisé est reconstruit par une combinaison linéaire des sinusoïdes de fréquences harmoniques avec la fréquence fondamentale. Les sons non-voisés peuvent être synthétisés avec le même modèle en utilisant des phases aléatoires. Les représentants les plus importants des codeurs sinusoïdaux sont le codeur STC (Sinusoidal Transform Coder) et le codeur MBE (Multiband Excitation).

Le dernier groupe de codeurs paramétriques comprend les codeurs par interpolation de forme d'onde (WI pour Waveform Interpolation). Dans ces codeurs, une forme d'onde caractéristique est extraite du signal à intervalles réguliers et ses paramètres sont interpolés d'une trame à l'autre. Deux formes d'ondes sont alors extraites et transmises, une forme d'onde représente la composante périodique et l'autre la composante aléatoire du signal de parole [3]. En général, les codeurs paramétriques sont utilisés pour les débits au dessous de 5 kb/s et on peut dire qu'ils ont une meilleure performance que les codeurs de type CELP en bas de 4 kb/s.

5.3 Présentation du codeur MELP [3]

Le MELP est un codeur LPC à excitation mixte, il a été développé par McCree dans sa thèse de PHD [95] et devenu en 1997 le nouveau standard DoD pour une haute qualité de la parole à 2400 bits/s [96], remplaçant les normes fédérales FS-1015 (LPC-10) et FS-1016 (CELP), qui, par les normes modernes, produisent une parole de basse qualité. MELP à 2400 bits/s fonctionne aussi bien voire mieux que la norme fédérale FS-1016 (CELP) à 4800 bits/s, qui est le codeur de référence pour la parole à bas débit, ce qui fait de MELP un excellent candidat pour la parole à bas débit et un excellent candidat pour les applications de voix sécurisée à bas débit pour les communications civiles et militaires. Avec des échantillons de parole à 8000 Hz, MELP opère sur des trames de la

parole de 22.5ms, produisant des trames codées et empaquetées de 54 bits chacune. Pour les 2400 bits/s de débit désiré, le taux résultant des trames est approximativement 44 trames par seconde.

Le MELP est basé sur un modèle de production de la parole plus sophistiqué que celui du LPC, avec des paramètres supplémentaires modélisant encore mieux les variations du signal de parole. L'idée de base est la génération d'un signal d'excitation mixte comme entrée du filtre de synthèse, c'est-à-dire formé de la combinaison d'une séquence filtrée d'impulsions périodiques avec une séquence filtrée de bruit. L'excitation mixte est formée de la somme d'une composante impulsionnelle et d'une composante de bruit. Ce codeur a d'autres caractéristiques supplémentaires permettant d'améliorer la représentation du signal de parole, comme les amplitudes de Fourrier, l'indicateur d'apériodicidé, le filtre adaptatif de renforcement spectral et le filtre de dispersion des impulsions. Ces paramètres supplémentaires modifient la structure de l'excitation du modèle LPC, et en même temps éliminent les tonalités courtes au niveau du signal de parole synthétisée, et par conséquent, le codeur MELP permet de mieux synthétiser la parole par rapport aux autres codeurs qui le précèdent.

5.3.1 Modèle de production de la parole

Comme dans le cas du codeur LPC10, la production de la parole par le codeur MELP est obtenue en excitant un filtre de synthèse par un signal d'excitation, ce dernier peut être un train d'impulsion périodique ou non périodique, ou un signal aléatoire. Le train d'impulsion apériodique est le résultat d'une perturbation aléatoire (jitter) appliquée à la période du pitch, il correspond aux zones de transition entre les segments voisés et non voisés.

L'excitation périodique et l'excitation aléatoire sont d'abord filtrées à l'aide du filtre de mise en forme d'impulsions et le filtre de mise en forme du bruit, respectivement ; l'addition des sorties des deux filtres forme l'excitation totale, connue sous le nom d'excitation mixte. A noter que les fonctions de transfert des deux filtres de mise en forme sont contrôlées par les intensités de voisement. Après filtrage de l'excitation mixte par le filtre de synthèse et la mise à l'échelle du signal, on obtient la parole synthétisée par ce modèle de production (figure 5.1).

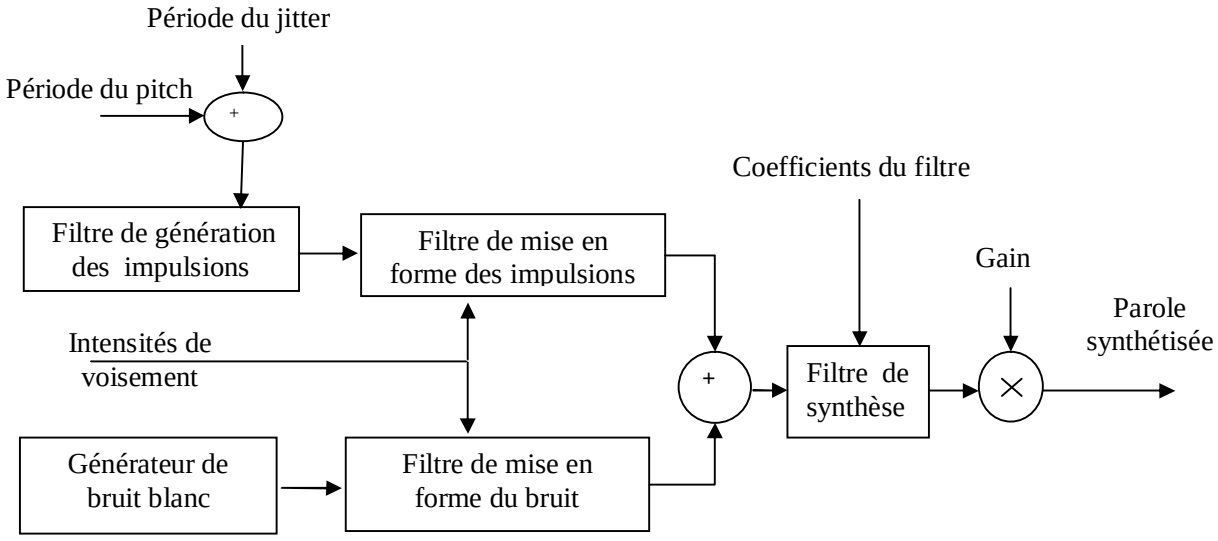


Figure 5.1 : Modèle de production de la parole du codeur MELP.

5.3.2 Etape d'analyse du codeur MELP

L'implémentation du codeur MELP se fait en deux phases : la phase d'analyse et la phase de synthèse. Durant cette première étape, les paramètres du codeur MELP sont estimés en exécutant chaque trame tous les 22.5 ms, soit un ensemble de 180 échantillons, pour une fréquence d'échantillonnage de 8 KHz, le dernier échantillon de la trame analysée est utilisé comme un point de référence : pour la plupart des paramètres (sauf le gain), cet échantillon représente le centre de la fenêtre d'analyse, donc l'estimation d'un paramètre est basée sur les échantillons de la demi trame traitée et celles de la demi trame suivante (figure 5.2).

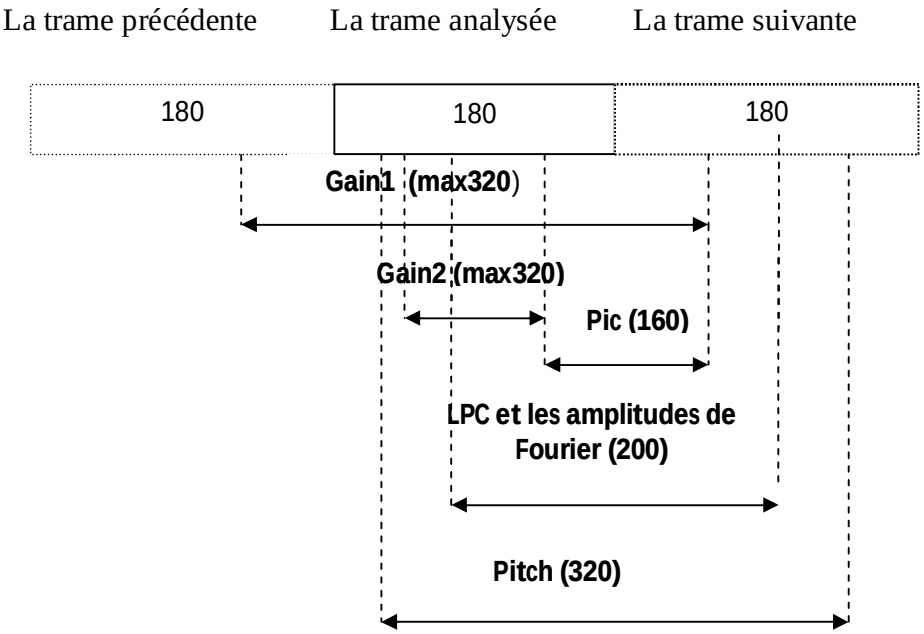


Figure 5.2 : Différents échantillons utilisés pour chaque paramètre [9].

.

Le codeur fait une première estimation de la fréquence fondamentale, puis il calcule les intensités de voisement dans les cinq bandes de fréquence adjacentes. Par la suite, chaque bande est classée voisée ou non voisée suivie d'une dernière estimation du pitch en exploitant l'erreur de la prédiction linéaire. Après l'analyse, le codeur peut positionner un indicateur appelé indicateur d'apériodicité (aperiodic flag) pour signaler au décodeur que la composante impulsionnelle doit être apériodique. Le codeur effectue par ailleurs une analyse spectrale par prédiction linéaire et calcule les amplitudes des dix premières harmoniques de la transformée de Fourier du signal résiduel. Les paramètres transmis par ce codeur dans le cas des trames voisées sont :

- la fréquence fondamentale.
- les cinq intensités de voisement.
- deux gains (correspondant aux énergies des deux demi trames).
- les dix coefficients de prédiction linéaire.
- les dix amplitudes de Fourier.
- le drapeau d'apériodicité.

Et dans le cas des sons non voisés, ces paramètres sont :

- Les dix coefficients de prédiction linéaire.
- Deux gains.

5.3.2.1 Amplitudes des coefficients de Fourier

Dans le codeur MELP, les amplitudes des coefficients de Fourier du signal d'erreur de la prédiction linéaire sont calculées, pour modéliser la forme de l'impulsion d'excitation. L'objectif est de créer dans le décodeur, une séquence périodique aussi proche que possible de l'excitation originale. De plus, ces amplitudes sont calculées seulement si la trame est voisée ou jittery voisée (transition).

a) Filtre de génération des impulsions

Le modèle MELP repose sur le filtre de génération des impulsions pour la génération de l'excitation périodique. L'idée est de trouver la réponse impulsionnelle du filtre durant le codage et de transmettre cette réponse au décodeur afin de générer le train d'impulsion, utilisé comme excitation périodique à l'entrée du filtre de synthèse. Durant le codage, l'amplitude de la TFD de l'erreur de prédiction est obtenue ; les sommets du spectre d'amplitude correspondants aux harmoniques ($\omega = 2\pi i / T,\ \ i = 1, 2, ...$) associés avec la période du pitch (T) sont mesurées ; les valeurs des sommets sont les amplitudes de Fourier qui seront transmises au décodeur pour générer l'excitation impulsionnelle, ou la réponse impulsionnelle du filtre de génération des impulsions.

b) Calcul et quantification des amplitudes de Fourier

La procédure utilisée pour le calcul des amplitudes de Fourier par le codeur MELP est illustrée par la figure (5.3).

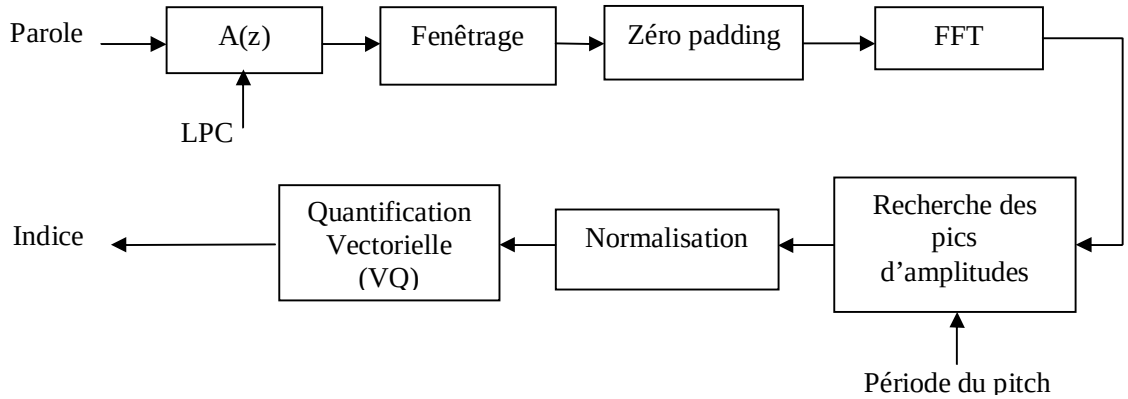


Figure 5.3 : Schéma bloc du calcul et quantification des amplitudes de Fourier.

Les entrées de la procédure sont : les 200 échantillons de la parole, les coefficients de l'analyse LPC et la période du pitch. Les deux dernières informations sont supposées déjà extraites à partir du signal. L'erreur de prédiction est d'abord calculée en passant les échantillons de la parole à travers le filtre d'analyse LPC. La séquence de l'erreur de prédiction de 200 échantillons résultante est multipliée par une fenêtre de Hamming. La séquence résultante sera augmentée à 512 échantillons par l'addition des zeros (zero padding), dans le but est d'augmenter la résolution fréquentielle. Suivie d'une FFT et un bloc de recherche des pics d'amplitudes. A la fin de la procédure, 10 pics d'amplitudes sont obtenus : $Fmag[i],\ i = 1\ :\ 10$ et seront normalisées comme suit :

$$Fmag'[i] = \alpha . Fmag[i] \qquad (5.1)$$

Où $$\alpha = \left(\frac{1}{10} \sum_{i=1}^{10} \left(Fmag\left[i \right] \right)^2 \right)^{-1/2} \quad (5.2)$$

A la fin, une quantification vectorielle (VQ) de 8 bits est appliquée.

5.3.2.2 Filtres de mise en forme

Le modèle MELP de production de la parole utilise deux filtres de mise en forme pour combiner l'excitation impulsionnelle et l'excitation bruitée de manière à former le signal d'excitation mixte. Les réponses de ces filtres sont contrôlées par un ensemble de paramètres appelés intensités de voisement ; ces paramètres sont estimés à partir du signal d'entrée. Ces filtres déterminent la quantité d'impulsions et la quantité de bruit dans l'excitation, aux différentes bandes de fréquences. Dans le FS MELP, chaque filtre de mise en forme est composé de cinq filtres, appelés filtres de synthèses, puisque ils sont utilisés pour synthétiser le signal d'excitation mixte durant le décodage. Chaque filtre contrôle une bande de fréquence particulière, les cinq bandes de fréquences adjacentes sont : 0-500, 500-1000, 1000-2000, 2000-3000, 3000-4000 Hz. Ces filtres connectés en parallèle déterminent les réponses en fréquence des deux filtres de mise en forme, la figure (5.4) présente le schéma bloc d'un filtre de mise en forme. Considérons $h_i\left[n\right],\ i = 1...5,$ les réponses impulsionnelles des filtres de synthèse, la réponse totale du filtre de mise en forme est :

$$h_p\left[n\right] = \sum_{i=1}^{5} vs_i h_i\left[n\right] \quad (5.3)$$

Avec $0 \le vs_i \le 1$ sont les intensités de voisement. D'autre part, le filtre de mise en forme du bruit a la réponse suivante :

$$h_n\left[n\right] = \sum_{i=1}^{5} (1 - vs_i) h_i\left[n\right] \quad (5.4)$$

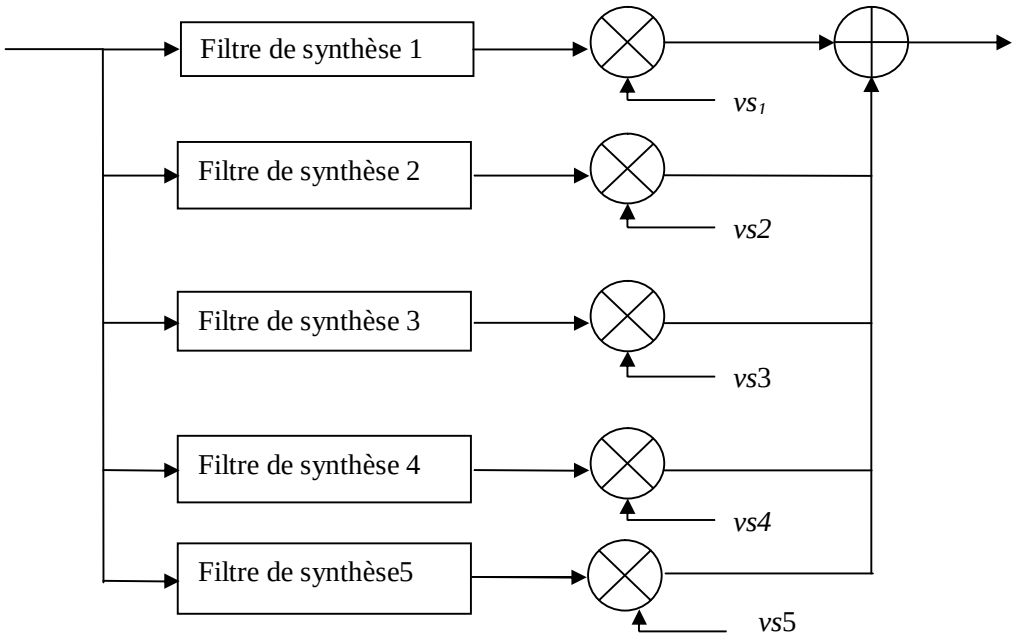


Figure 5.4 : Schéma bloc du filtre de mise en forme des impulsions.

Ces filtres de synthèse sont implémentés comme des filtres à réponse impulsionnelle finie (RIF) avec 31 points. Les figures (5.5) et (5.6) présentent les réponses impulsionnelles $\left(h_i[n],\ i=1...5,\ n=0...30\right)$ des filtres de synthèse et ses spectres d'amplitudes respectivement.

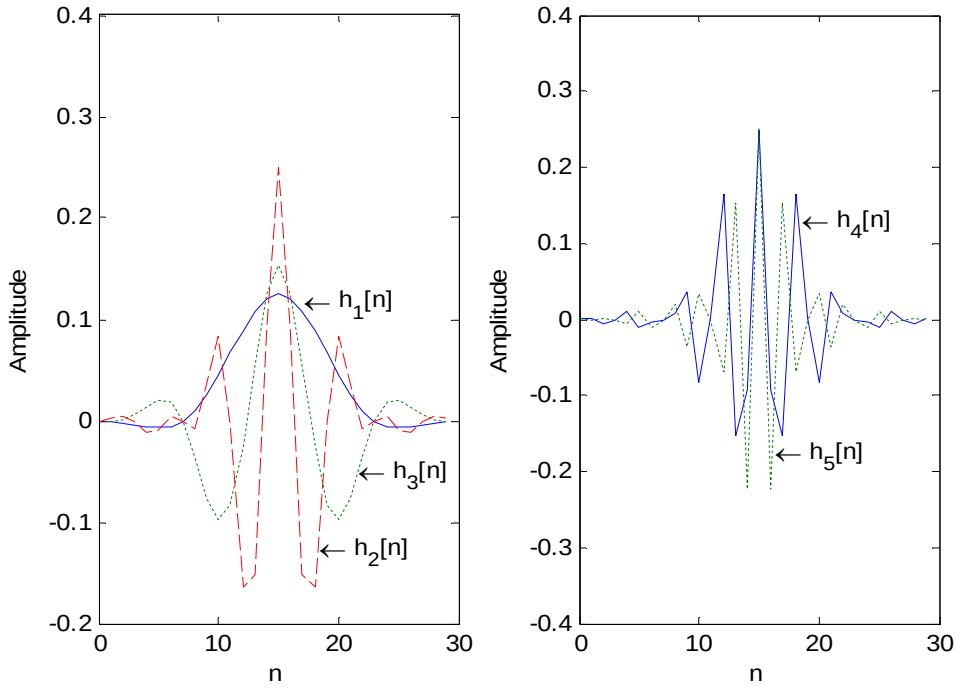


Figure 5.5 : Réponses impulsionnelles des filtres de synthèse du FS MELP.

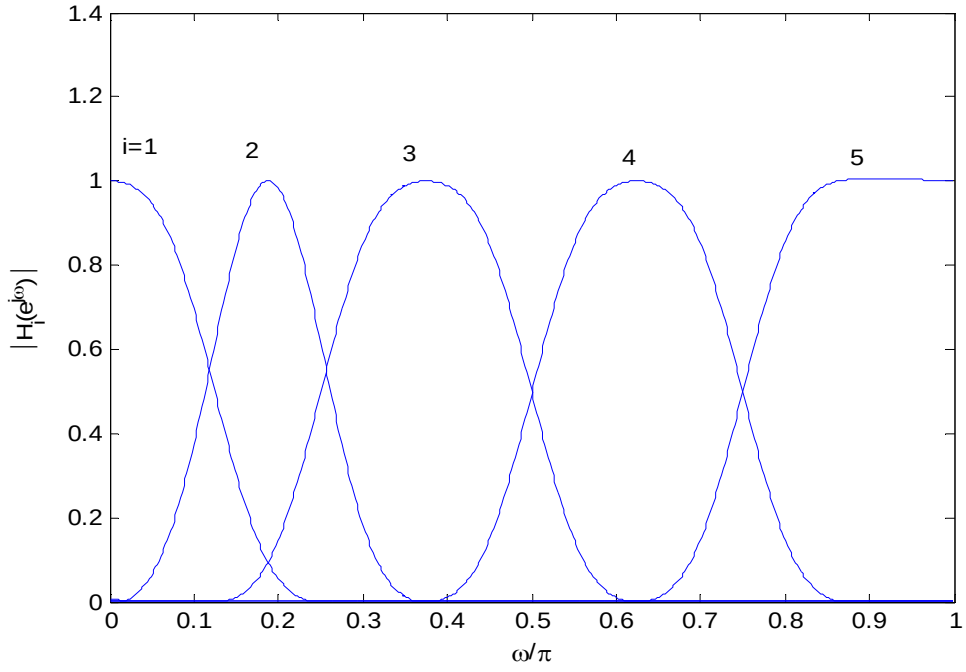


Figure 5.6 : Spectres d'amplitudes des filtres de synthèse du FS MELP.

5.3.2.3 Estimation du pitch et des intensités de voisement

Le codeur MELP utilise une procédure efficace pour augmenter la précision de l'estimation du pitch et des intensités de voisement, ces paramètres influents sur la qualité de la parole synthétisée. Un banc de filtre d'analyse est utilisé pour diviser le signal de parole d'entrée en cinq bandes, avec détermination d'intensité de voisement dans chaque bande**.** Les cinq filtres d'analyse possèdent les mêmes largeurs de bandes comme les filtres de synthèse déjà présentés. Mais implémentés comme des filtres à réponse impulsionnelle infinie (RII) de Butterworth d'ordre 6, donc un temps de calcul relativement faible par rapport à la configuration RIF avec 31 points.

a) Première estimation de la période du pitch

Le signal de la parole d'entrée est filtré par le premier filtre d'analyse, de bande passante allant de 0 Hz à 500 Hz**.** Pour chaque trame de 180 échantillons, l'autocorrélation normalisée

$$r[l] = \frac{c[0,l,l]}{\sqrt{c[0,0,l]c[l,l,l]}} \tag{5.5}$$

Où

$$c[l,m,k] = \sum_{n=-[k/2]-80}^{-[k/2]+79} s[n+l]s[n+m] \tag{5.6}$$

est calculée pour chaque échantillon « l » telle que : l = 40,…,160, la période de pitch $T^{(1)}$ est la valeur correspondante à « l » pour laquelle $r[l]$ est maximale.

b) Intensité de voisement de la bande basse

L'intensité de voisement vs_1 de la bande basse $(0-500\ \text{Hz})$ est la valeur d'autocorrélation normalisée associée à $T^{(1)}$, c'est-à-dire :

$$vs_1 = r\left(T^{(1)}\right) \tag{5.7}$$

La figure suivante illustre le calcul d'intensité de voisement dans la bande basse ainsi que la première estimation du pitch :

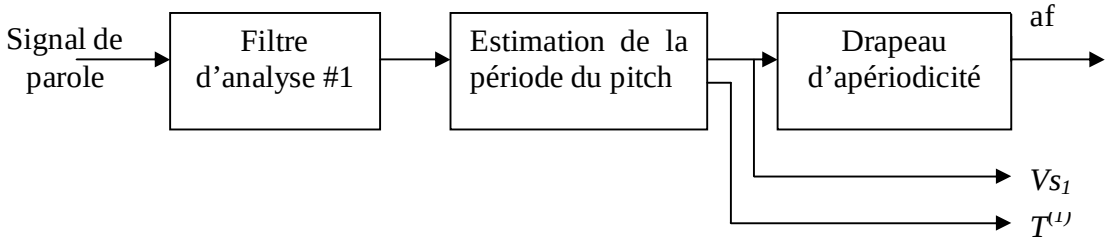


Figure 5.7 : Illustration de la première estimation de la période du pitch.

c) Drapeau d'apériodicité

L'indicateur d'apériodicité dépend de l'intensité de voisement de la bande basse, il est égale à 1 si vs_1<0.5 et égale à 0 autrement, cet indicateur signale au décodeur que la composante impulsionnelle doit être apériodique. L'indicateur d'apériodicité est quantifié sur un seul bit**.**

d) Intensités de voisement des quatre bandes restantes

La figure (5.8) illustre le système utilisé pour l'estimation des intensités de voisement des quatre bandes restantes. On calcule $r_1 = r\left(T^{(1)}\right)$ en utilisant pour l'évaluation de l'autocorrélation normalisée, le signal de sortie du filtre passe bande correspondant. Ensuite le calcul de $r_2 = r\left(T^{(1)}\right)$, mais cette fois le signal utilisé est l'enveloppe du signal obtenu à la sortie du filtre passe bande. Ainsi, l'intensité de voisement de cette bande est donnée par:

$$vs_2 = \max\left(r_1, r_2\right) \tag{5.8}$$

La même procédure se répète pour les trois bandes restantes, pour déterminer vs_3, vs_4 et vs_5.

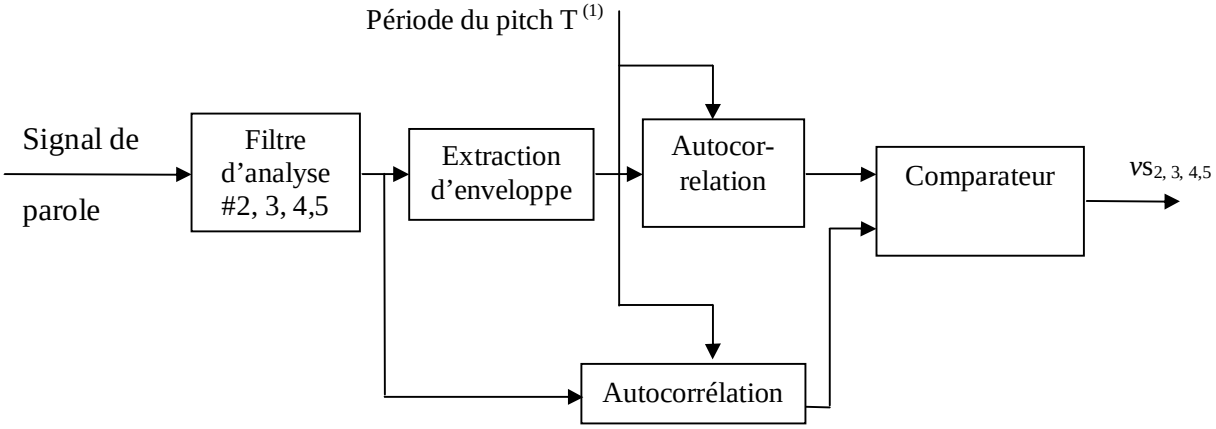


Figure 5.8 : Estimations des quatre intensités de voisement**.**

e) Analyse par prédiction linéaire et erreur de prédiction

Une analyse LPC est appliquée sur le signal de parole d'entrée fenêtré par une fenêtre de Hamming de 200 échantillons, centrée sur le dernier échantillon de la trame en cours. La méthode d'autocorrélation et l'algorithme de Levinson-Durbin sont utilisés pour déterminer les coefficients et l'erreur de prédiction. Les coefficients de prédiction résultants sont transformés en coefficients LSF puis quantifiés par une quantification vectorielle à multi- stage (MSVQ : Multi Stage Vectors Quantizer). Le vecteur des paramètres LSF d'ordre 10 est divisé en quatre parties : la première partie est assignée 7 bits, la deuxième partie 6 bits, la troisième partie 6 bits et la dernière partie 6 bits, donc il emploie 25 bits/trame.

f) Peakinness (pic)

Le pic du signal d'erreur de la prédiction est calculé sur une fenêtre de 160 échantillons centrée sur le dernier échantillon de la trame en cours. Cette valeur est définie par :

$$p = \frac{\sqrt{\frac{1}{160}\sum_{n=-80}^{79} e^2(n)}}{\frac{1}{160}\sum_{n=-80}^{79} \left|e(n)\right|} \tag{5.9}$$

Ce pic est plus élevé pour des trames voisées dues à la structure quasi-périodique du train d'impulsions. En particulier durant les transitions où sa valeur devient de plus en plus élevée. Alors, cette mesure est très utile dans la décision voisée / non voisée, ainsi que pour la détection des transitions. Les intensités de voisement pour les trois bandes les plus basses sont modifiées selon la valeur du pic, comme suit :

- Si $p > 1.34$ alors $vs_1 \leftarrow 1$.
- Si $p > 1.60$ alors $vs_i \leftarrow 1$ avec $i = 2,3$.

L'utilisation des deux mesures : le pic et l'autocorrélation est très efficace pour la classification des sons. Les trames non voisées ont un pic faible et une autocorrélation faible, conduisant à des intensités de voisement faibles. Pendant les transitions, la valeur du pic est élevée avec une valeur d'autocorrélation moyenne, ce qui ramène l'indicateur d'apériodicité à 1 signalant au décodeur de générer des périodes aléatoires. De l'autre coté, un pic élevé assure des valeurs maximales des intensités de voisement ; ce cas indique un état de transition (jittery voiced state). Dans le cas des trames voisées, la valeur du pic est moyenne avec une autocorrélation élevée ; ce qui ramène l'indicateur d'apériodicité à 0 avec des intensités de voisement élevées. Le codeur MELP utilise ces deux informations (l'indicateur d'apériodicité et les intensités de voisement) pour mieux classifier les trames de la parole.

g) Estimation finale de la période du pitch

L'erreur résiduelle est filtrée par un filtre passe bas avec une fréquence de coupure de 1 KHz, le signal résultant est utilisé pour l'estimation de la période de pitch, cette estimation est obtenue par une recherche au voisinage de la première estimation de la période du pitch $T^{(1)}$. La reéstimation du pitch à partir de l'erreur de prédiction augmente la précision de la première estimation, à cause de l'élimination de la structure formantique du signal de la parole original.

L'estimation finale du pitch T et l'intensité de voisement de la bande basse vs_1 sont quantifiés conjointement avec 7 bits. Si $vs_1 \leq 0.6$, donc la trame est non voisée et un code de 7 bits tout-zéro est envoyé. Sinon, $\log(T)$ est quantifié uniformément avec 99 niveaux de quantification entre log(20) et log(160). La valeur quantifiée de vs_1 notée qvs_1 est égale à 0 pour l'état non voisé et 1 pour l'état voisé. Les quatre intensités de voisement restantes sont quantifiées sur 4 bits.

5.3.2.4 Gain et allocation des bits

Deux gains sont mesurés pour chaque trame, la longueur de la fenêtre utilisée pour les deux mesures du gain est identique et dépend du paramètre vs_1.

- Lorsque $vs_1 > 0.6$, dans ce cas la longueur de la fenêtre est liée à la valeur de la première estimation de pitch. Elle est égale à deux fois cette valeur, ce qui correspond à une valeur minimale de 120 échantillons, et lorsque elle excède 320 échantillons, elle est divisée par 2. Ce cas correspond aux trames voisées.
- Lorsque $vs_1 \leq 0.6$, dans ce cas la longueur de la fenêtre est de 120 échantillons. Ce cas correspond aux trames non voisées ou de transition.

La fenêtre d'analyse est centrée sur l'échantillon 90 avant le dernier échantillon dans la trame en cours pour l'estimation du premier gain (G_1), et pour le deuxième gain (G_2) elle est centrée sur le dernier échantillon. L'équation pour le calcul du gain est :

$$G = 10\log_{10}\left(0.01 + \frac{1}{N}\sum_{n} s^2(n)\right) \tag{5.10}$$

Où : N est la longueur de la fenêtre, $s(n)$ le signal de parole d'entrée et les limites de n dépendent de la longueur de la fenêtre.
La mesure du gain suppose que le signal d'entrée est dans la gamme [-32768, 32767] (16 bits par échantillons). Le premier gain (G_1) est quantifié sur 3 bits par une quantification uniforme et le deuxième (G_2) sur 5 bits par une quantification uniforme de 32 niveaux.

La figure (5.9) résume les différentes étapes d'analyse MELP par le schéma bloc simplifié suivant :

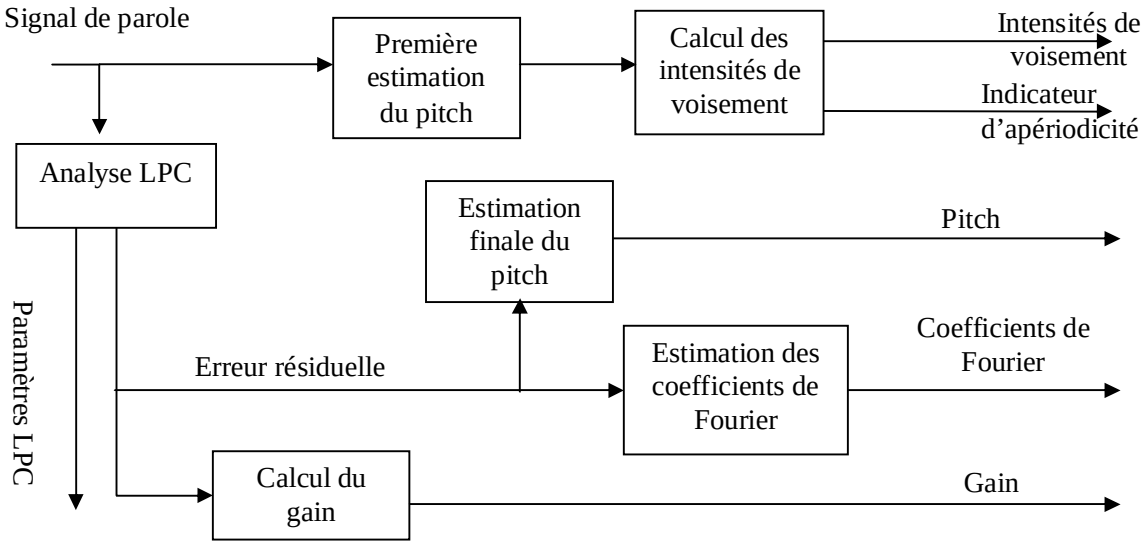


Figure 5.9 : Schéma bloc simplifié de l'analyse MELP.

Le tableau (5.1) présente l'allocation des bits des différents paramètres du codeur MELP dans le cas des trames voisées et non voisées. La protection contre les erreurs est assurée pour les trames non voisées seulement, en utilisant 13 bits. Ainsi, un total de 54 bits pour chaque trame de 22.5 ms donne un débit de 2.4 Kbit/s

Paramètres	**Voisés**	**Non voisés**
Les LSF	25	25
Les amplitudes de Fourier	8	-
Gain (2 à chaque trame)	8	8
Pitch/intensité de voisement vs_1	7	7
L'intensité de voisement	4	-
Indicateur d'apériodicité	1	-
Erreur de protection	-	13
Bit de synchronisation	1	1
Somme de bits/22.5 ms (trame)	54	54

Tableau 5.1 : Table d'allocation des bits du codeur MELP [96].

5.3.3 Etape de synthèse du codeur MELP

Le synthétiseur interpole linéairement les différents paramètres de manière synchrone au pitch. La composante impulsionnelle est obtenue sur une période de pitch par la transformée de Fourier inverse sur les 10 amplitudes de Fourier. Un nouveau paramètre « jitter » est introduit au niveau du décodeur pour contrôler l'excitation apériodique pour les trames de transition, la valeur du jitter est égale à 0.25 si l'indicateur d'apériodicité est activé, si non elle est égale à 0. Dans ce cas la période du pitch est donnée par la relation suivante :

$$T = T_0\left(1 + jitter.x\right) \tag{5.11}$$

Où T_0 est la valeur de la période du pitch, et x un nombre aléatoire distribué uniformément sur l'intervalle [-1,1].

La séquence impulsionnelle de T-échantillons générée à partir des amplitudes de Fourier à une variance unité. Cette séquence est filtrée par le filtre de mise en forme des impulsions et sommée avec la séquence filtrée du bruit pour former l'excitation globale mixte. Le générateur de bruit est une source aléatoire distribuée uniformément de moyenne nulle et de variance unité. Les coefficients des filtres transmis par le codeur sont interpolés pitch synchrone.

L'excitation globale est ensuite filtrée par un filtre adaptatif de renforcement spectral des formants basé sur les pôles du filtre de synthèse de la prédiction linéaire. Il est utilisé pour rehausser les formants de la parole synthétisée et d'augmenter la qualité perceptuelle de la parole synthétisée. Ce post- filtre est caractérisé par la fonction de transfert suivante :

$$H\left(z\right) = \left(1 - \mu z^{-1}\right)\frac{1 + \sum_{i=1}^{10} a_i \beta^i z^{-1}}{1 + \sum_{i=1}^{10} a_i \alpha^i z^{-1}} \tag{5.12}$$

Où les a_i sont les coefficients de la prédiction linéaires et les paramètres μ , α et β sont adaptatifs et dépendent du signal. Ce filtre est identique au post-filtre dans le codeur CELP.

Le signal à la sortie du filtre adaptatif de renforcement spectral traverse le filtre de synthèse, qui est un filtre de synthèse par formants avec des coefficients correspondants aux LSF interpolées.

La puissance du signal à la sortie du filtre de synthèse doit être égale au gain interpolé " g " de la période en cours. Comme l'excitation est générée avec un niveau quelconque, un facteur d'échelle g_0 est calculé afin de mettre la sortie du filtre de synthèse y(n) au niveau approprié. Ceci est donné par :

$$g_0 = \frac{10^{g/20}}{\sqrt{\frac{1}{T}\sum_n y^2\left[n\right]}} \tag{5.13}$$

Par la multiplication de $y\left[n\right]$ par g_0, la séquence résultante de T-échantillons aura comme puissance g dB.

Le dernier bloc dans la chaîne de décodage est le filtre de dispersion des impulsions. C'est un filtre passe tout RIF à 65 points, qui a pour rôle d'améliorer la qualité de la parole synthétisée dans les régions qui ne contient pas de structure formantique et d'étaler l'énergie des impulsions sur une période de pitch (pulse dispersive filter). Enfin, la figure (5.10) illustre un schéma bloc simplifié de la synthèse MELP et l'annexe C présente les deux schémas blocs complets du codeur et décodeur MELP.

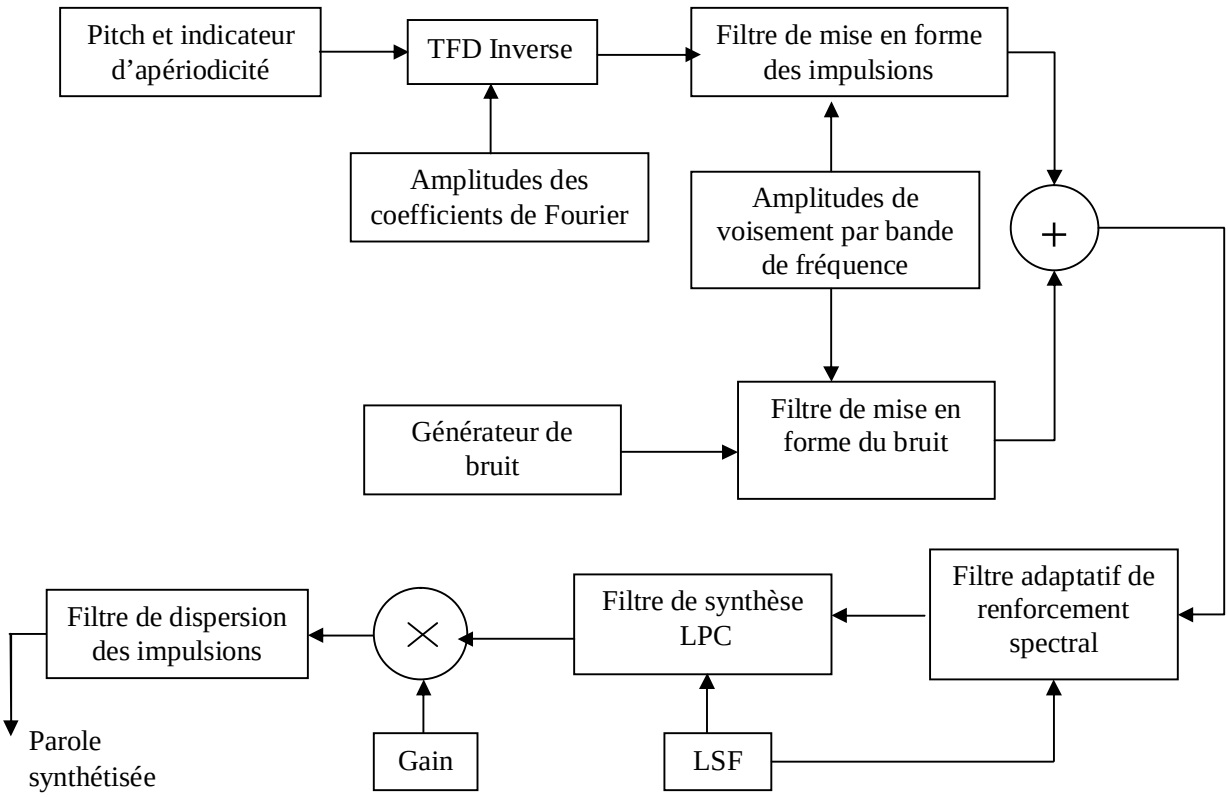


Figure 5.10 : Schéma bloc simplifié de la synthèse MELP.

5.3.4 Implémentation

L'évaluation des codeurs à bas et très bas débits ne peut pas se faire par des critères objectifs de rapport signal à bruit. Le signal décodé doit être perçu comme proche de l'original, mais les formes d'onde peuvent être très différentes. On évalue ces codeurs par des tests subjectifs, tels que le test ACR (Absolute Category Rating) délivrant un score MOS (*Mean Opinion Score*) ou le test d'acceptabilité DAM (*Diagnostic Acceptability Measure*) pour la qualité, et le test de rimes DRT (Diagnostic Rhyme Test) pour l'intelligibilité. Ces tests sont menés sous certaines conditions de bruit ambiant ou de taux d'erreurs canal. Pour qualifier la qualité d'un codeur, on utilise les termes anglais : « *broadcast* », « *toll* », « *telecommunication* », « *synthetic* ». Une qualité de type « *broadcast* » correspond à un codage large bande (audioconférence par exemple), la qualité de type « *toll* » est celle du téléphone analogique filaire. Pour une qualité de type « *telecommunication* », l'intelligibilité et le naturel sont conservés mais quelques distorsions sont audibles. Un codeur de qualité « *synthetic* » est intelligible mais le signal manque de naturel avec perte de reconnaissance du locuteur. Ainsi, une note MOS de 4 - 4.5 correspond à une qualité de type « *toll* », 3.5 - 4 : qualite de type « *telecommunication* », 2.5 - 3.5 : « *synthetic* ».

Le codeur MELP est un codeur paramétrique qui ne conserve pas la forme du signal d'entrée. En outre, l'information de phase n'est ni conservées ni utilisées dans le processus d'encodage. Donc, les résultats obtenus des mesures comme le SNR et le SNRseg restent limitées pour l'évaluation des performances. Par conséquent, l'accent est mis sur les tests subjectifs pour évaluer les performances du codeur MELP.

La mise en œuvre du codeur MELP sur un processeur Texas Instruments de la famille C3x, introduit un retard estimé à 122.5 ms (algorithmic delay), une complexité de l'algorithme = 40 MIPS et une qualité de parole reconstruite avec un MOS = 3.2 pour un débit de 2.4 Kbit/s. En plus, de la robustesse de l'algorithme aux erreurs de canal et à l'interférence acoustique [96][97]. Par ailleurs, les tests subjectifs standards cités dans les références [99][100] par exemple, révèlent des notes DRT = 93.00 et DAM = 64.9 dans un environnement sans bruit, et dans le cas d'un bruit HMMWV (High Mobility Multipurpose Wheeled Vehicle : véhicule de transport léger à roues de l'armée Américaine) des notes DRT = 67.3 et DAM = 38.9.

La version C standard [101] de l'implémentation du Texas Instruments du codeur MELP a été compilée, et le fichier exécutable est appelé à partir de l'environnement Matlab.

La figure suivante présente un exemple d'une phrase phonétiquement équilibrée, prononcée par un locuteur masculin, codée avec le codeur MELP.

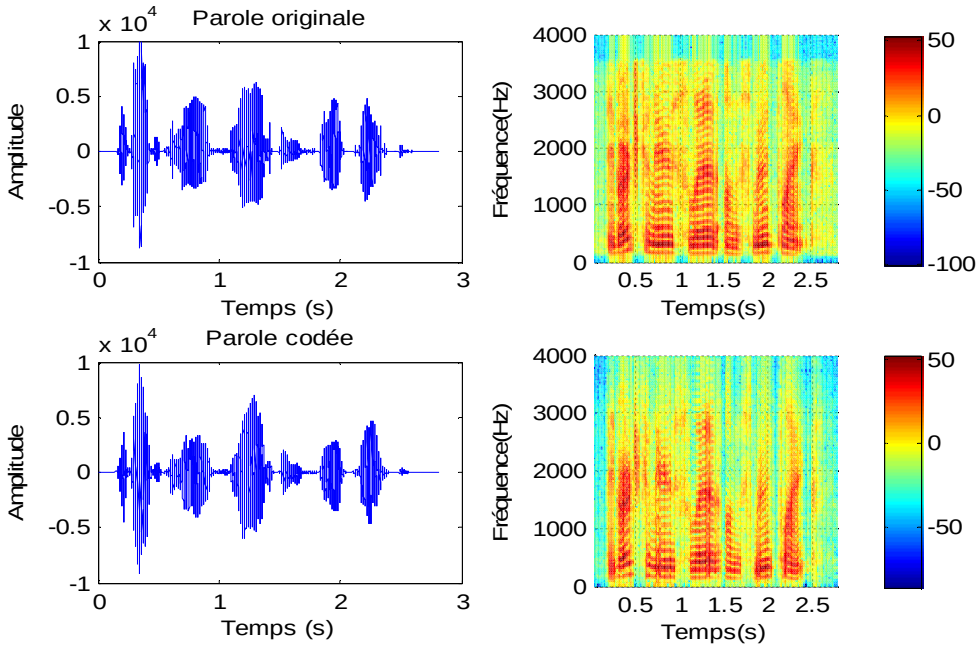


Figure 5.11 : Représentations temporelles et spectrogrammes des signaux de la parole originale et codée.

Les tests d'écoute effectués durant la simulation, la comparaison entre les représentations temporelles et les spectrogrammes du signal original et codé montrent que la parole reproduite par l'implémentation sous Matlab de ce codeur est intelligible et présente une qualité acceptable.

5.4 Description du codeur MELPe

La limitation des performances des codeurs bas débit dans des environnements bruités (y compris le codeur MELP), a donné lieu à l'idée de rehausser le signal bruité d'entrée avant son codage (figure 5.12).

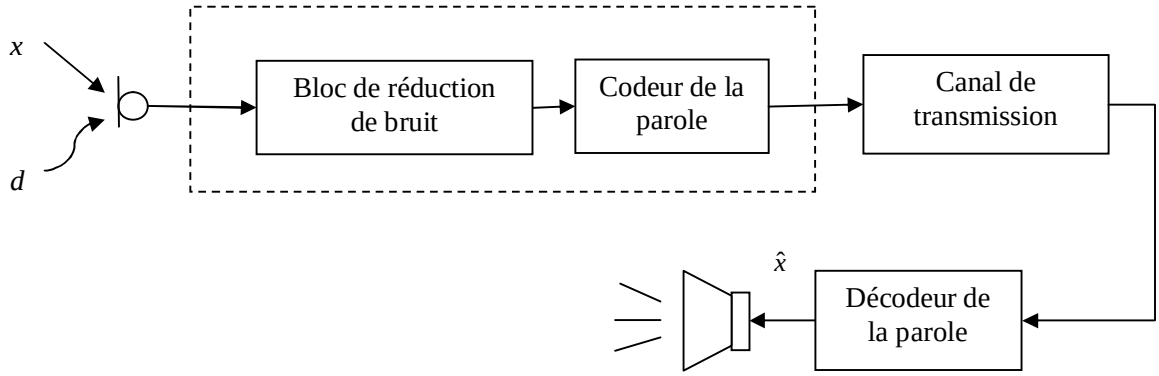


Figure 5.12 : Système de transmission de la parole avec bloc de réduction de bruit.

Un algorithme de rehaussement de la parole d'AT&T qui est l'aboutissement de plusieurs années de recherches [6][7][9] a été combiné avec le codeur MELP. Cet algorithme est basé sur la version MM-LSA comme règle de suppression du bruit, la méthode du MS pour l'estimation du bruit et quelques adaptations avec le codeur MELP. Cette combinaison entre l'algorithme de réduction de bruit et le codeur MELP est désignée par le codeur MELPe (enhanced MELP) [98]. Comme la plupart des notions sur le MM-LSA ont été détaillées dans le chapitre 4, seulement les améliorations adoptées de l'algorithme seront abordées dans les sections suivantes.

5.4.1 Bloc de réduction de bruit du codeur MELPe

Le schéma bloc de l'algorithme de réduction de bruit du codeur MELPe est illustré à la figure (5.13). Il est constitué d'un bloc d'estimation de l'amplitude spectrale à court terme (MM-LSA), un bloc d'estimation du SNR a priori ainsi que l'adaptation de ses limites et un bloc d'estimation de la puissance du bruit basée sur l'algorithme du minimum statistique.

Premièrement, le signal de la parole est divisé en trames de 22.5 ms, une FFT est appliquée par la suite pour fournir un accès au contenu fréquentiel du signal. Par la suite, un algorithme d'estimation basé sur la méthode du minimum statistique de Martin (annexe A), est utilisé pour modéliser le bruit dans les trames où la parole est absente. Cette partie de l'algorithme utilise un détecteur d'activité vocale pour permettre à l'algorithme de distinguer entre les trames composées de la parole + bruit, et celles composées du bruit seul. L'algorithme minimise l'erreur quadratique moyenne du logarithme de l'amplitude spectrale (MMSE-LSA), en plus de la version MM basée sur la probabilité de la présence de la parole. Une fois les règles de suppression sont appliquées sur la trame en cours, le signal rehaussé obtenu est utilisé comme entrée du codeur MELP.

Comme les codes sources du codeur MELPe et de son algorithme de réduction ne sont pas disponibles gratuitement, on a essayé d'implémenter sous MATLAB une version similaire à celle développée dans les articles consacrés au MELPe.

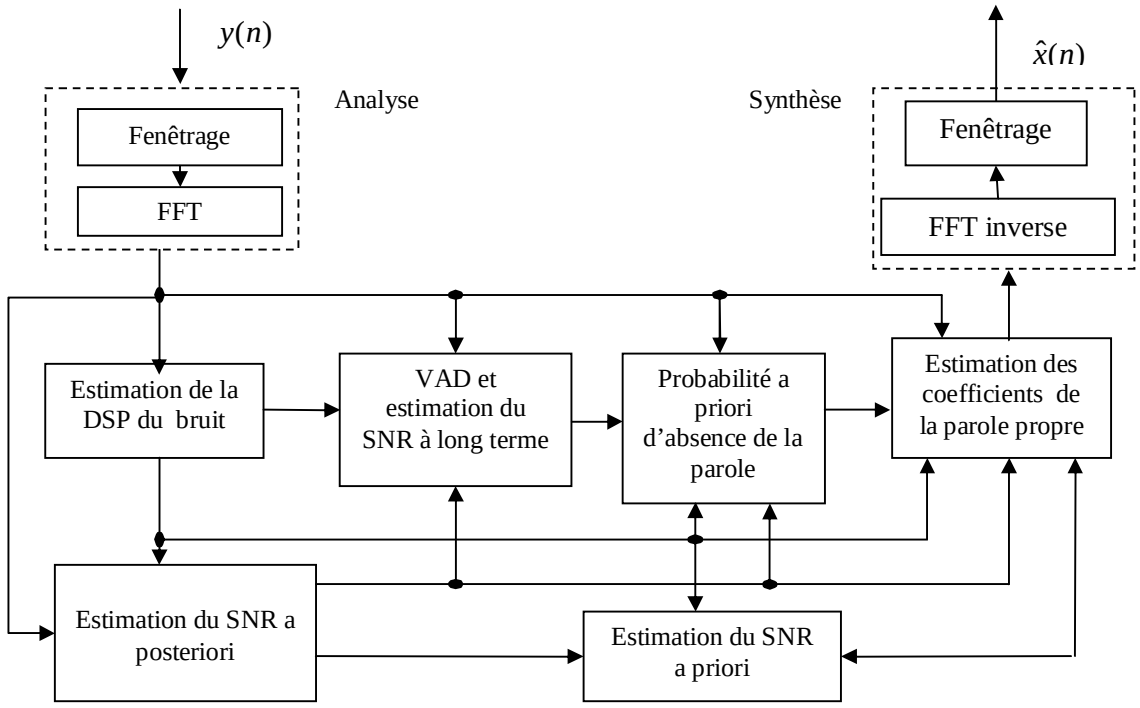


Figure 5.13 : Schéma bloc du prétraitement au codeur MELP [9].

L'essentiel des améliorations ajoutées à l'algorithme MM-LSA dans le bloc de réduction de bruit du codeur MELPe sont :

*** L'alignement des longueurs des trames**

Comme le codeur MELP utilise des trames de longueur 22.5 ms, c'est-à-dire M_C=180 échantillons pour une fréquence d'échantillonnage de 8 KHz, l'algorithme de rehaussement doit opérer sur une trame de M_E=M_C=180 échantillons. Ainsi, une FFT de taille 256 est utilisée avec un chevauchement de M_O=76 échantillons entre les trames adjacentes. Cette adaptation assure une résolution spectrale suffisante, des transitions lissées entre trames (moins de bruit), un faible retard et complexité moins pour le codeur.

*** Le calcul de la probabilité a priori de l'absence de la parole**

Pour calculer le gain (G_{MM}) de l'algorithme MM-LSA, une estimation fiable de cette probabilité est essentielle. L'article [89] donne les détails de cette estimation. Par ailleurs, l'utilisation des valeurs fixes (q=0.5) ou (q=0.3) [89] a donné des résultats étonnants et seront utilisées dans le cadre de ce projet.

*** La détection d'activité vocale et estimation du SNR à long terme**

L'estimation de la densité spectrale du bruit dans le cas du MELPe est basée sur la méthode du minimum statistique qui ne s'appuie pas sur le VAD. Néanmoins, il est utile d'avoir une VAD disponible pour contrôler certains aspects du pré-traitement, comme le contrôle de l'estimation de la probabilité a priori de l'absence de la parole et l'estimation à long terme du SNR.

Le SNR à long-terme (SNR_{LT}) caractérise le SNR du signal de parole bruité d'entrée moyenné sur plusieurs périodes de durée totale allant de 1 à 2 secondes. Il sera utilisé pour la limitation adaptative du SNR a priori, le calcul du SNR_{LT}(m) exploite la décision de la VAD, car la densité spectrale moyenne de la parole est mise à jours seulement si la parole est présente.

***Limitation adaptative du SNR a priori**

L'adaptation a le rôle d'améliorer l'estimation des paramètres LP et donc d'augmenter la précision de la robustesse du codeur. Pour éviter la structure de bruit musicale et pour avoir une bonne qualité et intelligibilité du signal de parole, il faut limiter ξ_k entre 0.1 et 0.2 [22][89], cela veut dire que moins d'atténuation du signal de parole est obtenue dans les zones où le SNR est faible (les vallées spectrales entre les formants). Cette atténuation a une mauvaise influence sur l'estimation des paramètres spectraux.

Martin et Cox ont présenté un système d'adaptation des limites du SNR a priori avec le quel une limite inférieur du SNR a priori est calculée [6], pendant les pauses de la parole, ξ prend une valeur minimale $\xi_{\min}$, typiquement, $\xi_{\min} = \xi_{\min P} = 0.15$; pendant l'activité vocale, cette valeur est changée suivant l'équation :

$$\xi_{\min 1}(m) = \xi_{\min P}\, 0.0067\left(0.5 + SNR_{LT}(m)\right)^{0.65} \tag{5.14}$$

Où : SNR_{LT} est le rapport signal sur bruit de la trame analysée.
Ainsi, ξ est limité par une valeur maximale $\xi_{\max}$, typiquement, $\xi_{\max}$ =0.25.

Avec des tests d'écoutes effectuées pour différentes valeurs du SNR, l'équation (5.14) peut être écrite par un système récursif de premier ordre suivant l'équation suivante :

$$\xi_{\min}(m) = 0.9\xi_{\min}(m-1) + 0.1\xi_{\min 1}(m) \tag{5.15}$$

Et donc le bruit de fond apparaît durant les pauses de la parole est naturel et de durée courte, cette méthode d'adaptation est efficace aussi pour les autres types de codeurs.

5.4.2 Description du bloc de réduction de bruit du codeur MELPeg

Lors de la conception d'un système de réduction de bruit pour le codeur MELP, il semble naturel d'utiliser différentes techniques de rehaussement de la parole. Comme déjà étudier dans le chapitre 4, les estimateurs basés sur des modèles super-Gaussiens assurent des performances meilleures en termes d'intelligibilité et de qualité par rapport à l'estimateur MM-LSA, surtout Les variantes de l'estimateur obtenues avec une modélisation des coefficients de la TFD du signal de parole propre avec une distribution Gamma généralisée. Ainsi, l'estimateur $\hat{A}^{(2)}$ obtenu avec $\gamma = 2$ sera utiliser dans un bloc de réduction de bruit en combinaison avec le codeur MELP, le système complet composé du bloc de pré-traitement et du codeur MELP sera désigné dans les simulations par le codeur MELPeg pour (enhanced gamma MELP). A noter, que les adaptations et les limites utilisées dans le codeur MELPe seront conservées dans le MELPeg. Donc, la seule différence réside dans les règles de suppression du bruit.

5.4.3 Tests et résultats

Dans cette section, nous présentons une évaluation des performances des codeurs MELP, MELPe et MELPeg. Les performances sont d'abord évaluées soit avec de la parole propre soit avec de la parole bruitée. Le codeur MELP est utilisé pour coder directement les signaux sans bloc de réduction de bruit et les deux algorithmes de réduction de bruit du MM-LSA et $\hat{A}^{(2)}$ seront exploités dans les blocs de pré-traitement des codeurs MELPe et MELPeg respectivement. La même base de données et les mêmes conditions de simulation utilisées dans le chapitre précédent sont utilisées avec une trame de 22.5 ms dans ce cas.

5.4.3.1 Evaluations des performances

Généralement, pour évaluer les performances du codeur MELP, il est préférable d'utiliser des tests subjectifs telle que DRT, DAM, AB et le test MOS, malheureusement, ces tests sont difficiles à mettre en œuvre puisque il nécessitent de mobiliser beaucoup de moyens (temps, personnes et argent), donc ils ne sont pas implémentés dans le cadre de notre travail. En se basant pour évaluer la qualité, sur des mesures objectives, des tests d'écoute et des représentations temporelles ainsi que les spectrogrammes.

Les mesures objectives qui sont utilisées sont le PESQ, IMPSNR (Improvement SNR) où le IMPSNR représente l'amélioration en SNRseg définit par :

$$\text{IMPSNR= SNRseg (rehaussé codé)-SNRseg (bruité codé)} \qquad (5.16)$$

Les tableaux suivants nous donnent les résultats de test des mesures IMPSNR et PESQ pour des signaux dégradés par un bruit blanc, babble, et de car pour le niveau du bruit 0 dB et 5 dB, rehaussés par l'algorithme MM-LSA et $\gamma^{(2)}$ et codés par le codeur MELP.

	Bruit Blanc		**Babble**		**Car**	
Type	0 dB	5 dB	0 dB	5 dB	0 dB	5 dB
MELP	4.7594	3.8720	4.6664	3.8084	4.7531	3.8873
MELPe	4.4975	3.9413	4.6250	4.1116	4.6327	4.0348
MELPeg	5.9786	5.0318	6.0505	4.9540	6.0902	5.1359

Tableau 5.2 : Résultats de la mesure IMPSNR pour des signaux dégradés par des bruits (Blanc, Babble, Car), rehaussés par l'algorithme MM-LSA et $\gamma^{(2)}$ puis codés par le codeur MELP.

		Bruit Blanc		**Babble**		**Car**	
Type	Parole propre	0 dB	5 dB	0 dB	5 dB	0 dB	5 dB
MELP	2.8196	1.3571	1.5367	1.3464	1.6546	1.3472	1.5698
MELPe		2.0732	2.3553	2.1705	2.4325	2.1229	2.3789
MELPeg		2.0989	2.3182	2.1055	2.3170	2.1056	2.3095

Tableau 5.3 : Résultats de la mesure PESQ pour des signaux dégradés par des bruits (Blanc, Babble, Car), rehaussés par l'algorithme MM-LSA et $\gamma^{(2)}$ puis codés par le codeur MELP.

5.4.3.2 Interprétations

Pour les deux codeurs, les types de bruit et le rapport signal sur bruit utilisé, on remarque une amélioration acceptable de la qualité en termes de mesures par rapport à celle obtenue par le codeur MELP, donc, ils présentent une qualité acceptable et robuste dans les environnements bruités.

La valeur de la mesure PESQ dans le tableau (5.2), confirme que les deux codeurs assurent des résultats plus proches.

Pour les faibles valeurs du SNR, le codeur MELPe présente des mauvais résultats de l'amélioration en SNRseg pour tous les types de bruit comparativement avec le codeur MELPeg.

De bons résultas de l'amélioration du SNRseg sont obtenus dans le cas du codeur MELPeg pour les différents types de bruit et les deux valeurs du SNR globale.

Ainsi, des représentations temporelles des formes d'onde et les spectrogrammes sont nécessaire pour confirmer les résultats de tests d'écoute et des mesures objectives, la figure suivantes illustre les représentations temporelles des formes d'onde des signaux de la parole propre codée par le codeur MELP et bruitée codée par le codeur MELPe et MELPeg sous certains conditions de bruit pour la 10ème phrase de la base de données.

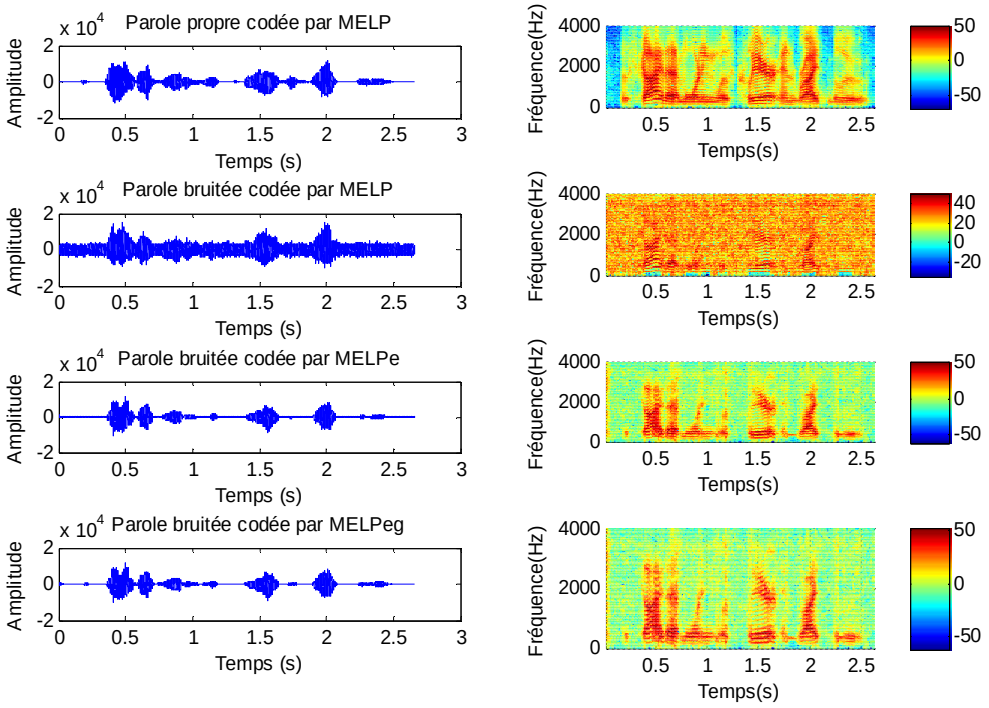


Figure 5.14 : Représentations temporelles et spectrogrammes des signaux de parole propre codés par le codeur MELP et bruités codés par le codeur MELPe et MELPeg (bruit blanc, SNR=0dB).

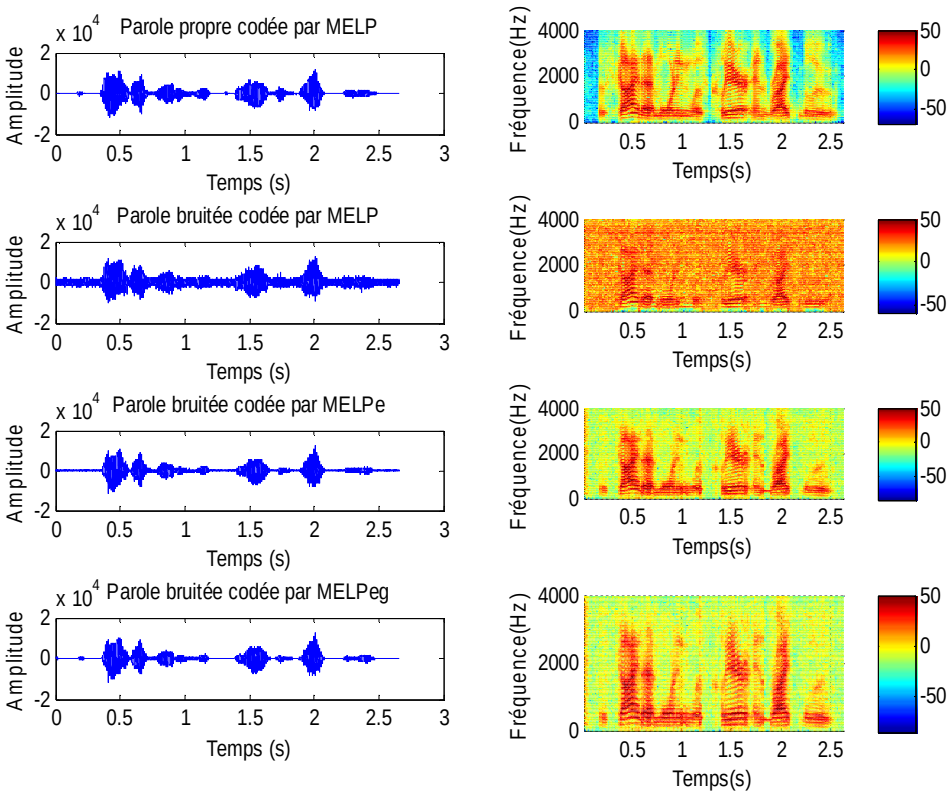


Figure 5.15 : Représentations temporelles et spectrogrammes des signaux de parole propre codés par le codeur MELP et bruités codés par le codeur MELPe et MELPeg (bruit blanc, SNR=5dB).

5.5 Conclusion

Nous avons exposé dans ce chapitre, au début une description détaillée du codeur MELP, le modèle d'excitation mixte utilisé ainsi que les différents paramètres supplémentaires spécifiques à ce codeur par rapport aux autres codeurs qui le précèdent. Par la suite, les deux variantes MELPe et MELPeg adaptées aux milieux hostiles ont été introduites.

Nous avons également évalué les performances des codeurs MELP, MELPe et MELPeg à 2.4 Kbit/s dans de nombreuses conditions opérationnelles, nous avons constaté que la qualité de la parole reproduite reste tout à fait acceptable, et que la version MELPeg présente des performances meilleures en terme de qualité et d'intégrabilité comparativement à celle du codeur MELPe.

CONCLUSION GENERALE & PERSPECTIVES

L'objet de cette thèse était d'explorer, d'approfondir, d'appliquer et d'évaluer différentes méthodes de rehaussement de la parole pour produire un signal de parole rehaussé qui sera codé dans un contexte de bas débit (autour de 2400 bits/s), où la parole synthétisée doit être de bonne qualité.

Cette thèse a permis de perfectionner le codeur MELP dans les environnements bruités. Les principales contributions ont porté sur la modélisation efficace du bruit en utilisant le filtre de Kalman, la réduction du bruit a été rendue plus efficace au moyen des méthodes STSA, en particulier celles basées sur des approches Bayesiennes, et enfin, la variante MELPeg proposée qui assure un codage à 2400 bits/s meilleur par rapport aux codeurs MELP et MELPe.

L'application du filtre de Kalman au rehaussement de la parole a été étudiée dans le cas d'un bruit blanc et les bruits réels colorés à des rapports signal sur bruit différents. Les mesures obtenues dans le cas du bruit blanc sont nettement supérieures aux mesures obtenues dans le cas des bruits réels, car les calculs de base du filtre de Kalman sont basés sur l'hypothèse que le bruit additif est un bruit blanc. De plus, pour un bruit coloré réel, en plus de la modélisation de la parole, une modélisation du bruit avec un modèle AR d'ordre q = 4 (bruit de voiture, bruit de train, bruit dans une station de train, bruit dans une rue), q = 6 (bruit à l'aéroport) et q = 8 (bruit au restaurant, bruit d'exhibition, bruit babble) est nécessaire pour assurer des meilleures performances.

Quelques résultats de l'application de l'algorithme EM au rehaussement de la parole sont présentés dans le cas d'un bruit blanc additif, pour la détermination des paramètres AR à partir du signal bruité. Les résultats obtenus confirment que plus le nombre d'itérations augmente, les résultats s'améliorent d'avantage, une taille de 10 ms permet d'avoir de bon résultats à l'itération k=2 mais avec un temps de calcul élevé et pour une taille de 20 ms, un nombre d'itérations k=4 est suffisant dans ce cas avec un temps de calcul moins par rapport au précédent.

Dans la catégorie des approches basées sur le modèle Gaussien, nous avons montré par simulation que l'estimateur MM-LSA est le plus performant parmi ces approches, en utilisant un corpus de parole avec différents types de bruits et à des SNRs variés. Par ailleurs, les nouveaux estimateurs au sens du MMSE basés sur les modèles super-Gaussiens ont été étudiés, présentés et adaptés à notre situation. Les résultats obtenus lors des différentes simulations, montrent bien que les deux variantes $\hat{A}_{C,K}^{(1)}$ et $\hat{A}^{(2)}$ présentent une bonne réduction du bruit comparativement aux estimateurs basés sur des modèles Gaussiens. De plus, l'estimateur $\hat{A}^{(2)}$ semble bien adapté aux codeurs de la parole dans un milieu bruité pour ses performances et son temps d'exécution réduit. Les mesures objectives, les tests d'écoute et les formes d'ondes confirment que cet estimateur est efficace à la réduction du bruit résiduel et à l'élimination du bruit musical.

Le travail réalisé a été entrepris dans le cadre d'une étude globale sur le codage paramétrique de la parole à bas débit d'une part, et de combiner ce bloc de codage avec un bloc de réduction de bruit, pour former un codeur robuste dans les environnements bruités d'autre part.

Au niveau du codage, nous avons d'écrit les concepts généraux du codage de parole, où une étude détaillée sur le codeur MELP a été présentée, surtout les différents blocs de la partie codage et de décodage ainsi que ses sous blocs utilisés par la version standard implémentée en langage C.

Nous avons testé par la suite un codeur (décodeur) de la parole « MELP » à bande étroite qui utilise une excitation mixte pour un débit de 2.4 Kbit/s. Les tests d'écoute et les représentations de la forme d'onde montrent que la qualité de la parole codée est de bonne qualité dans les conditions normales, dans le cas d'une seule phrase ou avec un corpus de parole. Par la suite, les deux variantes MELPe et MELPeg adaptées aux milieux hostiles ont été introduites.

L'évaluation des performances des codeurs MELP, MELPe et MELPeg à 2.4 Kbit/s dans de nombreuses conditions opérationnelles, nous a permet de constater que la qualité de la parole reproduite reste tout à fait acceptable, et que la version MELPeg présente des performances meilleures en terme de qualité et d'intégrabilité comparativement à celle du codeur MELPe.

Un autre travail sur les méthodes perceptuelles de réduction du bruit a été réalisé. Il n'a pas été présenté dans les chapitres de la thèse, mais il a fait l'objet d'une contribution dans un journal [102]. L'objectif du débruitage perceptuel est de réduire le bruit sans apporter plus de distorsion sur le signal de parole. En exploitant les propriétés des estimateurs au sens du MMSE basés sur les modèles super-Gaussiens et les améliorations assurées par l'utilisation des propriétés de masquage fréquentiel, nous avons obtenus des améliorations importantes en termes de mesures objectives.

Perspectives :

Le travail présenté dans cette thèse peut être approfondi et enrichi de plusieurs façons :

- L'incorporation de ces systèmes de réduction de bruit dans les blocs de pré-traitement de d'autres codeurs de la parole.
- Le test et l'évaluation du comportement du codeur en présence d'erreurs de canal, aléatoires ou en paquets, et la proposition des correctifs nécessaires.
- Détermination de d'autres règles de suppression de bruit sous l'hypothèse que le bruit additif est non stationnaire ou n'est pas Gaussien.
- Etude et implémentation des procédures de détermination des paramètres des codeurs plus efficaces directement de l'observation bruitée, sans utilisation des blocs de pré-traitement.
- Utilisation des nouvelles variantes non linéaires du filtre de Kalman (EKF, UKF), pour une meilleure estimation des paramètres non linéaires du codeur et dans le rehaussement.
- La recherche de nouvelles approches de VAD plus simples et efficaces.
- L'insertion d'une approche Bayesienne simple dans les prothèses auditives.
- Etude plus rigoureuse de l'influence des paramètres γ et ν sur les performances des estimateurs super-Gaussiens basés sur la distribution Gamma généralisée.
- …

Bibliographie

[1] René Boite, Hervé Bourlard, Thierry Dutoit, Joil Hancq et Henri Leich, *Traitement de la parole*, Presses Polytechniques et Universitaires Romandes, Lausanne, Suisse, 488p, 1999.

[2] A. M. Kondoz, *Digital speech-coding for low bit rate communications systems*, John Wiley & Sons, UK, 456p, 1994.

[3] Wai C. Chu, *Speech coding algorithms-foundation and evolution of standardized coders*, Wiley-Interscience, USA, 592p, 2003.

[4] A. S. Spanias, "Speech coding: A tutorial review," in *Proceedings of the IEEE*, vol. 82, no. 10, pp. 1541-1582, October 1994.

[5] M. R. Schroeder and B. Atal, "Code-Excited Linear Prediction (CELP): High quality speech at very low bit rates," in *Proceedings of ICASSP'85*, Tamp, Fla, USA, pp. 937-940, April 1985.

[6] R. Martin and R. Cox, "New speech enhancement techniques for low bit rate speech coding," in *Proceedings of the IEEE Workshop on Speech Coding*, Porvoo, Finland, pp. 165-167, May 1999.

[7] A. J. Accardi and R. V. Cox, "A modular approach to speech enhancement with an application to speech coding," in *Proceedings of ICASSP'99*, Phoenix, Arizona, USA, pp. 201-204, 15-19 March 1999.

[8] T. Agarwal, "Pre-processing of noisy speech for voice coders," Master Thesis, McGill University, Montreal, Canada, January 2002.

[9] R. Martin, D. Malah, R. V. Cox, and A. J. Accardi, "A noise reduction preprocessor for mobile voice communication," *EURASIP Journal on Applied Signal Processing*, pp. 1046-1058, 2004.

[10] P. Vary and R. Martin, *Digital speech transmission*, John Wiley & Sons, UK, 644p, 2006.

[11] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-27, no. 2, pp. 113-120, April 1979.

[12] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proceedings of ICASSP'79*, Washington, DC, USA, vol. 4, pp. 208-211, April 1979.

[13] P. Lockwood and J. Boudy, "Experiments with a Non-linear Spectral Subtractor (NSS), Hidden Markov Models and the projections, for robust recognition in cars," *Speech Communications*, vol. 11, nos. 2-3, pp. 215-228, June 1992.

[14] N. Virag, "Speech enhancement based on masking properties of the human auditory system," Master Thesis, Swiss Federal Institute of Technology, Swiss, 1996.

[15] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, "A Parametric formulation of the generalized spectral subtraction method," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 328-337, July 1998.

[16] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proceedings of ICASSP'2002*, Orlando, Florida, USA, vol. 4, pp. 4160-4164, May 13-17, 2002.

[17] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-28, no. 2, pp. 137-145, April 1980.

[18] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proceedings of ICASSP'96*, Atlanta, Georgia, USA, vol. 2, pp. 629-632, May 1996.

[19] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME - Journal of Basic Engineering*, vol. 82, pp. 35-45, 1960.

[20] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109-1121, December 1984.

[21] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-33, pp. 443-445, April 1985.

[22] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech, and audio Processing*, vol. 2, no. 2, pp. 345-349, 1994.

[23] R. Martin, "Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors," in *Proceedings of ICASSP'02*, Orlando, Florida, USA, pp. 253-256, May 13-17, 2002.

[24] R. Martin, "Statistical methods for the enhancement of noisy speech," in *Proceedings of IWAENC'2003*, Kyoto, Japan, pp. 1-6, September 8-11, 2003.

[25] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," *EURASIP Journal on Applied Signal Processing*, vol. 7, pp. 1110-1126, May 2005.

[26] R. Martin, "Speech enhancement based on minimum mean-square error estimation and super-Gaussian priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845-856, September 2005.

[27] I. Andrianakis and P. R. White, "MMSE speech spectral amplitude estimators with Chi and Gamma speech priors," in *Proceedings of ICASSP'06*, Toulouse, France, vol. III, pp. 1068-1071, May 2006.

[28] R. C. Hendriks, J. S. Erkelens, J. Jensen, and R. Heusdens, " Minimum mean-square error amplitude estimators for speech enhancement under the generalized Gamma distribution," in *Proceedings of IWAENC'06*, Paris, France, September 2006.

[29] J. Jensen, R. C. Hendriks, and J. S. Erkelens, "MMSE estimation of discrete Fourier coefficients with generalized Gamma priori," Technical Report, Delft University of Technology, Netherlands, 2006.

[30] J. Erkelens, J. Jensen, and R. Heusdens, "Improved speech spectral variance estimation under the generalized Gamma distribution," in *Proceedings of the SPS-DARTS, the third annual IEEE BENELUX/DSP valley signal processing symposium*, Antwerp, Belgium, pp. 43-46, 2007.

[31] R. C. Hendriks, "Advances in DFT-based single-microphone speech enhancement," PhD Thesis, Delft University of Technology, Netherlands, February 2008.

[32] J. S. Erkelens, R. C. Hendriks, and R. Heusdens, "On the estimation of complex speech DFT coefficients without assuming independent real and imaginary parts," *IEEE Signal Processing Letters*, vol. 15, pp. 213-216, 2008.

[33] R. C. Hendriks, R. Heusdens, and J. Jensen, "Log-spectral magnitude MMSE estimators under super-Gaussian densities," in *Proceedings of: INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association*, Brighton, United Kingdom, pp. 1319-1322, September 6-10, 2009.

[34] R. C. Hendriks, R. Heusdens, and J. Jensen, "On robustness of multi-channel minimum mean-squared error estimators under super-Gaussian priors," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA'09*, New Paltz, NY, USA, pp. 157-160, October 18-21, 2009.

[35] Chabane Boubakir et Daoud Berkani, "Approche mono voie de réduction de bruit basée sur la soustraction spectrale et un modèle statistique de la VAD," *TAIMA'05*, Hammamet, Tunisie, 26 Septembre 2005.

[36] Chabane Boubakir, Daoud Berkani and Francis Grenez, "A Frequency-Dependent speech enhancement methods," in *Proceedings of the Third International Summer School on Signal Processing and its Applications (I3SPA'06)*, JIJEL, ALGERIA, JULY 2006.

[37] Chabane Boubakir et Daoud Berkani, "Application du filtre de Kalman au rehaussement de la parole," in *Proceedings of the 2nd International Conference on Electrical and Electronics Engineering (ICEEE2008)*, LAGHOUAT, ALGERIA, April 21-23, 2008.

[38] Chabane Boubakir and Daoud Berkani, "On the use of Kalman filter for enhancing speech corrupted by colored noise," *the WSEAS Transactions on Signal Processing Journal*, Issue 12, Volume 4, December 2008.

[39] Chabane Boubakir and Daoud Berkani, "The estimation of line spectral frequencies trajectories based on unscented Kalman filtering," in *Proceedings of the 6th International Multi-Conference on Systems, Signals and Devices SSD'09, Conference on Communication and Signal Processing (CSP)*, DJERBA, TUNISIA, March 2009.

[40] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, Prentice-Hall, Inc, 1975.

[41] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314-323, Feb. 1988.

[42] E. Zwicker and H. Fastl, *Psychoacoustics: facts and models*, Springer, 2nd edition, Berlin, Heidelberg, New York, 475p, 1999.

[43] Laurent Buniet, "Traitement automatique de la parole en milieu bruité : étude de modèles connexionnistes statiques et dynamiques," Thèse de Doctorat, Université Nancy 1, France, 1997.

[44] Kotta Manohar, "Single channel enhancement of noisy speech," M.Tech. Credit Seminar Report, Electronic Systems Group EE Dept, ITT Bombay, Nov 2002.

[45] Signal Processing Information Base, "Noise data," March 2003, [Online], Available: http://spib.rice.edu/spib/select_noise.html.

[46] http://www.utdallas.edu/~loizou/speech/noizeus/

[47] D. Middleton and R. Esposito, "Simultaneous optimum detection and estimation of signal in noise," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 434-444, 1968.

[48] Marie Guéguin, "Évaluation objective de la qualité vocale en contexte de conversation," Thèse de Doctorat, Université de Rennes 1, France, Décembre 2006.

[49] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality*, Prentice-Hall, N.J, USA, 377p, 1988.

[50] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithm," in *Proceedings of ICSLP*, Sydney, Australia, pp. 2819-2822, December 1998.

[51] UIT-T Rec. P.862 (2001). Évaluation de la qualité vocale perçue : méthode objective d'évaluation de la qualité vocale de bout en bout des codecs vocaux et des réseaux téléphoniques à bande étroite.

[52] J. Berger, "Requirements for a new model for objective speech quality assessment P.OLQA. Contribution," UIT-T COM 12-D.75, 2005.

[53] B. Widrow and S. D. Stearns, *Adaptive signal processing*, Prentice Hall, Englewood Cliffs, N.J, 474p, 1985.

[54] J. S. Lim and A. V. Oppenheim, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 4, pp. 354-358, 1978.

[55] M. Sambur, "Adaptive noise canceling for speech signals," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-26, no. 5, pp. 419-423, October 1978.

[56] H. L. Van Trees, *Detection, Estimation and Modulation Theory*, New York: Wiley, 697p, 1968.

[57] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, pp. 197-210, June 1978.

[58] K. K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proceedings of ICASSP'87*, pp. 177-180, 1987.

[59] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Transactions on Signal Processing*, vol. 39, pp. 1732-1742, August 1991.

[60] M. Gabrea, "Adaptive Kalman filtering-based speech enhancement algorithm," in *Proceedings of Canadian Conference on Electrical and Computer Engineering*, Federicton, New-Brunswick, Canada, vol. 1, pp. 521-526, 2001.

[61] N. Ma, M. Bouchard, and R. A. Goubran, "Perceptual Kalman filtering for speech enhancement in colored noise," in *Proceedings of ICASSP'04*, pp. 717-720, 2004.

[62] M. Gabrea, "Robust adaptive Kalman filtering-based speech enhancement algorithm," in *Proceedings of ICASSP'04*, pp. 301-304, 2004.

[63] V. Grancharov, J. Samuelsson, and W. B. Kleijn, "Improved Kalman filtering for speech enhancement," in *Proceedings of ICASSP'05*, pp. 1109-1112, 2005.

[64] T. Painter and A. Spanias, "Perceptual coding of digital audio," in *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451-513, April 2000.

[65] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," in *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586-1604, December 1979.

[66] D. L. Wang and J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679-681, 1982.

[67] Mohamed Najim, "Filtrage Optimal," *Techniques de l'Ingénieur, Traité mesure et contrôle*, Article no. R7228.

[68] J. Sohn, N. S. Kim, and W. Sung, "A statistical model – based voice activity detection," *IEEE, Signal Processing letters*, vol. 6, no. 1, January 1999.

[69] Simon Robidas, "Comparaison de méthodes pour la détection d'activité vocale à bande large sous différents bruits," Maîtrise ès sciences, Ecole d'Ingénierie et de Technologie de l'Information (ÉITI), Université d'Ottawa, Canada, 2006.

[70] Thomas Fillon, "Traitement numérique du signal acoustique pour une aide aux malentendants," Thèse de Doctorat, Ecole Nationale Supérieure des Télécommunications de Paris, France, Décembre 2004.

[71] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. Eur. Signal Processing Conf. (EUSIPCO)*, pp. 1182-1185, 1994.

[72] H. G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proc. 20th IEEE. ICASSP'95*, Detroit, Michigan, USA, pp. 153-156, May 8-12, 1995.

[73] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on speech and audio processing*, vol. 9, no. 5, pp. 504-512, July 2001.

[74] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12-15, 2002.

[75] R. Sundarrajan, "Noise estimation algorithms for highly non stationary environments," Master Thesis, The University of Texas at Dallas, USA, August 2004.

[76] S. Rangachari, P. C. Loizou, and Y. Hu, "A noise estimation for highly non-stationary environments," *Speech communication*, pp. 220-231, August 2005.

[77] S. L. Lauritzen, "Thiele: Pioneer in Statistics," Oxford University Press, Oxford, 288p, 2002.

[78] R. E. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *Transactions of the ASME - Journal of Basic Engineering*, vol. 83, pp. 95-107, 1961.

[79] M. S. Grewal and A. P. Andrews, *Kalman filtering theory and practice using MATLAB*, 2nd edition, John Wiley & Sons, Canada, pp. 1-12, 2001.

[80] http://www.cs.unc.edu/~tracker/media/pdf/SIGGRAPH2001_CoursePack_08.pdf

[81] H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum likelihood estimates of linear dynamic systems," *AIAA Journal*, vol. 3, pp. 1445-1450, 1965.

[82] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *Journal of Time Series Analysis*, vol. 3, no. 4, pp. 253-264, 1982.

[83] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech Audio Process*. vol. 6, no. 4, pp. 373-385, 1998.

[84] "Objective measurement of active speech level," ITU-T Recommendation P.56, March 1993.

[85] S. J. Julier and J. K. Uhlmann, "A new extension of the Kalman filter to nonlinear systems," in *Proceedings of AeroSense: The 11th Int. Symp. On Aerospace/Defence Sensing, Simulation and Controls*, Orlando, Florida, USA, pp. 153-158, April 1997.

[86] R. Van der Merwe and E. A. Wan, "The square-root unscented Kalman filter for state and parameter estimation," in *Proceedings of ICASSP'01*, pp. 3461-3464, 2001.

[87] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," in *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401-422, March 2004.

[88] A. Errity, J. McKenna, and S. Isard, "Unscented Kalman filtering of line spectral frequencies," in *Proceedings of the Int. Conf. on Spoken Language Processing (Interspeech 2004 - ICSLP)*, Jeju, Korea, pp. 2697-2700, October 2004.

[89] D. Malah, R. Cox, and A. Accardi, "Tracking speech presence uncertainty to improve speech enhancement in non stationary noise environments," in *Proceedings of the IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Phoenix, Ariz, USA, vol. 2, pp. 789-792, Mar 1999.

[90] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, Academic Press, Inc., 6 edition, 1216p, 2000.

[91] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, Dover, New York, ninth dover printing, tenth gpo printing edition, 1046p, 1964.

[92] W. Rudin, *Real and complex analysis*, McGraw-Hill, New-York, 3 edition, 483p, 1987.

[93] Matlab routines for computation of special functions. http://ceta.mit.edu/ceta/comp_spec_func/

[94] J. Makhoul, V. Viswanathan, R. Schwarz, and A. W. F. Huggins, "A mixed source model for speech compression and synthesis," *Journal of the Acoustical Society of America*, vol. 64, pp. 1577-1581, December 1978.

[95] A. V. McCree and T. P. Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding,'' *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 242-250, July 1995.

[96] A. V. McCree, L.M. Supplee, R. P. Cohn, and J. S. Collura, "MELP: The new federal standard at 2400 bps," in *Proceedings of ICASSP'97*, vol. 2, pp. 1591-1594, April 1997.

[97] "Analog-to-Digital Conversion of voice by 2400 bit/second Mixed Excitation Linear Prediction (MELP)," MIL-STD-3005, Dec 1999.

[98] T. Wang, K. Koishida, V. Cuperman, A. Gersho, and J. Collura, "A 1200/2400 bps coding suite based on MELP," in *Proceedings of IEEE Workshop on Speech Coding*, Tsukuba, Japan, pp. 90-92, October 2002.

[99] J. Collura, "Speech enhancement and coding in harsh acoustic noise environments," in *Proceedings of the IEEE Workshop on Speech Coding*, Porvoo, Finland, pp. 162-164, June 1999.

[100] Department of Defence Digital Voice Processing Consortium, Specifications for the Analog to Digital Conversion of Voice by 2.400 Bits/Second Mixed Excited Linear Prediction, Internal report, May 1998.

[101] MELP source code link :
http://health.tau.ac.il/Communication%20Disorders/noam/speech/melp/Download/Download.htm

[102] Chabane Boubakir and Daoud Berkani, "An improved MMSE amplitude estimator under generalized Gamma distribution based on auditory perception", *Mathematical Problems in Engineering*, Volume 2013, Article ID821760, 7 pages.
http://www.hindawi.com/journals/mpe/2013/821760/

ANNEXE A

Méthode d'estimation du bruit basée sur les statistiques minimales de Martin

La méthode de base de la majorité des algorithmes d'estimation du spectre du bruit est sans doute celle proposée par Martin en 1994 [71] et sa version améliorée en 2001 [73]. Elle est appelée «MS : Minimum Statistics method », basée sur les statistiques minimales, c'est-à-dire à estimer la variance du bruit à partir des points de plus petite énergie, ceux-ci étant alors supposés ne provenir que du bruit. Il utilise la plus petite valeur du plan temps – fréquence après un lissage récursif.

A.1 Principes de l'algorithme MS

Comme la densité spectrale de puissance de la parole bruitée est égale à la somme de la puissance du bruit et celle de la parole propre, la variance du bruit est estimée par le suivi du minimum de la densité spectrale de puissance de la parole bruitée durant une fenêtre de longueur fixe. Cette longueur est choisie de manière à englober le plus large pic dans n'importe quel signal de parole. Expérimentalement, il a été admis que les fenêtres de tailles approximativement égales à 0.8s jusqu'à 1.4s ont donné de bons résultats.

Pour rechercher le minimum, une version récursive de premier ordre de la densité spectrale de puissance de la parole bruitée est utilisée, elle est donnée par :

$$S(k,l) = \alpha S(k,l-1) + (1-\alpha)\left|Y(k,l)\right|^2 \qquad \text{(A.1)}$$

Où α est la constante de lissage. Pour améliorer les performances du processus de suivi du minimum, des modifications ont été faites. Ces modifications consistent en :

- Remplacement du facteur de lissage constant dans l'équation (A.1) par un facteur de lissage dépendant du temps et de la fréquence.
- Dérivation d'un facteur de biais qui assure une estimation sans biais.
- Amélioration de la vitesse du suivi de l'algorithme, pour les changements des niveaux du bruit.

A.2 Dérivation du facteur de lissage optimal dépendant du temps et de la fréquence

Le paramètre de lissage utilisé dans l'équation (A.1) doit être faible pour permettre le suivi du non stationnarité du signal de la parole. D'autre part, il doit être presque égal à l'unité ($\alpha \cong 1$) pour garder la variance du minimum aussi petite que possible. Par conséquent, le facteur de lissage fixe sera remplacé par un facteur de lissage dépendant du temps et de la fréquence.

Ceci est dérivé durant les régions d'absence de la parole, où le spectre de puissance $S(k,l)$ est supposé égal à la variance du bruit $\lambda_d(k,l)$. Par conséquent, le paramètre de lissage sera dérivé en minimisant l'erreur quadratique moyenne entre $S(k,l)$ et $\lambda_d(k,l)$ comme suit :

$$\mathrm{E}\left\{(S(k,l)-\lambda_d(k,l))^2 / S(k,l-1)\right\} \qquad \text{(A.2)}$$

Où :

$$S(k,l) = \alpha(k,l)S(k,l-1) + (1-\alpha(k,l))\left|Y(k,l)\right|^2 \qquad \text{(A.3)}$$

Il est indispensable de noter que dans l'équation (A.3) le facteur de lissage dépendant du temps et de la fréquence. $\alpha(k,l)$ a été utilisé au lieu du facteur fixe α défini dans l'équation (A.1). En remplaçant (A.3) dans (A.2) et en annulant la dérivée première, cela donnera la valeur optimale de $\alpha(k,l)$:

$$\alpha_{opt}(k,l) = \frac{1}{1+(S(k,l-1)/\lambda_d(k,l)-1)^2} \tag{A.4}$$

Le terme $S(k,l-1)/\lambda_d(k,l) = \overline{\gamma}(k,l)$ est une version lissée du SNR a posteriori.

$$\gamma(k,l) = \frac{|Y(k,l-1)|^2}{\lambda_d(k,l)} \tag{A.5}$$

Comme la dérivation précédente est réalisée sous la supposition que la parole est absente, cela ne pose aucun problème principal. La procédure du facteur optimal réagit à l'activité vocale de la même manière que pour les bruits fortement non stationnaires. Dans le cas d'activité vocale, le facteur de lissage prend des valeurs faibles, qui permettent à $S(k,l)$ de suivre les variations de $|Y(k,l)|^2$. Dans l'implémentation pratique, on doit remplacer $\lambda_d(k,l)$ par son estimation précédente $\hat{\lambda}_d(k,l-1)$ pour calculer le facteur optimal, et on impose une limite maximale $\alpha_{\max} = 0.96$. Un autre facteur de correction $\alpha_C(l)$ est déterminé par l'auteur, qui est calculé à partir du rapport entre le périodogramme lissé moyenné et la variance du bruit estimé. Le paramètre de lissage final après correction est donné par :

$$\hat{\alpha}(k,l) = \frac{\alpha_{\max}\alpha_c(l)}{1+(S(k,l-1)/\hat{\lambda}_d(k,l-1)-1)^2} \tag{A.6}$$

Où :

$$\alpha_c(l) = 0.7\alpha_c(l-1) + 0.3\max(\tilde{\alpha}_c(l), 0.7) \tag{A.7}$$

Et :

$$\tilde{\alpha}_c(l) = \frac{1}{1+(\sum_{k=0}^{L-1} S(k,l-1)/\sum_{k=0}^{L-1}|Y(k,l)|^2 - 1)^2} \tag{A.8}$$

Une limite $\alpha_{\min}$ inférieure à α_{opt} est utilisée afin de réduire la variance du facteur de correction du biais et d'améliorer les performances de l'estimateur du bruit dans le cas des bruits non stationnaires, et elle est donnée par :

$$\alpha_{\min} = \min(0.3, SNR^{-\frac{L/2}{0.064 f_s}}) \tag{A.9}$$

Avec SNR est le rapport signal sur bruit global, L la taille de la FFT et f_s la fréquence d'échantillonnage.

A.3 Facteur de correction du biais

Pour une séquence infinie de périodogrammes $|Y(k,l)|^2$, la densité spectrale de puissance estimée à court terme $S(k,l)$ peut être écrite comme suit :

$$S(k,l) = (1-\alpha)\sum_{i=0}^{\infty}\alpha^i |Y(k,l-i)|^2 \tag{A.10}$$

La moyenne $\mathrm{E}\{S_{\min}(k,l)\}$ est proportionnelle à $\lambda_d(k,l)$ et la variance est proportionnelle à $\lambda_d^2(k,l)$. Afin d'avoir un biais nul, il est suffisant de calculer le facteur de correction du biais $B_{\min}(k,l)$ pour $\lambda_d(k,l) = 1$, qui est approximé par :

$$B_{\min}(k,l) \approx 1 + (D-1)\frac{2}{\tilde{Q}_{eq}(k,l)} \tag{A.11}$$

Où $\tilde{Q}_{eq}(k,l)$ est obtenue à partir de la variance normalisée inversée $Q_{eq}(k,l)$:

$$\tilde{Q}_{eq}(k,l) = \frac{Q_{eq}(k,l) - 2M(D)}{1-M(D)} \tag{A.12}$$

Avec : $Q_{eq}(k,l) = 2\lambda_d^2(k,l) \big/ \mathrm{var}\{S(k,l)\}$ et M(D) est une fonction de D de valeurs obtenues expérimentalement (Tableau A.1) :

D	M(D)	D	M(D)
1	0	30	0.762
2	0 .26	40	0.8
5	0.48	60	0.841
8	0.58	80	0.865
10	0.61	120	0.89
15	0.668	140	0.9
20	0.705	160	0.91

Tableau A.1 : Valeurs de M(D) pour différentes tailles de la fenêtre D.

A.4 Estimateur du bruit sans biais basé sur les statistiques minimales

D'après les sections précédentes, un estimateur sans biais de la densité spectrale de puissance du bruit est donné par :

$$\hat{\lambda}_d(k,l) = \frac{S_{\min}(k,l)}{E\{S_{\min}(k,l)\}_{|\lambda_d(k,l)=1}} = B_{\min}(D, Q_{eq}(k,l)) S_{\min}(k,l) \qquad \text{(A.13)}$$

L'estimateur sans biais exige la connaissance de la variance normalisée de l'estimation de la DSP lissée $S(k,l)$ pour chaque indice de temps et de fréquence.

Pour estimer la variance de $S(k,l)$, on utilise un lissage récursif du premier ordre pour l'approximation du moment d'ordre un : $E\{S(k,l)\}$, et du moment d'ordre deux : $E\{S^2(k,l)\}$ de $S(k,l)$, on a :

$$\overline{S}(k,l) = \beta(k,l)\overline{S}(k,l-1) + (1-\beta(k,l))S(k,l) \qquad \text{(A.14)}$$

$$\overline{S^2}(k,l) = \beta(k,l)\overline{S^2}(k,l-1) + (1-\beta(k,l))S^2(k,l) \qquad \text{(A.15)}$$

$$\hat{\mathrm{var}}\{S(k,l)\} = \overline{S^2}(k,l) - \overline{S}^2(k,l) \qquad \text{(A.16)}$$

De bons résultats ont été obtenus avec le choix du paramètre de lissage $\beta(k,l) = \alpha^2(k,l)$ et une limitation maximale à 0.8. Finalement $1/Q_{eq}(k,l)$ est estimé par :

$$1/Q_{eq}(k,l) \approx \frac{\hat{\mathrm{var}}\{S(k,l)\}}{2\hat{\lambda}_d^2(k,l-1)} \qquad \text{(A.17)}$$

Et cette estimation est limitée par une valeur maximale de 0.5 correspondant à $Q_{eq} = 2$. Le suivi d'une augmentation du niveau du bruit s'effectue seulement avec un petit retard, l'estimateur du minimum des statistiques à une tendance de sous-estimer les bruits non stationnaires élevés.

Donc, il est préférable d'augmenter le biais inverse $B_{\min}(k,l)$ par un facteur $B_c(l)$, proportionnel à l'écart type normalisé de l'estimation à court terme $S(k,l)$:

$$B_c(l) = 1 + a_v\sqrt{\overline{Q^{-1}}(l)} \qquad \text{(A.18)}$$

Avec a_v est une constante égale à 2.12 et $\overline{Q^{-1}}(l)$ la variance normalisée moyenne donnée par :

$$\overline{Q^{-1}}(l) = (1/L)\sum_{k=0}^{L-1} 1/Q_{eq}(k,l) \qquad \text{(A.19)}$$

Ce facteur de correction a une influence seulement quand l'estimation de la DSP à court terme a une large variance, et pour un bruit stationnaire ce facteur est proche de un. Donc, l'estimation finale du spectre du bruit est :

$$\hat{\lambda}_d(k,l) = B_{\min}(D, Q_{eq}(k,l)).B_c(l).S_{\min}(k,l) \qquad \text{(A.20)}$$

A.5 Implémentation efficace de la recherche du minimum

L'algorithme de cette méthode est basé sur la recherche du minimum chaque D valeurs consécutives de $S(k,l)$ calculées. Dans ce cas, une seule opération de comparaison est nécessaire pour chaque composante fréquentielle par trame, mais le retard du suivi d'une augmentation du niveau de bruit dans ce cas est 2D. Une implémentation efficace de l'algorithme a été proposée afin de réduire ce retard. La fenêtre de D échantillons est divisée en U sous fenêtres de V échantillons (VU=D). Chaque V échantillons, le minimum est mise à jour et stocké pour une utilisation ultérieure. Le minimum global est le minimum de tous les minimums des U sous fenêtres. Donc, on a : 1+ (U-1)/V opérations de comparaisons. Le retard dans ce cas est seulement D+V. Avec une fréquence d'échantillonnage de 8 kHz et une FFT de taille L=256, on utilise typiquement U = 8 et V = 12.

La figure (A.1) suivante présente le spectre de puissance du bruit réel et celui du bruit estimé, dans le cas d'un rapport signal sur bruit de 5 dB :

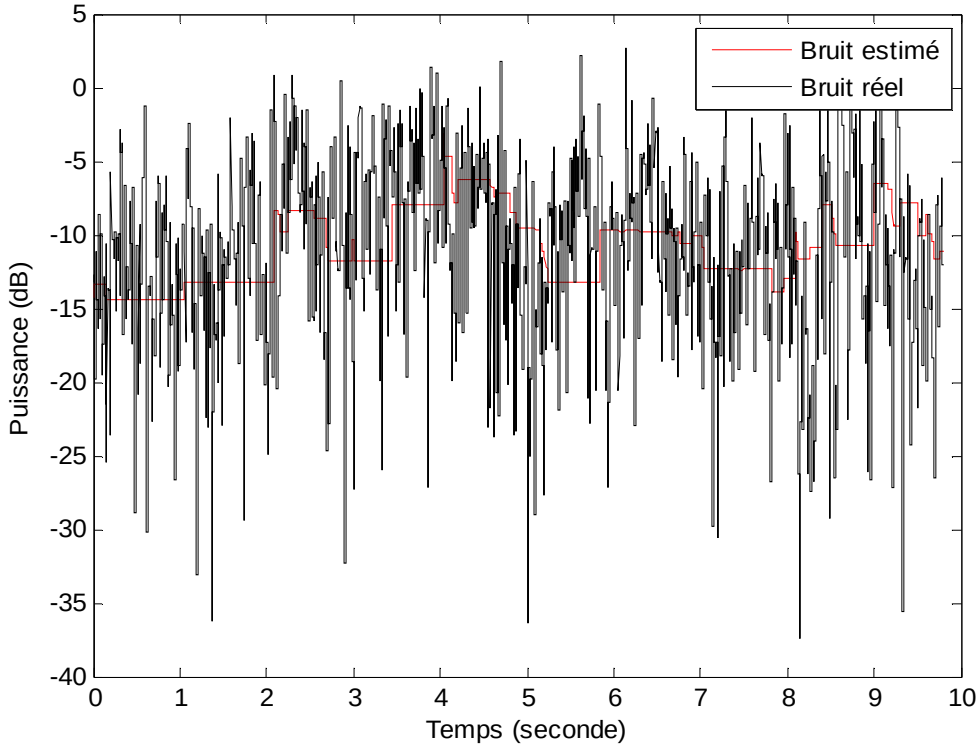


Figure A.1 : Spectre de puissance du bruit réel et celui du bruit estimé, cas d'un SNR = 5 dB à f = 500 Hz.

La figure A.2 résume l'algorithme (MS) et le tableau A.2 présente la liste des variables utilisées.

- Compute smoothing parameter $\hat{\alpha}(k,l)$, (A.6)
- Compute smoothed power $S(k,l)$, (A.3)
- Compute bias correction $B_{\min}(k,l)$ and $B_{\min\_sub}(k,l)$,(A.12), (A.11), (A.17)
- Compute $\overline{Q^{-1}}(l) = (1/L)\sum_{k=0}^{L-1} 1/Q_{eq}(k,l)$
- Set *k_mod*(*k*) =0 for all *k*
- If $S(k,l)B_{\min}(k,l)B_c(l) <$ *actmin* (*k,l*)
  - *actmin* (*k, l*) = $S(k,l)B_{\min}(k,l)B_c(l)$
  - *actmin_sub* (*k, l*) = $S(k,l)B_{\min\_sub}(k,l)B_c(l)$
  - set *k_mod* (*k*) =1;
- If *subwc* = =V
  - -if *k_mod* (*k*) = 1
    lmin_flag (*k, l*) =0
  - -store *actmin* (*k, l*)
  - -find $S_{\min\_u}$, the minimum of the last U stored values of *actmin*
  - -if $\overline{Q^{-1}}(l) < 0.03$, noise_slope_max =8;
  - -else if $\overline{Q^{-1}}(l) < 0.05$, noise_slope_max =4;
  - -else if $\overline{Q^{-1}}(l) < 0.06$, noise_slope_max =2;
  - -else *noise_slope_max* =1.2;
  - -if *lmin_flag* (*k, l*) & (*actmin_sub* (*k, l*) < *noise_slope_max* $S_{\min\_u}(k,l)$) & (*actmin_sub* (*k, l*) > $S_{\min\_u}(k,l)$)
    $S_{\min\_u}(k,l)$ = *actmin_sub* (*k, l*)
    replace all previously stored values of *actmin* (*k, l*) by *actmin_sub* (*k, l*)
  - -*lmin_flag* (*k, l*) = 0;
  - -set *subwc* = 1, and *actmin* (*k, l*) and *actmin_sub* (*k, l*) to their maximum values
- else
  - -if *subwc* > 1
    if *k_mod* = = 1
    set *lmin_flag* (*k, l*) = 1
    compute $\hat{\lambda}_d(k,l)$ = min (*actmin_sub* (*k, l*), $S_{\min\_u}(k,l)$)
    set $S_{\min\_u}(k,l) = \hat{\lambda}_d(k,l)$
  - -set *subwc*= *subwc*+1

Figure A.2 : Algorithme de la méthode MS [71][73].

Variable	Explication
l	Indice de trame
k	Composante fréquentielle
D	Nombre d'échantillons utilisés pour le calcul du minimum
$\|Y(k,l)\|^2$	Observation bruitée
$\alpha(k,l)$	Facteur de lissage
$S(k,l)$	Densité spectrale de puissance de la parole bruitée
$\hat{\lambda}_d(k,l)$	Variance du bruit
$B_{\min}(k,l)$	Facteur de correction du biais
$B_c(l)$	Facteur de graduation du biais
$Q_{eq}(k,l)$	Variance normalisée inversée
$\tilde{Q}_{eq}(k,l)$	Version mesurée de la variance normalisée inversée
$\overline{S}(k,l)$	Moment d'ordre un de $S(k,l)$
$\overline{S^2}(k,l)$	Moment d'ordre deux de $S(k,l)$
$\hat{\text{var}}\{S(k,l)\}$	Estimation de la variance de $S(k,l)$
$\overline{Q^{-1}}(l)$	Variance normalisée moyenne
$S_{\min}(k,l)$	Minimum pour le dernier échantillon lissé de D
sub	Quantité des sous fenêtres
subwc	Compteur des sous fenêtres
k_mod(*k*)	1 ou 0, drapeau pour indiquer la présence ou l'absence du bruit
actmin (*k, l*)	Le minimum actuel
lmin_flag (*k, l*)	1 ou 0, drapeau pour noter où le minimum de la sous fenêtre se trouve
noise_slope_max	Pente maximale du bruit

Tableau A.2 : Liste des variables utilisées dans la méthode MS.

ANNEXE B

Expressions des variances

Dans cette annexe, nous donnerons les expressions des variances de la variable aléatoire, qui suit la loi de probabilité Gamma généralisée (B.1). Pour les deux cas : $\gamma = 1$ et $\gamma = 2$.

$$f_A(a) = \frac{\gamma \beta^{\nu}}{\Gamma(\nu)} a^{\gamma \cdot \nu - 1} \exp\left(-\beta a^{\gamma}\right), \qquad a \geq 0, \beta > 0, \gamma > 0, \nu > 0, \tag{B.1}$$

B.1 Moment d'ordre deux de A dans le cas $\gamma = 1$

Avec [90, Eq. 3.381], il est simple de vérifier que $E\{A^2\}$ est donné par :

$$E\{A^2\} = \int_0^{\infty} a^2 \frac{\beta^{\nu}}{\Gamma(\nu)} a^{\nu-1} \exp(-\beta a)\, da \tag{B.2}$$

$$= \int_0^{\infty} \frac{\beta^{\nu}}{\Gamma(\nu)} a^{\nu+1} \exp(-\beta a)\, da \tag{B.3}$$

$$= \frac{\beta^{\nu}}{\Gamma(\nu)} \frac{\Gamma(\nu+2)}{\beta^{\nu+2}} \tag{B.4}$$

$$= \frac{\Gamma(\nu+2)}{\Gamma(\nu)\beta^2} \tag{B.5}$$

$$= \frac{\nu(\nu+1)}{\beta^2} \tag{B.6}$$

Donc :

$$\beta = \sqrt{\nu(\nu+1)/E\{A^2\}} = \sqrt{\nu(\nu+1)/\lambda_x(k)} \tag{B.7}$$

B.2 Moment d'ordre deux de A dans le cas $\gamma = 2$

Nous pouvons montrer avec [90, Eq. 3.381] et le changement de variable $a = \sqrt{x}$ que $E\{A^2\}$ est donné par :

$$E\{A^2\} = \int_0^{\infty} a^2 \frac{2\beta^{\nu}}{\Gamma(\nu)} a^{2\nu-1} \exp\left(-\beta a^2\right) da \tag{B.8}$$

$$= \int_0^{\infty} \frac{\beta^{\nu}}{\Gamma(\nu)} x^{\nu} \exp(-\beta x)\, dx \tag{B.9}$$

$$= \frac{\beta^{\nu}}{\Gamma(\nu)} \frac{\Gamma(\nu+1)}{\beta^{\nu+1}} \tag{B.10}$$

$$= \frac{\nu}{\beta} \tag{B.11}$$

Donc :

$$\beta = \nu / E\{A^2\} = \nu / \lambda_x(k) \tag{B.12}$$

ANNEXE C

Le codeur MELP

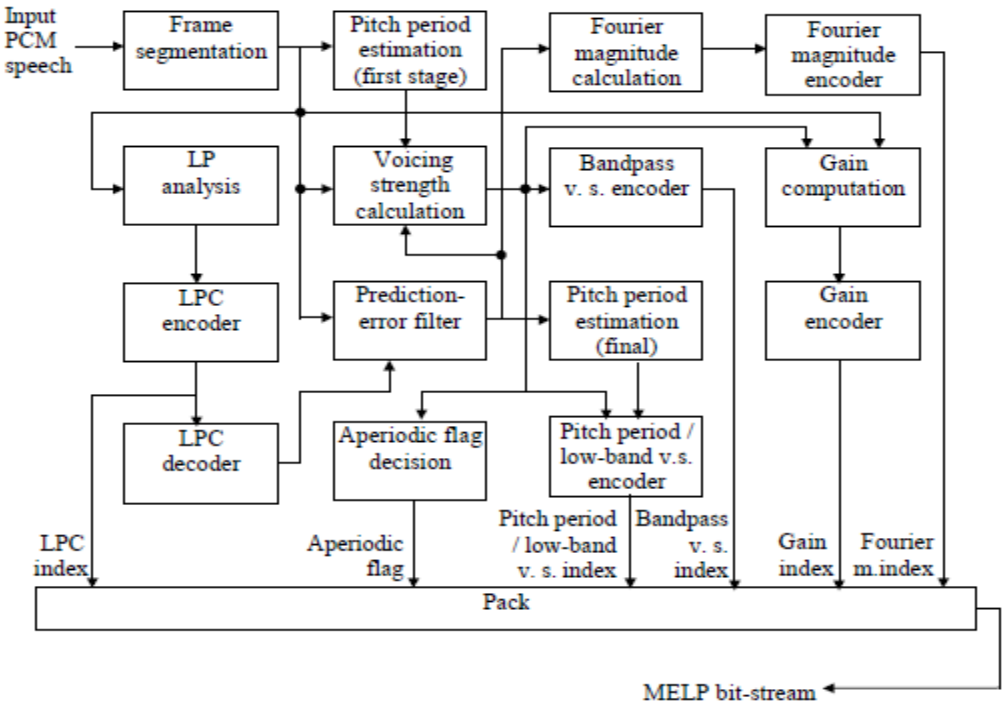


Figure C.1 : Schéma bloc du codeur MELP [3].

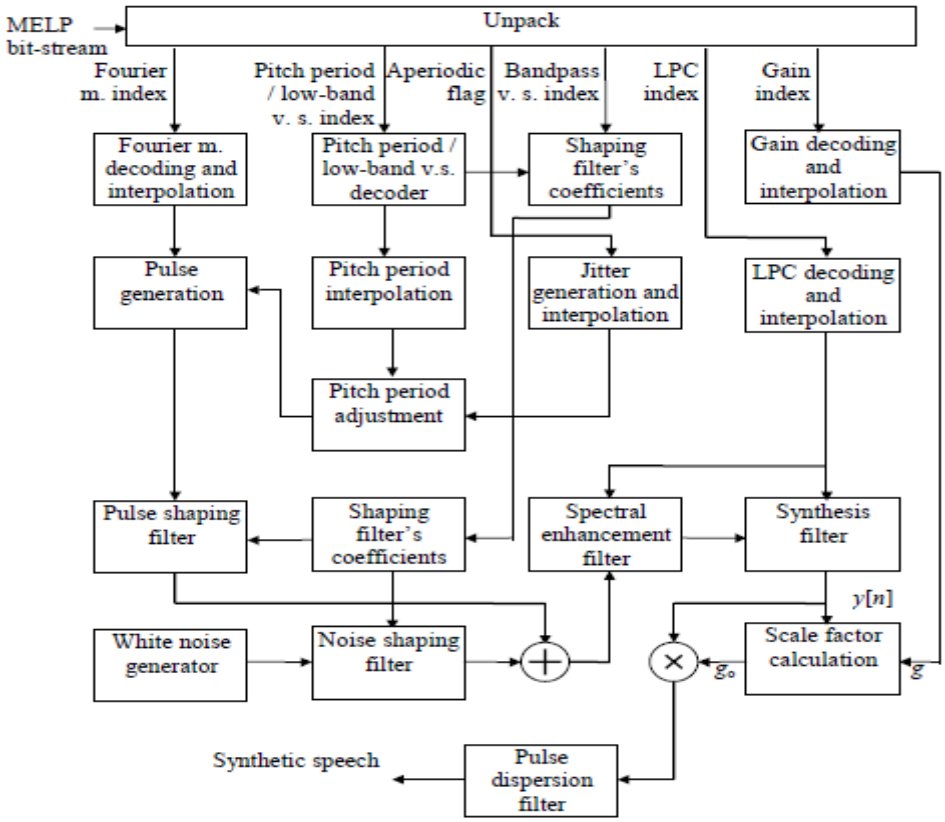


Figure C.2 : Schéma bloc du décodeur MELP [3].

ملخص

الاهتمام بمجال تحسين الكلام أزداد من الاستخدام الكبير في حياتنا اليومية لتطبيقات المعالجة الرقمية للكلام مثل الهاتف النقال، المعينات السمعية الرقمية وأنظمة الاتصالات بين الإنسان والآلات. نزعة تحويل هده التطبيقات في حالة المستخدم المتنقل من مكان إلى أخر زاد من تنوع المصادر المحتملة لتدهور جودة الكلام وضرورة اللجوء إلى طرق لتحسينه، والتي يمكن استخدامها لزيادة جودة ووضوح الكلام المعالج بهذه الأجهزة وجعلها أكثر ملائمة في ظل ظروف صاخبة، بحيث أصبح استخدام هده التقنيات قبل عملية الترميز شائعا في أنظمة الاتصالات اللاسلكية الحكومية والعسكرية على حد سواء. في هذه الأطروحة سوف نركز على طرق تخفيض الضجيج في حالة ميكروفون وحيد وبخاصة المعتمدة على تحويل فوريبه، ويكمن الهدف الرئيسي من البحث المقدم إلى تحسين طرق التخفيض لعدة أنواع من الضوضاء وفي مستويات مختلفة للضوضاء، لجعلها أكثر ملائمة لعمليات ترميز الكلام بتدفق رقمي منخفض في أماكن استخدام مملوءة بالضوضاء.

كلمات مفتاحيه :

تحسين الكلام، تخفيض الضوضاء، خوارزميات STSA، مرشح Kalman، تقدير نوع ومستوى الضوضاء، ترميز الكلام المتأثر بالضوضاء، الترميز دو التدفق المنخفض، الترميز MELP.

Abstract

The interest in the field of speech enhancement emerges from the increased usage of digital speech processing applications like mobile telephony, digital hearing aids and human-machine communication systems in our daily life. The trend to make these applications mobile increases the variety of potential sources for quality degradation. Speech enhancement methods can be used to increase the quality and intelligibility of these speech processing devices and make them more robust under noisy conditions. The use of speech enhancement techniques as a pre-processing stage to speech coders is being applied to governmental (military) wireless communication systems as well. In this thesis we will focus on single-microphone additive noise reduction and aim at methods that work in the discrete Fourier transform (DFT) domain. The main objective of the presented research is to improve on existing single-microphone schemes for an extended range of noise types and noise levels, thereby making these methods more suitable for low bit rate speech coder in noisy condition.

Keywords :

Speech enhancement, Noise reduction, Subtractive type algorithms, Kalman filter, Noise estimation, Noisy speech coding, Low bit-rate coders, MELP coder.

Résumé

L'intérêt du domaine de rehaussement de la parole à augmenter avec la large utilisation dans notre vie quotidienne des applications de traitement numérique de la parole, comme la téléphonie mobile, les prothèses auditives et les systèmes de communication homme-machine.

La tendance à rendre ces applications mobiles augmente la variété des sources potentielles de la dégradation de la qualité. Les méthodes de rehaussement de la parole peuvent être utilisées pour augmenter la qualité et l'intelligibilité de ces dispositifs de traitement de la parole et les rendre plus robustes dans des milieux bruités. L'utilisation des techniques de rehaussement de la parole en tant que blocs de pré-traitement aux codeurs de parole est devenue très utilisée dans les systèmes de communication sans fil civiles et militaires.

Dans cette thèse, on s'intéressera aux méthodes mono-voie de réduction de bruit, en particulier les méthodes STSA dans le domaine de la TFD. L'objectif principal du travail de recherche présenté est d'améliorer les systèmes mono-voie existants pour plusieurs types de bruits et à différents rapports signal sur bruit, rendant ainsi ces méthodes plus appropriées pour le codage de la parole à bas débit dans un milieu hostile.

Mots – clés :

Rehaussement de la parole, Réduction de bruit, Algorithmes STSA, Filtre de Kalman, Estimation du bruit, Codage de la parole bruitée, Codeurs à bas débit, Codeur MELP.