N° d'ordre : 09/ 2011-M / EL

République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université des Sciences et de la Technologie Houari Boumediène

Faculté d'Electronique et d'Informatique

جامعة هواري بومدين
للعلوم والتكنولوجيا
U S T H B

MEMOIRE

Présenté en vue de l'obtention du diplôme de MAGISTER

En : Electronique

Spécialité: Communication Parlée

Par : BOUCHAIR Asma

Thème

Amélioration de la Reconnaissance Vocale par Rehaussement de la Parole par le Filtrage de Kalman

Soutenu publiquement le 08 Décembre 2011, devant le Jury composé de :

Mr. A. HOUACINE	**Professeur à l'USTHB**	**Président**
Mr. A. AMROUCHE	**Maître de Conférences/A, à l'USTHB**	**Directeur de mémoire**
Mr. H. SAYOUD	**Professeur à l'USTHB**	**Examinateur**
Mme. L. FALEK	**Maître de Conférences/A, à l'USTHB**	**Examinatrice**

Remerciements

Mes remerciements les plus sincères vont à mon Directeur de mémoire, Mr A. AMROUCHE, Maitre de Conférences à la Faculté d'Electronique et d'Informatique de l'USTHB, pour sa disponibilité, sa simplicité, ses conseils avisés, son soutien et sa patience. Qu'il trouve ici mes respects et profonde gratitude.

Je remercie tout particulièrement Mr C.BOUBAKIR pour ses encouragements, ses conseils et son soutien technique.

Je tiens à exprimer mes remerciements à Mr A. HOUACINE, Professeur à la Faculté d'Electronique et d'Informatique de l'USTHB pour l'honneur qu'il me fait de présider le jury de soutenance.
Mes remerciement s'adresse également à Mr H. SAYOUD, Professeur, et Mme L.FALEK, Maitre de Conférences à la Faculté d'Electronique et d'Informatique de l'USTHB, membres examinateurs du jury, pour l'intérêt qu'ils ont porté à notre travail et pour leur disponibilité.

Grand merci à mes chers parents, qui m'ont inculqué la soif du savoir.
Je remercie ma famille qui m'a toujours soutenue, encouragée et qui a toujours été présente quand j'en avais besoin.
Je remercie également tous ceux qui m'ont aidé et encouragé de prés ou de loin.

Table des matières

Chapitre 2
Techniques de rehaussement de la parole

Chapitre 4
Amélioration de la reconnaissance automatique de la parole (RAP) par les techniques de rehaussement

Liste des figures

Liste des tableaux

Abréviations

ANN	Artificial Neural Network
AR	Auto-Régressif
BSD	Bark Spectral Distorion
CD	Cepstral Distance
CMOS	Comparaison Mean Opinion Score
DFT	Discret Fourier Transform
DMOS	Degradation Mean Opinion Score
DTW	Dynamique Time Warping
EM	Expectation Maximization
EQM	l'erreur quadratique moyenne
FFT	Fast Fourier Transform
HMM	Hidden Markov Model
HTK	Hidden Markov Model Toolkit
IDFT	Inverse Discret Fourier Transform
IFFT	Inverse Fast Fourier Transform
IS	Itakura Saito
LLR	Likelihood Linear Regression
LPC	Linear Predictf Coding
MBSD	Modified Bark Spectral Distortion
MFCC	Mel Frequency Cepsral Coefficient
MLP	Multi Layer Perceptron
MOS	Mean Opinion Score
PESQ	Perceptual Evaluation of Speech Quality
PLP	Perceptual Linear Predictive
PSQM	Perceptual Speech Quality Measure
RAP	Reconnaissance Automatique de la Parole
RASTA	RelAtive SpecTrAl technique
RBF NN	Radial Basis Neural Network
RNN	Recurrent Neural Networks

SNR	Signal to Noise Ratio
SNR_{seg}	segmental Signal to Noise Ratio
TDNN	Time Delay Neural Network
TR	Taux de Reconnaissance
VAD	Voice Activity Detector (détection d'activité vocale)
WSS	Weighted Spectral Slope

Introduction générale

Introduction générale

La parole est le support majeur d'expression de l'être humain, aussi les personnes peuvent partager des informations grâce à la voix. La communication verbale s'effectue soit dans des environnements le plus souvent acoustiquement pollués, ou bien à travers des réseaux de télécommunications non exempts de perturbations. Alors que les bruits environnementaux sont souvent de nature additive, les bruits générés dans les canaux de communication sont la plupart du temps de nature multiplicatifs, ou convolutifs. De plus, le bruit introduit par l'environnement peut être stationnaire ou non stationnaire et varier plus ou moins rapidement. Dans un milieu réel, on peut donc considérer que le signal observé résulte de la combinaison entre le signal de la parole propre et du bruit ambiant issu du milieu environnant, ce qui s'accompagne souvent de la dégradation de l'intelligibilité de la parole. L'amélioration de la qualité de la parole bruitée, ou rehaussement de la parole, est donc une nécessité primordiale dans le développement des systèmes multimédia utilisant la voix.

Le but du rehaussement de la parole est d'améliorer la qualité et l'intelligibilité de la parole, à partir du signal bruité, en utilisant les techniques de traitement du signal. On distingue deux grandes catégories des méthodes de rehaussement de la parole :

- Les méthodes paramétriques.
- Les méthodes non paramétriques.

Dans notre travail, nous avons étudié ces deux catégories et nous nous sommes focalisés sur les techniques de rehaussement de la parole basées sur la soustraction spectrale et le filtre de Wiener, considérées comme des méthodes non paramétriques, et le filtre de Kalman, assimilé à une technique paramétrique.

Les techniques de rehaussement de la parole sont exploitées dans plusieurs domaines comme : la reconnaissance automatique de la parole et du locuteur (RAP et RAL), les prothèses auditifs, la téléphonie mobile, la VoIP (Voice over Internet Protocol),etc. Dans notre cas, nous nous intéresserons à l'utilisation des techniques de rehaussement de la parole pour améliorer :

- L'intelligibilité de la parole bruitée, en perspective d'une application dans les réseaux de communications mobiles
- les performances des systèmes de la reconnaissance automatique de la parole (RAP) dans un milieu bruité.

A ce titre, la RAP est un domaine de recherche intéressant dont l'objectif est d'extraire l'information contenue dans un signal de la parole par un ordinateur pour décoder le message. Grâce à cette technologie, on peut communiquer oralement avec la machine, ce qui facilite considérablement l'interaction homme/machine. Les domaines d'application des systèmes de reconnaissance de la parole sont très variés : Commande et contrôle des machines à l'aide de la parole, dictée vocale, serveurs vocaux, etc. Les modèles les plus utilisés dans la plupart des applications en RAP sont : Les modèles de Markov cachés (HMM : Hidden Markov Model), les réseaux de neurones artificiels (ANN : Artificial Neural Network) et les modèles hybrides.

Toutefois, malgré les avancées très importantes dans le domaine de la RAP, les systèmes actuels sont encore en deçà des performances de notre système d'audition. Une des problématiques des systèmes de RAP actuels est la dégradation des performances des systèmes de reconnaissance dans les conditions réelles où l'environnement est bruité. Aujourd'hui, le défi majeur dans les recherches en reconnaissance automatique de la parole est de considérer toutes ces limitations et de les dépasser. Cela implique de prendre en compte les conditions de l'environnement. Aussi, un des objectifs de notre travail est d'étudier l'impact du rehaussement de la parole comme étape de prétraitement, sur les performances des systèmes de la reconnaissance vocale dans un milieu bruité.

Organisation du mémoire

Le premier chapitre de ce mémoire sera consacré à la présentation de l'état de l'art des techniques de rehaussement de la parole et ses applications. Ce chapitre abordera aussi la

caractérisation du signal de la parole et du bruit, ainsi que les différentes méthodes de codage du signal de la parole (LPC, MFCC, PLP, etc.)

Dans le deuxième chapitre, nous présenterons trois méthodes de rehaussement de la parole : la soustraction spectrale, le filtre de Wiener et le filtre de Kalman. Ces méthodes seront détaillées et appliquées sur la parole bruitée au moyen de la base de données bruitées NOISEUS.

Le troisième chapitre présentera tout d'abord les différents critères d'évaluation de la qualité et de l'intelligibilité de la parole rehaussée, suivi par les résultats de test des méthodes de rehaussement appliquées dans ce travail et décrites dans le second chapitre.

Dans le dernier chapitre, le principe et les différentes techniques de reconnaissance de la parole seront présentées, en particulier celles utilisant les modèles de Markov Cachés (HMM), les réseaux de neurones de type MLP (Multilayer Layer Perceptron) et les réseaux récurrents (RNN : Recurrent Neural Network). Une évaluation des résultats expérimentaux, suivie d'une discussion, a été effectuée afin de situer l'apport du rehaussement de la parole sur les performances de la RAP dans un milieu ambiant réel.

Enfin, nous conclurons ce mémoire par une conclusion générale et une présentation des perspectives pour des travaux futurs en vue d'améliorer la reconnaissance vocale en environnement réel.

Chapitre 1

Introduction au rehaussement de la parole

Chapitre 1 :

Introduction au rehaussement de la parole

1.1 Introduction

La parole est le vecteur de communication de l'être humain, le signal de parole est élaboré de façon à ce que le sens qu'il porte y soit robuste. Le bruit est le phénomène perturbateur qui dégrade les performances de la parole, il est très gênant car il peut masquer les caractéristiques spécifiques de la parole.

Le rehaussement de la parole consiste à améliorer la qualité et l'intelligibilité de la parole dégradée par un bruit. C'est un domaine de recherche très actif. Dans ce chapitre nous étudierons le principe de base du rehaussement de la parole et les différentes caractéristiques de la parole et du bruit.

1.2 Rehaussement de la parole

1.2.1 Principe et objectif de rehaussement de la parole

Le rehaussement, ou le débruitage de la parole, signifie l'amélioration de la qualité et/ou l'intelligibilité de signal de la parole dégradée.Le but principal du rehaussement de la parole,émise dans les environnements bruités, est d'améliorer la qualité et l'intelligibilité du signal de la parole restituée. Les domaines d'application des techniques de rehaussement de la parole sont vastes et englobent : la téléphonie mobile, la VoIP, la reconnaissance de la parole et du locuteur, les prothèses auditifs, etc.

Dans la plupart des cas on suppose que la parole est dégradée par un bruit additif, Dans le cas de bruits sonores issus de sources acoustiques présentes dans le milieu environnant (bruits ambiants), on suppose que la parole est dégradée par un bruit de type

additif. Pendant le rehaussement d'un signal de la parole dégradé on rencontre deux types de problèmes :

1. Premièrement, la nature et les caractéristiques du bruit peuvent changer dans le temps et/ou l'espace, d'où la difficulté d'élaborer des algorithmes de rehaussement qui opèrent dans des environnements différents.
2. Deuxièmement, les mesures de performance peuvent être définies différemment pour chaque application.

1.2.2 Classification des systèmes de rehaussement de la parole

Les systèmes de rehaussement de la parole peuvent être classés selon :

- Le nombre de canaux d'entrée (mono-voie / multi-voies).
- Le domaine du traitement (temps / fréquence).
- Le type d'algorithme (adaptatif / non adaptatif).

Les systèmes monovoie utilisent un seul microphone d'acquisition, donc ils sont moins chers que les systèmes multi-voies.Outre leurs simplicité d'implémentation, l'absence d'une référence pour le bruit est l'inconvénient major des systèmes mono-voies.

Les systèmes multi-voies sont des systèmes complexes qui utilisent plusieurs microphones d'acquisition, ils ont l'avantage de la disponibilité d'une référence pour le bruit.

1.2.3 Etat des recherches sur le rehaussement de la parole

Les recherches sur les techniques de rehaussement de la parole ont commencé il y a plus de 40 ans par les chercheurs du Bell Labs, une implémentation analogique de la méthode de la soustraction spectrale d'amplitude a été proposée. Plus de 15 ans après, ces méthodes sont réinventées dans le domaine numérique.

On distingue deux grandes catégories des méthodes de rehaussent de la parole mono-voie :

- Les méthodes non paramétriques.
- Les méthodes paramétriques.

➢ Pour la première catégorie, les méthodes de modifications spectrales à court terme constituent une famille d'algorithmes de référence**[1],[2],[3].**Une

importante approche basée sur la décomposition en sous-espaces signal et bruit est proposée par Ephraim et Van Trees en 1995 **[4].**

➢ Par opposition aux méthodes non paramétriques, les méthodes dites paramétriques, comme leur nom l'indique, se basent sur une paramétrisation ou un codage du signal de la parole.

Afin de ne conserver que le signal utile, les paramètres modélisant le signal de la parole sont estimés d'une manière robuste au bruit.

- Dans la première famille, le signal de la parole est modélisé par un modèle sinusoïdal **[5]** ou exponentiel **[4],[6],[7].**
- Dans la deuxième famille, le signal de la parole est modélisé par un modèle autorégressive (AR). Les algorithmes de rehaussement de la parole appartenant à cette catégorie fonctionnent en deux étapes :
    1. Estimation des cœfficients AR et des variances du processus générateur et du bruit.
    2. Puis, un filtrage est mis en œuvre pour rehausser le signal de la parole en utilisant les paramètres estimés, (filtre de Wiener ou filtre de Kalman).

Dans ce qui suit, on va présenter quelques méthodes proposées ces vingt dernières années**[8]**.En 1987, Paliwal et al**[9]** ont proposé de rehausser la parole par un filtre de Kalman en estimant les paramètres du modèle et la variance du bruit à partir du signal de la parole propre et du signal additif. Dans le cas réel cette méthode n'est pas applicable.

En 1991, Gibson et al**[10]** ont proposé d'estimer les paramètres AR et la variance du processus générateur à partir du signal bruité. Ensuite, on itère la procédure du filtrage sur le signal bruité et on estime les paramètres AR à partir du signal de la parole rehaussé.La VADest utilisée pour l'estimation de la variance du bruit durant les trames de silence.

En 1998, un algorithme itératif de type EM fondé sur un lissage de Kalman est mis en œuvre pour estimer à la fois les paramètres du modèle et le signal de la parole, cet algorithme est proposé par Gannot et al**[11]**. En 1999, Gabrea et al**[12]** ont proposé de calculer le gain de Kalman itérativement et sans connaissance explicite des variances des bruits.En 2002, Grivel et al **[13]** ont proposé d'estimer directement les matrices de la représentation dans l'espace d'état en utilisant des techniques d'identification en sous espace.Cette approche ne nécessite

pas l'utilisation d'un VAD.En 2004, Ma et al **[14]** ont proposé un filtre de Kalman perceptuel pour améliorer la qualité auditive du signal rehaussé.

Il est évident que ces recherches ne sauraient être conduites sans une connaissance approfondies du phénomène de la parole et des caractéristiques acoustiques du signal de parole.

1.3 Caractéristiques du signal de la parole

1.3.1 Le signal de la parole

1.3.1.1 Production et perception de la parole

La parole est un signal sonore engendré par un ensemble d'organes formant l'appareil phonatoire (Figure 1.1), c'est le moyen de communication privilégie des humains. Le processus de production de la parole est un mécanisme très complexe qui repose sur une interaction entre le système neurologique et physiologique.Les trois organes qui composent l'appareil phonatoire humain qui entrent en jeu dans la production de la parole sont :

1. **Les poumons :** fournit l'ai phonatoire pour la production de la parole.
2. **Le larynx :** l'air des poumons arrive au niveau des cordes vocales qui jouent le rôle de valve vis-à-vis de cet air.
3. **Le conduit vocal :** s'étend des cordes vocales jusqu'aux lèvres dans la cavité orale et jusqu'aux narines dans la cavité nasale.
    - **Cavité orale :** à géométrie variable en fonction d'articulateur (langue, mâchoire inférieure et lèvres).
    - **Cavité nasale :** à géométrie fixe, elle peut être couplée à la cavité orale par abaissement du voile du palais.

L'appareil phonatoire est l'organe responsable de la production de la parole. En effet, l'étude de ce système, notamment ses configurations volumiques variées, va nous permettre d'identifier les grandes classes de sons. Le mécanisme articulatoire influent directement sur de production différentiée des sons. D'ailleurs, les phonéticiens se basent essentiellement sur le lieu et le mode d'articulation pour caractériser les sons du langage.

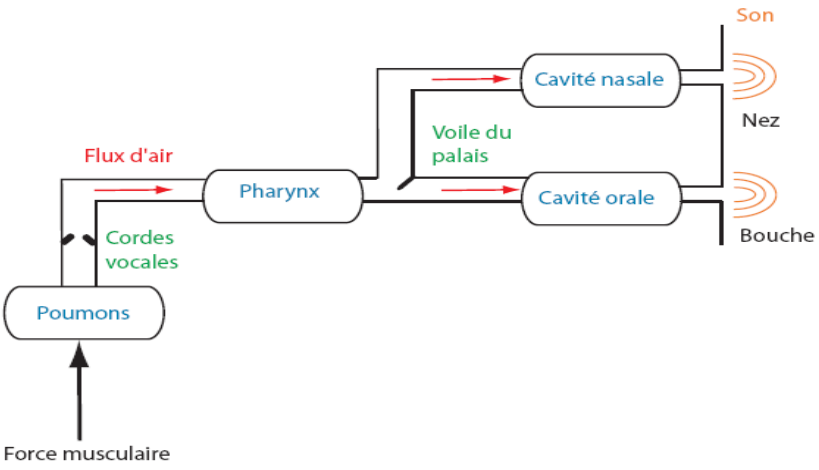

Figure 1.1 : Schéma synoptique de l'appareil phonatoire humain.

L'utilité du signal de la parole produit par l'appareil phonatoire est liée à l'existence d'un récepteur qui peut capter et analyser ce signal. L'appareil auditif humain (Figure 1.2)est divisé en deux parties : le système auditif périphérique et le système auditif central.

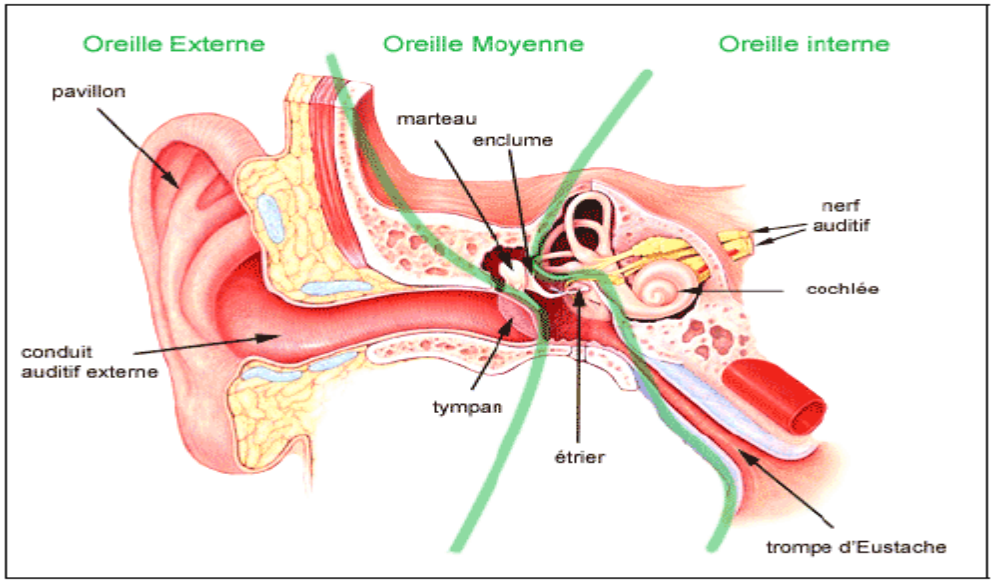

Figure 1.2 : Système auditif périphérique humain[15].

1. **Le système auditif périphérique** comporte trois parties : l'oreille externe, moyenne et interne.
    - **L'oreille externe** (pavillon, conduit auditif) **:** Elle transmet les vibrations acoustiques aériennes reçues à l'oreille moyenne.

- **L'oreille moyenne** (le marteau, l'enclume et l'étrier) **:** Elle assure la transmission de la vibration sonore d'un milieu aérien à un milieu liquidien.
- **L'oreille interne :** Elle transforme le signal acoustique en message nerveux.

2. **Le système auditif central** est composé de la partie nerveuse, au niveau de cette partie l'information auditive périphérique sera traitée.

1.3.1.2 Classification des sons de la parole

Les sons du langage sont classés d'après les critères suivants :

- L'opposition «sonore - sourde », selon que le son est voisé ou non voisé.
- Le mode d'articulation qui correspond à :
  - une constriction : pour la production des fricatives.
  - une occlusion : pour la production des occlusives.
- Le lieu (point) d'articulation : c'est l'endroit de la constriction maximale au niveau du conduit vocal.
- L'opposition « orale - nasale » : selon le chemin emprunté par l'air phonatoire, soit par les cavités buccale ou nasale.

Les sons du langage sont classés en consonnes et voyelles, bien que, rappelons le encore une fois, les phonéticiens arabes privilégient le système consonantique et estiment que les voyelles n'ont d'existence que pour permettre à la consonne de se produire. De ce fait, les réalisations vocaliques font appel à la notion de "haraka" que l'on traduit par consonne en mouvement vocalique, incluant donc un segment vocalique. Les phonéticiens arabes ont introduit la notion de "sukun" que l'on peut assimiler à un segment purement consonantique où la composante vocalique est absente.

Les consonnes

Ce sont des sons résultant d'une fermeture partielle (constriction) ou totale (occlusion) du conduit vocal lors du passage de l'air phonatoire. Elles peuvent être voisées ou non voisées, nasales ou orales. Les consonnes sont classées selon les trois principaux types suivants :

Occlusives (plosives)

Les occlusives sont caractérisées par un silence provenant de la fermeture complète du conduit vocal (occlusion) en un point précis : le point d'occlusion. L'écoulement de l'air généré par les poumons se trouve donc interrompu en un point particulier dans le conduit vocal, on parle de la « tenue » de l'occlusion. Cette tenue provoque une augmentation de

pression suivie d'un brusque relâchement, on parle alors « d'explosion ». La fin de l'occlusion provoque une perturbation acoustique, sous la forme d'une onde de pression due au relâchement de l'air qui était comprimé par l'occlusion. Cette perturbation (burst) est de courte durée (5 à 35 ms) mais peut être intense (sauf dans le cas des occlusives sonores). Elle est suivie de transitions formantiques vers le son vocalique suivant. Mise à part la zone de la tenue, les consonnes plosives peuvent donc être considérées comme des sons transitoires résultant de l'ouverture brusque du conduit vocal après son obstruction.

La durée de la tenue de la plosive (silence), influencée par l'entourage phonétique et par le débit de parole, est comprise entre 50 et 120 ms. Mais ce silence peut ne pas être total (cas des occlusives sonores), car il peut se former une " barre de voisement ", produite par une vibration des cordes vocales. Cette barre, de faible énergie, est concentrée dans les basses fréquences (100 à 300 Hz). Lorsque les cordes vocales vibrent lors du passage de l'air, on dit que l'occlusive est voisée, dans le cas contraire l'occlusive est sourde ou non voisée.

Exemple : | p | et | k | sont des occlusives non- voisées.

| b | et | d | sont des occlusives voisées.

Fricatives (constrictives)

Les fricatives (ou constrictives), sont des bruits produits par l'écoulement turbulent de l'air. Lorsque cet écoulement rencontre un rétrécissement, un lieu de constriction, il se produit un bruit de friction.

- Entre 4 et 8 kHz pour les consonnes | **s** | ou | ***z***|;
- Au fond du conduit vocal comme pour les palatales |ʃ|.

Nasales

Elles résultent de l'obstruction du conduit vocal et de l'ouverture de vélum qui permet l'échappement de l'air par les cavités nasales. On distingue deux consonnes nasales, toutes les deux voisées :

- | **m**| : dont le lieu d'articulation est labial.
- | **n**| : dont le lieu d'articulation est dental.

Sonnantes

Elles se caractérisent par une structure de formants et elles ne possèdent que peu ou pas de bruit. Plusieurs sous-classes existent : les vibrantes tel que le | ***r***|,les liquides tel que le | ***l***|.

Vibrantes

Il s'avère qu'il en existe une seule : le | ***r***| qui est produit par une vibration de la langue et qui est caractérisée par une structure de formants interrompus par des intervalles de silences très court, résultat du battement de la langue.

Liquides

Il en existe une seule **| *l*|**, produite par une obstruction partielle du conduit buccal et un écoulement latéral. Au plan spectral, elle est caractérisée par une structure de formant similaire à celle des voyelles.

Les voyelles

Elles sont caractérisées par le passage libre de l'air dans le conduit vocal, donc le conduit vocal présente dans ce cas une configuration quasi-stable, la source d'excitation du conduit vocal est la vibration laryngienne. La fréquence fondamentale de cette vibration est appelée fréquence fondamentale F0 ou pitch.

Elles se différentient par leur lieu d'articulation, leur aperture ou degré d'ouverture. Dans certaines langues, telle que le français, la nasalisation est également une caractéristique distinctive importante, car quatre voyelles de cette sont nasalisées.

Du point de vue acoustique, elles sont classées dans le plan des deux premiers formants F1-F2, dans une forme géométrique proche du triangle appelé triangle vocalique.

Les semi-voyelles

Elles sont voisées et se caractérisent par une affinité avec les voyelles qui se traduit par des structures de formants spécifiques telle que le |*w*| et le **| *j*|.**

1.3.1.3 Caractéristiques du signal de la parole

La structure du signal de la parole est très complexe. Ce signal est caractérisé par trois paramètres appelés aussi traits acoustiques :

1. **La fréquence fondamentale (pitch)**

C'est la fréquence de vibration des cordes vocales pour les sons voisés. La fréquence fondamentale peut varier, selon le genre et l'age du locuteur, de 100Hz à 500 Hz.

2. **L'énergie**

L'énergie correspond à la puissance du signal et elle est calculée sur plusieurs trames du signal. Par rapport aux segments voisés les segments non voisés ont une énergie plus forte.

3. Le spectre

Le spectre est obtenu par une analyse de Fourier à court terme. Une autre technique calcule le spectre à partir des coefficients LPC. Une représentation pseudo tridimentionnelle ou spectrogramme du signal de parole est donnée par la Figure 1.3.

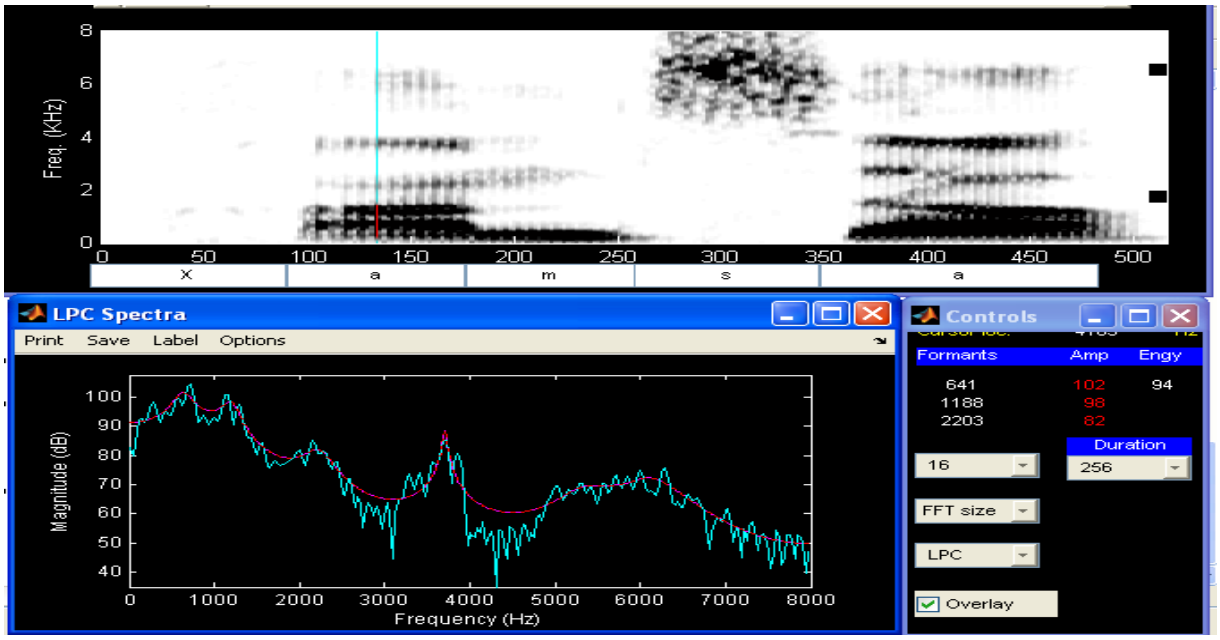

Figure 1.3 : Spectrogramme du mot arabe /Xamsa/ (chiffre 5), le spectre par LPC (en trait rouge) et le spectre par FFT (trait vert) calculés à partie d'une trame de 256 échantillons pris dans la partie stable de la première voyelle /a/ sont superposés. Le spectre LPC, qui a une forme lissée, permet de mettre en évidence les formants (pic de fréquences) caractéristiques des voyelles [d'après Amrouche 2007].

1.3 Le bruit

1.3.1 Origines et caractéristiques du bruit

On désigne par bruit, tout ensemble de sons nuisibles qui se superpose au signal utile.Il est perçu comme une sensation désagréable ou gênante. On distingue deux sources de bruits, dans le cas où les bruits sont générés à l'intérieur du système on a une **source interne,** et une **source externe** si les bruits sont générés à l'extérieur du système.Le bruit est caractérisé par les paramètres suivants :

1. **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.

2. L'intensité

Elle est définie en dB SPL (Sound Pressure Level) à partir d'un niveau 0 dB correspondant au seuil d'audibilité. Le seuil d'audibilité est pris à un niveau de pression sonore de 2.10^{-5} Pa/m^2.

Pour l'homme, un son d'une intensité supérieure à 65 dB SPL est perçu comme un son désagréable. A partir de 85 dB environ, nous avons un seuil de gêne. Il existetoute une graduation de 80 à 120 décibels, voire au-delà, qui conduit à des lésions irrémédiables dans le mécanisme de l'oreille.

3. 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.

1.3.2 Types de bruit

La classification des bruits est liée aux propriétés caractérisant le bruit. Le tableau (1.1) représente les différentes classes du bruit :

Tableau 1.1 : Les différentes classes du bruit [16].

Propriété	Types
Structure	Continu / Impulsif/ Périodique
Type d'interaction	Additif / Multiplicatif / Convolutif
Comportement temporel	Stationnaire /Non-stationnaire
Bande de fréquence	Etroite / Large
Dépendance	Corrélé / Décorrélé
Propriétés statistiques	Dépendant / Indépendant
Propriétés spatiales	Cohérent / Incohérent

Les spectrogrammes des Figures 1.4 et 1.5 montrent les différences significatives dans les composantes spectrales de deux bruits parmi les plus présents dans les milieux environnants.

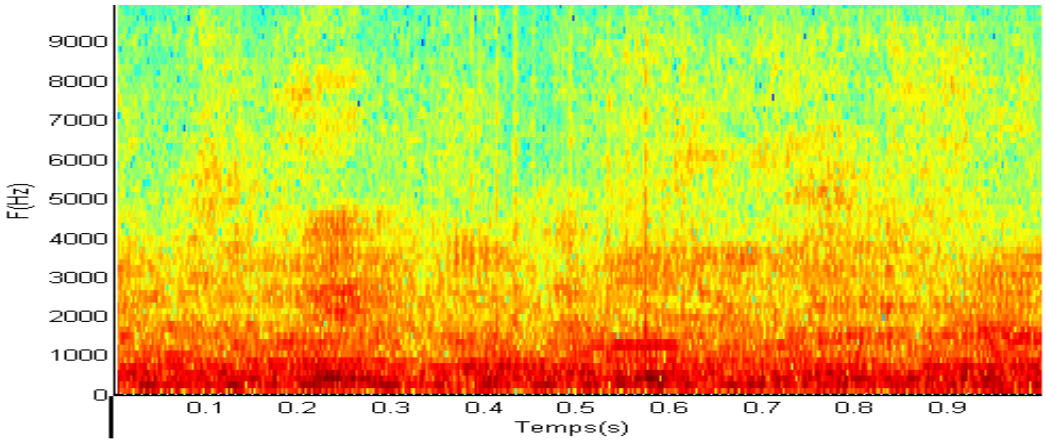

Figure 1.4: Spectrogramme du bruit de chahut dans une cantine.

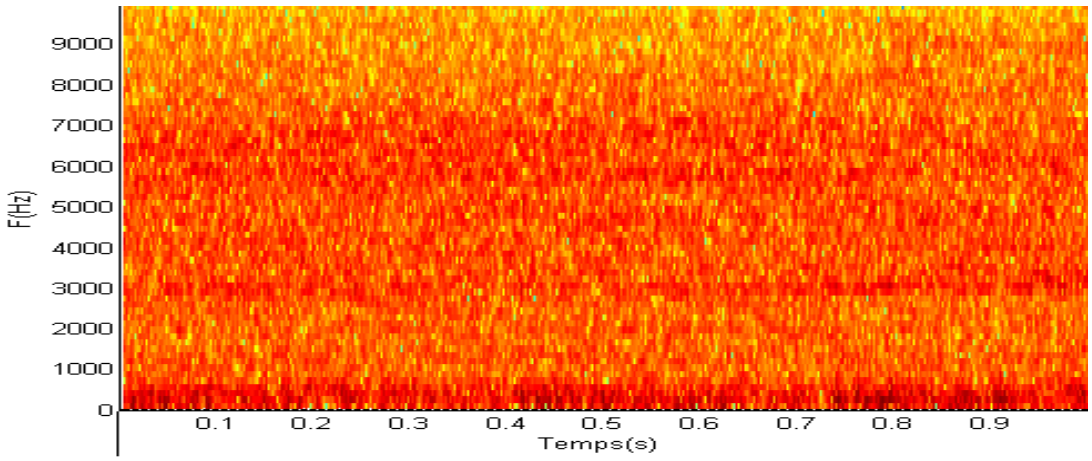

Figure 1.5: Spectrogramme du bruit d'avion de combat buccaneer.

1.4 Analyse et traitement du signal de la parole

Le signal de la parole est un signal non stationnaire et très redondant, pour cela il faut l'analyser avant qu'il soit utilisé dans n'importe quel système.L'analyse de la parole consiste à la mise en forme de ce signal et l'extraction des paramètres, elle se fait sur des fenêtres temporelles de l'ordre de 20 à 30 ms. On peut considérer que le signal de la parole est stationnaire sur ces courtes durées.Cependant, l'analyse (Figure 1.6) de la parolerequiert deux étapes : Le prétraitement et la paramétrisation du signal de la parole.

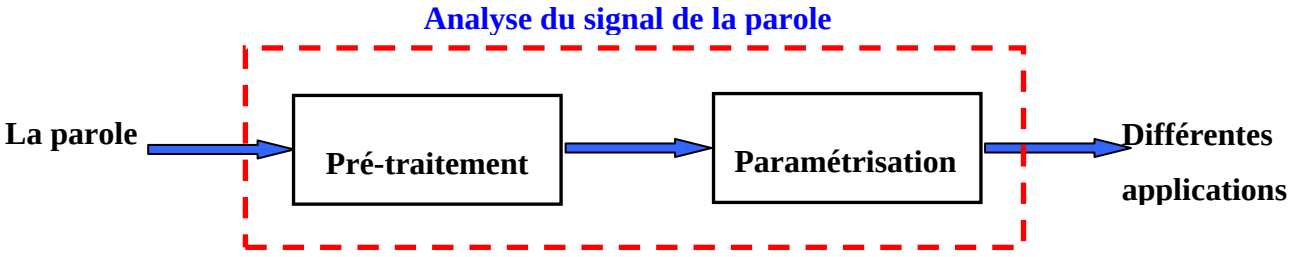

Figure 1.6 : Les étapes d'analyse du signal de la parole.

1.4.1 Le prétraitement

Le prétraitement (Figure 1.7) du signal de la parole permet de compenser les déformations de la mise en forme du signal.

- **L'échantillonnage**

L'échantillonnage transforme le signal à temps continu x (t) en un signal à temps discret x (n). Pour le signal de la parole, il faut choisir une fréquence d'échantillonnage satisfaisant le théorème de Shannon :

$$f_e \geq 2.f_{\max} \tag{1.1}$$

Avec :

f_e : La fréquence d'échantillonnage.

$f_{\max}$: La fréquence maximale du signal a traité.

- **La Pré-accentuation**

La pré-accentuation permet d'enlever les composantes continues du signal et d'amplifier les hautes fréquences. Elle consiste à passer le signal de la parole dans un filtre passe-haut, dit filtre de Pré-accentuation.

Ce filtre est du premier ordre et a pour fonction de transfert :

$$H(z) = 1 - a.z^{-1} \tag{1.2}$$

α : étant un coefficient de pondération compris entre 0.95 et 0.98**.**

- **Segmentation en trame**

Le changement des propriétés statistiques du signal de la parole est lié au fait que le signal de la parole n'est pas stationnaire. Cependant, ce signal peut être considéré comme quasi-stationnaire sur des courts intervalles.Pour palier ce problème, on divise le signal de la parole en segments successifs stationnaires de courtes durées (10 à 30ms).

- **Fenêtrage**

Le but du fenêtrage est d'éviter la distorsion du signal de la parole au début et à la fin de la trame.Le signal final est obtenu par la multiplication du signal segmenté x (n) par une fenêtre de pondération w (n) :

$$s(n) = \sum_{k=1}^{N} x(n).w(k) \qquad (1.3)$$

Avec :

s (n) : La trame courante.

x (n) : Signal segmenté.

$w(k)$: Fenêtre d'analyse, $k = 1,\ldots, N$.

Un bon choix de la fenêtre d'analyse est très important. Généralement, on utilise la fenêtre de Hamming, cette fenêtre est avantageuse par le fait quelle entraîne un minimum de distorsion spectrale du signal de parole. Elle est donnée par :

$$w(k) = 0.54 - 0.46 \cdot \cos\left(\frac{2 \cdot \Pi \cdot k}{N - 1}\right) \qquad (1.4)$$

Avec : $1 \leq k \leq N$

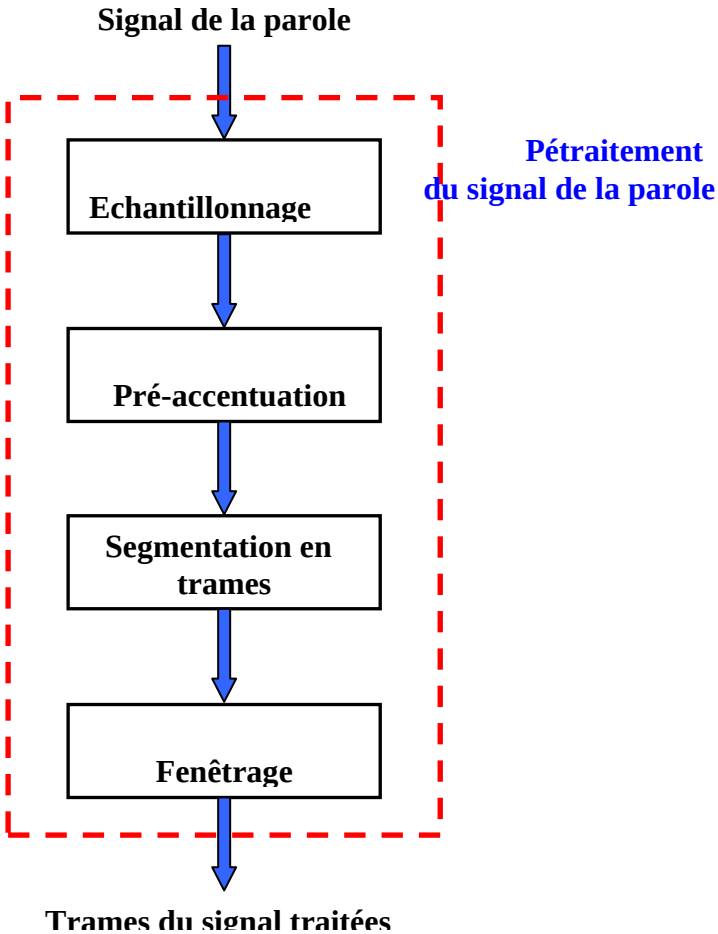

Figure 1.7 : Les étapes du prétraitement du signal de la parole.

1.4.2 Paramétrisation du signal de la parole

Le signal de la parole numérique n'est pas exploitable directement car beaucoup de données sont inutiles ou redondante.L'objectif de la paramétrisation du signal de la parole est de réduire la redondance et de supprimer les informations inutiles en donnant une représentation du signal de la parole adaptée à l'application.

Différentes paramètres peuvent être utilisés pour la paramétrisation du signal de tels que :

- L'énergie.
- Le taux de passage par zéros.
- Le voisement.
- La fréquence fondamentale (pitch).
- Le vecteur code : LPC, MFCC, PLP,…etc.
- Les dérivées premières et secondes du vecteur code.

1.5 Extraction des paramètres du signal de la parole

1.5.1 Le codage par prédiction linéaire LPC

Le codage par prédiction linéaire LPC : Linear Predictive Coding) repose sur la connaissance du modèle de production de la parole, le conduit vocal est modélisé par un filtre tous –pôle qui à une fonction de transfert d'un modèle autorégressif (AR).

On ne peut pas utiliser la même modélisation de la source d'excitation pour les sons voisés ou non voisés. Pour cela, le signal d'excitation est une suite d'impulsion d'amplitude unité pour les sons voisés, et un bruit blanc de moyenne nulle et de variance unité pour les sons non voisés.La fonction de transfert du conduit vocal est donnée par :

$$H(z) = \frac{1}{A(z)} \tag{1.5}$$

Avec : $A(z)$: Le prédicteur linéaire donné par :

$$A(z) = \sum_{i=1}^{p} a_i z^{-i} \tag{1.6}$$

Avec :

p : L'ordre du modèle.

a_i : Les coefficients LPC.

Le signal de la parole prédit est donné par :

$$\hat{s}(n) = \sum_{i=1}^{p} a_i s(n-i) \tag{1.7}$$

D'après cette relation, on remarque que chaque échantillon du signal de la parole est une combinaison linéaire des p échantillons qui le précédent.Pour calculer les coefficients de prédiction a_i on minimise la variance de l'erreur de prédiction $E\left[e^2(n)\right]$.

Le codage LPC est facile à mettre en œuvre et très utilisé en traitement de la parole. Outre sa simplicité, il permet d'éviter la redondance et de représenter le signal de la parole par des paramètres pertinents.

1.5.2 Le cepstre

La représentation cepstrale est basée sur une connaissance du modèle de production de la parole.Pour le calcul du cepstre, le signal vocal est défini comme le résultat de la convolution de la fonction de transfert du conduit vocal par un signal d'excitation :

$$s(n) = g(n) * b(n) \tag{1.8}$$

Avec :

$s(n)$: Signal vocal.

$g(n)$: Signal d'excitation.

$b(n)$: Fonction de transfert du conduit vocal.

Le passage au domaine spectral est fait via l'application de la transformée de Fourier sur l'équation (1.8), donc on obtient :

$$S(f) = G(f) \cdot B(f) \tag{1.9}$$

$S(f)$: Transformée de Fourier du signal $s(n)$.

$G(f)$: Transformée de Fourier du signal $g(n)$.

$B(f)$: Transformée de Fourier du signal $b(n)$.

Le logarithme de l'amplitude de $S(f)$ transforme le produit $G(f) \cdot B(f)$ en une somme :

$$\log|S(f)| = \log|G(f)| + \log|B(f)| \tag{1.10}$$

Parmi les avantages envisageables dans la représentation cepstrale le fait que les bruits convolutifs deviennent additifs grâce à la fonction logarithmique, donc le logarithme permet

d'éliminer les effets convolutifs dans le domaine temporel.La Figure 1.8montre les étapes de calcul du cepstre :

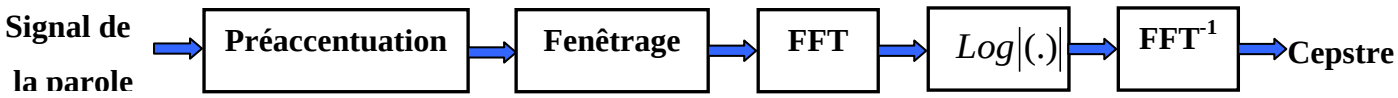

Figure 1.8 : Etapes de calcul du cepstre.

1.5.3 Le codage MFCC

Les cœfficients cepstraux à échelle Mel, dits MFCC (Mel Frequency Cepsral Coefficient), sont les paramètres les plus utilisés en reconnaissance de la parole.

L'extraction des paramètres MFCC est basée sur les variations des bandes critiques de fréquence de l'oreille humaine. En effet, le calcul de ces paramètres est basé sur l'échelle Mel ce qui permet de capturer les caractéristiques phonétiques que l'oreille perçoit.

Les étapes de calcul des cœfficients MFCC sont décrites dans le schéma suivant donné par la Figure 1.9. L'échelle de transformation Hertz-Mel, est sensiblement linéaire entre 0 et 1kHz. Au delà de cette fréquence, elle suit une loi de conversion qui peut être exprimée par la relation :

$$Mel = 2595\log_{10}(1+\frac{f_{Hertz}}{700}) \qquad (1.11)$$

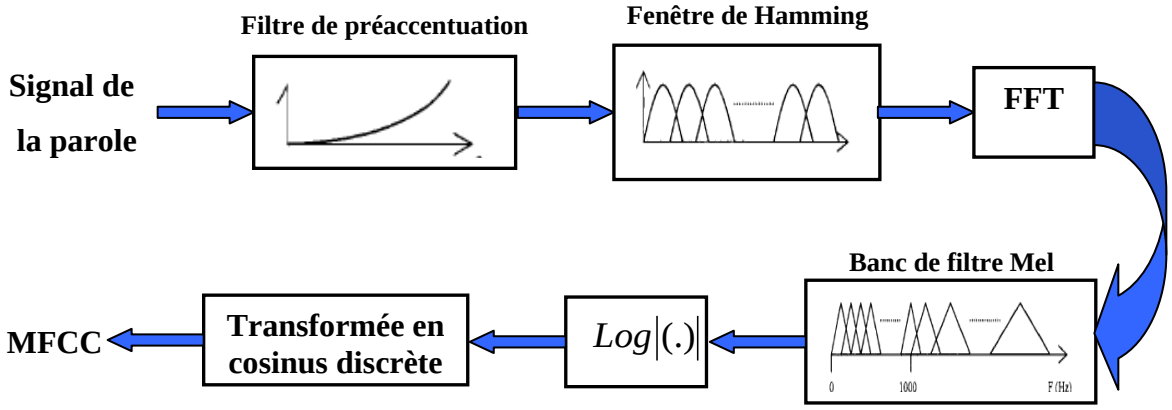

Figure 1.9 : Processus général de production des cœfficients MFCC.

1.5.4 Le codage PLP

Le codage PLP (PerceptualLinear Predictive) est basé sur l'intégration des connaissances sur la perception humaine pour l'extraction des paramètres.Les paramètres PLP

sont calculés à partir d'un spectre représentant le contenu fréquentiel du signal suivant l'échelle de Bark (correspondant à l'échelle des bandes critiques du système auditif humain). La procédure de calcul des cœfficients PLP est représentée par la Figure 1.10, qui nous donne aussi une comparaison des procédures de calcul des coéfficients PLP et LPC.

Une amélioration intéressante au codage PLP est apportée par l'utilisation du filtrage RASTA (RelAtive SpecTrAl technique) qui à pour but d'éliminer les distorsions des bruits additifs ou convolutifs.

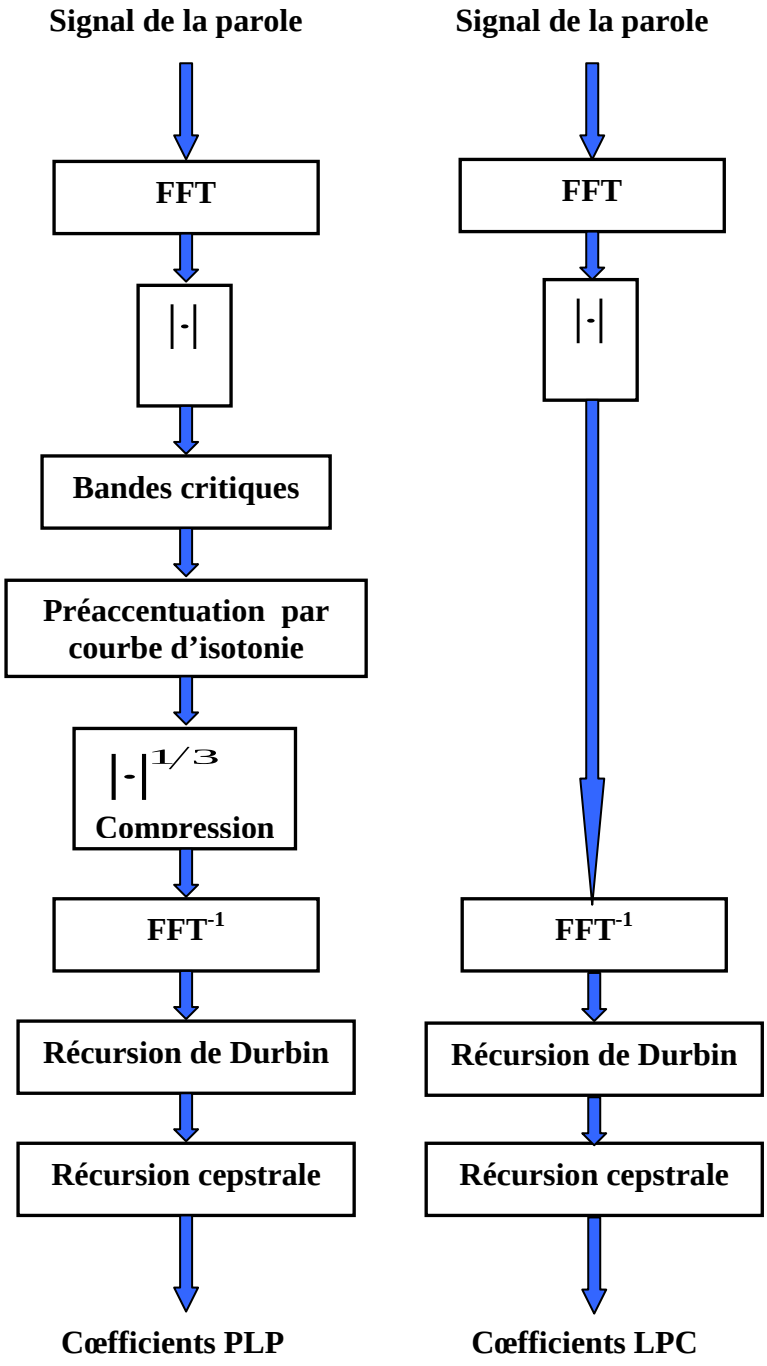

Figure 1.10 : Comparaison entre les méthodes de calcul des cœfficients PLP et LPC.

1.6 Conclusion

Dans ce chapitre nous avons dans un premier temps cité les notions de base de la parole, le bruit et le rehaussement de la parole. Dans une secondeétape, nous avons présentés les traitements numériques appliqués au signal de la parole et les paramètres basés sur les deux principaux modèles exploités en traitement de la parole: les modèles de production et de perception. On peut citer par exemple, le codage prédictif linéaire (LPC:Linear Predictive Coding) qui est directement lié au modèles de production, les codages MFCC (Mel Frequency Cepsral Coefficient) et PLP (Perceptual Linear Predictive) qui s'inspirent eux du modèle de perception. Ces paramètres sont utilisés dans l'analyse, le codage et la reconnaissance de la parole. Ces notions de base et paramètres sont nécessaire à la compréhension des algorithmes développés dans les prochains chapitres.

Chapitre 2

Techniques de rehaussement de la parole

Chapitre 2 :

Techniques de rehaussement de la parole

2.1 Introduction

Les méthodes classiques de rehaussement de la parole se basent sur la soustraction spectrale, qui a comme inconvénient majeur l'apparition d'un bruit résiduel gênant à la perception. Dans ce chapitre nous présentons quelques méthodes de rehaussement de la parole, qui réduisent de façon significative le bruit additif, et qui s'affranchissent de l'effet musical.

La simplicité d'application, l'efficacité et le critère temps réel sont les plus importants critères qui entre en jeu dans le choix de la méthode de rehaussement de la parole. Dans le cadre de notre travail, nous avons également choisi de nous tourner vers le cas des méthodes mono-voie puisque c'est le contexte le plus courant.

2.2 La soustraction spectrale

La soustraction spectrale est une méthode largement implémenté dans les systèmes de codage et de reconnaissance de la parole dans un milieu hostile, à cause de sa simplicité et son efficacité à réduire le bruit de fond. Cette méthode propose de calculer une estimée du bruit sur des portions du signal ne contenant pas de parole, donc elle demande un détecteur de parole/non parole robuste. Sous l'hypothèse que le bruit soit stationnaire, l'estimée du bruit est soustraite du spectre de puissance du signal bruité.

2.2.1 Principe 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 :

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

Le principe de la soustraction spectrale à court- terme est résumé par l'organigramme de la Figure 2.1.

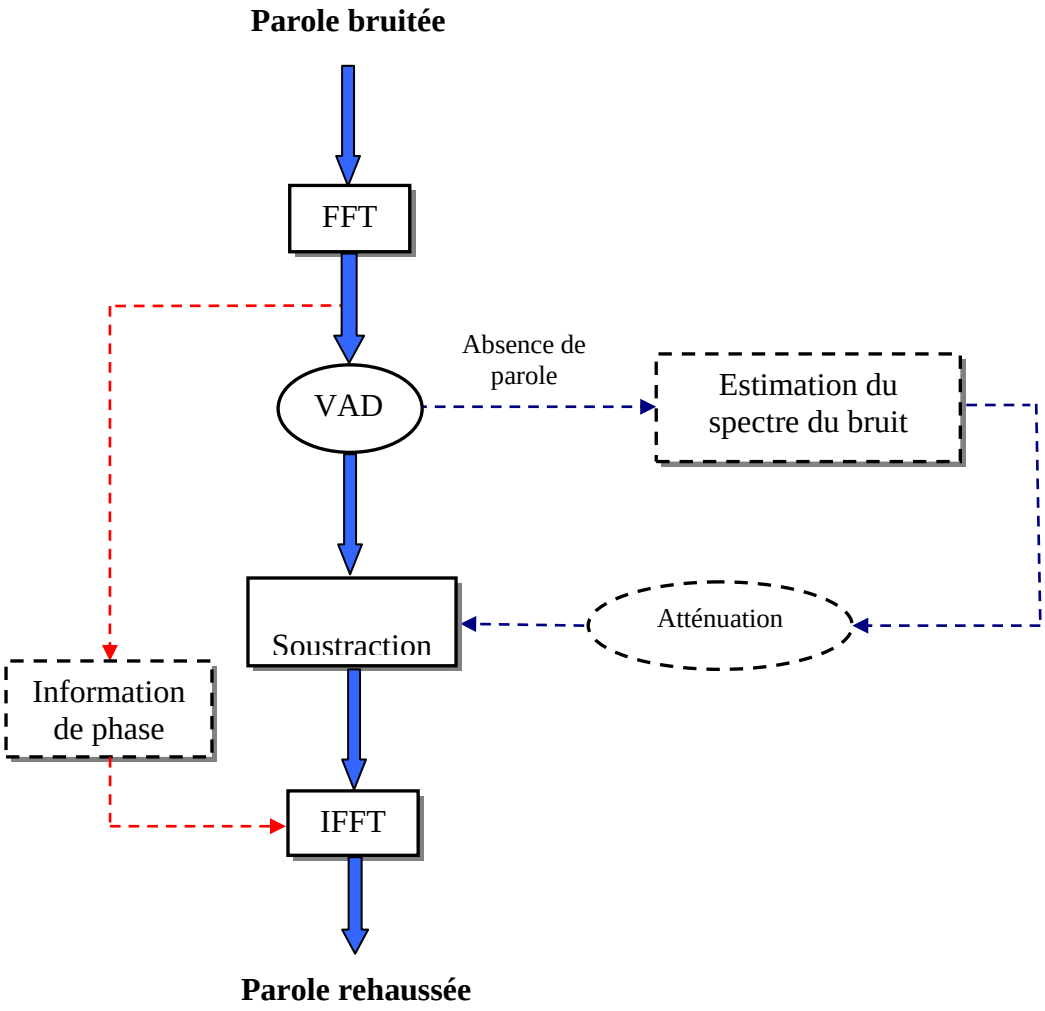

Figure 2.1 : Principe de la soustraction spectrale.

2.2.2 La soustraction spectrale d'amplitude

La soustraction spectrale d'amplitude est une technique de la soustraction spectrale proposée par M. Boll et al en 1979 [1]. En appliquant la transformée de Fourier à court terme sur l'équation (2.1) on obtient :

$$Y(k) = X(k) + D(k) \tag{2.2}$$

$$|Y(k)|\, e^{j\theta_Y(k)} = |X(k)|\, e^{j\theta_X(k)} + |D(k)|\, e^{j\theta_D(k)} \tag{2.3}$$

Dans les régions du bruit seul, l'amplitude du bruit est remplacée par sa valeur moyenne $\mu(k)$. Puisque l'oreille est insensible aux modifications de la phase des signaux, on peut remplacer la phase $\theta_D(k)$ du bruit par la phase $\theta_Y(k)$ du signal bruité. L'estimation du signal propre est donnée par :

$$\hat{X}(k) = \left[|Y(k)| - \mu(k)\right] e^{j\theta_Y(k)} \tag{2.4}$$

En remplaçant $|Y(k)|$ par sa moyenne $|\overline{Y(k)}|$ on aura :

$$\hat{X}(k) = \left[|\overline{Y(k)}| - \mu(k)\right] e^{j\theta_Y(k)} \tag{2.5}$$

L'estimation $\hat{X}(k)$ peut avoir des valeurs négatives, pour résoudre ce problème deux méthodes sont proposées :

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

Dans notre cas, on utilise la rectification demi trame

Pendant l'inactivité vocale l'estimation du signal propre est atténuée, on a donc :

$$\hat{X}(k) = \begin{cases} \hat{X}(k) & T \geq -12dB \\ c\,Y(k) & T \leq -12dB \end{cases} \tag{2.6}$$

Avec :

T : le rapport signal sur bruit :

$$T = 20Log_{10}\left[\frac{1}{2\Pi}\int_{-\Pi}^{\Pi}\left|\frac{\hat{X}(k)}{\mu(k)}\right| dk\right]$$

c : un facteur d'atténuation avec $20Log_{10}(c) = -30dB$.

Le seuil T=-12 dB est calculé expérimentalement.

2.2.3 La soustraction spectrale de puissance

Parmi les méthodes mono-voie de la soustraction spectrale, la méthode de la soustraction spectrale de puissance a pour le but d'améliorer la qualité de la parole dégradée par un bruit additif. Tout d'abord comme il est admis ci-dessus, le spectre de puissance du signal bruité est égal à la somme du spectre de puissance de la parole propre et du spectre de puissance du bruit.

$$\left|\mathrm{Y}(k)\right|^2 = \left|X(k)\right|^2 + \left|D(k)\right|^2 \tag{2.7}$$

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 $\mathrm{E}\left[\left|D(k)\right|^2\right]$, il est aussi noté $\left|\hat{D}(k)\right|$. L'estimation du signal propre $\left|\hat{X}(k)\right|^2$ est liée à l'estimation du spectre du bruit par la relation suivante :

$$\left|\hat{X}(k)\right|^2 = \left|\mathrm{Y}(k)\right|^2 - \left|\hat{D}(k)\right|^2 \tag{2.8}$$

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.

La surestimation du bruit provoque le problème de l'apparition des valeurs négatives pour l'estimation du spectre de puissance du signal propre. La phase initiale est conservée pendant le traitement du signal.

Dans le but d'obtenir de bons résultats, la rectification demi-trame est utilisée comme suit :

$$\left|\hat{X}(k)\right|^2 = \begin{cases} \left|Y(k)\right|^2 - \left|\hat{D}(k)\right|^2 & \text{si} \quad \left|\mathrm{Y(k)}\right|^2 > \left|\hat{D}(k)\right|^2 \\ 0 & \text{ailleurs} \end{cases} \tag{2.9}$$

La généralisation de l'équation (2.8) nous donne :

$$\left|\hat{X}(k)\right|^{\alpha} = \left|Y(k)\right|^{\alpha} - \left|\hat{D}(k)\right|^{\alpha} \tag{2.10}$$

Dans le cas où α =2, la méthode est la soustraction spectrale de puissance, quand α =1, c'est la soustraction spectrale d'amplitude proposée par Boll [1].

Dans ce qui suit, on va présenter trois techniques de la soustraction spectrale de puissance : la soustraction spectrale de Berouti [2]**,** la soustraction spectrale paramétrique [20] et la soustraction spectrale multi bande [21] [22]**.**

2.2.3.1 La soustraction spectrale de Berouti

Berouti et al [2] ont révolutionné la méthode de la soustraction spectrale de puissance en apportant des modifications à l'algorithme de base de cette dernière.

Avec ces modifications apportées, la méthode de la soustraction spectrale devient plus efficace et diffère des autres méthodes par deux applications :

- en soustrayant un facteur α multiplié avec le spectre du bruit où α est un nombre supérieur à l'unité.
- en prévenant les composantes spectrales du signal d'aller au-dessous d'une certaine limite.

Les équations peuvent être exprimées de la manière suivante :

$$\left|\hat{X}(k)\right|^{2} = \left|\mathrm{Y}(k)\right|^{2} - \alpha\left|\hat{D}(k)\right|^{2} \tag{2.11}$$

$$\left|\hat{X}(k)\right|^{2} = \begin{cases} \left|Y(k)\right|^{2} - \alpha\left|\hat{D}(k)\right|^{2} & \text{si} \quad \left|\hat{\mathrm{X}}(\mathrm{k})\right|^{2} > \beta\left|\hat{D}(k)\right|^{2} \\ \beta\left|\hat{D}(k)\right|^{2} & \text{ailleurs} \end{cases} \tag{2.12}$$

Avec :

α : Le facteur de soustraction (surestimation), (α >1).

β : Le paramètre de lissage spectral, (0< β <<1).

La méthode de Berouti est représentée par le diagramme de la Figure 2.2 suivant:

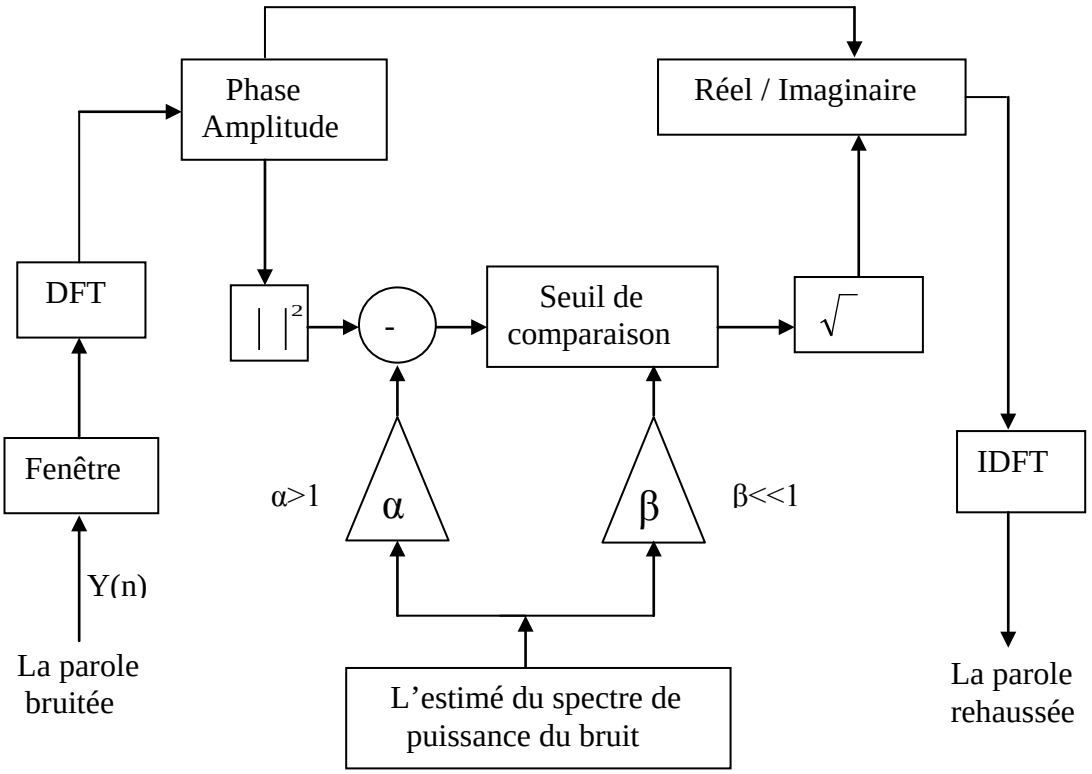

Figure 2.2 : La soustraction spectrale proposée par Berouti et al [2].

- **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.
- 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 assurée 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 se fait, mais cela provoque une distorsion du signal de la parole.

Afin d'obtenir des valeurs optimales du facteur de soustraction α , il faut prendre en compte que α est une fonction du rapport signal sur bruit segmental, dont sa valeur réelle est donnée par la relation suivante :

$$\alpha = \begin{cases} 5 & \text{SNR} < \text{-5dB} \\ \alpha_0 - (SNR/s) & \text{-5dB} < \text{SNR} < \text{20dB} \\ 1 & \text{SNR} > \text{20dB} \end{cases} \qquad (2.13)$$

Avec :

α_0 : La valeur de α pour un SNR =0. Dans la pratique $3 < \alpha_0 < 6$.

1/s : la pente de la droite dans la Figure 2.3.

SNR : le rapport signal sur bruit segmental estimé.

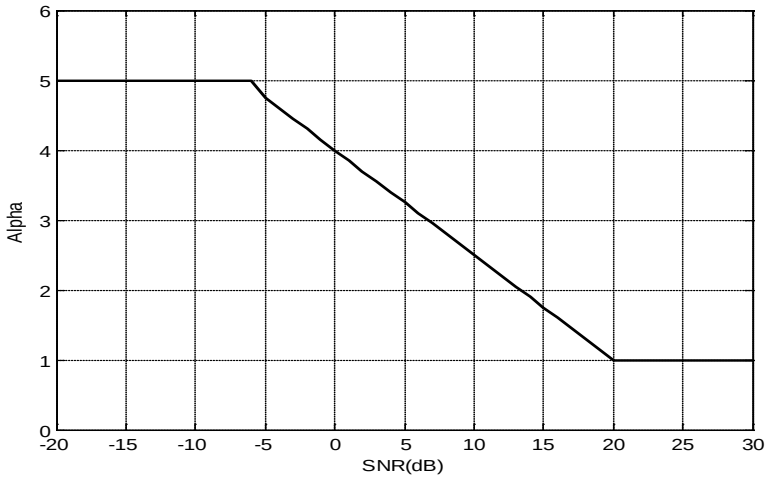

Figure 2.3 : Les valeurs de α en fonction du SNR.

Outre la réduction des pics, il y a le problème du remplissage des vallées d'où la réduction du bruit musical. Cela est effectué 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 musical. Toutefois, malgré cet inconvénient du bruit musical, la méthode de la soustraction spectrale reste performante en termes d'atténuation du bruit.

2.2.3.2 La soustraction spectrale paramétrique

J .S. Chang et al [20] utilisent une formulation paramétrique de la méthode originale donnée par (2.10) avec la contrainte $a_k = b_k$:

$$\left|\hat{X}(k)\right|^{\alpha} = a_k \left|Y(k)\right|^{\alpha} - b_k \left|\hat{\mathrm{D}}(k)\right|^{\alpha} \tag{2.14}$$

On à :

$\left|\hat{X}(k)\right|^{\alpha}$: L'estimé paramétrique de $\left|X(k)\right|$.

a_k et b_k : des paramètres dans la formulation paramétrique.

L'estimé paramétrique du signal de la parole est optimisée par la minimisation de l'erreur quadratique moyenne entre le spectre idéal et le spectre de la parole rehaussée.

Le spectre d'amplitude estimé est donné par :

$$\left|\hat{X}(k)\right| = \left\{ \frac{\xi^2(k)}{\xi^2(k) + 0.5} \left(\left|Y(k)\right|^2 - \left|\hat{\mathrm{D}}(k)\right|^2 \right) \right\}^{1/2} \tag{2.15}$$

Où la moyenne du SNR a priori $\xi(k)$ est donnée par :

$$\xi(k) \approx (1-\eta) \underbrace{\frac{\left|X(k)\right|^2_{est}}{\left|\hat{D}(k)\right|^2}}_{\approx SNR\ courant} + \eta \underbrace{\frac{\left|\hat{X}(k)\right|^2_{prev}}{\left|\hat{D}(k)\right|^2_{prev}}}_{\approx SNR\ précédent} \tag{2.16}$$

Avec :

$$\left|X(k)\right|^2_{est} = Max\left(\left|Y(k)\right|^2 - \left|\hat{D}(k)\right|^2, 0 \right) \tag{2.17}$$

Où :

η : Constante de lissage.

$\left|\hat{X}(k)\right|^2_{prev}$: L'estimé de la trame précédente.

$\left|Y(k)\right|^2 - \left|\hat{D}(k)\right|^2$: Le spectre de puissance de la parole estimé dans la trame courante.

Parfois une très grande réduction du bruit peut affecter les segments de parole à faible énergie, donc une limite inférieure lissée $\mu\overline{|Y(k)|}$ est nécessaire pour limiter l'atténuation du signal.

μ : Gain de lissage spectral.

L'estimé paramétrique final avec contrainte est donné comme suite :

$$|\overline{X}(k)| = \begin{cases} |\hat{X}(k)| & si\ |\hat{X}(k)| \geq \mu |Y(k)| \\ \overline{\mu|Y(k)|} & ailleurs. \end{cases} \quad (2.18)$$

2.2.3.3 La soustraction spectrale multi bande

Le signal de parole n'est pas affecté uniformément par le bruit réel, donc quelques fréquences sont plus affectées par rapport aux autres. En effet, la soustraction spectrale multi bande [21], [22] propose d'estimer pour chaque fréquence un facteur qui soustraira juste le niveau nécessaire du spectre du bruit.

Le spectre de la parole est divisé en N bandes non-chevauchantes et la soustraction spectrale est appliquée séparément dans chaque bande comme suit:

$$|\hat{X}_i(k)|^2 = |Y_i(k)|^2 - \alpha_i\, \delta_i\, |\hat{D}_i(k)|^2 \qquad b_i \leq k \leq e_i \quad (2.19)$$

b_i et e_i : les bornes inférieures et supérieures de la bande de fréquence i.

δ_i : 'Tweaking factor' un facteur qui peut être mis pour chaque bande de fréquence pour personnaliser la réduction du bruit.

α_i : Le facteur de soustraction de la bande i. Ce facteur est calculé en fonction du SNR segmental de la bande i par la relation :

$$\alpha_i = \begin{cases} 5 & SNR_i \leq -5 \\ 4 - \dfrac{3}{20} SNR_i & -5 \langle SNR_i \leq 20 \\ 1 & SNR_i \rangle 20 \end{cases} \quad (2.20)$$

Avec :

$$SNR_i\,(dB) = 10\log_{10}\left(\frac{\sum_{k=b_i}^{e_i}\left|Y_i(k)\right|^2}{\sum_{k=b_i}^{e_i}\left|\hat{D}_i(k)\right|^2}\right) \tag{2.21}$$

Les valeurs négatives dans le spectre de parole rehaussée sont remplacées par les valeurs du spectre bruité pondérées par un facteur de lissage β :

$$\left|\hat{X}_i(k)\right|^2 = \begin{cases}\left|\hat{X}_i(k)\right|^2 & si \quad \left|\hat{X}_i(k)\right|^2 \rangle 0\\ \beta\left|Y_i(k)\right|^2 & ailleurs.\end{cases} \tag{2.22}$$

2.2.4 Limitation des méthodes basées sur la soustraction spectrale

Rappelons que le problème majeur des méthodes de réduction de bruit basées sur la soustraction spectrale est l'apparition d'un bruit sous forme de pics dans le spectre du signal rehaussé appelé bruit musical. Comme le spectre à court terme du bruit fluctue autour des valeurs moyennes, son amplitude atteint à certains instants des valeurs largement supérieures à la moyenne, la bande de fréquence correspondante est traitée comme du signal utile et relativement moins atténuée que les composantes fréquentielles voisines.

Donc pour mieux réduire ce bruit musical d'autres méthodes plus complexes ont été proposées. Parmi ces techniques, le filtrage de Wiener.

2.3 Le filtre de Wiener

2.3.1 Principe du filtrage de Wiener

Le filtre de Wiener est basé sur la minimisation de l'erreur quadratique moyenne (EQM) entre la sortie de filtre et une sortie désirée. Il est entièrement déterminé par les caractéristiques statistiques des signaux, donc c'est un filtre optimal et le problème de filtrage optimal consiste à déterminer le filtre qui donne la meilleure estimation du signal désiré. Ce filtre est employé pour les situations dans lesquelles le signal et le bruit sont stationnaires.

Le principe de base du filtre de Wiener [27] est d'obtenir une estimation du signal utile (signal désiré) $x(n)$ à partir d'un signal $y(n)$ dégradé par un bruit additif.

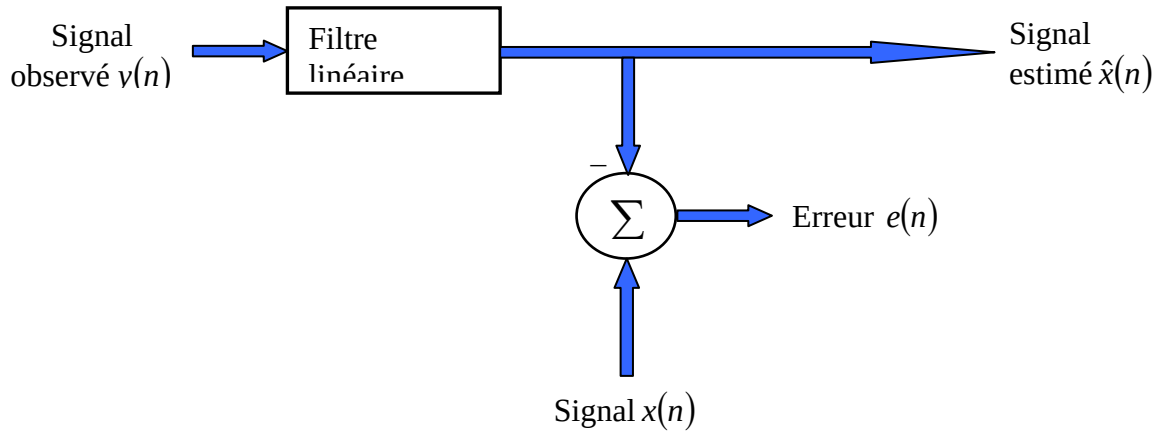

Figure 2.4 : Schéma général du filtrage de Wiener.

Cette estimation est obtenue par le filtrage de $y(n)$ de telle sorte que la sortie $\hat{x}(n)$ soit la plus proche possible de $x(n)$. La qualité de l'estimation est définie par :

$$e(n) = x(n) - \hat{x}(n) \tag{2.23}$$

On peut noter que, plus $e(n)$ sera faible plus l'estimation sera meilleure.

2.3.3 Critère d'optimisation

Dans la synthèse des filtres optimaux, le problème majeur est de déterminer le filtre qui minimisera l'erreur. Pratiquement, il suffit de minimiser $e^2(n)$, car c'est une fonction quadratique facilement dérivable. Donc l'erreur quadratique moyenne (EQM) est la fonction qui sera minimisée, elle est définie par :

$$J = E(e^2(n)) \tag{2.24}$$

Une condition nécessaire pour que $J(n)$ ait un minimum est que son gradient soit nul :

$$\frac{\partial J}{\partial \mathbf{h}} = 0 \tag{2.25}$$

$\mathbf{h}$ est la réponse impulsionnelle du filtre H que nous recherchons, elle est donnée par une notation matricielle de longueur M par :

$$\mathbf{h} = [h_0 \quad h_1 \quad \ldots \quad h_{M-1}]^T \tag{2.26}$$

2.3.4 L'équation de Wiener-Hopf

Le signal estimé $\hat{x}(n)$ est donné par :

$$\hat{x}(n) = \sum_{i=0}^{M-1} h_i y(n-i) \tag{2.27}$$

La notation matricielle est la suivante :

$$\hat{x}(n) = h^T \mathrm{y}(n) \Leftrightarrow \hat{x}(n) = \mathrm{y}^T(n) h \tag{2.28}$$

Avec :

$$\mathbf{y}(n) = \begin{bmatrix} y(n) & y(n-1) & \ldots & y(n-(M-1)) \end{bmatrix}^T \tag{2.29}$$

Sous l'hypothèse que $x(n)$ et $y(n)$ sont stationnaires et par introduction de (2.24) et (2.29) dans l'équation (2.25), on trouve la fonction suivante :

$$\begin{aligned} J &= E\left[\left(x(n) - \mathbf{h}^T \mathbf{y}(n)\right)^2\right] \\ &= E\left[x^2(n) - 2\mathbf{h}^T \mathbf{y}(n) x(n) + \mathbf{h}^T \mathbf{y}(n) \mathbf{y}^T(n) \mathbf{h}\right] \end{aligned} \tag{2.30}$$

$$J = E\left[x^2(n)\right] - 2\mathbf{h}^T \mathbf{r}_{yx}^{-} + \mathbf{h}^T \mathbf{R}_{yy} \mathbf{h} \tag{2.31}$$

Avec :

$$\mathbf{R}_{yy} = E\left[\mathbf{y}(n) \mathbf{y}^T(n)\right] \tag{2.32}$$

$$\mathbf{r}_{yx}^{-} = E\left[\mathbf{y}(n) x(n)\right] = E\left[\left(\mathbf{y}(n)\, \mathbf{y}(n-1) \ldots \mathbf{y}(n-N+1)\right) x(n)\right] \tag{2.33}$$

Où :

$\mathbf{R}_{yy}$: La matrice d'autocorrélation $(M \times M)$ de y.

$\mathbf{r}_{yx}^{-}$: Le vecteur d'intercorrélation $(M \times 1)$ de x et y.

L'indice supérieur dans $\mathbf{r}_{yx}^{-}$ est utilisé pour indiquer que le m^{ème} élément du vecteur d'intercorrélation est en fait $\mathbf{r}_{yx}(-m)$.

Pour 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. Cette dérivée est donnée par :

$$\frac{\partial J}{\partial \mathbf{h}} = -2\mathbf{r}_{yx}^{-} + 2\mathbf{h}^T \mathbf{R}_{yy} \tag{2.34}$$

La minimisation de J conduit à l'équation :

$$\frac{\partial J}{\partial \mathbf{h}} = 0 \Leftrightarrow \frac{\partial J}{\partial \mathbf{h}} = -2\mathbf{r}_{yx}^{-} + 2\mathbf{h}^T \mathbf{R}_{yy} = 0 \tag{2.35}$$

On arrive à la fonction suivante :

$$\mathbf{h}^* = \mathbf{R}_{yy}^{-1}\mathbf{r}_{yx}^{-} \tag{2.36}$$

C'est l'équation de Weiner-Hopf dans le cas de filtres. Avec $\mathbf{h}^*$ représente le vecteur optimum qui annule le gradient du critère.

Cette équation peut être représentée sous forme de matrice comme suit :

$$\begin{bmatrix} h_0 \\ h_1 \\ h_2 \\ \vdots \\ h_{M-1} \end{bmatrix} = \begin{bmatrix} r_{yy}(0) & r_{yy}(1) & r_{yy}(2) & \cdots & r_{yy}(M-1) \\ r_{yy}(1) & r_{yy}(0) & r_{yy}(1) & \cdots & r_{yy}(M-2) \\ r_{yy}(2) & r_{yy}(1) & r_{yy}(0) & \cdots & r_{yy}(M-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{yy}(M-1) & r_{yy}(M-1) & \cdots & r_{yy}(1) & r_{yy}(0) \end{bmatrix}^{-1} \begin{bmatrix} r_{yx}(0) \\ r_{yx}(-1) \\ r_{yx}(-2) \\ \vdots \\ r_{yx}(-M+1) \end{bmatrix} \tag{2.37}$$

2.3.5 Application du filtre de Wiener pour le débruitage de la parole

- **Principe de la méthode**

Soit le signal bruité :

$$y(n) = x(n) + d(n) \tag{2.38}$$

Où $x(n)$ représente le signal de parole et $d(n)$ le bruit additif, qui sont deux processus aléatoires stationnaires.

$$\mathbf{R}_{yy} = E\left[\mathbf{y}\mathbf{y}^T\right] = E\left[(x+d)(x+d)^T\right] = E\left[xx^T\right] + E\left[dd^T\right] + E\left[xd^T\right] + E\left[dx^T\right] \quad (2.39)$$

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

$$\mathbf{R}_{yy} = \mathbf{R}_{xx} + \mathbf{R}_{dd} \quad (2.40)$$

On remplace l'équation (2.40) dans l'équation (2.36) on trouve :

$$\mathbf{h}^* = \left(\mathbf{R}_{xx} + \mathbf{R}_{dd}\right)^{-1} \mathbf{r}_{yx}^- \quad (2.41)$$

- **Dans le domaine fréquentiel**

On a :

$$\hat{x}(n) = h(n) * y(n) \xrightarrow{TF} \hat{X}(k) = H(k)Y(k) \quad (2.42)$$

Donc :

$$E(k) = X(k) - \hat{X}(k) = X(k) - H(k)Y(k) \quad (2.43)$$

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

$$E\left[|E(k)|^2\right] = E\left[|X(k)|^2\right] - H(k)E\left[X^*(k)Y(k)\right] - H^*(k)E\left[Y^*(k)X(k)\right] + |H(k)|^2 E\left[|Y(k)|^2\right] \quad (2.44)$$

$$J_2 = E\left[|E(k)|\right] = E\left[|X(k)|^2\right] - H(k)P_{yx}(k) - H^*(k)P_{xy}(k) + |H(k)|^2 P_{yy}(k) \quad (2.45)$$

Où :

$P_{xy}(k)$: La densité interspectrale de puissance de x et y.

$P_{yy}(k)$: La densité spectrale de puissance de y.

$$\frac{\partial J_2}{\partial H(k)} = H^*(k)P_{yy}(k) - P_{yx}(k) = \left[H(k)P_{yy}(k) - P_{xy}(k)\right]^* = 0 \quad (2.46)$$

$$H(k)=\frac{P_{xy}(k)}{P_{yy}(k)} \tag{2.47}$$

$$Y(k)=X(k)+D(k) \tag{2.48}$$

$$P_{xy}(k)=E\left[X(k)\{X(k)+D(k)\}^{*}\right]=E\left[X(k)X^{*}(k)\right]+E\left[X(k)D^{*}(k)\right]=P_{xx}(k) \tag{2.49}$$

$$\begin{aligned} P_{yy}(k) &= E\left[\{X(k)+D(k)\}\{X(k)+D(k)\}^{*}\right] \\ &= E\left[X(k)X^{*}(k)\right]+E\left[D(k)D^{*}(k)\right]+E\left[X(k)D^{*}(k)\right]+\mathrm{E}\left[D(k)X^{*}(k)\right] \\ &= P_{xx}(k)+P_{dd}(k) \end{aligned} \tag{2.50}$$

Avec :

$P_{xx}(k)$: La densité spectrale de puissance de x.

$P_{dd}(k)$: La densité spectrale de puissance de d.

En remplaçant les équations (2.49) et (2.50) dans (2.47), on trouve :

$$H(k)=\frac{P_{xx}(k)}{P_{xx}(k)+P_{dd}(k)} \tag{2.51}$$

On pose :

$$\xi_{k}=\frac{P_{xx}(k)}{P_{dd}(k)} \tag{2.52}$$

ξ_{k} : Le rapport signal sur bruit a priori (SNR a priori).

Si on introduit l'équation (2.52) dans l'équation (2.51), on arrive à la fonction de transfert du filtre suivante :

$$H(k)=\frac{\xi_{k}}{\xi_{k}+1} \tag{2.53}$$

Donc, la fonction du gain de Wiener est :

$$G_w(\xi_k, \gamma_k) = \frac{\xi_k}{\xi_k + 1} = H(k) \tag{2.54}$$

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).

2.3.6 Limitation du filtre de Wiener

Le filtre de Wiener vise à atténuer le spectre à court-terme des observations bruitées et il parvient de cette manière à réduire efficacement le niveau de bruit de fond. En contrepartie, ce filtre offre les limitations suivantes :

- Le calcul de $\mathbf{R}_{yy}$ et $\mathbf{r}_{yx}$ est nécessaire, ce qui augmente le temps de calcul.
- Pour résoudre l'équation de Wiener-Holf il faut inverser la matrice $\mathbf{R}_{yy}$ et cette opération demande beaucoup de calcul et d'espace mémoire.
- Dans le cas où les signaux ne sont pas stationnaires, $\mathbf{R}_{yy}$ et $\mathbf{r}_{yx}$ évoluent au cours du temps, donc il faut résoudre l'équation de Wiener-Holf à chaque instant.

Le filtrage de Wiener est inadéquat pour les situations dans lesquelles le signal ou le bruit sont non stationnaires. Dans de telles situations, le filtre optimal doit être variable dans le temps. La solution à ce problème est fournie par le filtrage de Kalman.

2.4 Le filtre de Kalman

Le filtre de Kalman a été développé par Rodolph Kalman en 1960, bien que Peter Swerling ait développé un algorithme très semblable en 1958. Ce filtre a été développé pour résoudre le problème d'estimation de la trajectoire pour le programme Appolo.

Dans sa célèbre publication en 1960 '' A new approch to linear filtering and prediction problems '' [36] R.Kalman a basé la construction de filtre d'estimation d'état sur la théorie des probabilités. Le filtre de Kalman est un ensemble d'équations mathématiques qui permet une meilleure estimation de l'état futur d'un système malgré l'imprécision des mesures et de modélisation.

2.4.1 Filtrage de Kalman

Le filtre de Kalman peut être vu comme un estimateur récursif et optimal. Il consiste à chercher une estimation de l'état x du système à chaque instant, en minimisant la variance de l'erreur d'estimation.

Le filtre de Kalman est appliqué à des systèmes qui peuvent être modélisés par des équations différentielles linéaires stochastiques. Nous nous intéressons au cas discret qui est le plus simple et surtout le plus utilisé [37] [38] [39].

L'estimation de l'état $x \in \Re^n$ d'un système avec une mesure $y \in \Re^m$ à chaque pas k nous donne ce qui suit :

- Nous commençons par l'équation d'état :

$$x_k = Ax_{k-1} + Bu_k + w_{k-1} \tag{2.55}$$

Avec :

x_k : L'état du système.

$A(n \times n)$: La matrice de prédiction relie l'état x_{k-1} à l'état x_k.

$B(n \times 1)$: Vecteur de consigne reliant l'état x_k à un signal de contrôle (consigne) u_k.

u_k : Consigne appliquée à l'entrée du système.

w_k : Bruit de système.

A cause du bruit, on ne peut pas résoudre cette équation et connaître exactement x. Pour résoudre ce problème, nous utilisons une équation de mesure qui nous fournisse une approximation de l'état réel.

- L'équation de mesure (ou d'observation) est donnée par :

$$y_k = Hx_k + v_k \tag{2.56}$$

Avec :

y_k : Vecteur de mesure.

$H(m \times n)$: Équation de mesure reliant l'état x_k à la mesure y_k.

v_k : Bruit de mesure.

Bien que toute la théorie du filtre de Kalman soit valable dans le cas non-stationnaire, nous supposons que le système et les bruits sont non stationnaires.

Les matrices A, B et H sont stationnaires et déterministes. Par ailleurs, on admet que w_k et v_k sont des bruits blancs avec une distribution qui suit une loi normale de

moyenne nulle et de matrice de covariance non nulle, et non corrélée entre eux. Leurs matrices de covariances ont pour expressions :

$$E[w_k] = 0. \qquad E[w_k w_i^T] = Q, pour 1 \le i \le k \ .$$

$$E[v_k] = 0. \qquad E[v_k v_i^T] = R, pour 1 \le i \le k$$

$$E[w_k v_i^T] = 0, pour 1 \le i \le k$$

Cette relation traduit l'indépendance stockastique des bruits w_k et v_k. Cette hypothèse est introduite pour alléger les calculs qui vont suivre. On pourrait représenter les densités de probabilité des variables w_k et v_k par :

$$\mathrm{P}(w) \approx N(0, Q) \tag{2.57}$$

$$\mathrm{P}(v) \approx N(0, R) \tag{2.58}$$

Avec :

Q et R les matrices de covariance de bruit de système et du bruit de mesure respectivement.

2.4.1.1 Caractérisation des différents estimateurs

Nous avons :

$$e^{-}_k = x_k - \hat{x}^{-}_k \tag{2.59}$$

$$e_k = x_k - \hat{x}_k \tag{2.60}$$

Telle que :

e^{-}_k : Estimation a priori de l'erreur.

e_k : Estimation a posteriori de l'erreur.

$\hat{x}^{-}_k$: Estimateur a priori.

$\hat{x}_k$: Estimateur a posteriori.

Les covariances des erreurs estimés a priori et a posteriori sont données par :

$$P^{-}_k = E[e^{-}_k e^{-T}_k] \tag{2.61}$$

$$P_k = E[e_k e_k^T] \tag{2.62}$$

D'après les équations précédentes, les estimateurs a priori et a posteriori sont ainsi liés par l'équation du filtre de Kalman appelée aussi équation de correction donnée par :

$$\hat{x}_k = \hat{x}^{-}_k + K_k\left(y_k - H\hat{x}^{-}_k\right) \tag{2.63}$$

Avec :

K_k : Gain de Kalman

$\left(y_k - H\hat{x}_k^-\right)$: La différence entre la mesure y_k et la prédiction $H\hat{x}_k^-$, appelée innovation (résiduel).

2.4.1.2 Choix du gain de Kalman K_k

Le but ici est de minimiser la covariance de l'erreur a posteriori P_k . La matrice de gain de Kalman obtenue à la fin sert à corriger la matrice de covariance à posteriori P_k .

Le gain de Kalman est donné par la relation suivante :

$$K_k = \frac{P_k^- H^T}{H\ P_k^- H^T + R} \tag{2.64}$$

Où R représente la covariance du bruit de mesure.

- Plus la covariance de l'erreur de mesure R approche de zéro plus le gain augmente et favorise le résiduel $\left(y_k - H\hat{x}_k^-\right)$ donc on obtient :

$$\lim_{R\to 0} K_k = H^{-1} \tag{2.65}$$

Et l'équation du filtre de Kalman est donnée comme suit :

$$\hat{x}_k = H^{-1} y_k \tag{2.66}$$

On remarque que l'état de l'étape précédente $\hat{x}_k^-$ est éliminé.

- Plus la covariance de l'erreur d'estimation a priori $P^-{}_k$ approche de zéro, plus le gain diminue et moins le résiduel a d'importance, donc on a :

$$\lim_{P^-\to 0} K_k = 0 \tag{2.67}$$

L'équation du filtre de Kalman est donnée comme suit :

$$x_k = \hat{x}_k^- \tag{2.68}$$

2.4.1.3 Etapes du filtre de Kalman

Le problème du filtre de Kalman est de trouver la meilleure estimation de l'état x_k , à partir de l'observation effectuée jusqu'à l'instant k .

Cette estimation est obtenue en utilisant un contrôle avec retour d'état, c'est-à-dire on à une rétroaction sous forme d'observations. Cependant, le filtre de Kalman est constitué de deux étapes (Figure 2.5) :

- **Etape de prédiction (mise à jour du temps)**

Dans cette étape, l'estimation de l'état courant est obtenue en utilisant l'état estimé à l'instant précédent. En plus, une estimation prédite de la covariance de l'erreur est effectuée.

- **Etape de correction (mise à jour de mesure)**

Pour améliorer la précision de l'estimation, l'état prédit est corrigé en utilisant les observations de l'instant courant. Au cour de cette étape, on calcule le gain de Kalman et on effectue une mise à jour de la matrice de covariance de l'état et de l'estimation de l'état.

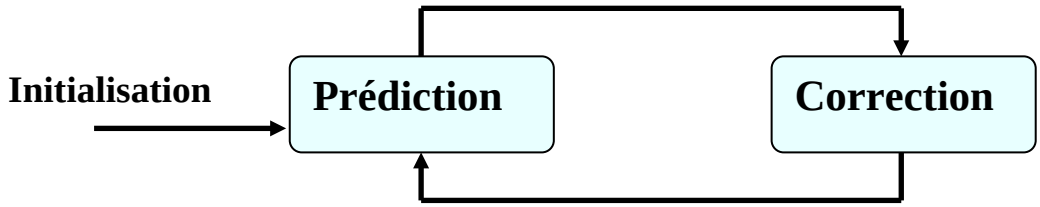

Figure 2.5: Etapes du filtre de Kalman.

2.4.1.4 Algorithme du filtre de Kalman

L'algorithme du filtre de Kalman ressemble à un algorithme de prédiction correction pour la résolution des problèmes numériques. Pour le filtre de Kalman on a deux types d'équations ; les équations de prédiction (propagation ou mise à jour du temps) et les équations de correction (mise à jour de la mesure). On peut représenter l'algorithme du filtre de Kalman comme suit :

Initialisation

Les valeurs initiales sont : $\hat{x}_{k-1}^{-}$ et P_{k-1}

Etape de prédiction

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

$P_k^- = AP_{k-1}A^T + Q$ (Estimation prédite de la covariance)

Etape de correction

$K_k = P_k^- H^T (HP_k^- H^T + R)^{-1}$ (Gain de Kalman)

$\hat{x}_k = \hat{x}_k^- + K_k (y_k - H\hat{x}_k^-)$ (Mise à jour de l'état)

$P_k = (I - K_k H)P_k^-$ (Mise à jour de la covariance)

Pour mieux comprendre le fonctionnement du filtre de Kalman on peut représenter l'algorithme de Kalman sous la forme suivante (voir Figure 2.6) :

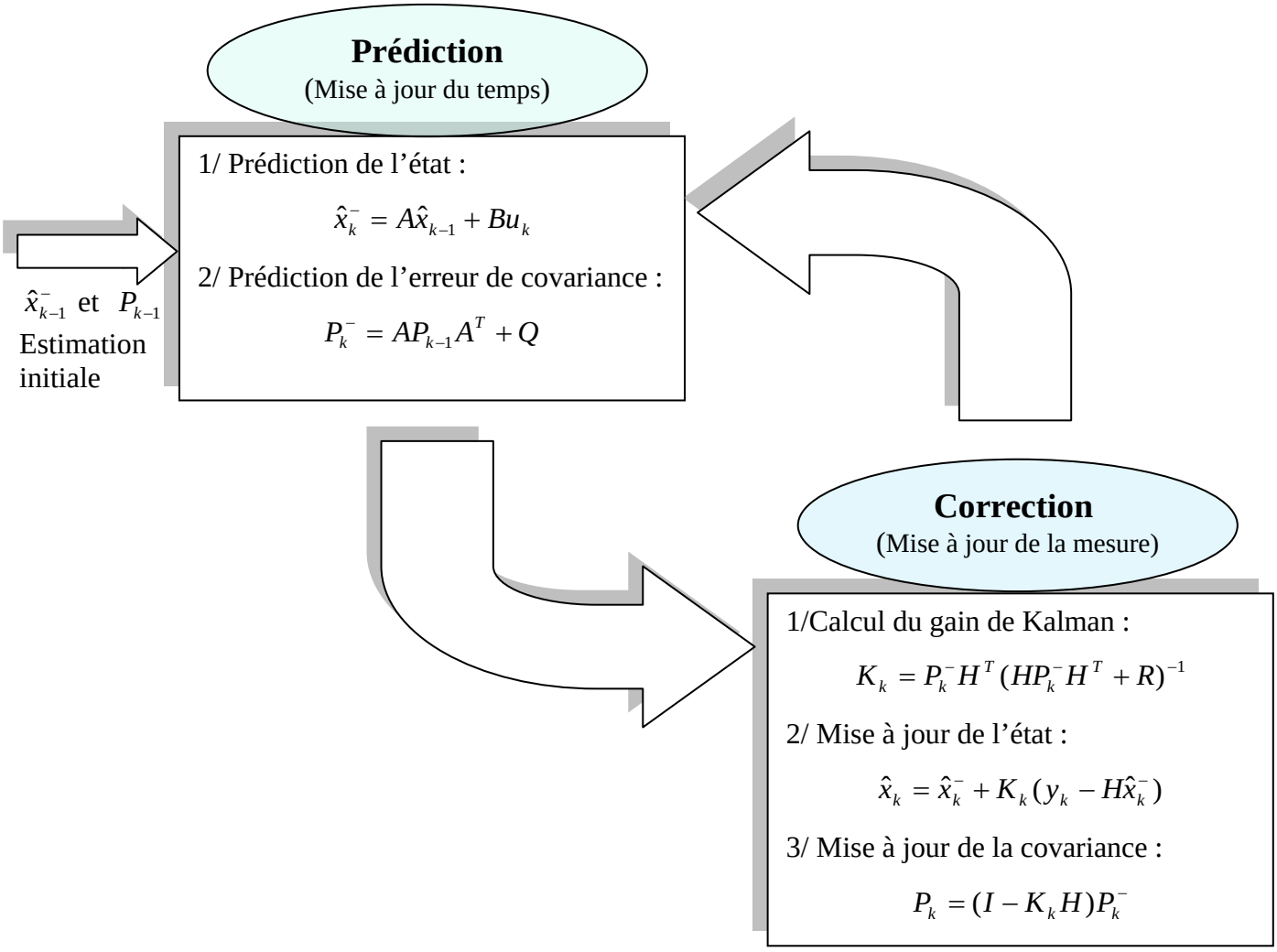

Figure 2.6: Schéma complet des opérations du filtre de Kalman.

Une estimation initiale de la covariance du bruit de mesure R est effectuée pendant la phase de silence au début de chaque signal, cette covariance est mise à jour au cour de l'opération du filtrage. Cependant, la détermination de la covariance du bruit de processus (Q) est très difficile. Pour pallier ce problème, on utilise un modèle AR simple et une analyse LPC où (Q) représente le gain de la prédiction.

La stabilité de P_k (matrice de covariance de l'erreur) et K_k (gain de Kalman) est liée à la stabilité de (Q) et (R).

2.4.2 Rehaussement de la parole par filtre de Kalman

Le signal de la parole est modélisé comme un processus autorégressif (AR) et représenté dans l'espace d'état. Le filtre de Kalman est le meilleur estimateur linéaire au sens de l'erreur quadratique moyenne, il diffère des autre méthodes de rehaussement de la parole par :

- La modélisation du signal de la parole par un modèle autorégressive.
- Le problème du bruit musical ne se pose pas dans les signaux rehaussés par le filtre de Kalman.
- Le filtre de Kalman est valable pour des signaux de parole non-stationnaire.

Dans ce qui suit on va représenter l'application du filtre de Kalman pour deux types de bruit différent le bruit blanc et le bruit coloré.

2.4.2.1 Filtrage de Kalman pour la parole dégradé par un bruit blanc

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

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + u(n) \tag{2.69}$$

Ce signal est bruité par un bruit blanc [9] [40] [41], et le signal de la parole bruité est donné par :

$$y(n) = s(n) + v(n) \tag{2.70}$$

Avec :

a_i : Cœfficients de la prédiction linéaire (LPC).

$u(n)$: Bruit de processus.

$v(n)$: Bruit additif.

On peut représenter le système dans l'espace d'état comme suit :

$$\begin{cases} x(n) = Fx(n-1) + Gu(n) \\ y(n) = Hx(n) + v(n) \end{cases} \tag{2.71}$$

Avec :

$x(n)$: Vecteur d'état $(p \times 1)$, constitué par les p dernières valeurs du signal $s(n)$.

$$x(n) = [s(n-p+1).....s(n)]^T \tag{2.72}$$

F : Matrice de transition $(p \times p)$ donnée par :

$$F = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 1 \\ a_p & a_{p-1} & a_{p-2} & \cdots & a_1 \end{bmatrix} \tag{2.73}$$

G : Vecteur d'entrée $(p \times 1)$ donné par :

$$G = [0,.....0,1]^T \tag{2.74}$$

H : Vecteur d'observation $(1 \times p)$ donné par :

$$H = [0,.....,0,1] = G^T \tag{2.75}$$

$u(n)$ et $v(n)$ sont des bruits blancs, Gaussien, indépendants de moyennes nulles et de matrices de covariance Q et R respectivement.

Le filtre de Kalman calcule $\hat{x}_n$ l'estimation du vecteur d'état x_n en utilisant les équations récursives suivantes :

$$\hat{x}_n^- = F.\hat{x}_{n-1}^- \tag{2.76}$$

$$P_n^- = F.P_{n-1}^-.F^T + G.Q.G^T \tag{2.77}$$

$$K_n = P_n^-.H[H.P_n^-.H^T + R_n]^{-1} \tag{2.78}$$

$$\hat{x}_n = \hat{x}_n^- + K_n[y_n - H.\hat{x}_n^-] \tag{2.79}$$

$$P_n = [I - K_n.H].P_n^- \tag{2.80}$$

Avec :

$\hat{x}_n^-$: L'estimation du vecteur d'état x_n au sens de la minimisation de l'erreur quadratique moyenne donné par les observations passée $y(1),....., y(n-1)$.

P_n^- : Matrice de covariance de l'erreur de l'état prédit.

K_n : Gain de Kalman.

$\hat{x}_n$: L'estimation filtrée de vecteur d'état x_n.

P_n : Matrice de covariance de l'erreur de l'état filtré.

L'estimation de la parole propre peut être obtenue à partir de l'estimation filtrée du vecteur d'état par la relation :

$$\hat{s}(n) = H\hat{x}_n \tag{2.81}$$

Bien que la matrice de transition $H = [0,.....,0,1]$ on peut dire que la dernière composante du vecteur d'état estimé représente l'estimation filtrée de Kalman du signal de la parole $s(n)$.

Plusieurs méthodes d'estimations peuvent être utilisé pour estimer les cœfficients AR $\{a_1.....a_p\}$ et les variances des bruits Q et R pour chaque segment stationnaire de la parole.

2.4.2.2 Filtrage de Kalman pour la parole dégradé par un bruit coloré

Le signal de la parole bruité est donné par :

$$y(n) = s(n) + v(n) \tag{2.82}$$

$s(n)$: Signal de la parole propre.

$v(n)$: Bruit additif.

Le filtre de Kalman a pour but de calculer à partir de l'observation bruitée $y(n)$ l'estimation $\hat{s}(n)$ de la parole propre au sens de l'erreur quadratique moyenne.

L'application du filtre de Kalman dans le cas d'un bruit coloré [10] [44] [45], nécessite la modélisation du signal de la parole $s(n)$ et de bruit additif $v(n)$ par deux modèles autorégressifs (AR) d'ordre p et q respectivement :

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + u(n) \tag{2.83}$$

$$v(n) = \sum_{j=1}^{q} b_j v(n-j) + w(n) \tag{2.84}$$

Avec :

$s(n)$: Le n^{eme} échantillon du signal de la parole.

$v(n)$: Le n^{eme} échantillon de bruit.

a_i : Le i^{eme} paramètre du modèle AR de la parole.

b_j : Le j^{eme} paramètre du modèle AR du bruit.

La représentation de ce système dans l'espace d'état est la suivante :

$$\begin{cases} x(n) = Fx(n-1) + Gu(n) \\ y(n) = Hx(n) \end{cases} \tag{2.85}$$

Avec :

$x(n)$: Le vecteur d'état $(p+q)\times 1$:

$$x(n) = [s(n-p+1).....s(n), v(n-q+1), ..., v(n)]^T \tag{2.86}$$

G : Le vecteur d'entrée $(p+q)\times 1$:

$$G = [\underbrace{0,.....,0,1}_{p}, \underbrace{0,.....,0,1}_{q}]^T \tag{2.87}$$

F : La matrice de transition $(p+q)\times(p+q)$:

$$F = \begin{bmatrix} F_s & 0 \\ 0 & F_v \end{bmatrix} \tag{2.88}$$

$$F_s = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 1 \\ a_p & a_{p-1} & a_{p-2} & \cdots & a_1 \end{bmatrix}, \; F_v = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 1 \\ b_q & b_{q-1} & b_{q-2} & \cdots & b_1 \end{bmatrix} \tag{2.89}$$

H : La matrice d'observation $1\times(p+q)$:

$$H = \left[\underbrace{0,.....,0,1}_{p}, \underbrace{0,.....,0,1}_{q} \right] \tag{2.90}$$

$u(n)$ et $w(n)$ sont des bruits blanc, Gaussien, indépendants, de moyennes nulles et de matrices de covariance (Q) et (R) respectivement. On peut noter que les équations de l'algorithme du filtre de Kalman reste les même pour les deux types de bruits blanc et coloré.

Dans l'application du filtre de Kalman au rehaussement de la parole, des paramètres doivent être estimés (telle que les cœfficients LPC et les matrices de covariance). La qualité du signal rehaussé par filtre de Kalman dépend de la fiabilité de l'estimation de ces paramètres.

Dans le cas idéal, l'estimation des différents paramètres est effectuée à partir de la parole propre, cependant dans la pratique ce n'est pas possible car le parole propre est inconnue a priori, donc seule la parole bruitée est disponible pour l'extraction des différents paramètres.

Une estimation imprécise des paramètres dégrade les performances du filtre de Kalman, pour pallier ce problème on utilise l'algorithme EM [54] qui sert à estimer et corriger ces paramètres d'une manière récursive.

2.5 Conclusion

Il existe de nombreuses méthodes de rehaussement de la parole, ces méthodes sont choisies selon des critères bien définis. La complexité de la tache de rehaussement réside dans la nature du signal de la parole et le type du bruit.

Dans les trois dernières décennies, l'intérêt a été axé sur le développement de méthodes de plus en plus performantes et rigoureuses en matière d'élimination de bruit tout en préservant la qualité et l'intelligibilité du signal de la parole.

Nous avons décrit dans ce chapitre trois méthodes de rehaussement de la parole : La soustraction spectrale, le filtre de Wiener et le filtre de Kalman. Dans le prochain chapitre, nous présenterons les résultats expérimentaux de rehaussement de parole bruitée obtenus avec les méthodes mono-voie, puisque c'est le cas le plus adapté aux applications réelles d'annulation du bruit.

Chapitre 3

Application des techniques de rehaussement de la parole : Résultats expérimentaux

Chapitre 3 :

Application des techniques de rehaussement de la parole : Résultats expérimentaux

3.1 Introduction

Les critères d'évaluation de la qualité de la parole ont le point commun de donner une idée globale de la qualité du signal vocal. Les mesures subjectives sont les plus fiables pour évaluer la qualité de la parole. Les mesures objectives se présentent comme une alternative aux méthodes subjectives et permettent d'automatiser l'évaluation de la qualité de la parole.

Dans ce chapitre, nous avons implémenté et mis en œuvre différents algorithmes de rehaussement de la parole étudiés précédemment. Des tests et des essais sont appliqués sur chacun des algorithmes pour une évaluation objective de la qualité.

3.2 Evaluation de la qualité et l'intelligibilité de la parole

L'évaluation de la qualité du signal de la parole peut être effectuée en utilisant des tests subjectifs ou des tests objectifs.

L'évaluation subjective est basée sur la comparaison entre le signal original et le signal rehaussé à travers des tests d'écoute par un groupe d'auditeurs. Cette évaluation est précise mais coûteuse.

L'évaluation objective est basée sur une comparaison mathématique entre le signal original et le signal rehaussé. Elle est pratique, moins coûteux et moins précise par rapport à l'évaluation subjective.

Une forte corrélation entre les tests objectifs et les tests subjectifs est nécessaire pour que ces tests objectifs soient valides.

3.2.1 La qualité et l'intelligibilité de la parole

La qualité est l'une des attributs du signal de la parole. Elle est liée à l'aspect agréable de l'écoute du signal rehaussé par l'auditeur. Pour évaluer la qualité on utilise des tests subjectifs ou des tests objectifs qui ont simulé notre perception, tel que le PESQ (Perceptual Evaluation of Speech Quality).

L'intelligibilité évalue la signification des mots parlés elle est mesurée en comptant le nombre de mots ou de phonèmes identifiés correctement par l'auditeur. Les tests d'écoute restent le meilleur moyen pour évaluer l'intelligibilité.

La parole peut être fortement intelligible mais d'une faible qualité et contrairement, elle peut avoir une bonne qualité mais complètement n'est pas intelligible. Donc on peut dire que la qualité et l'intelligibilité de la parole sont deux choses différentes.

3.2.2 Evaluation subjective

L'évaluation subjective est basée sur des tests d'écoute effectués par un groupe d'auditeurs. Pendant l'évaluation subjective il faut prendre en compte le degré de perception de l'auditeur, son humeur et l'environnement. Les tests subjectifs les plus utilisés sont le MOS (Mean Opinion Score), le DMOS (Degradation Mean Opinion Score) et le CMOS (Comparaison Mean Opinion Score).

Le MOS (Mean Opinion Score) : est le test subjectif le plus utilisé, dans lequel les auditeurs évaluent la qualité du signal de la parole selon une échelle de cinq note (Tableau 3.1), où '5' indique une excellente qualité et '1' indique une mauvaise qualité. La qualité du signal de la parole est obtenue en faisant la moyenne des notes obtenues par tous les auditeurs. Cette moyenne est connu sous le nom: Note Moyenne D'opinion (MOS : Mean Opinion Score). L'avantage du MOS est qu'elle est une évaluation réelle et fiable.

Tableau 3.1: Echelle MOS.

Note MOS	Qualité MOS
5	Excellent
4	Bon
3	Passable
2	Médiocre
1	Mauvaise

3.2.3 Evaluation objective

L'évaluation objective est basée sur des calculs mathématiques entre le signal rehaussé et le signal original. Les tests objectifs les plus utilisés sont classés dans le Tableau 3.2 suivant :

Tableau 3.2: Classification des tests objectifs.

Domaine temporel	Domaine fréquentiel	Domaine perceptuel
SNR	IS	BSD, MBSD
SNR_{seg}	CD	PSQM
	WSS	PESQ
	LLR	

Les tests objectifs dans le domaine temporel et fréquentiel se basent sur le calcul de la distorsion de forme entre le signal original et le signal rehaussé. Le problème qui se pose est que ces tests sont limités par le fait qu'on peut avoir une perception différente pour deux signaux de même forme. Pour pallier ces limitations et avoir une bonne corrélation avec les tests subjectifs, des tests objectifs perceptuels sont développés tel que le PESQ.

Dans ce qui suit, nous allons détailler deux tests objectifs, le SNR_{seg} pour évaluer le rapport signal sur bruit du signal rehaussé et le PESQ qui a une bonne corrélation avec les tests subjectifs.

- **SNR segmental (SNR_{seg})**

Dans le domaine temporel le SNR segmental (segmental Signal to Noise Ratio) est un test qui présente une forte corrélation avec les tests subjectifs par rapport au rapport signal sur bruit global. Il est obtenu en faisant la moyenne de tous les rapports signal sur bruit de chaque trame comme suit :

$$SNR_{seg} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Nm}^{Nm+N-1}x^2(n)}{\sum_{n=Nm}^{Nm+N-1}\left[x(n)-\hat{x}(n)\right]^2} \qquad (3.1)$$

Avec :

$x(n)$: Signal propre.

$\hat{x}(n)$: Signal rehaussé.

N : Taille de la trame.

M : Nombre de trame.

Pendant les segments de silence, la valeur du SNR en dB sera négative, ce qui affecte et dégrade le SNR total. Pour résoudre ce problème, il faut choisir un seuil d'énergie au-delà duquel le SNR sera calculé.

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

Le test PESQ est une amélioration du test PSQM (Perceptual Speech Quality Measure) Il est basé sur la comparaison entre le signal rehaussé et le signal de référence pour obtenir une note de la qualité d'écoute que pourrait attribuer un auditeur au signal rehaussé.
Ces notes seront appliquées à une échelle du type MOS (Mean Opinion Score) sous forme d'un scalaire compris entre -0.5 et 4.5. On peut dire que ce test donne une bonne corrélation avec les tests subjectifs.

3.3 Mise en œuvre des techniques de rehaussement de la parole

3.3.1 La soustraction spectrale de Berouti

L'estimation de la densité spectrale de puissance de bruit est l'un des éléments critiques de toute technique de rehaussement de la parole. Ainsi, l'efficacité des méthodes de rehaussement par soustraction spectrale repose sur une estimation préalable correcte du niveau de bruit. Pratiquement un système de rehaussement de la parole se compose de deux composantes essentielles, l'estimation du spectre de puissance du bruit et l'estimation de la parole [23]. Le calcul du spectre moyen de quelques trames du début du signal nous donne une estimée initiale du spectre de puissance de bruit.

La mise à jour de l'estimation de bruit est réalisée durant des périodes de l'inactivité vocale. Ces périodes sont déterminées en utilisant un détecteur d'activité vocale (VAD : Voice Activity Detection). L'objectif d'un détecteur d'activité vocale (VAD) est de détecter la présence ou l'absence de la parole. Le détecteur de VAD qui a été incorporé dans cette méthode est un système basé sur un modèle statistique [49], [50]. Le rapport de vraisemblance de la parole présente ou absente dans la trame courante est donnée par :

$$\frac{1}{N}\sum_{n=0}^{N-1}\left\{\log\left[\frac{1}{1+\xi_n^{DD}}\exp\left(\frac{\gamma_n\xi_n^{DD}}{1+\xi_n^{DD}}\right)\right]\right\} \underset{silence}{\overset{parole}{\gtrless}} \eta \qquad (3.2)$$

Avec :

η : Seuil de préréglage

γ_n : Le SNR a posteriori.

ξ_n^{DD} : Le SNR a priori estimé par la méthode de la décision dirigée.

Pendant l'absence de la parole, la mise à jour de bruit est effectuée en utilisant la méthode de Welch. Elle est donnée par la relation suivante :

$$\left|\hat{D}_m(k)\right|^2 = \lambda_N.\left|\hat{D}_{m-1}(k)\right|^2 + (1-\lambda_N).\left|Y_m(k)\right|^2 \qquad 0.5 \leq \lambda_N \leq 0.9 \tag{3.3}$$

Avec :

m : L'indice de la trame.

λ_N : Facteur d'oubli, il a été pris $\lambda_N = 0.9$.

$\left|Y_m(k)\right|^2$: Spectre de puissance de la trame en cours d'analyse.

$\left|\hat{D}_{m-1}(k)\right|^2$: Estimé du spectre de puissance du bruit de la trame précédente.

Les différents paramètres de la soustraction spectrale de Berouti [2] ont été choisis d'une manière a assurer de bons résultats.

$\alpha \geq 1$ et dépend du SNR dans chaque trame

Et $\beta = 0.002$

3.3.2 Le filtre de Wiener

La mise en œuvre du filtre de Wiener nécessite la segmentation du signal $y(n)$ en blocs de N échantillons. En tenant compte de la spécifité du signal de parole et dans la mesure où l'enregistrement commence par un silence, donc seul le bruit est présent dans ce cas, on peut avoir une estimation de sa densité spectrale de puissance.

Le SNR a priori d'une composante spectrale est estimé pour chaque trame par une méthode dite décision dirigée [3]. Cet estimateur est combiné avec l'estimateur d'amplitude de Wiener. L'estimation du SNR a priori par la méthode dite décision dirigée est donnée par :

$$\xi_m(k) = \alpha \frac{\left|\hat{X}_{m-1}(k)\right|^2}{\left|\hat{D}_m(k)\right|^2} + (1-\alpha)P[\gamma_m - 1] \tag{3.4}$$

Où :

m : Indice de la trame.

$\left|\hat{X}_{m-1}(k)\right|^2$: Estimation du spectre de puissance de la parole pour la trame précédente.

$\left|\hat{D}_m(k)\right|^2$: Estimée du spectre de puissance du bruit.

γ_m : SNR à posteriori, il est donné par :

$$\gamma_m = \frac{|Y_m(k)|^2}{|\hat{D}_m(k)|^2} \tag{3.5}$$

Une valeur élevée de α permet de limiter le bruit musical. En outre, pour éviter la distorsion du signal de la parole il faut choisir des valeurs basses. Les meilleurs résultats sont obtenus d'après [51] pour une valeur de α égale à 0.98.

3.3.3 Le filtre de Kalman

Pendant l'application du filtre de Kalman, une mauvaise estimation des paramètres AR résulte en une mauvaise qualité du signal de la parole rehaussé. Donc, la qualité du signal rehaussé par le filtre de Kalman est liée à la robustesse des méthodes d'estimation des paramètres AR.

- **Filtre de Kalman simple : Sans modélisation du bruit**

L'implémentation de l'algorithme de Kalman s'est effectuée sous Matlab. Dans ce cas, le filtre de Kalman est appliqué sur les trente phrases de la base de données NOISEUS et le signal de la parole est modélisé par un modèle AR d'ordre p=10.

On applique une analyse LPC sur le signal de la parole propre à chaque trame d'analyse de durée 5ms pour obtenir les cœfficients AR. Dans cette approche le bruit additif est considéré comme un bruit blanc, même dans le cas des bruits réels.

Pour les mêmes caractéristiques type et niveau du bruit, la mesure objective obtenue est une moyenne effectuée sur les trente phrases de la base de données.

- **Filtre de Kalman simple : Avec modélisation du bruit**

Dans cette deuxième approche, le signal de la parole est modélisé par un modèle AR d'ordre p=10 et le bruit réel est modélisé par un modèle AR d'ordre q variable : 2 ,4 , 6 , 8 , 10 , avec une trame d'analyse de durée 5ms.

3.4 Résultats expérimentaux

3.4.1 Base de données utilisée

Dans le but de faciliter la phase de test et de comparaison des algorithmes de rehaussement de la parole que nous avons étudié on utilise la base de données ‘’NOISEUS’’ qui est une des bases de données standards utilisées dans les travaux de recherche consacrés au rehaussent de la parole.

Cette base comporte trente phrases phonétiquement équilibrées, produites par trois locuteurs et trois locutrices, l’enregistrement de ces phrases s’est effectué au niveau des salles acoustiquement isolées avec une fréquence d’échantillonnage égale à 25 KHz , re-échantillonnées à 8 KHz .

Ces phrases seront bruitées artificiellement par différents bruits pris de la base de donnée NOISEX 92 avec des rapports signal sur bruit variant de 0 dB jusqu’à 20 dB.

3.4.2 Evaluation des performances

Nous présentons dans cette section, une évaluation objective des performances des différentes techniques de rehaussement de la parole étudiées précédemment, le but de cette évaluation est de comparer leurs performances en présence de différents bruits pour plusieurs rapports signal sur bruit. L’étude expérimentale est menée sur les 30 phrases du corpus. Les fichiers parole sont bruités additivement par trois types de bruits réels tel que le bruit babble, le bruit factory et le bruit buccaneer de la base de donnée NOISEX’92, avec des rapports signal sur bruit variant de 0 dB jusqu’à 20 dB. Deux types de mesures sont utilisées afin de comparer l’efficacité des méthodes de rehaussement de la parole, soit :

SNR_{seg} : Rapport signal sur bruit segmental.

PESQ : Evaluation perceptuelle de la qualité de la parole.

Les Figures 3.1(a,b) -3.6(a,b) représentent les formes d’ondes et les spectrogrammes d’un fichier de la base de donnés dans les cas de la parole propre, bruitée par un bruit babble à 5dB , rehaussée par la soustraction spectrale de Berouti , rehaussée par le filtre de Wiener , rehaussée par le filtre de Kalman simple sans modélisation du bruit et rehaussée par filtre de Kalman simple avec modélisation du bruit respectivement :

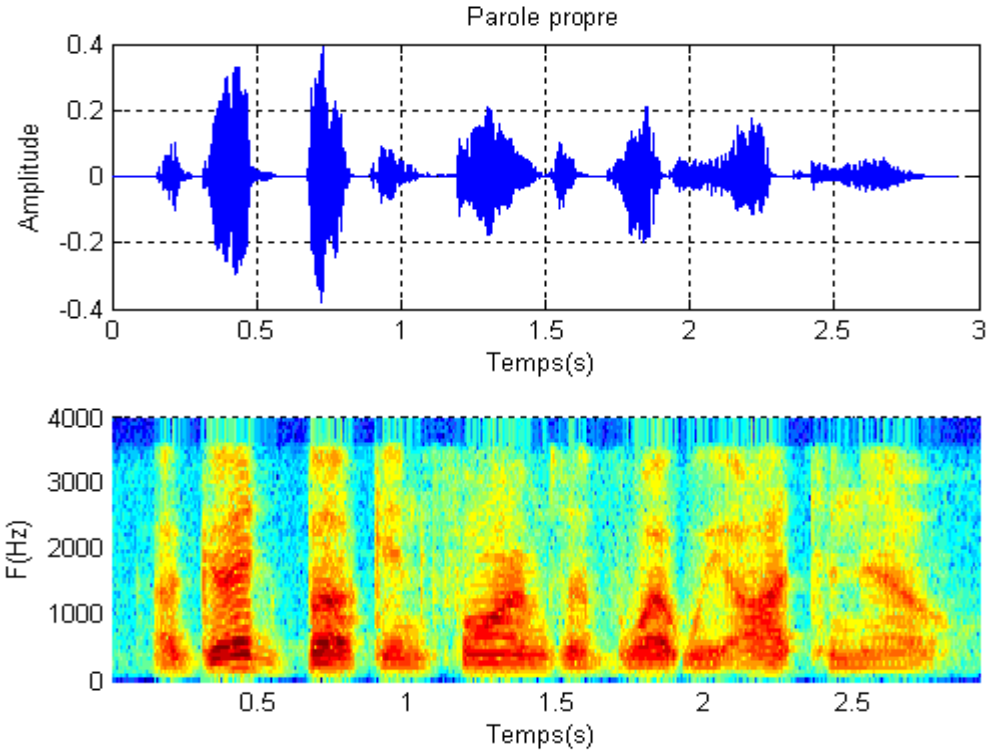

Figure 3.1 (a,b) : La forme d'onde et le spectrogramme de la parole propre.

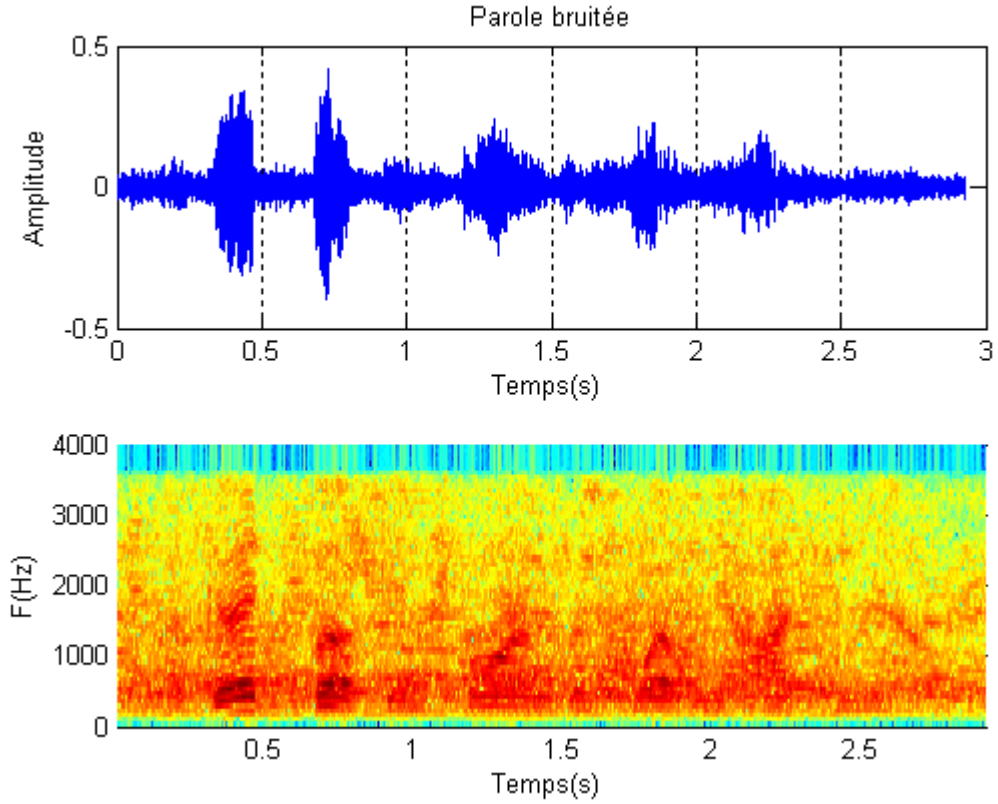

Figure 3.2 (a,b) : La forme d'onde et le spectrogramme de la parole bruitée par un bruit babble à 5dB.

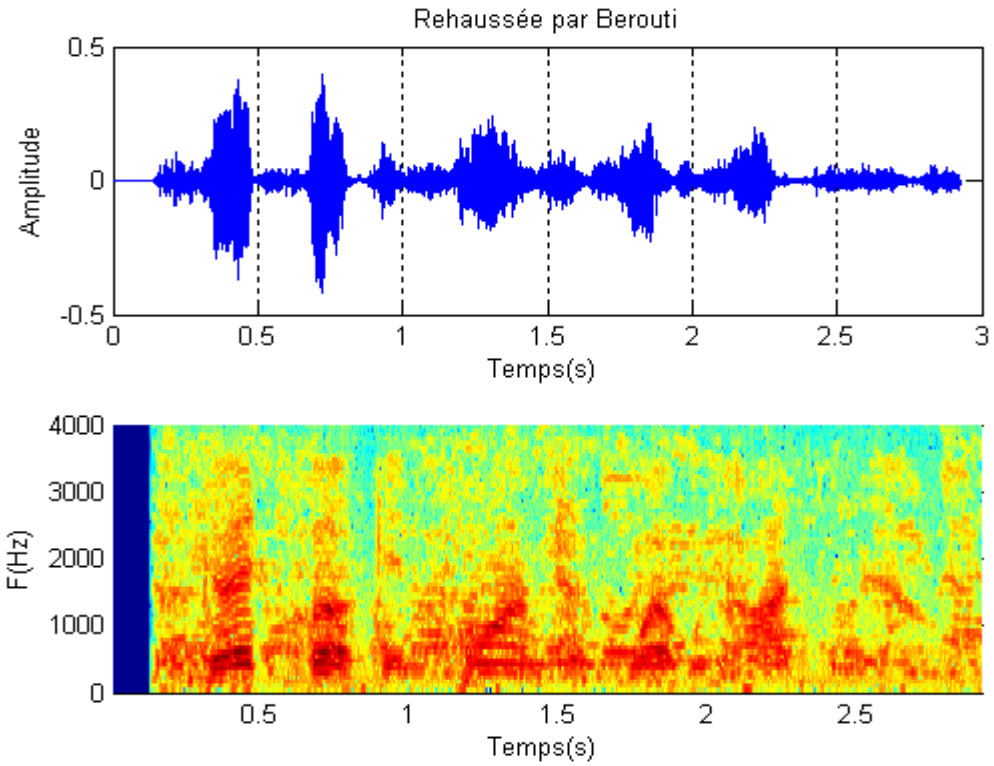

Figure 3.3 (a,b) : La forme d'onde et le spectrogramme de la parole bruitée, rehaussée par la méthode de Berouti.

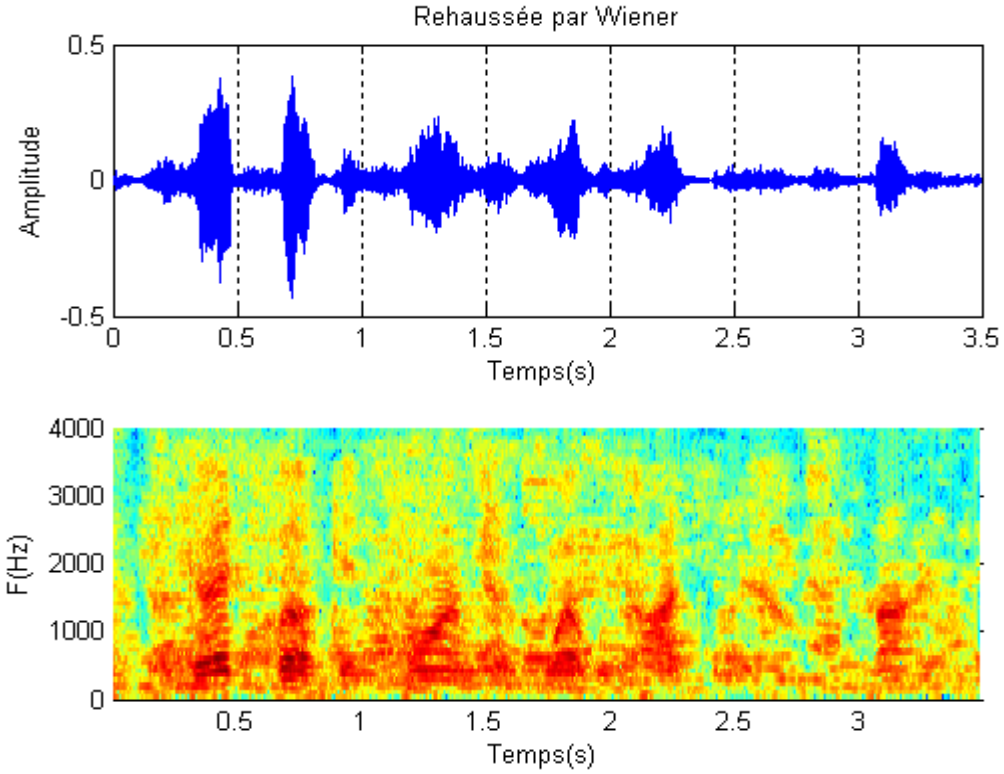

Figure 3.4 (a,b): La forme d'onde et le spectrogramme de la parole bruitée, rehaussée par le filtre de Wiener.

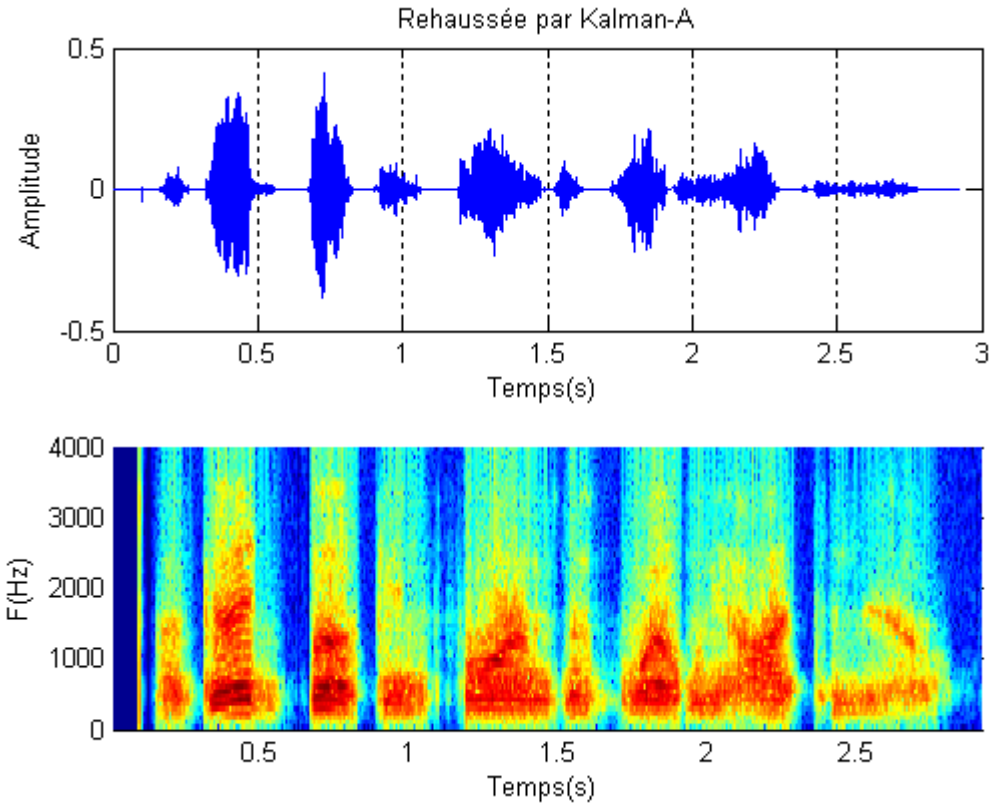

Figure 3.5 (a,b) : La forme d'onde et le spectrogramme de la parole bruitée, rehaussée par le filtre de Kalman sans modélisation du bruit.

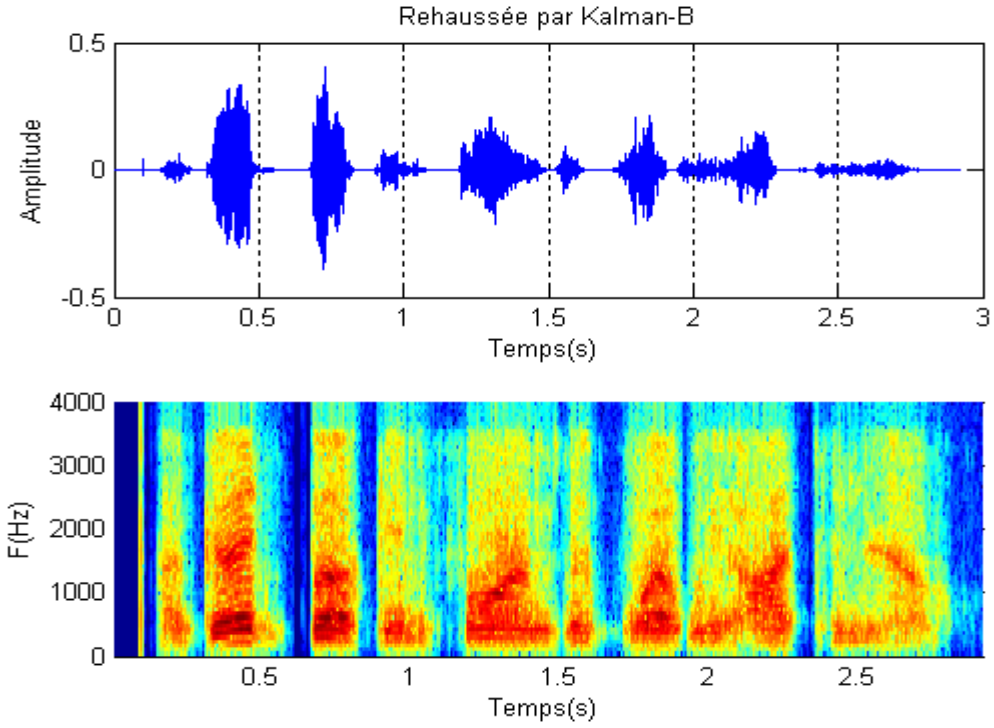

Figure 3.6 (a,b) : La forme d'onde et le spectrogramme de la parole bruitée, rehaussée par le filtre de Kalman avec modélisation du bruit.

- **Résultats expérimentaux avec l'ensemble des 30 phrases**

Chaque valeur obtenue correspond à la moyenne effectuée sur les 30 fichiers du corpus présentant le même type et niveau de bruit. On va effectuer une comparaison entre les résultats des tests **(SNR_{seg}, PESQ)** des différentes méthodes de rehaussement de la parole : la soustraction spectrale de Berouti, le filtre de Wiener, le filtre de Kalman simple sans modélisation du bruit et le filtre de Kalman simple avec modélisation du bruit. Avant de faire cette comparaison, on doit choisir dans un premier temps la valeur q qui nous donne la meilleure modélisation du bruit pour chaque type de bruit réel pour le filtre de Kalman simple avec modélisation du bruit. A ce stade, les résultats des tests **(SNR_{seg}, PESQ)** pour le filtre de Kalman avec modélisation du bruit sont obtenus avec une modalisation du signal de la parole propre d'ordre p=10, et une modélisation du bruit réel avec un ordre q variable : 2, 4,6 ,8 ,10. Les résultats sont représentés dans les tableaux suivants :

-Le tableau (3.3) pour le cas des signaux bruités par un bruit babble.

-Le tableau (3.4) pour le cas des signaux bruités par un bruit factory.

-Le tableau (3.5) pour le cas des signaux bruités par un bruit buccaneer.

Tableau 3.3 : Les résultats de test pour un bruit de chahut dans une cantine (babble).

		0 dB		5 dB		10dB		15dB		20dB	
	(p, q)	SNRseg	PESQ	SNRseg	PESQ	SNRseg	PESQ	SNRseg	PESQ	SNRseg	PESQ
Dégradé		- 4.6321	1.7056	1.7838-	2.0061	1.4197	2.3213	4.8522	2.6529	8.3619	3.0701
babble	(10,2)	2.1691	2.1500	3.8963	2.4317	6.1964	2.7294	8.7115	3.0617	11.7450	3.4622
babble	(10,4)	2.2427	2.1638	3.9816	2.4488	6.2746	2.7439	8.7708	3.0749	11.7230	3.4600
babble	(10,6)	2.3013	2.1655	4.0381	2.4570	6.3242	2.7507	8.8060	3.0819	11.6901	3.4531
babble	**(10,8)**	**2.3067**	**2.1663**	**4.0370**	**2.4603**	**6.3158**	**2.7514**	**8.7986**	**3.0787**	**11.6814**	**3.4541**
babble	(10,10)	2.2924	2.1641	4.0206	2.4552	6.3057	2.7501	8.7889	3.0765	11.6853	3.4558

Tableau 3.4 : Les résultats de test pour un bruit d'usine (factory 1).

		0 dB		5 dB		10dB		15dB		20dB	
	(p, q)	SNRseg	PESQ	SNRseg	PESQ	SNRseg	PESQ	SNRseg	PESQ	SNRseg	PESQ
Dégradé		- 4.8181	1.9425	- 1.9908	2.2755	1.1446	2.6110	4.5531	2.9151	8.4009	3.2469
factory	(10,2)	3.4095	2.4483	5.0467	2.7231	7.0938	3.0248	9.4621	3.3387	12.3542	3.6444
factory	(10,4)	3.4576	2.4412	5.1171	2.7225	7.1734	3.0277	9.5298	3.3442	12.4052	3.6510
factory	**(10,6)**	**3.4444**	**2.4354**	**5.1207**	**2.7223**	**7.1901**	**3.0280**	**9.5320**	**3.3422**	**12.4205**	**3.6551**
factory	(10,8)	3.4164	2.4282	5.1077	2.7165	7.1778	3.0223	9.5137	3.3384	12.4154	3.6559
factory	(10,10)	3.3946	2.4253	5.0875	2.7101	7.1557	3.0251	9.4937	3.3345	12.3960	3.6540

Tableau 3.5 : Les résultats de test pour un bruit d'avion de combat (buccaneer).

		0 dB		5 dB		10dB		15dB		20dB	
	(p, q)	**SNRseg**	**PESQ**	**SNRseg**	**PESQ**	**SNRseg**	**PESQ**	**SNRseg**	**PESQ**	**SNRseg**	**PESQ**
Dégradé		- 5.0476	1.5820	2.2830 -	1.8784	0.8232	2.1988	4.2501	2.5360	8.0440	2.8726
buccaneer	(10,2)	2.4761	2.2589	4.0635	2.5328	6.0455	2.8017	8.4643	3.1065	11.3492	3.4110
buccaneer	(10,4)	2.3862	2.2387	3.9955	2.5202	6.0122	2.7931	8.4418	3.0939	11.3392	3.4024
buccaneer	(10,6)	**2.5514**	**2.2566**	**4.1810**	**2.5415**	**6.1651**	**2.8114**	**8.5767**	**3.1207**	**11.4514**	**3.4232**
buccaneer	(10,8)	2.5027	2.2418	4.1405	2.525	6.1310	2.8006	8.5478	3.1092	11.4301	3.4115
buccaneer	(10,10)	2.4761	2.2374	4.1109	2.5223	6.1014	2.7969	8.5246	3.1071	11.4080	3.4088

Le tableau (3.3) montre qu'un ordre q=8 pour le bruit babble, nous donne les meilleurs résultats pour les deux tests ($SNR_{seg,}$ PESQ). Pour les deux bruits factory et buccaneer, on remarque qu'ils sont bien modélisés par un ordre q=6 d'après les deux tableaux (3.4) et (3.5).

On peut dire que le bruit babble nécessite une modélisation plus élevée par rapport aux autres types de bruit à cause de sa forme qui ressemble au signal de la parole.

Dans un second temps, nous présentons une comparaison entre les résultats des tests **($SNR_{seg,}$ PESQ)** pour les différentes méthodes de rehaussement de la parole.

-Les figures (3.7a,b) représentent les résultats des tests **SNR_{seg}** et **PESQ** pour le cas des signaux bruités par un bruit babble et rehaussés par les différentes méthodes.

-Les figures (3.8a,b) représentent les résultats des tests **SNR_{seg}** et **PESQ** pour le cas des signaux bruités par un bruit factory et rehaussés par les différentes méthodes.

-Les figure (3.9a,b) représentent les résultats des tests **SNR_{seg}** et **PESQ** pour le cas des signaux bruités par un bruit buccaneer et rehaussés par les différentes méthodes.

Les deux notations Kalman-A et Kalman-B dans les figures correspondent respectivement avec filtre de Kalman simple sans modélisation du bruit et filtre de Kalman simple avec modélisation du bruit.

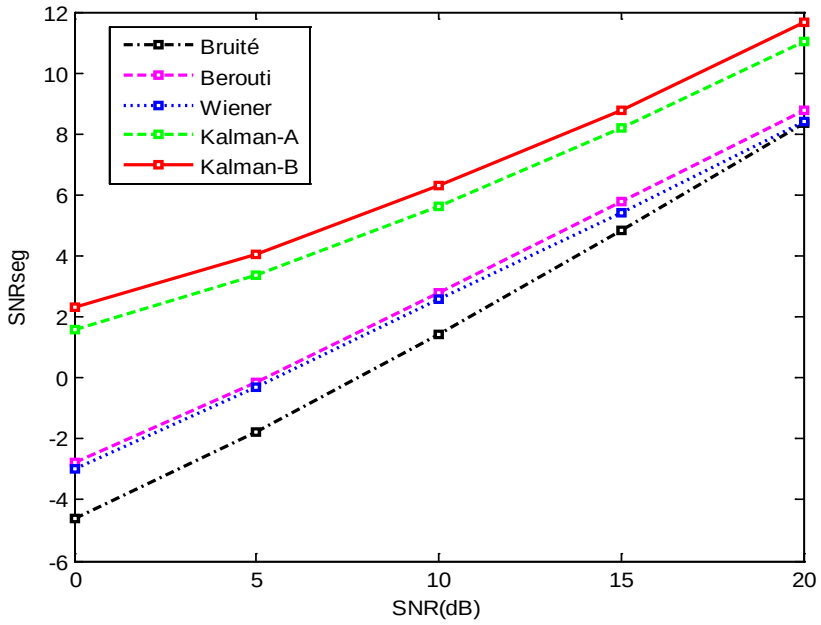

Figure 3.7a : Comparaison des résultats des tests SNR_{seg} avec différentes méthodes de rehaussement dans le cas de la parole bruitée par un bruit babble.

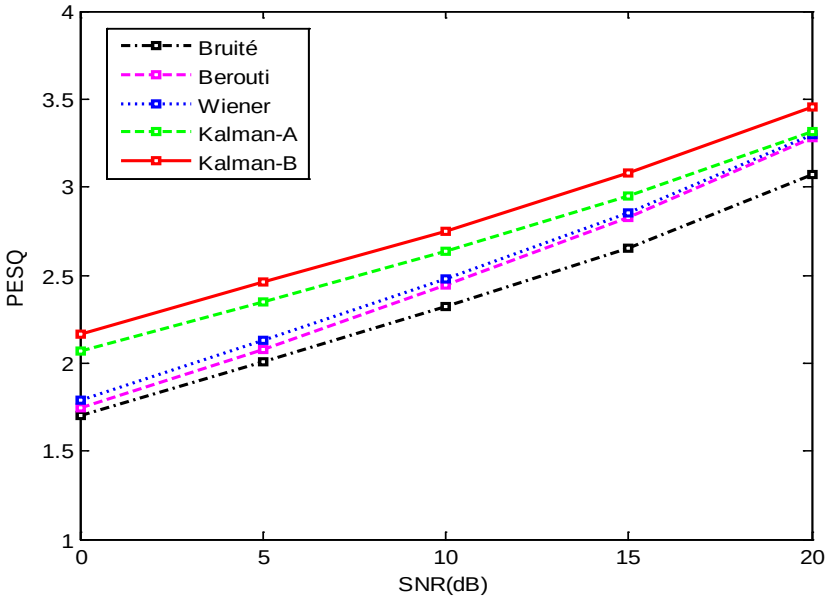

Figure 3.7b : Comparaison des résultats des tests PESQ avec différentes méthodes de rehaussement dans le cas de la parole bruitée par un bruit babble.

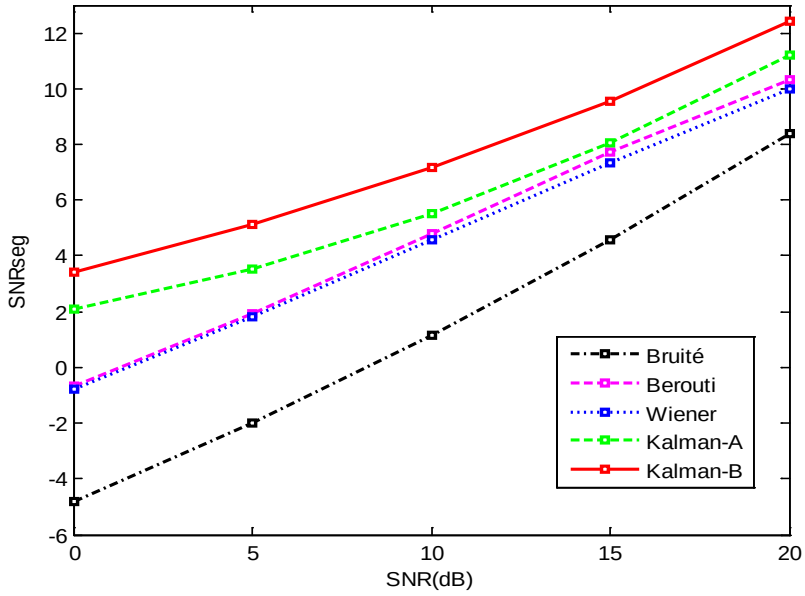

Figure 3.8a : Comparaison des résultats des tests SNR_{seg} avec différentes méthodes de rehaussement dans le cas de la parole bruitée par un bruit factory.

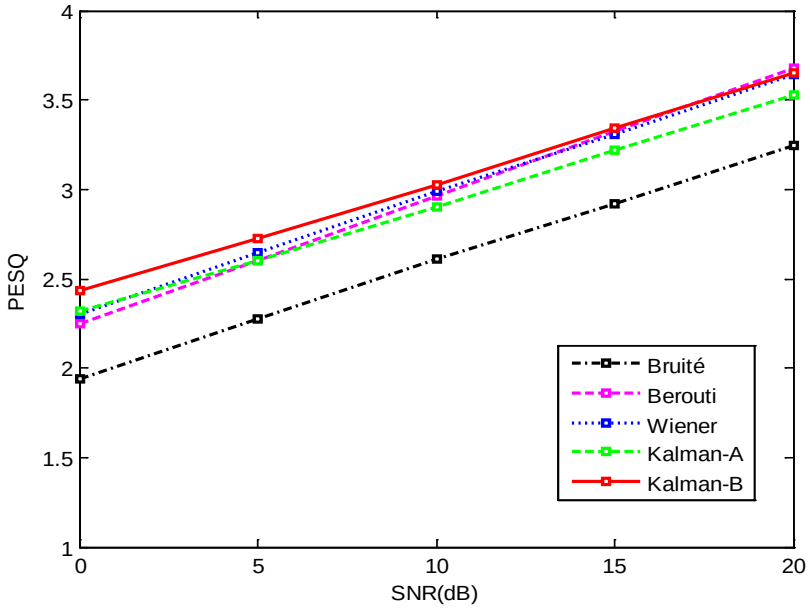

Figure 3.8b : Comparaison entre les résultats des tests PESQ avec différentes méthodes de rehaussement dans le cas de la parole bruitée par un bruit factory.

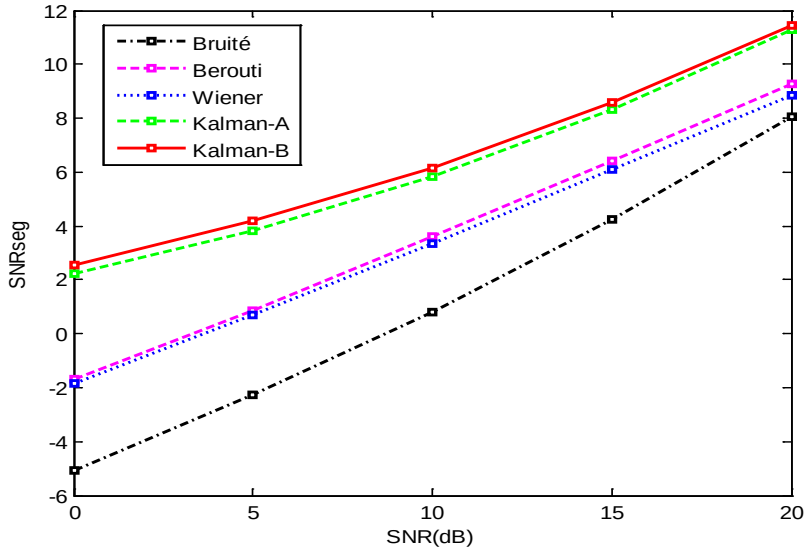

Figure 3.9a : Comparaison des résultats des tests SNR_{seg} avec différentes méthodes de rehaussement dans le cas de la parole bruitée par un bruit buccaneer.

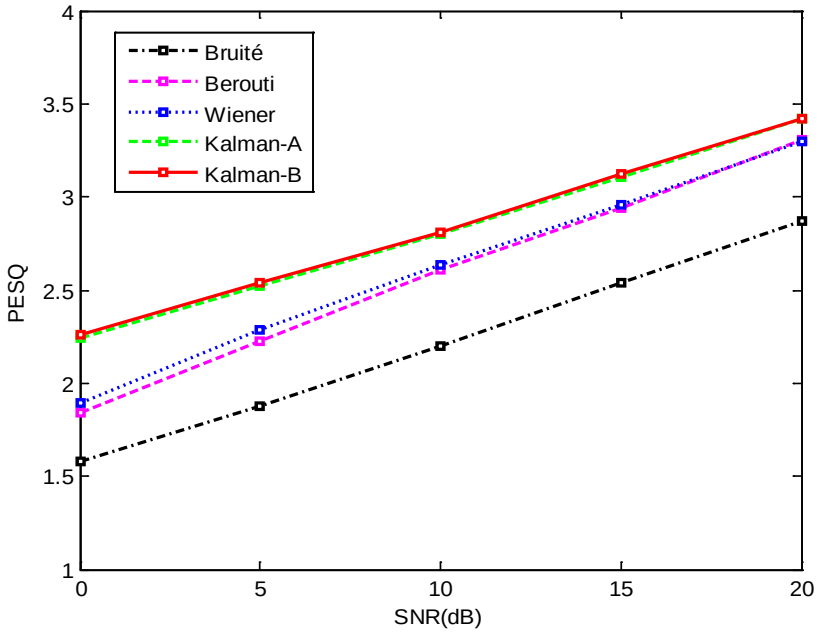

Figure 3.9b : Comparaison entre les résultats des tests PESQ avec différentes méthodes de rehaussement dans le cas de la parole bruitée par un bruit buccaneer.

3.4.3 Interprétation des résultats

- On peut dire que les résultats obtenus sont satisfaisants, une amélioration importante en terme des mesures SNR_{seg} et PESQ est donnée par l'application des différentes méthodes de rehaussement de la parole.

- Nous observons que les résultats sont meilleurs pour les valeurs élevées du rapport signal sur bruit où le niveau du bruit est faible par rapport au cas où le rapport signal sur bruit est faible.
- Par rapport aux autres types de bruit, l'application des différents algorithmes de rehaussement de la parole dans le cas du bruit babble nous donne des résultats faible a cause de sa forme qui ressemble à la forme d'un signal de la parole.
- On peut voir que les deux méthodes, filtre de Kalman simple avec modélisation du bruit et filtre de Kalman sans modélisation du bruit sont alors plus performantes en terme de SNR_{seg} et PESQ par rapport au filtre de Wiener et à la soustraction spectrale de Berouti pour tous types et niveaux du bruit.
- Les résultats des tests SNR_{seg} et PESQ pour le filtre de Kalman simple avec modélisation du bruit sont plus élevés par rapport au filtre de Kalman simple sans modélisation du bruit, ceci est du au fait que les calculs de base du filtre de Kalman sans modélisation du bruit sont basés sur l'hypothèse que le bruit additif est un bruit blanc, donc il prend toujours en considération que le bruit est un bruit blanc.
- En général, pour tous les types et les niveaux de bruit les résultats en terme du SNR_{seg} pour le filtre de Wiener sont presque les mêmes où légèrement supérieures par rapport aux résultats obtenus par la soustraction spectrale de Berouti. En termes de PESQ les résultats obtenus par le filtre de Wiener sont meilleurs par rapport à celles obtenus par la méthode de Berouti.

3.5 Conclusion

D'après les résultats des tests effectués, on peut dire que les méthodes de rehaussement de la parole apportent une amélioration des performances du signal de la parole. Les différentes méthodes de rehaussement de la parole étudiées sont appliquées à des signaux pris de la base de données « Noizeus », des résultats objectives et des formes d'ondes sont représentés, suivis d'une interprétation des résultats.

Tout d'abord on peut affirmer que le filtre de Kalman est plus performant que le filtre de Wiener et la soustraction spectrale de Berouti. Du point de vue de la réduction du bruit, on ne peut que constater son efficacité et sa robustesse, et même du point de vue préservation du signal on remarque qu'il est meilleur.

Chapitre 4

Amélioration de la reconnaissance automatique de la parole (RAP) par les techniques de rehaussement

Chapitre 4 :

Amélioration de la reconnaissance automatique de la parole (RAP) par les techniques de rehaussement

4.1 Introduction

Le domaine de la reconnaissance de la parole est devenu l'un des sujets de recherches les plus intéressants en traitement du signal. Les systèmes de reconnaissance automatique de la parole voient leurs performances diminuer de manière significative lorsque les environnements dans lesquels ils ont été entrainés et ceux dans lesquels ils sont utilisés diffèrent.

Dans ce chapitre, nous présentons le principe général des systèmes de reconnaissance de la parole et ses différentes méthodes et techniques. Le but de ce chapitre est d'étudier l'influence du rehaussement de la parole comme prétraitement sur les performances du système de reconnaissance.

4.2 La reconnaissance automatique de la parole

4.2.1 Historique

Les premiers travaux sur la reconnaissance de la parole remontent aux années 1950, avec l'arrivée des méthodes numériques la capacité des systèmes de reconnaissance est augmentée.

Au milieu des années 1970, une modélisation statistique des différentes prononciations a été proposée. Le principe de base est de remplacer l'ensemble des références acoustiques représentant les différentes prononciations d'un même mot, par un modèle de ces prononciations. Donc, chaque mot du vocabulaire est représenté par un modèle qui peut faire aux modèles de Markov cachés. Avec l'arrivée des années 1980, une introduction des réseaux de neurones artificiels dans la reconnaissance de la parole a été lancée. Pendant les années 1990, une combinaison est faite entre les méthodes Markovienne et les méthodes connexionniste, ce qui nous donne les systèmes hybrides qui ont pour but d'éviter les inconvénients des différentes approches et de profiter de ses avantages.

4.2.2 Principe de la reconnaissance de la parole

La reconnaissance automatique de la parole (RAP) consiste à transcrire un signal vocal en informations numériques, compréhensibles et reconnaissables par l'ordinateur. Le principe de la reconnaissance automatique de la parole peut être résumé par la figure (4.1) :

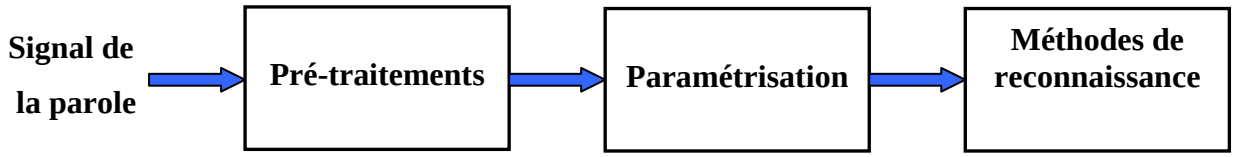

Figure 4.1 : Schéma bloc d'un système de reconnaissance de la parole.

a. Le signal de la parole est numérisé puis modélisé.
b. Extraction des paramètres du signal acoustique.
c. Finalement, le module de reconnaissance utilise ces paramètres pour reconnaître le message véhiculé par le signal vocal.

4.2.3 Les méthodes de la reconnaissance de la parole

Le but de la parole est la communication en langage naturel avec une machine. Dans le cadre de la reconnaissance automatique de la parole il existe deux méthodes :

4.2.3.1 Les méthodes analytiques

Les approches analytiques consistent à décomposer le signal de la parole en unités phonétiques courtes et discrètes, comme les phonèmes, c'est l'étape de décodage acoustico-phonétique, ensuite un système linguistique est utilisé pour une interprétation linguistique.

Ces méthodes (Figure 4.2) se basent sur l'utilisation des connaissances linguistiques pour augmenter les performances de la reconnaissance.

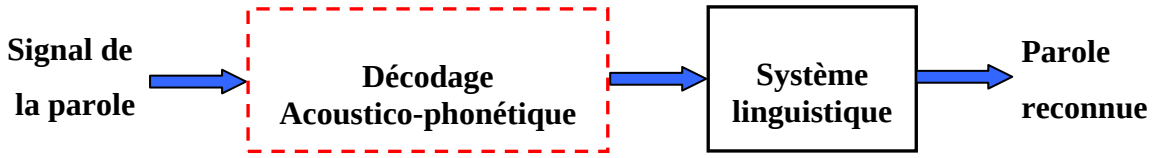

Figure 4.2 : Schéma synoptique d'un système de reconnaissance de la parole utilisant l'approche analytique.

4.2.3.2 Les méthodes globales

Dans ce cas, les mots ou les phrases sont considérés comme des entités élémentaires qu'on compare à des références enregistrées sans décomposition préalable. L'idée de base des méthodes globales (Figure 4.3) est de donnée au système au moins une image acoustique de chacune des unités qu'il devra identifier par la suite.

On distingue deux grandes catégories des méthodes globales de reconnaissance de la parole :

- Les approches basées sur la reconnaissance par comparaison à des exemples.
- Les approches probabilistes.

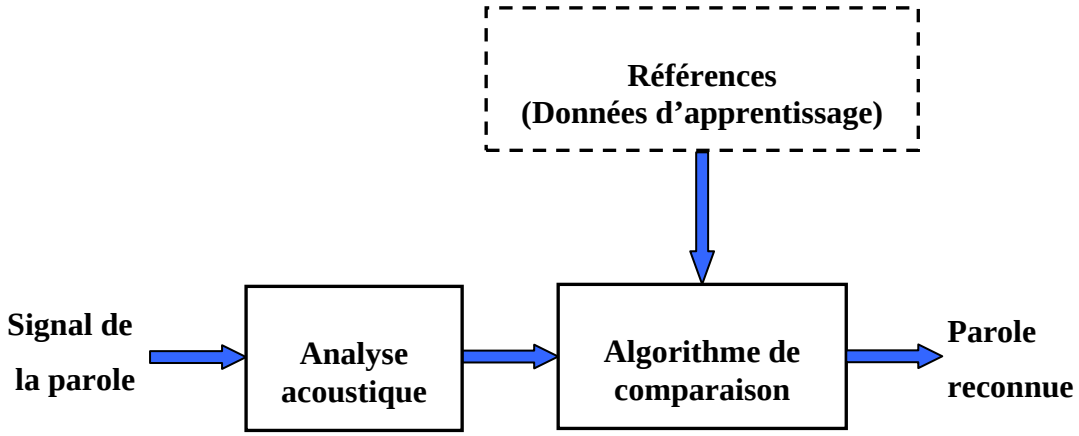

Figure 4.3 : Schéma synoptique d'un système de reconnaissance de la parole utilisant l'approche globale.

- **Reconnaissance par comparaison à des exemples**

L'idée de base consiste à faire prononcer un ou plusieurs exemples de mots susceptibles d'être reconnus. Ensuite ces exemples sont convertis sous forme de séquences de paramètres acoustiques, c'est la phase d'entraînement du système.

Dans la phase de test, le signal de la parole est encodé sous forme de séquence de vecteur acoustique. La phase de reconnaissance consiste à faire une comparaison entre la séquence obtenue et toutes les séquences obtenues lors de la phase d'entraînement, donc le mot qui correspond à la séquence d'entraînement se rapprochant le plus de la séquence de test sera le mot reconnu.

Pour faire la comparaison entre les séquences on utilise l'algorithme d'alignement temporel dynamique 'DTW' (Dynamique Time Warping) [56]. L'algorithme DTW est basé sur la définition d'un indice de dissemblance entre deux mots en mettant en correspondance optimale les échelles temporelles des deux mots. Cette technique est applicable sous la condition que la taille du vocabulaire soit faible, ce qui implique :

- Le nombre maximal de mots à reconnaître soit faible à cause du vocabulaire qui est limité.
- L'application est réservée pour un seul locuteur qui aura lui-même entraîné le système.

Le problème majeurr de cette technique est qu'elle nécessite un espace mémoire et un temps de calcul très importants.

- **Les approches probabilistes**

La plupart des systèmes actuels de la RAP sont fondés sur une approche probabiliste. Cette approche a pour but de construire un message (M) inconnu à partir d'une séquence d'observation (O). Donc, il s'agit de trouver la séquence de mot (M') qui correspond au message le plus probable connaissant la suite des observations acoustiques (O) :

$$M' = \arg\max_{M} P(M/O) \tag{4.1}$$

L'utilisation de la règle de Bayes permet de décomposer la probabilité a posteriori $P(M/O)$ en deux composantes :

$$P(M/O) = \frac{P(M)P(O/M)}{P(O)} \tag{4.2}$$

Avec :

$P(M)$: Probabilité a priori d'observer la séquence M indépendamment du signal, déterminée par le modèle langage.

$P(O/M)$: Probabilité d'observer la séquence des vecteurs acoustiques O sachant la séquence de mots spécifique M . Cette probabilité est déterminée lors de l'étape de reconnaissance des unités acoustiques.

Puisque $P(O)$ ne dépendant pas des paramètres du modèle, l'équation (4.1) sera donc :

$$M^{'} = \arg\max_{M} P(M)P(O/M) \tag{4.3}$$

On peut dire que la qualité d'un tel système de reconnaissance de la parole peut être caractérisée par la robustesse des modèles qui permettent de calculer les deux termes $P(M)$ et $P(O/M)$. Donc, il s'agit d'une intégration des niveaux acoustiques et linguistiques dans un seul processus de décision permettant de trouver le message prononcé.

La RAP peut réalisée par les réseaux de neurones artificiels (ANN : Artificial Neural Network), des modèles de Markov cachés (HMM : Hidden Markov Model) ou des modèles hybrides.

a. Reconnaissance de la parole par Modèles de Markov caches (HMM : Hidden Markov Model)

En reconnaissance de la parole des modèles de Markov cachés sont apparus dans les années 70. Ces modèles sont des automates probabilistes à états finis qui permettent de calculer la probabilité d'émettre une séquence d'observation.

Un HMM peut être représenté comme un ensemble discret de nœuds (ou états) reliés entre eux par des arcs de transition, les transitions d'un état à un autre sont régies par des probabilités définies lors de l'apprentissage des modèle. La figure (4.4) présente un modèle de Markov caché gauche-droit ou de Bakis :

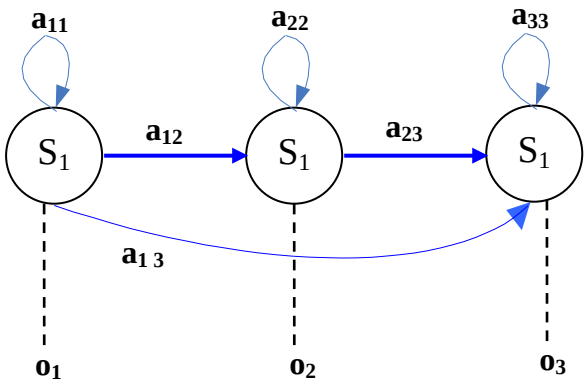

Figure 4.4 : Exemple de HMM à 3 états gauche-droit.

Les HMM peuvent modéliser toutes unités acoustique (mots, phonèmes,…etc.) selon l'application. Un modèle de Markov caché est défini par l'ensemble des paramètres $\Phi = (A, B, \pi)$ avec :

- $S = \{S_1, S_2, ..., S_N\}$: L'ensemble des états du modèle avec N le nombre d'état et q_t l'état à l'instant 't'.
- $A = (a_{ij})$: La matrice des probabilités de transition et a_{ij} la probabilité de passer de l'état i à l'état j .
- $B = (b_i(o_t))$: La matrice des probabilités d'émission des observations dans chaque état et $b_i(o_t)$ est la probabilité d'émission de l'observation o_t dans l'état S_i.
- $\pi = (\pi_i)$: La matrice de distribution de l'état initial et π_i est la probabilité d'être dans l'état S_i à l'instant initiale.

$o = (o_1, o_2, ..., o_M)$: Une suite des observations avec M le nombre fini de symboles d'observation par états.

La reconnaissance de la séquence d'observation o est effectuée en trouvant l'ensemble des paramètres $\Phi = (A, B, \pi)$ qui maximise la probabilité qu'un modèle Φ génère une séquence d'observation 'o'. L'utilisation d'un modèle de Markov caché dans la reconnaissance de la parole se base sur la résolution des trois problèmes suivants :

- **Evaluation :** Comment calculer $P(o/\Phi)$? la probabilité que la séquence d'observation $o=(o_1,o_2,...,o_M)$ ait été émise par le modèle $\Phi=(A,B,\pi)$. Ce problème est résolu par l'algorithme Forward.

- **Décodage :** Comment déterminer la séquence d'état $S=\{S_1,S_2,...,S_N\}$ qui est la plus probable pour générer la séquence d'observation $o=(o_1,o_2,...,o_M)$. Dans cette étape l'algorithme de Viterbi est utilisé comme solution.

- **Apprentissage :** Comment déterminer les paramètres du modèle $\Phi=(A,B,\pi)$ pour maximiser la probabilité $P(o/\Phi)$. Pour résoudre le problème d'apprentissage on peut utiliser l'algorithme de Bam-Welch (Forward_Backward).

Les systèmes de RAP à base des HMM reposent ainsi sur les postulats suivants :

1-La parole est une suite d'états stationnaires, représentés par des vecteurs caractéristiques et leurs dérivées premières et secondes.

2-L'émission d'une séquence de ces vecteurs est générée par un HMM respectant l'hypothèse Markovienne d'ordre 1.

Reconnaissance de la parole par HTK

Dans ce travail, l'un des systèmes de reconnaissance que nous avons implémentés repose sur la plateforme HTK. La plate forme HTK (Hidden Markov Model Toolkit) ou, 'boite à outils de modèle de Markov caché' a été développée à l'Université de Cambridge par S.J.Young [57]. La boite à outils HTK permet de construire la chaîne complète : Apprentissage et reconnaissance. Ces modèles peuvent représenter des mots ou tout type d'unité sub-lexicale (phonème, triphone). L'ensemble des outils écrit en langage 'C'. En annexe A, on représente les étapes de conception de la reconnaissance de la parole par HTK.

b. Reconnaissance de la parole par les modèles connexionnistes

Un réseau de neurone est composé de plusieurs couches de neurones, l'assemblement des neurones augmente la capacité d'apprentissage. L'organisation des neurones entre eux au sien d'un même réseau dépend du problème à résoudre. Le neurone formel (Figure 4.5) est une unité élémentaire qui concentre plusieurs signaux d'entrée en une seule sortie.

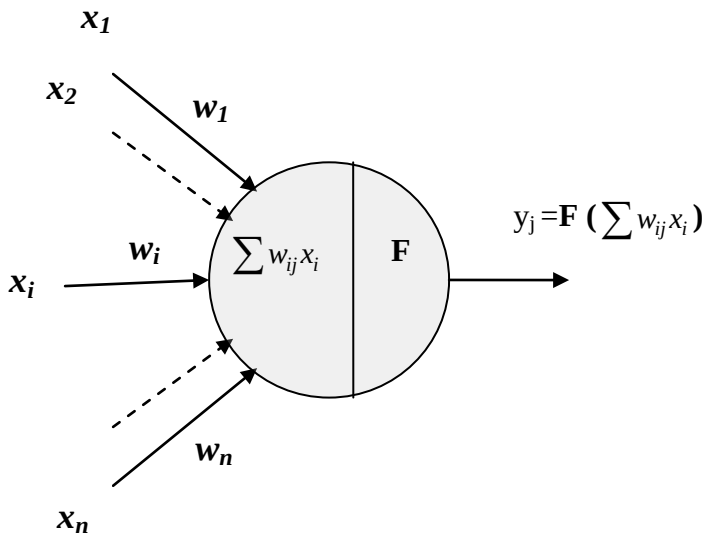

Figure 4.5 : Un modèle du neurone formel.

Avec :

$X = [1, x_1, x_2, ..., x_n]$ **:** Vecteur d'entrée de la cellule.

w_{ij} **:** Les poids synaptique.

$F(.)$ **:** La fonction d'activation.

y_j **:** La sortie de la cellule.

Les modèles les plus utilisés des réseaux de neurones artificiels (ANN : Artificial Neural Network) sont :

1) Le perceptron multicouches (MLP : Multi Layer Perceptron) [58]

Les perceptrons multicouches (Multi Layers Perceptron : MLP) appartiennent au réseau de neurones multicouche, ils ne possèdent pas de boucle de retour c'est-à-dire uniquement des réseaux à propagation directe. Ces réseaux (Figure 4.6) sont constitués d'une couche d'entrée, une couche de sortie et une ou plusieurs couches intermédiaires dites couches cachées.

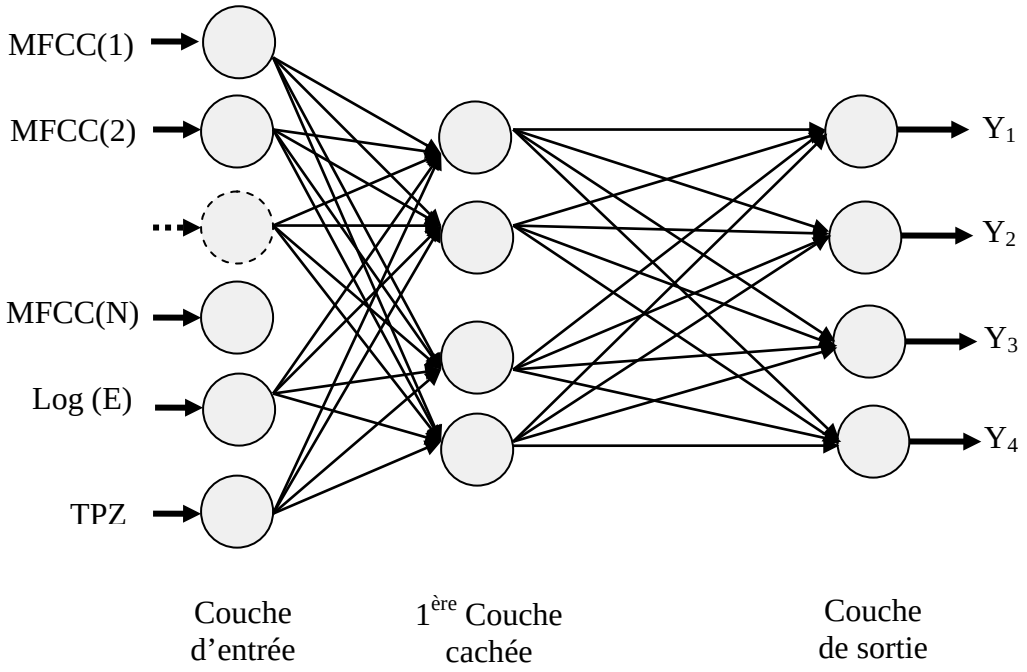

Figure 4.6 : Architecture de perceptron multicouche (MLP).

Pour la reconnaissance de la parole par un réseau MLP, la dimension temporelle du signal de la parole est le problème majeur, pour pouvoir utiliser ce réseau, une modélisation des entrées est nécessaire. Cette opération est effectuée pour que tous les fichiers aient le même nombre d'échantillons, donc pour qu'on puisse extraire le même nombre de paramètre.

2) Le réseau de neurones à délai (TDNN : Time Delay Neural Network) [61]

Le réseau de neurones à délais (TDNN) est un réseau dynamique qui utilise la structure de base du MLP. Il prend en compte le contexte temporel de l'unité élémentaire basse vers l'unité élémentaire haute en intégrant des retards, le nombre de retard de chaque couche est déterminé par l'application recherchée. La figure (4.7) illustre la structure d'un réseau TDNN.

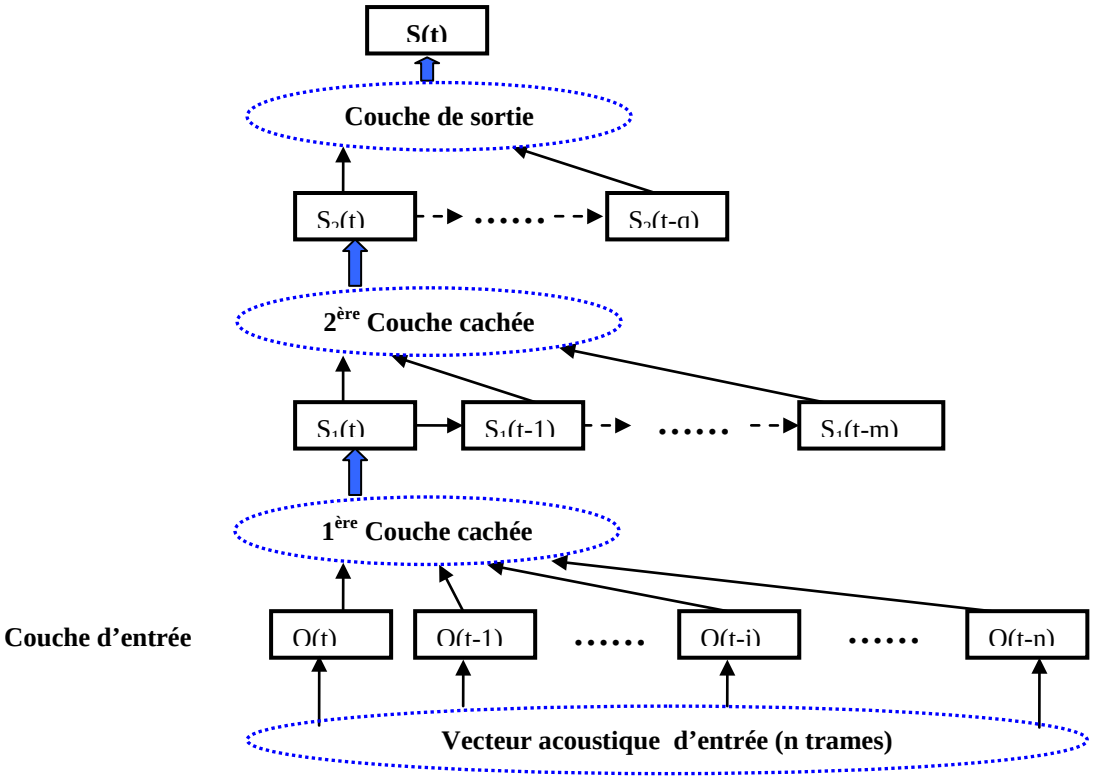

Figure 4.7: Architecture d'un réseau de neurones à temps de retard (TDNN).

3) Le réseau de neurones récurrent (RNN : Recurrent Neural Networks) [64]

Les réseaux de neurones récurrents (Figure 4.8) sont caractérisés par la présence d'au moins une boucle de rétroaction au niveau des neurones ou entre les couches, dans ce cas on peut dire que l'aspect temporel du phénomène est pris en compte. L'utilisation de ce réseau nous donne presque les mêmes résultats obtenus par MLP.

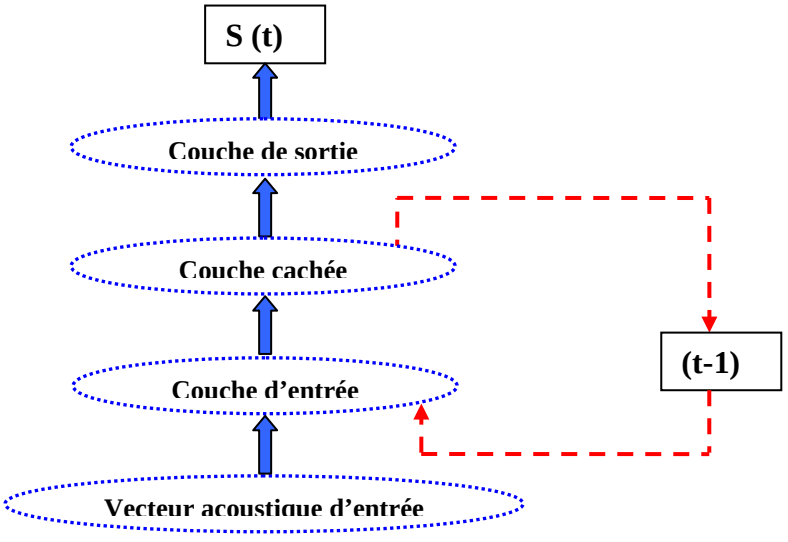

Figure 4.8: Architecture d'un réseau de neurones récurrent (RNN) : le modèle d'Elman.

4) Le réseau de neurone à fonction de base radiale (RBF NN : Radial Basis Neural Network)

Les réseaux RBF (Figure 4.9) sont des réseaux à une couche cachée, ils possèdent la même architecture que les MLP, mais la fonction non linéaire des neurones de la couche cachée est une Gaussienne. Les neurones de la couche de sortie n'ont pas d'activation (somme pondéré uniquement). Les résultats obtenus avec le réseau RBF sont inférieurs par rapport à ceux de la méthode MLP.

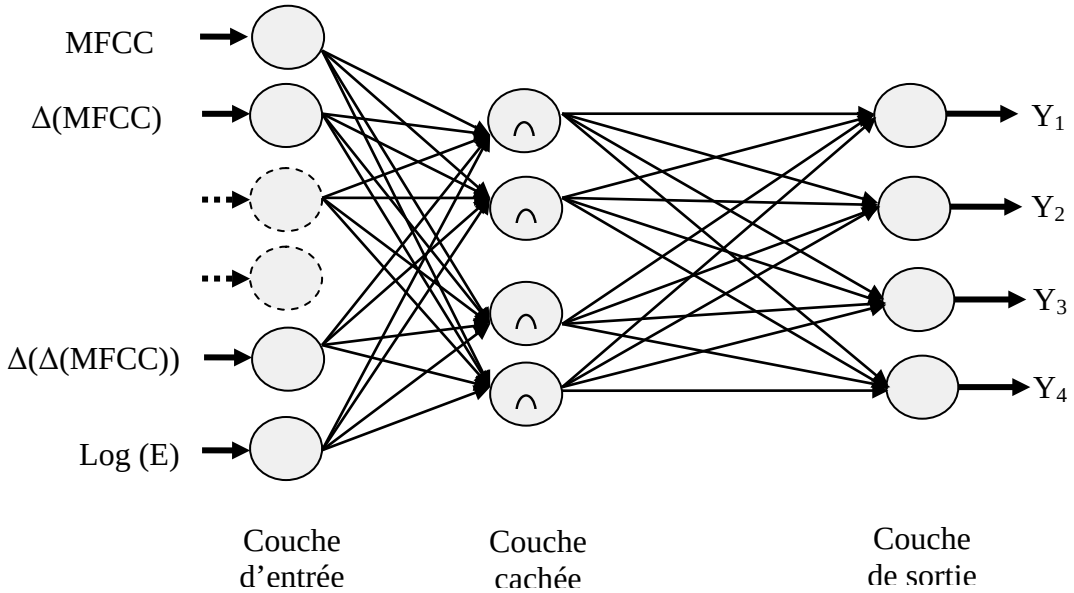

Figure 4.9: Architecture du réseau de neurone à fonction de base radiale (RBF).

c. Reconnaissance de la parole par les modèles hybrides

Une combinaison entre les ANN et les HMM nous donne les modèles hybrides. Les modèles de neurones ne peuvent pas modéliser l'évolution temporelle d'un message de la parole en raison de leur relative inadéquation à traiter des signaux séquentiels. Donc, un modèle hybride nous donne une solution adéquate pour palier les inconvénients des modèles ANN et HMM.

4.2.4 Les techniques de reconnaissance de la parole

Les systèmes de reconnaissance de la parole peuvent être classé selon :

4.2.4.1 Reconnaissance de mots isolés

Dans ce cas, on commence par faire effectuer une analyse acoustique du signal de manière à extraire des vecteurs acoustiques caractérisant ce signal, puis les comparer à un dictionnaire pour extraire le mot le plus proche.

4.2.4.2 Reconnaissance de la parole continue

Ce type de reconnaissance entraîne les problèmes plus complexes avec la nécessité de trouver la frontière entre les mots. Pour résoudre ce problème, on peut envisager plusieurs solutions :

- Définition des règles de segmentation pour un vocabulaire donné en se basant sur quelques critères acoustiques.
- Ne pas faire une segmentation préalable, et rechercher à identifier l'image acoustique de toute la phrase.

4.2.4.3 Taille du vocabulaire

On parle de vocabulaire limité si le nombre de mots à reconnaître est inférieur à 100, c'est le cas pour la reconnaissance de mots isolé. Afin d'obtenir plus de confort, le vocabulaire pourra être étendu pour des applications plus larges.

4.3 Influence du milieu réel sur la reconnaissance vocale

La robustesse des systèmes de reconnaissance vocale est liée à leur résistance aux différentes formes de pollution sonore (bruit de toute nature). La plupart des systèmes de RAP fonctionnent mal en milieu bruité, car les contraintes posées par de tels environnements n'ont pas été prises en compte dès le départ. L'enregistrement de la parole était fait dans des conditions optimales, pour cela les premiers systèmes de la RAP ne pouvaient plus fonctionner convenablement lorsque la parole est confondue à un bruit de fond, ou modifiée lors de sa transmission. Le bruit fait partie intégrante des environnements réels et pour atténuer l'effet du bruit dans un système de RAP il faut prendre en compte l'influence du bruit sur les deux aspects suivants :

- **Sur le signal :** Le bruit change la moyenne et l'écart type des cepstres.
- **Sur le locuteur :** Le bruit peut introduire une :
  - Modification de la façon de parler (effet de Lombard). Lorsqu'un locuteur est placé dans un environnement bruité, il modifie sa voix et son effort vocal, de manière à ce que la parole produite conserve un bon SNR. Cela pose un problème aux systèmes de RAP car les spectres de tous les phonèmes peuvent être modifiés, ce qui nous donne des faible taux de reconnaissance.
  - Augmentation de pitch.
  - Raccourcissement des consones et allongement des voyelles.

Aujourd'hui, la recherche se poursuit notamment vers la reconnaissance de la parole pour de grands vocabulaires dans des milieux bruités. Pour augmenter la robustesse des systèmes de reconnaissance de la parole en milieu bruité, ils existent trois différentes approches :

1) Approche agit sur le signal

Dans ce cas, le signal de la parole est rehaussé par différentes techniques de rehaussement, ensuit on utilise ce signal rehaussé dans la phase de reconnaissance.

2) Approche agit sur les modèles

Cette approche se base sur l'adaptation des modèles acoustiques d'apprentissage pour qu'ils soient plus proches des modèles acoustiques de test. Cela est effectué par le bruitage de ces modèles. Donc, on utilise directement le signal bruité pour la reconnaissance.

3) Approche agit sur la paramétrisation

Le but de cette approche est de trouver des paramètres adéquats qui ne soient pas influencés par l'environnement acoustique. La paramétrisation robuste ne transforme pas le signal de la parole et utilise des paramètres résistant au bruit.

Il faut noter que dans notre travail on s'intéresse à la première approche.

4.4 Protocole expérimental

4.4.1 Méthodologie

La représentation la plus communément utilisée pour évaluer les performances des systèmes RAP est le taux de reconnaissance (TR) défini par :

$$TR = \frac{R}{N} \times 100\% \tag{4.4}$$

Avec :

N : Le nombre total de mots dans la base de tests.

R : Le nombre de mots correctement reconnus.

Dans le cadre de notre travail, la procédure expérimentale consiste à calculer les performances des différents systèmes de RAP en utilisant une base de test bruitée par

différents types et niveaux de bruit ensuite. Cette base est rehaussée par les différentes techniques de rehaussement de la parole que nous avons élaborées et testées dans le chapitre précédent. La procédure expérimentale est représentée par la figure (4.10) :

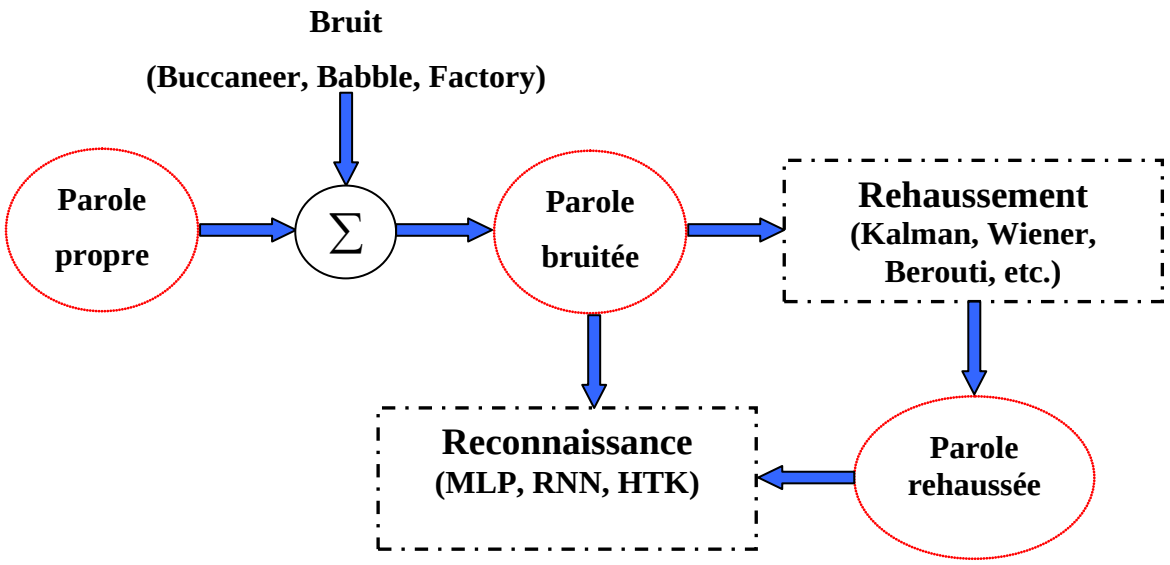

Figure 4.10 : Procédure expérimentale utilisée.

4.4.2 Base de données utilisée

Nous évaluons les performances des systèmes RAP en utilisant la base de données arabe ARADIGIT. Cette base est conçue pour l'évaluation des algorithmes de reconnaissance de la parole, elle contient des mots et chiffres arabes parlés prononcés par 110 locuteurs et locutrices âgés entre 18 et 50 ans de niveau universitaire. Les sons enregistrés à 22050 Hz ont été ensuite sous échantillonnés à 16000 Hz.

Dans la phase d'évaluation des performances, nous avons utilisé seulement la partie chiffre dans le mode broadcast (ARADIGIT) de la base.

Pour la phase d'apprentissage, nous avons utilisé 60 locuteurs des deux sexes prononçant 1800 chiffres.

Pour la phase de test, nous avons utilisé 8 locuteurs masculins prononçant 80 chiffres. La base de données test est bruitée par des bruits de la base de données NOISEX-92 [66], [67] à des niveaux du bruit varié de 0 jusqu'à 20 dB.

Un système de reconnaissance de la parole nécessite une paramétrisation du signal de la parole. Le processus d'extraction des paramètres est appliqué pour chaque trame du signal de la parole. Chaque trame est représentée par un vecteur acoustique de 36 coefficients.

Douze (12) coefficients MFCC sont calculés à partir d'un banc de 24 filtres répartis dans l'échelle fréquentielle Mel. Pour chaque coefficient, on attribue une dérivée première (12 dérivées premières au total) ainsi qu'une dérivée seconde (12 dérivées secondes) pour prendre en compte la dynamique du signal.

4.5 Résultats expérimentaux

Dans l'évaluation des performances des systèmes de reconnaissance, nous considérons que la base de test est bruitée additivement par trois types de bruit (buccaneer, babble, factory) de la base bruitée NOISEX-92 à différents rapports signal sur bruit. Cette base est par la suite rehaussée par différentes méthodes de rehaussement de la parole que nous avons implémentées.

Nous calculons les taux de reconnaissance avec la base bruité puis avec celle rehaussée en considérant les trois systèmes de reconnaissance : MLP, RNN, HTK. Dans l'étape de rehaussement on utilise le filtre de Kalman et le filtre de Wiener.

Pour le système de RAP basé sur la méthode MLP, on utilise un réseau de neurones à couche cachée composé de 300 neurones avec un seuil de l'erreur fixé à $\varepsilon = 10^{-4}$.

Dans le cas d'un système de RAP basés sur la méthode RNN, on utilise un réseau de neurones de type Elman avec rétroaction des sorties de la couche cachée vers l'entrée. La couche cachée est composée de 80 neurones.

Les résultats obtenus sont représentés dans les figures suivantes :

-Les figures (4.3), (4.4) et (4.5) représentent respectivement les taux de reconnaissance par MLP des signaux bruités par les bruits buccaneer, babble et factory et rehaussés par le filtre de Kalman.

-Les figures (4.6), (4.7) et (4.8) représentent respectivement les taux de reconnaissance par RNN des signaux bruités par les bruits buccaneer, babble et factory et rehaussés par le filtre de Kalman.

-Les figures (4.9), (4.10) et (4.11) représentent respectivement les taux de reconnaissance par HTK des signaux bruités par les bruits buccaneer, babble et factory et rehaussés par le filtre de Kalman.

-Les figures (4.12), (4.13) et (4.14) représentent respectivement une comparaison entre les taux de reconnaissance par MLP des signaux bruités par les bruits buccaneer, babble et factory et rehaussés par le filtre de Kalman et le filtre de Wiener.

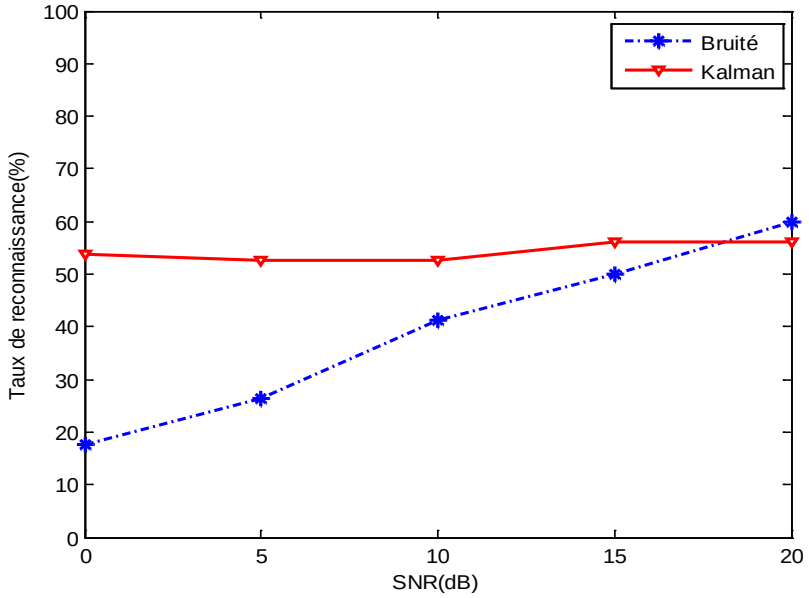

Figure 4.11 : Taux de reconnaissance par MLP des signaux bruités par un bruit buccaneer et rehaussé par filtre de Kalman.

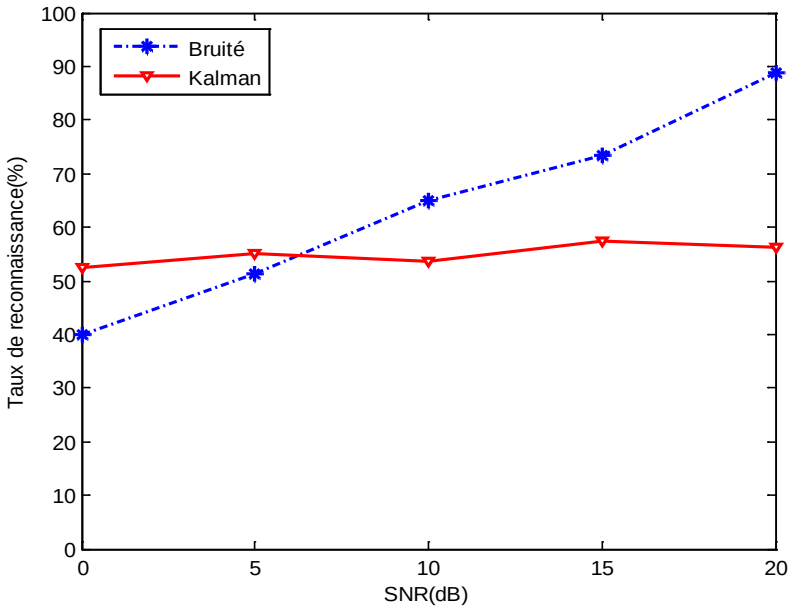

Figure 4.12 : Taux de reconnaissance par MLP des signaux bruités par un bruit babble et rehaussé par filtre de Kalman.

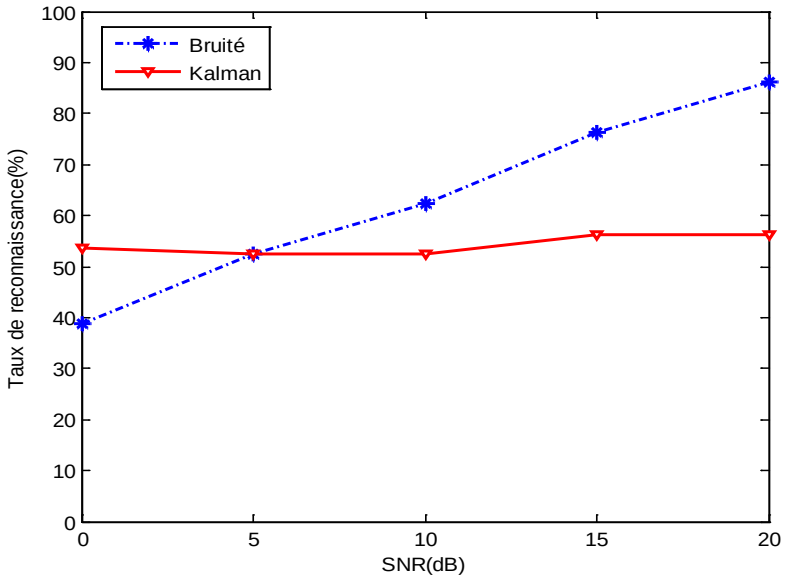

Figure 4.13 : Taux de reconnaissance par MLP des signaux bruités par un bruit factory et rehaussé par filtre de Kalman.

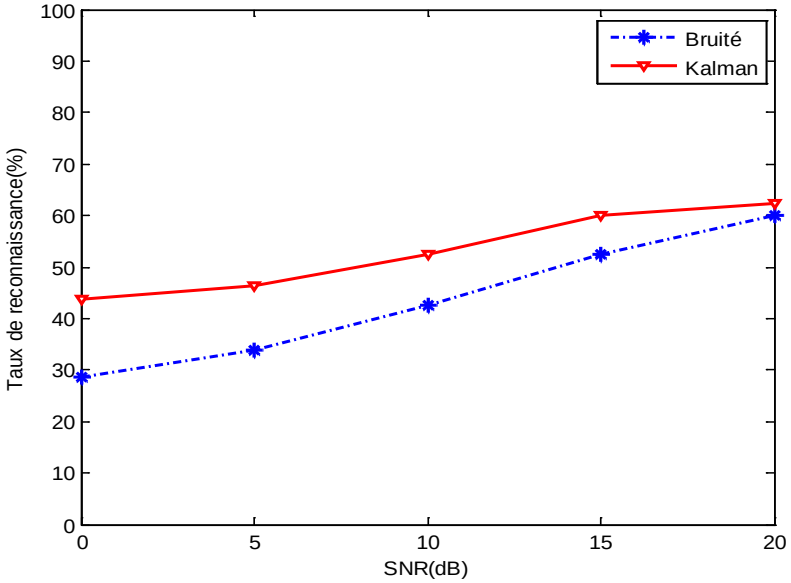

Figure 4.14 : Taux de reconnaissance par RNN des signaux bruités par un bruit buccaneer et rehaussé par filtre de Kalman.

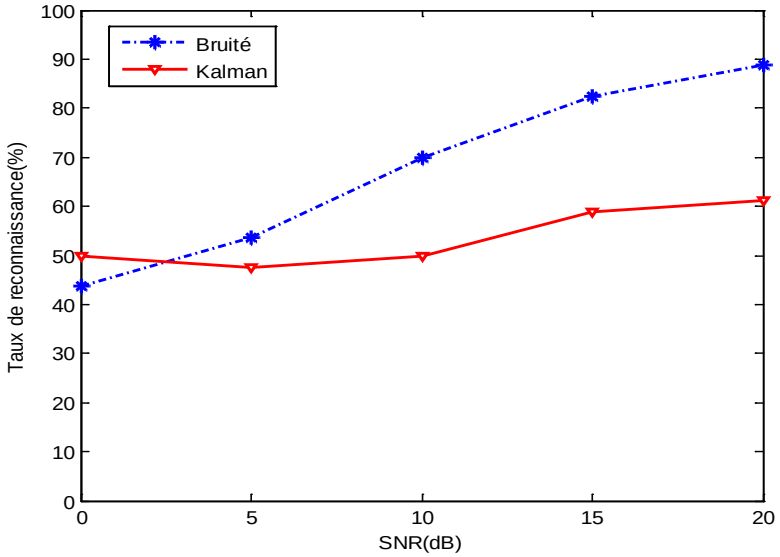

Figure 4.15 : Taux de reconnaissance par RNN des signaux bruités par un bruit babble et rehaussé par filtre de Kalman.

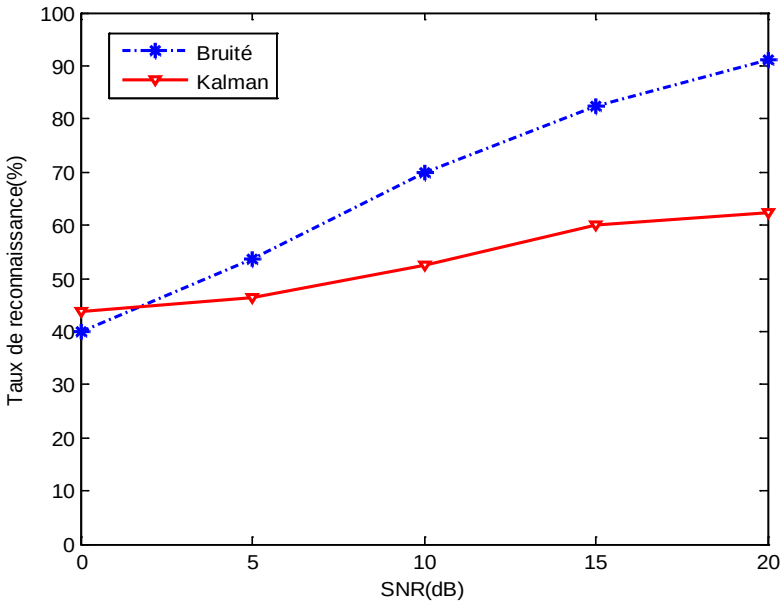

Figure 4.16 : Taux de reconnaissance par RNN des signaux bruités par un bruit factory et rehaussé par filtre de Kalman.

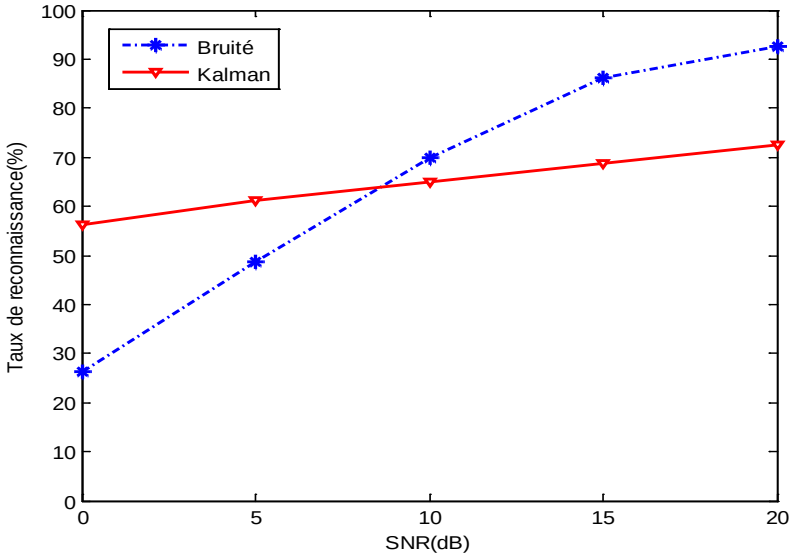

Figure 4.17 : Taux de reconnaissance par HTK des signaux bruités par un bruit buccaneer et rehaussé par filtre de Kalman.

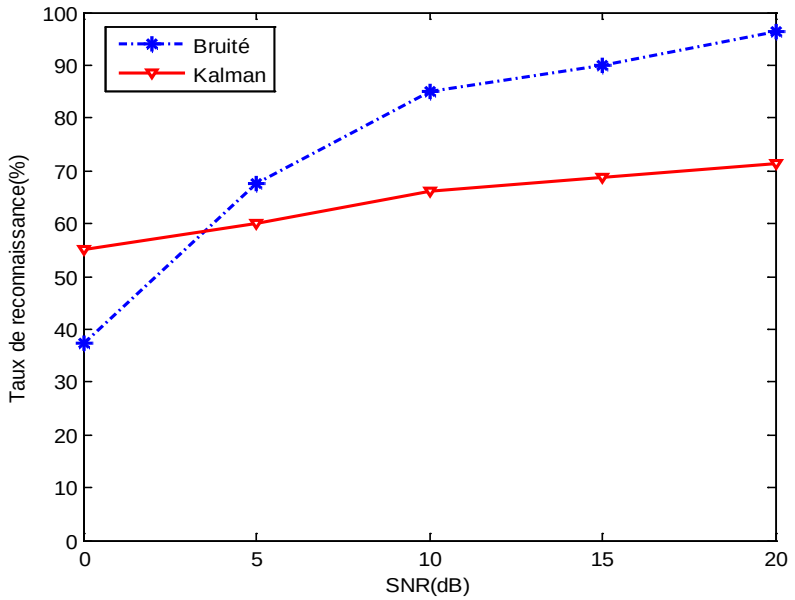

Figure 4.18 : Taux de reconnaissance par HTK des signaux bruités par un bruit babble et rehaussé par filtre de Kalman.

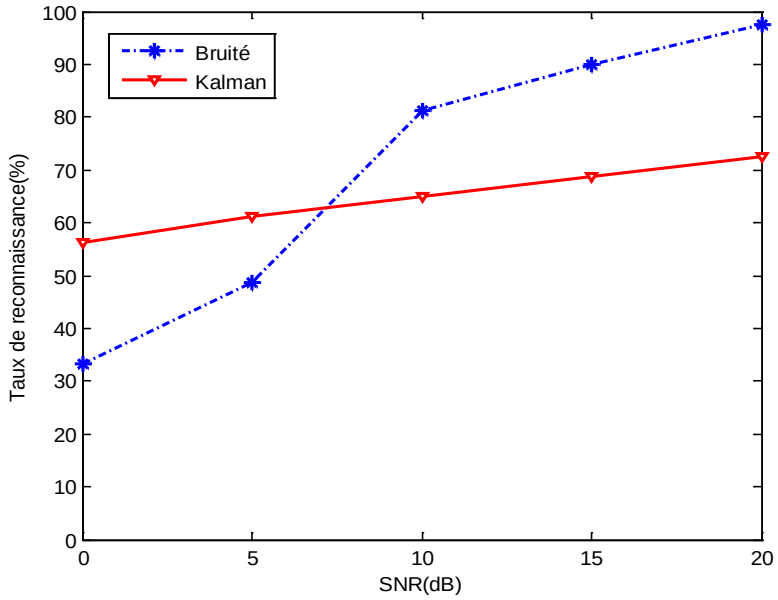

Figure 4.19 : Taux de reconnaissance par HTK des signaux bruités par un bruit factory et rehaussé par filtre de Kalman.

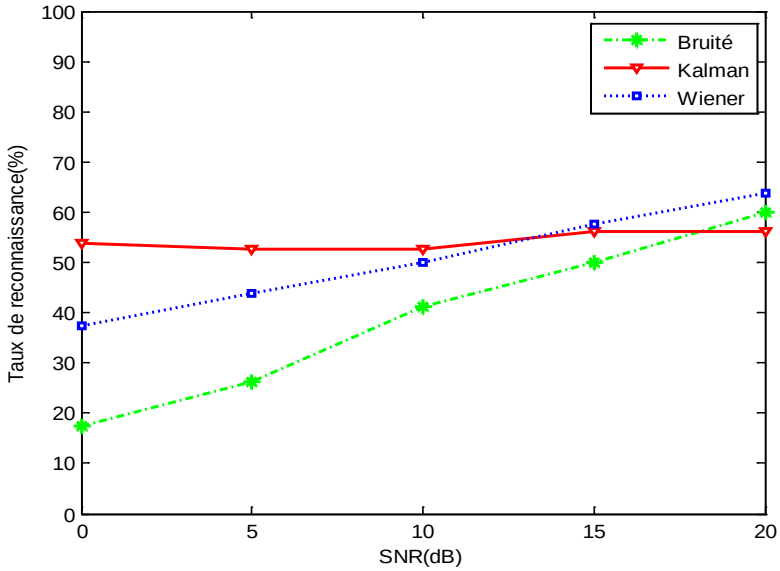

Figure 4.20 : Comparaison entre les taux de reconnaissance par MLP des signaux bruités par les bruits buccaneer et rehaussés par le filtre de Kalman et le filtre de Wiener.

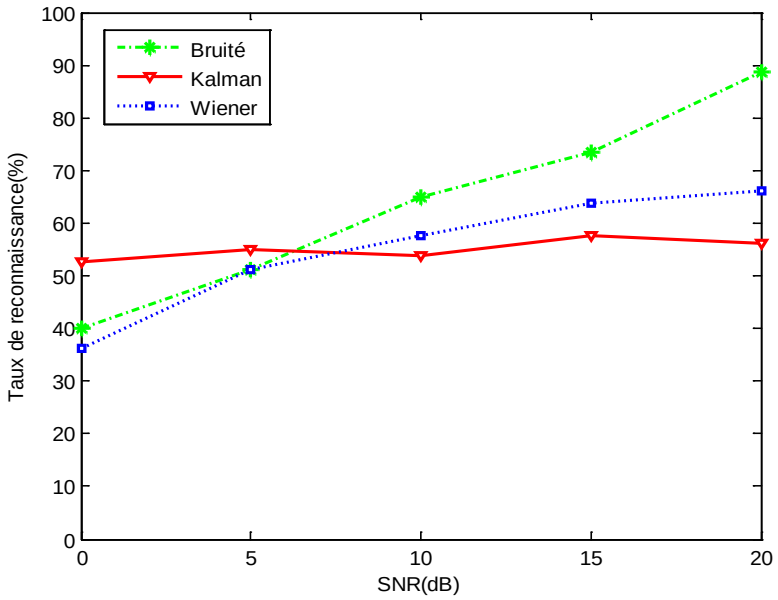

Figure 4.21 : Comparaison entre les taux de reconnaissance par MLP des signaux bruités par les bruits babble et rehaussés par le filtre de Kalman et le filtre de Wiener.

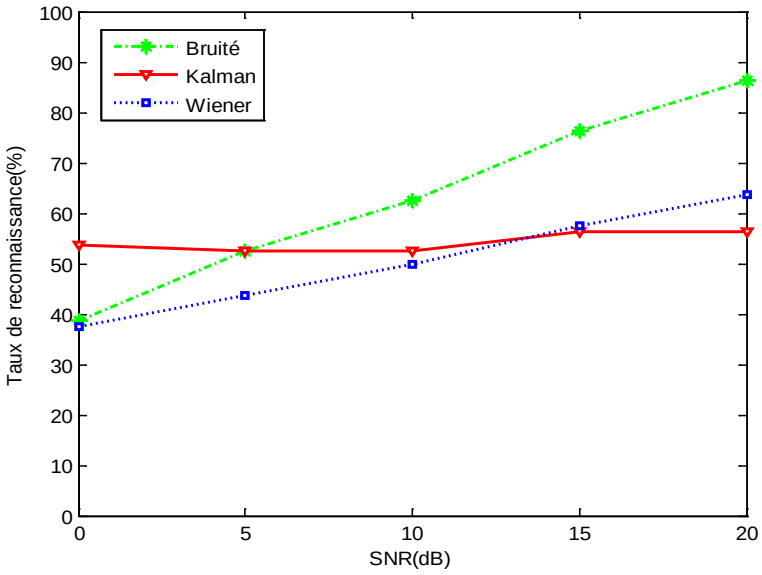

Figure 4.22 : Comparaison entre les taux de reconnaissance par MLP des signaux bruités par les bruits factory et rehaussés par le filtre de Kalman et le filtre de Wiener.

4.6 Interprétation des résultats

- D'après les résultats obtenus on peut dire que le filtre de Kalman peut porter une amélioration de taux de reconnaissance pour les rapports signal sur bruit faibles avec les différents systèmes de reconnaissance et pour différents types de bruit.

- Une exception pour le bruit buccaneer avec les deux méthodes de RAP, MLP et RNN on remarque que l'amélioration et aussi pour les rapports signal sur bruit élevés.

- Une comparaison entre les résultats obtenus par l'utilisation du filtre de Kalman et le filtre de Wiener, nous permet de dire que la méthode la plus performante dans la phase de rehaussement peut nous donne des meilleurs résultats dans la phase de reconnaissance par rapport aux autres méthodes de rehaussement.

- A partir des résultats obtenus, on remarque que l'adaptation des systèmes de rehaussement en fonction de leurs objectifs final est nécessaire, ceci est expliqué par le fait que les conditions d'application d'un système de rehaussement dans une application audio ne sont pas celles dans une application de reconnaissance vocal.

- Par ailleurs, les expériences simulant le fonctionnement en environnement bruité montrent que les classificateurs sont plus sensibles aux bruits à large bande spectrale, notamment pour les niveaux de RSB en dessous de 15 dB.

4.7 Conclusion

Ce chapitre est dédié à l'étude des différentes techniques de la reconnaissance vocale et l'impact de rehaussement de la parole sur les performances de ces systèmes dans un milieu bruité par différents types et niveaux du bruit.

D'après les résultats obtenus on peut dire que dans la plupart des cas les techniques de rehaussement apportent une amélioration aux performances des systèmes de RAP pour des rapports signal sur bruit faibles.

Donc, on note qu'il est nécessaire d'améliorer les techniques de rehaussement de la parole pour les adapter à l'application de reconnaissance vocale et voir une amélioration des performances pour tous types et niveaux du bruit.

Conclusion générale

Conclusion générale

Le rehaussement de la parole qui consiste à améliorer la qualité et l'intelligibilité de la parole dégradée par les bruits ambiants est un chalenge intéressant pour le développement des systèmes de communications. Dans ce mémoire, nous avons implémenté des techniques de rehaussement paramétriques basées sur le filtrage de Kalman avec et sans modélisation du bruit, ainsi que des techniques non paramétriques telles que la méthode de soustraction spectrale de Bérouti et celle fondée sur le filtrage de Wiener. Nous nous sommes restreint au rehaussement du signal en situation mono-voie, car c'est le contexte le plus courant en traitement du signal quel que soit le champ d'application concerné, mais surtout le plus réaliste.

Nous avons effectué une évaluation objective de la qualité des différents algorithmes de rehaussement de la parole implémentés. Les différentes méthodes de rehaussement de la parole étudiées ont été appliquées à des signaux pris de la base de données « Noizeus ». D'après les résultats des tests effectués, on peut dire que les méthodes de rehaussement de la parole apportent une amélioration des performances du signal de la parole et on peut affirmer aussi que le filtre de Kalman est le plus performant en terme du SNR_{seg} et PESQ par rapport au filtre de Wiener et à la soustraction spectrale de Berouti pour tous types et niveaux du bruit. Le filtrage de Kalman a montré donc son efficacité, notamment pour de faible valeur de SNR et en particulier dans le filtrage avec modélisation du bruit.

Dans une seconde partie, nous avons appliqué le rehaussement à la reconnaissance automatique de la parole. Ainsi, nous avons étudié l'impact du rehaussement de la parole comme étape de prétraitement pour les systèmes de la reconnaissance vocale. Le but est d'améliorer les performances des systèmes de reconnaissance vocale dans un environnement réel. Pour cela, nous avons implémenté la reconnaissance de la parole suivant les modèles stochastiques (HMM) avec la plateforme HTK et neuromimétiques avec les MLP et RNN.

Les résultats obtenus illustrent bien la fiabilité des techniques de rehaussement de la parole quand à l'amélioration des performances des systèmes de reconnaissance vocale. Le filtre de Kalman peut apporter une amélioration importante à la reconnaissance de la parole pour les rapports signal sur bruit faibles. Donc, il peut s'intégrer dans la partie prétraitement acoustique des systèmes de reconnaissance de la parole.

D'importantes améliorations restent cependant à accomplir, parmi les perspectives envisagées, nous citons :

- ✓ L'adaptation des techniques de rehaussement de la parole au champ d'application, dans notre cas avec les systèmes de la reconnaissance vocale.
- ✓ L'utilisation des autres variantes pour les méthodes de rehaussement de la parole et les systèmes de reconnaissance de la parole.
- ✓ L'intégration du rehaussement de la parole comme une phase de prétraitement dans les systèmes de RAP et les systèmes RAL.
- ✓ L'incorporation de ces techniques de rehaussement et de reconnaissance vocale dans les systèmes de communication.

Bibliographie et Webographie

Bibliographie et Webographie

[1] S.F. Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction,” IEEE Trans. On Acoust., Speech, signal process, Vol. Assp.27, no. 2, pp. 113-120, April 1979.

[2] M. Berouti, R. Schwartz, and J. Makhoul. “Enhancement of Speech Corrupted by Acoustic Noise,” Proc ICASSP, pp. 208-211, April 1979.

[3] Y. Ephraim, D, Malah, “Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator,” IEEE Trans. On Acoust., Speech, signal process, Vol.32, no. 6, pp. 1109-1121, December 1984.

[4] Y. Ephraim, H. L. Van Trees, “A Signal Subspace Approach for Speech Enhancement,” IEEE Trans .on Speech audio processing, Vol.3, no. 4, pp. 255-266, July 1995.

[5] J. Jensen, J. H. L. Hansen,“Speech Enhancement Using a Constrained Iterative Sinusoidal Model ,” IEEE Trans .on Speech audio processing , Vol.9, no. 7, pp. 731-740, October 2001.

[6] J. Jensen, P .C .Hansen, S. D. Hansen, J. A. Sorensen, “Reduction of Broad Band Noise in Speech by Truncated QSVD, ” IEEE Trans .on Speech audio processing , Vol.3, no. 6, pp. 439-448, November 1995.

[7] S. Doclo, M. Moonen, “GSVD-Based Optimal Filtering for Signal and Multimicrophone Speech Enhancement,” IEEE Trans. on Signal Processing, Vol.50, no.9, pp. 2230-2244, September 2002.

[8] D. Labarre, “Du Filtrage de Kalman aux techniques H_0 et Particulaires Application au Traitement du Signal de Parole et à L’analyse de Signaux Biomédicaux, ” Doctorat, Université de BORDEAUX I, Décembre 2005.

[9] K. K. Paliwal, A. Basu, “A Speech Enhancement Method Based on Kalman Filtering,” proc. of IEEE-ICASSP, Vol.1, pp. 177-180, April 1987.

[10] J. D. Gibson, B. Koo, S. D. Gray, “Filtering of Colored Noise for Speech Enhancement and Coding,” IEEE Trans. on Signal Processing, Vol.39, no. 8, pp. 1732-1742, August 1991.

[11] S. Gannot, D.Buechtein, E. Weinstein, "Iterative and Sequential Kalman Filter-Based speech Enhancement Algorithms," IEEE Trans. on Signal Processing, Vol.6, no. 4, pp. 373-385, July 1998.

[12] M. Gabrea, E. Grivel, M. Najim, "A Single Microphone Kalman Filter-Based Noise Canceller, " IEEE Signal Processing Letters , Vol.6, no. 3, pp. 53-55, March 1999.

[13] E. Grivel, M. Gabrea, M. Najim, "Speech Enhancement as a Realization Issue," Signal Processing, Vol.82, no.12, pp. 1963-1978, December 2002.

[14] N. Ma, M. Bouchard, R. A. Goubran, "Perceptual Kalman Filtering for Speech Enhancement in Colored Noise," proc. of IEEE-ICASSP, Vol.2, pp. 717-720, May 2004.

[15] http://www.iframsurdite.com/thematique.html

[16] N. Virag, "Single Channel Speech Enhancement Based on Masking Properties of the Human Auditory System," IEEE Trans .Speech and Audio Processing, Vol.7, pp. 126-137, 1999.

[17] A, Amehraye, "Débruitage Perceptuel de la Parole, " Doctorat, Ecole Nationale Supérieure Des Télécommunications de Bretagne, 15 Mai 2009.

[18] Kotta Manohar, "Single Channel Enhancement of Noisy Speech," M. Tech. Credit Seminar Report, Electronic Systems Group EE Dept, IIT Bombay, Nov 2002.

[19] K. Manohar and Preeti Rao , "A Comparison Study of Spectral Subtraction Speech Enhancement Methods," National conference on communication, NCC 2004, IISc Bangalore, 2004

[20] J.S. Chang, B.L. Sim, Y.C. Tong and C.T. Tan, "A Parametric Formulation of the Generalized Spectral Subtraction Method," IEEE Tans on speech and audio proc, Vol. 6, no. 4, july 1998.

[21] S. Kamath, "A Multi-band Spectral Subtraction Method for Speech Enhancement," Master Thesis, the University of Dallas, 2001.

[22] S. Kamath, "A Multi-band Spectral Subtraction Method for Speech Enhancement," Master thesis, the University of Dallas, 2001.

[23] Mukul Bhatnagar, B. E, "A Modified Spectral Subtraction Method Combined with Perceptual Weighting for Speech Enhancement," Master Thesis University of Dallas, 2002.

[24] Nathalie Virag, "Speech Enhancement Based on Masking Properties of the Human Auditory System," Doctorat, Ecole Polytechnique Fédérale de LAUSANNE, 1996.

[25] L. Buniet, “Traitement Automatique de la Parole en Milieu Bruité : Etude de Modèles Connexionnistes Statiques et Dynamiques,” Thèse de Doctorat, Université Nancy1, 1997.

[26] C. Manohar, P. Rao, “Speech Enhancement in Non Stationary Noise Envirennement Using Noise Properties,” Speech communication, no. 48, pp. 96-108, 2006.

[27] P. Loizou, Speech Enhancement: Theory and Practice, Taylor and Francis, 2007.

[28] Ningping Fan, “Low Distortion Speech Denoising Using an Adaptative Parametric Wiener Filter,” IEEE. Siemens Corporate Research Inc, pp .I-309-I-312, 2004.

[29] C. Beaugeant, p. Scalart, “Noise Reduction Using Perceptual Spectral Change,” Processing of EUROSPEECH'99, pp. 2543-2546, Septembr 1999.

[30] L. Lin, W. H. Holmes and E. Ambikairajah, “Subband Noise Estimation for Speech Enhancement Using a Perceptual Wiener Filter,” Acoustics, Speech, and Signal Processing. IEEE-ICASSP, Vol. 1, pp. I80-I83, May 2003.

[31] T. Fillon, “Traitement Numérique du Signal Acoustique pour une Aide aux Malentendants,” Doctorat, École Nationale Supérieure des Télécommunications, Décembre 2004.

[32] B. Milner and I. Almajai, “Noisy Audio Speech Enhancement Using Wiener Filters Derived from Visual Speech,” In: Auditory-Visual Speech Processing, Kasteel Groenendaal, Hilvarenbeek, The Netherlands. August 31 - September 3, 2007

[33] V. Stahl, A. Fischer and R. Bippus, “Quntile Based Noise Estimation for Spectral Substraction and Wiener Filtering,” in Proc, IEEE. Acoustics, Speech and Signal Processing, Vol. 3, pp. 1875 - 1878. juin 2000.

[34] F. De coulon, Théorie et Traitement des Signaux, Presse Polytechniques Romandes, Lausanne, Suisse, 1996.

[35] L, Quarty, “Discrete Time Speech Signal Processing,” Upper Saddle Rive, NY, Prentice Hall, 2002.

[36] 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.

[37] G. Welch and G. Bishop, “An Introduction to the Kalman Filter,” Course8, University of North Carolina at Chapel Hill, August 2001.

[38] M. Najim, “ Filtrage Optimal, ” Technique de l'ingénieur, traité mesure et contrôle, article no. R7228.

[39] M. S. Grewal and A. P. Andrews, “Kalman Filtering Theory and Practice Using MATLAB 2nd edition,” John Wiley & Sons, Canada, pp. 1-12, 2001.

[40] M. Gabrea, “Robust Adaptative Kalman Filtering-Based Speech Enhancement Algorithm,” in Proc. IEEE Int. Conf. Acoustics Speech Signal Processing, pp.301-304, 2004.

[41] C. H. You, S. Rahardja, S. N. Koh, “Autoregressive Parameter Estimation for Kalman Filtering Speech Enhancement, ”in Proc IEEE ICASSP, pp. 913-916 , 2007.

[42] N. Ma, “Speech Enhancement Algorithms Using Kalman Filtering and Masking Properties of Human Auditory Systems,” Doctorat, University of Ottawa CANADA, August 2005.

[43] S. So, K. K. Paliwal, “A long State Vector Kalman Filter of Speech Enhancement,” in Proc ISCA, pp. 391-394, September 22-26. 2008.

[44] D. C. Popescu , I. Zeljkovic, “ Kalman Filtering of Colored Noise for Speech Enhancement,” in Proc. IEEE ICASSP, Vol. 2, pp. 997-1000, Seattle, USA, May 1998.

[45] M. Gabrea, “ An Adaptative Kalman Filter For Enhancement of Speech Signals in Colored Noise,” IEEE Workshop on applications of Signal Processing to Audio and Acoustics , pp. 45-48, October 16-19, 2005.

[46] V. Grancharov, J. Samuelsson, and W. B. Kleijn, “Improved Kalman Filtering for Speech Enhancement,” in Proc. IEEE ICASSP, pp. 1109-1112, 2005.

[47] W. Du, P. Driessen, “Speech Enhancement Based on Kalman Filtering and EM Algorithme, ” IEEE Pacific Rim Conference on Communications, Computers and signal Processing, pp. 142-145, May 9-10, 1991.

[48] 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.

[49] J. Sohn, N. S. Kim and W. Sung, “A Voice Activity Detector Employing Soft Decision Based Noise Spectrum Adaptation,” Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 365-368, 1998.

[50] 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.

[51] O. Cappé, “Elimination of the Musical Noise Phenomenon with the Ephraïm and

Malah Noise Suppressor", IEEE Trans. on ASSP, Vol. 2, pp. 345-349, 1994.

[52] V. Barreau, "Reconnaissance Automatique de la Parole Continue : Compensation des Bruits par Transformation de la Parole," Doctorat Université Henri Poincaré-Nancy1, 9 Novembre 2004.

[53] J. Razik, "Mesures de Confiance Trame-Synchrones et Locales en Reconnaissance Automatique de la Parole," Doctorat Université Henri Poincaré-Nancy1, 9 Octobre 2007.

[54] H. Glotin, "Elaboration et Comparaison des Systèmes Adaptatifs Multi-Flux de Reconnaissance Robuste de la Parole : Incorporation des Indices de Voisement et de Localisation," Doctorat Institut National Polytechnique De Grenoble, 13 Juin 2001.

[55] A. Amrouche, "Reconnaissane Automatique de la Parole par les Modèles Connexionnistes," Doctorat Université des Science et de la Technologie Houari Boumediène, 18 Décembre 2007.

[56] T. Hoya, Artificial Mind System, Kernel Memory Approach, Series: Studies in Computational Intelligence, Springer, Verlag, 2005.

[57] S. J. Young, HTK Version 1.4: Reference Manual and User Manual, Cambridge University, Engineering Department-Speech Group, 1992.

[58] A. Amrouche, M. Debyeche, A. Adoul, K. Amrouch and J.M. Rouvaen, "Reconnaissance des Phonèmes par Réseau de Neurones et Normalisation Temporelle: Application aux Consonnes Pharyngales et Glottale Arabe," in Proc. XXIIèmes J.E.P., S.F.A, pp. 397-400, June 15-19 1998, Martigny, Switzerland.

[59] A. Amrouche, M. Debyeche and J. M. Rouvaen, "Identification des Phonèmes par les Réseaux de Neurones pour la Reconnaissance de la Parole," in Proc. Conf. on Soft Computing and Applications, CSCA'99, pp. 131-136, October 18-19 1999, Algiers.

[60] Y. A. Alotaibi, "Investigation of Spoken Arabic Digits in Speech Recognition Setting," Informatics and Computer Sciences Vol. 173, no. 1, pp. 105-139, 2005.

[61] A. Waibel, T. Harazawa, G. Hinton, K. Shakano and K.J. Lang, "Phoneme recognition using Time-"Delay Neural Networks, " IEEE Trans. on ASSP, Vol. 37, no. 3, pp. 328–339, 1989.

[62] R. O. Duda, R.O. P.E. Hart and D. G. Stork, Pattern Classification, John Wiley and Sons, 2nd. Ed. New York, USA, 2001.

[63] D. F. Specht, "A General Regression Neural Networks," IEEE Trans. on Neural Networks, Vol. 2, no. 6, pp. 568–576, 1991.

[64] Y. A. Alotaibi, "Spoken Arabic digits Recognizer Using RNNs," in Proc. of the 4th IEEE Int. Symp. On Signal Proc. and Information Technology, pp. 195-199, December 18-21 2004, Roma, Italy.

[65] L. Rutkowski, "Generalized Regression Neural Networks in Time Varying Environments," IEEE Trans. on Neural Networks, Vol. 15, no. 3, pp. 576-596, 2004.

[66] A. P. Varga and H. J. M. Stenekken, "Assesment for ASR: II NOISEX-92: A Database and Experiment to Study the Effect of Additive Noise on Speech Recognition Systems," Speech Com. Vol. 12, no. 3, pp. 247-251, July 1993.

[67] A. P. Varga, H. J. M. Stenekken, M. Tomlessen and D. Jones, "The NOISEX-92, Study on the Effect of Additive Noise on ASR," Technical Report, DRA Speech Research Unit, Malvern, Worcestershire, UK.

Annexe A

Annexe A

Un système de reconnaissance de la parole sous HTK

Dans cette annexe nous présentons les étapes de conception d'un système de reconnaissance de la parole en utilisant l'outil HTK. Les étapes de la reconnaissance d'un mot isolé par HTK sont :

1. Représentation acoustique du signal

Dans cette phase, une paramétrisation des fichiers son de la base de données d'apprentissage est effectuée (Figure A.1). Dans notre cas nous avons utilisé la partie chiffre dans le mode broadcast de la base de données ARADIGIT.

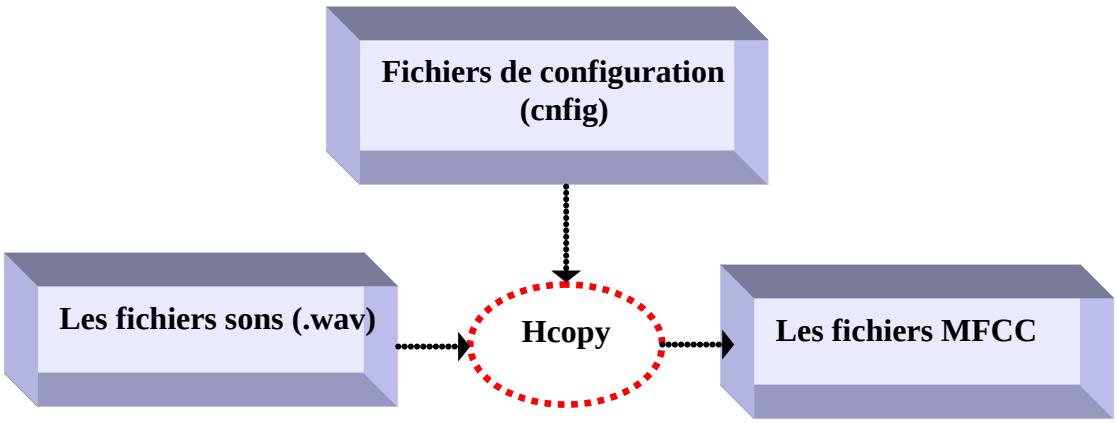

Figure A.1 : Représentation acoustique du signal.

On commence par définir le modèle de la langue, appelé aussi grammaire (Figure A.2), ensuite, on construit le dictionnaire (Figure A.3).

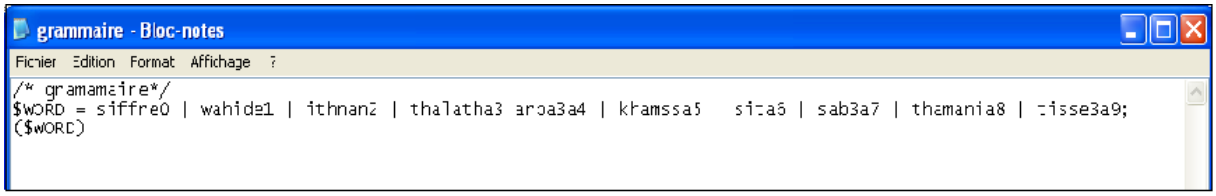

Figure A.2 : Grammaire de la base ARADIGIT.

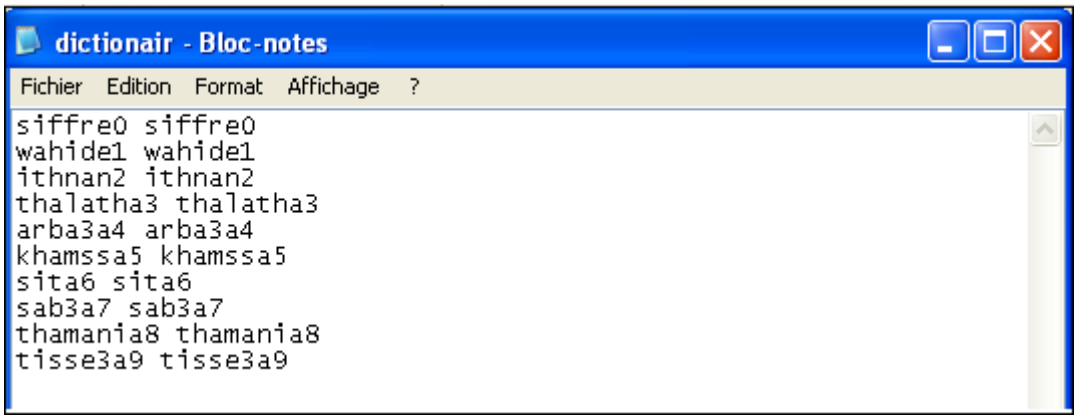

Figure A.3 : Dictionnaire de la base ARADIGIT.

Les coefficients MFCC sont extraits des fichiers wav et sur des fenêtres de 25ms grâce à l'outil **Hcopy** (ligne de commande A.1) en se servant du fichier de configuration **config** comme paramètre d'entrée :

Hcopy -D -C $config -S $liste_fich (A.1)

Le fichier de configuration **config** (Figure A.4) peut définir les paramètres indispensables pour la phase d'analyse acoustique. Le choix s'est porté sur les 12 premiers coefficients MFCC. Pour chaque coefficient, on attribue une dérivée première (12 dérivées premières au total) ainsi qu'une dérivée seconde (12 dérivées secondes) pour prendre en compte la dynamique du signal. En somme on obtient un vecteur acoustique de 36 coefficients correspondant à chaque trame du signal.

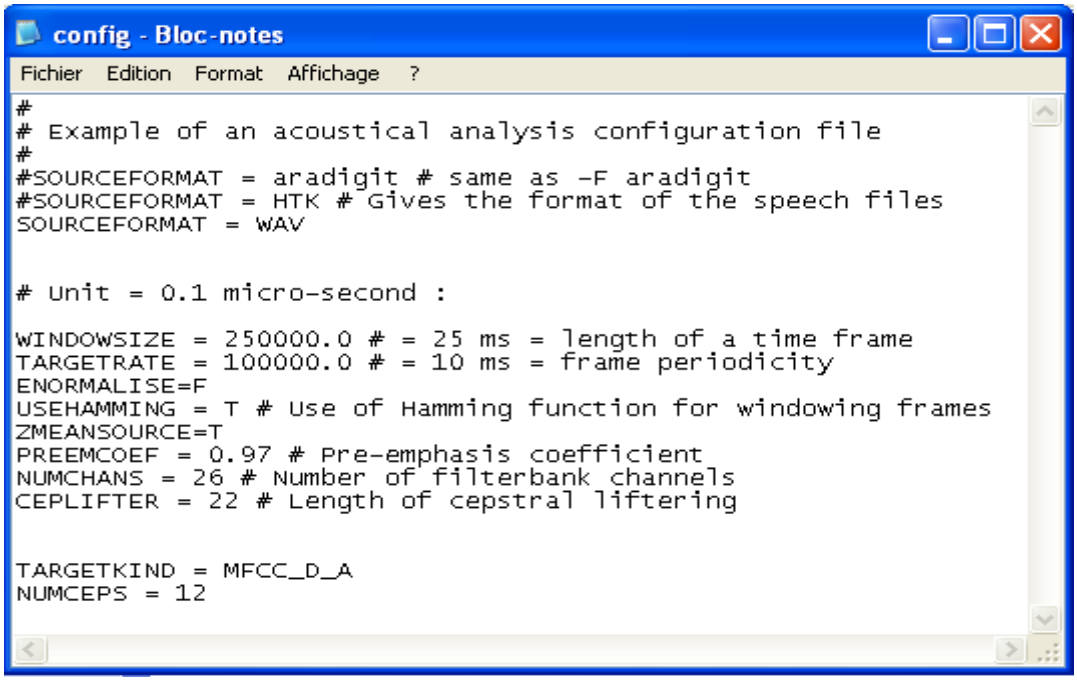

Figure A.4 : Fichier de configuration pour la phase de l'analyse acoustique.

Le résultat de cette phase (Figure A.5) est un ensemble de fichiers .mfcc dans le dossier MFCC contenant les coefficients.

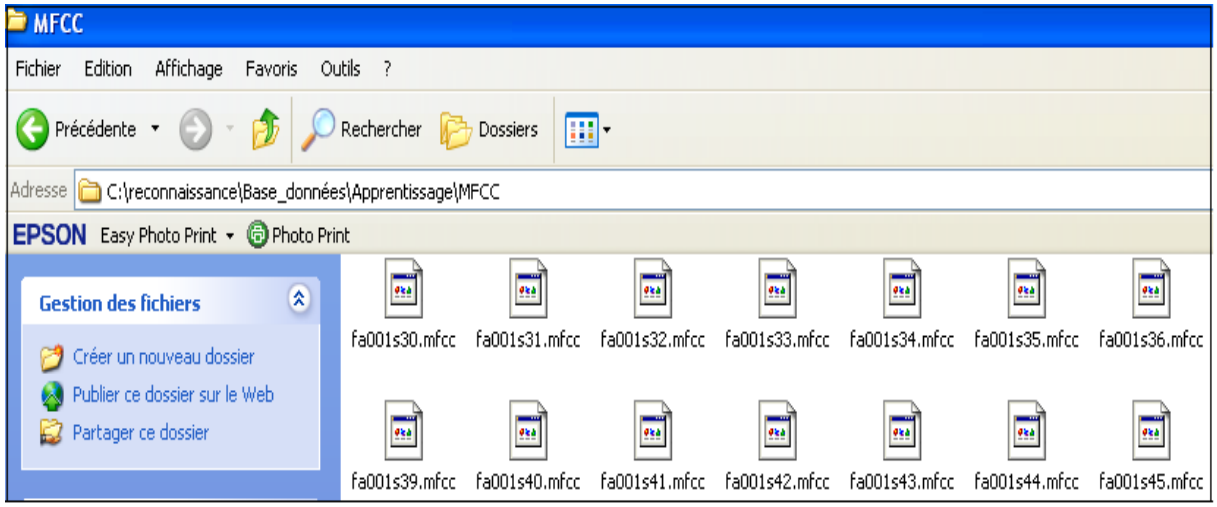

Figure A.5 : Dossier MFCC qui contient les coefficient mfcc.

2. Création des modèles HMM prototypes

Une fois qu'on a défini le dictionnaire et la grammaire, on passe à la description des modèles de Markov cachés. Pour chaque entité lexicale, on définira le modèle associé. Pour cela, on donnera la topologie de chaque modèle, le nombre d'états et les probabilités de transition entre les états.

La topologie HMM choisie (Figure A.6) est de type gauche-droit à 5 états dont les transitions autorisées sont initialisées sont dans la matrice de transition. La moyenne est initialisée à 0 et la variance à 1.

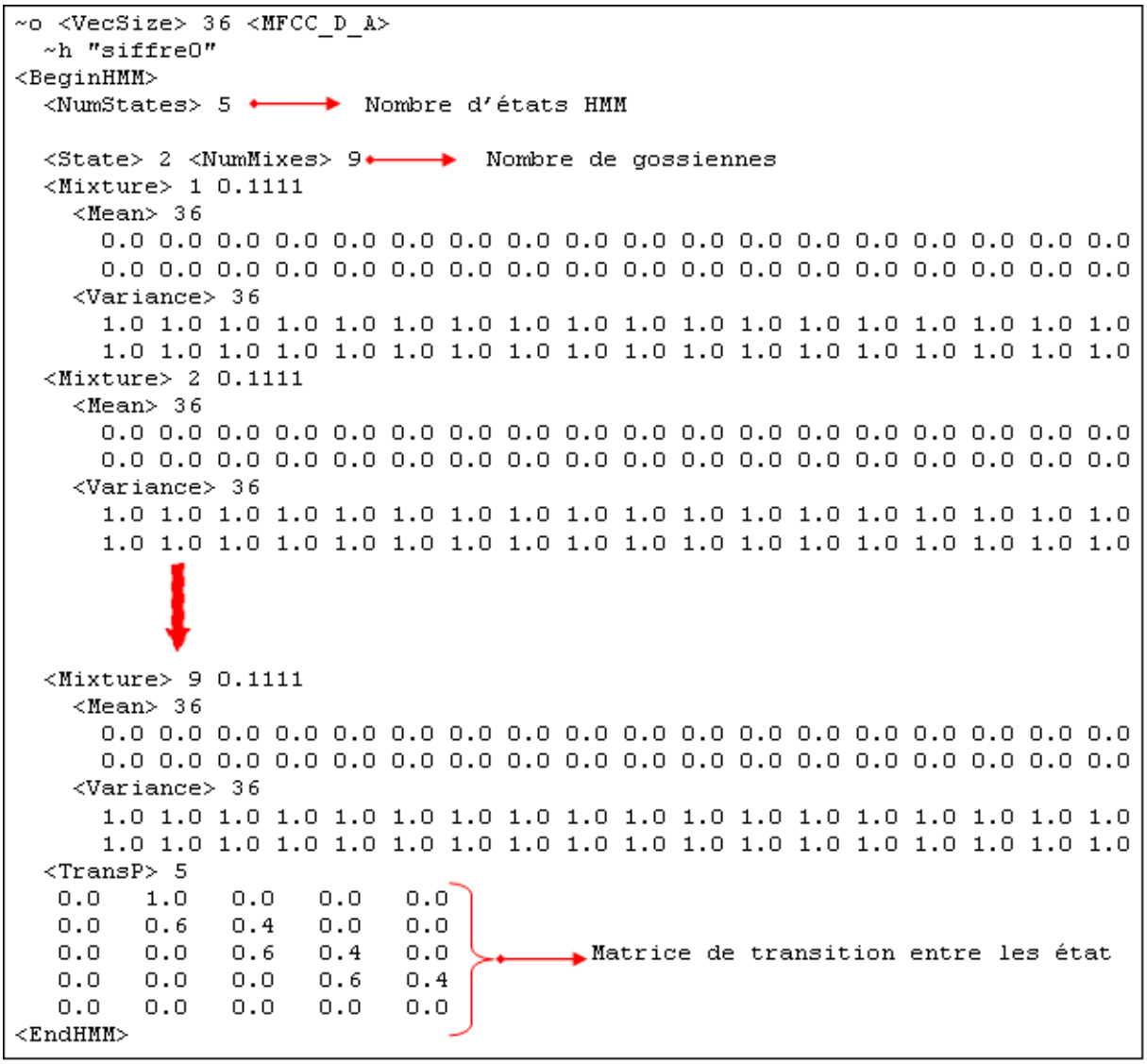

Figure A.6 : Fichier prototype d'initialisation.

3. Apprentissage

Lors de la phase d'apprentissage les paramètres de modèle HMM seront ré estimés. L'apprentissage est réalisé en deux étapes majeures: l'initialisation et la ré estimation.

La phase d'initialisation permet de mettre à jour les moyennes, les variances et les probabilités de transition entre états jusqu'à ce qu'un seuil de convergence ou qu'un nombre maximum d'itération soient atteint. Ceci est fait par l'algorithme de Viterbi.

Pour initialiser les modèles HMM, on applique l'outil **HInit** (ligne de commande A.2), la figure (A.7) représente le fichier résultat de cette commande.

HInit ($liste_hmm,$rep_hmms_inities,$rep_prototypes,$rep_mfcc) (**A.**2)

```
~o
<STREAMINFO> 1 36
<VECSIZE> 36<NULLD><MFCC_D_A><DIAGC>
~h "siffre0"
<BEGINHMM>
<NUMSTATES> 5
<STATE> 2
<NUMMIXES> 9
<MIXTURE> 1 2.397940e-001
<MEAN> 36
 -2.397482e+001 6.240237e+000 -9.567173e+000 -5.160815e-001 1.091027e+000 -
7.482425e+000 3.581874e+000 -2.392719e+000 -2.383030e-001 -1.851186e+000
1.485509e-001 -7.120632e-001 -3.781525e-001 -1.078587e-001 -5.510669e-001
2.213123e-001 -2.889897e-001 2.064435e-001 -6.569157e-002 1.278690e-001 -
1.297795e-001 -7.797478e-002 -4.155184e-002 7.515262e-002 1.075563e-001 -
6.334087e-002 1.813320e-001 -4.523319e-002 2.730190e-003 -8.489572e-003 -
5.944809e-002 -3.756697e-003 8.114487e-003 6.749070e-002 -3.317951e-002 -
5.902759e-002
<VARIANCE> 36
 5.887239e+000 7.396028e+000 1.922451e+001 2.742626e+001 2.578146e+001
2.196786e+001 2.048685e+001 2.855649e+001 2.127349e+001 1.811445e+001
1.650461e+001 1.521184e+001 2.100464e-001 5.253052e-001 9.369449e-001
1.052856e+000 1.173436e+000 1.718456e+000 1.386771e+000 1.390820e+000
1.597484e+000 1.476302e+000 1.658795e+000 1.352036e+000 4.013453e-002
8.279419e-002 1.705637e-001 1.808772e-001 2.186327e-001 2.674730e-001
2.330260e-001 2.666900e-001 3.012339e-001 2.558426e-001 2.889346e-001
2.468890e-001
```

```
<TRANSP> 5
 0.000000e+000 1.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000
 0.000000e+000 9.337992e-001 6.620081e-002 0.000000e+000 0.000000e+000
 0.000000e+000 0.000000e+000 8.924088e-001 1.075912e-001 0.000000e+000
 0.000000e+000 0.000000e+000 0.000000e+000 9.433606e-001 5.663939e-002
 0.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000
<ENDHMM>
```

Figure A.7 : Fichier résultat de la commande HInit.

Par la suite, les modèles HMM sont ré estimés de façon indépendante avec l'algorithme de Baum Welch en utilisant la commande **HRest** (ligne de commande A.3), la figure (A.8) représente le fichier résultat de cette commande.

HRest($liste_hmm,$rep_hmms_entraines,$rep_hmms_inities,$rep_mfcc) (A.3)

```
~o
<STREAMINFO> 1 36
<VECSIZE> 36<NULLD><MFCC_D_A><DIAGC>
~h "siffre0"
<BEGINHMM>
<NUMSTATES> 5
<STATE> 2
<NUMMIXES> 9
<MIXTURE> 1 2.385591e-001
<MEAN> 36
 -2.382454e+001 6.243543e+000 -9.406921e+000 -4.304787e-001 9.406039e-001 -
7.422508e+000 3.416749e+000 -2.420051e+000 -2.677918e-001 -1.906896e+000
2.117597e-001 -7.676847e-001 -3.858622e-001 -1.017566e-001 -5.605260e-001
2.319177e-001 -2.873726e-001 2.061277e-001 -4.657719e-002 1.285235e-001 -
1.356969e-001 -7.260159e-002 -3.559946e-002 8.181649e-002 1.004514e-001 -
6.826609e-002 1.627952e-001 -3.942966e-002 3.565171e-003 2.548283e-003 -
5.026236e-002 -3.222687e-003 1.011455e-002 6.793239e-002 -3.640525e-002 -
5.621304e-002
<VARIANCE> 36
 6.267688e+000 7.484953e+000 1.900351e+001 2.788663e+001 2.668456e+001
2.189226e+001 2.055404e+001 2.802755e+001 2.096082e+001 1.794263e+001
1.662189e+001 1.535888e+001 2.120706e-001 5.265909e-001 9.350460e-001
1.024293e+000 1.152183e+000 1.685674e+000 1.401438e+000 1.380390e+000
1.573262e+000 1.521523e+000 1.644706e+000 1.337630e+000 3.754331e-002
8.002968e-002 1.654512e-001 1.767221e-001 2.123558e-001 2.596314e-001
2.340834e-001 2.606094e-001 3.029055e-001 2.503834e-001 2.881733e-001
2.423906e-001
```

```
<TRANSP> 5
 0.000000e+000 1.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000
 0.000000e+000 9.341850e-001 6.581503e-002 0.000000e+000 0.000000e+000
 0.000000e+000 0.000000e+000 8.902658e-001 1.097342e-001 0.000000e+000
 0.000000e+000 0.000000e+000 0.000000e+000 9.436573e-001 5.634271e-002
 0.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000
<ENDHMM>
```

Figure A.8 : Fichier résultat de la commande HRest.

4. Reconnaissance

La reconnaissance consiste à comparer l'image de l'unité à identifier avec celles de la base de référence en utilisant successivement les modèles issus de **HInit** et ceux issus de **HRest.** Pour cela nous allons utiliser la commande **HVite** (ligne de commande A.4)

HVite -d$rep_hmms_entraines-i $reco-w $reseau_syntaxique-S$liste_test_mfcc$dictionnaire $liste_hmm (A.4)

Les résultats sont évalués par alignement dynamique avec les données de référence par l'outil **HRESULT** (ligne de commande A.5)

HRESULTS -p -I $fich_ref $liste_hmm $reco> $resultats (A.5)