وزارة التعليم العالي والبحث العلمي

BADJI MOKHTAR-ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR-ANNABA

جامعة باجي مختـــار- عنابـــــة

Faculté des Sciences de l'ingénieur

Département d'électronique

THÈSE

Présentée en vue de l'obtention du diplôme de **MAGISTER**

La recherche des paramètres de la trace acoustique et son application dans la reconnaissance de la parole

Option
Systèmes intelligents

Par
HADRI Cherif

DIRECTEUR DE THÈSE : **BOUGHAZI Mohamed** **M.C** **U.Annaba**

DEVANT LE JURY

PRESIDENT	**BOUSBIA SALAH MOUNIR**	**M.C**	**U.Annaba**
EXAMINATEURS	**HAMDI RACHID**	**M.C**	**U.Annaba**
	MESSADEG DJAMIL	**M.C**	**U.Annaba**

Année 2007/08

Résumé

La première étape dans un système de reconnaissance automatique de la parole (RAP) est l'analyse acoustique et le traitement du signal qui transforment le signal parole en une séquence de vecteurs acoustiques, Cette représentation doit être adaptée pour la reconnaissance, on conservons dans les vecteurs acoustiques que l' information lexicale et de supprimer toutes autres informations, telles que variabilité intra et interlocuteur, les bruit ambiants etc.

La représentation utilisée généralement en reconnaissance est basée sur des coefficients cepstraux (LPCC, MFCC, PLP). Bien que les coefficients cepstraux soient utilisés en raison de leurs propriétés de représentation, notamment la décorrélation des coefficients, ils souffrent de plusieurs limitations. En particulier ils sont sensibles aux conditions d'acquisition du signal et à l'environnement acoustique (problème de robustesse). A cause de cette sensibilité, la performance d'un système RAP est dégradée, elle est encore plus dégradée quand les conditions de l'apprentissage et de l'utilisation du système sont différentes. Le but de ce travail est d'étudier et de mettre en œuvre des paramètres (représentations) robustes aux différences entre les conditions acoustiques d'apprentissage et d'évolution. Ces paramètres seront évalués sur un système de reconnaissance automatique des chiffres arabe. Une attention particulière sera prêtée aux méthodes d'extractions des paramètres robustes (CMS, CGN, RASTAPLP, MBLPCC, LPC MFCC).

Summary

The first stage in an automatic speech recognition system (ASR) is the acoustic analysis and the treatment of the signal which transforms the speech signal into a sequence of acoustic vectors, This representation must be adapted for the recognition, i.e. to keep in the acoustic vectors any lexical information and to remove all other information, such as variability will intra and interlocutor, ambient noise etc.

The representation used generally in recognition is based on cepstrals coefficients (LPCC, MFCC, PLP). Although the cepstral coefficients are used because of their representation properties, in particular the decorrelation of the coefficients, they suffer from several limitations. In particular they are sensitive to signal acquisition conditions and to acoustic environment (robustness problem). Because of this sensitivity, the performances of ASR systems are degraded, even more when the conditions of training and testing are different. The goal of this work is to study and implement features (representations), which are robust to the differences between the acoustic conditions of training and evolution. These features will be evaluated in an Automatic Arab digits recognition system. A particular attention will be lent to the methods of robust features extractions (CMS, CGN, RASTAPLP, MBLPCC, and LPC MFCC).

REMERCIEMENTS

En premier lieu je tiens à remercier Mr Bedda mouldi pour être à l'origine de ce projet et bien sûr pour en avoir assumé la direction. Ma gratitude va également à Mr Boughazi Mohamed pour sa direction et ses conseils avisés.

Je voudrais également remercier Mr Bousbia salah mounir pour l'honneur qu'il me fait en présidant ce jury de thèse. Je remercie très sincèrement Mr Messadeg djamil et Mr Hamdi rachid pour avoir accepté d'évaluer ce travail.

Je tiens à remercier tous les membres passés et présents du Laboratoire d'automatique et signaux d'Annaba LASA et du Laboratoire d'automatique et informatique de Guelma LAIG avec qui j'ai partagé bien plus qu'un lieu de travail et plus particulièrement Toufik, Lotfi, Naim, Faiçal, Laid, Debha, Karim, sofiane, Abd anour, Riad, Djalil.

Un merci particulier va à Rachid et Farouk pour leurs soutiens .

Enfin je voudrais remercier du fond du cœur mes frères particulièrement Sofiane ma mère et à toute ma famille pour avoir été courageux et de m'avoir attendu si longtemps...

La Table des Matières

Chapitre4 :

Le Choix D'une Nouvelle Représentation Robuste D'un Signal Vocal

Chapitre 5 :

Résultats expérimentaux

Liste des figures

LISTE DES TABLEAUX

Introduction Générale

De nos jours, il n'y a aucun doute que les machines et les ordinateurs sont largement répandus presque par chaque personne pour faciliter la gestion et le stockage de l'information. Pour cette raison, les outils comme les ordinateurs, et la microélectronique ont connu une évolution considérable dans ces dernières années. Cette évolution a permis de faciliter la communication entre l'homme et la machine par l'usage de la parole où l'information transmise à la machine est un signal vocal. Après traitement la machine répond par un autre signal vocal adéquat.

Il est clair que cette opération nécessite des traitements sur le signal vocal telles que reconnaissance - synthèse. L'application de la reconnaissance est multiple l'exemple suivant montre une de ces applications qui peut être implémentée réellement.

Où en remarque la présence de tous les champs d'un système TLH (technologie de langue humaine).

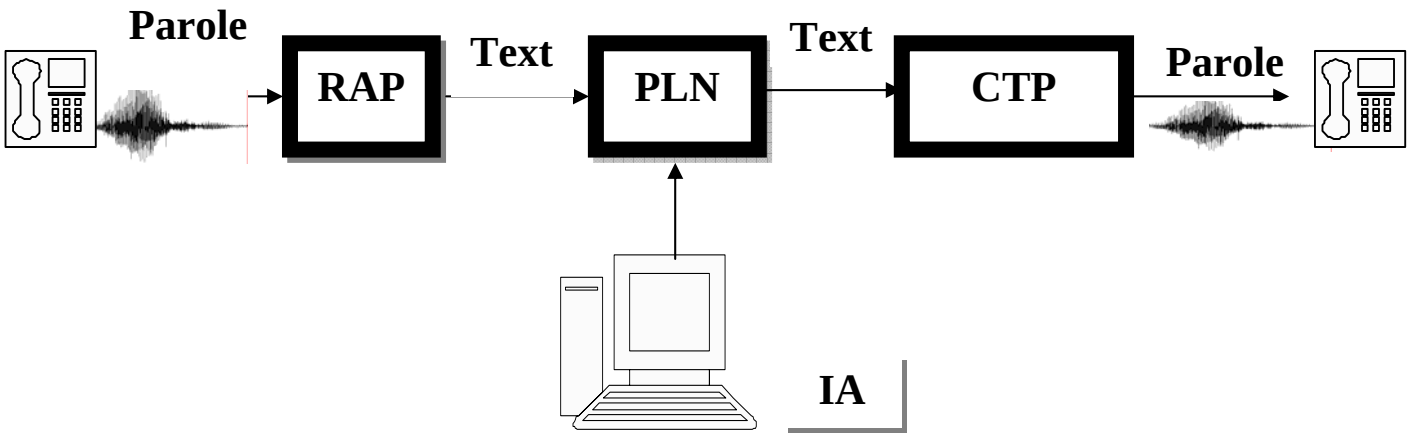

Schéma fonctionnel d'une application de technologie de la langue humaine

La reconnaissance automatique de la parole (RAP) traite le fait d'identifie le discours humain, processus de langage naturel (PLN) avec extraction d'information des phrases ou la gestion du dialogues connexes aussi avec l'intelligence artificielle (IA) et les techniques de conversions texte parole (CTP).

Il n'est pas difficile d'imaginer une situation où TLH peut être appliqué avec succès, et par conséquent RAP, mais les applications les plus importantes peut être la dictée (transcriptions médicales, légales ou d'affaires), applications de téléphone (téléphone opérations bancaires, audio messagerie) et applications conçue pour rendre quelques services accessible aux

personnes handicapées (distributeur automatique de billets pour personnes aveugles, une conversation téléphonique assistée par ordinateur pour sourd ou sourd-muet ou paramètres commandés par voix pour des personnes avec des problèmes musculaires).

La reconnaissance automatique de la parole (RAP) est un exemple typique d'un problème de classification automatique des modèles (formes), le but d'un RAP est de déterminé la séquence des mots la plus probable pour un flux d'information acoustique. Un système RAP se réalise en deux phases apprentissage et classification.

Mais, avant que les modèles soient appris ou classés. Le signal parole a besoin d'être codé dans un vecteur des paramètres acoustiques représentatif

Extraction des paramètres de la trace acoustique

La première étape dans tous les algorithmes RAP (reconnaissance automatique de la parole) est l'extraction des paramètres acoustiques. Le but de la procédure d'extraction des paramètres est de traduire l'information contenue dans les signaux acoustiques vers une représentation des données qui convient au calcul statistique de vraisemblance et de probabilité.

Un système RAP exige des paramètres acoustiques qui représentent l'information phonétique fiable, c.-à-d les paramètres qui décrivent les propriétés distinctives des sons de la parole efficacement. Dans le meilleur des cas, les paramètres acoustiques devraient clairement indiquer les différences linguistiques appropriées entre différents sons de la parole, tout en masquant la variation acoustique du signal qui ne représente pas des différences phonétiques ou qui n'est pas lié aux événements de la parole.

L'extraction acoustique des paramètres est un composant essentiel dans un système RAP, parce que pendant le processus une transition est faite à partir des signaux continus aux éléments discrets les plus fondamentaux de la reconnaissance de la parole. Une multitude de technique ont été proposées pour produire l'extraction acoustique de paramètre efficacement, par exemple LPC, LPCC, MFCC, PLP., Le but commun de presque toutes ces représentations des paramètres est de décrire les signaux acoustiques en termes de leur distribution spectrale d'énergie courte terme. L'information sur la distribution spectrale d'énergie d'un signal est habituellement extraite dans des intervalles réguliers de temps. En plus de ces derniers appelés des coefficients statiques les premiers et second dérivés de temps des paramètres statiques désignés sous le nom paramètre des dérivés des paramètres

statiques également visés comme des paramètres dynamiques – sont presque toujours inclus dans le vecteur acoustiques des paramètres.

Plan de mémoire

Le but de ce travail est d'étudier et de mettre en œuvre des paramètres (représentations) robustes de la trace acoustique aux différences entre les conditions acoustiques d'apprentissage et d'évolution. Ces paramètres seront évalués sur un système de reconnaissance automatique des chiffres arabe. Une attention particulière sera prêtée aux méthodes d'extractions des paramètres robustes (CMS, CGN, RASTAPLP, MBLPCC, LPC MFCC).

Ce mémoire est organisé en 5 chapitres :

- Le premier chapitre est consacré à la théorie de production et perception d'un signal vocal.
- Le chapitre 2 donne une vision globale sur les outils utilisés dans le traitement des signaux vocaux (acquisition, filtrage, segmentation, transformation, etc.).
- Dans le chapitre 3 nous présentons les techniques de représentations du signal les plus utilisées en reconnaissance de la parole (LPCC, PLP, MFCC).
- Le chapitre 4 présente une série des techniques qui ont été envisagées pour rendre les systèmes de reconnaissance de la parole plus robustes aux différentes variabilités acoustiques.
- Le chapitre 5 montre l'influence, en termes de taux d'erreur sur les mots, dans le cas de la présence de bruit en phase de test, sur les différentes techniques de paramétrisation présentées dans les chapitres 3 et 4. L'objectif principal de ce chapitre expérimental est de montrer que quel que soit le degré de bruitage des données de test, le système de reconnaissance de la parole se trouve sérieusement perturbé. Il est donc indispensable de mettre en œuvre des techniques permettant de compenser les effets du bruit pour garantir la robustesse du système de reconnaissance lors d'une application dans des conditions acoustiques difficiles.

Nous terminons ce travail par une conclusion et les perspectives espérées.

Production Acoustique et Perception de la Parole

La parole est généralement classée dans trois perspectives différentes [01]:
Perspectives articulatoires, acoustiques *et* perceptuelles. Dans l'approche articulatoire, on essaye de décrire comment les humains produisent des sons de la parole par une près étude de la l'anatomie et la physiologie des organes de production de la voix.
Dans l'approche acoustique, le son articulé acoustique est lui-même l'objet d'intérêt.
Dans l'approche perceptuelle, on examine l'anatomie et la physiologie du mécanisme et d'audition humains pour trouver les modèles qui relient les mesures acoustiques aux caractéristiques perceptuels.

1.1 Approche articulatoire

La production de la parole est un processus complexe qui dans un modèle simplifié consiste les tâches successives suivantes [02]:

1. Formulation de message,
2. Codage du message dans un code de langue,
3. Traçage du code de langue dans des commandes neuro--musculaires,
4. Réalisation des commandes neuraux—musculaires.

Le résultat final des commandes neuro--musculaires complexes est exposé par les mouvements physiques dans les organes de production de la parole Fig.1.1 [03].

Dans une classification commune trois composants physiologiques de production de la parole sont reconnus [04]:

- *composant sous glotte* qui comprend les poumons et les muscles respiratoires associés,
- *le larynx* qui inclut les cordes vocales,
- *conduit vocal sup. laryngienne* qui comprend les cavités : pharyngienne, oral, et nasale.

Tous les trois composants, particulièrement le conduit vocal, sont des systèmes complexes avec une nature variable en temps c à d lorsque en parle, la configuration de ces composants change sans interruption.

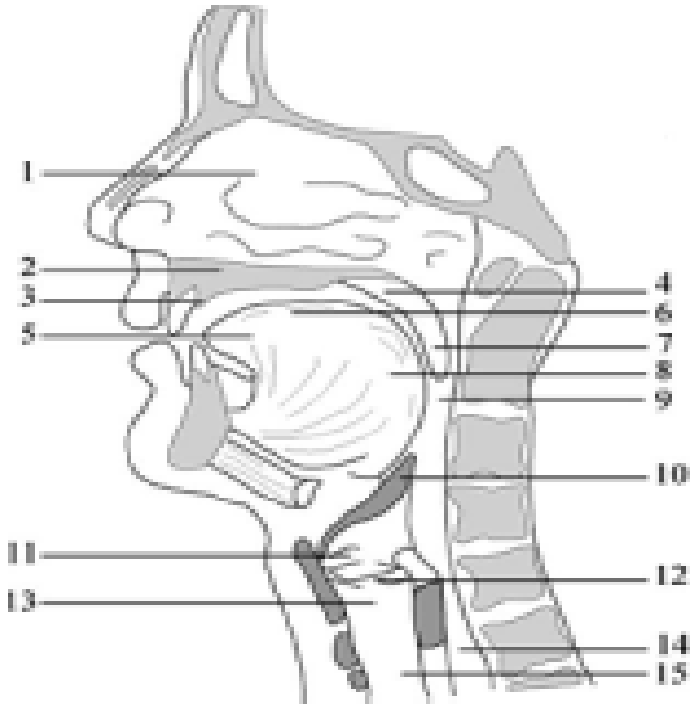

Figure 1.1: Les organes vocaux humains: (1) cavité nasale, (2) palais dur, (3) arête alvéolaire, (4) palais mou (voile), (5) bout de la langue (apex), (6) Dorsum, (7) luette, (8) base, (9) pharynx, (10) épiglotte, (11) cordes vocales fausses, (12) cordes vocales, (13) larynx, (14) œsophage et (15) trachée.

1.1.1 Le Système Respiratoire Sous glotte

Le composant sous glotte produit un courant d'air qui actionne le procédé de production de la parole. Pendant l'inspiration, la force musculaire est employée en remplissant les poumons. Les poumons augmenteront en leur volume comme ce qui arrive à un ballon en caoutchouc quand on souffle l'air dans ce dernier, et l'énergie est stocké dans les expansions élastiques de chaque poumon. Pendant l'expiration, cette énergie est spontanément libérer à une prétendue *force élastique de reflux* [04]. Le courant d'air résultant traverse *la trachée* au larynx.

1.1.2 Larynx

Larynx est responsable des différents *mécanismes de phonation* [04]. Ceci se rapporte à produire l'énergie acoustique qui sert d'entrée à la région vocale. Plus spécifiquement, la glotte et *les cordes vocales* sont des pièces intéressantes d'un point de vue de production de la parole. La glotte est un petit espace entre les cordes vocales d'une forme triangulaire [04]. Le courant d'air agressif des poumons passe par la glotte au conduit vocal. L'action des cordes vocales détermine le type de phonation, dont les types principaux sont non *voisés* et *voisés*.

Pendant la phonation non voisée, les cordes vocales sont libres et le courant d'air des poumons traverse la glotte ouverte.

La phonation voisée est un mécanisme plus complexe que la phonation non voisée. Le son voisé est un résultat des répétitions périodiques d'ouverture et de fermeture des cordes vocales. Ceci est représenté sur la fig.1.2. Pendant la phase d'ouverture, l'effort respiratoire accumule la pression sous glotte jusqu'à ce qu'il surmonte la force musculaire qui garde les cordes vocales ensemble. La glotte s'ouvre, et Le courant d'air compressé éclate dans le pharynx. Ceci cause relativement à grande vitesse une baisse locale de pression atmosphérique dans la glotte, et comme conséquence de ce prétendu *effet de Bernoulli* les cordes vocales commencent à se fermer.

L'effort combiné de l'effet de Bernoulli et de la tension musculaire surmonte très rapidement la force de la pression respiratoire, et les cordes vocales sont tirées ensemble. Le résultat d'accouplement continu des phases d'ouverture et de fermeture est un jet périodique des souffles d'air qui sert de signal acoustique de source pour les sons voisés.

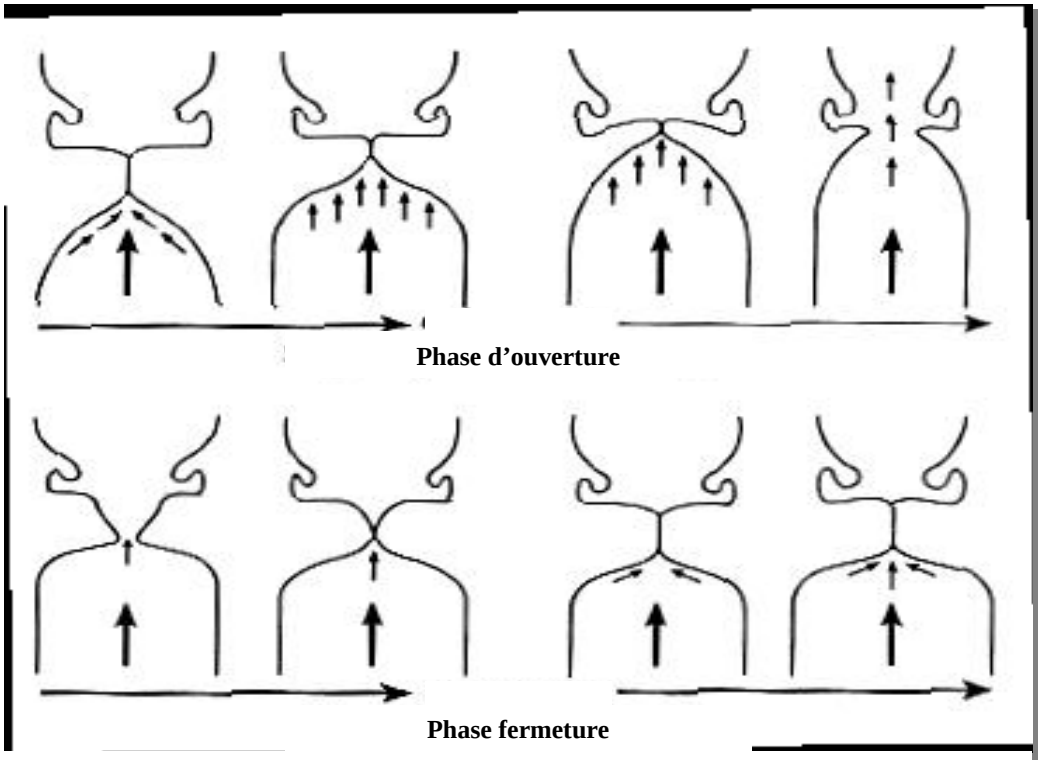

Figure 1.2 illustration d'un cycle glottal

Les phases d'ouverture fermeture ne forment pas un cycle parfaitement périodique dans le sens mathématique, et donc la limite *quasi-périodique est* utilisée dans ce contexte.

Le temps d'un cycle ouverture -fermeture est désigné sous le nom *période fondamental* l'inverse de cette période est signalée sous le nom *fréquence- fondamental* et *abrégé* F_0.

La fréquence fondamentale diffère entre les femelles, les mâles et les enfants [04]. Ceci est le résultat des différences anatomiques; habituellement les femelles ont de plus petites cordes vocales comparées aux mâles, et en raison de leur tension plus élevée, elles vibrent à plus haut taux. Les enfants ont aussi des cordes vocales plus petites. Les valeurs *moyennes* de F_0 dans un discours conversationnel pour des mâles, les femelles et les enfants ont lieu approximativement 120 hertz, 220 hertz, et 330 hertz, respectivement [02]. C'est important de maintenir dans l'esprit que *F*0 est défini seulement pour la phonation voisée.

1.1.3 Le conduit vocal

Le conduit vocal est également le plus important du système complexe dans le procédé de production de la parole. Le conduit vocal est une limite générique qui se rapporte aux organes de production de la voix au-dessus du larynx.

Les parties principales du conduit sont montrées dans un schéma de la fig. 1.3. Les trois cavités principales du conduit sont les cavités pharyngienne, orale et cavité nasale. *Le palais ou* le voile *mou* commande la quantité de flux d'air à la cavité nasale.

Les parties du conduit, particulièrement ceux de la cavité orale, servir comme *articulateurs.* Chaque *geste articulatoire* par exemple, un mouvement de langue, vise une certaine cible *phonétique idéale.* L'événement *acoustique réalisé* rapproche la cible phonétique. Les gestes articulatoires en général se chevauchent dans le temps. Dans un autre terme, la cible phonétique précédente affecte la prochaine cible (et donc, ses paramètres acoustiques également). Le phénomène est connu sous le nom de *coarticulation.* En raison de la coarticulation, les cibles phonétiques ne sont pas codées dans le son articulé en tant que segments linéaires simples qui se suivent dans le temps, tel comme lettres dans un texte écrit. Coarticulation est l'un des raisons pour lesquelles la segmentation automatique de la parole dans des événements phonétiques demeure un problème difficile.

L'articulateur le plus flexible est la langue, qui peut avoir des positions et orientations diverses. Il peut être fait, par exemple, pour former un passage étroit (une prétendue *réduction* dans le conduit, par lequel le courant d'air passe.

En raison de la réduction, le courant d'air devient turbulent et fait la caractéristique " bruit de sifflement " de certains phonèmes. Un exemple de ceci est le non voisé la fricative [s] où la restriction est constituée par un réglage du corps de la langue contre le palais dur.

D'autre part dans la production des voyelles, les écoulements du courant d'air passent librement par le conduit. Cependant, aussi dans ce cas-ci, il y a une constriction dans la cavité orale. La section de la constriction est de manière significative plus grand comparé aux fricatives, et donc turbulence n'est pas formée. Au lieu de cela, *une vague debout* surgit. L'endroit et la section du la constriction orale détermine dans la plupart du temps quelle voyelle est produite [02]. Pour cette raison, les voyelles sont souvent classifiées en tant que *l'avant, l'arrière* ou *mi* voyelles basés dessus l'endroit de la constriction orale. Par exemple, [a] est une voyelle arrière et [i] est

une voyelle avant. L'arrondi des lèvres affecte également la qualité phonétique de certaines voyelles.

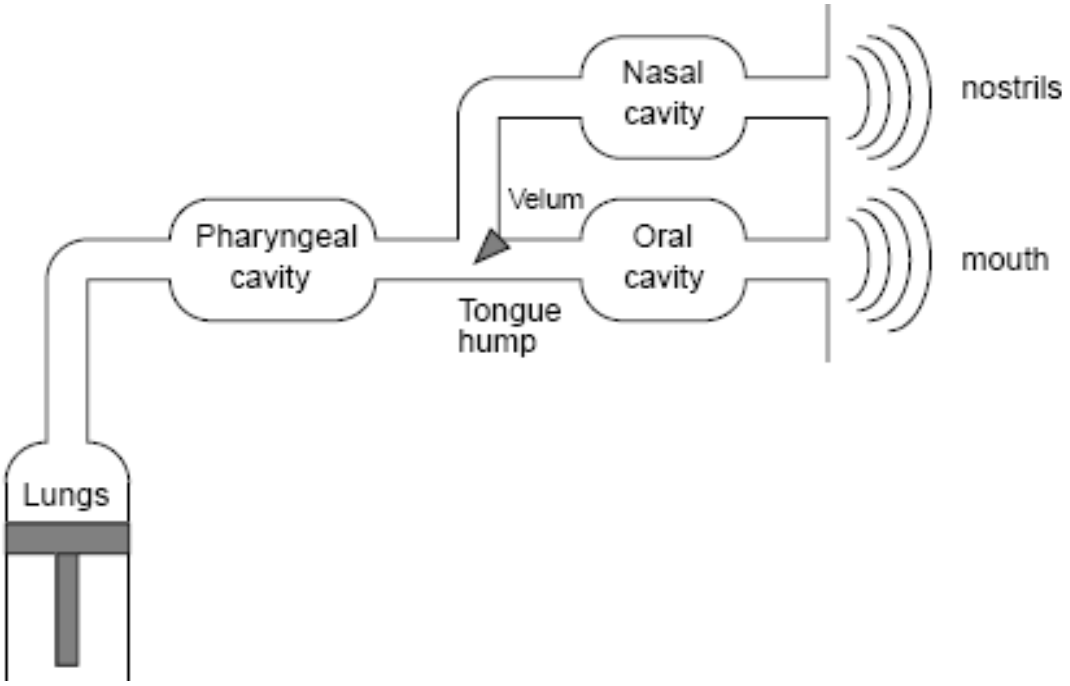

Figure 1,3: Un modèle simplifié du conduit vocal.

Les sons nasaux sont produits par la cavité nasale avec le voile ouvert et une fermeture dans la cavité buccale. On peut considérer le voile comme valve qui commande la quantité de courant d'air à la cavité nasale. Par exemple, pendant la production de l'initiale [m] dans le mot " maman" on peut noter que les lèvres sont fermées. En conséquence, le courant d'air coule par l'intermédiaire de la cavité nasale et sortie des narines. Un autre exemple d'un bruit nasal est l'initiale [n] dans le mot " non". Ce bruit est également produit en faisant accomplir la fermeture, cette fois par le corps de la langue. Ainsi, l'endroit oral de la constriction joue le rôle de la qualité phonétique des différents bruits nasaux.

1.2 Approche acoustique

La phonétique articulatoire essaye de décrire comment des sons de la parole sont produits en termes de gestes articulatoires, tandis que l'approche acoustique vise une conclusion des corrélations acoustiques de la physiologie et des aspects comportementaux des organes de production de la voix. Le son articulé acoustique ne porte pas une agrafe visuelle des mouvements de lèvre.

Cependant, certains paramètres acoustiques ont plus où moins dirigent des corrélations avec l'anatomie et la physiologie des organes de production de parole. Des sons articulés peuvent être analysés dans *le domaine temps* ou *fréquence.* Un exemple d'une forme d'onde *dans le domaine temps* et *du spectre court terme* du même segment est montré dans fig. 1.4.

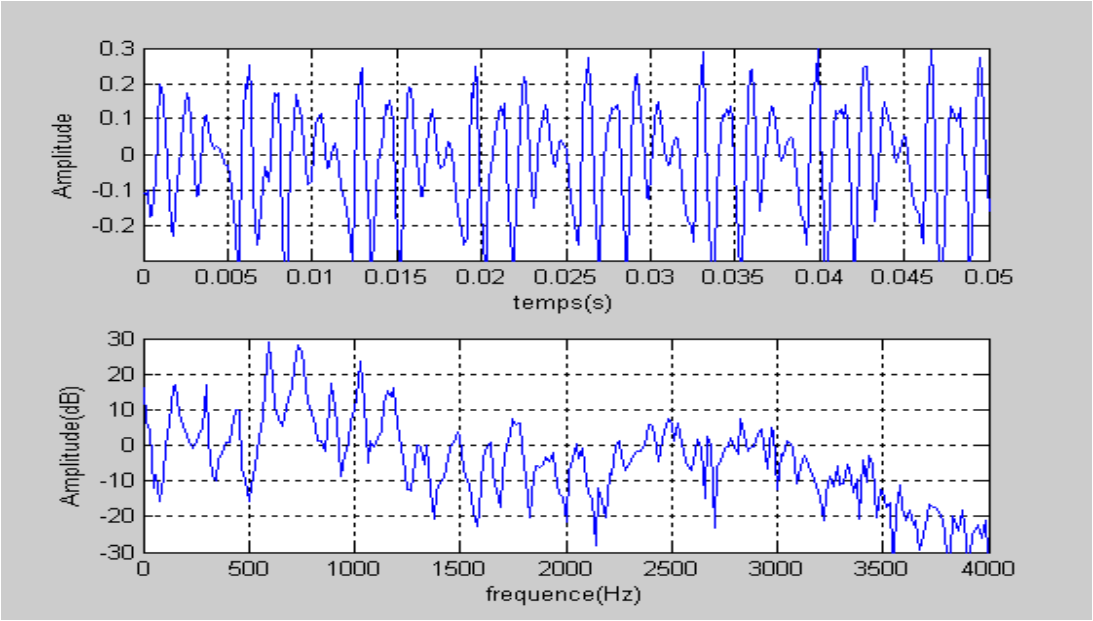

Fig.1.4: La forme temporelle et spectrale de la voyelle [a] prononcé par un locuteur masculin.

1.2.1 Analyse spectrographique de la parole

Dans la recherche phonétique acoustique, deux représentations utiles sont la forme d'onde et une parcelle de terrain temps- fréquence appelée le *spectrogramme.* La forme d'onde montre les variations de pression d'air, tandis que le spectrogramme montre l'importance des différentes

fréquences en fonction de temps. Des exemples des spectrogrammes sont montrés sur les figures 1.5 et 1.6. Le niveau de gris en position (*t ;* f) montre l'importance de la fréquence *f* au temps *t* de sorte que des régions plus foncées correspondent à des grandeurs plus élevées.

Il y a une compensation entre les résolutions temps et fréquence. Si la résolution de temps est haute, c.-à-d., une fenêtre courte d'analyse est employée, la résolution fréquence devient plus mauvaise et vice versa. Les résolutions sont approximativement inversement proportionnelles entre eux. Par exemple, une résolution de temps de 20 millisecondes donne approximativement un espacement de fréquence de 1/0,02 = 50 hertz. Dans le traitement des signaux et la physique, le rapport entre le moment et la résolution fréquence est connu comme *principe d'incertitude* [05].

Il y a deux types de spectrogrammes: *spectrogrammes* à large bande *et* à bande étroite (figure 1.5 et 1.6). Dans les spectrogrammes à large bande, la largeur de bande du filtre d'analyse est autour de 300 hertz et l'espacement temps est ainsi approximativement de 1/300 s = 0.0033 s pour l'analyse à bande étroite, la largeur de bande est autour 50 hertz et l'espacement temps est ainsi autour de 1/50 s = 0.020 s [51]. Spectrogrammes à large bande conviennent au cheminement des formants des voyelles tandis que les spectrogrammes à bande étroite peuvent être employés dans *l'évaluation* de F_0.

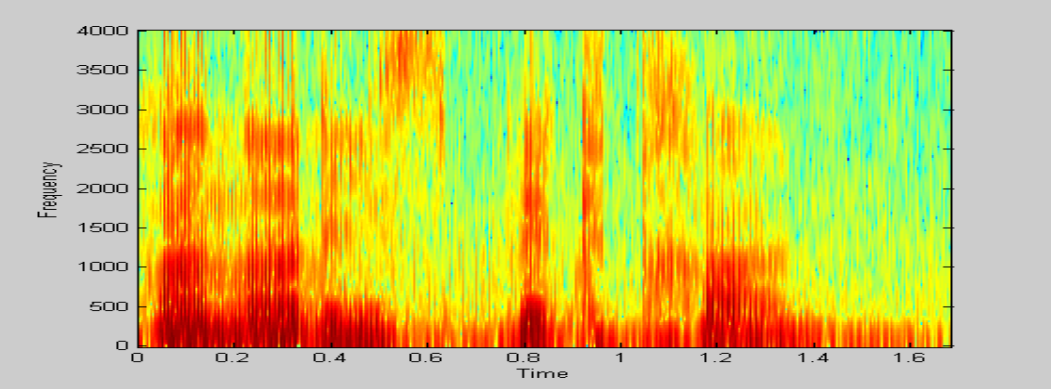

Figure 1.5 exemple d'un spectrogramme à large bande

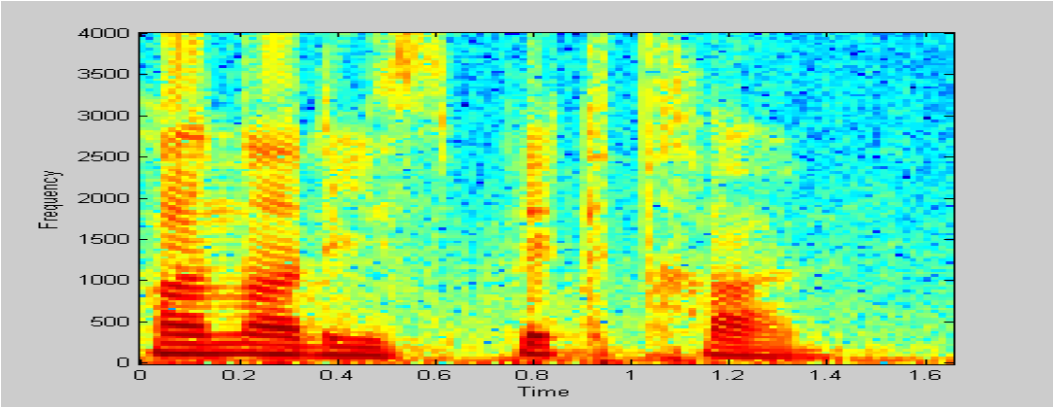

Figure 1.6 exemple d'un spectrogramme à bande étroite

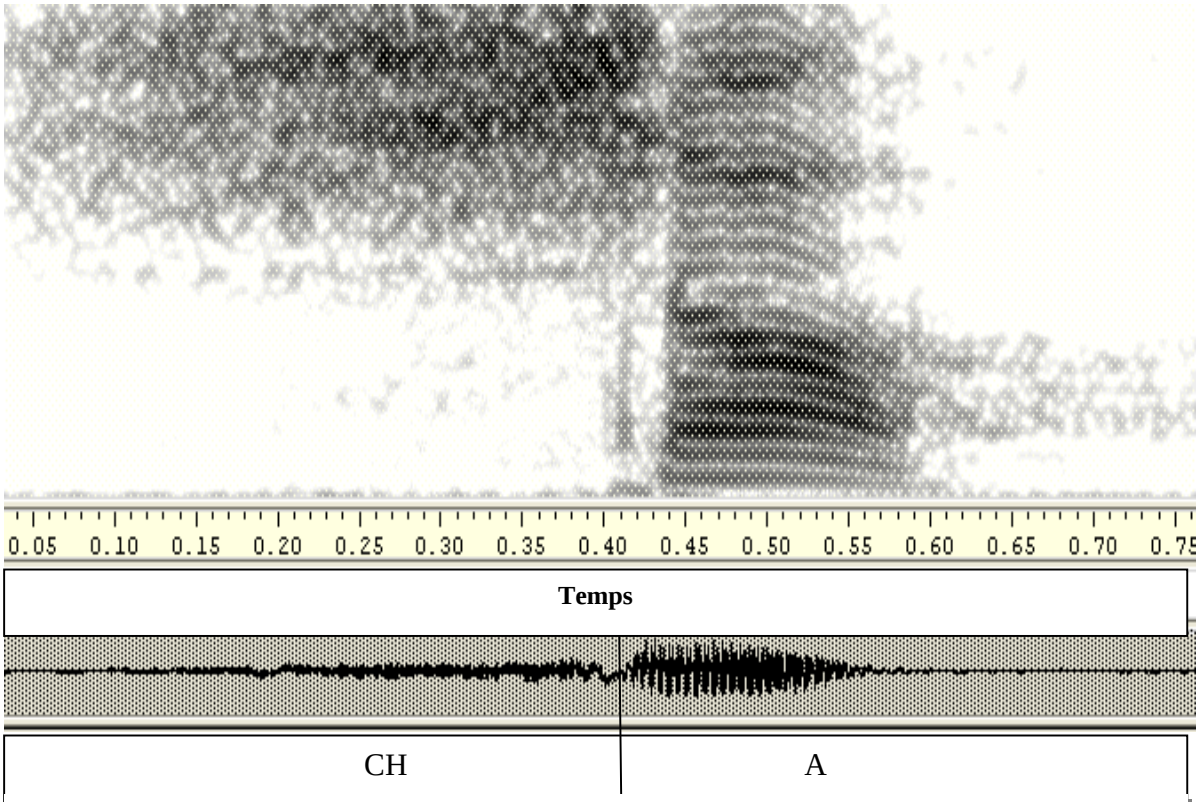

Fig.1.7 : Exemple d'un son non voisé et voisé prononcé par un locuteur masculin

1.2.2 Le Modèle filtre source

La production de la parole peut être modelée par le prétendu *modèle filtre source* [01,04]. Comme le nom suggère, le modèle considère le mécanisme de production de la parole comme

une combinaison de deux composants: *la source de la voix* et *le filtre acoustique* La " source " se rapporte au courant d'air produit par le larynx et le " filtre "se rapporte au conduit vocale. Tous les deux composants varient en temps et assumé d'être indépendant l'un de l'autre.

La Source de la voix :

Retournons au mécanisme de phonation. Selon le modèle filtre source, deux sources de voix sont possibles:

- un jet périodique des souffles d'air ce qui émerge de la vibration des cordes vocales comme il est décrit dans la section 1.1.2
- flux d'air turbulent non périodique qui résulte quand les cordes vocales sont ouvertes.

La source périodique de la voix est une caractéristique pour toutes les voyelles et nasale. La turbulence non périodique, est l'entrée acoustique pour les sons comme les fricatives [f] et [s]. Une classification brute des sons de la parole peut être basée sur le type de phonation, au lequel est référé comme *voisé* ou *non voisé* pour les deux types de source de la voix, respectivement.

Un exemple des sons voisé et non voisé est montré dans la fig. 1.7. Pour les sons voisés une structure quasi-périodique est apparemment présente dans la forme d'onde. La périodicité a comme conséquence *un spectre harmonique* dans lequel la plupart de l'énergie acoustique est distribuée sur des multiples de la fréquence fondamental, c.-à-d. KF_0 k = 1, 2, 3,…, .Par exemple, les harmoniques *de* F_0 peuvent être vu clairement dans le spectre de la fig. 2.4. Les sons non voisés n'ont pas la structure périodique, et leurs spectres sont non harmoniques et souvent répartissent dans les hautes fréquences.

Le Filtre acoustique

Dans le modèle de filtre source, le conduit est considéré comme filtre acoustique caractérisé par ses fréquences normales de résonance [01,04]. Dans le voisinage des résonances, les fréquences du signal de source sont amplifiées. Les sons voisés, ces maximums locaux du spectre sont appelés *des formants* et peuvent être vus dans les spectrogrammes en tant que secteurs foncés. Les formants sont numérotés comme F_1, F_2 et ainsi de suite. Pour la partie majeure des voyelles, les deux premiers formants soutiennent la partie majeure d'information phonétique [04].

Souvent le filtre acoustique est modelé comme tube résonateur *dur voilé* Fig. 1.8. Dans ce *modèle* le conduit vocal est considéré comme une cascade *de N tubes* avec changement des

sections. Pour ce genre de résonateur, les résonances peuvent être calculées analytiquement. Dans le cas d'un tube simple (N = 1), les résonances du tube (fréquences des formants) sont données par l'équation (1.1) [41]:

$$F_n = (2n -1)*c / 4*l \quad (1.1)$$

Où F_n est la nième fréquence de formant [hertz], c est la vitesse du son dans l'air [m/s], et l est toute la longueur du tube [m]. En effet le modèle tube simple prévoit assez les fréquences de formant de la voyelle neutre, puisque pendant son articulation, la section sur la long du conduit est approximativement constante. Pour un locuteur masculin adulte moyen (l=17.5 centimètre) les formants de la voyelle neutre seraient prévus par l'équation se produisant à 500 hertz, 1500 hertz, 2500 hertz et ainsi de suite.

La production d'autres voyelles peut être modelée par un modèle de trois tubes [02], où le premier tube de la glotte correspond à la cavité pharyngienne, le dernier tube correspond à la cavité buccale, et le tube moyen représente l'endroit de la constriction principale. L'endroit de la constriction définit la qualité phonétique de la voyelle.

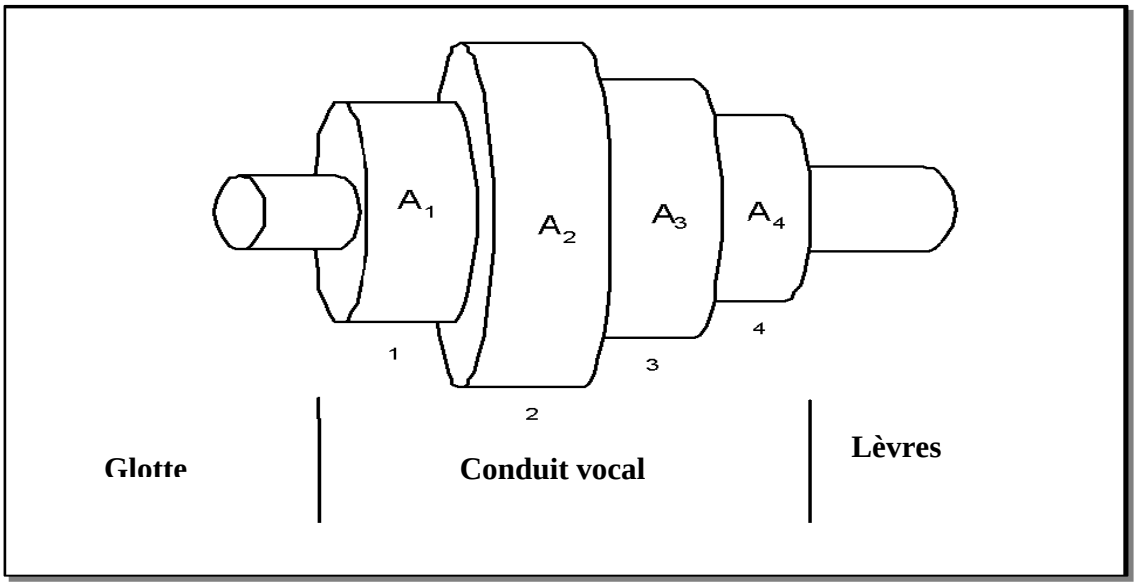

Figure 1.8 : Modèle multitube du conduit vocal

Le modèle de tube de la fig. 1.8 est au meilleur une approximation brute de la physiologie réelle du conduit, puisque :

- le conduit n'est pas une cascade des tubes durs voilés discrets mais des sections qui changent sans arrêt,
- des affaiblissements d'énergie dans le conduit dû à la vibration des murs de cavité, conduction de frottement et de chaleur.
- En plus, dans les bruits nasaux, il y a un tube latéral (cavité buccale fermée) qui n'est pas pris en considération dans le modèle

Spectres du modèle source filtre

Selon la théorie de source filtre, le spectre résultant d'une cascade de la source et le filtre est le produit de leurs spectres:

$$S(z) = U(z)\, H(z) \tag{1.2}$$

Où $S(z)$ est le spectre de la parole, $U(z)$ est le spectre de la source et $H(z)$ est la fonction de transfert du filtre de conduit vocal. En d'autres termes, le filtre souligne les fréquences de la source autour des résonances du conduit vocale. Ceci est illustré dans la fig. 1.9. Notez que le spectre de la source est responsable pour produire de la structure harmonique, tandis que la fonction de transfert du conduit modifie l'enveloppe du spectre global. Le point clé dans le modèle de source filtre est que la source et le filtre soient *indépendants* l'un de l'autre. Le désaccouplement rend la séparation des deux composants possible. Bien que cette solution soit raisonnable dans certains cas, ceci n'est pas vrai en général. Pendant la phonation, le rendement du larynx est affecté par le conduit, et cette interaction de la source et du filtre peut être vue particulièrement dans le cas quand le premier formant est bas [02]. Voyelles aux les quelles faites tendre *un bas* F_1 et une fréquence fondamentale plus élevée.

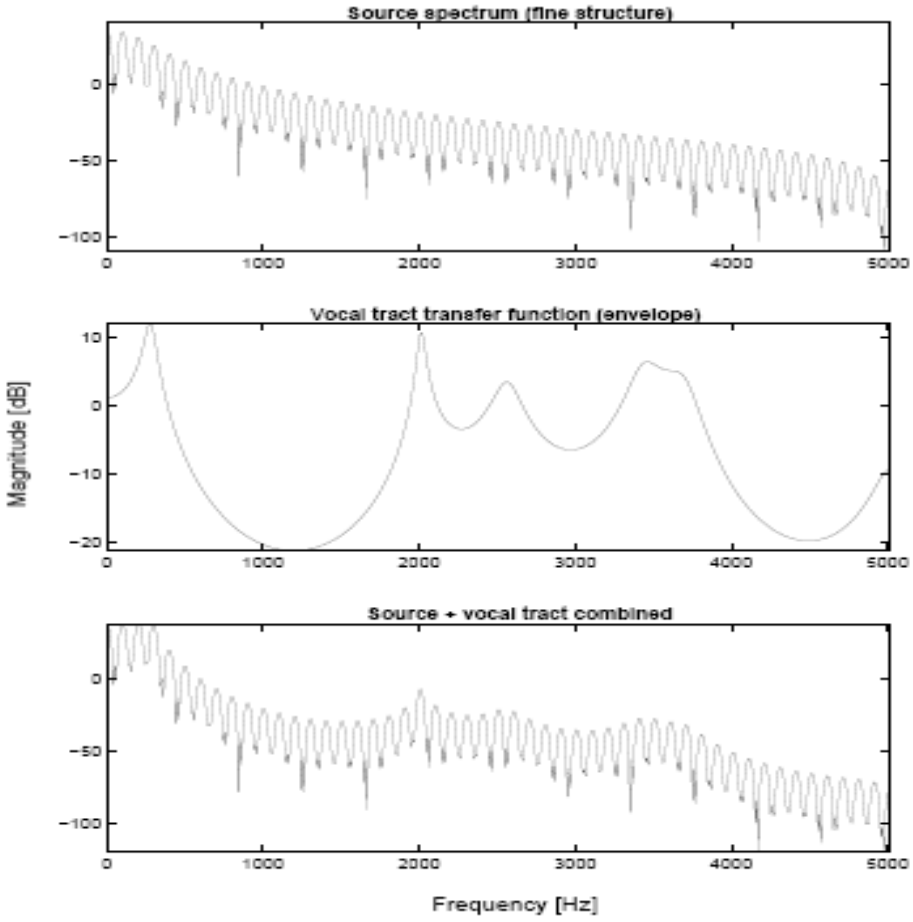

Figure 1.9 Spectre de la source et du filtre

1.3 Approche Perceptuelle

Le dernier point de vue de communication de la parole étudie *la perception* de la parole, c.-à-d. comment le système auditif de l'auditeur humain traite les sons de la parole. La discipline de la perception est désignée en général sous le nom *la psycho acoustique.*

Des techniques adoptées de la psycho acoustique sont intensivement employées dans l'acoustique et systèmes de traitement de la parole pour réduire la quantité des données perceptuelles non pertinent. La psycho acoustique vise à trouver des raccordements entre l'excitation auditifs objectivement mesurables, et l'impression individuelle approximativement ce que l'auditeur a au sujet des excitations.

Le volume d'un son n'est pas linéairement proportionnel à l'intensité mesurée. Par exemple, si l'intensité du son est doublée, elle n'est pas perçue " deux fois aussi fort " en général. *L'échelle décibel (dB)* est une manière plus commode pour décrire ce rapport. L'échelle décibel est l'un des moyens pour comparer l'intensité de deux sons [04]:

$$10 * \log 10 (I / I_0) \tag{1.3}$$

Où I_0 est l'intensité du son de référence étant comparés, par exemple, si l'intensité I est deux fois l'intensité I_0 de référence son niveau de dB est approximativement +3 dB. La fréquence fondamentale (F_0) est définie comme le taux auquel les cordes vocales vibrent pendant la phonation voisée. L'appel Psycho acoustique de F_0 est pitch. Même, si un son articulé est filtré de sorte que la région de fréquence fondamentale n'est pas présente dans le signal, le système auditif humain peut le percevoir [04].

L'oreille humaine traite la fréquence fondamentale sur une échelle logarithmique plutôt qu'une échelle linéaire, c –à- d F_0 doit changer plus pour qu'un auditeur humain puisse entendre une différence entre deux tonalités. *Le Mel* est une unité de la fréquence fondamentale perçue. Il était à l'origine déterminé par des essais d'écoute, et plusieurs modèles analytiques aient proposé pour rapprocher le Mel-mesuré. Par exemple, Fant a proposé la formule suivante :

$$F_{mel} = 1000 \log_2 (1 + \frac{F_{Hz}}{1000}) \tag{1.4}$$

Les amplitudes relatives des différentes fréquences déterminent la *forme du spectre global.* Si la fréquence fondamentale est maintenue et les amplitudes relatives des harmoniques supérieurs sont changées, le son sera perçues en tant que timbre *différent* Ainsi, le timbre est l'attribut perceptuel du forme spectrale, qui est connue pour être un dispositif important dans la reconnaissance de la parole Par exemple, le dispositif largement répandues pour la mesure de la forme spectrale perceptuelle est le mels-cepstre. Les études du mécanisme humain d'audition montrent que l'excitation d'entrée est fractionnée dans des plusieurs bandes de fréquence dans lesquelles deux fréquences ne sont pas distinguables. Ces bandes de fréquence désignées sous le nom des bandes *critiques* L'oreille fait la moyenne des énergies des fréquences dans chaque

bande critique et ainsi forme une représentation comprimée de l'excitation originale. Cette observation a donné l'impulsion pour concevoir des bancs de filtres perceptuels motivés comme entrées pour le système d'identification de la parole.

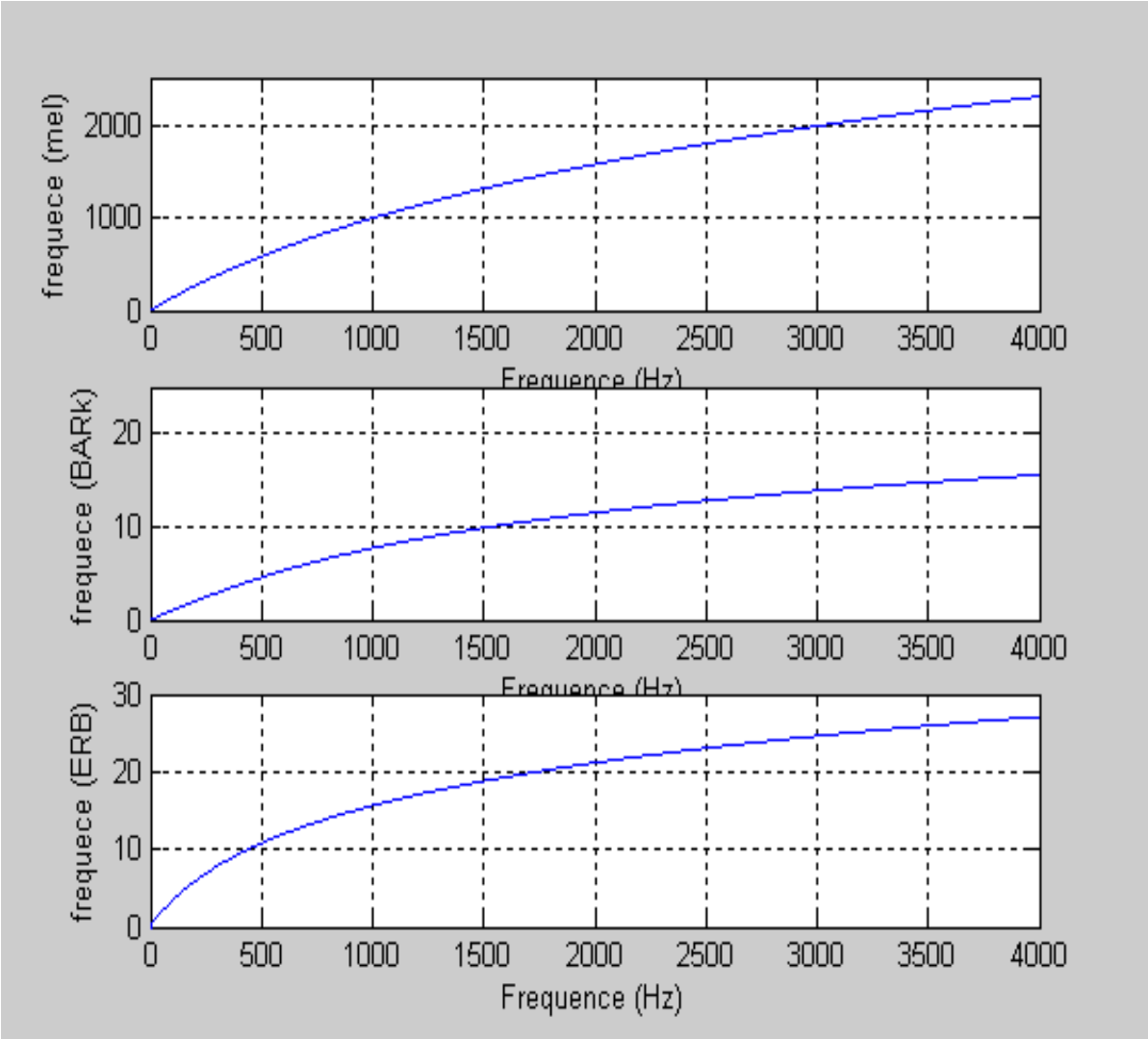

Figure 1.10 Les échelles mel, Bark et ERB

Beaucoup d'échelles d'approximations à la bande critique ont été proposé. Une échelle bien connue est *l'échelle Bark*. Qui est donné par plusieurs formules analytiques. L'une d'elles est proposée par Zwicker et Terhardt :

$$F_{Bark} = 13\tan^{-1}(\frac{0.76F_{Hz}}{1000}) + 3.5\tan^{-1}(\frac{F_{Hz}}{7500})^2 \quad (1.5)$$

Un autre exemple d'approximation Bark est le suivant:

$$F_{Bark} = 6\sinh^{-1}(\frac{F_{Hz}}{600}) \quad (1.6)$$

En plus de l'échelle Bark une autre approximation de bande critique est l'échelle *ERB* (ERB =Equivalent Rectangular Bandwidth of the auditory filter), qui est défini comme suit:

$$ERB = 21.4\log_{10}(1 + \frac{4.37F_{Hz}}{1000}) \quad (1.7)$$

Les échelles Mels, Bark et ERB sont tracés dans la Fig. 1.10. Les formes des courbes sont différentes, mais le message de chacune des trois est identique. Dans la région des hautes fréquences, deux excitations différentes doivent avoir une plus grande différence pour que l'oreille humaine les distingue. Dans les fréquences plus basses, la résolution spectrale de l'oreille humaine est plus haute.

L'exigence qui nous pousse à employer les représentations psycho acoustique est que l'oreille humaine est un système de reconnaissance optimal.

Conclusion :

Dans ce chapitre on a vu les trois perspectives articulatoire, acoustique et perceptuelle. Dans l'approche articulatoire on essaye de savoir le rôle de chaque organe pour décrire comment les sons de la parole sont produits en termes des gestes articulatoires, tandis que l'approche acoustique cherche des corrélations entre le signal acoustique et les organes de production de la parole pour trouver des modèles qui approche au mieux le système articulatoire humain . Dans l'approche perceptuelle on essaye de savoir comment le système humain traite les signaux acoustiques par l'utilisation des techniques psychos acoustiques comme les échelles Mel, Bark et ERB.

Traitement et Analyse du Signal Vocal

Ce chapitre donne une brève vue d'ensemble des méthodes de transformation des signaux numériques utilisées dans l'extraction des dispositifs. Un traitement complet du sujet peut être trouvé en général dans les livres de DSP comme [06]

2.1 Conversion Analogique Numérique

Un son articulé est une forme de mouvement d'onde portée par un milieu (par exemple les particules d'air) [04], et peuvent être capturé par un microphone, qui convertit le changement continu de la pression atmosphérique en changements continu de tension. Le signal analogique s_a (t) est alors échantillonné à une forme numérique s [n] par *un convertisseur A/D*. Le convertisseur prélève le signal analogique uniformément avec *la période d'échantillonnage T :*

$$s[n] = s_a(nT) \qquad (2.1)$$

L'inverse *de T* est *la fréquence d'échantillonnage* (ou *taux d échantillonnage)* est notée par $F_s = 1/T$. Étant donné que le signal original s (t) contient des fréquences seulement jusqu' *à* $F_s/2$, il peut être entièrement reconstruit a partir des échantillons s[n] [05]. La fréquence $F_s/2$ est la limite supérieure pour des fréquences présentent dans le signal numérique. Par exemple, si on veut préserver des fréquences jusqu' à 4 kHz, le tau d'échantillonnage doit être choisi *tel que* $F_s > 8$ kHz. En plus de l'échantillonnage, le CAN *quantifie* les échantillons dans une précision finie. Le nombre de bits utilisés par échantillon détermine la gamme dynamique du signal.

2.2 L'analyse De Fourier

L'analyse de Fourier fournit une manière d'analyser *les propriétés spectrales d'un* signal donné dans le domaine fréquentiel.

Les outils d'analyse de Fourier considèrent un signal comme superposition *des fonctions* sinusoïdales de base de différentes fréquences, phases et amplitudes. L'exemple de la fig.2.1

montre trois sinusoïdes et leur superposition (somme). L'analyse de Fourier fournit un outil pour trouver les paramètres des sinusoïdes fondamentaux (Transformée directe) ou pour synthétiser le signal original domaine - tems d'après la présentation du domaine – fréquence (*Transformée inverse).*

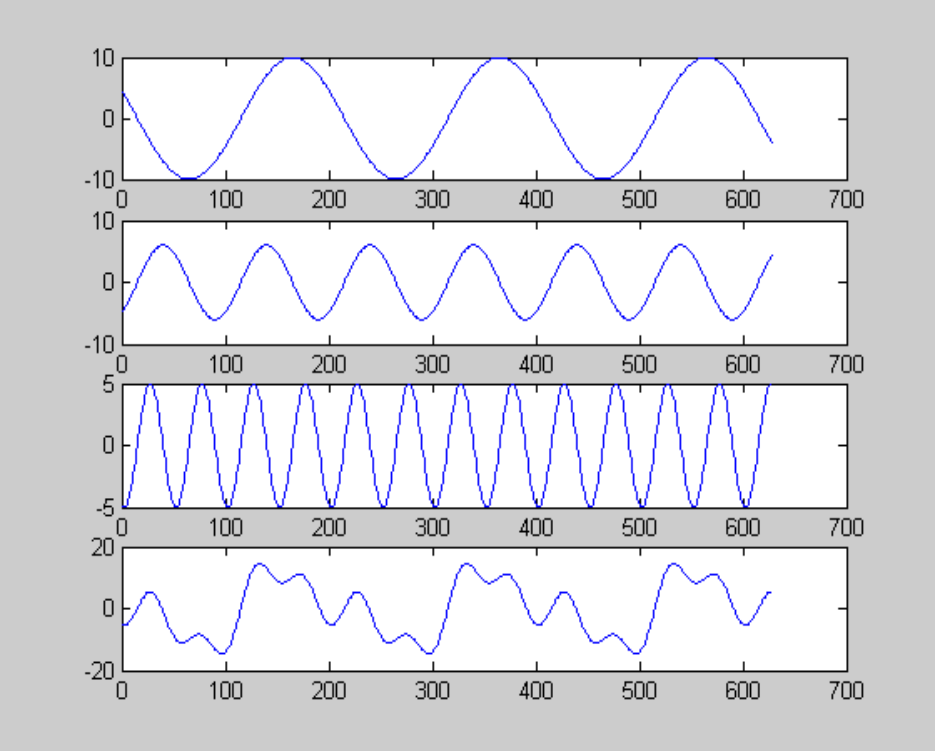

Fig. 2.1 Trois sinusoïdes et leur somme

2.2.1 La Transformée de Fourier Discrète (TFD ou DFT)

Supposons que $s[n]$, $n = 0,1,\ldots..N-1$ une séquence de temps discret *de N échantillons.*
La transformée de Fourier discrète ou TFD de s [n*]* est définie comme suite :

$$\hat{S}[k] = f\{s[n]\} = \sum_{n=0}^{N-1} s[n] e^{-j2\pi nk/N}, 0 \le k \le N-1, \qquad (2.2)$$

Où k représente la variable discrète de fréquence et j l'unité imaginaire. Le résultat du TFD est un nombre complexe de longueur *N. la* valeur $k = 0$ ($w_k = 0$) correspond *à la fréquence nulle* ou la *composante continue du* signal. L'inverse de la TFD ou ITFD est défini comme suite :

$$s[n] = f^{-1}\{\hat{S}[n]\} = \frac{1}{N}\sum_{k=0}^{N-1} \hat{S}[k] e^{j2\pi nk/N}, 0 \le n \le N-1, \quad (2.3)$$

Où n représente la variable discrète du temps. En d'autres termes, le signal original peut être reconstruit de sa transformée de Fourier par la transformée inverse.

DFT et IDFT sont des transformations linéaires, c.-à-d.

$$f\{\alpha s_1[n] + \beta s_2[n]\} = \alpha f\{s_1[n]\} + \beta f\{s_2[n]\}$$

$$f^{-1}\{\alpha s_1[n] + \beta s_2[n]\} = \alpha f^{-1}\{s_1[n]\} + \beta f^{-1}\{s_2[n]\}$$

3.2.2 Les spectres d'amplitude et de phase

Le $k^{ième}$ composant harmonique de la TFD est un nombre complexe $\hat{S}[k] = \hat{S}_{\text{Re}}[k] + j\hat{S}_{\text{Im}}[k]$. Il peut être exprimé sur un forme polaire comme $\hat{S}[k] = |\hat{S}[k]| e^{j\sphericalangle \hat{S}[k]}$, ou

$$|\hat{S}[k]| = \sqrt{\hat{S}_{\text{Re}}[k]^2 + \hat{S}_{\text{Im}}[k]^2} \qquad \sphericalangle \hat{S}[k] = \tan^{-1}\left(\frac{\hat{S}_{\text{Im}}[k]}{\hat{S}_{\text{Re}}[k]}\right).$$

$\hat{S}[k]$ Est *l'amplitude* et $\sphericalangle \hat{S}[k]$ est *la phase du $k^{ième}$ composant* harmonique. TFD est périodique avec N c.-à-d. $\hat{S}[k+N] = \hat{S}[k]$ Pour des signaux réels comme la parole, le spectre d'amplitude est symétrique par rapport à la fréquence $N/2$. En outre, le spectre de phase pour des signaux réels est antisymétrique par apport à la fréquence $N/2$. En raison de cette redondance, tout signal à valeurs réelles est entièrement représenté par les composants harmoniques jusqu' à $N/2$. Dans l'analyse de la parole, le spectre de phase est habituellement négligé, puisqu'il a généralement moins d'effet sur la perception de la parole [02].

2.2.3 Transformée de Fourier Rapide (TFR ou FFT)

De la définition (2.2) il est facile de voir que la complexité de temps de DFT est (N^2). Cependant, la DFT peut être calculé par l'intermédiaire d'un algorithme plus rapide appelé la *transformée de Fourier rapide* ou *FFT* [05]. Une condition pour FFT est que le signal d'entrée l (vecteur) doit

avoir une longueur de 2^M tel que $M \in$ N + c.-à-d une puissance de deux. Dans le pratique, le signal d'entrée est d'abord chargé par des zéros jusqu à la prochaine puissance de deux et le signal chargé par des zéros est donné comme entrée pour l'algorithme FFT. Par exemple, si la longueur du signal est 230 échantillons, il est chargé par des zéros à la longueur N = 256 pour que l'algorithme FFT puisse le calculé. Des zéros peuvent être ajoutés au commencement ou l'extrémité du signal, sans que le résultat du DFT soit affecté. La complexité du FFT est ($N \log_2$ N). Le gain dans le temps de calcul dans le pratique est très important Par exemple, pour N = 1024, le rapport des multiplications de DFT aux multiplications de FFT est environ 200 et le rapport des additions environ 100 [05].

2.3 Filtres Numériques

Un filtre est un système qui modifie le signal d'entrée s [n] en un signal de sortie y [n] [06]. Il y a plusieurs manières d'indiquer un filtre numérique. Dans le domaine temporel, le filtre est caractérisé par sa *réponse* impulsionnelle h[n] qui peut être fini (Filtre *RIF*) ou infini (filtre *RII)*. Dans le domaine fréquentiel, un filtre est spécifié par sa *fonction de transfert H* (z), où z est une variable complexe. Dans le domaine temporel, le filtrage est présenté comme *convolution* entre le signal d'entrée et la réponse impulsionnelle h [n] [05,06]:

$$y[n] = s[n] * h[n] = \sum_{k=-\infty}^{\infty} s[k]h[n-k] \tag{2.4}$$

Dans la pratique, ceci est mis en application en utilisant un rapport récursif [44]:

$$y[n] = \sum_{k=0}^{N} a[k]s[n-k] - \sum_{k=1}^{M} b[k]y[n-k] \tag{2.5}$$

Où les coefficients a [k], b [k] sont déterminés à partir des caractéristiques du filtre.

La dernière somme dans (2.5) représente la partie de rétroaction du filtre et il est nul pour les filtres RIF (b [k] = 0 pour tout k). *La fonction de transfert H* (z) de (2.5) est obtenu par prendre la transforme Z des deux côtés et la résoudre pour H (z)=Y (z) / S (z):

$$H(z) = \frac{Y(z)}{S(z)} = \frac{\sum_{k=0}^{N} a[k] z^{-k}}{1 + \sum_{k=0}^{M} b[k] z^{-k}} \tag{2.6}$$

On appelle les racines du numérateur de (2.6) *les zéros* du système, et les racines du *les pôles* du système. Un pole cause *une résonance* (crête) dans la réponse d'amplitude du filtre, tandis qu'un zéro cause *une anti-résonance* (vallée). Par exemple, la fonction de transfert du conduit vocal d'un son de voyelle peut être bien caractérisée seulement par des pôles, qui correspondent aux endroits des formants. D'autre part, les sons nasaux comme [n] ayant en plus des résonances, des anti-résonances dans leur spectre, et donc les pôles et les zéros sont nécessaires dans la modélisation [04].

Dans le domaine fréquentiel, le filtrage est effectué en multipliant point à point la DFT du signal d'entrée par la fonction de transfert du filtre. S'accorder *au théorème de convolution* [06], multiplication dans le domaine fréquentiel correspond à la convolution dans le domaine temporel, et vice versa:

$$s[n] * h[n] \leftrightarrow S(z).H(z) \tag{2.7}$$

$$s[n].h[n] \leftrightarrow S(z) * H(z) \tag{2.8}$$

3.4 L'analyse Spectrale courte terme

Puisque le son de la parole change en continu en raison des mouvements articulatoires des organes vocaux de production, le signal doit être traité avec des petits segments, dans lesquels les paramètres demeurent quasi stationnaires (Fig.2.2). Le calcul de la DFT du signal entier jetterait les propriétés spectrales locales qui présentent des réalisations de différents phonèmes. Au lieu d'exécuter la DFT pour le signal entier, une fenêtre *DFT* est calculée. Un segment en général autour de 10-30 millisecondes, est multipliée par *une fonction fenêtre* et la DFT du segment fenêtré est alors calculée Ce processus est répété jusqu à la fin du son articulé de sorte que le segment soit décalé en avant par une quantité fixe des points, en général autour 30 à 75 % de la longueur du segment.

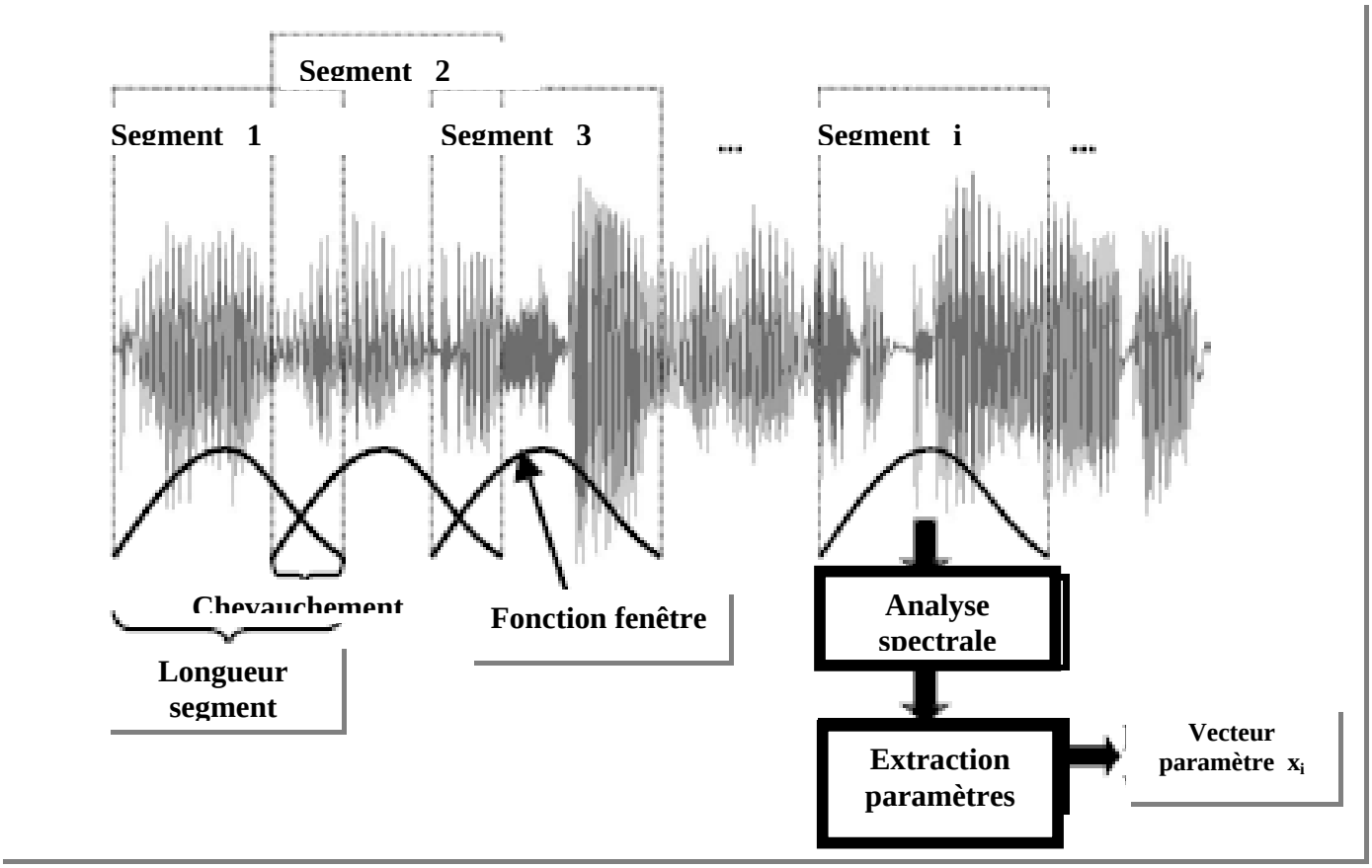

Figure 2.2 : l'analyse spectrale courte terme

2.4.1 La Fonction Fenêtre

Le but du fenêtrage est de réduire l'effet résultant du processus de segmentation [04,05]. Fenêtrage dans le domaine temporel est une multiplication point par point entre le segment et la fonction fenêtre. Selon *le théorème de convolution* [06], ceci correspond à une convolution du spectre courte terme avec la réponse d'amplitude de la fonction fenêtre. En d'autres termes, la fonction de transfert de la fenêtre sera présente dans le spectre observé. Une bonne fonction fenêtre a un *lobe principal étroit* et des petits lobes secondaires [05] dans sa fonction de transfert. Il y a une compensation entre ces deux conditions : rendre le lobe principal plus étroit augmente le niveau des lobes secondaires, et vice versa. En général, une fonction fenêtre appropriée diminue aux bords de segment de sorte que l'effet des discontinuités est diminué.

Intuitivement le fenêtrage le plus simple est " aucun fenêtrage ", ou *fenêtre rectangulaire* définie comme suit:

$$w[n]=\begin{cases}1, & 0\le n\le N-1\\ 0, & autrement\end{cases} \tag{2.9}$$

Bien que la fenêtre rectangulaire préserve la forme d'onde originale sans changement, elle est rarement utilisée en raison de ses effets *spectraux*. Généralement dans le traitement de la parole la fonction fenêtre la plus utilisée est *la fenêtre de Hamming* définie comme suite [07,08]:

$$w[n]=\begin{cases}0.54-0.46\cos\dfrac{2\pi n}{N}, & 0\le n\le N-1\\ 0, & autrement\end{cases} \tag{2.10}$$

Les formes du domaine temporel et réponses d'amplitude des fenêtres rectangulaire et Hamming (calculées en utilisant DFT) sont montrées dans la Fig. 2.3.

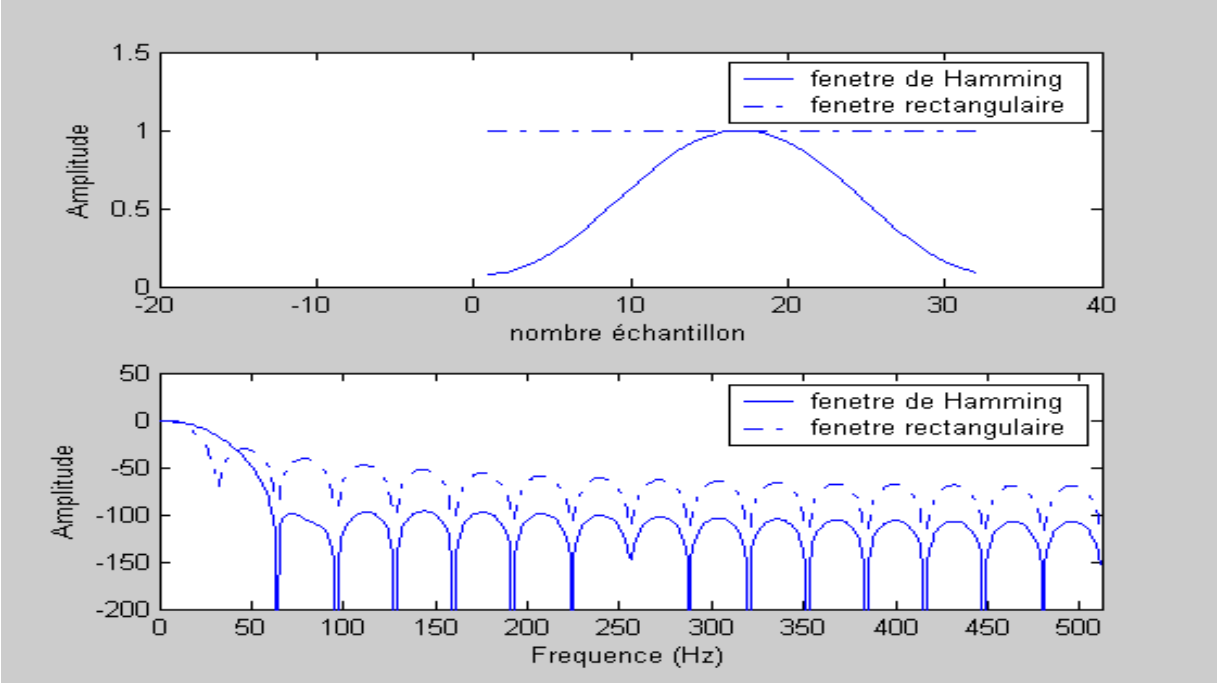

Figure.2.3 formes du domaine temporelle et réponses fréquentielle des fenêtres rectangulaire et Hamming

Des exemples du fenêtrage sont montrés sur les figures 2.4 et 2.5 qui montre des segments voisés et non voisé de parole prononcé par le même locuteur. On peut voire d'après le segment voisé, que la fenêtre de Hamming donne moins de fuite spectrale. La fenêtre rectangulaire cause aux harmoniques *de* F_0 des harmoniques voisins, et comme résultat, on lisse les différents harmoniques.

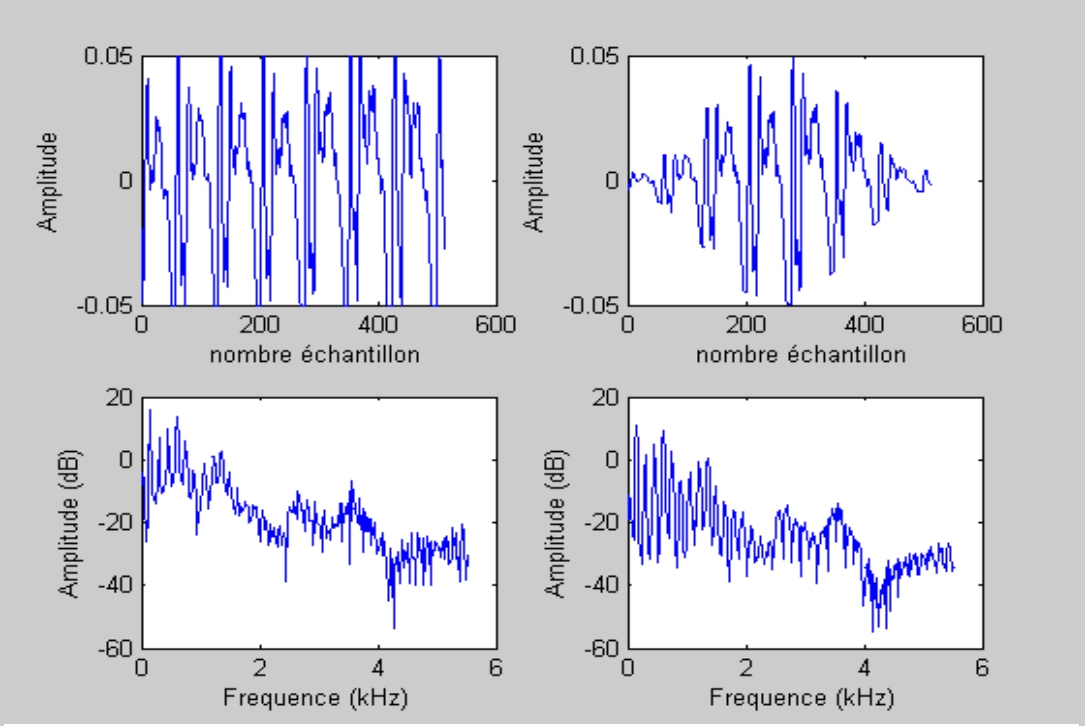

Figure2.4 : Segment d'un son voisé [voyelle a] fenêtré à gauche par une fenêtre rectangulaire et à droite par une fenêtre de Hamming

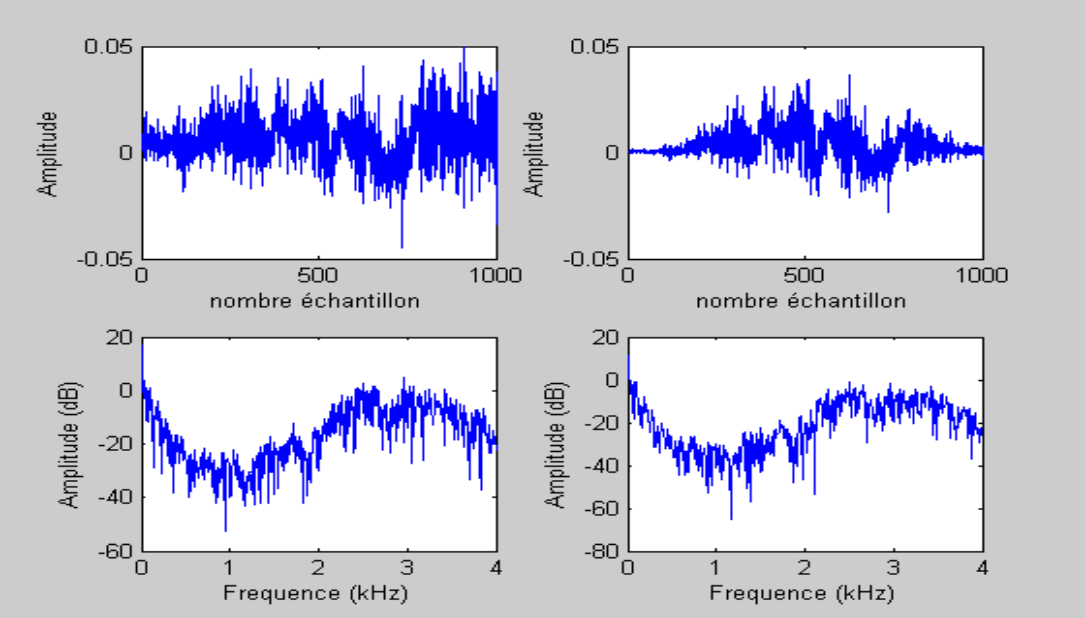

Figure2.5 : Segment d'un son non voisé [ch] fenêtré à gauche par une fenêtre rectangulaire et à droite par une fenêtre de Hamming

Pour la fenêtre de Hamming, les différents Harmoniques peuvent être vus facilement. Le segment non voisé dans la fig. 2,5 ne contient pas des harmoniques, mais la fuite spectrale peut encore être vue.

3.4.2 Longueur et chevauchement des fenêtres

Le choix de longueur pour la fenêtre est un paramètre décisif pour une bonne analyse spectrale, due à la compensation entre les résolutions temps et fréquence. La fenêtre devrait être assez longue pour une résolution fréquentielle stable, mais d' autre part, elle devrait être assez courte de sorte qu'elle capture les propriétés spectrales locales. Typiquement une durée de 10-30 millisecondes est utilisée [09]. Pour des femelles et des enfants, le pitch tend à être plus haut, et une fenêtre plus courte employé que pour les locuteurs masculins qui ont un pitch moins petit [04]. Habituellement les fenêtres adjacentes recouvrent par une certaine quantité. Un chevauchement typique des fenêtres est autour de 30 à 50 % de la taille fenêtre. Le but de l'analyse par recouvrement est pour que chaque son parole de la séquence d'entrée serait approximativement centré à une certaine fenêtre.

2.5 La Préaccentuation :

Habituellement la parole subit une *préaccentuation* avant une transformation plus ultérieure. Préaccentuation se rapporte à un filtrage qui renforce les hautes fréquences. Son but est d'équilibrée les sons voisés qui ont un tendu glottale dans la région de haute fréquence. Pour les sons voisés, la source glottale a approximativement une pente de -12dB/octave [02]. Cependant, quand l'énergie acoustique rayonne des lèvres, ceci cause une poussée d'approximativement +6 dB/octave au spectre. Un résultat net, un son voisé une fois enregistré avec un microphone, a approximativement une pente de -6 dB/octave en bas comparée au spectre vrai (du conduit). Par conséquent, la préaccentuation enlève les effets glottaux des paramètres du conduit. Puisque le spectre des sons non voisés est déjà plat, il ni pas raison de la préaccentuation [02].

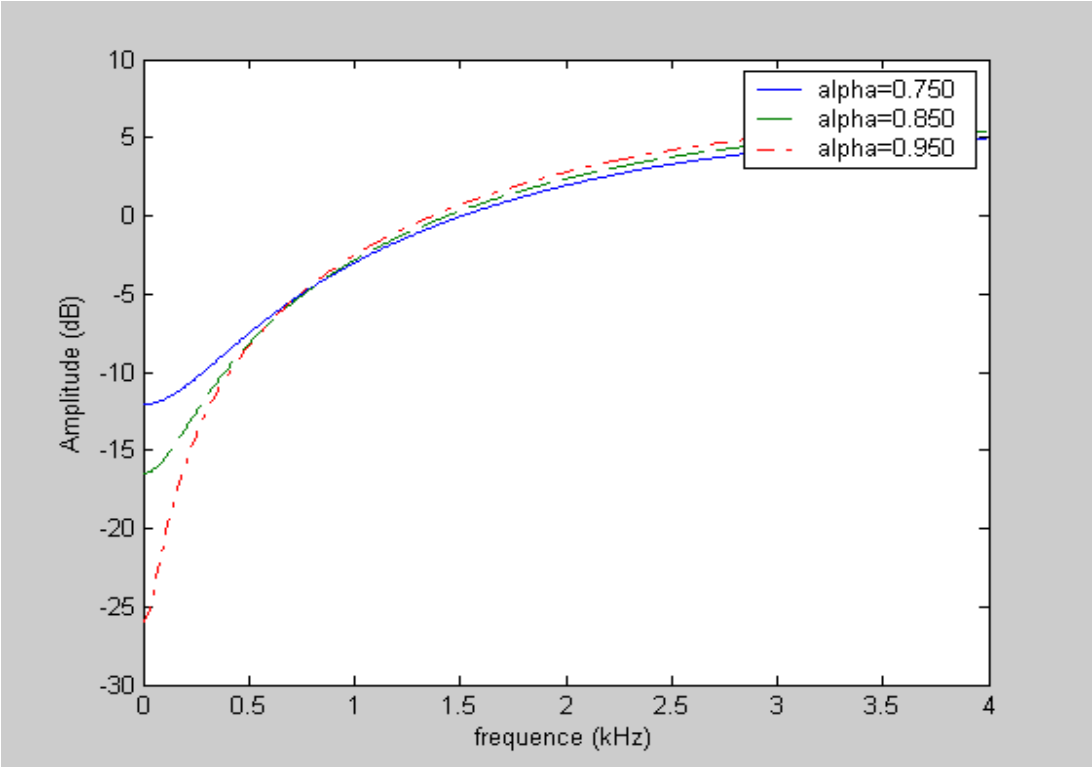

Figure 2.6 : Réponse d'un filtre de préaccentuation pour des déférentes valeurs d'alpha

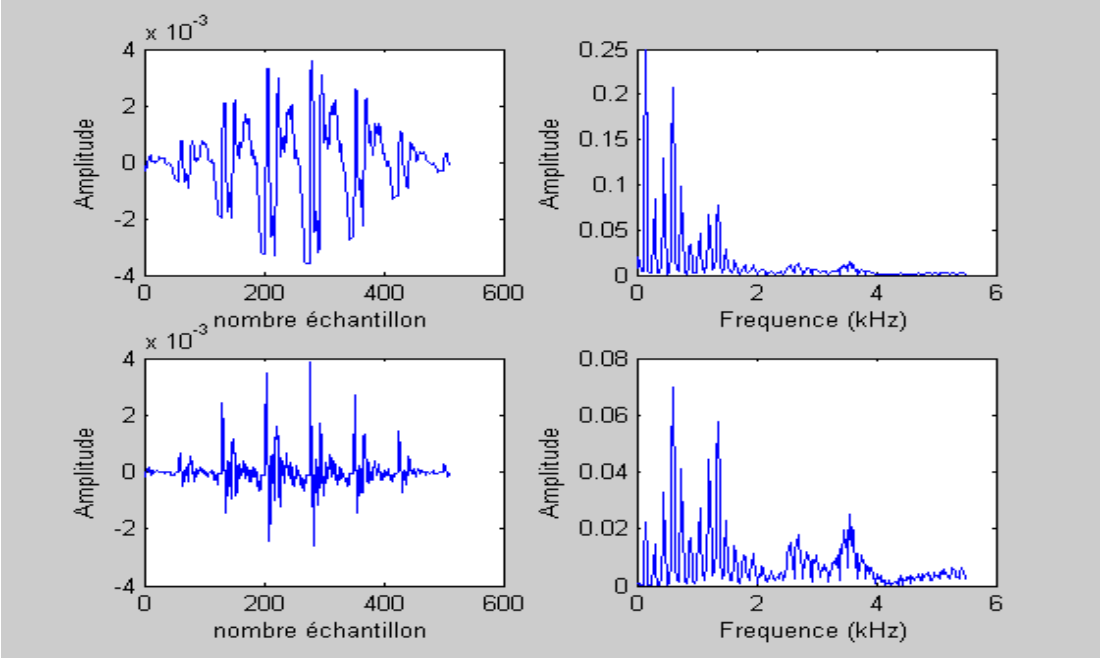

Figure 2.7 : Exemple de préaccentuation d'un segment de signal parole

La préaccentuation a d'autres avantages [10]. Par exemple la stabilité numérique de l'analyse prédictive linéaire (LP) est inversement proportionnelle à la gamme dynamique du spectre analysé par LPC. Par conséquent, un filtre qui écrase le spectre devrait être utilisé avant le LPC pour évitez les problèmes numériques, et est ce que le filtre de préaccentuation le réalise.

Le filtre de préaccentuation le plus généralement utilisé est donné par la fonction de transfert suivante [04, 07,08] :

$$H(z) = 1 - \alpha z^{-1} \tag{2.11}$$

Où $\alpha \rangle 0$ commande la pente du filtre. La réponse impulsionnelle du filtre est $h[n] = \{1, -\alpha\}$ et le filtre est simplement mis en application comme un différentiateur de premier ordre :

$$y[n] = s[n] - \alpha s[n-1] \tag{2.12}$$

La réponse fréquentielle du filtre est :

$$\begin{aligned} H(e^{-j\omega}) &= 1 - \alpha e^{-j\omega} \\ &= 1 - \alpha(\cos\omega - j\sin\omega) \end{aligned}$$

Par conséquent, la réponse du carrée d'amplitude :

$$\begin{aligned} \left|H(e^{j\omega})\right|^2 &= (1 - \alpha\cos\omega)^2 + \alpha^2\sin^2\omega \\ &= 1 - 2\alpha\cos\omega + \alpha^2\cos^2\omega + \alpha^2\sin^2\omega \\ &= 1 - 2\alpha\cos\omega + \alpha^2(\cos^2\omega + \sin^2\omega) \\ &= 1 - 2\alpha\cos\omega + \alpha^2 \end{aligned} \tag{2.13}$$

Les réponses d'amplitude dans l'échelle dB pour des différentes valeurs *de* α sont montrées dans la Fig.2.6, un exemple d'un segment préaccentué dans le domaine temporelle et fréquentielle est montré dans la Fig.2.7. Notez que la préaccentuation rend les hautes harmoniques de F_0 plus distinctes, et la distribution de l'énergie à travers la gamme de fréquence plus équilibrée

2.6 Banc de filtre

Banc de filtre est une limite générique qui se rapporte à la classe des méthodes qui traitent le signal sur des bandes de fréquence multiples. Banc de filtre et traitement sous bande se réfèrent plus ou moins au même concept, et nous les emploierons l'un pour l'autre. Une autre branche du

traitement des signaux utilise le traitement sous bande parallèle est l'analyse par wavelet [11]. L'extraction des paramètres à base du Wavelet- a été exercée dans la reconnaissance de la parole [12,13]. Bien que les wavelets puissent fournir une meilleure représentation de signal que la DFT à courte terme, plusieurs questions sont en pause concernant le choix *de la wavelet* mère *et de la structure de décomposition wavelet* et de l'extraction des paramètres de la transformée wavelet [14]. Des exemples de deux réponses d'amplitude différentes de banc des filtres sont montrés dans la Fig.2.8 Dans les deux cas, les filtres sont linéairement espacés sur la gamme fréquence 0-4 kHz. Dans les deux figures, le filtre qui analyse la bande 2000-2500 hertz est souligné pour expliquer qu'un filtre donné dans le banc des filtres à une réponse zéro en dehors de son passe bande. Pour créer un banc des filtres on doit décider si le banc est mis en application dans le domaine temporel par un ensemble d'équations récursives (2,5) ou dans le domaine fréquentielle en multipliant le spectre du signal avec l'amplitude de la réponse du filtre. Dans la réalisation du domaine temporel, l'extraction des paramètres peut se faire pour chaque bande du signal on utilisant un traitement régulier segment par segment. Un avantage de ceci est que chaque sou bande peut être traitée par les mêmes techniques comme pour toute la bande du signal. Ceci à une conséquence pratique que la résolution de chaque sou bande peut être commandé plus facilement que dans le traitement sur la bande entière. Supposez qu'*un N* points du spectre d'amplitude S[j], j = 1,…N est produit par DFT à courte terme. Et Supposez une chaîne de M filtre, dont les échantillons de sa réponse d'amplitude est rangées dans des tableaux H_i [j], i = 1...M ; j=1,…N

La sortie du $i^{ème}$ *filtre* $Y[i]$ est donné par :

$$Y[i] = \sum_{j=1}^{N} S[j] H_i[j] \qquad (2.14)$$

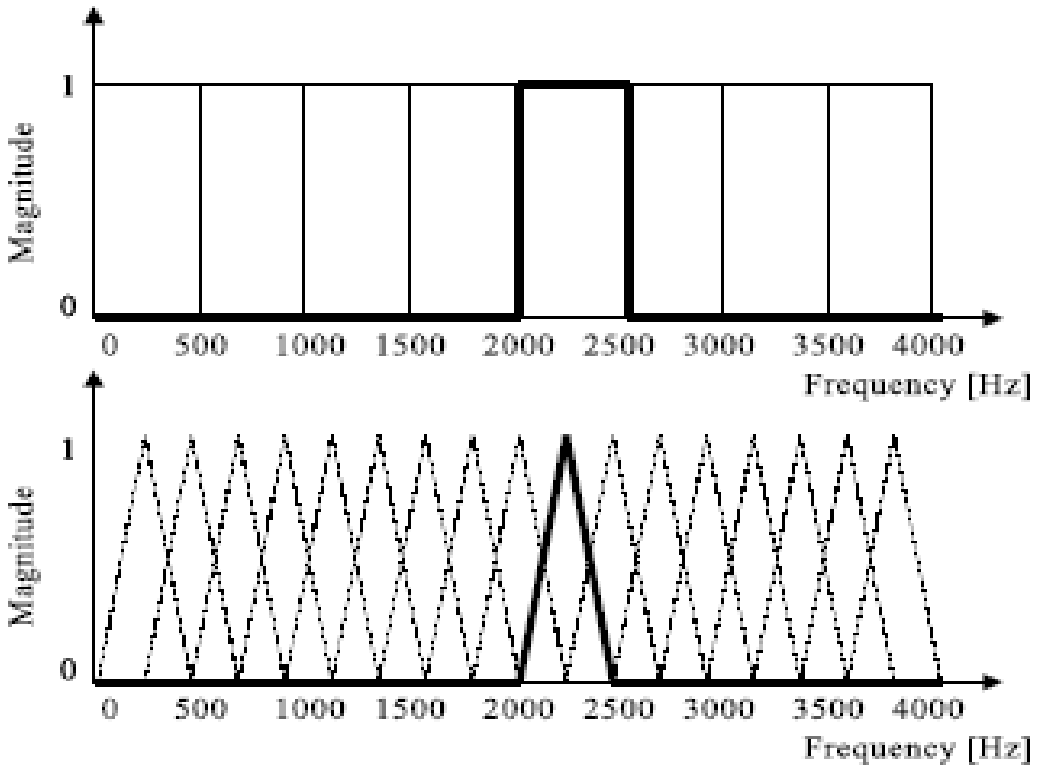

Figure 2.8 deux bancs de filtre linéaire d'une forme rectangulaire et triangulaire

En d'autres termes, le rendement *du* $i^{ème}$ *canal* est le rendement des amplitudes de la DFT dans cette région de fréquence pesée par la réponse du filtre. Par cette manière, le banc de filtre fournit une version douce du spectre original avec M<N *Composants*.

Notez que la DFT du segment d'entrée nécessite un calcule en ligne, les réponses du banc de filtre doivent être calculées seulement quand le système de reconnaissance est initialisé. Il est également intéressant de noter que dans la pratique seulement les éléments différents de zéro des réponses de filtre doivent être stockés.

Une propriété souhaitable d'un banc de filtre est que la somme de ses réponses d'amplitude égale l'unité à chaque bande de fréquence,

C.-à-d. Pour tout *j.*

$$\sum_{i=1}^{M} H_i[J] = 1 \qquad (2.15)$$

Ceci assure que la gamme de fréquence entière d'intérêt est traitée avec une signification égale. En raison de la précision finie et du nombre fini des échantillons du banc il pourrait être difficile d'atteindre (2.15) dans la pratique.

Les fréquences centrales des filtres sont également souvent espacées sur un certain axe de fréquence. L'axe peut être linéaire comme dans la Fig. 2.8, ou *déformé* selon certains fonction non linéaires comme les échelles Mels, BARK ou ERB montrés dans la Fig.1.10. Par la déformation de fréquence, on peut ajuster la quantité de résolution qui est désirée autour d'une certaine fréquence. L'idée dans l'analyse par wavelet [14] est presque la même. Il y a deux approches pour utiliser la fréquence déformée dans l'analyse spectrale: *approches paramétriques* et *non paramétriques*. Dans le 1^{ier}cas, la fonction de déformation est directement branchée à un modèle paramétrique de signal tel que la *prédiction linéaire généralisé.* Dans le 2ème cas, la fonction de déformation est prélevée dans plusieurs points de façon finie, et les points prélevés représentent les localisations du filtre. Cette approche est employée avec les bancs de filtres qui utilisent une implémentation DFT. L'idée de la déformation non paramétrique dans la conception des bancs de filtres est illustrée dans la Fig. 2.9. Un nombre désiré de fréquences centrales de filtre sont placés linéairement sur *un* axes w' (tel que des mels). Puis, un replacement inverse est employée w' → w pour résoudre les fréquences centrales dans l'axe fréquence (hertz). Ainsi, la fonction de déformation doit être bijective ainsi qu'elle peut être inversée uniquement.

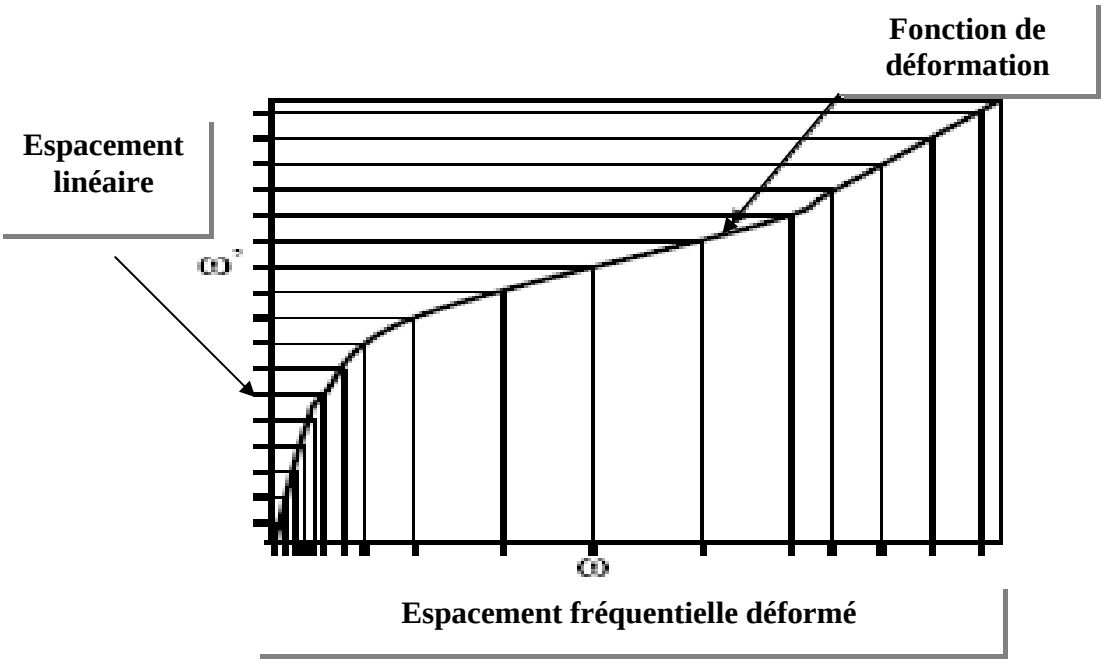

Figure 2.9 Principe de la déformation fréquentielle

2.7 Classification de modèle statistique

Le but de la classification automatique est de déterminer automatiquement l'identité des données non classifiées. La classification automatique de modèle prend place en deux étapes, apprentissage et classification de modèle. La figure 2.10 donne une vus graphique globale du processus [15].

Avant que l'apprentissage ou la classification de modèle puisse avoir lieu, les données doivent être codées en termes de paramètres représentatifs. Le processus de codage est désigné habituellement sous le nom mesure de paramètre ou l'extraction des paramètres. Typiquement, les paramètres sont choisis tels qu'ils représentent les différences appropriées entre les classes des données tout en masquant les variations dans la même classe.

Pendant l'apprentissage du modèle, le classificateur apprend les propriétés des données d'un modèle spécifique qui sont fiables et répétées à travers les données d'apprentissage. Dans la classification statistique automatique, la " connaissance " qui est rassemblée pendant l'apprentissage est représentée sur des modèles statistiques. Un modèle est formé pour chacune

des classes dans un ensemble spécifique de donné. Typiquement, une classe dans les données est connue a priori et chaque empreinte dans l'ensemble d'apprentissage est marquée selon son identité de classe. Afin d'établir le bon classificateur, les données d'apprentissage devraient contenir un nombre stable d'empreinte d'apprentissage pour chaque classe.

Pendant la phase de classification, les modèles inconnues de test sont comparées avec les modèles construits dans l'apprentissage, une mesure de similarité (distance statistique) entre le modèle de teste et chaque modèle dans la base des connaissances est calculée. La décision logique est utilisée pour décider qui parmi les modèles donne le meilleur score pour le modèle de test. La classe du meilleur modèle et donc assignée au modèle de test.

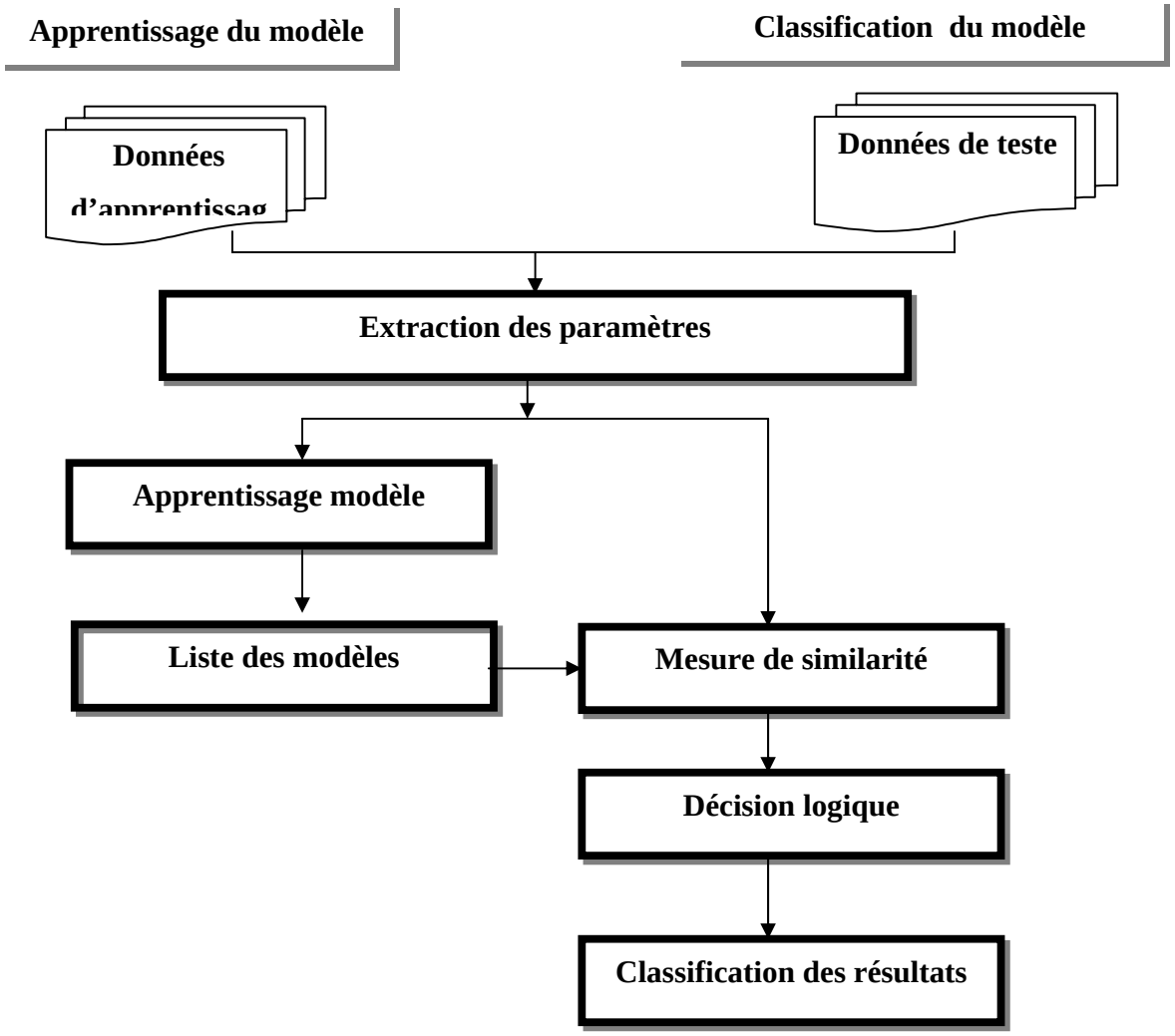

Fig.2.10 : Un graphique globale d'un processus de classification automatique

Conclusion

- Dans ce chapitre on a vu quelques outilles utilisés dans le traitement des signaux vocaux (acquisition, filtrage, segmentation, transformation, classification etc.). Ces outilles sont nécessaire dans le prétraitement et l'analyse du signale qui sont des étapes essentiel dans un système RAP. Aussi on a donné une idée globale sur les principes utilisés et les opérations de base dans l'extraction des paramètres, comment ces paramètres sont utilisés dans un système RAP et les obstacles techniques qu'on peut les rencontrer.

Représentation du Signal Vocal par les Méthodes classiques

Dans le contexte de la reconnaissance vocale le but essentiel de la phase d'extraction de paramètre est de calculer une séquence des vecteurs de paramètres qui donnent une représentation compacte pour un signal vocal d'entrée donné. La question de base est comment faire pour passer le problème de redondance et la variabilité causée par le signal vocale.

L'extraction des paramètres est habituellement réalisée dans trois phases. La première phase est appelée l'analyse acoustique est prévue pour produire une représentation de l'enveloppe spectre d'énergie court terme. Ce spectre est une version lisse du spectre d'énergie détaillé et montre ainsi une variabilité sensiblement plus petite que le spectre origine. La $2^{ème}$ étape et de compiler un vecteur étendu composé de paramètre statique est dynamique. La troisième phase qui ni pas tous jour présente transforme ces vecteurs étendu en des vecteurs plus compacte qui sont alors fournis au système de reconnaissance.

3.1 La Prédiction Linéaire

3.1.1 Interprétation temporelle

Le raisonnement dans *l'analyse* par prédiction *linéaire* (LP) est que les échantillons adjacents d'un signal de parole sont fortement corrélés et ainsi, le comportement de signal peut être prévu jusqu' à certain degré basé sur des échantillons passés. Le modèle LP suppose que chaque échantillon peut être rapproché par une combinaison linéaire de quelques échantillons du passé [04]:

$$s[n] \approx \sum_{k=1}^{p} a[k]s[n-k] \tag{3.1}$$

Où *p* est *l'ordre* de prédiction. Le but de l'analyse LP est à déterminez *les coefficients prédicateur* $\{a[k] \quad k = 1,....., p\}$ de sorte que l'erreur moyenne de prédiction (ou *le résidu)* est le plus petit possible. L'erreur de prédiction pour n échantillon est donnée par la différence entre l'échantillon réel et sa valeur prévue:

$$e[n] = s[n] - \sum_{k=1}^{p} a[k]s[n-k] \tag{3.2}$$

Plus l'erreur est faible mieux est le prédicateur (3.1) .L'énergie de l'erreur est donnée par :

$$\begin{aligned} E &= \sum_{n} e[n^2] \\ &= \sum_{n} \left(s[n] - \sum_{k=1}^{p} a[k]s[n-k] \right)^2 \end{aligned} \tag{3.4}$$

Selon le choix de l'intervalle d'erreur de minimisation dans (3.4), deux méthodes sont utilisées: *méthode de covariance* et *méthode d'autocorrélation* [04] [16]. La méthode d'autocorrélation est la méthode préférée puisqu'elle est plus efficace et garantit toujours la stabilité des filtres. Les équations AR pour la méthode d'autocorrélation sont de la forme :

$$Ra = r \tag{3.5}$$

Où *R* est une matrice *appelée* matrice de Toeplitz $\boldsymbol{a}$ est le vecteur des coefficients LPC et $\boldsymbol{r}$ est l'autocorrélation. La matrice $\boldsymbol{R}$ et le vecteur $\boldsymbol{r}$ sont complètement définis par **p** échantillons *d'autocorrélation*. La séquence d'*autocorrélation de* $s[n]$ est défini comme suite :

$$R[k] = \sum_{n=0}^{N-1-k} s[n]s[n-k] \tag{3.6}$$

En raison de la redondance dans les équations AR, il existe un algorithme efficace pour résoudre le problème, connu par la *récursion de Levinson-Durbin* [04]. La procédure de Levinson prend la séquence d'autocorrélation comme entrée, et produit les coefficients $\{a[k] \quad k = 1,....., p\}$ La complexité en temps est de *O* (p^2). Les étapes utilisées dan la méthode d'autocorrélation pour calculer les coefficients sont résumés dans la figure.3.1

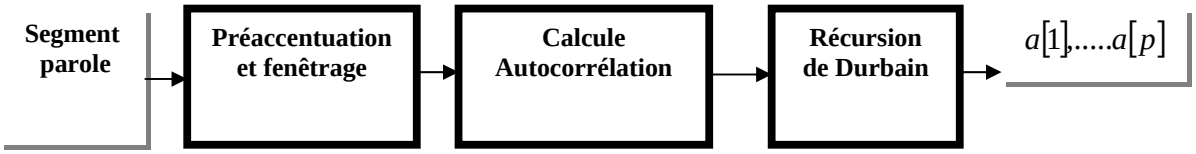

Figure 3.1 Calcul des coefficients LPC par la méthode d'autocorrélation

N'importe quel signal peut être rapproché avec le modèle LP avec une petite erreur de prédiction. L'ordre optimal du modèle dépend de quelle type d'information on veut extraire à partir du spectre. Ceci peut être vu en considérant l'interprétation de LP dans le domaine fréquentielle.

La minimisation de (3.4) réduire au minimum l'erreur carrée entre le spectre amplitude de signal et la réponse en amplitude du modèle.

3.1.2 Interprétation fréquentielle

L'équation (3.1) peut être transformée en égalité comme suit [02] :

$$s[n] = \sum_{k=1}^{p} a[k]s[n-k] + Gu[n] \tag{3.7}$$

Où *u* [n*]*, G représente l'excitation et son gain respectivement. Considérer ceci comme l'équation récursive d'un filtre RII, la somme présente la partie de rétroaction, et $le\,Gu[n]$ représente le signal d'entrée. La fonction de transfert est donnée par:

$$H(Z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^{p} a[k]z^{-k}} \tag{3.8}$$

Chaque pôle du filtre H(z) représente une crête locale dans le spectre d'amplitude, le modèle est limité à modéliser seulement les crêtes du spectre (résonances du conduit vocal). Les sons et voyelles nasaux incluent, en plus des résonances, des *antirésonances* qui résultent du côté tube fermé formée par la cavité orale [02]. Modeler ces antirésonances exige des zéros dans le filtre. Cependant, comme l'ordre *p* du modèle tout pôle est suffisamment haut, les résonances nasales sont également modelées arbitrairement. Ceci signifie que pour une valeur donnée d'erreur, les sons et voyelles nasales exigent un prédicteur d'ordre plus supérieur que les sons non nasaux.

Si nous comparons Eqs. (3.3) et (3.7), nous pouvons remarquer que le système qui a produit $s[n]$ est près du modèle (3.7), $e[n] \approx Gu[n]$. Dans un autre mot, le signal résiduel $e[n]$peut être employé en estimant le signal d'excitation. Les pôles de la fonction transfert (3.8), d'autre part, modélise l'enveloppe du spectre court terme. Les pôles *de* $H(z)$ sont situés aux fréquences des formants quand la proposition tout pôle est valide. Ceci suggère que l'enveloppe spectrale lissée obtenue par l'intermédiaire de l'analyse LP peut être employé dans l'estimation des formants [04].

Un exemple d'emploi d'analyse par LP dans l'extraction d'enveloppe spectrale est montré dans la figure.3.2 Le panneau supérieur montre le spectre original FFT, et le panneau inférieur montre trois enveloppes d'ordre LPC différents. On peut voir que pour p =8 le spectre original est fortement lissé. D'autre part, si l'ordre de prédiction est haut (p = 120), le modèle de LP est capable de reproduire même les harmoniques. Un compromis (p = 16) fournit les informations sur la structure spectrales produite par le filtre di conduit vocal. Les formants sont visibles dans l'enveloppe spectrale, montrant les positions approximatives des trois premiers formants *F1 F2 F3*.

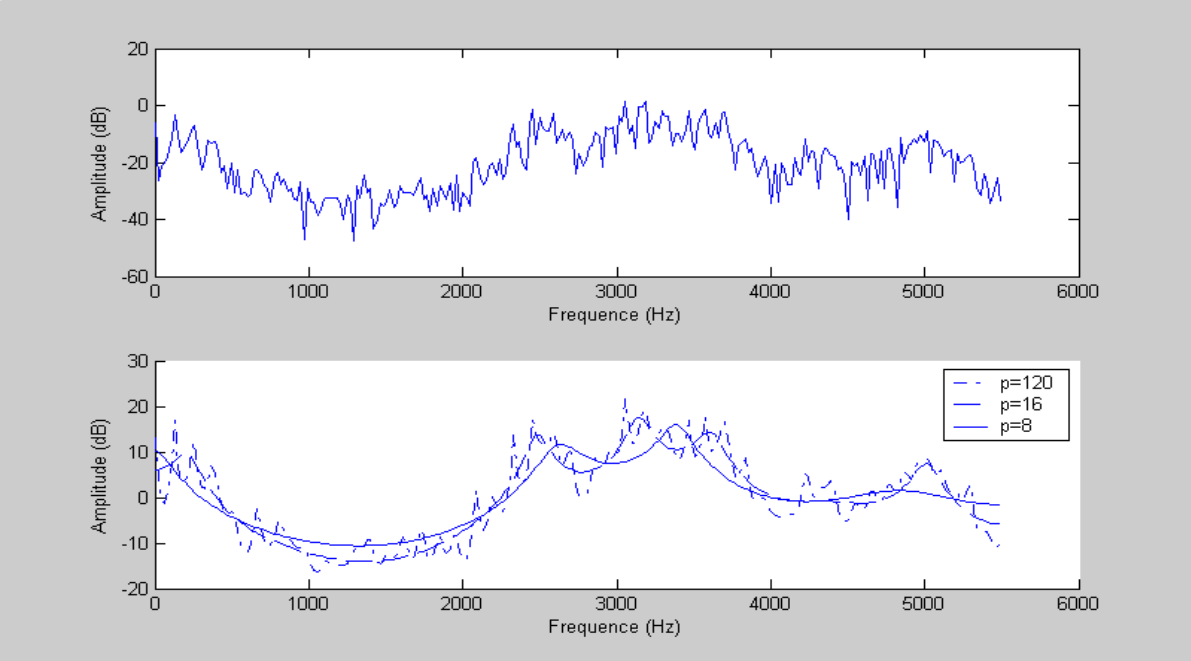

Figure 3.2 Estimation de l'enveloppe spectrale de la voyelle [i] par différent ordre de prédiction LPC

Une des règles utilisée pour le choix d'ordre de prédiction est de prendre un pôle complexe pour chaque intervalle de 01 kilohertz plus 2-4 pôles pour modeler les rayonnements des lèvres et les effets glottaux [02]. Par exemple, pour les communications téléphoniques la gamme de fréquence efficace de la parole est 0-4 kHz. Nous pouvons donc estimer le besoin à 4 pôles + 2-

4 pôles s = 6-8 pôles. De plus les pôles doivent être réels ou des paires complexes conjuguées pour s'assurer que les coefficients de filtre sont réels, l'ordre du modèle est deux fois le nombre de pôle. Ainsi, nous choisirions l'ordre de $p = 12$ à $p = 16$.

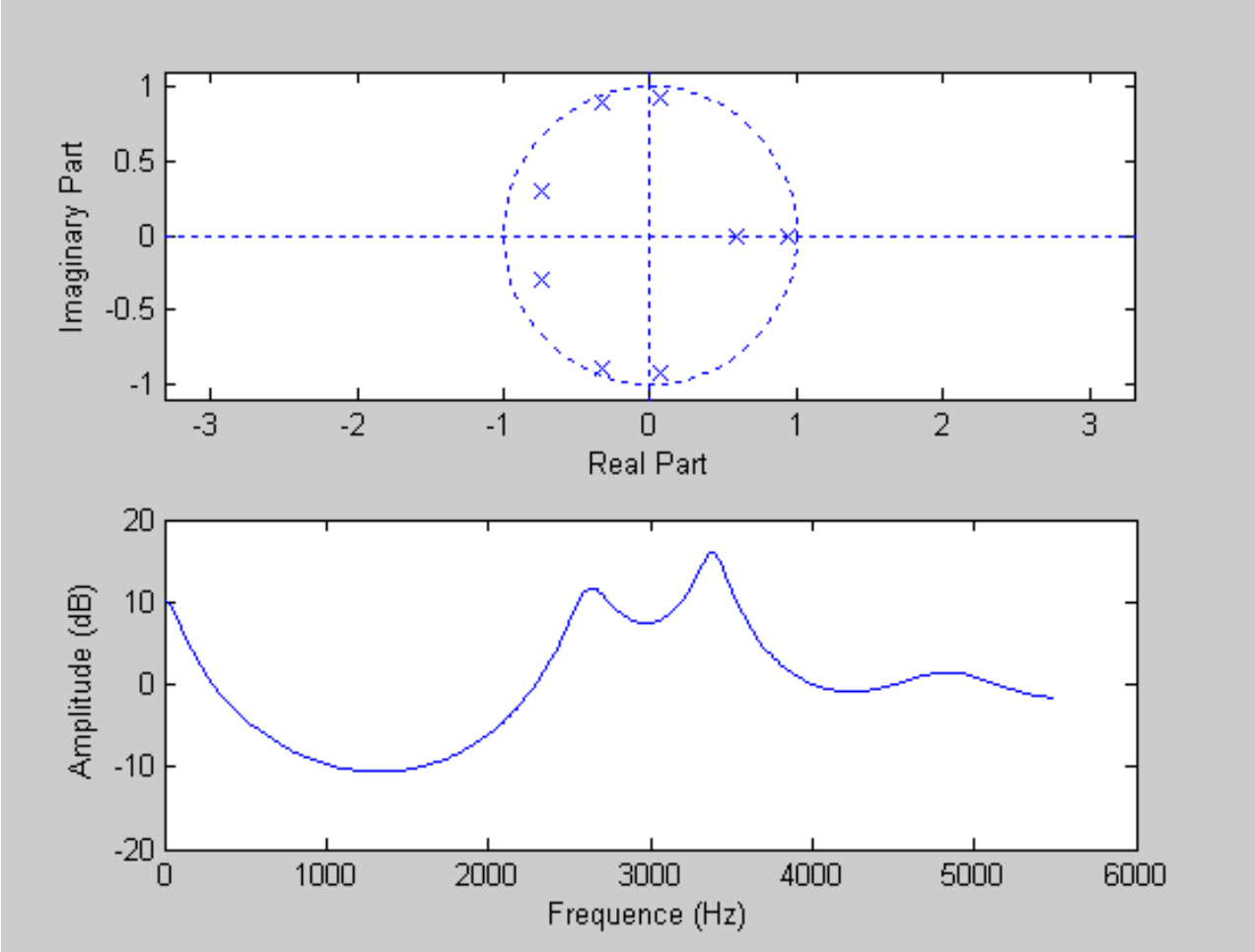

Figure 3.3 Les pôles LPC dans le plan Z et le spectre d'amplitude correspondant

3.2 Bancs de filtre

Les bancs de filtre ont l'avantage par rapport à la plupart des autres représentations spectrales parce que les paramètres ont une interprétation physique directe. Ceci permet, par exemple, l'utilisation de *la connaissance* a priori des puissances de discrimination des sous bandes pour pondérer les sous bandes [17]. En outre, si des sous bandes sont bruitées, les sous bandes propres peuvent toujours être employées [18].

Employer des fonctions de déformations expliquées psycho acoustiquement dans la construction des bancs de filtre (particulièrement les échelles Mel et Bark) est bien connue dans les systèmes RAP.

En général deux approches basées sur les bancs de filtre sont utilisées dans les systèmes RAP [18]:

- Fusion au niveau des paramètres (fusion en entrée).
- Fusion au niveau du classificateur (fusion en sortie).

Dans le premier cas, les sorties des sous bandes sont combinées dans un seul paramètre de dimension M .le vecteur représente une unité de son de parole. Dans l'autre cas, chaque sous bande est considérée comme indépendante et un modèle est généré séparément pour chaque sous bande.

3.2.1 Fusion au niveau des paramètres

L'approche la plus simple à l'extraction des paramètres est de considérer les sorties des sous bandes (éventuellement comprimées en utilisant le logarithme ou toute autre fonction non linéaire) directement comme paramètres. La prolongation normale de cette approche est de pondérer chacune des sous bande par une masse représentant la puissance de discrimination de chaque sou bande (figure 3.4).

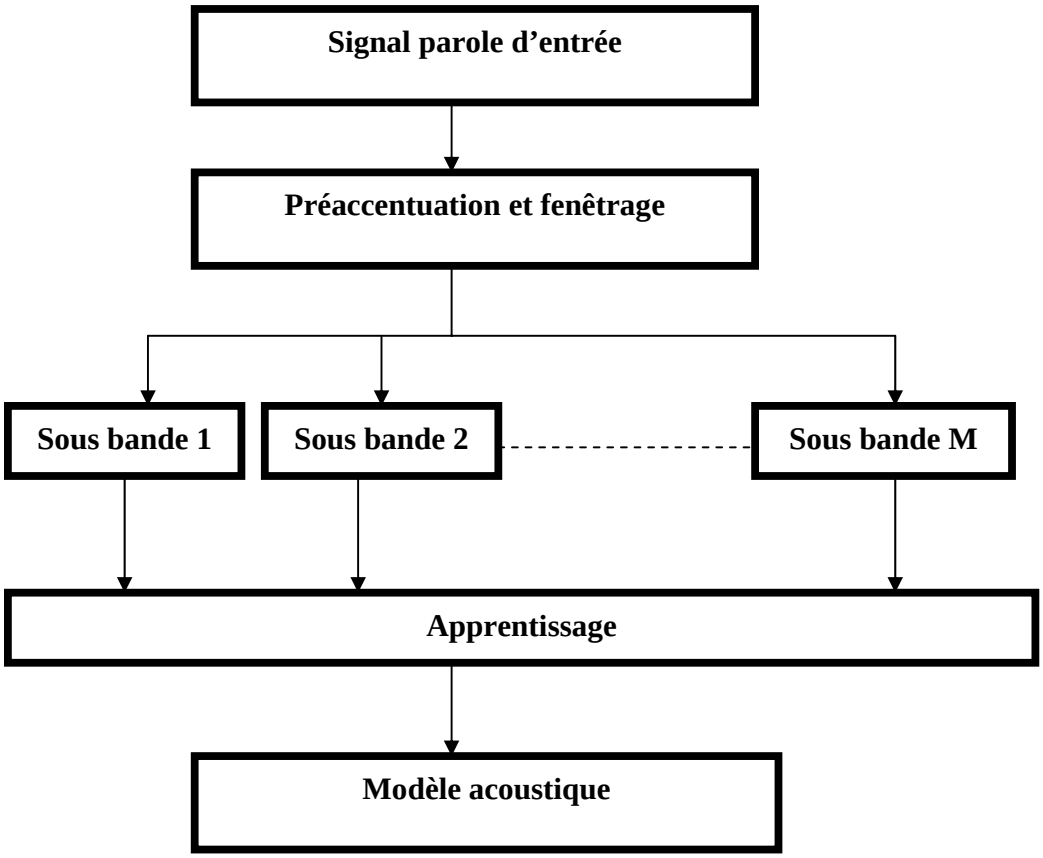

Figure 3.4 Extraction des paramètres par banc de filtre (fusion au niveau des paramètres)

3.2.2 Fusion au niveau du classificateur

Une autre approche pour combiner l'information de plusieurs sous bandes est de modeler chaque sous bande indépendamment des autres en suite faire une combinaison des points des modèles des sous bandes dans le classificateur (figure 3.5). Le résultat de combinaison au niveau sortie du classificateur est flexible.

La conception de tels systèmes inclut des entrées de type banc de filtre, des architectures de classificateur, et une méthode de combinaison des points ce qui n'est pas une tâche exacte. La règle simple pour combiner les points est de faire la *somme*. Les inconvénients principaux des systèmes de fusion en sortie du classificateur sont le temps de calcul et l'espace mémoire réservé. Pour chaque sous bande de chaque classe, un modèle séparé doit être stocké et dans

l'étape d'identification, chaque classificateur doit calculer ses propres points. Le temps global augmente avec le nombre de sous bandes et la complexité des classificateurs. La figure 3.6 montre un des principes utilisés dans l'architecture des systèmes à base d'un banc de filtre et la fusion au niveau du classificateur.

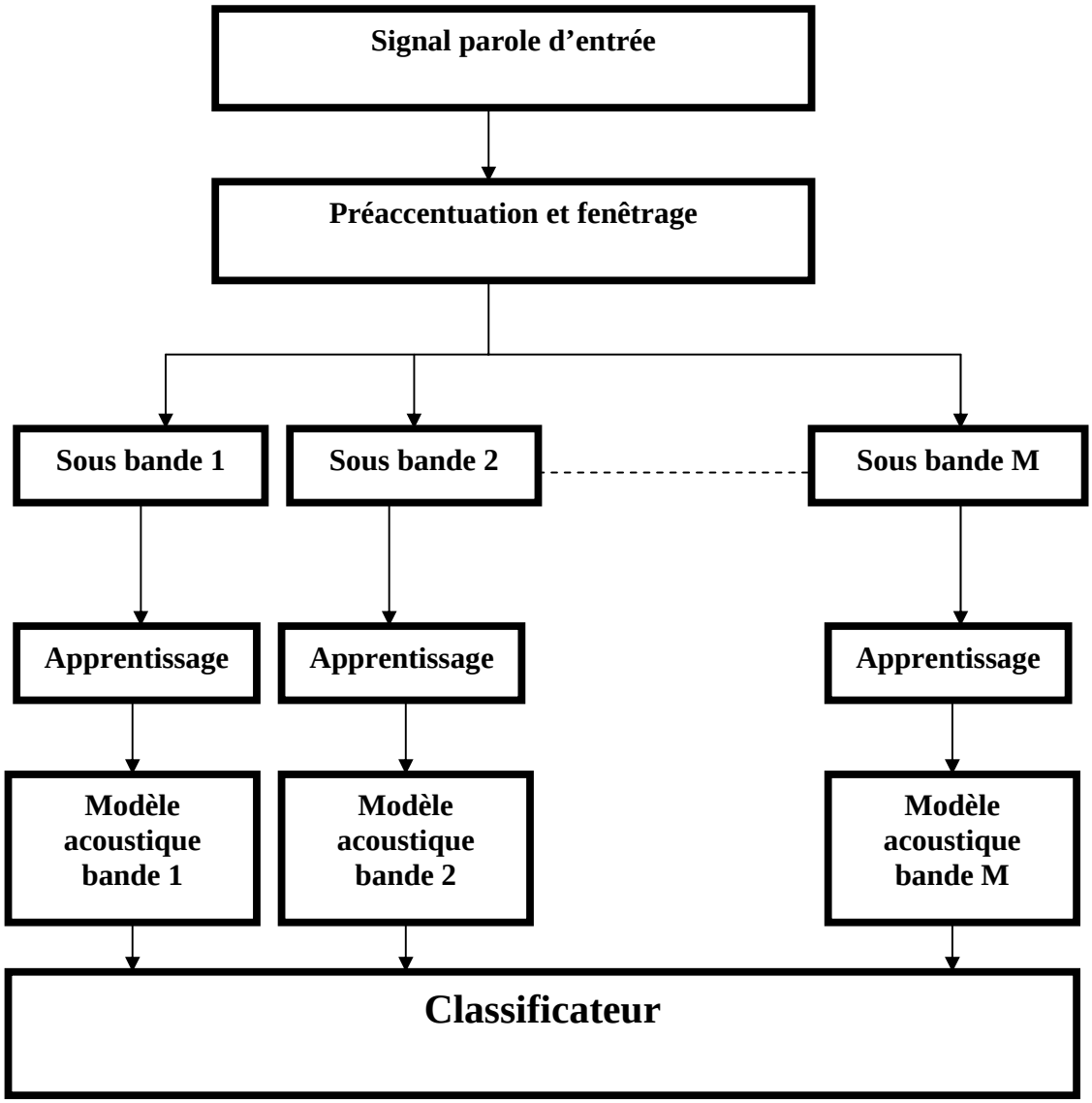

Figure 3.5 Extraction des paramètres par banc de filtre (fusion au niveau du classificateur)

3.3 L'analyse cepstrale

La prédiction linéaire emploie le modèle tout pôle du spectre. Une méthode alternative au LPC est la prétendue *analyse cepstral* [04]. Dans l'analyse cepstral, le spectre d'amplitude est représenté comme combinaison des fonctions de base cosinus avec des fréquences variables. Les coefficients cepstral sont les amplitudes des fonctions de base. La figure 3.6 montre une comparaison de l'évaluation de l'enveloppe spectrale en utilisant LPC (modèle de tout-pôle) et la représentation cepstrale.

Notons que les crêtes dans le modèle de LPC sont très claires, tandis que le cepstre présente une enveloppe plus lisse. Dans ce sens, le modèle de LPC préserve plus de détails au sujet du spectre avec le même nombre de coefficients.

Formellement, *le vrai cepstre* du signal numérique $s[n]$ est défini comme l'inverse du Transformée de Fourier du logarithme du spectre d'amplitude [06]:

$$c[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} C_s(\omega) e^{j\omega\pi} d\omega \tag{3.9}$$

Où $C_s = \log\left|S\left(e^{j\omega}\right)\right|$ dénote le logarithme du spectre d'amplitude, les coefficients $c[n]$sont les coefficients de série de Fourier du Log Spectre. Le Log spectre est représenté comme une infinité d'addition des cosinus de différentes fréquences, et les coefficients cepstral sont les amplitudes des fonctions de base. Les petits coefficients cepstral représentent les changements lents du spectre et les coefficients plus élevés représentent les composants rapidement variables du spectre. Dans les sons voisés, il y a des composants périodiques dans le spectre d'amplitude, la structure fine harmonique résulte de la vibration des cordes vocales. Les variations lentes résultent du filtrage du conduit vocal, et de la descente spectrale de la source. Un exemple de spectre modélisé en employant le cepstre réel est utilisé dans la figure (3.7).

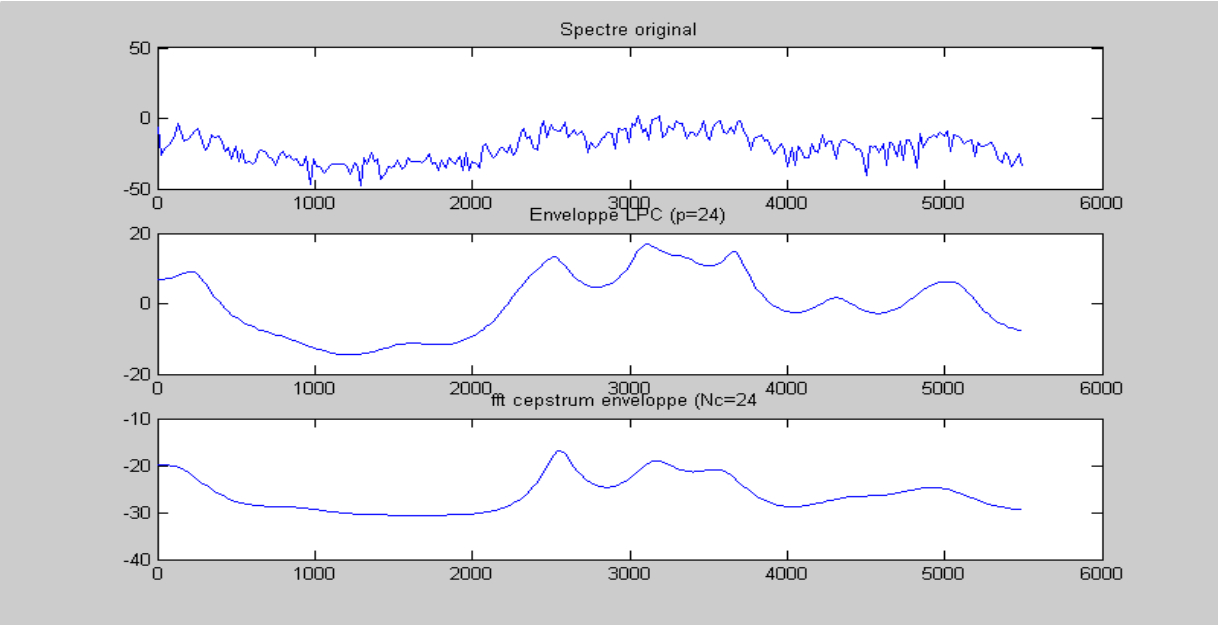

Figure 3.6 Exemple d'estimation d'enveloppe spectrale par LPC est FFT cepstre

Similaire à l'analyse de LPC, l'augmentation du nombre de coefficients a comme conséquence plus de détails. La raison de prendre le logarithme du spectre peut être expliquée comme suit [06]. Selon la théorie du filtre source,

$$\left|S\left(e^{j\omega}\right)\right| = \left|U\left(e^{j\omega}\right)\right|.\left|H\left(e^{j\omega}\right)\right|, \tag{3.10}$$

Où *S; U* et *H* correspondent au son articulé, source et au filtre, respectivement. En prenant le logarithme, les composants multiplicatifs sont convertis dans des composants additifs:

$$\log\left|S\left(e^{j\omega}\right)\right| = \log\left|U\left(e^{j\omega}\right)\right| + \log\left|H\left(e^{j\omega}\right)\right|. \tag{3.11}$$

Prendre le logarithme correspond à exécuter *une transformation homomorphique* [06], les séquences multiplicatives sont converties en un nouveau domaine, où ils deviennent additifs.

La formule pratique pour calculer le cepstre réel est obtenue par l'emploie DFT et IDFT:

$$c[n] = f^{-1}\left\{\log\left|f\left\{segment\ signal\right\}\right|\right\} \tag{3.12}$$

Donc le cepstre réel est obtenu en appliquant la DFT inverse au logarithme de l'amplitude du DFT d'un segment du signal parole à analyser.

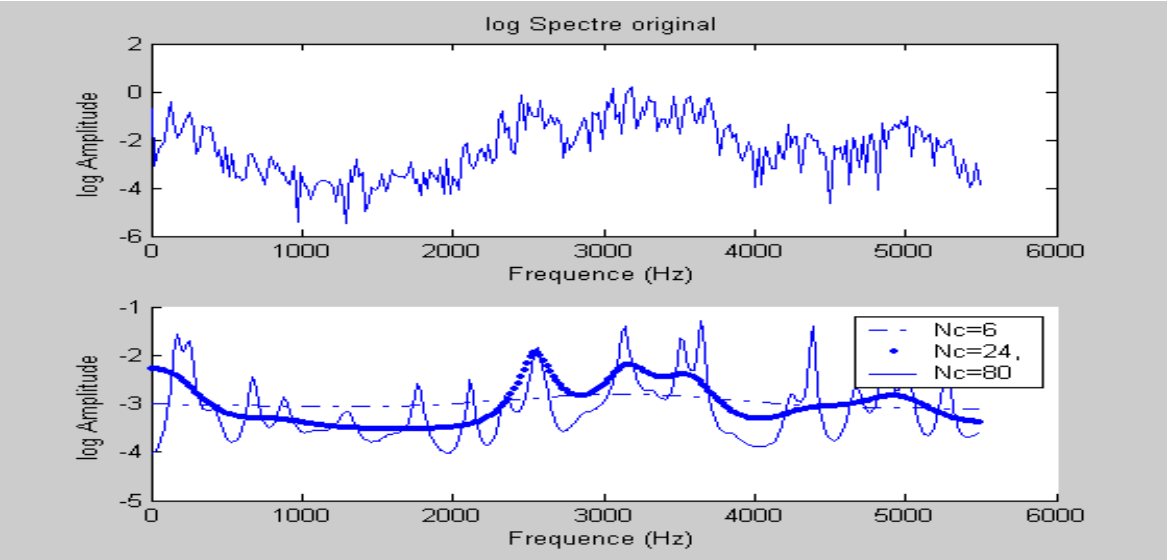

Figure 3.7 Enveloppe spectral reconstruit on utilisant des différents nombre de coefficients Cepstraux (Nc=6, 24,80).

3.2.1 Coefficients LPCC

Les coefficients de LPC sont rarement employés comme paramètres eux-mêmes, les coefficients adjacents de ce prédicteur sont fortement corrélés [02], et donc, les représentations avec des paramètres moins corrélés seraient plus efficaces. Un ensemble populaire de paramètre est LPCC (linear predictive cepstral coefficients).

Etant donné les coefficients :

$$LP\{a[k] \quad k = 1,....., p\},$$

Les coefficients cepstrals sont calculés pare la formule récursive suivante [04]:

$$c[n] = \begin{cases} a[n] + \sum_{k=1}^{n-1} \frac{k}{n} c[k] a[n-k], & 1 \le n \le p \\ \sum_{k=n-p}^{n-1} \frac{k}{n} c[k] a[n-k], & n \succ p \end{cases} \tag{3.13}$$

Une chose apparente est que le nombre des coefficients LP (p) est fini, mais le nombre des coefficients LPC cepstre $c[n]$est infini. Cependant, les amplitudes $|c[n]| \to 0$ rapidement avec l'augmentation de n et relativement un nombre petit de coefficients est ainsi nécessaire pour modeler le spectre.

Il est important de noter que les coefficients LPC cepstral sont dérivés des coefficients de LPC, donc un modèle tous pôle est employé. Par conséquent, en général les coefficients LPC cepstral ne sont pas les mêmes que les coefficients cepstral dérivés directement des amplitudes du spectre.

3.3.2 Coefficients MFCC

Les paramètres les plus généralement utilisés dans la reconnaissance de la parole est MFCC (*mel-frequency cepstral coefficients)* [17]. Le calcul de cepstre est semblable à ce lui décrit dans la section précédente. Mais, le banc de filtre mel (ou tout autre) est appliqué dans le domaine fréquentielle avant le logarithme et IDFT. Le but du banc mel est de simuler les filtres des bandes critiques du mécanisme d'audition. Les filtres sont également espacés sur une échelle mel, et habituellement ils ont une forme triangulaire [07] [08]. Les sorties du filtre triangulaire sont comprimées en utilisant le logarithme et la transformée cosinus discrète (DCT) est appliquée [19]:

$$c[n] = \sum_{i=1}^{M} \log Y(i) \cos\left[\frac{\pi n}{M}(i - 1/2)\right] \tag{3.14}$$

Typiquement $c[0]$ est exclu. Une des propriétés importantes des coefficients cepstrals est qu'ils sont particulièrement non corrélés. Cette propriété a quelques conséquences pratiques importantes. Par exemple, si une distance de Mahalanobis est employée en tant que métrique dans le classificateur, on aura un gain faible en employant une matrice de covariance pleine. Par contre une matrice de covariance diagonale présente un gain plus considérable [02].

3.3.3 Coefficients PLP

Un autre ensemble des paramètres les plus utilisés dans le traitement de la parole est PLP prédiction linéaire perceptuelle (perceptuel linear predictive) Hermansky [20]. Ces paramètres sont presque similaires à celle des (MFCC). La différence principale entre les deux paramètres est la nature de l'enveloppe spectrale utilisée pour les calculer.

PLP combine plusieurs techniques pour approximer les caractéristiques de l'oreille humaine.

- **Intégration des bandes critiques (Bark):** l'analyse LPC donne la même approximation de l'enveloppe de la densité spectrale pour toute la zone des fréquences utiles, ceci n'est pas en accord avec le fonctionnement de l'oreille ou chaque fréquence porte une force

sonore résultant de l'intégration de l'information sur une zone de fréquences appelée bande critique. Les bandes critiques sont réparties suivant l'échelle fréquentielle de Bark sur toute la zone de fréquence utilisée. Un Bark correspond à une augmentation de fréquence d'une quantité égale à une bande critique. Pour simuler ce fonctionnement dans le cadre de l'analyse PLP la densité spectrale est une bande critique. Cette fonction est la même pour toutes les fréquences dans l'échelle de Bark (à une translation prés). Ce qui réduit largement la résolution, surtout dans les hautes fréquences.

- **Préaccentuation par courbe d'isosonie**: des expériences de psycho acoustique ont montré que l'oreille possède des caractéristiques non linéaires. En effet, il faut mètre en évidence que l'intensité perçue, lorsqu'on écoute un son pur d'intensité acoustique constante, varie avec la fréquence de ce son pur. Pour simuler ce phénomène dans le cadre de l'analyse PLP, on multiplie la densité spectrale résultante de l'étape précédente par une fonction de pondération. Il est possible d'estimer cette fonction en reportant sur un abaque des lignes isosoniques le long desquelles un son pur donne à l'oreille une sensation égale d'intensité. Ceci est à l'origine de ce qu'on appelle la sonie considérée comme l'intensité subjective des sons.
- **Loi de Stevens :** les deux traitements précédents ne sont pas suffisants pour avoir une correspondance entre l'intensité mesurée et l'intensité subjective (la sonie). La loi de Stevens affirme qu'après avoir réalisé l'intégration des bandes critiques et la préaccentuation, la relation entre l'intensité et la sonie devient:

$$\textbf{Sonie} = (\textbf{intensité})^{0.33}$$

En fin le spectre audible est rapproché à un modèle tout pôle la figure 3. 8 illustre les étapes principales pour le calcule des coefficients PLP :

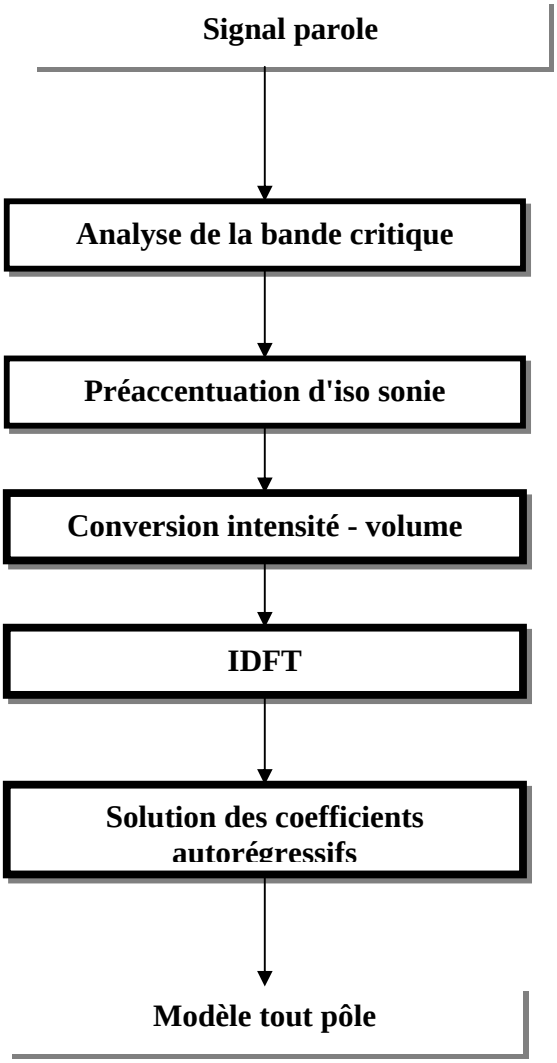

Figure 3.8 Extraction des paramètres par l'analyse PLP

PLP est plus proche du comportement de l'oreille humaine que la technique LPC. L'analyse PLP prend le spectre de puissance cout terme de la parole et fait une convolution avec un modèle de bande critique. Puis la bande critique est au sujet prélevée dans des intervalles d'une échelle Bark, puis une opération de préaccentuation est effectuée avec une courbe **d'iso sonie**. Finalement le spectre résultant est comprimé avec une fonction non linéaire de racine –cubique

simulant la loi de conversion intensité - volume et un modèle tout pôle est calculé après une IDFT appliquée sur le signal comprimé.

3.4 Les paramètres dynamiques

Les paramètres discutés dans les sections précédentes sont appelés paramètres statiques parce que on assume que chaque vecteur spectral est une représentation d'un signal stationnaire court terme, il n'y a aucune information de temps codée dans ces paramètres.

Donc nous avons supposé que chaque vecteur spectral de paramètre représente à un certain intervalle temps une évolution du spectre.

Lorsqu'on parle, les articulateurs changent sans interruption leurs positions avec un certain taux. Les mouvements articulatoires sont alors représentés dans le spectre mesuré. Le taux de ces changements spectraux dépend du style, taux et contexte de la parole. Certains de ces paramètres spectraux dynamiques sont les indicateurs de l'information du message parlé.

Une méthode largement utilisée pour coder une partie d'information dynamique des paramètres spectraux est connue par le nom *delta paramètres.* Les dérivés de temps des paramètres sont estimés par une certaine méthode, et puis l'approximation du dérivé est collée au vecteur paramètre, rendant ce dernier d'une dimension plus grande. Comme exemple, si les coefficients cepstral de 12 mel-fréquences sont regroupés avec leurs approximations dérivées, la dimensionnalité du nouveau vecteur paramètre est 12 + 12 = 24. On peut voir que la dimensionnalité de l'espace a augmenté, donc plus de données d'entraînement sont exigées pour des évaluations fiables de modèle. Généralement, les dérivés de temps des paramètres delta sont estimés aussi, rapportant des nouveaux *paramètres appelés delta-delta.* Ces derniers sont de nouveau collés au vecteur paramètres et comme résultat un espace dimensionnel plus élevé de paramètre.

Parfois les paramètres delta et delta-delta se nomment comme *coefficients vitesse* et *accélération* respectivement [02]. C'est un phénomène physique, c.-à-d. vitesse est la dérivée de temps du déplacement, et accélération est le dérivé de temps de la vitesse.

Les paramètres delta et delta-delta- sont employés avec plusieurs formes de paramètres, particulièrement le cepstre et ses variantes [07, 08 ,03]. L'ensemble des paramètres statiques sont des descripteurs courts terme du spectre, les paramètres dynamiques représentent les

changements spectraux dans le temps. Un exemple des paramètres de delta et delta-delta est montré dans la figure.3.9.

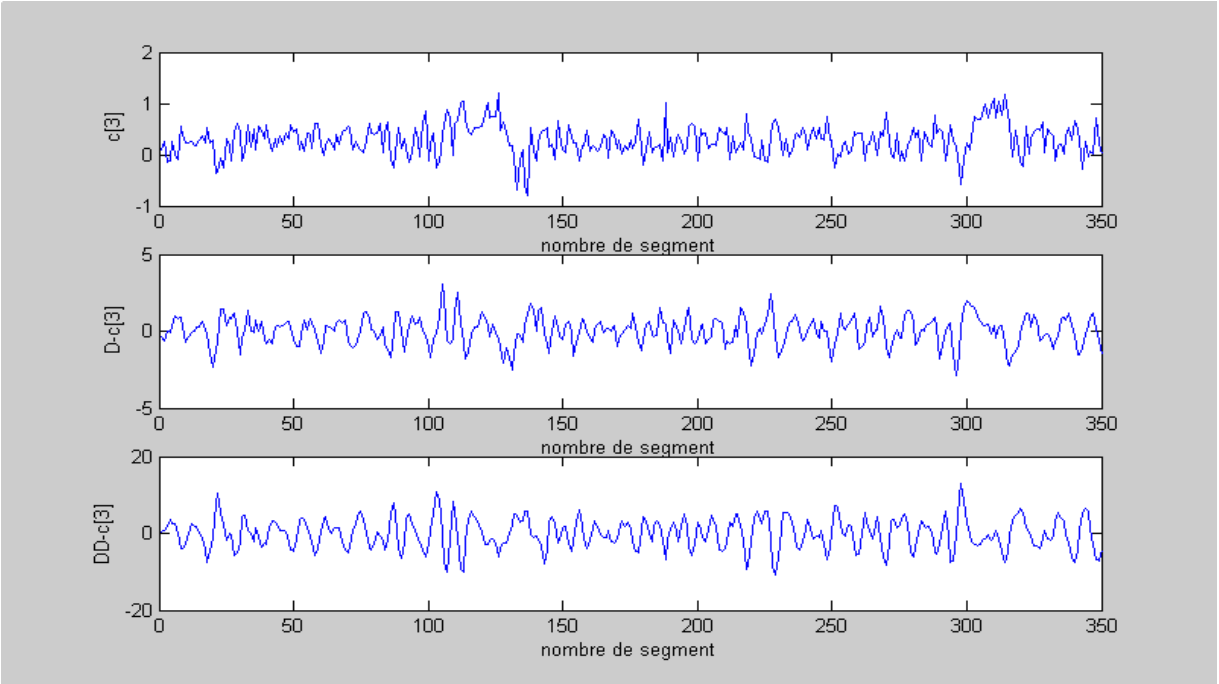

Figure 3.9 Suivi d'un coefficient mfcc c [3] et ses dérivées Delta et Delta -Delta

Le 1ier panel montre les variations du 3ème coefficient *c* [3] dans le temps. Les panneaux 2et3 montrent l'évaluation du premier et deuxième dérivé. Il y a deux principes généraux pour estimer ces dérivés [02]: (1) différenciation, et (2) utilisation d'une expansion polynomiale. Soit $f_k[i]$ dénote *le* $i^{ème}$ *paramètre* dans k^{ème} *intervalle* de temps.

Dans la méthode de différenciation le delta paramètre *du* $i^{ème}$ paramètre est défini comme suite :

$$\Delta f_k[i] = f_{k+M}[i] - f_{k-M}[i] \tag{3.15}$$

Où *M* est en général 2 ou 3 segments. La différentiation est faite séparément pour chaque paramètre *i* qui résulte dans un vecteur paramètre delta. La méthode des différenciations est simple, mais puisqu'elle agit comme un filtrage passe-haut sur le domaine paramètre, elle tend à amplifier le bruit [02]. Pour cette raison, l'adaptation d'une courbe polynomiale pour la trajectoire temps du paramètre peut avoir comme conséquence des meilleures évaluations. Ce

problème s'appelle *l'analyse de régression.* Assemblé une courbe polynomial sur plusieurs échantillons représente l'évaluation d'une trajectoire passe-bas filtrée dans le temps.
De même comme avec LPC, l'ordre du polynôme est d'abord fixé. Puis, la solution des moindres carrés pour des coefficients du polynôme est obtenue. Comme exemple, pour *la régression linéaire* c.-à-d. le polynôme de premier ordre, la solution des moindres carrés est montrée sous la forme suivante:

$$\Delta f_k[i] = \frac{\sum_{m=-M}^{M} m f_{k+m}[i]}{\sum_{m=-M}^{M} m^2} \tag{3.16}$$

Notez que le dénominateur est constant et peut être interprété simplement comme un facteur de graduation qui peut être remplacé par des autres constants. Une augmentation d'ordre des polynômes peut être employée pour obtenir des approximations plus douces, pratiquement un polynôme de premier ordre est satisfaisant. La figure 3.10 montre une comparaison entre les méthodes de différentiateur et de régression linéaire utilisées pour l'estimation des paramètres Delta du paramètre c [3] de la figure.3.9. On peut voir que l'augmentation du nombre d'armatures (m) lisse les évaluations avec les deux méthodes, mais la méthode de régression produit des évaluations plus douces

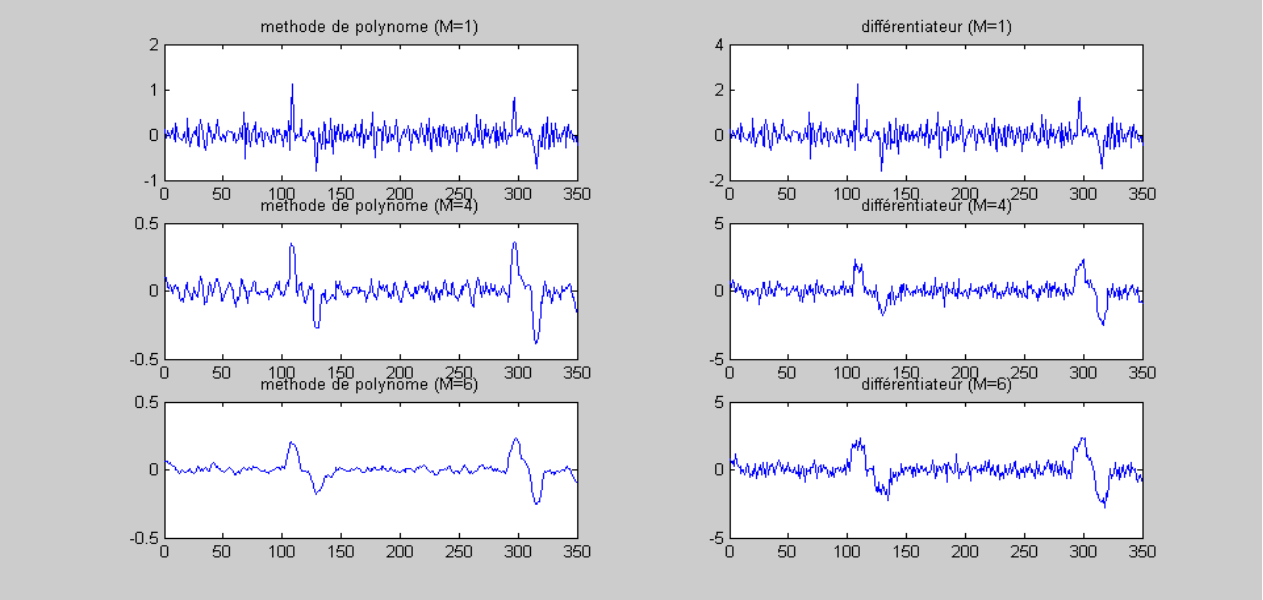

Figure.3.10 Comparaison entre estimation polynomiale et différentielle des paramètres Delta

Dans les deux méthodes, le calcul de $\Delta f_k[i]$ exige M *segments* passé et futur. Par conséquent, les points finaux d'expression doivent être traités comme cas spéciaux. Les méthodes simples incluent le remplissage des armatures supplémentaires de M dans les deux extrémités avec des zéros, des nombres aléatoires, ou des copies des armatures adjacentes. Si les delta-deltas ou les dérivés d'ordre plus supérieur sont estimés, les frontières devraient être manipulés avec plus de soin puisque l'erreur accumule chaque fois des paramètres deltas sont calculés des deltas précédents. Pour un paramètre statique donné, il y a une longueur optimum pour la taille de la fenêtre de régression M. Comme on a vu déjà, une fenêtre trop courte mène des évaluations bruyantes. Si la fenêtre est trop longue, la résolution temporelle diminue.

3.5 Paramètre d'énergie

Après la phase de numérisation et surtout de quantification, le paramètre intuitif pour caractériser le signal ainsi obtenu est l'énergie [02]. Cette énergie correspond à la puissance du signal. Elle est souvent évaluée sur plusieurs trames de signal successives pour pouvoir mettre en évidence des variations. La formule de calcul de ce paramètre est :

$$\log(E) = \log \sum_{fenetre} s(n)^2 \tag{3.17}$$

Conclusion

L'extraction des caractéristiques est une étape importante dans le processus de reconnaissance de la parole. En effet, cette étape permet d'extraire des caractéristiques qui seront ensuite utilisées par le classificateur. La phase d'extraction des caractéristiques doit être faite avec soin, car elle contribue directement aux performances du système global. Nous avons vu dans ce chapitre les paramètres couramment utilisés qui sont les paramètres LPC (Linear Predictive Coefficients), les paramètres MFCC (Mel Frequency Cepstral Coefficients), les paramètres LPCC (Linear Predictive Coefficients) et les paramètres PLP (Perceptuel Linear Predictive Coefficients). On n'a vu aussi que le vecteur acoustique est renforcé par les paramètres dynamiques et énergie. Généralement le vecteur final contient:

(12 statiques + log E) + D +DD = 39 paramètres.

Le Choix D'une Nouvelle Représentation Robuste D'un Signal Vocal

4.1 Introduction

Bien que les systèmes de reconnaissance de la parole basés sur des modèles statistiques présentent une performance acceptable dans les environnements bien contrôlés (les laboratoires, les bureaux). Cette performance dégrade énormément dans les applications du monde réel. Le problème essentiel de cette dégradation revient à la dissemblance entre les expériences d'apprentissages (généralement réalisées dans des endroits sévères) et les tests réalisés dans le monde réel. La communication dans le monde réel implique transformation, corruption par écho, bruit de fond, distorsion causé par microphone, chaîne de transmission et codage,…, etc. qui sont des exemples des phénomènes qu'il faut prendre en considération.
Plusieurs techniques ont été développées pour trouver un remède à ce problème qui réside à notre avis dans les extracteurs de paramètre, quelques techniques qui sont jugées efficaces pour la reconnaissance sont proposées dans ce mémoire, ce chapitre donne une description générale sur ces techniques.

4.2 Dissemblance dans un système RAP

La stratégie de classification discutée avant est basée sur la supposition que les échantillons d'apprentissage et de test viennent des populations de même propriétés statistiques. Si cette condition est fausse, il y aura une dissemblance entre les propriétés des données

d'apprentissage et de test. La dégradation de performance de classification est déterminée par le degré de dissemblance entre les propriétés statistiques des données d'apprentissage et de test, en d'autre terme un degré de dissemblance élevé donne un taux bas de classification.
Dans les applications RAP, une multitude de facteurs peut provoquer une dissemblance entre les propriétés statistiques des données acoustiques d'apprentissages et de test.
La figure 5.1 illustre une situation typique d'acquisition des données dans laquelle on trouve un certain nombre des facteurs qui déterminent les propriétés acoustiques des données lors d'une acquisition d'un signal parole, par exemple:

- Le son de la parole qui est produit par les locuteurs.
- Les caractéristiques spécifiques du locuteur telles que: sexe, âge, accent (étranger), taux de la parole, état émotif, effort.
- L'environnement acoustique dans lequel le son articulé est produit, par exemple un environnement silencieux, un bureau, une station de train.
- La manière dont la quelle les sons articulés sont capturés et transmis par exemple à l'aide d'un microphone ou un téléphone.
- La numérisation des signaux en plus des erreurs de quantification qui sont inséparables dans n'importe quel procédé de numérisation, la fréquence à laquelle les signaux sont prélevés également limite l'information qui peut être représentée clairement dans les données
- Les propriétés des canaux et des réseaux de transmission pour les applications de ligne fixe et de ligne mobile, respectivement.

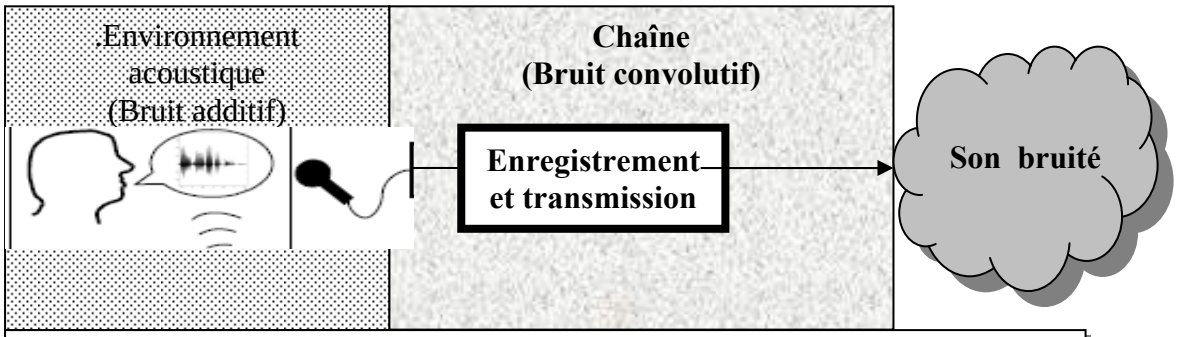

Figure 4.1 représentation d'une situation d'acquisition de données

Si les situations dans lesquelles les données d'apprentissage et de test sont acquis différemment dans un ou plusieurs aspects illustrés sur le schéma 4.1 les propriétés

acoustique des deux ensemble seront différents, si des modèles statistiques sont employés pour décrire les propriétés acoustiques des données, les différences dans les propriétés acoustiques d'apprentissage et de test se révéleront comme dissemblance entre les propriétés statistiques des deux ensembles .

4.3. Utilisation de l'analyse multi - résolution

Comme nous l'avons discuté dans le chapitre précédent l'analyse de Cepstre fournit des moyens efficaces pour manipuler les interférences convolutives (c.-à-d. la séparation entre la source d'excitation et le conduit vocal). Cependant pour les interférences additives le domaine cepstral n'est pas adéquat. L'opération de DCT (due à la localisation insuffisante des propriétés temporelles de l'analyse de Fourier) distribue les composantes fréquentielles locales entièrement sur le cepstre. Par conséquent les dispositifs cepstraux ont des propriétés faibles de la localisation des fréquences. Une des méthodes propose des dispositifs cepstraux de Multi résolution pour compléter les dispositifs cepstraux classiques. Le principe de cette technique est de calculer les dispositifs cepstraux dans des différentes résolutions de sous bandes. Dans cette méthode on peut considérer les problèmes suivants [14] :

- La superposition du cepstre plein bande et sous-bande cause une redondance dans l'espace paramètre. Par conséquent les paramètres résultants sont fortement corrélés. Pour la reconnaissance de la parole nous n'avons pas besoin de la même propriété de localisation à toutes les bandes de fréquence, seulement les fréquences des trois premières formant sont suffisantes dans le cas de la classification des phonèmes.

À cette considération, un autre algorithme [12] peut être utilisé pour calculer de nouveau paramètres du cepstre par l'analyse multi résolution MFDWCs. Dans cet algorithme On remplace la transformé DCT utilisée dans le calcule MFCCs classique par un calcul DWT dyadique appliqué sur 32 sorties log – énergie d'un banc de filtre Mel. On, ne garde que les 15 derniers coefficients (8 pour le niveau 2 + 4 pour le niveau 3 + 2 pour le niveau 5 + 1 pour le niveau 5 =15).

La transformation par ondelettes emploie des fenêtres courtes pour mesurer le contenu du signal dans les hautes fréquences et des fenêtres longues pour mesurer le contenu dans les basses fréquences du signal. Cette propriété rend la transformation par ondelettes différente de la transformée de Fourier à courte terme et de la transformée de Fourier.

Une ondelette est une fonction $\psi(t) \in L^2(\Re)$ (l'espace des fonctions carrées intégrables) d'une norme égale à l'unité et d'une moyenne zéro de sorte que :

$$\int_{-\infty}^{+\infty} \psi(t)dt = 0 \tag{4.1}$$

$$\left\| \psi(t) \right\| = 1. \tag{4.2}$$

La fonction d'analyse de la transformation par ondelettes à l'échèle **s** et la translation **u** est donnée par :

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right). \tag{4.3}$$

La transformation par ondelettes d'une fonction $f(t) \in l^2(\Re)$ dans le temps **u** et à l'échelle s est donné par :

$$WF(u,s) = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{s}} \psi^*\left(\frac{t-u}{s}\right) dt = \int_{-\infty}^{+\infty} f(t) \psi_{u,s}^*(t) dt. \tag{4.4}$$

Théoriquement n'importe quelle fonction d'une énergie finie et moyenne nul peut être une ondelette. Cependant, il y a beaucoup de critères pour choisir une ondelette. Puisque nous ne pouvons pas mettre en application une ondelette d'une durée infinie, nous avons besoin des ondelettes de manière compacte soutenues pour l'application pratique. L'affaiblissement de l'ondelette dans les domaines temporelle et fréquentielles est important. Nous voulons que l'ondelette se délabre rapidement en temps et en fréquence pour avoir de bonne localité en temps et en fréquence. Les ondelettes basées sur des bancs de filtres peuvent être mises en application efficacement. Puisque les signaux sont de longueur finie, les coefficients

d'ondelette auront de grandes variations non désirées aux extrémités en raison des discontinuités aux extrémités. Nous pouvons employer les ondelettes pliées qui demandent les ondelettes symétriques ou antisymétriques telles que l'ondelette spline pour diminuer l'effet des **discontinuités aux** extrémités. La figure 4.2 donne une idée globale sur le principe d'extraction des paramètres par l'algorithme MFDWCs.

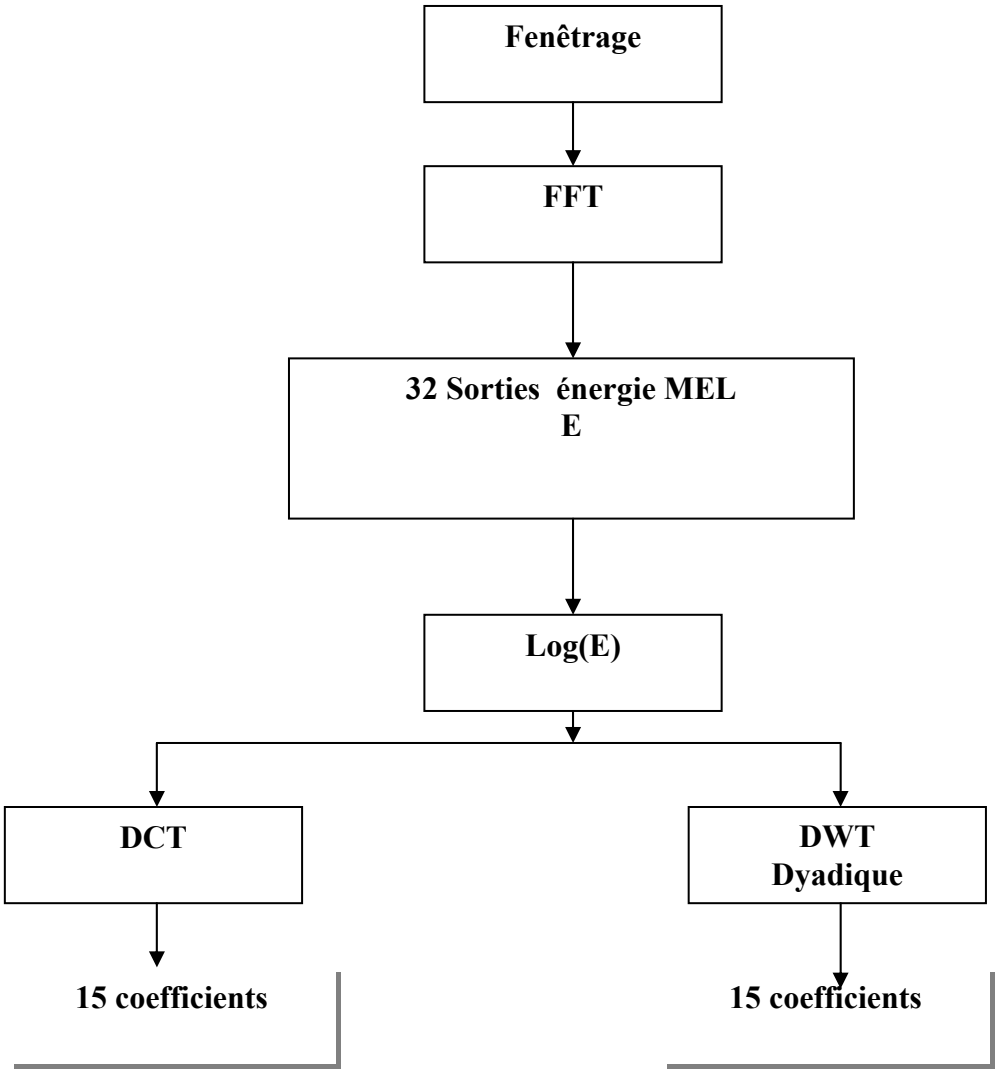

Figure 4.2. Algorithmes MFCC et MFDWC

4.3.1 Algorithme d'extraction de paramètre MBLPCCs :

La figure (4.4) montre un organigramme illustratif de l'algorithme d'extraction de paramètre MBLPCCs [18] [21]. Après le calcul de LPCCs de toute la bande fréquentielle du signal vocal, la DWT [11] (discrete Wavelet transform) est appliquée pour décomposer le signal en deux sous - bande de fréquence (approximation et détail). L'extraction de LPCCs est seulement réalisé sur la bande des basses fréquences (approximation) figure (4.3). La transformée en ondelettes peut être réalisées on utilisant une paire de filtre RIF (h filtre passe bas et g filtre passe haut)

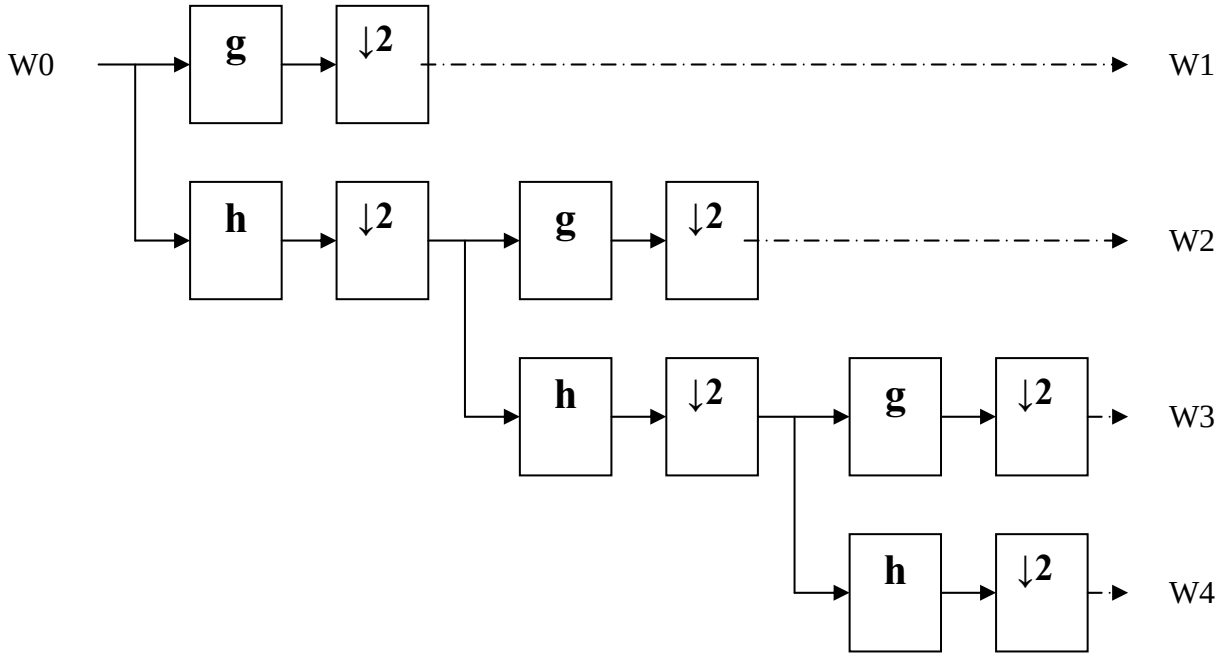

Figure 4.3 L'arbre d'analyse multi- bande pour une transformation discrète d'ondelettes

Une procédure récursive est employée pour mieux extraire les paramètres multi – bandes du signal vocal. Basé sur ce concept le nombre des paramètres MBLPCCs dépend du niveau de décomposition de la procédure implantée. Si le signal vocal est limité entre 0 à 5500 Hz. La décomposition en deux niveaux génère les trois bandes (0-5500), (0 -2250) et (0 -1125). On peut remarquer que le spectre des trois bandes se chevauche dans la région des basses

fréquences. Cette méthode vise sur le spectre des basses fréquences ce qui ressemble un peut à l'extraction des paramètres MFCCs.

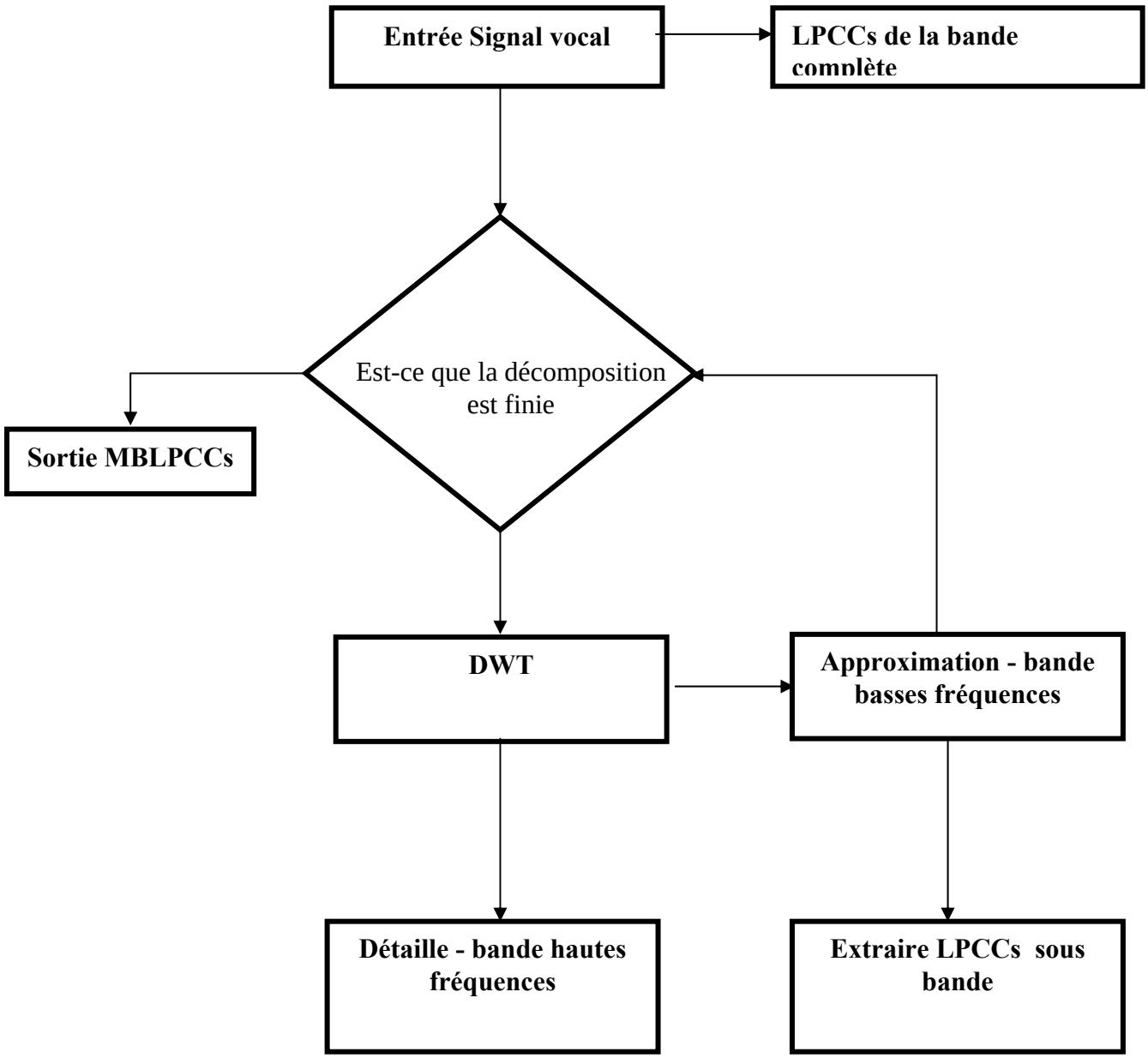

Figure 4.4 algorithme d'extraction des paramètres MBLPCCs

4.4 Normalisation du cepstre

4.4.1 Normalisation de la moyenne

Si les caractéristiques fréquentielle de la chaîne de transmission ne sont pas plat on peut considère que le signal transmit est une version du signal original convolué avec la réponse impulsionnelle de la chaîne, donc on peut dire que le signal observé est corrompu par un bruit convolutif. Si les caractéristiques de la chaîne varient lentement par apport à celle du signal vocal on peut considérer ce bruit multiplicatif dans le domaine spectral et additif dans le domaine log – spectral, ainsi on peut le soustraire dans n'importe quel espace de paramètre qui représente le log – spectre du signal. La méthode la plus utilisé est CMS (cepstral mean substraction) [09]. Donc comme le cepstre représente le log –spectre du signal une normalisation sur la valeur moyenne des paramètres court terme est suffisante [21] :

$$\hat{X}_K = X_k(n) - \mu_k$$

Ou $X_k(n)$ est le k^{ème} composant du vecteur paramètre a l'instant segment n et μ_k est la moyenne du k^{ème} composant de tous les vecteurs paramètres pour la phase d'apprentissage ou test d'un son vocal spécifique.

4.4.2 Normalisation du gain

Modèle approximatif dans un bruit additif : Quand le signal vocal est corrompu par un bruit stationnaire additif inconnu et distorsion multiplicative inconnue, on peut supposer un modèle du signal dans l'environnement bruyant donné par l'équation suivante [22]:

$$X(n, e^{jw}) = S(n, e^{jw}).H(e^{jw}) + A(e^{jw}) \quad (4.5)$$

Où $X(n,e^{jw})$ est le spectre de puissance du son bruité à la fréquence e^{jw} et segment(fenêtre) du temps n, $S(n, e^{jw})$ est le spectre de puissance du son propre, $H(e^{jw})$ est la distorsion et $A(e^{jw})$ est le spectre de puissance du bruit additif. On assume que le bruit additif est non corrélatif avec le son articulé. Si on pose $E(n, e^{jw})=S(n, e^{jw}).H(e^{jw})$, on peut écrire La transformation log de l'Eq.1 comme suit :

$$\text{Log}\,(X\,(n, e^{jw})) = \text{Log}(E\,(n, e^{jw}) + A\,(e^{jw})) \tag{4.6}$$

La forme des trajectoires log (spectre) dans le son bruité $log(E\,(n, e^{jw}) + A\,(e^{jw})$ diminue en gain (distance entre les valeurs maximum et minimum) par contre une augmentation de la composante continue DC avec celui du son non bruité.

On tien considération que :

- Le spectre de puissance du son non bruité a $E\,(n, e^{jw})$ à une valeur minimale (les fréquences doivent être différentes de zéro.
- Le spectre de puissance du bruit additif $A\,(e^{jw})$ est plus grand que la valeur minimale $min(E(n, e^{jw}))$, c'est $A(e^{jw}) >> min(E(n, e^{jw}))$.

Dans le meilleur des cas, $min(E\,(n, e^{jw}))$ est une valeur nulle dans l'environnement propre. Cependant, la valeur nulle est convertie en infini dans le domaine log (spectre). Ce qui provoque des problèmes dans les étapes d'apprentissage ou de test du système de reconnaissance. Cette contrainte est fatale et la valeur minimale devient nécessairement un facteur important qui provoque un gain dans le log (spectres).

Le changement du gain G (e^{jw}) peut être exprimé comme suit :

$$G(e^{jw}) = \frac{G_N\left(e^{jw}\right)}{G_C e^{jw}} \tag{4.7}$$

$$G_N\left(e^{jw}\right) = \log\left(\frac{\max\left(E\left(n, e^{jw}\right)\right) + \left(A\left(e^{jw}\right)\right)}{\min\left(E\left(n, e^{jw}\right)\right) + \left(A\left(e^{jw}\right)\right)}\right) \tag{4.8}$$

$$G_C\left(e^{jw}\right) = \log\left(\frac{\max\left(E\left(n, e^{jw}\right)\right)}{\min\left(E\left(n, e^{jw}\right)\right)}\right) \tag{4.9}$$

Où G_N (e^{jw}) et G_C (e^{jw}) sont des facteurs de gain dans des environnements bruité et propre respectivement. Le gain de la parole propre est généralement déterminé par la valeur minimum $min(E\,(n, e^{jw}))$. D'autre part, le gain du son vocal bruité est déterminé par la valeur A (e^{jw}) *du bruit* et devient beaucoup plus petit que celui du signal propre a cause de :

$$A(e^{jw}) >> min(E(n, e^{jw}))$$

Les centrages de DC des sons vocaux propre et bruité peuvent être exprimés comme *suit:*

$$D_N\left(e^{jw}\right)=\frac{\sum_{n=1}^{L}\log\left(E\left(n,e^{jw}\right)+A\left(e^{jw}\right)\right)}{L} \qquad (4.10)$$

$$D_C\left(e^{jw}\right)=\frac{\sum_{n=1}^{L}\log\left(E\left(n,e^{jw}\right)\right)}{L} \qquad (4.11)$$

Où $D_N\left(e^{jw}\right)$ et $D_C\left(e^{jw}\right)$ sont des centrages de DC du son vocal bruité et propre respectivement et L est le nombre des trames de la parole. En conséquence, en ajustant les gains et les centrages DC., le log -spectre su son vocal bruité peut être exprimée par l'équation approximative suivante:

$$\log\left(X\left(n,e^{jw}\right)\right)=G\left(e^{jw}\right).\left[\log E\left(n,e^{jw}\right)-D_c\left(e^{jw}\right)\right]+D_N\left(e^{jw}\right) \qquad (4.12)$$

On outre

$$\log\left(X\left(n,e^{jw}\right)\right)\approx G\left(e^{jw}\right).\left[\log E\left(n,e^{jw}\right)\right]+B\left(e^{jw}\right) \qquad (4.13)$$

$B\left(e^{jw}\right)$ Et le biais de DC donner par :

$$B\left(e^{jw}\right)=D_N\left(e^{jw}\right)-G\left(e^{jw}\right).D_C\left(e^{jw}\right) \qquad (4.14)$$

Normalisation Quand l'apprentissage est exécuté dans un environnement propre et le test est évalué dans un environnement bruyant, les différences d'environnement dans log -spectres apprentissage et test peut être enlevée en ajustant le gain et les centrages D C. Le log -spectre ajusté est obtenu par l'équation suivante:

$$\log S'\left(n,e^{jw}\right)=\log E\left(n,e^{jw}\right)-D_C\left(e^{jw}\right) \qquad (4.15)$$

En environnement bruyant, $\log S'\left(n,e^{jw}\right)$ peut être obtenu en supprimant le gain $G\ (e^{jw})$ et le DC $D_N\ (e^{jw})$ selon Eq. (4.12). Le $D_N\ (e^{jw})$ peut être calculé à partir de la moyenne du log -spectres. Le gain $G\ (e^{jw})$ peut être éliminé en normalisant le gain propre et bruyant $G_N\ (e^{jw}) = 1$,

$G_C\ (e^{jw}) = 1$. Ces opérations peuvent être appliquées avec les paramètres cepstraux approximativement.

Une série de procédures est résumée dans les deux étapes suivantes, qui doivent être appliquées dans la phase d'apprentissage et de test.

Étape1 : Soustraire la moyenne des coefficients cepstraux. L'opération est connue en tant que normalisation moyenne cepstral (CMS) (cepstral mean substraction).

$$C'(n,k) = C(n,k) - \left(\sum_{n=1}^{L} C(n,k) \right) / L \quad for\ 1 \le k \le M \tag{4.16}$$

Etap2: Normalisez les gains en calculant les valeurs maximum et minimum des coefficients cepstraux. .

CGN est donné par l'équation :

$$C''(n,k) = C'(n,k) \Big/ \left(\max_{1 \le n \le L} C'(n,k) - \min_{1 \le n \le L} C'(n,k) \right) for\ 1 \le k \le M \tag{4.17}$$

Où k est la séquence paramètre dans un cepstre d'ordre M. Ces étapes sont aussi bien appliqués au delta cepstre et au delta-delta cepstre. CGN est semblable à la normalisation cepstral de variance (CVN). Tandis que CGN est basé sur un modèle approximatif.

La figure 4.5 montre l'influence de la normalisation sur la séquence du 5ème coefficient MFCC du mot 'thamanya' prononcé par un locuteur masculin.

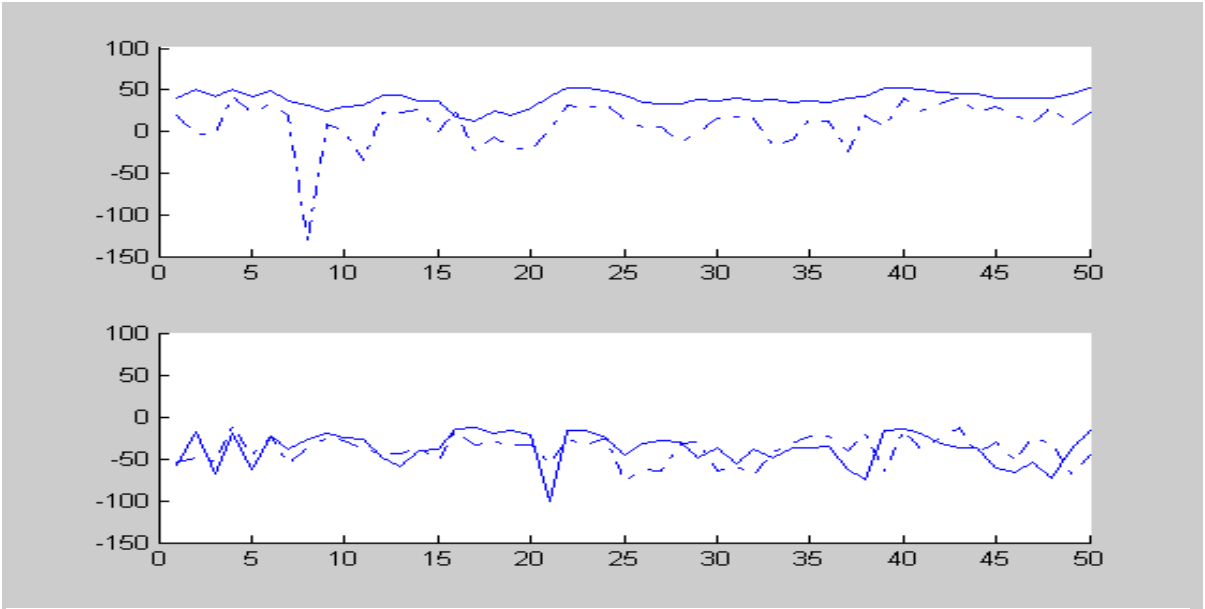

Figure 4.5 le 5ème coefficient MFCC du mot 'thamanya' propre et bruité (Snr=10) sans et avec normalisation

4.5 Prédiction linéaire perceptive RASTA

L'analyse RASTA PLP proposé par H.Hermansky [23] [24] en remplaçant le spectre absolu court –terme par un spectre estimé dont le quel chaque canal fréquentiel est filtré par un filtre passe bande d'amplitude nulle pour la fréquence zéro pour éliminer les variations lentes.
Les étapes de l'analyse RASTAPLP sont comme suite :

1. calcul le spectre d'amplitude en bandes critiques (comme pour la PLP).
2. compression de l'amplitude à l'aide d'une transformation non linéaire (log).
3. filtrage des trajectoires temporelles de chaque composante spectrale par un filtre RII ARMA.
4. expansion de l'amplitude à l'aide d'une transformation non linéaire (expo).
5. préaccentuation à l'aide du contour d'égale intensité sonore et prise en compte de l'échelle sonore par élévation à la puissance 0.33.
6. calcul du modèle tout pôle du spectre selon la méthode PLP classique.

L'idée principale ici est de supprimer les facteurs constants dans chaque composante spectrale courte terme comme le spectre auditif avant l'estimation du modèle tout pôle.
Les questions les plus importantes sont dans les étapes **2)** et **3),** c.-à-d. :

- dans quel domaine se fait le filtrage
- quel est le type du filtre à employer.

La fonction de transfert du filtre RII est donnée par [24]

$$H(z) = 0.1.\frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{4}.(1 - 0.98z^{-1})}$$

La fréquence de coupure basse du filtre détermine le changement spectral le plus rapide du log- spectre, qui est ignoré dans le rendement, tandis que la fréquence de coupure élevée détermine le changement spectral le plus rapide qui est préservé dans les paramètres de rendement.
On s'attend à ce que la partie passe-haut du filtre passe-bande allège l'effet du bruit convolutionnaire présenté dans le canal. Le filtrage passe-bas aide à lisser une partie de l'évaluation spectrale courte terme qui se trouve dans les changements spectraux rapides

entre segments due aux objets crées par l'analyse. En (4.6), la basse fréquence de coupure est de **0.26 hertz.** La pente de filtre diminue **6** DB /oct de 12.8 hertz avec des zéros pointus à 28.9 et à 50 hertz.

Le filtre **RASTA** a une constante de temps longue pour l'intégration (environ 500 ms). Il signifie que le résultat courant d'analyse dépend de son antécédent (c.-à-d., sur les sorties précédentes stockées dans la mémoire du filtre récursif **RASTA**).

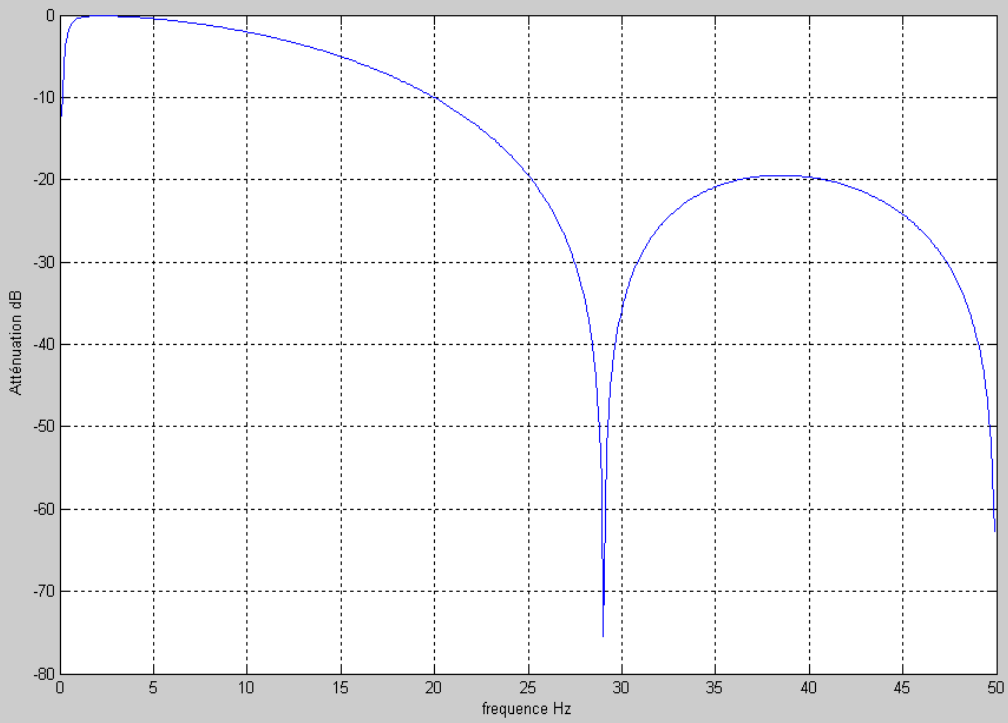

Figure 4.6 La réponse fréquentielle du filtre RASTA

4.6 Accouplement entre l'analyse LPC et MFCC

D'une vue d'ensemble des techniques d'analyse utilisées dans l'extraction des représentations (paramètres) il est clair que deux estimateurs spectraux sont couramment employés, c.-à-d la Transformée de Fourier rapide (FFT) et le codage linéaire de prévision (LPC). En effet le bruit additif avec un spectre assez doux près du spectre moyen du signal de parole tendra à compléter les vallées entre les formants dans le spectre du parole exprimé, mais il ne décalera pas les endroits des formants, c.-à-d. la gamme dynamique dans le spectre est réduite. Prendre le raisonnement donnée par [25] au sujet du comportement des formants en présence du bruit additif et de tenir compte que l'évaluation de LPC d'un spectre portant une enveloppe spectrale avec une description relativement précise des crêtes dans le spectre, bruit additif devrait avoir peu d'effet sur des évaluations spectrales de LPC. Le rapport donnée part [25] pourrait donc être une indication que, dans certains cas, les évaluations de LPC sont bien adaptées pour décrire les sons articulés bruités.

Cependant, il faudrait maintenir dans l'esprit que les coefficients de LPC décrivent les propriétés spectrales des sons articulés en termes de modèle (AR). Si le bruit est ajouté aux sons articulés propres, leurs propriétés spectrales peuvent ne plus être décrites exactement par les modèles (d'ordre réduit) d'AR. Des modèles évolués d'AR ou les modèles autorégressifs de la moyenne mobile (ARMA) sont exigés pour estimer exactement les propriétés spectrales des signaux bruités. L'avantage de la description précise de LPCs des crêtes spectrales de la parole se tiendra seulement tant que le modèle d'AR fournit une approximation raisonnable des données. Employer les dispositifs acoustiques basés sur LPC de faible SNR pourrait donc avoir comme conséquence des évaluations spectrales pauvres en information, parce que les modèles (d'ordre réduit) d'AR ne peuvent plus décrire les données d'une manière suffisamment exacte.

Une autre étude réalisée par [15], les auteurs ont rapporté que les dispositifs acoustiques dérivés des spectres de puissance de LPC sont meilleur que les dispositifs semblables calculés à partir des spectres de puissance de FFT. A leur avis, l'exécution supérieure des dispositifs basés sur LPC peut être attribuée au fait que le lissage spectral inséparable au modèle linéaire de prévision fournit un ensemble doux de paramètres qui ne représente pas les variations fines provoquées par les changements d'excitation.

Le lissage spectral semble être une propriété souhaitable des estimateurs spectraux, car le lissage spectral traduit en réduction du désaccord spectral qui alternativement, devrait

rapporter une réduction du désaccord des modèles acoustiques et donc une séparabilité meilleure de classe [15].

Cependant, l'application des bancs de filtres (Mel) réduit largement le désaccord dans des spectres de FFT, parce que les différents coefficients de FFT dans chaque bande de Mels sont ramenés à une moyenne. D'autre part, les évaluations spectrales de LPC, sont garanties pour être lisses dans le domaine fréquentiel par leur nature même. Cependant, en conditions bruyantes, la contribution relative du bruit peut être si grande que les propriétés spectrales du signal d'entrée soient principalement déterminées par le bruit. Dans ces circonstances, le spectre de deux segments adjacents peut différer largement, par ce qu'un estimateur LPC d'ordre fixe peut rapporter des modèles très différents des deux segments. Une partie du désaccord segment -a- segment présenté par ce phénomène peut être allégée par la Mel-fréquence en faisant la moyenne, mais il reste à voir si la compensation est suffisante.

Pousser par ces études l'algorithme LPCMFCC utilise l'analyse LPC pour remplacer le spectre court terme calculé par FFT par un nouveau spectre lisse calculé à base des paramètres LPC [15]. Comme nous l'avons vu l'analyse LPC offre une estimation du conduit vocal basé sur un système tous pole, tandis que l'analyse avec les paramètres MFCC approche le système auditif humain en plus de l'utilisation du cepstre qui sépare les signaux convolutifs dans le domaine temporel, donc un accouplement entre LPC et MFCC permet de rassembler plusieurs connaissances sur le signal parole analysé, la figure 5.4 nous donne une idée globale sur l'algorithme LPCMFCCs.

.

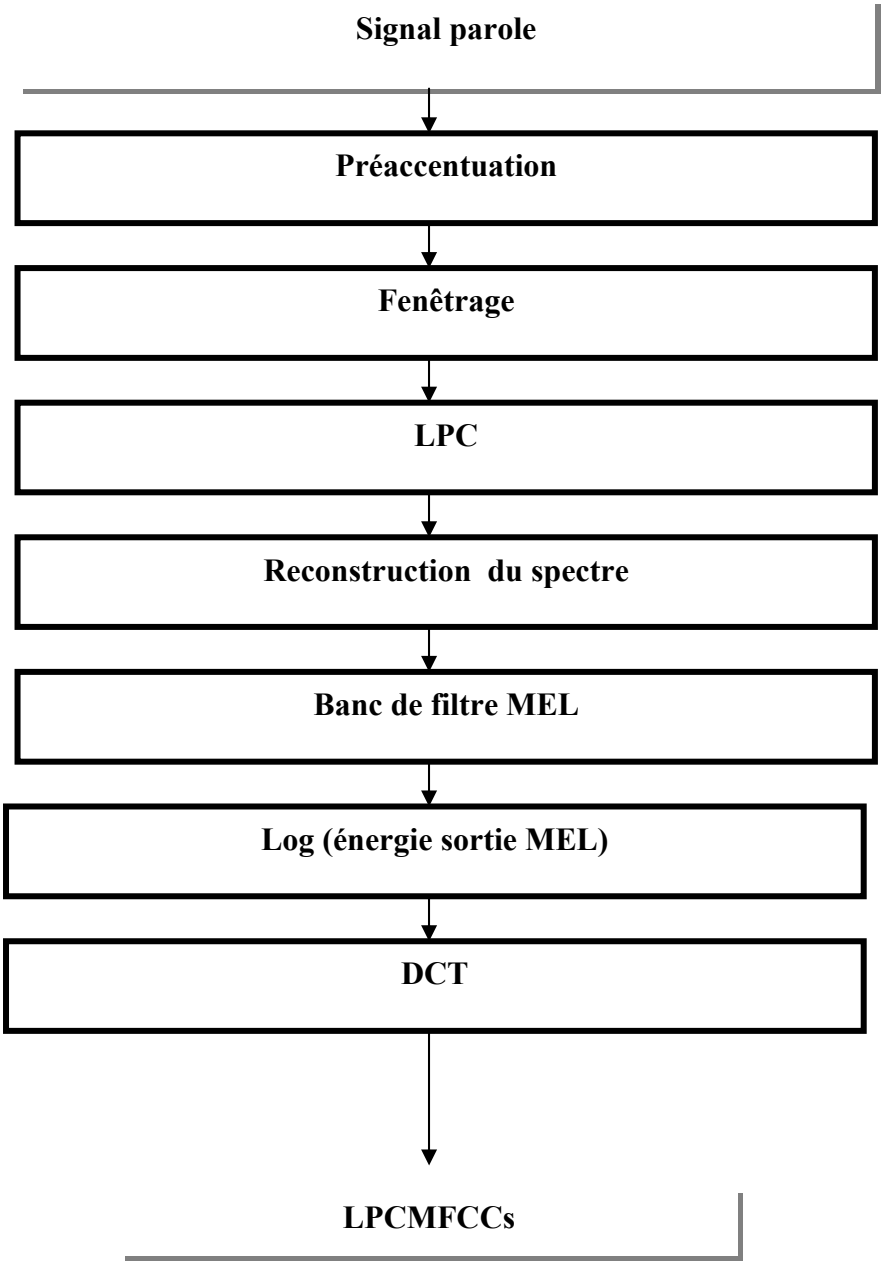

Figure 4.6 L'algorithme LPCMFCCs

Conclusion

Par la fin de ce chapitre nous avons vu une série de méthode de paramétrisation d'un signal vocale on s'appuyons sur des techniques comme la normalisation par (CGN,CGM), l'analyse multi résolution en ondelettes par (MBLPCC), le filtrage par (RASTAPLP) et le lissage spectrale par(LPCLFCC) . Les nouveaux paramètres ont été envisagés pour rendre les systèmes de reconnaissance de la parole plus robustes aux différentes variabilités acoustiques.

Résultats expérimentaux

5.1 Introduction

Ce chapitre se situe dans le cadre d'une étude expérimentale permettant de choisir parmi les paramétrisations utilisées actuellement dans les systèmes RAP celle qui donne les meilleures performances dans le cas de signaux propres et qui en même temps robuste aux dégradations causées par la présence de bruit ambiant. Au-delà de cet aspect, nous montrons dans ce chapitre, la dégradation des performances de reconnaissance en fonction de différents nivaux de bruit additif. Les différentes expériences réalisées dans ce mémoire, sont organisées dans trois étapes:

- **Etape 1** : Une étude comparative entre trois méthodes classiques connues dans le système RAP (LPCC, MFCC et PLP) afin de fixer le choix des différents variables utilisés en extraction des paramètres .

- **Etape 2** : Une étude réalisée pour montrer la dégradation de performance des systèmes de reconnaissance a la présence du bruit additif.

- **Etape 3** : On présente une série des techniques qui peuvent être utilisées comme solutions pour augmenter la performance du système RAP.

5.2 La reconnaissance automatique de la parole

Un système RAP se réalise en deux phases c –à –d apprentissage et classification.

5.2.1 Apprentissage des modèles

Pendant l'apprentissage, les modèles des signaux paroles sont pratiquement obtenus à partir d'une large base de données d'apprentissage. Les signaux paroles qui sont produits par

l'appareillage de production de la parole humain ont des propriétés acoustiques bien définies. Cependant, il y a de nombreux facteurs qui causent la variation dans les propriétés acoustiques des signaux de la parole typiques. Par exemple, deux personnes différentes ne produiront exactement jamais la même expression de la même manière, c'est en raison des différences évidentes entre eux, par exemple, des voix femelles de fin masculine, des voix des enfants, des adultes et de vieilles personnes, etc.

Il est également vrai que le même locuteur puisse lui même ne pas pouvoir produire la même expression exactement de la même manière. Pour que la parole puisse être reconnue correctement, il est donc nécessaire de modeler leur caractéristique acoustique (valeur moyenne) aussi bien que la variation observée dans ces propriétés caractéristiques (désaccord). Les modèles statistiques sont bien convenus pour décrire un tel comportement. C'est probablement la raison pour la quelle la plupart des systèmes RAP sont des machines statistiques d'assemblage de modèle.

Modèles de Markov cachés (HMMs).

Les modèles statistiques qui sont le plus souvent employés dans les moteurs des systèmes d'identification 'RAP sont les modèles de Markov cachés (HMMs). HMMs constituent un cadre pour caractériser l'évolution en temps des modèles statistiques. Dans le domaine 'RAP, HMMs sont devenus particulièrement populaires en raison de l'efficacité avec laquelle ils modèlent la variation des propriétés statistiques de la parole dans les deux domaines temporel et fréquentiel simultanément.

Un HMM peut être décrit comme collection d'états reliés par des transitions [07, 08,19]. Chaque état modélise un processus stochastique et les transitions entre les états correspondent à une séquence de temps dans lequel les processus stochastiques se produisent. La théorie des HMMs exige des processus stochastiques qui sont modelés par des états de HMM pour être stationnaires. C.-à-d. leurs propriétés statistiques ne devraient pas changer en fonction du temps. Le discours humain n'est pas un processus stochastique stationnaire. Il est donc possible de supposer des intervalles de temps pendant lesquels les sons articulés sont approximativement stationnaires. Dans ces derniers ainsi - appelé intervalles quasi - stationnaires. Il est raisonnable de supposer que les propriétés acoustiques des sons articulés peuvent être modelées en termes de distributions statistiques

Deux processus stochastiques se produisent concurremment dans un HMM. Le premier processus traite la séquence dans laquelle les états HMM se produisent et modélise la

structure temporelle des sons prononcés. Les statistiques de ce processus sont modelées par des probabilités de transition. Le seconde ou processus de sortie modélise les propriétés quasi –stationnaires des sons articulés et il est décrit par une fonction de densité de probabilité conditionnelle de sortie (FDP). Le FDP de sortie définit la probabilité conditionnelle d'émettre une observation (c.-à-d. un vecteur acoustique de paramètres), étant donné que le modèle est dans l'état spécifique.

. Un modèle Markov caché peut être caractérisé par les éléments suivants :

- N, le nombre d'état du système. On peut noter ici l'ensemble des états du modèle : S= {S1, S2,..., SN}, et qt comme l'état du système au temps t.
- M, le nombre de symboles d'observations distingués par état. Les symboles d'observations correspondent à chaque sortie physique du système réel qu'on veut modéliser. On peut noter ici l'ensemble des symboles d'observations du modèle : V= {V1, V2,..., VM}.
- La matrice de probabilité de transition d'état A = {aij} telle que :
  aij = P[qt+1 = Sj | qt = Si] avec $1 \leq i, j \leq N$ et $aij \geq 0$.
- La distribution à l'état j de probabilité des symboles d'observation B = {bj(k)} dont : bj(k) = P [vk à t | qt = Sj], avec $1 \leq j \leq N$ et $1 \leq k \leq M$.
- La distribution d'état initial $\pi = \{\pi i\}$ dont : $\pi i = P [q1 = Si]$, avec $1 \leq i \leq N$.

Etant donné les valeurs de N, M, A, B et πi le modèle de Markov Caché peut générer la série d'observations O = O1, O2,..., OT avec Ot Є V et T le nombre d'observations faites. On appelle $\lambda = (A, B, \pi)$ les paramètres complètes du modèle HMM.

Topologie du modèle: la parole est un exemple d'un signal dont les propriétés changent d'une façon successive avec le temps. De tels signaux sont mieux modelés par le prétendue HMMs gauche –à– droit [15] dans lequel l'ordre fondamental d'état a la propriété que pour une augmentation d'index temps, l'index d'état augmente ou il reste le même. Les HMMs gauche - à - droit ont été employées dans toutes les expériences discutées dans ce travail.

Si on veut modéliser V classes (phonèmes, mots ou autres unités acoustiques), chaque

Classe sera modélisée par un modèle de Markov Caché distinct. Donc on doit accomplir les procédures suivantes :

- Pour chaque classe ***v***, on doit déterminer un modèle de Markov Caché $\lambda = (A, B, \pi)$ qui est optimal pour des vecteurs d'observation de la classe ***v***. C'est la phase d'apprentissage du système.

- Pour chaque occurrence dont on veut reconnaître la classe, on détermine d'abord les vecteurs acoustiques et puis on calcule toutes les probabilités des modèles possibles $P(O|\lambda^v)$ avec $1 \leq \boldsymbol{v} \leq V$ et on choisit la classe dont le modèle donne la probabilité Maximum :

$$v^* = \arg\max_{1 \leq v1 \leq V} \left[\left(P\left(O/\lambda^v\right) \right) \right]$$

On utilise l'algorithme de Viterbi pour calculer ces probabilités et la complexité et O ($V*N^2*T$) avec V, le nombre de classes, N le nombre d'états et T, le nombre de symboles d'observations.

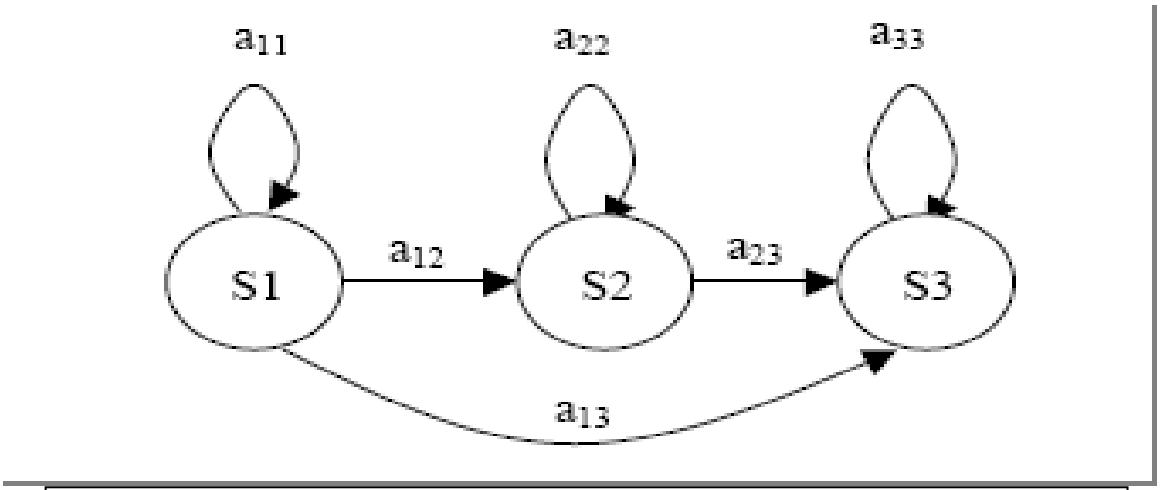

Figure 5.1 Exemple d'un HMM gauche-droit

5.2.2 Classification des modèles

Dans les systèmes RAP, une distinction est souvent faite entre la classification et l'identification. Par exemple, toutes les expériences dans ce travail sont limitées à la classification des chiffres arabe, c.-à-d. un son parole non classifié doit être assigné à un des 10 classes possibles de chiffre. La classification est donc limitée à assigner des étiquettes de classe aux données qui ont été déjà segmentées, tandis que l'identification implique la segmentation et la classification.

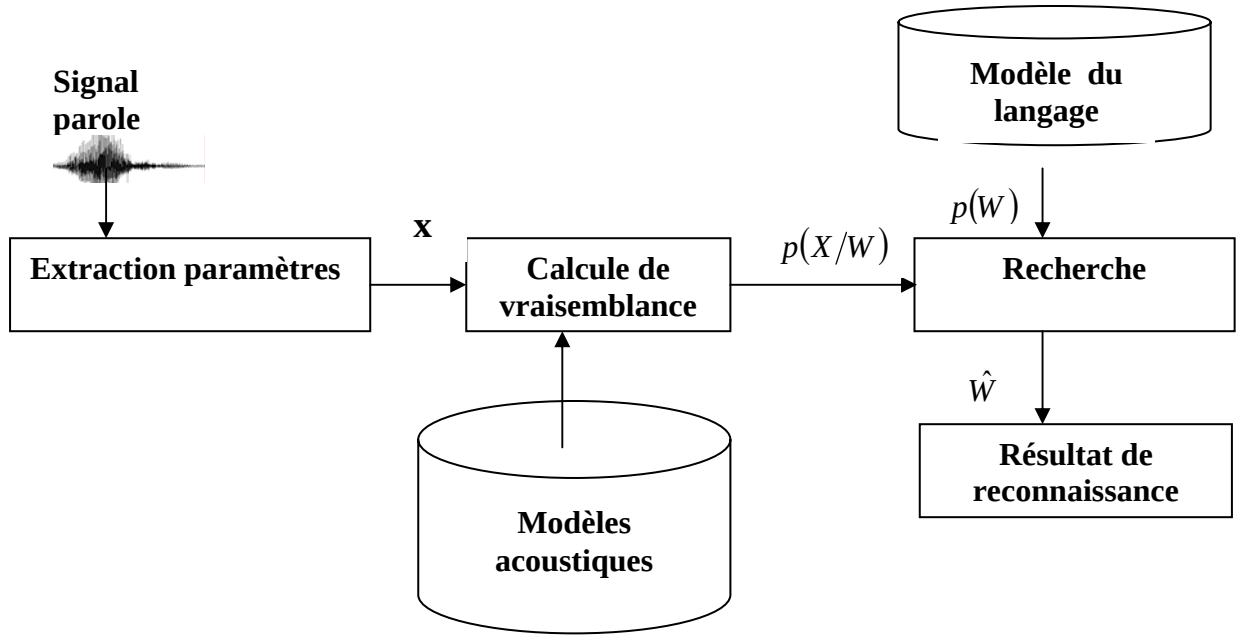

Figure 5.2 Vue graphique globale d'un système RAP

Le schéma de la figure 5.2 donne une vue générale sur un système RAP. Comme est illustré sur ce schéma le système se compose d'un module d'extraction de paramètre, des modèles acoustiques, et d'un modèle de langage.

Les HMMs présentés dans les sections précédentes ont été employés en tant que modèles acoustiques dans toutes les expériences qui ont été réalisé. Cette section fournit une brève vue de la manière de laquelle les HMMs, et le modèle de langage sont employés pour identifier automatiquement la parole.

Statistiquement parlant, la tâche de reconnaissance d'un système RAP peut être décrite comme la recherche d'un mot $\hat{W}$ qui optimise la probabilité a posteriori p (W/x) dans une séquence des mots W a été produite et que la séquence des vecteurs acoustiques de paramètre X a été observée:

$$\hat{W} = \arg\max_{W \subset w} P(W/X) \qquad (5.1)$$

Où w est l'ensemble de toutes les séquences possibles de mot, W = w_1, w_2..,w_n...,w_N est la chaîne de N mots à reconnaître et X = x_1,x_2,..,x_n...,x_T est la séquence des vecteurs paramètre acoustiques observés pendant l'intervalle T.

En langage naturel, le nombre des séquences possibles de mot qui peuvent se produire est extrêmement grand. D'ailleurs, les propriétés acoustiques des sons prononcés qui sont produits quand des différents locuteurs prononcent la même séquence des mots peuvent changer énormément. Dans des applications pratiques il est donc non faisable de modeler p(W/x) directement. Cependant, en utilisant la règle de bayes, Eq. 1.1 peut être récrit comme:

$$\hat{W} = \arg\max_{W \subset w} \left(\frac{P(X/W)P(W)}{P(X)} \right) \tag{5.2}$$

Où P(x) est la probabilité a priori d'un ensemble de vecteurs acoustiques de paramètre, X. pendant l'identification d'une expression donnée. On assume que P(x) est constant pour tout le W. Par conséquent Eq 1,2 se simplifie à:

$$\hat{W} = \arg\max_{W \subset w} P(X/W)P(W) \tag{5.3}$$

Selon Eq. 1,3, $\hat{W}$ est déterminé par deux probabilités: P(W) et P(X/W). P (W) est la probabilité a priori d'observer la séquence de mot W est modelée par le modèle de langage dans le système. P(X/W) est la probabilité de vraisemblance qu'on observera l'ensemble de vecteurs acoustiques X quand la séquence de mot W est produite. La probabilité P(X/W) peut être estimée à partir des bases volumineuses des données d'apprentissages et elle est représentée par des modèles acoustiques dans le système, par exemple. HMMs.

- **Modèle du langage** La signification du modèle langage a priori P(W), dans Eq5.3 dépend de la complexité de la tache de recognition .Dans la reconnaissance des chiffres arabe qui est le sujet de cet mémoire , les digits étaient tout l'également probables, ce qui s'explique par une probabilité unique $p(W_i)$ pour tout les digits.
- **Algorithme de recherche** Les systèmes RAP basés sur des méthodes de classification du modèle statistique utilisent essentiellement un processus de décodage acoustique qui recherche $\hat{W}$ pour maximiser Eq 5.3.Dans toutes les études qui sont rapportées dans ce travail, l'algorithme de Viterbi qui est basé sur des principes de programmation dynamiques est appliqué pour trouver le chemin le plus probable par l'espace de recherche engendré par X et W

5.3 Base des données

La base des données employée pour l'évaluation des différentes expériences est celle présentée dans [08]. A partir de cette base nous avons construit deux sous bases une pour l'apprentissage et l'autre pour le test.

1. Base1 construite Pour la phase d'apprentissage contenant 42 locuteurs (21 femmes et 21hommes) chaque locuteur prononce 10 occurrences pour chaque mot.

2. Base2 construite Pour la phase test contenant 42 locuteurs (21 femmes et 21 hommes) chaque locuteur prononce 10 occurrences pour chaque mot.

Le Tableau 5.1 donne quelque information sur le vocabulaire utilisé dans la base

Le chiffre	Signification En arabe	Signification En français	symbole
SIFR	صفر	ZERO	'0'
OUAHID	واحد	UN	'1'
ITHNAN	اثنان	DEUX	'2'
THALATHA	ثلاثة	TROIS	'3'
ARABAA	أربعه	QUATRE	'4'
KHAMSSA	خمسه	CINQ	'5'
SETAA	ستة	SIX	'6'
SABAA	سبعه	SEPT	'7'
THAMANYA	ثمانية	HUIT	'8'
TISAA	تسعه	NEUF	'9'

Tableau 5.1 le vocabulaire utilisé dans la base

Les caractéristiques d'enregistrement sont comme [08] :

- **Type de ressource :** enregistrement de la parole (acoustique).
- **Nature de parole :** lus normalement (préparée n'est pas spontanée)
- **Langue :** arabe
- **Environnement :** pièce calme (bureau)
- **Pc compatible :** processeur intel Pentium III fréquence d'horloge 825Hz
- **Carte son :** carte intégrée
- **Plat forme :** Windows xp
- **Microphone :** casque microphone GALAXY AP-860
- **Fréquence d'échantillonnage :** 11025hz (choix qui vérifier le théorème de Shannon)

- **Logiciel d'enregistrement :** Goldwave.
- **Support électronique** : CD ROM
- **Le format de fichier :** des échantillons de son sont stockés sous un FORMAT WAV.
- **Résolution :** 16bits.

Le modèle HMMs 16 états de R.Bakis gauche -droit présenté dans la figure 5.1 est employé pour toutes les expériences réalisées au niveau de ce travail.

On raison du temps de calcul le nombre de Mixture =1(une seule Gaussienne pour modeler chaque état).

Pour ce qui concerne le calcule du taux de reconnaissance on a simplement sommé la reconnaissance des différents digits dans la base.

Taux _Rec _glob = REC_0 + REC_1 + ……. + REC_9.

5.4 Résultats et expériences

- **Etape 1 :**

Exp. 1 :

Cette expérience est faite pour fixer le coefficient de préaccentuation α, une fenêtre d'analyse Hamming 20 ms de largeur pour MFCC, 25ms pour LPCC et PLP avec 10 ms de chevauchement pour les trois ensembles, 12 paramètres statiques (LPCC, MFCC, PLP) sont utilisés. Pour ce qui concerne le choix des valeurs de test pour le coefficient α on a choisi des valeurs beaucoup utilisés dans les systèmes RAP.

	12 LPC Cs	12 MFC Cs	12 PLP
Sans préaccentuation	92.1739	91 ,9565	93,6957
α =0.935	89,3478	93,6957	91.9565
α =0.950	89,3478	94.5652	92,1739
α =0.970	88,4783	94,7826	91,7391
α =0.980	88,6957	94,7826	92,3413

Tableau 5.2 influence du coefficient de préaccentuation sur la reconnaissance

D'après les résultats obtenus dans le tableau 5.1 on voit clairement que l'utilisation d'une opération de préaccentuation augmente le taux de reconnaissance de 1 à 3%.

On remarque aussi que $\alpha = 0.98$ est un choix judicieux pour les deux types de paramètre (MFCCs, PLP).

Les coefficients LPCCs donnent un taux de reconnaissance maximal lorsque la préaccentuation n'est pas utilisée.

Les coefficients MFCCs au contraire des coefficients LPCCs donnent un taux de reconnaissance max lorsque $\alpha = 0.98$ ou $\alpha = 0.97$ et la non utilisation d'une opération de préaccentuation diminue le taux de reconnaissance de 3% par rapport à ce lui marqué lorsque $\alpha = 0.98$, les coefficients PLP sont moins touchés par le changement, l'utilisation ou non de l' opération de préaccentuation

Exp.2

Dans cette expérience on fixe le coefficient de préaccentuation α =0.95 et on essaye de voir l'influence des paramètres de chevauchement et la longueur de fenêtre d'analyse utilisés sur le taux de reconnaissance globale, la fenêtre de Hamming est --utilisée pour réaliser l'opération du fenêtrage, les valeurs de chevauchement et longueur de la fenêtre testés sont les plus utilisées dans les systèmes RAP. Les résultats sont rassemblés dans le tableau 5.3.

	12 LPCCs			12 MFCCs			12 PLP		
	20 ms	25 ms	30 ms	20 ms	25ms	30ms	20ms	25ms	30ms
10 ms	87.826	89.3478	88,9130	94,7826	94,5652	93,0435	92,1739	93,6957	92,1739
12,5ms	87,8261	89,3478	85,8696	94.7826	94.5652	93.6957	91.7391	93.6957	91.7391
15 ms	88,269	88.0435	88,0435	94.5652	94.7826	94.5652	92.1739	92.8261	92.1739

Tableau 5.3 influence des paramètres de chevauchement et la longueur de la fenêtre

Les résultats du tableau 5.3 montrent que les deux paramètres ont une influence positive sur le taux de reconnaissance globale si un bon choix de largeur et chevauchement est pris en compte.

Donc avant toute analyse du signal de parole, la procédure communément utilisée est la suivante:

- Filtrage passe – bas anti –repliement à Fe/2.
- Echantillonnage à Fe et quantification sur 16 bits.
- Préaccentuation par un filtre d'équation $1-\alpha z^{-1}$ pour aplatir le spectre moyen du signal qui décroit de 6 dB par octave.
- Fenêtrage.

La taille de la fenêtre d'analyse doit être supérieure ou égale au double du pas d'analyse. Dans nos expériences suivantes, nous avons utilisé un pas d'analyse de 10ms: à une fréquence d'échantillonnage de 11025 Hz. Ce la correspond à un pas d'analyse de 11025*0.010 ≈110 points.

Nous avons choisi une fenêtre d'analyse de taille L=11025*0.025 ≈ 275 points, c'est –à- dire 5/2 le pas d'analyse (25ms). Le choix d'une telle taille est justifié par le fait que la fenêtre d'analyse doit pouvoir contenir au moins deux périodes de la fréquence fondamentale des voix les plus basses, on a ainsi L >= 0.0235 ms pour une fréquence fondamentale à 85 Hz.

Le gabarit de la fenêtre appliquée au signal est choisi de manière à minimiser la discontinuité aux extrémités de chaque trame. La fenêtre la plus utilisée est celle de Hamming (voire Eq 3.10).

➢ Etape : 2

Comme nous l'avons discuté précédemment cette étape est réservée pour prouver la dégradation de performance du système de reconnaissance en présence du bruit, le bruit utilisé pour tester les approches de compensations présentées dans ce mémoire est de type additif blanc centré, additionné artificiellement aux signaux de la parole. Ce ci est dû au fait qu'il est difficile d'avoir les mêmes mots prononcés par des locuteurs avec différents rapports signal sur bruit. Mais additionner artificiellement du bruit sur les signaux, simule très bien le cas réel. Le SNR est utilisé pour ajuster le niveau de bruit, 39 paramètres sont utilisés :

12+ log E + D +DD = 39 paramètres.

On plus dans cette étape on peut voir l'influence des paramètres log -énergie et les dérivés première et seconde.

	39 LPCCs	39 MFCCs	39 PLP
Clean	96.7391	99.3478	99.3478
SNR=30 dB	90.4348	99.1304	93,9130
SNR=20 dB	86.0870	92.3913	88.6957
SNR=10 dB	78.0435	58.4783	67.3913
SNR=05 dB	56.3048	25.6522	37.3913

-

Tableau 5.4 L'influence du bruit sur le taux de reconnaissance

Une comparaison des résultats obtenus en utilisant la LPCC avec ceux en utilisant la MFCC ou la PLP (Tableaux 5.2, 5.3 et 5.4), montre clairement la supériorité de la MFCC et la PLP par rapport à la LPCC. Ce ci certainement du au fait que la LPCC ne tient pas compte des caractéristiques non linéaires de l'oreille, notamment de la résolution fréquentielle qui doit être une fonction logarithmique de la fréquence linéaire (Hertz). En effet, l'information contenue dans les basses fréquences est plus discriminante que celle qui se trouve dans les hautes fréquences.

Nous remarquons aussi que la différence, en termes de taux de reconnaissance, entre la LPCC et les autres techniques de paramétrisation, est plus importante dans les conditions propres ou faiblement bruitées:

- 2.6 % quand les données de la reconnaissance sont propre.
- De 2 à 9 % quand les données de la reconnaissance sont bruitées. (SNR entre 20 et 30dB).

Lorsque les données d'apprentissage et test sont propres, les résultats obtenus par la MFCC et la PLP sont pratiquement identique. En revanche des données d'apprentissages propres et des données bruitées, la MFCC donne de meilleurs résultats:

- quand le SNR dans le test est de 30 dB, la MFCC donne un taux de reconnaissance 99.1304 et la PLP donne un taux de reconnaissance de 93.9130.
- quand le SNR dans le test est de 20 dB, la MFCC donne un taux de reconnaissance 92.3913 et la PLP donne un taux de reconnaissance de 88.6957.

Les résultats obtenus montre aussi clairement que le gain résultant de l'utilisation des paramètres logE et dérivés notamment pour les MFCCs et PLP (99.3478 dans des conditions propres) est rapidement perdu lorsque les signaux de test sont bruités et aucune des trois paramètres (LPCCs, MFCCs, PLP) ne présentent une résistance au bruit additif, cette chute du taux de reconnaissance revient à la dissemblance entre les données d'apprentissage et de test. Ainsi on peut dire que les trois groupes de paramètres (LPCCs, MFCCs, PLPs) sont sensibles aux environnements non stricts qui approchent un peu le monde réel, pour résoudre ce problème deux méthodes sont proposés :

- Utilisation des techniques de correction appliquées après l'extraction des paramètres classiques (LPCCs, MFCCs, PLP), dans ce cas deux techniques sont évaluées dans ce travail (CMS et CGN).
- Extraire des nouveaux ensembles de paramètre plus sable et ont une résistance plus puissante contre les changements d'environnement, par ce principe trois autres ensemble de paramètres sont utilisées (RASTAPLP, MBLPCCs, LPCMFCCs)

➢ Etape 3 :

Les études précédentes sur la robustesse des techniques de paramétrisation (MFCC, PLP, LPCC), nous permet d'affirmer qu'une paramétrisation de type MFCC ou PLP donne de meilleurs résultats qu'une paramétrisation de type LPCC. Cependant, quel que soit le niveau de bruit affectant les données d'apprentissage, et quel que soit le type de paramètres acoustiques choisis, le système de reconnaissance de la parole se trouve sérieusement perturbé, surtout pour des rapports signal sur bruit faible. L'étude de la mise en œuvre de techniques permettant de compenser les effets du bruit est donc indispensable pour garantir une bonne robustesse d'un système de reconnaissance lors d'une application dans des conditions acoustiques d'utilisation difficiles.

Dans cette étape on va utiliser une série de techniques pour surmonter les problèmes rencontrés dans les expériences précédentes (taux de reconnaissance faible + dégradation de performance en présence de bruit additif).

Techniques utilisées :

- Normalisation de la moyenne des coefficients cepstraux avec CMS.
- Normalisation du Gain cepstral avec CGN.

- Le filtrage RASTA avec RASTAPLP
- L'analyse multi -résolution par ondelette MBLPCCs.
- Accouplement entre l'analyse LPC et MFCC (LPCMFCCs).

Pour toutes les expériences qu'on va réaliser la fenêtre de Hamming, 25ms de longueur et 10ms de chevauchement sont utilisés pour la segmentation. La préaccentuation $\alpha = 0.98$ est utilisée seulement avec les coefficients MFCC et PLP, LPMFCC, les coefficients LPCCs, MBLPCCs et RASTAPLP ne nécessite aucune opération de préaccentuation

- **Utilisation de la normalisation de la moyenne CMS**

Dans cette expérience une normalisation de la moyenne des coefficients cepstraux est utilisée sur un ensemble de 12 paramètres statiques en suite sur 39 paramètres.

	12 LPCC	12 MFCCs	12 PLP
clean	95.6522	98.4783	98.0435
SNR=30 dB	87.8261	96.3043	95.0000
SNR=20 dB	79.3478	87.6087	83.4783
SNR=10 dB	49.5652	46.7391	53.2609
SNR=05 dB	16.9565	20.2174	33.4783

Tableau 5.5 Résultats de l'utilisation d'une normalisation de la moyenne CMS sur 12 paramètres

	39 LPCC	39 MFCCs	39 PLP
clean	95.6522	99.3478	99.3478
SNR=30 dB	95.8696	98.0435	98.0435
SNR=20 dB	94.3478	95.2174	95.2174
SNR=10 dB	82.3913	58.2609	68.9130
SNR=05 dB	49.7826	21.7820	33.3913

Tableau 5.6 Résultats de l'utilisation d'une normalisation CMS sur 39 paramètres

D'après les résultats des deux tableaux 6.5 et 6.6 il est très clair que l'utilisation d'une opération CMS au niveau des paramètres du cepstre donne une amélioration remarquable sur le taux de reconnaissance global pour les trois ensembles de paramètres même si nous utilisons que 12 paramètres statique, la résistivité contre le bruit additif est aussi augmentée et le système de reconnaissance garde sa performance lorsque un SNR >=20 dB est utilisé, pour SNR<20dB le système de reconnaissance perd sa signification , cette dégradation explique clairement que la technique CMS est utilisée seulement pour éliminer les effets convolutifs tel que le filtrage des canaux de transmission et pour éliminer les effets additifs il faut intégré des autres moyens ou techniques.

- **Utilisation de la normalisation du GAIN**

Dans cette expérience une normalisation du Gain des coefficients cepstraux **CGN** indiquée dans le chapitre précédent est utilisée sur un ensemble de 39 paramètres + **CMS**
(12 statiques +log -énergie +D+DD =39 + CMS) les résultats sont rassemblées dans le tableau suivante :

	39 LPCCs	39 MFCCs	39 PLP
Clean	96.9565	98.6057	98.0435
SNR=30 dB	96.5217	97.6087	97.8261
SNR=20 dB	93.4783	93.0435	94.1304
SNR=10 dB	83.2609	76.5217	78.6957
SNR=05 dB	74.7826	63.2609	65.0000

Tableau 5.7 Résultats de l'utilisation d'une normalisation CMS+ CGN sur 39

On voit très bien que CGN n'apporte aucune amélioration dans les conditions propres ou en présence d'un bruit faible. Cependant en présence d'un bruit fort SNR < 20 dB une amélioration du taux de reconnaissance jusque 25% pour LPCCs, 33% pour PLP et
42 % pour MFCCs dans le cas de SNR=05 dB comparés à ceux marqués dans les expériences précédentes.

Les résultats des trois Tableaux (5.5, 5.6, 5.7) prouvent que l'utilisation d'une opération de normalisation (CMS+CGN) on une influence directe sur la performance du système de reconnaissance.

- **Utilisation du filtrage RASTA :**

La 3ème technique évaluée est celle introduite par H.hermansky (1994) pour améliorer le taux de reconnaissance donné par les coefficients PLP (voire chapitre précédant), la préaccentuation n'est pas nécessaire, 12 coefficients RASTA PLP + log(E) +D+DD =39 sont utilisées pour valider l'efficacité de cette technique, les résultats sont rassemblés au Tableau 5.8

	39 PLP	39 PLP+CMS	39 PLP+CMS+CGN	39 RASTA PLP
Clean	99.3478	99.3478	98.0435	98.6957
SNR=30dB	93.9130	98.0435	97.8261	99.1304
SNR=20dB	88.6957	95.2174	94.1304	97.1739
SNR=10dB	67.3913	68.9130	78.6957	86.0870
SNR=05dB	37.3913	33.3913	65.0000	59.5652
Moyenne	77.3478	78.9826	86.7391	88.1304

Tableau 5.8 Résultats d'utilisation d'un filtrage RASTA sur les coefficients PLP

Les résultats du tableau 5.7 montrent la performance rapportée par RASTAPLP dans les milieux propre et bruités comparé aux PLP, PLP+CMS et PLP+CMS+CGN. Par exemple RASTA PLP donne une amélioration de 2% pour un SNR =20 dB comparé à ce lui marqué par PLP+CMS et 8% pour un SNR=10dB comparé à ce lui marqué par PLP+CMS+CGN, pour un niveau de bruit élevé (SNR=05) les coefficients RASTAPLP perdrent leurs puissance devant les paramètres PLP normalisées.

- **Utilisation de l'analyse multi résolution :**

Comme nous l'avons vus il est très claire que les résultats de la reconnaissance obtenus par l'utilisation des paramètres LPCCs sont presque les inferieures, afin d'amélioré le taux de reconnaissance l'algorithme multi résolution MBLPCCs discuté dans le chapitre précédant est utilisé.

Les coefficients du filtre passe bas utilisés pour l'analyse par DWT sont :

L= [-0.0001 0.0007 -0.0004 -0.0049 0.0087 0.0140 -0.0441 -0.0174 0.1287 0.0005 -0.2840 -0.0158 0.5854 0.6756 0.3129 0.0544] ;

Les coefficients du filtre passe haut sont calculés comme suite :

$$H_k = (-1)^k h_{n-1-k} \quad k=0,1\ldots, n,$$

H= [-0.0544 0.3129 -0.6756 0.5854 0.0158 -0.2840 -0.0005 0.1287 0.0174 -0.0441 -0.0140 0.0087 0.0049 -0.0004 -0.0007 -0.0001].

Ces deux filtres sont utilisés simplement dans MATLAB on appelle l'ondelette mère 'db8' par la fonction DWT intégrée aussi dans MATLAB.

MBLPCCs=13 (pour la bande 0-5500 Hz) + 13 (pour la bande 0-2250 Hz +13 (pour la bande 0 -1125 Hz) =39 coefficients sont utilisés suivi d'une normalisation de la moyenne CMS, la préaccentuation n'est pas introduite, les résultas sont résumés dans le tableau 5.8

	39 LPCCs +CMS	39 LPCCs +CMS+CGN	39 MBLPCCs +CMS
Clean	95.6522	96.9565	98.0435
SNR=30dB	95.8696	96.5217	97.1739
SNR=20dB	94.3478	93.4783	95.6522
SNR=10dB	82.3913	83.2609	88.6957
SNR=05dB	49.7826	74.7826	70.6522
Moyenne	83.6087	89	90.0435

Tableau 5.9 Comparaison entre LPCCs simple et MBLPCCs

D'après les résultats du tableau 5,9 l'utilisation de l'analyse multi résolution n'améliore pas seulement le taux de reconnaissance (95.6522 ----> 98.0435) dans les conditions propre mais aussi augmente la résistivité lorsque les signaux de test sont bruités, par exemple dans le cas SNR=05 dB une amélioration de 21% est atteinte.

- **Accouplement entre l'analyse LPC et MFCC (LPCMFCCs).**

La dernière technique évaluée dans ce travail est LPCMFCC, cette méthode est une combinaison entre les techniques (LPC, MFCC, CMS, CGN), les résultats sont rassemblés

dans le tableau 5.10 pour le cas d'une préaccentuation $\alpha = 0.98$ et 5.11 dans le cas d'aucune préaccentuation.

	LPCMFCC+CMS	LPCMFCC+CMS+CGN
Clean	98.2609	97.3913
SNR=30dB	98.2609	96.9565
SNR=20dB	96.9565	95.4348
SNR=10dB	87.8261	90.5652
SNR=05dB	70.8696	78.2609

Tableau 5.10 Résultats rapportées par la méthode LPCMFCCs $\alpha = 0.98$

	LPCMFCC+CMS	LPCMFCC+CMS+CGN
Clean	98.6957	98.4783
SNR=30dB	98.2609	98.2609
SNR=20dB	97.8261	97.3913
SNR=10dB	82.6087	92.1739
SNR=05dB	61.0870	83.4783
Moyenne	87.6957	93.9565

Tableau 5.11 Résultats rapportées par la méthode LPCMFCCs $No \quad \alpha$

Il est très clair que la combinaison entre les paramètres (LPC, MFCC, CMS, CGN) donne aux coefficients extractées par LPCMFCC une puissance qui garde la performance du système presque dans toutes les situations, on peut remarquer aussi que l'utilisation d'une opération de préaccentuation ne donne pas une amélioration comme le cas des coefficients MFCCs, par contre la préaccentuation dégrade la performance du système

5.5 Comparaison et discussion générale :

Le but de ce travail est de trouver le vecteur paramètre le plus fiable qui assure la performance du système de reconnaissance dans les différents environnements, trois ensembles de vecteur paramètre les plus utilisés dans les systèmes RAP (LPCC, MFCC, PLP) ont été utilisés, ensuite on a exécuté plusieurs techniques de renforcements afin d'obtenir de nouveaux vecteurs paramètres robustes.

- **Opérations sur le vecteur LPCC :**

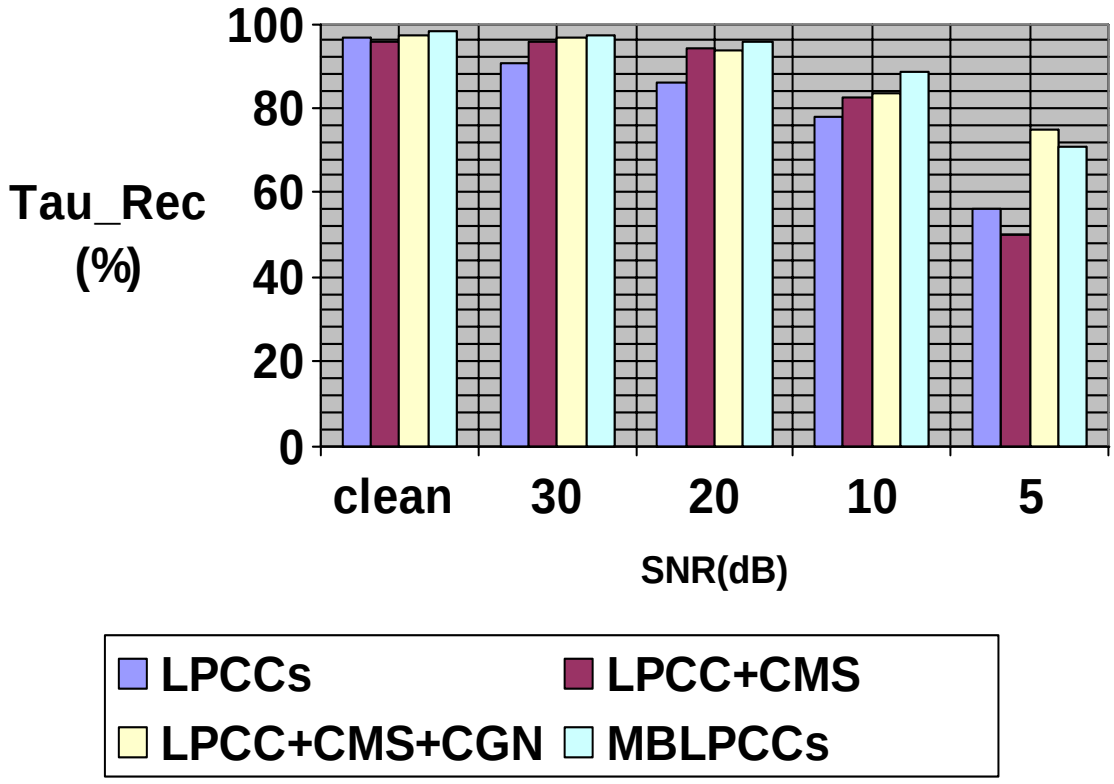

Figure 5.3 Améliorations sur le vecteur LPCC

La figure 5.3 montre que le vecteur MBLPCC+CMS est la meilleure version des paramètres LPCCs dans touts les conditions sauf pour le cas ou SNR<=05dB, le vecteur LPCC+CMS+CGN donne aussi un taux de reconnaissance acceptable en présence d'un bruit élevé, mais dans les autres situations est toujours classé après le vecteur MBLPCC.

- **Opération s sur le vecteur MFCC :**

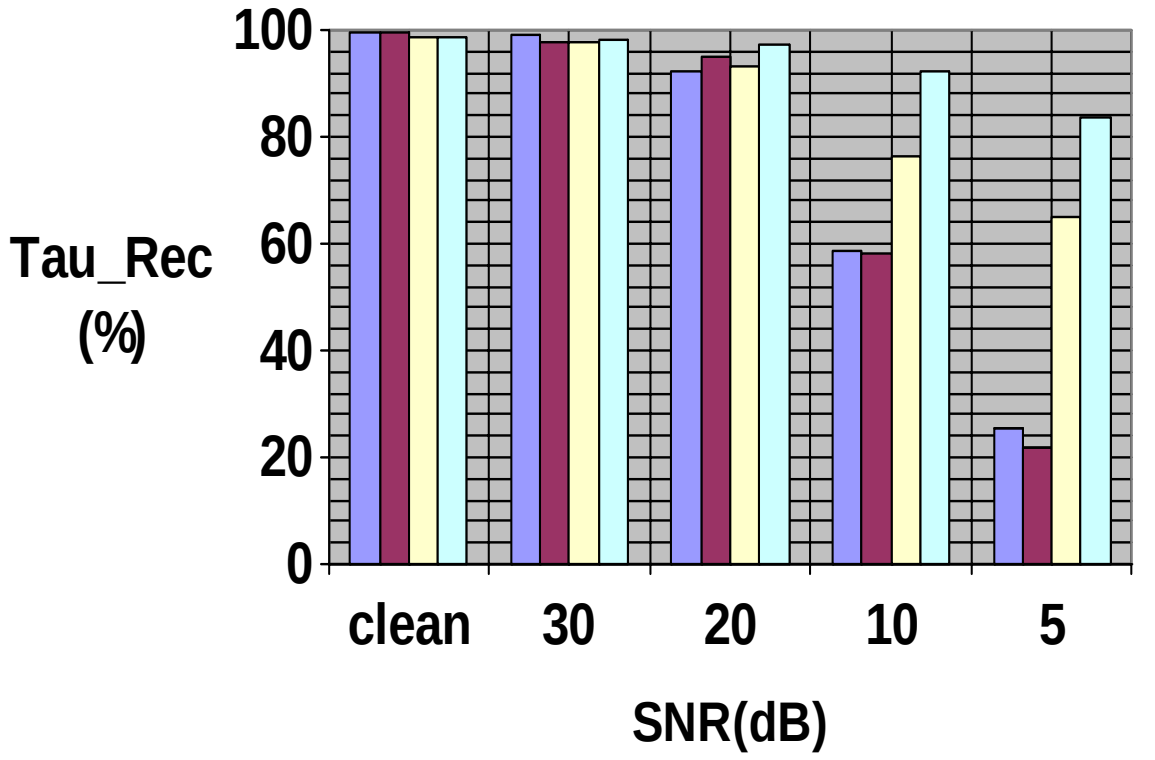

Figure 5.4 Améliorations sur le vecteur MFCC

La figure 5.4 montre que le vecteur LPCMFCC avec normalisation présente des taux de reconnaissance presque égal (<1%) a ce lui donné par MFCC (+CMS ou classique) dans les conditions propre, mais dans les conditions ou SNR<=20dB l'amélioration du taux de reconnaissance est remarquable (plus 60% lorsque SNR=05dB). La résistance des coefficients MFCCs (version LPCMFCC) est devenue meilleure que celle de MFCCs classique est la variation du taux de reconnaissance est très lente ce qui se traduit par une stabilité dans le système RAP.

- **Opération s sur le vecteur PLP :**

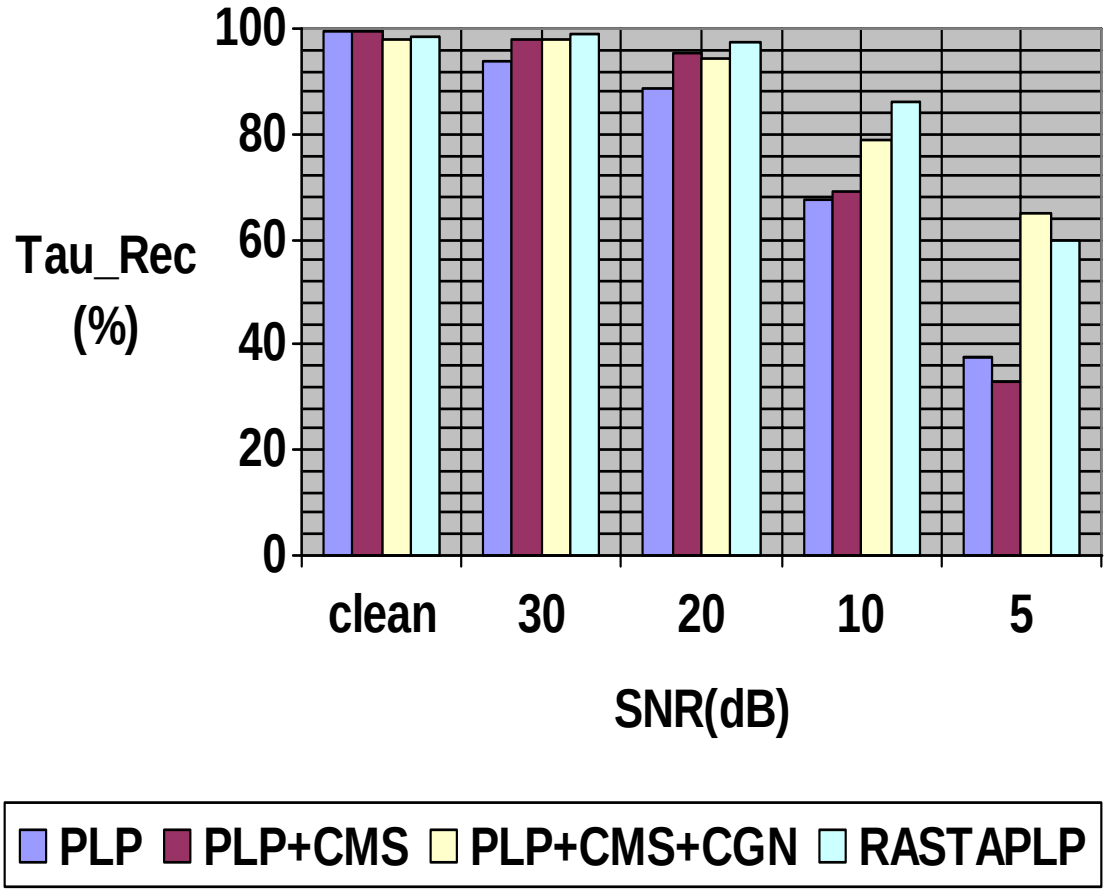

Figure 5.5 Améliorations sur le vecteur PLP

Les résultats représentés sur la figure 5.3 montrent que le vecteur RASTAPLP est la meilleure version du vecteur PLP en plus d'un bon taux de reconnaissance dans le milieu propre, il résiste mieux que les autres versions PLP lorsque 30dB<=SNR<=10dB, ce qui garde le système RAP sable dans les situations propre, faiblement ou moyennement bruité.

- **Comparaison entre déférentes versions des paramètres Robustes :**

Le graphe représenté sur la figure 5.6 juge les performances rapportées par les nouveaux groupes LPCMFCC, MBLPCC, RASTAPLP, on remarque que les trois vecteurs ont presque le même rendement en taux de reconnaissance particulièrement lorsque on utilise un SNR>=20 dB, pour le cas d'un SNR<20 dB la résistance du vecteur LPCMFCC est la plus considérable et ce la se vois très claire sur la partie 10dB<=SNR<= 05dB du graphe.

Si on prend en considération la moyenne du taux de reconnaissance dans des déférentes situations (dans le cas d'un bruit additif ajouté artificiellement) donnés par les tableaux 5.11, 5.9, 5.8 (93.9565pour LPCMFCC, 90.0435 pour MBLPCC, 88.1304 pour RASTAPLP), le vecteur LPCMFCC est le plus fiable ensuite celui du MBLPCC et enfin celui du RASTAPLP. Cette classification se rapporte seulement avec les conditions indiquées dans ce travail et une classification ou jugement réel ne peut être effectué que dans le monde réel (implémentation dans un système RAP de la vie courante).

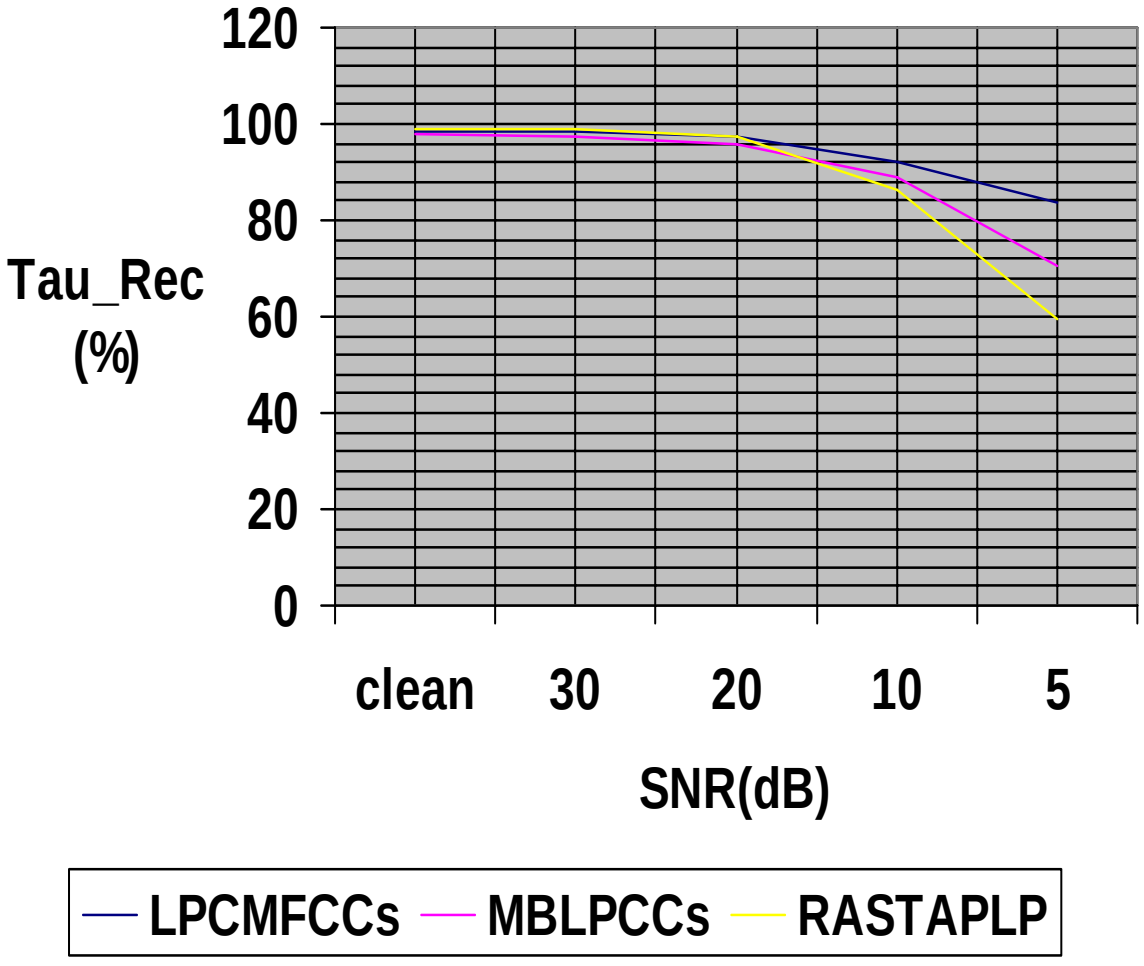

Figure 5.6 Comparaison entre déférentes versions Robustes

Conclusion et perspectives

1. Conclusion

Une étude comparative des représentations robustes aux différences entre les conditions acoustiques d'apprentissage et d'évaluation en reconnaissance de la parole a été réalisée.

Les résultats des études sur la robustesse des techniques de paramétrisation (MFCC, PLP, LPCC), nous permettent d'affirmer qu'une paramétrisation de type MFCC ou PLP donne des meilleurs résultats qu'une paramétrisation de type LPCC. Cependant, quel que soit le niveau de bruit affectant les données d'apprentissage, et quel que soit le type des paramètres acoustiques choisis, le système de reconnaissance de la parole se trouve sérieusement perturbé, surtout pour des rapports signal sur bruit faible.

Pour assurer la robustesse dans les systèmes RAP on a donné une attention particulière aux techniques qui peuvent réduire l'effet du bruit sur la reconnaissance tel que (CMS, CGN, RASTAPLP, MBLPCC, LPCMFCC). Ces approches simples au niveau de l'implémentation ont donnée des bons résultats.

Les expériences réalisées montrent que le bruit additif et la distorsion par les canaux de transmission peuvent être éliminés par ces techniques : CMS et le filtrage RASTA pour réduire l'effet du bruit des canaux de transmission et CGN, MBLPCC, LPCMFCC contre le bruit additif. On remarque aussi qu'une meilleure robustesse peut être aboutie par la combinaison adéquate entre ces techniques comme RASTAPLP, MBLPCC+CMS, LPCMFCC+CMS+CGN. L'expérience nous permet d'affirmer qu'une combinaison entre LPC, MFCC, CMS et CGN qui se traduit par le vecteur paramètre LPCMFCC+CMS+CGN donne au système de reconnaissance une grande fiabilité qui se traduit par un tau moyen de reconnaissance **> 93%.**

2. Perspectives

A la suite de ce travail nous pouvons penser à quelques améliorations potentielles et quelques compléments, à savoir :

1. Utiliser la fusion des paramètres au niveau du classificateur dans le calcule du modèle statistique.
2. Utiliser les techniques de réduction des dimensionnalités comme LDA (Linear Discriminant Analysis) et PCA (Principal Component Analisis) pour réduire la taille énorme des données manipulées dans les phases d'apprentissage et de test.
3. Evaluer ces représentations sur des autres types de bruit additif (bruit de voiture, d'offices etc.). Ajouter en plus les bruits convolutifs. Enfin tester dans les conditions réelles.
4. Utiliser des nouveaux techniques comme les réseaux de neurone, les algorithmes génétiques, etc., pour les calculs des paramètres de la trace acoustiques et les comparées avec les résultats obtenus dans ce travail.
5. Tester ces représentations sur une autre tache telle que la reconnaissance automatique du locuteur (RAL). A priori elles ne doivent pas fonctionner de la même façon, puisque l'information discriminante à garder dans la représentation cepstrale en RAL n'est pas la même qu'en RAP, ce qui va demander des améliorations à faire avant l'implémentation dans le système RAL.

Bibliographie

[01] Dr. J.Picone, FUNDAMETALS OF SPEECH RECOGNITION Institute for Signal and Information Processing, Department of Electrical and Computer Engineering Mississippi State University.

[02] K. Tomi , Spectral Features for Automatic Text-Independent Speaker Recognition Thèse University *of Joensuu Department of Computer Science Finland* December 21, 2003

[03] D. Moreno Moreno, Harmonic Decomposition applied to Automatic Speech Recognition Report_RGB, University of Birmingham (UoB) 29th December, 2002

[04] R. Boite, H. Bourlad, T. Dutoit, J. Hancq, and H. Leich , *Traitement de la parole*, presses polytechniques et universitaires romandes . , France, décembre 99.

[05] The Scientist and Engineer's Guide to Digital Signal Processing.

[06] M. Kunt,: *Techniques modernes de traitement numérique des signaux*, Presses polytechniques et universitaires Romandes, 1991

[07] A. Sakina Reconnaissance de la parole par HMM , Mémoire de magister, institut d'électronique , université d'Annaba, 2004

[08] C. Snani , Conception d'un système de reconnaissance de mots isolés a base de l'approche stochastique en temps réel. Application: commande vocale d'une calculatrice Mémoire de magister, institut d'électronique, université d'Annaba, 2004

[09] H . Hermansky, "Speech Beyond 10ms (Temporal Filtering in feature Domain)", Invited Keynote Lecture, in Proceedings of the International Workshop on Human Interface Technology 1994, Aizu, Japan, Sept. 1994.

[10] Lecture 5. Linear Prediction (LPC), E.4.14 – Speech Processing.

[11] S. Mallat, A WAVELET TOUR OF SIGNAL PROCESSING Second Edition &cole Polytechnique, Paris Courant Institute, New York University

[12] Z. Tufekci, J.N. Gowdy, FEATURE EXTRACTION USING DISCRETE WAVELET TRANSFORM FOR SPEECH RECOGNITION, 0-7803-63 124/00/$10.00@2000 EEE 116

[13] J.N. Gowdy, Z. Tufekci MEL-SCALED DISCRETE WAVELET COEFFICIENTS FOR SPEECH RECOGNITION , 0-7803-6293-4/00/$10.00 02000 IEEE.

[14] Murat Deviren Systèmes de reconnaissance de la parole revisites : Réseaux Bayesiens dynamiques et nouveaux paradigmes, THESE Doctorat informatique Université Henri Poincaré 20/10/2004.

[15] F. de Wet, Automatic speech recognition in adverse acoustic conditions. PhD Departement of Language et Speech of the University of Nijmegen in the Netherlands.

[16] M. Djemili, " *reconnaissance de mots isolés arabes par DTW & HMM*". Mémoire de Magister, Institut d'électronique, Université d'Annaba, 2001.

[17] C. Becchetti and L. Prina Ricotti, *Speech Recognition Theory and C++ Implementation*, John Wiley et Sons, England, 1999.

[18] W –Chen Chen*[+], C .T . Hsieh [*] and E. lai[*]. Multiband approach to robust text – independent speaker identification

[19] LE Viet Bac Reconnaissance automatique de digits en anglais en conditions bruitées DEA D'INFORMATIQUE SYSTÈMES ET COMMUNICATIONS ÉCOLE DOCTORALE MATHEMATIQUES ET INFORMATIQUE.

[20] H. Hermansky, B, Hanson and H. Wakita. Perceptually based linear predictive analysis of speech. Ch 211-81851000-0509 1985 IEEE

[21] C .T. Hsieh, E . lai and Y. C . wang. Robust Speaker identification system based on wavelet transform and Gaussian Mixture Model , journal of information science and engineering A-267 -282(2003)

[22] S. YOSHIZAWA, N. HAYASAKA, N. WADA and Y. MIYANAGA, CEPSTRAL GAIN NORMALIZATION FOR NOISE ROBUST SPEECH RECOGNITION
ICASP 2004

[23] H. Hermansky* , N. Morgan**, A. baya* , Phil Kohn**, The challenge of Inverse –E: THE RASTA –PLP Method, 1058-6393/911991 IEEE

[24] H. Hermansky , Member, IEEE, and N. Morgan, signor Member, IEEE. RASTA processing of speech, IEEE TRANSECTION ON SPEECH AND AUDIOU PROCESSING , VOL .2.NO 4, October 1994

[25] J.de veth,, B. cranenn , F.de wet , et L.boves (2002), A comparison of LPC and FFT –based acoustic feature for noise Robust automatic speech recognition, European project on Speech driven Multi –modal Automatic Directory Assistance (SMADA).

[26] Q .Zhu, A ..Alwan *, Non –linear feature extraction for robust speech recognition in stationary and non –stationary noise, Computer Speech and Language.

[27] P. vary and U –Hente. A short –Time spectrum analyzer with polyphase –Network and DFT. Signal processing 2.(1980) .55 -65

[28] John D .Markel. Digital inverse filtering a new tool for formant trajectory estimation, IEEE Transaction on audio and electro acoustic Vol 20-2 June 1972

[29] J .Scroeter. Techniques for Estimating Vocal –Tract Shapes from the speech signal.
IEEE Transactions on speech and Audio processing Vol .2, No 1 part II, January 1994

[30] S. Stephane, An algorithm for automatic formant extraction using linear prediction spectra , IEEE transaction on acoustics, speech, and signal processing, volassp -22. April 1974

[31] T. Funada, A method for extraction of spectral peaks and its application to Fundamental frequency estimation of speech signal, signal processing (1987) 15 -28

[32] R. Sarikaya and J. H. L. Hansen, High Resolution Speech Feature Parameterization for Monophone-Based Stressed Speech Recognition IEEE SIGNAL PROCESSING LETTERS, VOL. 7, NO. 7, JULY 2000

[33] William J. J. Roberts and S. Furui, Fellow, IEEE Maximum Likelihood Estimation of K-Distribution Parameters via the Expectation–Maximization Algorithm
IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 48, NO. 12, DECEMBER 2000

[34] R. Hariharan, Member, IEEE, I. Kiss, Member, IEEE, and O. Viikki, Member, IEEE NOVEMBER 2001Noise Robust Speech Parameterization Using Multiresolution Feature Extraction IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 9, NO. 8,

[35] S. FURUI, MEMBER, IEEE Cepstral Analysis Technique for Automatic Speaker Verification, IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, VOL. ASSP-29, NO. 2, APRIL 1981

[36] Z. Tufekci and S. Gurbuz** Noise Robust Speaker Verification Using Mel-Frequency Discrete Wavelet Coefficients and Parallel Model Compensation, ICASSP 2005

[37] F. GRANDIDIER UN NOUVEL ALGORITHME DE SÉLECTION DE CARACTÉRISTIQUES – APPLICATION À LA LECTURE AUTOMATIQUE DE L'Écriture MANUSCRITE, PH.D. ÉCOLE DE TECHNOLOGIE SUPÉRIEURE MONTRÉAL, LE 24 JANVIER 2003

[38] T. Islam Interpolation of Linear Prediction Coefficients for Speech Coding, Department of Electrical Engineering McGill University Montréal, Canada April 2000

[39] C. Demars1 REPRESENTATIONS BIDIMENSIONNELLES D'UN SIGNAL DE PAROLE ELEMENTS DE MONOGRAPHIE Version 2000, revue et augmentée Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur. B.P.133 - 91403 Orsay cedex France.11 décembre 2000

[40] H. Hermansky Human Speech Perception: Some Lessons from Automatic Speech Recognition OGI School of Engineering, Oregon Health and Sciences University, Portland, Oregon and International Computer Science Institute, Berkeley, California, hynek@ece.ogi.edu

[41] J.de veth,, F.de wet , B. cranenn , et L.boves (2001), Acoustics features and a distance measure that reduce the impact of training –test mismatch in ASR, Speech communication 34. pp .57 -44

[42] F. de wet, K. Weber, L. boves, B. cranenn , et (2003). Evaluation of formant –like features for automatic speech recognition. Journal of the acoustical society of America.

[43] C. Lévy, Reconnaissance de chiffres isolés embarquée dans un téléphone portablle, Laboratoire Informatique d'Avignon, France.

[44] C. Ping Chen, J. Bilmes . Speech feature smoothing for robust ASR, Department of Electrical Engineering, and University of Washington.

[45] P. Premarkanthan , W. B . Mikhael. Speaker verification /recognition and the importance of selective feature extraction: review. Department of Electrical Engineering, University of central Florida, Orlando.