وزارة التعليم العالي و البحث العلمي

BADJI MOKHTAR-ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR-ANNABA

جامعة باجي مختار - عنابة

Faculté des Sciences de l'Ingénieur

Département d'Electronique

Année 2009

THESE

Présenté en vue de l'obtention du diplôme de Doctorat

Reconnaissance Automatique de la Parole par les HMM en Milieu Bruité : Contribution par paramétrisation acoustique robuste

Option

Systèmes Intelligents

Par

AMARA KORBA Mohamed Cherif

DIRECTEUR DE THÈSE : MESSADEG Djemil Maître de Conférences U. ANNABA

DEVANT LE JURY

PRESIDENT :	**DOGHMANE Noureddine** Professeur	U. ANNABA
EXAMINATEURS		
Mr. **TEBBIKH Hicham**	Professeur	U. GUELMA
Mme. **ROUAINIA Mounira**	Maître de Conférences	U. SKIKDA
Mr. **BOUGHAZI Mohamed**	Maître de Conférences	U. ANNABA
Mr. **SBAA Salim**	Maître de Conférences	U. BISKRA

REMERCIEMENTS

Je voudrais remercier tout particulièrement mon encadreur monsieur MESSADEG DJEMIL, maître de conférences au département d'électronique de l'université de Badji-Mokhtar Annaba, pour sa confiance en moi et en notre projet. Son support moral et scientifique a été indispensable tout au long de cette thèse et je dois avouer que j'ai eu la chance de travailler avec un homme extrêmement compétent mais surtout un homme formidable.

Je remercie mon co-encadreur Monsieur DJEMILI RAFIK, Maître de conférences au département d'électronique de l'université de Skikda, pour avoir été présent aux instants importants de prise de décision ainsi que pour les discussions riches et vives que nous avons eues durant ces dernières années. Je le remercie aussi pour ses remarques et critiques qui ont contribué à l'élaboration de ce travail.

Je remercie monsieur DOGHMANE NOUREDINNE, professeur au département d'électronique de l'université de Badji Mokhtar Annaba, pour m'avoir fait l'honneur de présider mon jury de thèse et pour l'intérêt qu'il a porté à mon travail.

Mes remerciements s'adressent à mes rapporteurs, monsieur TEBBIKH HICHAM professeur au département d'électronique de l'université de Guelma, madame ROUAINIA MOUNIRA maître de conférences à l'université de Skikda, monsieur BOUGHAZI MOHAMED Maître de conférences au département d'électronique de l'université de Annaba et Monsieur SBAA SALIM maître de conférences au département d'électronique de l'université de Biskra qui ont bien voulu accepter d'évaluer le présent travail et ce malgré toutes les responsabilités qu'ils assument. Je les remercie pour le temps qu'ils consacreront à la lecture de cette thèse et je souhaite qu'ils y trouvent entière satisfaction.

Mes vifs remerciements s'adressent à monsieur BEDDA MOULDI, mon encadreur de thèse de Magister.

Mes plus sincères remerciements à l'équipe parole, et plus particulièrement à monsieur BOUROUBA HOUCINE pour les riches discussions.

Pour finir je tiens à remercier toute ma famille et plus particulièrement mes parents pour leur soutien durant ces nombreuses années d'études.

مـلـخـص

إن الأعمال المنجزة في مذكرة البحث تدخل في الإطار العام لنظم التعرف الأوتوماتيكي على اللغة عن طريق الحاسوب.
أغلب نظم التعرف على اللغة تعمل بشكل جيد في المحيط ذو الخصائص الصوتية القريبة من المحيط الذي ثم فيه التمرن، إلا أن قدرات نظام التعرف تتدهور بشكل ملموس إذا كانت خصائص المحيط متغيرة جداً. إن هذا التأثر بالتشويشات من أكبر المعيقات التي تمنع استعمال نظم التعرف في التطبيقات اليومية.

إن الهدف من هذا البحث, وهو جعل نظام التعرف أقل حساسية بالمؤثرات التي تطرأ على المحيط. وهدا باقتراحنا تقنية جديدة صوتية قادرة على تحسين إشارة الصوت عند مدخل نظام التعرف على الصوت. لتمكننا من معالجة التشويش بطريقة عامة (التشويش الأبيض، الوردي، الصناعي...إلخ).
في بادئ الأمر قمنا بمقرنة مختلف التقنيات النموذجية الصوتية الأكثر استعمالا في هدا الميدان و التي تقوم بطريقة فعالة في حل مشكلة التعرف على اللغة في المحيط النقي الخالي من التشويشات، لكن قدرات النظام تبقى ضعيفة و بعيدة عن القدرات المرغوبة في حضور التشويش، إن أكثر المعاير الصوتية تعتمد على طيف الطاقة.

لقد قمنا بدراسة المعايرالصوتية القوية المقترحة من طرف 'دونغلي زهو و بالي وال' المبنية على جداء طيف الشدة بطيف فرق الصفحة. إن المعاملات المحصل عليها أعطت نسبة تعرف على الصوت جيدة حتى حدود إشارة على الصوت تقدر ب5 دسي بال في مختلف التشويشات مما يبين أنه من الممكن وضع نظام تعرف أكثر مقاومة لتغيرات المحيط دون أن يتمرن عليه.

إن النتائج المحصل عليها تتدهور تدريجيا عندما يصبح مستوى التشويش أكثر حدة. لقد اقترحنا مرحلة ابتدائية تتم فيه معالجة الإشارة الصوتية حيث نقوم بنزع ملائم و فعال لتشويش و لا يؤثر كثيراً على مكونات الطيف النافعة للإشارة الصوتية و هذا باستعمال طريقة تفكيك الإشارة عن طريق التقسيم

الحسي الموجي، حيث أن هده الطريقة تعتمد على الحس البسيكو- صوتي والمتعلق بتحسس الأذن للأصوات. باستعمالنا لتقنيتي التحديد: تقنية التحديد الناعمة و تقنية التحديد الناعمة المحولة تمكنا من مراعاة عدم إلغاء مركبات الإشارة ذات الترددات العالية والأقل طاقة من المركبات ذات الترددات المنخفضة، مثل ما هو الحال بالنسبة للأصوات الغير متناغمة. إن الحد الملائم نتحصل عليه عن طريق الحد بالمخالفة.

إن كافة الدراسات التي قمنا بها كانت عن طريق البنك المعلومات الصوتي المطور على مستوى مخبر الأتوماتيك والإشارة بعنابة, بجامعة عنابة. حيث يحتوي هذا الأخير على 9000 كلمة (أرقام اللغة العربية) منطوقة بطريقة منعزلة من طرف90 شخص (46 رجل و44 امرأة). كما أن مختلف التشويشات التي عملنا عليها كانت مأخوذة من بنك التشويشات الصوتية المطور من طرف معهد TNO بهولندة.

إن كافة التجارب المحدثة في هده المذكرة كانت قائمة حول نظام التعرف المركوفي المعلمي المبني على نماذج مركوف المستمرة. لتخفيف من حدة البرمجة، قمنا باستعمال البنية البرمجية HTK (علبة الوسائل لنمادج مركوف المخفي) الموزعة من جامعة كمبريدج، لقد ثم اختيار هده البنية البرمجية لمرونتها و لقابليتها لتغير بكافة أجزائها عند مختلف مراحل إنجاز النظام التعرف على اللغة.

إن المقارنة المجرات بين المعاير الصوتية المقترحة و المعاير الصوتية MFCC تبين أن , المعاير الصوتية المقترحة تحسن من طريقة أداء نظام التعرف ب ٪ 44.71 من أجل SNR يقدر ب (-5dB) , و بنسبة متوسطة تقدر ب ٪14.8 متحصل عليها عن طريق 7 مستويات من SNR للإشارة الصوتية المشوشة بتشويش الأبيض.

Abstract

The work completed in this report lies within the general scope of the robust automatic speech recognition (ASR).
The majority of ASR function correctly in an environment with the characteristics acoustic and sound close to the environment in which the training was done but the performances will be degraded notably if the environmental conditions are very different. This sensitivity to the noise is one of the major brakes to the use of the automatic speech recognition in applications known as general public.

Our objective, in this report, is to make recognition system insensitive, i.e. robust, with the changes of environmental conditions, by proposing a novel method of acoustic modelling able to improve speech signal at the entry of recognition system. This technique being based on the exploitation of the perceptual indices of the speech signal, allowing to treat noises of more general nature (white noise, pink noise, industrial noise….etc).

In the first time we compared various of acoustic modelling techniques, the most used in this discipline, and which effectively solve the problem of the RAP in the clean environment, but the performances of the system are far from being satisfactory in the presence of noise, more the share of these acoustic parameters are based on the study of the power spectrum.

We studied the robust acoustic parameters suggested by Donglai Zhu and K.K. Paliwal which is founded on the product of the spectrum of amplitude by the spectrum of phase. These coefficients having allowed to obtain good rates of recognition rate until to signal to noise ratio 5dB with varied noise conditions, which prove that it is possible to implement a system resistant to different sound environments and who were not contributed at the training stage.

The results obtained are degraded however as the sound level increases, we proposed pre-processing stage to enhance speech signal by adaptive denoising who affects little a useful spectral components of the speech signal by the perceptual wavelet packet (PWP) based denoising algorithm with both type of thresholding procedure, soft and modified soft thresholding

procedure. A penalized threshold was selected.

In the experiments reported in this paper, isolated digit recognition experiments were performed using the Arabic digit corpus database from the national laboratory of automatic and signals (LASA) of University of Annaba, which were designed to evaluate the performance of automatic speech algorithms. This database contains 90 speakers (46 male and 44 female). The studies were made on the corpus of preregistered noise Noisex-92 developed at TNO institute in Soesterberg, Netherlands. The corrupted speech is obtained by adding noise to clean speech at different SNR.

All the experiments performed in this report are evaluated by the same Markovian recognition system of reference based on the continuous HMM. To reduce the task of programming to the minimum, we have used software platform HTK (Hidden Markov Model Toolkit) distributed by the university of Cambridge, we have chose this platform for its user-friendliness, its flexibility and its great choice throughout various stage of the recognition system.

Comparison of the proposed approach with the MFCC-based conventional (baseline) feature extraction method shows that the proposed method improves recognition accuracy rate by 44.71%, with an average value of 14.80 % computed on 7 SNR level for white Gaussian noise conditions.

Résumé

Le travail réalisé lors de cette thèse s'inscrit dans le cadre général de la reconnaissance automatique de la parole (RAP) robuste.
La plupart des systèmes RAP fonctionnent correctement dans un environnement aux caractéristiques acoustiques et sonores proches de l'environnement dans lequel s'est fait l'entraînement mais les performances vont se dégrader notablement si les conditions environnementales sont très différentes. Cette sensibilité au bruit est un des freins majeurs à l'emploi de la reconnaissance automatique de la parole dans des applications dites grand public.

Notre objectif, lors de cette thèse, est de rendre le système de reconnaissance insensible, c'est-à-dire robuste, aux changements de conditions environnementales, en proposant une nouvelle technique de modélisation acoustique capable d'améliorer le signal de parole à l'entrée du système de reconnaissance. Cette technique se fondant sur l'exploitation des indices perceptuels de la parole, permettant ainsi de traiter des bruits d'ordre plus général (bruit blanc, rose, industriel ….etc.).

En un premier temps nous avons comparé différentes techniques de modélisation acoustique, les plus utilisées dans cette discipline, et qui résolvent efficacement le problème de la RAP dans le milieu non bruité, mais les performances du système sont loin d'être satisfaisantes en présence de bruit, la plus part de ces paramètres acoustiques sont basés sur l'étude du spectre d'énergie.

Nous avons étudié les paramètres acoustiques robustes proposés par Donglai Zhu et K.K. Paliwal qui sont fondés sur le produit du spectre d'amplitude par le spectre de phase. Ces coefficients nous ayant permis d'obtenir de bons taux de reconnaissance jusqu'à des rapport signal-sur-bruit (SNR : signal to noise ratio) de 5dB avec des conditions de bruits variées qui prouvent qu'il est possible de mettre en oeuvre un système résistant à des environnements sonores différents et qui n'ont pas été rencontrés lors de la phase d'apprentissage.

Les résultats obtenus se dégradent cependant à mesure que le niveau du bruit augmente, nous avons proposé une phase de prêt traitement du signal parole qui permet un débruitage adaptatif

efficace et qui affecte peu les composantes spectrales utiles du signal parole par l'introduction de la *décomposition en paquet d'ondelettes perceptuel* (PWP : Perceptual wavelet packet), cette décomposition psycho acoustique dépend de la perception de l'oreille humaine. Deux techniques de seuillages ont été envisagées : le seuillage doux et le seuillage *doux modifié* à fin de ne pas éliminé les composantes de haute fréquence qui sont moins énergétique que les basses fréquences, tel que le cas pour les consones. Le seuil adaptatif a été obtenu par la méthode de *seuillage pénalisé*.

Toutes les expériences ont été effectuées à l'aide d'une base de données vocale acquise au niveau du laboratoire LASA, à l'université de Annaba, cette base contient 9000 mots (chiffre arabes) prononcés par 90 locuteurs (46 hommes et 44 femmes) de façon isolée. Les études ont été faites sur le corpus de bruit préenregistré Noisex-92 développé par l'institut TNO à Soesterberg aux Pays-Bas. Les séquences bruitées sont obtenues en additionnant des segments de bruit à la parole propre avec différents (SNR).

Toutes les expériences menues dans cette thèse ont été évaluées par le système de reconnaissance Markovien de référence fondé sur les HMM continus. A fin de réduire au minimum la tache de programmation, nous avons utilisé la plate-forme logicielle HTK (Hidden Markov Model Toolkit) distribuée par l'université de Cambridge, nous avons choisie cette plate forme pour sa convivialité, sa souplesse et sa grande liberté de choix laissée tout au long de la construction des différentes parties du système de reconnaissance.

Les comparaisons effectuées entre les paramètres acoustiques robustes proposés et les paramètres de références qui sont les MFCC ont montré, que nos paramètres améliorent le taux de reconnaissance du système de référence de **44,71 %** pour un SNR de -5dB, et avec une valeur moyenne de **14,8 %** calculée sur 7 niveau de SNR pour le signal parole affecté par le bruit blanc gaussien.

Liste des Tableaux

Liste des Figures

Liste des Symboles

ASR : Automatic speech recognition

RAP : Reconnaissance automatique de la parole

SRAP : Système de reconnaissance automatique de la parole

HMM : Hidden Markov Models

EM : Expectation-Maximisation

ARPA : Advanced Research Projects Agency

DAP : Décodage Acoustico- Phonétique

RAL : Reconnaissance Automatique du Locuteur

LPC : Linear Predictive Coefficients

LPCC : Linear Predictive Cepstral Coefficients

MFCC : Mel Frequency Cepstral Coefficients

MLP : Multi Layer Perceptron

TDNN : Time Delay Neural Network

RBF : Radial Basis Function

VQ : Vector Quantization

LVQ : Learning Vector Quantization

GMM : Gaussian Mixture Model

RASTA : RelAtive SpecTrAl

MAP : Maximum A Posteriori

MLE : Maximum Likelihood Estimation

MFCC : Mel frequency Cepstral Coefficient

MFSCC : Mel Frequency Product spectrum Cepstral Coefficient

PNRF_Soft : Proposed Noise Robust feature with Soft Thresholding

PNRF_Mst : Proposed Noise Robust feature with Modified Soft Thresholding

WP : Wavelet Packet

WPC : Wavelet Packet Coefficient

WPT : Wavelet Packet Transform

PWP : Perceptual Wavelet Packet

PWPC : Perceptual Wavelet Packet Coefficient

PWPT : Perceptual Wavelet Packet Transform

QV : Quantification Vectorielle

Table des matières

Chapitre I : Caractéristiques du signal parole

Chapitre II : Paramètrisation acoustique du signal parole

Chapitre III : Systèmes de reconnaissance automatique de la parole

Chapitre IV : Application des ondelettes au signal de la parole

Chapitre V : Nouveau paramètre acoustique pour la reconnaissance robuste

Introduction générale

La parole est le principal moyen de communication dans toute société humaine. Son apparition peut être considérée comme concomitante à l'apparition des outils, l'homme ayant alors besoin de raisonner et de communiquer pour les façonner.
L'importance de la parole fait que toute interaction homme-machine devrait plus ou moins passer par elle. D'un point de vue humain, la parole permet de se dégager de toute obligation de contact physique avec la machine, libérant ainsi l'utilisateur qui peut alors effectuer d'autres tâches. Ces applications peuvent être regroupées en quatre catégories :

- Commande et contrôle.
- Accès à des bases de données.
- Dicté vocale.
- Transcription automatique de la parole (sous-titrage et transduction automatique).

Notre étude s'intéresse à la conception d'un système de reconnaissance de la parole robuste et efficace, pour obtenir un outil s'intégrant naturellement dans la communication entre l'homme et la machine et facilitant son interaction.

Le principal but de notre travail est d'étudier les moyens qui peuvent rendre le système de reconnaissance insensible aux changements des conditions environnementales en précédant le moteur de reconnaissance par des modules de traitement du signal capables d'améliorer le signal de parole à l'entrée du système de reconnaissance.
Ces modules de prêt-traitement sont capables d'améliorer le signal parole par débruitage adaptatif basé sur les connaissances perceptuelles de l'oreille humaine, et exploitant le module et la phase du spectre de fréquence. Une multi-résolution temps/fréquence est nécessaire afin de mieux analyser les différentes composantes spectrales, pour cet effet nous avons introduit les ondelettes orthogonales qui s'avèrent un outil incontournable pour l'analyse des signaux non stationnaires.
Notre nouveau paramètre acoustique a été évalué par les modèles de Markov cachés continus HMMc. Et cela dans des milieux affectés par différents types de bruits (Blanc, rose, industriel,…etc.). Le système de reconnaissance est sensé reconnaître les chiffres arabes prononcé d'une manière isolée.

Le mémoire de ce travail est réparti en cinq chapitres, Dans le premier chapitre nous

présentons les mécanismes de production et de perception de la parole. Ainsi que les principales caractéristiques de ce signal et les difficultés qui peuvent être rencontrées lors de sa modélisation. Nous présentons ainsi les indices acoustiques pour distinguer et identifier les différents phonèmes arabes.

Dans le deuxième chapitre, nous présentons les paramètres acoustiques les plus utilisés en reconnaissance de la parole, les paramètres qui dépendent de l'appareil de production de la parole (LPC, LPCC) et les paramètres qui dépendent de l'appareil de perception humaine (MFCC, PLP, MFPSCC). Une évaluation de ces paramètres est envisagé afin de déterminer quel sont les paramètres acoustiques les plus adapté à la tache de reconnaissance en milieu bruité.

Le troisième chapitre consacré à l'étude des moteurs de reconnaissance qui sont la programmation dynamique et les modèles de Markov cachés discrets et continus, et présente leurs théories et les algorithmes d'apprentissages. Puis nous présenterons la plate forme logiciel HTK (Hidden Markov Model Tool Kits), ainsi q'une brève description sur les logiciels de développement de systèmes à bases des HMM.

Dans le quatrième chapitre, nous commençons par présenter ce sur quoi tous ses travaux sont basés : les ondelettes, qui offrent une décomposition multi-echelles et une analyse efficace pour les signaux non stationnaires. Ainsi q'une brève description des familles d'ondelettes les plus utilisées pour l'analyse du signal parole.

Dans le dernier chapitre, nous présentons les détails de notre système de paramétrisation acoustique robuste. Ensuite, les différentes expériences sur les corpus de développement et de validation sont présentées. Par l'introduction de d'ondelette de daubechies et par l'application de différentes techniques de seuillage.

Nous terminons ce manuscrit, par une conclusion générale de notre travail sur la paramétrisation acoustique robuste. Nous résumons les résultats importants obtenus au cours de diverses expérimentations. Enfin, nous présentons nos perspectives concernant la paramétrisation robuste de la parole et concernant les améliorations à apporter à notre système.

Chapitre 1

Caractéristiques du signal parole

Introduction

Le présent chapitre a pour intention de présenter les notions élémentaires et les termes relatifs à la description de la parole. Nous présentons les appareils auditif et phonatoire de l'être humain. Nous présenterons ensuite les problèmes dus à la complexité du signal parole : variabilité, non stationnarité, redondance et coarticulation. Nous présenterons l'alphabet arabes et une classification phonétique de la langue arabe.

1.1 Mécanismes de production de la parole

Le processus de production de la parole est un mécanisme très complexe qui repose sur une interaction entre les systèmes neurologique et physiologique. La parole commence par une activité neurologique. Après que soient survenues l'idée et la volonté de parler, le cerveau dirige les opérations relatives à la mise en action des organes phonatoires. Le fonctionnement de ces organes est bien, quant à lui, de nature physiologique.

Une grande quantité d'organes et de muscles entrent en jeu dans la production des sons des langues naturelles. Le fonctionnement de l'appareil phonatoire humain repose sur l'interaction entre trois entités : les poumons, le larynx, et le conduit vocal.

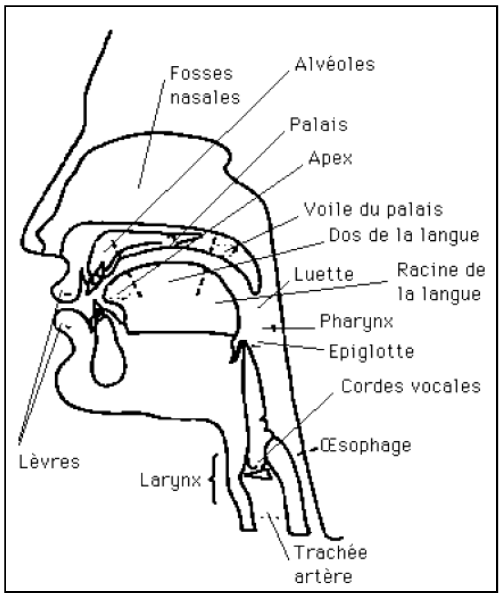


Figure 1.1 Organes de production de la parole

La figure 1.1 représente une vue globale de l'appareil de production de la parole. Le larynx est une

structure cartilagineuse qui a notamment comme fonction de réguler le débit d'air via le mouvement des cordes vocales. Le conduit vocal s'étend des cordes vocales jusqu'aux lèvres dans sa partie buccale et jusqu'aux narines dans sa partie nasale.

La parole apparaît physiquement comme une variation de la pression de l'air causée et émise par le système articulatoire. L'air des poumons est comprimé par l'action du diaphragme. Cet air sous pression arrive ensuite au niveau des cordes vocales. Si les cordes sont écartées, l'air passe librement et permet la production de bruit. Si elles sont fermées, la pression peut les mettre en vibration et l'on obtient un son quasi-périodique dont la fréquence fondamentale correspond généralement à la hauteur de la voix perçue. L'air mis ou non en vibration poursuit son chemin à travers le conduit vocal et se propage ensuite dans l'atmosphère. La forme de ce conduit, déterminée par la position des articulateurs tels que la langue, la mâchoire, les lèvres ou le voile du palais, détermine le timbre des différents sons de la parole. Le conduit vocal est ainsi considéré comme un filtre pour les différentes sources de production de parole telles que les vibrations des cordes vocales ou les turbulences engendrées par le passage de l'air à travers les constrictions du conduit vocal.

Le son résultant peut être classé comme voisé ou non voisé selon que l'air émis a fait vibrer les cordes vocales ou non. Dans le cas des sons voisés, la fréquence de vibration des cordes vocales, dite fréquence fondamentale ou pitch, noté F_0, s'étend généralement de 70 à 400 hertz. L'évolution de la fréquence fondamentale détermine la mélodie de la parole. Son étendue dépend des locuteurs, de leurs habitudes mais aussi de leurs états physique et mental.

1.2 Mécanismes d'audition de la parole

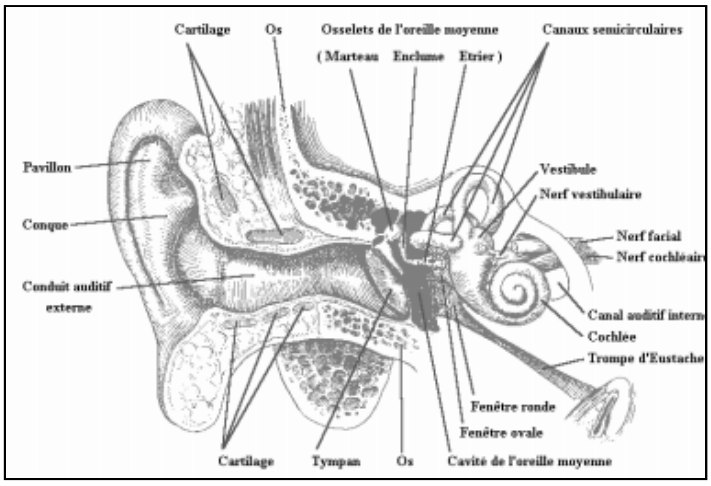
La parole est un vecteur de transmission d'information d'une grande complexité. En tant que récepteur de ce vecteur, l'appareil auditif de l'être humain se caractérise par une grande finesse d'analyse de cette complexité et par une grande robustesse à l'environnement. Pour cette raison, de nombreux systèmes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

Les mécanismes physiologiques qui permettent l'audition d'un message oral sont classiquement séparés en deux parties : l'appareil auditif périphérique et le système auditif central. Dans ce qui suit, nous présentons succinctement l'appareil l'auditif périphérique chez l'être humain pour

introduire la description d'importantes propriétés perceptives du système auditif en relation avec la psycho-acoustique.

L'oreille est divisée en trois parties distinctes, cette division se faisant en fonction de la distance par rapport à l'environnement aérien, porteur des sons. Une première partie, l'oreille externe, correspond à la partie visible de l'organe, pavillon et lobe, à laquelle est rattaché le conduit auditif externe qui permet de propager le son jusqu'au tympan.

Le tympan marque la frontière entre l'oreille externe et l'oreille moyenne. Les organes de l'oreille moyenne permettent de transformer les sons en vibrations grâce au contact qu'ils ont avec le tympan. Ces vibrations, une fois générées, sont transmises à la cochlée qui constitue l'organe majeur de l'oreille interne. La cochlée permet de transformer les vibrations en influx nerveux par le biais de cellules ciliées qui captent les vibrations produites dans le fluide de la membrane basilaire par l'étrier, le dernier os de l'oreille moyenne. Cet influx nerveux est alors transmis au cerveau en charge du traitement. Une description détaillée de l'oreille (figure 1.2) permettra au lecteur de mieux appréhender les différents organes la constituant et de mieux visualiser leur répartition. Il faut noter que la présence de deux oreilles permet d'effectuer, au niveau du cerveau, des traitements plus complexes que le simple décodage d'une scène auditive. Le positionnement des oreilles de chaque côté du crâne permet en effet de profiter des capacités de la binauralité. Cette faculté permet de calculer la provenance d'un son en fonction du retard d'arrivée de ce son dans une oreille par rapport à l'autre. Il est à noter que cette binauralité permet à l'homme de discerner la position horizontale de l'émetteur d'un son mais pas sa position verticale.



L'oreille réagit à des sons de diverses fréquences qui peuvent être regroupées sur des échelles linéaires ou non linéaires.

1.3 Propriétés psycho-acoustiques du système auditif

La psycho-acoustique a pour objet l'étude des relations quantitatives entre les stimuli acoustiques et les réponses du système auditif de l'être humain

Les résultats les plus marquants de cette science sont les suivants :

1.3.1 Échelle d'intensité

Le système auditif ne présente pas une sensibilité à l'intensité sonore identique à toutes les fréquences. En effet, des sons d'intensité sonore égale n'auront pas la même sonie (l'intensité perçue) selon qu'ils soient de haute fréquence 10kHz, de basse fréquence 100Hz, ou de fréquence moyenne 1kHz. Ainsi, si ces trois sons ont une même intensité de 40dB, les sons de fréquence 100Hz et 10kHz seront plus faiblement perçus que le son de fréquence 1kHz.

Les courbes d'isosonie représentent les niveaux d'intensité sonore générant une perception auditive d'égale intensité en fonction de la fréquence du son stimulant (figure. 1.3).

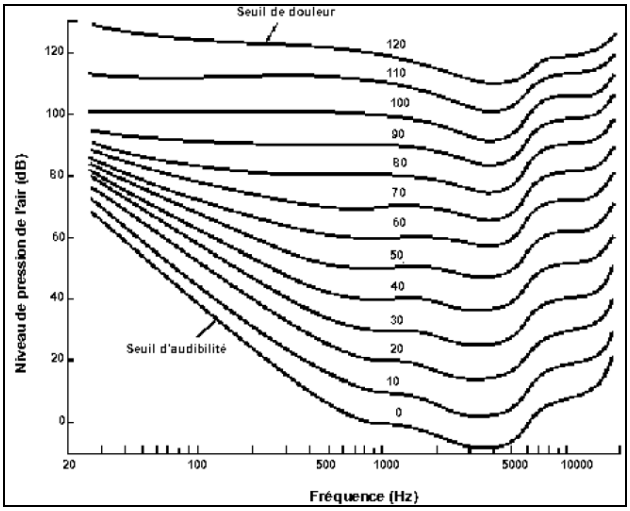


Figure 1.3 Courbes d'isosonie

1.3.2 Échelle de hauteur

La tonie (la hauteur) d'un son est la qualification subjective de sa fréquence. Des études psycho-acoustiques ont en effet montré que la perception humaine du contenu fréquentiel des sons ne suit pas une échelle linéaire mais une échelle fréquentielle de Mel. Cette échelle est approximativement linéaire de 20 Hz jusqu'à 1kHz et logarithmique de 1kHz jusqu'à 20kHz.

Certains chercheurs utilisent échelle Bark. Mais les différences entre les deux échelles sont peu importantes. La figure1.4 montre les différentes échelles naturelles de la membrane basilaire [81].

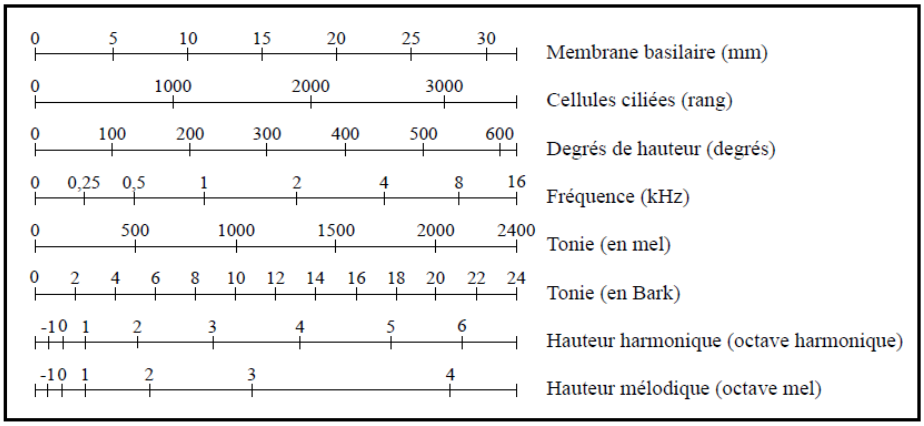


Figure.1.4 les échelles naturelles de la membrane basilaire

1.4 Complexité du signal parole

La complexité du signal parole provient de la combinaison de plusieurs facteurs, Principalement la redondance du signal acoustique la grande variabilité qui peut être due à l'environnement, la variabilité intera-locuteur et inter-locuteur, et les effets de coarticulation en parole continue, qui doivent être pris en compte lors de la conception d'un system de RAP. Nous allons maintenant voir les problèmes liés à la parole, ceux-ci sont relatifs à la différence innée de prononciation vis-à-vis de un ou plusieurs locuteurs.

1.5 Continuité et coarticulation

Tout discours peut être retranscrit par des mots, qui peuvent à leur tour être décrits comme une suite de symboles élémentaire appelés phonèmes par les linguistes. Cela laisse supposer que la parole est un processus séquentiel, au cours du quel des unités indépendantes se succèdent. La parole est en réalité un flux continu, et il n'existe pas de pause entre les mots qui pourrait faciliter leur localisation automatique par les systèmes de reconnaissance.

De plus, les contraintes introduites par les mécanismes de production créent des phénomènes de coarticulation. La production d'un son est fortement influencée par les sons qui le précédent mais aussi qui le suivent en raison de l'anticipation du geste articulatoire. Ces effets s'étendent sur la durée d'une syllabe, voire même au-delà, et sont amplifiés par une élocution rapide.

1.6 Redondance du signal parole

Le signal acoustique présente, dans le domaine temporel, une redondance qui rend indispensable un traitement préalable à toute tentative de reconnaissance. Il existe en effet une grande disproportion entre le débit du signal enregistré et la quantité d'information cherchée pour une tache de reconnaissance. Un signal échantillonné à 16 kHz sur 16 bits représente un débit de 256k bit/s, alors qu'une tache de reconnaissance phonétique recherche typiquement une dizaine de phonèmes à la seconde, soit une compression de prés de 10^4 du débit initial.

1.7 Variabilité

1.7.1 Variabilité intra-locuteur

La variabilité intra-locuteur identifie les différences dans le signal produit par une même personne. Cette variation peut résulter de l'état physique ou moral du locuteur. Une maladie des voies respiratoires peut ainsi dégrader la qualité du signal de parole de manière à ce que celui-ci devienne totalement incompréhensible, même pour un être humain. L'humeur ou l'émotion du locuteur peut également influencer son rythme d'élocution, son intonation ou sa phraséologie. Il existe un autre type de variabilité intra-locuteur lié à la phase de production de parole ou de préparation à la production de parole. Cette variation est due aux phénomènes de coarticulation [67]. Il est possible de voir la phase de production de la parole comme un compromis entre une minimisation de l'énergie consommée pour produire des sons et une maximisation des scores d'atteinte des cibles que sont les phonèmes tels qu'ils sont théoriquement définis par la phonétique.

1.7.2 Variabilité inter-locuteur

La variabilité inter-locuteur est un phénomène majeur en reconnaissance de la parole. La cause principale des différences inter-locuteurs est de nature physiologique. La parole est principalement produite grâce aux cordes vocales qui génèrent un son à une fréquence de base, le fondamental. Cette fréquence de base sera différente d'un individu à l'autre et plus généralement d'un genre à l'autre, une voix d'homme étant plus grave qu'une voix de femme, la fréquence du fondamental étant plus faible. Ce son est ensuite transformé par l'intermédiaire du conduit vocal, délimité à ses extrémités par le larynx et les lèvres. Cette transformation, par convolution, permet de générer des sons différents qui sont regroupés selon les classes que nous avons énoncées

précédemment. Or le conduit vocal est de forme et de longueur variables selon les individus et, plus généralement, selon le genre et l'âge. Ainsi, le conduit vocal féminin adulte est, en moyenne, d'une longueur inférieure de 15% à celui d'un conduit vocal masculin adulte. Le conduit vocal d'un enfant est bien sûr inférieur en longueur à celui d'un adulte. Les convolutions possibles seront donc différentes et, le fondamental n'étant pas constant, un même phonème pourra avoir des réalisations acoustiques très différentes.

La variabilité inter-locuteur trouve également son origine dans les différences de prononciation qui existent au sein d'une même langue et qui constituent les accents régionaux.

1.7.3 Variabilité due à l'environnement

La variabilité liée à l'environnement peut, parfois, être considérée comme une variabilité intra-locuteur mais les distorsions provoquées dans le signal de parole sont communes à toute personne soumise à des conditions particulières. La variabilité due à l'environnement peut également provoquer une dégradation du signal de parole sans que le locuteur ait modifié son mode d'élocution.

Les moyens de transport peuvent entraîner des déformations du signal, d'origine psychologique. Le bruit ambiant peut ainsi provoquer une déformation du signal de parole en obligeant le locuteur à accentuer son effort vocal. Enfin, le stress et l'angoisse que certaines personnes finissent par éprouver lors de longs voyages peuvent également être mis au rang des contraintes environnementales susceptibles de modifier le mode d'élocution.

1.8 Description acoustique

Il est possible de classer les différents sons visibles sur un spectrogramme selon leurs classes respectives en très peu de temps et sans aucune écoute de la phrase correspondante. Le travail des phonéticiens est à ce titre très intéressant et parfois fort impressionnant.

La figure1.5 nous montre une transcription du mot (**"سبيل"**). L'axe des abscisses du spectrogramme représente le temps, l'axe des ordonnées représentant la fréquence qui est, ici, comprise entre 0 et 8Khz. Les nuances de grisé du spectrogramme représentent l'énergie du signal pour une fréquence et à un instant donné. L'énergie minimale des spectrogrammes présentés est de 30 décibels (correspondant au gris le plus clair), l'énergie maximale étant, elle, de 100 décibels (correspondant au noir).

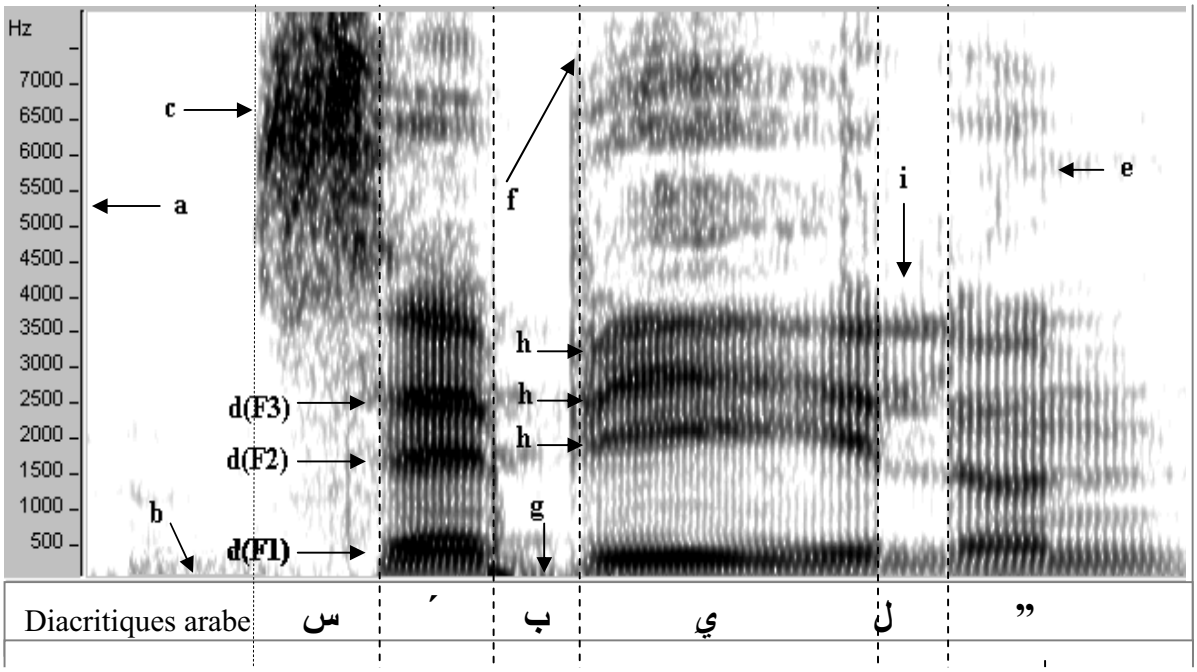


Figure 1.5 Spectrogramme du mot ”سبيل / [sαbīlun] échantillonnée à 16KHz (calculer avec une fenêtre de hamming de 256 points)

(a) l’axe des fréquence
(b) produit par l’appareil en absence du signal.
(c) Bruit de friction (س / [s])
(d) Formants : la mesure se fait au centre de la bande noire.
(e) Modulation de l’énergie à la fréquence fondamentale
(f) Barre d’explosion de l’occlusion.
(g) Barre de voisement
(h) Transitions formantiques.
(i) Formant faible de la sonnante (ل / [l]).

1.9 L’Alphabet Arabe

L'alphabet arabe comprend vingt-huit lettres fondamentales, et s'écrit de droite à gauche. Il n'y a pas de différence entre les lettres manuscrites et les lettres imprimées ; les notions de lettre capitale et lettre minuscule n'existent pas (l'écriture est donc monocamérale). En revanche, la plupart des lettres s'attachent entre elles, même en imprimerie, et leur graphie diffère selon qu'elles sont précédées et/ou suivies d'autres lettres ou qu'elles sont isolées (on parle de variantes contextuelles). L'alphabet arabe est un *abjad*, terme technique décrivant les écritures dans lesquels les voyelles ne sont pas implicitement notées ; le lecteur doit donc connaître la langue pour les restituer. Dans les éditions du Coran ou les ouvrages didactiques, cependant, on utilise

une notation vocalique sous forme dédia critique.

API	FORME	NOM	VALEUR	API	FORME	NOM	VALEUR
[a]	أ	alif	ā	[ḍ]	ض	ḍād	ḍ emph
[b]	ب	bā’	b	[ţ]	ط	ţā’	ţ emph
[t]	ت	tā’	t	[ẓ]	ظ	ẓā’	ẓ emph
[θ]	ث	thā’	th , angl	[‘a]	ع	‘ayn	‘
[∞]	ج	djīm	dj	[ɣ]	غ	ghayn	rh, gh
[h]	ح	hā’	h	[f]	ف	fā’	f
[ξ]	خ	khā’	kh, ch	[q]	ق	qāf	q
[d]	د	dāl	d	[k]	ك	kāf	k
[δ]	ذ	dhāl	dh, angl	[l]	ل	lām	l
[r]	ر	rā’	r roulé	[m]	م	mīm	m
[z]	ز	zāy	z	[n]	ن	nūn	n
[s]	س	sīn	s	[h]	ه	hā’	h
[ʃ]	ش	chīn	ch	[w]	و	wāw	ū, w
[ṣ]	ص	ṣād	ṣ emph	[j]	ي	yā’	ī, y

Tableau 1.1 L’Alphabet arabe

1.10 Les classes phonétiques arabes

Les phonèmes, le cas échéant, sont notés par paire, sourd d'abord puis sonore. Toutes ces classes peuvent se retrouver dans le tableau ci-dessous :

Phonétique	Bilab	Labiodent	Dent	Alvéol	Post-alvéol	Palat	Vélaire	uvulaire	Pharyng	Glot
Occlusives	ب		د ~ ت				ك	ق		ء
Nasales	م		ن							

Fricatives	ف	ذ ~ ث	ز ~ س	ش	ج	غ ~ خ		ح	ه	
Spirantes						ي	و			
Affriquées					ج [dj]					
Liquides					ل					
Vibrantes					ر					

Tableau 1.2 Classification des phonèmes arabes.

Définition des différentes abréviations :

Vélaire : Se dit des voyelles ou des consonnes articulées près du voile du palais.
Uvulaire : consonne dont le lieu d'articulation se situe à l'extrémité postérieure du palais mou, au niveau de la luette.
Pharyngal : Se dit d'une consonne articulée en rapprochant la racine de la langue et la paroi arrière du pharynx.
Glottal : Emis par la glotte.
Alvéole : Consonne articulée avec la pointe de la langue au niveau des alvéoles des dents.
Dentale : Consonne dentale que l'on prononce en appuyant la langue sur les dents.
Bilabiale : Consonne labiale réalisée avec la participation des deux lèvres.
labiodentale : Se dit d'une consonne réalisée avec la lèvre inférieure et les incisives supérieures.
Palatale : Se dit d'une voyelle ou d'une consonne qui a son point d'articulation situé dans la région du palais dur .

1.11 Classification Phonétiques

Les différents sons de la parole sont regroupés en classes phonétiques en fonction de leurs caractéristiques principales. Ces caractéristiques représentent des différences qui sont suffisamment importantes pour qu'il soit possible de classer les différents sons visibles sur un spectrogramme selon leur classe respective en très peu de temps et sans aucune écoute de la phrase correspondante.

Les différentes classes phonétiques existantes, dont nous donnons ci-après la liste, correspondent à des regroupements qui suivent, dans les grands principes, les catégories de l'alphabet. Il existe ici aussi une différence entre voyelles et consonnes par exemple. Mais l'étude des sons de la

parole a obligé à nuancer cette répartition et à créer d'autres classes subdivisant l'ensemble des consonnes.

Les différentes classes phonétiques présentes en Arabe, Anglais et Français sont :

1.11.1 Les Voyelles

Cette classe correspond, à quelques nuances supplémentaires près, aux voyelles de l'écrit. Elles se caractérisent principalement par le voisement qui crée des formants. Ces formants, qui sont des zones fréquentielles de forte énergie, correspondent à une résonance dans le conduit vocal de la fréquence fondamentale produite par les cordes vocales. Ces formants peuvent s'élever jusqu'à des fréquences de 5 kHz mais ce sont principalement les formants en basses fréquences qui caractérisent les voyelles. Cette caractéristique permet d'ailleurs de distinguer grossièrement les voyelles en fonction de leur premier et deuxième formant.

1.11.2 Les Occlusives

Les phonèmes de cette classe se caractérisent oralement par la fermeture du conduit vocal, fermeture précédant un brusque relâchement. Les occlusives sont donc constituées de deux parties successives : une première partie de silence, correspondant à l'occlusion effective, et une deuxième partie d'explosion, au moment du relâchement. Les occlusives peuvent être voisées, à la manière des voyelles, ou sourdes, c'est à dire non voisées. Les occlusives voisées peuvent également être appelées occlusives sonores.

1.11.3 Les Fricatives :

Dans cette classe sont regroupés les sons produits par la friction de l'air dans le conduit vocal lorsque celui-ci est rétréci au niveau des lèvres, des dents ou de la langue. Cette friction produit un bruit de hautes fréquences et peut être voisée ou sourde.

1.11.4 Les Sonnantes

Cette classe est en fait constituée, pour simplification, du regroupement des trois sous-classes que sont les semi-consonnes, les liquides et les nasales.

1.11.5 Les Semi-voyelles

Elles ont la structure acoustique des voyelles mais ne peuvent en jouer le rôle car elles ne sont que des transitions vers d'autres voyelles qui sont les véritables noyaux syllabiques. D'un point de vue syntaxique, une règle stricte de la langue française veut que deux voyelles ne puissent jamais se suivre. Cette règle est très largement respectée dans la construction des mots mais présente, comme toute règle, quelques exceptions. La classe des semi-consonnes a été créée pour pallier ces exceptions de manière gracieuse. Les semi-consonnes sont évidement sonores.

1.11.6 Les Liquides

Les liquides sont très similaires aux voyelles et aux semi-consonnes mais leur durée et leur énergie sont généralement plus faibles. Elles sont sonores.

1.11.7 Les Nasales

Les phonèmes sont formés par passage de l'air dans le conduit vocal depuis les cordes vocales. Ce passage exclut normalement toute connexion du conduit normal, le conduit buccal, avec le conduit nasal. Ce dernier peut cependant être employé, dans un nombre limité de cas puisque sa physionomie ne permet pas de créer des sons autrement qu'en modifiant le volume de la caisse de résonance qu'il constitue par l'intermédiaire de la langue, faisant occlusion dans le conduit buccal. Les nasales sont donc produites de la même manière que les occlusives nasales mais l'air n'est pas, cette fois, comprimé dans le conduit vocal. Le vélum est en effet abaissé pour permettre à l'air d'être expiré. Les nasales sont voisées. Il est à noter que certaines voyelles possèdent également un caractère de nasalité.

1.11.8 Les Diphtongues

Cette classe phonétique est propre à l'anglo-américain et l'arabe. Les phonèmes qui composent cette classe se caractérisent par deux états stables formantiques et par la transition entre ces deux états.

1.11.9 Les Affriquées

Cette classe est, elle aussi, propre à l'anglo-américain et l'arabe mais les affriquées peuvent également être observées dans le français québécois. Les affriquées sont composées d'un occlusive immédiatement suivie par une fricative de durée cependant plus faible que celle des

véritables fricatives.

1.11.10 Les Emphatiques

L'arabe connaît une série de consonnes complexes, dites « emphatiques », qui comprennent, simultanément au phonème, un recul de la racine de la langue (créant ainsi une augmentation du volume de la cavité buccale) vers le fond de la bouche (recul noté en API au moyen « ̠ » de souscrit et une pharyngalisation (API : « ˠ » adscrit), c'est à dire une prononciation simultanée du phonème au niveau du pharynx, là où s'articule [ħ]. On note même une certaine vélarisation, ou prononciation simultanée du phonème au niveau du palais mou, le *velum* ou « voile du palais ».

Les consonnes emphatiques sont les suivantes :

[*tā'*]	[*ẓ ā'*]	[*ṣ ād*]	[*ḍ ād*]	[*'ayn*]	[*qāf*]
ط	ظ	ص	ض	ع	ق

Tableau 1.3 les consonnes emphatiques

1.12 L'alphabet arabe n'a pas de voyelles

Toutes les lettres des tableaux précédents sont des consonnes (ou des lettres muettes). Les voyelles ne sont que rarement notées, et si elles le sont, c'est sous la forme de diacritiques.

- Ainsi, *'alif* n'est pas la voyelle *ā* mais une lettre de prolongement pour la voyelle /a/ (voir à la section « voyelles simples ») ou un support pour divers diacritiques, dont un transcrit une consonne, la *hamza* (voir plus bas). Il est donc improprement transcrit par *ā* ;
- de même, la lettre ى *'alif maqṣūra*, qui ne s'utilise qu'en fin de mot, est une autre lettre de prolongement pour la voyelle /a/. Son nom indique le son obtenu, «*'alif* de prolongement », et non sa forme, puisque la lettre ressemble à un ي *yā'* ;
- enfin, le ة *tā' marbūṭa* est aussi une consonne, à savoir un /t/ ; toutefois, elle ne se trouve qu'en fin de mot et toujours précédée de /a/. Le son /t/, cependant, n'est prononcé que si

les voyelles casuelles finales qui suivent la lettre le sont aussi ; or, ces voyelles sont souvent omises dans la prononciation courante.

1.12.1 Voyelles simples

Les voyelles (qui peuvent être brèves ou longues) ne sont généralement pas écrites, sauf parfois dans les textes sacrés et didactiques, auquel cas l'on dit de ces textes qu'ils sont « vocalisés ». Les brèves sont des diacritiques placés sur ou sous la consonne qui les précède dans la syllabe, tandis que les longues sont notées par le diacritique de la brève équivalente suivie d'une consonne de prolongement :

- ‘ا‘ ‘*alif* ou ‘ى‘ ‘*alif maqs̲ūra* (seulement en fin de mot) pour l'allongement de /a/ ;
- ‘ي‘ *yā’* pour celui de /i/ : *iy* = *ī* ;
- ‘و‘ *wāw* pour celui de /u/ : *uw* = *ū*.

Voyelles simples	Nom	TIMIT
تَ	Fatha	ae / aa
تُ	Damma	ux / uh
تِ	Kasra	ih / ix / ih
تَا	Fatha + alif	ah
تَى	Fatha + maqsora	ay
تُو	Damma+ waw	ux / uw
تِي	Kasra + yaa	iy

Tableau 1.4 les voyelles simples et longues

Note : Les voyelles changent légèrement de timbre selon le contexte dans lequel elles se trouvent.

1.12.2 Sukūn

Une syllabe arabe peut être ouverte (elle est terminée par une voyelle) ou fermée (par une consonne) :

- ouverte : C[onsonne]V[oyelle] ;
- fermée : CVC ; la voyelle en question est le plus souvent brève

Quand la syllabe est fermée, On peut indiquer que la consonne qui la ferme ne porte aucune voyelle en plaçant au dessus un signe nommé *sukūn*, de la forme « ° », pour lever tout ambiguïté .

exemple : قَلب (qalb) ainsi les *sukūn* permettent de savoir où ne pas placer une voyelle .

Chapitre 2

Paramètrisation acoustique du signal parole

2.1 Introduction

Tout Les système de reconnaissance de la parole sont divisés en deux parties, une première partie qui représente la phase d'extraction des paramètres, et une deuxième partie qui est le moteur de reconnaissance. Les performances des systèmes de reconnaissance de la parole dépendent de façon considérable des paramètres acoustiques utilisés.

Dans le présent chapitre nous présentons les paramètres acoustiques les plus utilisés en reconnaissance automatique de la parole, quand on peut les décomposés en deux types, ce qui dépendent de la modélisation du système de production de la parole tel que les paramètres LPC et les paramètres LPCC, et ce qui dépendent de la perception de l'oreille humaine tel que les paramètres acoustiques MFCC, PLP et PLP-RASTA. Á la fin de ce chapitre une évaluation des paramètres acoustiques étudiés est faite.

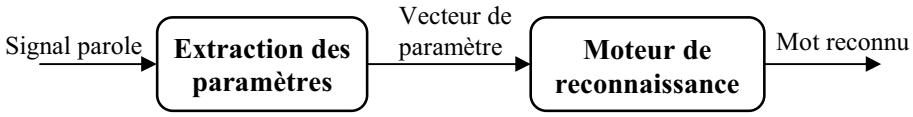


Figure 2.1 Schéma bloc d'un système de reconnaissance de la parole

Ce chapitre est organisé de la façon suivante. Dans la prochaine section, nous présentons les représentations non paramétriques. Dans la troisième section nous présentons les représentations paramétriques les plus utilisés dans la discipline (LPC, LPCC, MFCC, PLP, PLP-Rasta), en fin dans la dernière section nous définirons les paramètres acoustiques MGDCC, MFGDCC et MFPSCC, ces derniers qui dépendent du produit du spectre d'amplitude et de phase.

2.2 Représentations non paramétriques

Le signal de parole peut être analysé dans le domaine temporel ou dans le domaine spectral par des méthodes non paramétriques, sans faire l'hypothèse d'un modèle pour rendre compte du signal observé. Les représentations le plus souvent retenues sont l'énergie du signal et les sorties d'un banc de filtres numériques.

2.2.1 Analyse temporelle

L'énergie du signal est un indice qui peut par exemple contribuer à la détection du voisement d'un segment de parole. L'énergie totale E_0 est calculée directement dans le domaine temporel sur une trame de signal $\{S_n\}$ $0 \le n \le N\text{-}1$ comme :

$$E_0 = \sum_{n=0}^{N-1} S_n^2 \qquad (2.1)$$

L 'énergie ainsi obtenue est sensible au niveau d'enregistrement; on choisit en général de la normaliser, et d'exprimer sa valeur en décibels par rapport à un niveau de référence. D'autres paramètres peuvent être calculés dans le domaine temporel, comme les coefficients d'auto-corrélation, le taux de passage par zéro, ou encore la fréquence fondamentale. L'estimation des coefficients d'auto-corrélation $\{r_k\}$ est calculée par :

$$r_k = \sum_{n=0}^{N-1} s_n \cdot s_{n-k} \qquad 0 \le k \le N-1 \qquad (2.2)$$

Ces coefficients sont utilisés dans le cadre de la modélisation auto-régressive. Cependant, la production de la parole rend souhaitable une analyse du signal dans le domaine spectral pour la reconnaissance.

2.2.2 Analyse spectrale

La transformée de Fourier et l'implantation algorithmique efficace qui y a été associée à la transformée de Fourier rapide, présente de nombreux avantages en tant que méthode l'analyse temps-fréquence. La rapidité de sa mise en œuvre l'a propulsé au rang d'élément incontournable des systèmes de traitement de signal. Mais, après la naissance de la notion de représentation temps-fréquence, qui fait suite à l'utilisation de représentations spectrographiques.

Des spectrogrammes ont été utilisés pour représenter la parole dès les années 40 [79], en utilisant des bancs de filtres analogiques. Actuellement, les spectres sont obtenus numériquement par Transformée de Fourier Discrète, en particulier grâce à l'algorithme de la Transformée de Fourier Rapide (FFT) [75].

Le spectre à court terme $\{S_k\},\ k = 0 \cdots N\text{-}1$ est calculé à partir des N échantillons $\{S_n\},\ n = 0 \cdots N\text{-}1$ comme :

$$S_k = \sum_{n=0}^{N-1} s_n e^{-j2\pi n \frac{k}{N}} \quad , \quad 0 \le k \le N-1 \qquad (2.3)$$

L'intensité en décibels du spectre est directement visualisable sous la forme d'un spectrogramme pour une évaluation qualitative du signal.

Le nombre de paramètres spectraux calculés sur une trame par FFT reste trop élevé pour un traitement automatique ultérieur. L'énergie du spectre est calculée à travers un banc de filtres

numériques couvrant la bande passante, ce qui permet de ne conserver qu'une vingtaine de valeurs d'énergie par exemple sur une bande passante de 8 kHz. Des filtres triangulaires sont préférés pour leur simplicité et leur effet de lissage sur le spectre.

2.2.3 Représentation graphique temps/fréquence (Spectrogramme)

Le spectrogramme est un outil de visualisation utilisant la technique de la transformée de Fourier et donc du calcul de spectres. Il a commencé à être largement utilisé en 1947, à l'apparition du sonagraphe [54], et est devenu l'outil incontournable des études en phonétique pendant de nombreuses années. Il est largement utilisé du fait de sa simplicité de mise en œuvre et du grand nombre d'études qui ont déjà été réalisées.

Le spectrogramme permet de mettre en évidence les différentes composantes fréquentielles du signal à un instant donné, une transformée de Fourier rapide [73] étant régulièrement calculée à des intervalles de temps rapprochés.

L'ensemble du processus de calcul d'un spectrogramme est résumé dans la figure suivante.

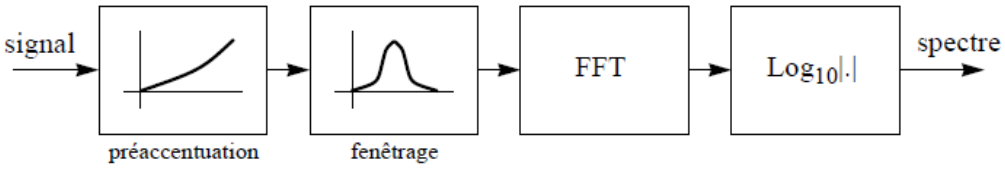


Figure 2.3 Méthode de calcul d'une transformée de Fourier rapide

L'axe des abscisses du signal temporel représente le temps alors que l'axe des ordonnées représente l'amplitude du signal. L'axe des abscisses du spectrogramme représente également le temps, l'axe des ordonnées représentant la fréquence qui est, ici, comprise entre 0 et 5512 Hz. Les nuances de grisé du spectrogramme représentent l'énergie du signal pour une fréquence et à un instant donné. L'énergie minimale des spectrogrammes présentés est de 30 décibels (correspondant au gris le plus clair), l'énergie maximale étant, elle, de 100 décibels (correspondant au noir).

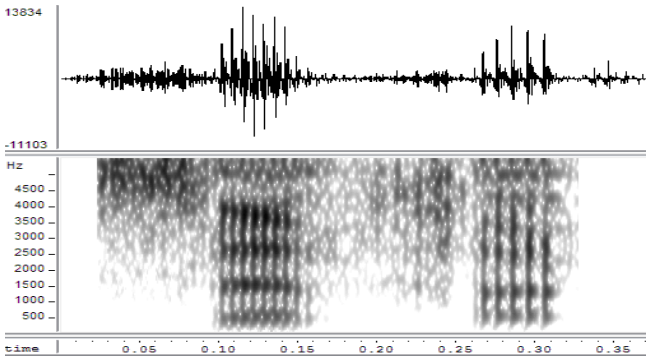


Figure 2.4 Représentation temporelle (en haut), spectrogramme (en bas) du mot 'zéro' en arabe

2.3 Représentations paramétriques

Pour résoudre les problèmes liés à la complexité de la parole, il est possible de calculer des coefficients représentatifs du signal traité. Ces coefficients sont calculés à l'intervalle temporel régulier. En simplifiant les choses, le signal de parole est transformé en une série de vecteurs de coefficients.

Ces coefficients doivent représenter au mieux le signal qu'ils sont censés modéliser, et extraire le maximum d'informations utiles pour la reconnaissance.

Un système de paramétrisation du signal, se décompose en deux blocs (figure 2.5), le premier de mise en forme (figure 2.6) et l'autre de calcul de coefficients.

Le signal analogique est fourni en entrée et une suite discrète de vecteurs, appelée trame acoustique est obtenue en sortie.

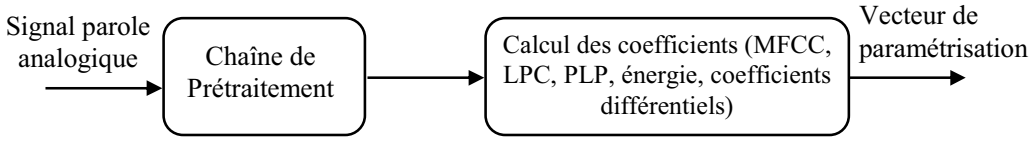


Figure 2.5 Phase de paramétrisation acoustique

2.3.1 Chaîne de prés-traitement

Il est nécessaire de mettre en forme le signal de parole. Pour cela, quelques opérations sont effectuées avant tout traitement. La (figure 2.6) illustre l'ensemble de ces opérations. Le signal est tous d'abord filtré puis échantillonné à une fréquence donnée. Une pré-accentuation est effectuée afin de relever les hautes fréquences. Qui sont moins énergétiques que les basses

fréquences; la pré-accentuation $s_n^{'}$ de l'échantillon s_n à l'instant *n* est calculée pour une valeur α comprise entre 0,9 et 1 comme :

$$S_n^{'} = S_n - \alpha S_{n-1} \qquad (2.4)$$

Puis le signal est segmenté en trames. Chaque trame est constituée d'un nombre *N* fixe d'échantillons de parole. En général, *N* est fixé de telle manière que chaque trames corresponde à environs 25 ms de parole (durée pendant laquelle la parole peut être considérée comme stationnaire). Enfin une multiplication par une fenêtre de pondération W_n est effectuée, afin de réduire les effets de bords. Le choix se porte généralement sur les fenêtres de Hamming ou de Hanning:

$$Ham\min g(n) = 0{,}54 - 0{,}64\cos\left(2\pi\frac{n}{N-1}\right),\quad 0 \le \mathrm{n} \le \mathrm{N}\text{-}1 \qquad (2.5)$$

$$Hanning(n) = 0{,}5 - 0{,}4\cos\left(2\pi\frac{n}{N-1}\right),\quad 0 \le \mathrm{n} \le \mathrm{N}\text{-}1 \qquad (2.6)$$

avec

$$S_n^{''} = W_n S_n^{'} \qquad (2.7)$$

Après cette mise en forme du signal (commune à la plupart des méthodes d'analyse de la parole), une transformée de Fourier discrète DFT en particulier FFT (Transformé de Fourier Rapide) est appliquée pour passer dans le domaine fréquentiel.

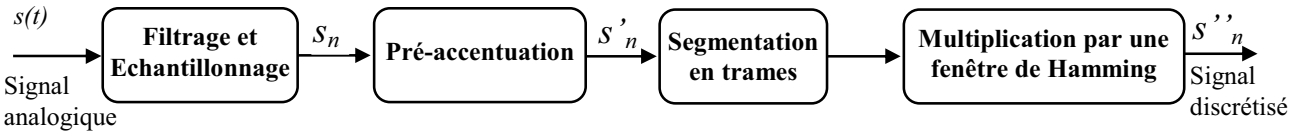


Figure 2.6 Chaîne de prétraitement du signal parole

2.3.2 L'analyse LPC

Le principe du modèle autorégressif du signal de parole est de modéliser le processus phonatoire par un système de synthèse élémentaire comprenant un module d'excitation à gain variable *G*, suivi par un filtre tout-pôles d'ordre p (LPC: Linear Predictive Coding). Les coefficients du filtre sont considérés constants pendant des intervalles de temps réduits de l'ordre de 25 ms (hypothèse de quasi-stationnarité). L'excitation u est soit périodique (train d'impulsions, ou plus généralement signal périodique dont le spectre d'amplitude est un train d'impulsions, ce qui permet de modéliser les déphasages entre les différentes harmoniques), soit

stochastique (bruit blanc), et éventuellement mixte, de façon à pouvoir modéliser les sons voisés ainsi que les sons non-voisés.

Remarquons que pour le cas des sons purement voisés, l'excitation du système représentera l'action opérée par la vibration des cordes vocales, alors que le filtre représentera l'action du conduit vocal. Pour le cas de sons partiellement non voisés par contre, le signal acoustique est le résultat d'un processus plus complexe faisant intervenir la frication, c'est à dire les perturbations créées par le passage de l'air au travers des constrictions du conduit vocal ou des lèvres. L'interprétation du modèle n'est donc plus aussi simple. Ce modèle reste cependant très utilisé en pratique car, quel que soit la nature périodique ou apériodique du signal, la fonction de transfert du filtre sera un bon modèle de l'enveloppe spectrale du signal, caractéristique essentielle pour la distinction des sons linguistiques.

La prédiction linéaire [71] permet la modélisation d'un signal $s(n)$ comme une combinaison linéaire de ses valeurs passées et des valeurs d'un signal d'excitation $u(n)$.

Un échantillon $s(n)$ est calculé comme suit :

$$s(n) = \sum_{i=1}^{p} a_i \, s(n-i) + Gu(n) \qquad (2.8)$$

En effectuant la transformation en Z, on obtient

$$S(z) = \sum_{i=1i}^{p} a_i \, z^{-i} S(z) + GU(z) \qquad (2.9)$$

La fonction de transfert du filtre est bien évidemment exprimée par :

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i \, z^{-i}} \qquad (2.10)$$

et devra idéalement avoir un ordre suffisamment élevé pour modéliser avec précision la structure en formants du spectre du signal. L'ordre ne sera cependant pas trop élevé, et ce pour éviter la modélisation de détails spectraux au contenu linguistique négligeable. On estime en général avoir besoin d'une paire de pôles par kHz de bande passante, plus 3 ou 4 pôles pour l'excitation glottique et la radiation des lèvres. Pour une fréquence d'échantillonnage de 8 kHz, on choisira

donc un ordre de 11 ou 12. Les expériences de reconnaissance vocale montrent que ces valeurs sont raisonnables.

Les paramètres de ce modèle, à savoir le gain, l'excitation et les coefficients a_i peuvent être estimées par des méthodes d'analyse. Une interprétation de ces méthodes d'analyse est de séparer la source et la structure, et donc d'obtenir des paramètres de structure a_i relativement "propres".

A partir du modèle qui vient d'être décrit, une estimation de l'échantillon $s(n)$ peut-être calculée de la sorte:

$$\hat{s}(n) = \sum_{i=1}^{p} a_i \, s(n-i) \qquad (2.11)$$

L'erreur de prédiction $\hat{s}(n) - s(n)$ vaut donc :

$$s(n) - \sum_{i=1}^{p} a_i \, s(n-i) \qquad (2.12)$$

Une estimation des paramètres a_i peut être obtenue par minimisation de la somme des carrés des erreurs de prédiction sur une trame de parole provenant des étapes de traitement précédentes, ce qui conduit à un système linéaire de p équations à p inconnues faisant intervenir la fonction de covariance du signal s. En limitant l'ordre de la somme des erreurs de prédiction par définition d'une fenêtre de signal de durée limitée, on peut montrer que les éléments intervenant dans le système d'équation sont les $p+1$ premiers éléments de la fonction d'autocorrélation du signal. De plus, la matrice du système est une matrice de Toeplitz (les éléments de toutes les diagonales sont égaux) symétrique. Cette particularité permet l'utilisation d'une méthode de résolution particulièrement efficace appelée récursion de Durbin. Une description de cette méthode peut être trouvée dans [46].

2.3.3 LPCC (Linear Prediction Cepstral Coefficients)

Les paramètres LPCC sont calculés à partir d'une modélisation auto-régressive du signal. Si un modèle auto-régressif $A(1, a_1 .... a_p)$ d'ordre p a été estimé sur une trame du signal, les d premiers coefficients cepstraux C_n sont obtenus par :

$$C_n = -a_n + \frac{1}{n}\sum_{i=1}^{n-1}(n-i)a_i C_{n-i} \quad 1 \le n \le d \qquad (2.13)$$

Ces coefficients sont utilisés à AT&T [60].

Ensuite un liftrage est effectué pour augmenter la robustesse des coefficients cepstraux, ce liftrage consiste en une multiplication par la fenêtre de poids (représenté par la formule (2.13)) par des coefficients cepstraux augmentant l'amplitude des coefficients connus pour être moins sensibles au canal de transmission et au locuteur

$$\forall i \in [1, L] \quad w(i) = 1 + \frac{L}{2}\sin\left(\frac{\pi \cdot i}{L}\right) \qquad (2.14)$$

$$c_i = \left(1 + \frac{L}{2}\sin\left(\frac{\pi \cdot i}{L}\right)\right) a_i \qquad (2.15)$$

Où L est le nombre de coefficients. Cette méthode de prédiction linéaire est beaucoup Plus utilisée en reconnaissance de la parole que celle de l'analyse spectrale.

2.3.4 Les coefficients PLP (Perceptual Linear Predictive)

PLP est une technique d'analyse de la parole [57] fondée sur la modélisation du spectre par un modèle tout pôle suivant un principe identique à la technique de prédiction linéaire (LP). Cependant, la différence réside dans le fait que les paramètres d'un filtre auto-régressif tout pôle sont estimés en modélisant au mieux le spectre auditif. Ceci est fondé sur trois effets auditifs : sélectivité spectrale de bande critique, courbe d'intensité égale et loi de puissance.

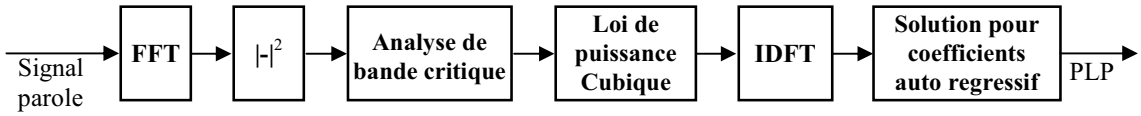


Figure 2.7 Étapes de calcule des coefficient PLP

La figure ci-dessus représente le processus de calcul des coefficients PLP. Pour obtenir un spectre auditif, la courbe de masquage $\Psi(\Omega)$ est tout d'abord utilisée

$$\Psi(\Omega) = \begin{cases} 0 & si \ \ \Omega \prec -1{,}3 \\ 10^{2{,}5(\Omega+0{,}5)} & si \ \ -1{,}3 \le \Omega \le -0{,}5 \\ 1 & si \ \ -0{,}5 \le \Omega \le 0{,}5 \\ 10^{-1{,}0(\Omega-0{,}5)} & si \ \ 0{,}5 \le \Omega \le 2{,}5 \\ 0 & si \ \ \Omega \succ 2{,}5 \end{cases} \qquad (2.16)$$

Où Ω est la fréquence de Bark calculée à partir de la fréquence angulaire ω par la définition :

$$\Omega(\omega) = 6\ln\left(\frac{\omega}{1200\pi} + \left(\left(\frac{\omega}{1200\pi}\right)^2 + 1\right)^{\frac{1}{2}}\right) \qquad (2.17)$$

Le spectre de puissance du signal $P(\omega)$ (pair et périodique) est convolué avec la courbe de masquage:

$$\theta(\Omega_k) = \sum_{\Omega=-1,3}^{\Omega=+2,3} P(\Omega - \Omega_k)\Psi(\Omega) \qquad (2.18)$$

Puis, l'algorithme tente de faire l'approximation de la sensibilité de l'oreille humaine à différentes fréquences par l'intermédiaire d'une fonction de transfert $E(\omega)$:

$$\Xi(\Omega(\omega)) = E(\omega)\Theta(\Omega(\omega)) \qquad (2.19)$$

La non-linéarité entre l'intensité d'un son et son niveau de perception par l'oreille est réalisée en l'approchant par une loi de puissance :

$$\Phi(\omega) = \Xi(\Omega)^{\frac{1}{3}} \qquad (2.20)$$

Enfin le spectre auditif est modélisé par un modèle tout-pôle. Une transformée de Fourier inverse discrète est appliquée sur le spectre auditif $\Phi(\omega)$ pour obtenir les valeurs d'autocorrélation. *M+1* premiers coefficients d'autocorrélation sont utilisés pour calculer les coefficients auto régressifs du modèle tout pôle d'ordre *M* qu'on appelle les coefficients PLPs.

Comme la méthode LPC, les coefficients cepstraux peuvent être obtenus à partir des coefficients PLPs.

2.3.5 Rasta PLP

La méthode PLP [57], dont l'algorithme repose sur des spectres à court terme de la parole, résiste difficilement aux contraintes qui peuvent lui être imposées par la réponse fréquentielle d'un canal de communication. Pour atténuer les effets de distorsions spectrales linéaires, [53] propose de modifier l'algorithme PLP en remplaçant le spectre à court terme par un spectre estimé où chaque canal fréquentiel est modifié par passage à travers un filtre. Cette modification est à la base de la méthode RASTA PLP, RASTA étant l'acronyme de *RelAtive SpecTrAl* [53]. La mise en place de ce filtrage permet, lorsqu'il est effectué dans le domaine spectral logarithmique, de supprimer les composantes spectrales constantes, supprimant ainsi les effets de convolution du canal de communication.

Nous décrivant dans ce qui suit Algorithme de calcule des coefficients RASTA PLP

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.
3. Filtrage des trajectoires temporelles de chaque composante spectrale.
4. Expansion de l'amplitude à l'aide d'une transformation non linéaire.
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 pole du spectre selon la méthode PLP classique.

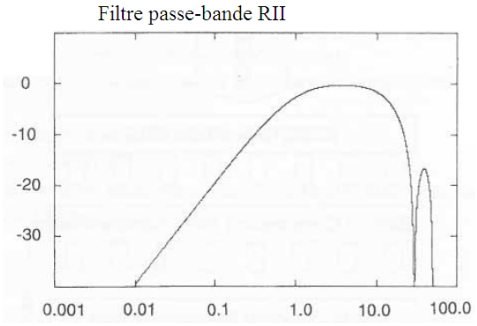


Figure 2.8 Réponse fréquentielle du filtre passe bande RASTA

2.3.6 Analyse cepstrale

L'analyse cepstrale permet, dans le cas d'un signal de parole, la séparation des deux composantes de ce signal que sont : l'excitation de la source et la réponse du conduit vocal [68]. Comme la modélisation linéaire prédictive, cette analyse suppose que l'appareil de production de la parole se comporte comme un modèle source-filtre. Le signal de parole résulte donc du produit de convolution de l'excitation et de la réponse impulsionnelle du filtre

$$s(n) = e(n) * h(n) \qquad (2.21)$$

Un traitement appelé déconvolution, ou traitement homomorphique de la convolution, permet de séparer les signaux $e(n)$ et $h(n)$.

Ce traitement consiste à calculer d'abord la transformée de Fourier du signal $s(n)$, c'est-à-dire

$$S(\omega) = E(\omega) \cdot H(\omega) \qquad (2.22)$$

En prenant le logarithme de cette expression, puis en faisant une transformée de Fourier inverse, on obtient le cepstre :

$$\hat{s}(n) = TF^{-1}(\log|E(\omega)|) + TF^{-1}(\log|H(\omega)|) \qquad (2.23)$$

où TF^{-1} désigne le transformée de Fourier inverse.

On appellera le signal $\hat{s}(n)$ obtenu par cette opération cepstre complexe associé au signal $s(n)$. On a donc [68] :

$$\hat{s}(n) = \hat{e}(n) + \hat{h}(n) \qquad (2.24)$$

Si, comme le signal de parole, le signal $s(n)$ est un signal réel, alors $\hat{s}(n)$ sera aussi réel et on pourra le calculer à partir du module de la transformée de Fourier.

Le cepstre réel est ainsi défini comme la transformée de Fourier inverse du logarithme du module de la transformée de Fourier d'une fenêtre à court terme du signal de parole :

$$c(n) = TF^{-1}\left(\log(S(k))\right) \qquad (2.25)$$

$$= \frac{1}{k}\sum_{k=0}^{K-1} \log(S(k) e^{j2\pi n\frac{k}{K}} \qquad (2.26)$$

avec $S(k) = |s(k)|^2$ la densité spectrale de puissance du signal

Pour estimer la contribution du conduit vocal dans le signal de parole, on ne conserve que les premiers échantillons du cepstre $c(n)$ qui correspondent en particulier aux informations sur les formants. Les échantillons du cepstre d'ordre plus élevé correspondent en général aux caractéristiques de la fréquence fondamentale des cordes vocales.

Une des propriétés du cepstre est qu'il effectue un filtrage passe-bas du spectre du signal et tend donc à lisser les irrégularités du spectre. De ce fait, les amplitudes des harmoniques ne sont pas conservées. Pour palier ce problème et obtenir une enveloppe spectrale passant par les amplitudes des harmoniques du signal, Galas et Rodet ont proposé une méthode dite du cepstre discret [55].

2.3.7 L'analyse MFCC

L'analyse acoustique MFCC est l'une des techniques les plus utilisées pour la paramétrisation du signal en segmentation markovienne de parole.

Cette technique est basée sur deux idées clés [65] [47] [39]. La première consiste à exploiter les propriétés du système auditif humain par la transformation de l'échelle linéaire des fréquences en échelle de Mel. Et la deuxième consiste à effectuer une transformation cepstrale qui permet la décorrélation des composantes spectrales du signal de parole.

Pour transformer une fréquence linéaire en une fréquence Mel, on utilise la formule de transformation suivante:

$$B(f) = 2595\log_{10}\left(1 + \frac{f}{700}\right) \qquad (2.27)$$

où *f* est la fréquence en Hz, *B(f)* est la fréquence mel-échelle de *f*.

Les bandes-passantes sont de même taille dans l'échelle Mel.

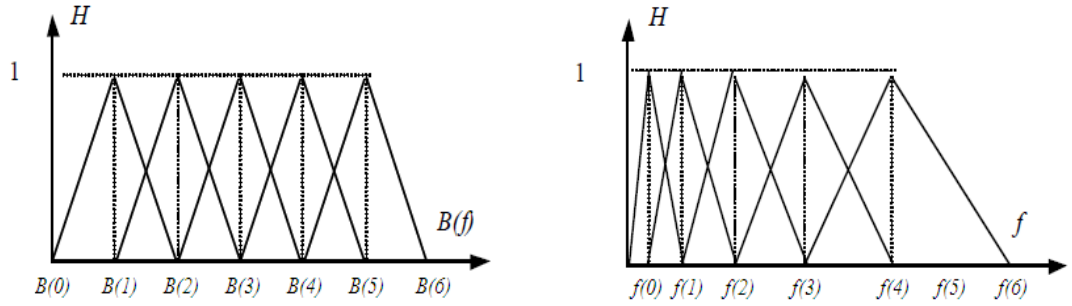


Figure 2.9 Les filtres triangulaires passe-bande en Mel-Fréq *(B(f))* et en fréquence *(f)*

On peut calculer les points frontières $B(j)$ des filtres en mel-fréquence ainsi :

$$B(j) = B(f_l) + j\frac{B(f_h) - B(f_l)}{J+1} \quad 0 \le j \le N+1 \qquad (2.28)$$

N est le nombre de filtres ($N = 22$).

On doit calculer les points $f(j)$ correspondants dans le domaine de fréquence réelle :

$$f(j) = \frac{N}{F_s} B^{-1} B(j) \qquad (2.29)$$

Puis on détermine tous les coefficients de chaque filtre :

$$H_j(k) = \begin{cases} 0 & k \le f(j-1) \\ \dfrac{k - f(j-1)}{f(j) - f(j-1)} & f(j-1) \le k \le f(j) \\ \dfrac{f(j+1) - k}{f(j+1) - f(j)} & f(j) \le k \le f(j+1) \\ 0 & k \ge f(j+1) \end{cases} \qquad (2.30)$$

L'analyse MFCC comporte plusieurs étapes représentées dans la (figure 2.12). Le pré-traitement consiste à effectuer sur le signal de parole, échantillonné à 11025 Hz et quantifié sur 16 bits, les opérations suivantes :

- Toutes les 10ms (110 échantillons), une trame acoustique de 25ms (275 échantillons) est extraite du signal.
- La composante continue des échantillons constituant cette trame est enlevée.
- Afin de compenser l'atténuation naturelle du spectre du signal de parole, la séquence des échantillons constituant la trame subit une pré-accentuation avec le filtre du premier ordre

$$H(Z) = 1 - 0{,}97Z^{-1} \qquad (2.40)$$

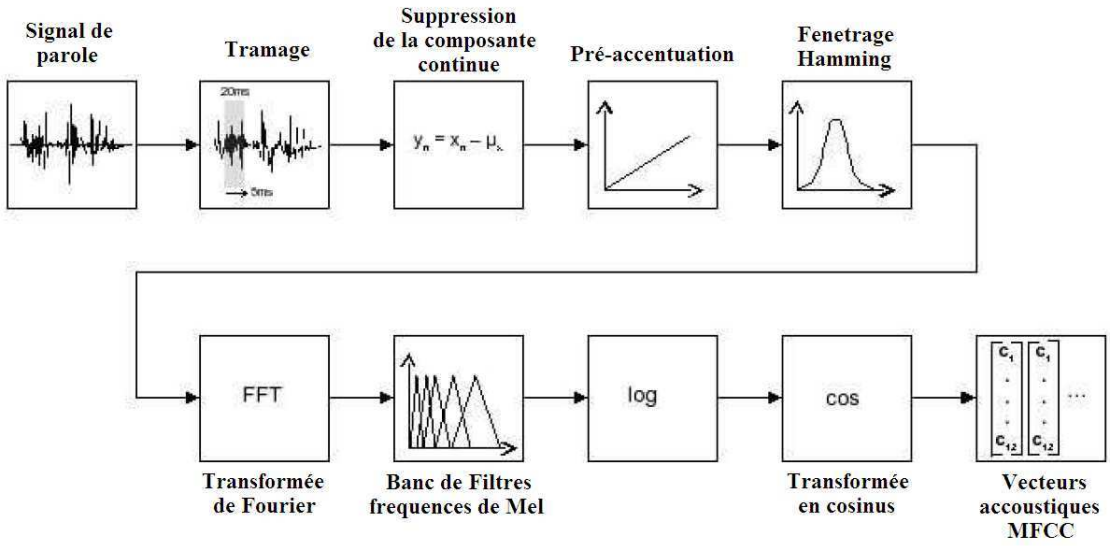


Figure 2.10 Schéma en blocs de l'analyse acoustique permettant le calcul des vecteurs MFCC.

Pour atténuer les distorsions spectrales introduites par l'extraction de la trame du signal de parole, on pondère les échantillons de cette trame par la fenêtre de Hamming.

L'analyse MFCC proprement dite consiste à effectuer sur chacune des trames résultantes du pré-traitement les opérations suivantes :

- La transformation de Fourier permet de calculer le spectre d'amplitude de la trame.
- Pour chacun des 22 filtres triangulaires répartis sur l'échelle des fréquences de Mel, l'énergie du spectre d'amplitude en sortie de ce filtre est calculée. Cette opération donne un vecteur de 22 valeurs énergétiques E_j.

$$E_j = \sum_{k=0}^{N-1} |S(k)|^2 H_j(k) \qquad (2.31)$$

- Les logarithmes de ces 22 valeurs sont alors transformés en 12 coefficients MFCC par l'inverse de la transformée en cosinus discrète :

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} \log_{10}(E_j) \cos\left(\frac{\pi i}{N}(j+0{,}5)\right) \qquad (2.32)$$

où c_i est le i^{eme} coefficient mel-cepstral, E_j est l'énergie du spectre calculée sur la bande passante du j^{eme} filtre, et N est le nombre de filtres ($N = 22$).

- Afin d'augmenter la robustesse de ces coefficients pour le calcul des distances cepstrales, une pondération en sinus (liftering) est appliquée sur les coefficients MFCC c_i [46] :

$$\hat{c}_i = \left(1 + \frac{L}{2}\sin\frac{i\pi}{L}\right)c_i \quad 1 \le i \le 12 \qquad (2.33)$$

où $\hat{c}_i$ est le $i^{\text{ème}}$ coefficient mel-cepstral liftré et L est le coefficient du liftering (L = 22). Ces pondérations corrigent la décroissance rapide des coefficients MFCC d'indice élevé et permet l'utilisation d'une distance euclidienne.

2.3.8 Produit spectral et la fonction de temps de groupe

Les coefficients MFCC sont les paramètres acoustiques les plus largement utilisés en reconnaissance de la parole. Ils se dérivent du spectre de puissance du signal parole.

Récemment, des études ont été menés par Murthy et Gadde dans [21] sur les paramètres cepstraux basés sur la *fonction de temps de groupe* (GDF : Group Delay Function). Dans ce qui suit nous présentons des paramètres acoustiques robustes proposés par Donglai et Paliwal dans [18] qui dépendent du produit du spectre de puissance et de la fonction de temps de groupe, et qui résultent des coefficients cepstrals de Mel du produit spectral. Le spectre résultant est une combinaison entre le spectre d'amplitude et le spectre de phase.

2.3.8.1 Définition de la fonction de temps de groupe

Soit la trame du signal parole $x(n), n = 0 \cdots N-1$ sa transformée de Fourier est donnée par :

$$X(\omega) = |X(\omega)|e^{j\theta(\omega)} \qquad (2.34)$$

La fonction de temps de groupe (GDF : group Delay function) est définie par [72] :

$$\tau_p(\omega) = -\frac{d\theta(\omega)}{d\omega} \qquad (2.35)$$

L'équation (2. 35) peut être simplifiée comme suit [72]:

$$\tau_p(\omega) = -\operatorname{Im}\frac{d(\log(X(\omega))}{d\omega} \qquad (2.36)$$

$$= \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2} \qquad (2.37)$$

Où $Y(\omega)$ est la transformée de Fourier de $nx(n)$, I et R pour designer respectivement la partie imaginaire et la partie réelle. La figure 2.13 (a), (b) et (c) montre une trame (de durée T=30ms) de la voyelle (i), son spectre de puissance et sa GDF respectivement. Avant le calcul de la transformée de Fourier, le signal parole a subit un filtrage de pre-accentuation ensuite une multiplication par la fenêtre de Hamming. Dans le spectre de puissance les formants sont

clairement visibles, cependant, il y a seulement des pics sans signification dans la GDF. Ils se produisent en raison du spectre de puissance dans le dénominateur dans l'équation (2.37). Afin de rendre la GDF significative, une modification a été proposée pour la GDF en remplaçant le spectre de puissance $|X(\omega)|^2$ par le spectre de puissance cepstral lissé $(S(\omega))^2$ dans l'équation (2.37) [50]. Il donne la MGDF comme suit

$$= \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{(S(\omega))^2} \qquad (2.38)$$

La figure 2.13 (d) montre que la MGDF du signal a des valeurs négatives, qui doivent être limité par un seuil non négatif avant le calcul des valeurs en dB. Nous adoptons le seuil dynamique proposé dans [47], c.-à-d., rejet des valeurs au-dessous d'un certain seuil de la crête dans le spectre. Dans notre cas le seuil a été placé à - 60dB.

En outre, la MGDF a une enveloppe plutôt plate, qui est obtenu par la présence du spectre de puissance lissé dans le dénominateur de l'équation (2.38).

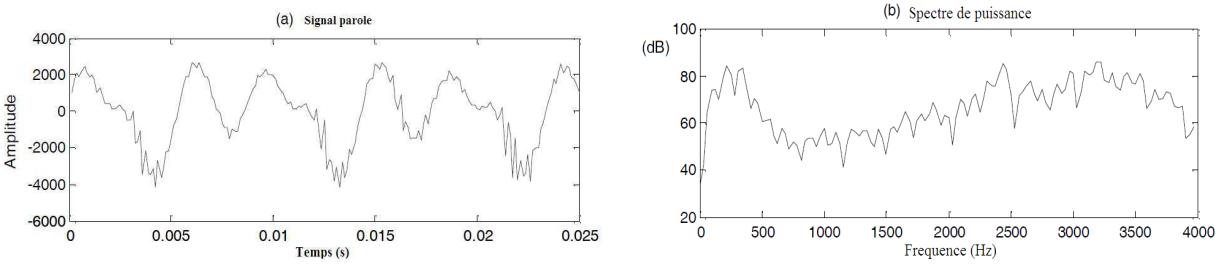
2.3.8.2 Produit spectral

Donglai et Paliwal dans [18] ont défini le produit spectrale par $Q(\omega)$ comme le produit du spectre de puissance par la GDF, et il est défini comme suit :

$$Q(\omega) = |X(\omega)|^2 \tau_p(\omega) \qquad (2.39)$$

$$= X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega) \qquad (2.40)$$

Le spectre du produit est influencé par les deux spectre, spectre d'amplitude et spectre de phase. La figure 2.13 (e) montre le spectre de produit spectral du signal. Il améliore les régions aux formants au-dessus du MGDF et il a une enveloppe comparable à celle du spectre de puissance.



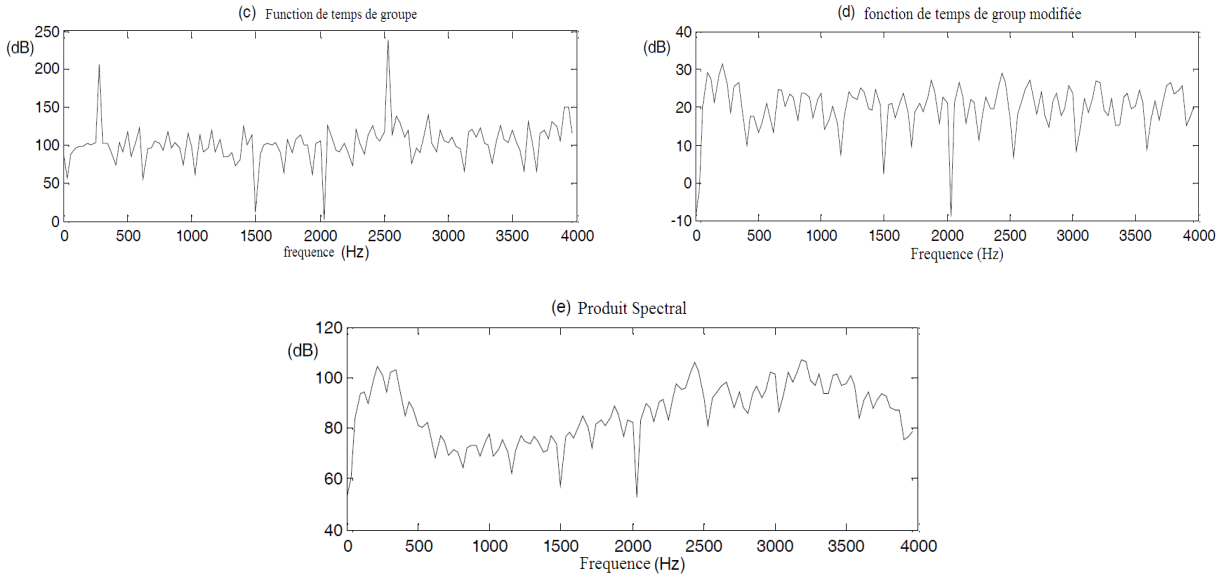


Figure 2.11 une trame de la voyelle (i), son spectre de puissance, fonction de temps de groupe, fonction de temps de groupe modifiée et le produit spectral.

2.3.8.3 Paramètres acoustiques cepstrals

Nous présentons en bref les différents algorithmes pour extraire des paramètres robustes pour la tâche de la reconnaissance proposés par Donglai et Paliwal dans [18], dans ces algorithmes il y a des paramètres qui dépendent du spectre de phase, du spectre de module et ainsi d'autre qui dépendent du produit spectral.

2.3.8.4 Les coefficients cepstrals de la fonction de temps de groupe (MGDCC)

Les coefficients MGDCC sont calculés avec les quatre étapes suivantes [21]:

1- Calcul du spectre de Fourier de $x(n)$ et $nx(n)$ dénoter respectivement par $X(k)$ et $Y(k)$.

2- Calcul du cepstre du spectre de $|X(k)|$ denoter par $S(k)$

3- Calcul de la MGDF comme suit :

$$\tilde{\tau}_p(k) = sign \cdot \left| \frac{X_R(k)Y_R(k) + X_I(k)Y_I(k)}{(S(\omega))^{2\gamma}} \right|^{\alpha} \text{ avec } (\alpha = 0{,}4 \text{ et } \gamma = 0{,}9) \qquad (2.41)$$

où sign est la fonction signe de $\dfrac{X_R(k)Y_R(k) + X_I(k)Y_I(k)}{(S(k))^{2\gamma}}$ (2.42)

4- enfin nous appliquons la cosinus discrète DCT à $\tilde{\tau}_p(k)$ pour obtenir les coefficients MGDCC.

2.3.8.5 Les coefficients cepstrals de la fonction de temps de groupe modifiée (MFGDCC)

Les coefficients MFGDCC sont calculés par les cinq étapes suivantes :

1- Calcul du spectre de Fourier de $x(n)$ et $nx(n)$ dénoter respectivement par $X(k)$ et $Y(k)$.

2- Calcul de la MGDF comme suit :

$$\tilde{\tau}_p(k) = \max\left(\frac{X_R(k)Y_R(k) + X_I(k)Y_I(k)}{(S(\omega))^2}, \rho \right) \qquad (2.43)$$

où $\rho = 10^{\frac{\sigma}{10}} \cdot \max\left(\frac{X_R(k)Y_R(k) + X_I(k)Y_I(k)}{(S(k))^2} \right)$ (2.44)

3- Application de la cosinus discrète DCT à $\tilde{\tau}_p(k)$ pour obtenir les coefficients MGDCC.

σ est le seuil en dB

4- Appliquer le banc de filtre de Mel sur $\tilde{\tau}_p(k)$ pour obtenir l'énergie issue de chaque banc de filtre (E_j)

5- Calcul de la DCT de du logarithme de E_j.

2.3.8.6 Les coefficients cepstrals du produit spectral (MFPSCC)

Les coefficients MFPSCC sont calculés par les quatre étapes suivantes :

1- Calcul du spectre de Fourier de $x(n)$ et $nx(n)$ dénoter respectivement par $X(k)$ et $Y(k)$.

2- Calcul du produit spectral

$$Q(k) = \max\left(X_R(k)Y_R(k) + X_I(k)Y_I(k), \rho\right) \qquad (2.45)$$

où $\rho = 10^{\frac{\sigma}{10}} \cdot \max\left(X_R(k)Y_R(k) + X_I(k)Y_I(k)\right)$ (2.46)

σ est le seuil en dB

3- Appliquer le banc de filtre de Mel sur $Q(k)$ pour obtenir l'énergie issue de chaque banc de filtre (E_j)

4- Calcul de la DCT du logarithme de E_j pour l'obtention des coefficients MFPSCC.

2.4 Les coefficients différentiels

Il est possible d'introduire dans ces systèmes une information sur la dynamique temporelle du signal en utilisant, en plus des paramètres initiaux, des coefficients différentiels du premier ordre issus des coefficients cepstraux ou de l'énergie. Soit $C_k(t)$ le coefficient cepstral d'indice *k* de la trame *t*, alors le coefficient différentiel $\Delta C_k(t)$ correspondant est calculé sur $2\Delta_n + 1$ trames d'analyse par :

$$\Delta C_k = \frac{\sum_{i=-n_\Delta}^{n_\Delta} C_k(t+i)}{\sum_{i=-n_\Delta}^{n_\Delta} i^2} \qquad (2.47)$$

Des coefficients de second ordre peuvent aussi contribuer à l'amélioration du système surtout dans le cas de la parole bruitée soumise à l'effet lombard. Ces coefficients $\Delta\Delta C_k$ et $\Delta\Delta E$ sont calculés par la régression linéaire des coefficients delta sur $n_{\Delta\Delta}$.

2.5 Evaluation des paramètres acoustiques étudiés par le système de RAP de référence

Une séries d'expériences est effectuée afin chercher quel sont les paramètres acoustiques les plus adaptés à la RAP par les HMMc, les évaluations ont été faites en présence de quatre types de bruits additifs réels (blanc, rose, industriel, cockpit F16). La base de données utilisée dans nos expériences contient 90 locuteurs (46 hommes et 44 femmes), chaque locuteur prononce 10 fois le même chiffre arabe (0-9). 6 locutions ont été utilisées pour l'apprentissage du système de référence (chapitre5) et les quatre restantes sont utilisées pour les tests.
Une description détaillée de la base de données vocale et de la base de bruit ainsi que le système de référence est présentée dans le chapitre5, section3.

Nous présentons les différentes configurations des paramètres acoustiques par 4 tableaux ci-dessous :

Longueur de la fenêtre :	25 ms
Pas de traitement	10 ms
Fenêtre utilisée	Hamming
Coefficient de pre-accentuation	0,97
LPC ordre	12
Ajout du logarithme d'énergie	E_j
Ajout des coefficients dynamiques	Δ, Δ Δ

Tableau 2.1 configuration du paramètre LPC

Longueur de la fenêtre	25 ms
Pas de traitement	10 ms
Fenêtre utilisée	Hamming
Coefficient de pre-accentuation	0,97
Nombre de coefficients cepstrals	12
Nombre de filtre dans le banc de filtres	22
Nombre de coefficients de lifting	26
σ	-60dB
Ajout du logarithme d'énergie	E_j
Coefficients dynamiques ajoutés	Δ, Δ Δ

Tableau 2.2 configuration du paramètre MFCC

Longueur de la fenêtre	25 ms
Pas de traitement	10 ms
Fenêtre utilisée	Hamming
Coefficient de pre-accentuation	0,97
Nombre de coefficients PLP	12
Elévation de l'échelle à la puissance	0.33
Ajout du logarithme d'énergie	E_j
Coefficients dynamiques ajoutés	Δ, Δ Δ

Tableau 2.3 configuration du paramètre PLP

Longueur de la fenêtre	25 ms
Pas de traitement	10 ms
Fenêtre utilisée	Hamming
Coefficient de pre-accentuation	0,97
Nombre de coefficients cepstrals	12
Nombre de filtre dans le banc de filtres	22
Nombre de coefficients de lifting	26
Ajout du logarithme d'énergie	E_j
Coefficients dynamiques ajoutés	Δ, Δ Δ

Tableau 2.4 configuration du paramètre MFPSCC

2.5.1 Evaluation des performances du système ASR en présence du bruit blanc

	clean	20	15	10	5	0	-5	T_{moyen}
LPC	88,61	77,24	65,46	45,62	25,28	17,37	14,34	47,70
PLP	**98,72**	97,14	94,19	84,47	64,02	33,37	16,45	69,76
MFCC	98,55	97,55	96,03	90,78	76,69	48,04	22,70	75,76
MFPSCC	98,61	**98,33**	**98,08**	**96,44**	**92,47**	**75,85**	**34,04**	**84,83**

Tableau 25 Taux de reconnaissance (%) obtenus avec les différents paramètres en présence du bruit blanc.

2.5.2 Evaluation des performances du système ASR en présence du bruit rose

	clean	20	15	10	5	0	-5	T_{moyen}
LPC	88,61	71,85	56,18	37,68	24,28	19,51	15,03	44,73
PLP	**98,72**	97,33	94,50	85,19	67,41	42,73	23,03	72,70
MFCC	98,55	96,55	91,94	80,30	61,79	35,76	16,00	68,69
MFPSCC	98,61	**89,60**	**97,75**	**96,33**	**91,05**	**71,13**	**42,01**	**85,06**

Tableau 2.6 Taux de reconnaissance (%) obtenus avec les différents paramètres en présence du bruit rose.

2.5.3 Evaluation des performances du système ASR en présence du bruit industriel

	clean	20	15	10	5	0	-5	T_{moyen}
LPC	88,61	68,02	51,10	34,90	24,40	19,84	14,37	43,03
PLP	**98,72**	96,55	92,00	80,44	60,82	38,65	22,78	69,99
MFCC	98,55	95,11	88,77	75,44	57,57	35,59	20,06	67,29
MFPSCC	98,61	**98,59**	**97,36**	**95,94**	**90,28**	**71,69**	**40,54**	**84,71**

Tableau 2.7 Taux de reconnaissance (%) obtenus avec les différents paramètres en présence du bruit industriel.

2.5.4 Evaluation des performances du système ASR en présence du bruit de cockpit de l'avion de chasse F16

	clean	20	15	10	5	0	-5	T_{moyen}
LPC	88,61	67,69	50,93	35,01	24,48	18,92	11,92	42,50
PLP	**98,72**	95,55	88,22	73,13	52,99	30,01	20,28	65,55
MFCC	98.55	94.28	85.94	72.55	54.29	34.04	17.09	65,24
MFPSCC	98,61	**98,60**	**97,17**	**94,69**	**85,79**	**63,02**	**30,90**	**81,25**

Tableau 2.8 Taux de reconnaissance (%) obtenus avec les différents paramètres en présence du bruit du cockpit de l'avion de chasse F16.

2.5.5 Discussion des résultats

Les tableaux ci-dessus présentent les taux de reconnaissance obtenus à partir du système de reconnaissance de référence, dans le cadre de nos expériences, nous constatons que les coefficients MFCC et PLP pressentent de bonnes performances dans le milieu non bruité. Mais les performances du système sont loin d'être satisfaisantes en présence de bruit, les coefficients MFCC se dégradent de façon considérable et plus rapidement par rapport aux coefficients PLP.

Les coefficients MFPSCC offrent de bonnes performances au système et présentent un apport majeur sur le taux de reconnaissance de 10 à 15% en valeur moyenne et sur 7 niveaux de SNR par rapport aux coefficients MFCC et PLP.

Les coefficients LPC ne sont pas adaptés à la tache de reconnaissance en milieu bruité.

2.6 Conclusion

Nous constatons que le paramètre acoustique MFPSCC présent de bonnes performances au système RAP, et offre une robustesse par rapport aux paramètres acoustiques étudiés, Mais son inconvénient est la dégradation pour les SNR (SNR $\leq$ 5dB) surtout en présence de bruit blanc et rose. Pour remédier à ce problème, qui présente un handicap et un frein pour la RAP, nous proposons un nouveau paramètre acoustique basé sur le paramètre MFPSCC est plus adapté au bruit et offrant des meilleurs performances en des faible SNR. Nous présenterons au chapitre 5 les détails et les différentes étapes de développement de notre nouveau paramètre.

Chapitre 3

Systèmes de reconnaissance automatique de la parole

3.1 Introduction

La reconnaissance automatique de la parole est un domaine d'étude très actif depuis le début des années cinquante. Vu la complexité de cette tache Plusieurs méthodes de reconnaissance ont été développées, Nous l'avons principalement restreint aux méthodes de reconnaissance des mots isolés et nous l'avons encore plus particulièrement restreint aux méthodes stochastiques. Ces différentes restrictions nous ont d'ailleurs poussé à focaliser le titre sur les modèles de Markov cachés (HMM) bien qu'il ne soit pas ici le seul sujet de dissertation.

Ce chapitre nous permet de présenter en détail les deux grandes techniques de reconnaissance des formes qui sont utilisées en reconnaissance automatique des mots isolés : programmation dynamique DTW et les modèles de Markov cachés. Ensuite nous donnons un aperçu sur les logiciels de développement de systèmes à bases de HMM, plus particulièrement sur la plate forme logicielle HTK choisie pour le développement de notre système ASR. A la fin de ce chapitre nous survolons les méthodes de reconnaissance hybrides (HMM/ANN, HMM/SVM…..etc.) les plus efficaces dans cette discipline.

3.2 L'alignement temporel

L'alignement temporel, plus connu sous l'acronyme de DTW, *Dynamic Time Warping*, est une méthode fondée sur un principe de comparaison d'un signal à analyser avec un ensemble de signaux stockés dans une base de référence. Le signal à analyser est comparé avec chacune des références et est classé en fonction de sa proximité avec une des références stockées. Le DTW est en fait une application au domaine de la reconnaissance de la parole [74] de la méthode plus générale de la programmation dynamique [77]. Elle peut ainsi être vue comme un problème de cheminement dans un graphe [78], [44].

Ce type de méthode pose deux problèmes : la taille de la base de référence, qui doit être importante, et la fonction de calcul des distances, qui doit être choisie avec soin. La taille de la base contenant les signaux de référence est directement liée aux capacités, variables, de reconnaissance du système d'alignement temporel. Chacun des signaux de référence est en effet stocké dans son état brut, sans compression d'aucune sorte. Ce stockage permet de disposer d'un vocabulaire dont la taille correspond au nombre de mots du vocabulaire multiplié par le nombre de locuteurs et le nombre des éventuelles répétitions des mots. Cette base de référence permet d'effectuer une mise en correspondance entre le signal stocké, d'une part, et sa retranscription

symbolique d'autre part.

La taille de la base de référence est importante et implique une charge de travail non négligeable puisque la classification de chaque forme à analyser impose de la comparer à chaque forme de la base de référence. Donc, si la constitution de la base de référence est assez rapide et si le processus d'apprentissage est inexistant dans la méthode de l'alignement temporel, la phase d'utilisation nécessite une puissance de calcul non négligeable pour chaque référence atomique de signal à analyser. Le schéma de principe de la méthode est présenté dans la figure 3.1.

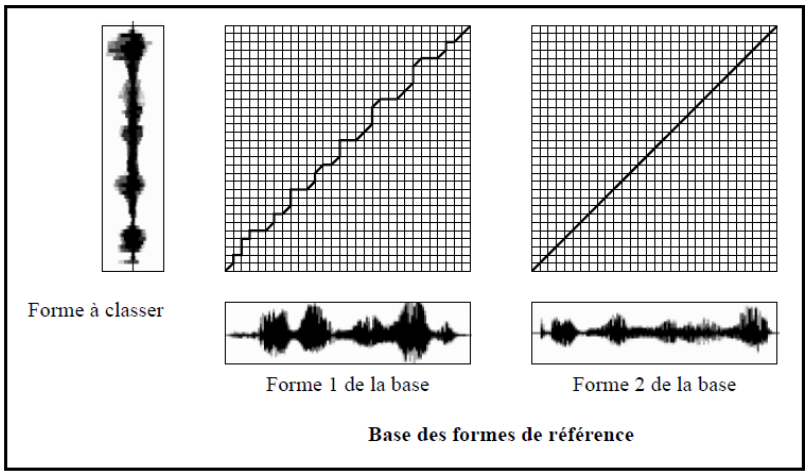


Figure 3.1 : Visualisation du cheminement de l'alignement temporel pour des formes de la base de référence.

Comme le montre le schéma de la figure 3.1, la forme choisie sera celle pour laquelle le chemin de mise en correspondance est le plus court, cette taille minimale marquant le peu de différences entre la forme à analyser et la forme de référence.

L'autre partie importante de l'alignement temporel est la définition de la fonction de recalage qui permet de calculer, selon certaines contraintes, la distance entre la forme à comparer et la forme de référence. La forme à analyser est mise en correspondance dans le plan temporel par l'algorithme d'alignement qui essaie de trouver le plus court chemin dans le graphe ainsi constitué. Cette fonction de mise en correspondance définit une valeur pour chaque arc du graphe, ces valeurs favorisant l'axe médian qui correspond à une parfaite mise en relation de la forme à analyser et d'une forme de référence comme le montre la figure 3.1.

La fonction de recalage suit typiquement le schéma présenté dans la figure 3.2. La fonction *d(i,j)*

est la fonction de calcul de la distance entre deux points successifs du graphe. Les valeurs α, β et γ permettent de définir une partie du comportement de la fonction *d* qui peut être soit symétrique ($\alpha = \gamma$) soit asymétrique ($\alpha \neq \gamma$). Ce calcul de distance entre deux noeuds successifs du graphe n'est cependant pas suffisant pour calculer la longueur totale du chemin parcouru dans le graphe. Une fonction supplémentaire, *G*, calcule une longueur totale qui permettra, après le calcul de cette longueur des chemins sur toutes les formes de la base de référence, de savoir à quel mot du vocabulaire préenregistré correspond la forme à classer. D'un point de vue mathématique, *M* et *N* étant les longueurs respectives de la forme à classer et de la forme de référence, on cherche sur l'ensemble du corpus la distance globale minimale *D(M,N)*.

3.2.1 Distance globale

Les distances cumulées représentent la dissemblance entre les références et les tests. Alors que les distances locales représentent la dissemblance entre les deux signaux en un instant donné, la distance cumulée en un point est la somme des distances locales depuis l'origine en suivant le chemin optimal, c'est à dire de moindre coût. Pour préserver une certaine cohérence dans le calcul du chemin optimal, les transitions autorisées entre les points du graphe de coïncidence sont limitées à quelques uns des points les plus proches.

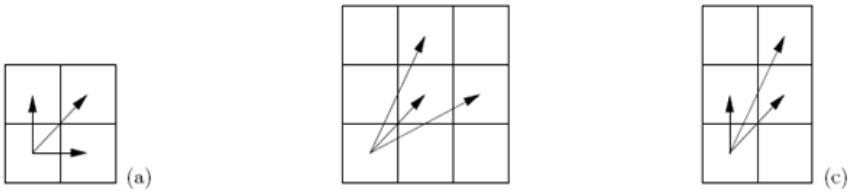


Figure 3.2 les transitions autorisées entre les points du graphe.

La distance cumulée au point *(i,j)* est obtenue de manière récursive par la formule suivante :

$$G(i,j) = \min\begin{Bmatrix} G(i-1,j) + d(i,j) \\ G(i-1,j-1) + d(i,j) \\ G(i-1,j-2) + d(i,j) \end{Bmatrix} \qquad (3.1)$$

La distance globale entre l'observation *T* et la référence *R* est alors donnée par

$$D(R,T) = G(M,N)/(M+N) \qquad (3.2)$$

Où *M* et *N* sont respectivement les nombres de trames des signaux *R* et *T*.

Le calcul de cette fonction *G* répond au même principe que le principe général énoncé par Bellman pour la programmation dynamique : toute sous-partie du chemin optimal est lui-même un chemin optimal. Des exemples de fonctions *d* et *G* de calcul de distance, qui peuvent être bien plus complexes que la fonction de recalage présentée en figure 3.3, pourront être trouvées dans [70] ou [69]. Dans ces références, les fonctions présentées peuvent analyser jusqu'à 9 chemins différents pour *d*, la fonction *G* étant de complexité égale à celle de *d*.

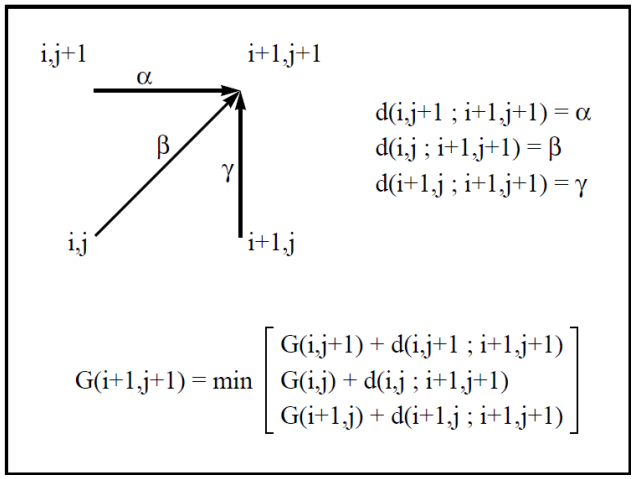


Figure 3.3 : Schéma typique d'une fonction de recalage en alignement temporel.

Cette méthode de reconnaissance des formes est, initialement, bien adaptée à la reconnaissance de mots isolés mais des extensions ont été développées pour permettre de l'appliquer à la parole continue [64] et [66].

D'autres méthodes complémentaires ont par ailleurs été développées pour tenter de réduire la taille de la base des formes de référence par sélection optimale des formes à conserver. Ces méthodes reposent surtout sur une exploration statistique de la base des formes de référence et permettent d'obtenir une caractérisation des différents ensembles la constituant, ces ensembles correspondant aux différents symboles référencés dans la base. Une des techniques qu'il est possible d'employer pour ce faire est, par exemple, la méthode des plus proches voisins. Certaines méthodes permettent de réduire ce temps de calcul à l'utilisation par apprentissage a

priori de coefficients qui permettent de compacter la connaissance présente dans la base de référence qui devient ainsi un corpus d'apprentissage. Une première méthode mettant en oeuvre ce principe de compactage de la connaissance est le modèle de Markov.

3.3 Les Modèles de Markov cachés

3.3.1 Définitions

Un modèle de Markov λ est un automate probabiliste d'états finis constitué de *N* états. Un processus aléatoire se déplace d'état en état à chaque instant, et on note q_t le numéro de l'état atteint par le processus à l'instant *t*. L'état réel q_t du processus n'est pas directement observable — on dit qu'il est "caché" — mais le processus émet après chaque changement d'état un symbole discret *ot* qui appartient à un alphabet fini de n_v symboles $V = \{v_k\}$, $1 \le k \le n_v$. Dans le cas d'un processus markovien du premier ordre, la probabilité de passer de l'état *i* à l'état *j* à l'instant *t* et d'émettre le symbole v_k ne dépend ni du temps, ni des états aux instants précédents. Un modèle de Markov caché ou HMM est alors défini par:

- $S = \{s_i\}$, $1 \le i \le N$ l'ensemble des *N* états, en sachant que le processus part de l'état initial s_1 à l'instant *t*=0 et arrive à l'état final s_N à l'instant *t*=*T*
- $A = \{a_{ij}\}$ $1 \le i, j \le N$ l'ensemble des probabilités de transition entre les états *i* et *j* :

  $$A_{ij} = P(q_t = j / q_{t-1} = i) \qquad (3.3)$$

- $V = \{v_k\}$ $1 \le k \le n_v$ l'ensemble des n_v symboles observables,
- $B = \{b_j(k)\}$ $1 \le j \le N,\ 1 \le k \le n_v$ l'ensemble des probabilités d'émission du symbole v_k lors de l'arrivée dans l'état *j*, avec :

  $$B_j = P(O_t = v_k / q_t = j) \qquad (3.4)$$

Des variantes existent cependant. La probabilité d'émission est parfois notée $b_{ij}(k)$ dans le cas où l'on associe l'émission du symbole à la transition plutôt qu'à l'état d'arrivée:

$$B_{ij} = P(O_t = v_k / q_{t-1} = i, q_t = j) \qquad (3.5)$$

La réalisation par la machine d'un processus markovien de durée *T* est décrite par :

- $Q = (q_1 \ldots q_T)$ un chemin *a priori* caché parmi les *N* états; on pose de plus par convention $q_0 = 1$ puisque tous les processus partent de l'état initial à l'instant *t* = 0, et on impose l'arrivée à l'état final par $q_T = N$.
- $O = (O_1 \ldots O_T)$ une suite d'observations appartenant à l'alphabet de n_v symboles.

Formellement, un model de Markov caché peut être défini par l'ensemble des paramètres λ

$$\lambda = (A, B, \pi) \qquad (3.6)$$

la distribution initial des états π :

$$\pi_i = P(q_1 = S_i), \quad 1 \le i \le N \qquad (3.7)$$

3.3.2 Les trois problèmes de base en HMMs

Une fois choisie la topologie d'un HMM, sa mise en œuvre nécessite la résolution des trois problèmes :

3.3.2.1 Evaluation de la vraisemblance

L'évaluation de la probabilité que la suite des observations ait été émise par un modèle. Lorsque plusieurs modèles existent, cette évaluation permet le choix du modèle le plus probable.

3.3.2.2 Le décodage

La recherche de la séquence d'états d'un modèle ayant produit les observations. La séquence cachée de plus forte probabilité est déterminée par l'algorithme de Viterbi.

3.3.2.3 L'apprentissage

L'apprentissage des paramètres d'un modèle. A partir d'un modèle donné *a priori* et d'observations supposées émises par ce modèle, on cherche les probabilités de transition et d'émission maximisant la vraisemblance des observations.

La solution au problème de l'évaluation de la vraisemblance nous donne un moyen de mesurer l'adéquation d'une séquence d'observation à un modèle. Ainsi nous pouvons décider du meilleur modèle selon la règle de bayes, Résoudre le problème du décodage permettra de segmenter les séquences par la recherche de la séquence d'états de vraisemblance maximale. Enfin, l'apprentissage doit permettre d'adapter automatiquement un HMM à un ensemble de données particulier.

3.3.3 Résolution des trois problèmes

3.3.3.1 Problème 1 : Estimation des probabilités

Le problème de l'estimation des probabilités peut être énoncé de la façon suivante : étant donné un modèle de Markov *M*, comment calculer la probabilité $P(O/M)$ qu'il génère la séquence de d'observation O ?

On considère la séquence d'état Q

$$Q = q_1 \dots q_T \qquad (3.8)$$

où q_1 est l'état initial, la probabilité de la séquence d'observation O pour une séquence d'état Q est :

$$P(O/Q,\lambda) = \prod_{t=1}^{T} P(O_t/q_t,\lambda) \qquad (3.9)$$

On considère que les observations sont statistiquement indépendantes, cela nous donne :

$$P(O/Q,\lambda) = b_{q_1}(O_1)\cdot b_{q_2}(O_2)\cdots b_{q_T}(O_T) \qquad (3.10)$$

La probabilité de la séquence d'état Q peut être écrite de la façon suivante :

$$P(Q/\lambda) = \pi_{q_1}\cdot a_{q_1q_2}\cdots a_{q_{T-1}q_T} \qquad (3.11)$$

La probabilité jointe de *O* et *Q*, est la probabilité de la production de O et Q simultanément, elle peut être décomposé simplement en deux termes :

$$P(O/Q,\lambda) = P(O/Q,\lambda)\cdot P(Q,\lambda) \qquad (3.12)$$

La probabilité de O est obtenu par la somme des probabilité jointe par rapport à tous les état possible de la séquence Q donne

$$P(O/\lambda) = \sum_{Q} P(O/Q,\lambda)\cdot P(Q/\lambda) \qquad (3.13)$$

$$= \pi_{q_1} b_{q_1}(O_1)\cdot a_{q_1q_2} b_{q_2}(O_2)\cdot a_{q_2q_3} b_{q_3}(O_3)\cdots a_{q_{T-1}q_T} b_{q_T}(O_T) \qquad (3.14)$$

Le calcule de $P(O/\lambda)$, donnée par la définition directe, nécessite trop de calcul , pour cela il existe une procédure récurrente de calcul de cette probabilité que nous nous proposons de décrire, c'est l'algorithme 'Forward-Backward' qui fournit un solution exacte à ce problème faisant intervenir tous les chemins dans le modèle HMM.

a) Algorithme forward :

On définir la variable α comme suit :

$$\alpha_t(i) = P(O_1\cdot O_2\cdots O_t\ ,\ q_t = S_i/\lambda) \qquad (3.15)$$

La variable α est définie comme la probabilité partielle de la séquence d'observation $O_1 \cdot O_2 \cdots O_t$ et l'état S_i à l'instant *t*, et donnée par le modèle $\lambda \cdot \alpha_t(i)$ peut être calculer comme suit :

$$\alpha_1(t) = \pi_i b_i(O_1), \quad 1 \le i \le N \qquad (3.16)$$

Pour *t* de *1* à *T*

Pour *j* de *1* à N

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_{t+1}(j) a_{ij} \right] b_j(O_{t+1}), \quad 1 \le i \le T\text{-}1 \quad 1 \le j \le N \qquad (3.17)$$

pour arriver finalement, à l'état final à l'instant *T*

$$P(O/\lambda) = \sum_{i=1}^{N} \alpha_T(i) \qquad (3.18)$$

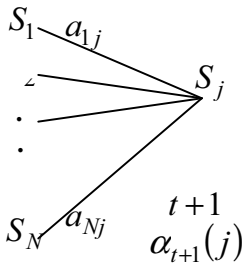


Figure 3.4 Illustration de l'utilisation des récurrences Forward pour le calcule de $\alpha_{t+1}(j)$.

Cette méthode est une formulation simple de l'exploration de la matrice temps / états sous la contrainte des transitions autorisées entre états.

L'estimation direct est suffisante pour obtenir la probabilité définie par toutefois l'apprentissage des modèle sera facilité par (3.7) l'introduction de la probabilité rétrograde.

b) Algorithme backward

Cette dernière est définie comme la probabilité que les trames suivant O_t aient été émises sachant que O_t a été émise par *i* :

$$\beta_t(i) = P(O_{t+1} \cdots O_T / q_t = i / \lambda_i) \qquad (3.19)$$

Le calcul de β est opéré par une récurrence sur le temps en partant de l 'état final *F* au temps T :

Initialement

$$\beta_T(i) = 1 \quad , \quad 1 \leq i \leq N \qquad (3.20)$$

Pour t de T à 1

Pour i de 1 à N

$$\beta_{t-1}(i) = \sum_{j=1}^{N} a_{ij} b_j(O_t) \cdot \beta_t(j) \qquad (3.21)$$

Pour arriver finalement, à l'état initial au temps 1 :

$$P(O, \lambda) = \beta_0(1) \qquad (3.22)$$

L'utilisation conjointe des variables directe et rétrograde permet de calculer la probabilité de l'émission d'une trame sur un état par rapport à tous les chemins possibles :

$$P(O, q_t = i/\lambda) = \alpha_t(i)\beta_0(1) \qquad (3.23)$$

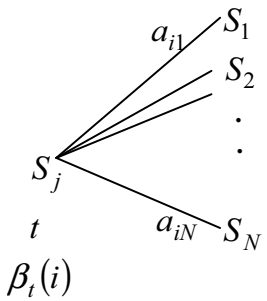


Figure 3.5 Illustration de l'utilisation des récurrences backward pour le calcule de $\beta_t(i)$.

Cette probabilité est utilisée lors de l'apprentissage des modèles par l'algorithme de Baum Welch. Ces variables permettent l'estimation de la vraisemblance de la séquence sur le modèle en tenant compte de l'ensemble des chemins.

3.3.3.2 Probleme2 : le décodage

La procédure d'estimation directe ou rétrograde fournit la probabilité d'émission des observations cumulée sur toutes les séquences d'états possibles, sans choisir un chemin particulier. Il est parfois utile de connaître la séquence d'états qui a émis les observations. L'algorithme de Viterbi cherche la séquence d'états cachés la plus probable et calcule la probabilité d'émission le long de ce chemin. La probabilité ainsi estimée néglige les chemins moins probables.

3.3.3.2.1 Algorithme de Viterbi

La variable $\delta_t(i)$ est définie comme la probabilité maximale que les observations observées jusqu'à l'instant t aient été émises par le modèle λ en suivant un chemin qui arrive à l'état d'indice i:

$$\delta_t(i) = \max_{q_1 \cdots q_{t-1}} P(O_1 \cdots O_t, q_1 \cdots q_{t-1} \cdot q_t = i/\lambda) \quad (3.24)$$

Alors une récurrence similaire à celle suivie pour le calcul de la probabilité d'émission s'applique, à laquelle s'ajoute la mémorisation du meilleur chemin:

Le processus est initialement dans l'état d'indice 1:

$$\delta_1(i) = \pi_i b_i(O_1) \ , \quad 1 \le i \le N \quad (3.25)$$

$$\psi_1(i) = 0 \quad (3.26)$$

Pour t de T à 1

Pour i de 1 à N

$$\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1} a_{ij}] b_j(O_t) \ , \quad 2 \le t \le T \quad 1 \le j \le N \quad (3.27)$$

arrivée du processus dans l'état final :

$$P^* = \max_{1 \le i \le N} [\delta_T(i)] \quad (3.28)$$

$$q_T{}^* = \max_{1 \le i \le N} [\delta_T(i)] \quad (3.29)$$

La probabilité d'émission sur le meilleur chemin peut être utilisée pour la reconnaissance comme une approximation de la probabilité d'émission par le modèle; mais cette méthode de résolution est sous-optimale puisqu'elle néglige les chemins de plus faible probabilité.

La segmentation du signal fournie par l'algorithme de Viterbi sert principalement à l'initialisation des modèles à l'apprentissage et à la reconnaissance de la parole continue.

3.3.3.3 Problème d'estimation des paramètres et entraînement des modèles

Le but de l'entraînement d'un HMM est de trouver l'ensemble des paramètres λ maximisant sur l'ensemble des données d'entraînements O_j la vraisemblance des données étant donné les modèles associés M_j. soit :

$$\arg\max_{\lambda} \prod_{j=1}^{J} P\left(O_j / M_j, \lambda_j\right) \qquad (3.30)$$

3.3.3.3.1 Entraînement Baum-Welch

L'algorithme de baum Welch est un processus itératif, où à chaque itération, de nouvelles valeurs des paramètres a_{ij} des modèles sont estimées à partir des anciennes valeurs. L'entraînement des modèles est effectué à partir de l'estimation de $P(O/M)$ en tenant compte de tous les chemins possibles. La ré-estimation des paramètres du modèle xx est basé sur le comptage du nombre moyen de transitions observées entre les états *i* et *j*. la probabilité w_{ij} de suivre cette transition à l'instant *t* peut s'exprimer au moyen des variables discrètes et rétrogrades introduites aux paragraphes précédents

$$w_t(i,j) = P(q_{t-1} = i, q_t = j/O, \lambda) = \frac{\alpha_{t-1} a_{ij} b_j(O_t) \beta_t(j)}{P(O/\lambda)} \qquad (3.31)$$

le nombre moyen de transition entre i et j est donc :

$$\gamma_{ij} = \sum_{t=1}^{T} w_t(i,j) \qquad (3.32)$$

et la probabilité de transition est ré-estimée par :

$$\overline{a}_{ij} = \frac{\gamma_{ij}}{\sum_{k=1}^{N} \gamma_{ik}} \qquad (3.33)$$

L'estimation de la probabilité d'émission associée à un état nécessite le décompte des observations correspondant à chaque catégorie de symbole :

$$\overline{b}_j = \frac{1}{\gamma_j} \sum_{t=1}^{T} w_t(j), \; 1 \le k \le n_v \qquad (3.34)$$

avec

$$w_t(j) = P(q_t = j/O, \lambda) = \frac{\alpha_t(j) \beta_t(j)}{P(O/\lambda)} \qquad (3.35)$$

et $$\gamma_j = \sum_{t=1}^{T} w_t(j) \qquad (3.36)$$

3.3.4 Cas des modèles continus

Le principe de l'émission de symboles discrets peut se généraliser au cas continu. Les probabilités d'émission discrètes $b_j(k)$ sont alors remplacées par des densités de probabilité continues dans l'espace de représentation. Cette solution évite les distorsions introduites par la QV, mais pose le problème du choix des densités de probabilité et de la robustesse de leur estimation. L'utilisation d'une combinaison linéaire de gaussiennes dans l'espace R^d est fréquente:

$$b_j(O) = \sum_{k=1}^{G} g_k N(O, \mu_k, \Sigma_k) \qquad (3.37)$$

où μ_k et Σ_k sont respectivement la moyenne et la matrice de covariance de la gaussienne, et g_k la pondération qui lui est affectée. Nous rappelons que la densité de probabilité d'une loi normale de moyenne μ et de matrice de covariance Σ en dimension *d* est:

$$N(O, \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}} [\Sigma]^{\frac{1}{2}}} e^{-\frac{1}{2}(O-\mu)' \Sigma^{-1} (O-\mu)} \qquad (3.38)$$

La ré-estimation des probabilités d'émission est différente pour des modèles continus. Nous détaillons le cas de densités de probabilité continues représentées par une gaussienne multi-dimensionnelle, mais ces formules peuvent être généralisées au cas de multi-gaussiennes. Le vecteur de moyenne et la matrice de covariance de la densité de probabilité associée à l'état *i* sont recalculés comme:

$$\overline{\mu}_i = \frac{1}{\gamma_i} \sum_{t=1}^{T} w_t(i) O_t \qquad (3.39)$$

et

$$\overline{\mu}_i = \frac{1}{\gamma_i} \sum_{t=1}^{T} \left\{ w_t(i) (O_t - \mu)(O_t - \mu)' \right\} \qquad (3.40)$$

3.4 Plate-forme logicielle HTK

A fin de réduire au minimum la tache de programmation des différentes parties du système de reconnaissance. Nous avons choisi de mener cette étude en utilisant un logiciel de développement de systèmes de bases de modèles de Markov cachés le plus complet possible au sens des taches à réaliser et dont le programme sources et ouvert pour d'éventuelles fonctions à mettre en œuvre. Le tableau 3.1 est une liste des logiciels libres.

Après une analyse des caractéristiques de chacun de ces logiciels, notre choix s'est finalement porté sur la plate-forme logicielle HTK (Hidden Markov Model Toolkit). elle a été développée à l'Université de Cambridge par S.J. Young et son équipe. Elle est constituée d'un ensemble d'outils logiciels qui permettent de construire des systèmes de reconnaissance de la parole continue à base de modèles de Markov cachés.

	HTK	**Sphinx**	**ISIP (ASR)**	**CSLU (ASR)**
Organisme	Microsoft et Combridge University	Carnegie Mellon University	Mississipi University	Oregon Graduate Institut
URL	htk.eng.cam.ac.uk	fife.speech.cs.cmu.edu	www.isip.msstate.edu	www.cslu.ogi.edu
Langage	C	C, perl, Java	C++	C, Tcl/Tk
Environnement	Unix, Linux, Windows	Unix, Linux, Windows	Unix	Windows
Support	Excellent	Moyen	Bon	Moyen
Date de la première version	1993	1987	1997	1992
Disponibilité du source	Sous licence	Sous licence	Domaine publique	Sous licence

Tableau 3.1 Quelques caractéristiques des logiciels libres de développement de systèmes de reconnaissance de parole à base des HMM.

Contrairement aux autres logiciels figurant dans le tableau 3.1, HTK a connu une période de commercialisation. De ce fait, HTK est passé par les différents cycles de perfectionnement nécessaires au logiciel commercial. Il est par conséquent plus documenté, plus convivial et plus souple que les autres logiciels.

HTK est remarquable par la très grande liberté de choix laissée tout au long de la construction du système de reconnaissance. Les modèles peuvent représenter des mots ou tout type d'unité sub-lexicale, et leur topologie est librement configurable. Les densités de probabilité d'émission, qui sont associées aux états, sont décrites par des multi-gaussiennes.

Les modèles sont initialisés avec l'algorithme de Viterbi, puis ré-estimés par l'algorithme optimal de Baum-Welch. Le décodage est réalisé par l'algorithme de Viterbi, sous la contrainte d'un réseau syntaxique défini par l'utilisateur, et le résultat est enfin évalué par alignement dynamique avec la chaîne phonétique ou lexicale de référence.

L'ensemble de ces outils est écrit en langage C, et la documentation détaille leur utilisation et les principes de leur implémentation, ce qui permet d'intégrer de manière efficace les modifications souhaitées dans le système de reconnaissance. De plus, HTK est un système largement répandu dans le monde de la recherche; en 1992, ses concepteurs revendiquaient déjà plus d'une centaine d'utilisateurs.

Les outils de base manipulent des fichiers de différents types: signaux, étiquettes, paramètres, description des modèles, définition de réseaux. Les formats des fichiers de signaux et d'étiquettes des bases de données les plus répandues sont reconnus. Les autres fichiers sont dans un format particulier à HTK, décrit dans le manuel de référence. En particulier, les modèles et les réseaux sont définis dans des fichiers texte, ce qui facilite leur création et leur modification par l'utilisateur. Les options d'utilisation des outils sont transmises en argument sur la ligne de commande, ce qui facilite la tache l'automation des processus d'apprentissage et de décodage avec des scripts écrits dans le langage de commande du système d'exploitation.

3.4.1 Utilisation d'HTK

Les principaux outils de base de HTK s'enchaînent naturellement pour réaliser les différentes étapes d'un système de reconnaissance. Toutes les fonctionnalités de HTK sont définies dans des modules constituant la librairie qui assure l'interfaçage avec les objets extérieurs et constitue la ressource commune aux outils permettant :

- l'analyse du signal de parole.
- la manipulation des transcriptions orthographiques et phonétiques.
- la définition de dictionnaires de prononciation.
- la définition de modèles du langage.
- l'apprentissage et l'adaptation des modèles acoustiques.
- le décodage acoustico-phonétique de parole.
- l'alignement de parole sur des transcriptions linguistiques.

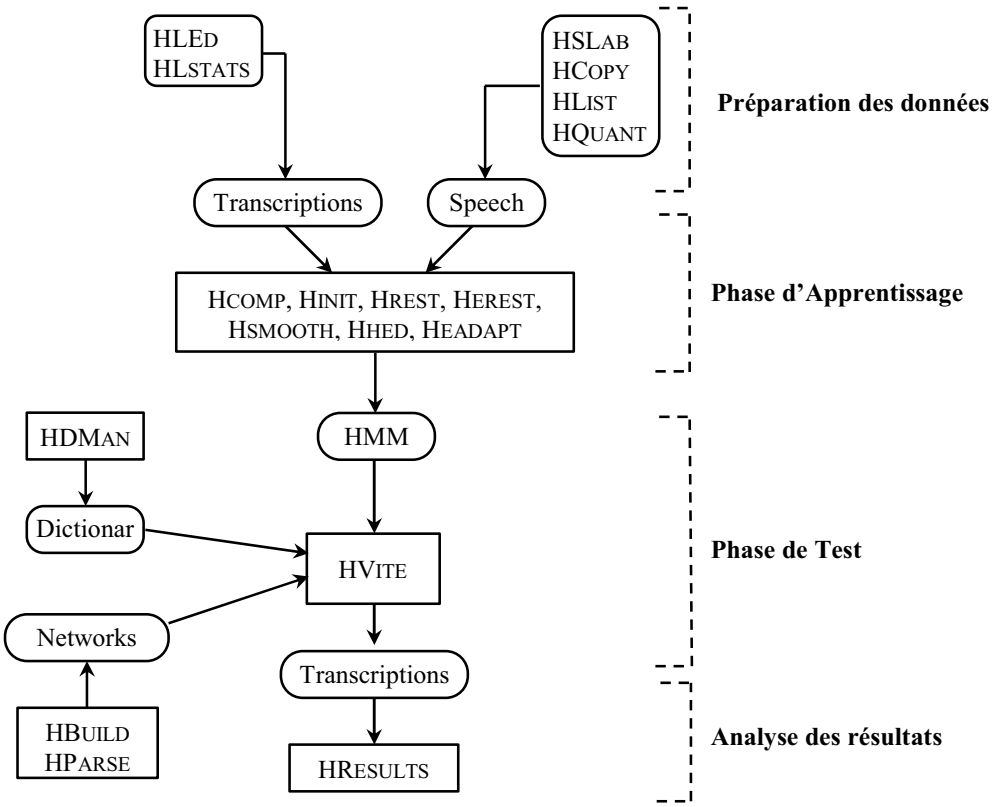


Figure 3.6 Structure d'un système de reconnaissance avec HTK.

3.5 Autres méthodes de reconnaissance

3.5.1 Les réseaux de neurones: le perceptron multi-couches (PMC)

Les réseaux de neurones (RN) constituent un domaine de recherche très intéressant et sont très couramment utilisés lorsque l'on parle de classification. Ils ont été notamment appliqués à des problèmes tels que: la reconnaissance de visage, le contrôle de robot, la reconnaissance de la parole, l'identification du locuteur etc.

Les RN réalisent un traitement d'informations distribué et sont composés d'unités de calcul primitives (les neurones formels) fonctionnant en parallèle et reliées entre elles par des connexions. Un neurone formel reçoit un nombre variable d'entrées en provenance de neurones en amont. A chacune de ces entrées est associé un poids représentant la force de la connexion. Il est aussi doté d'une sortie unique qui se ramifie ensuite pour alimenter les neurones en aval. Le principe de fonctionnement du neurone est simple, il calcule la somme pondérée de ses entrées et

passe cette valeur à une fonction d'activation qui détermine l'excitation de ce neurone. La figure 3.7 illustre l'architecture d'un neurone formel. La sortie du neurone $y = F\left(\sum_{i=1}^{n} w_i x_i\right)$ dépend de la fonction d'activation choisie: fonction seuil, linéaire par morceaux, sigmoïde, gaussienne etc.

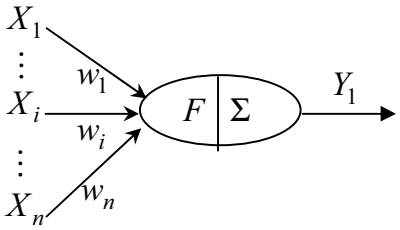


Figure 3.7 Architecture d'un neurone formel à *n* entrées

Dans un réseau, la connaissance se trouve dans la topologie même du réseau et dans les poids des connexions. L'apprentissage d'un RN est réalisé à l'aide de méthodes d'apprentissage automatique utilisant la descente du gradient de l'erreur et se fait par modification des poids des connexions du réseau en fonction des données d'apprentissage. Aucune hypothèse sur la distribution des données n'est nécessaire.

Enfin, les RN ont de nombreuses propriétés très intéressantes telles que leur robustesse au bruit, leur flexibilité et leur capacité importante de généralisation. Nous allons présenter rapidement le réseau de neurones le plus souvent utilisé dans le domaine de la reconnaissance

Automatique de la parole: le perceptron multi-couches (PMC).

3.5.2 Le Perceptron Multi-Couches (PMC)

Le perceptron Multi-Couches est issu des travaux de F. Rosenblatt sur le perceptron monocouche [76]. Un PMC est un réseau dont les neurones sont disposés en plusieurs couches successives et où chaque neurone d'une couche est connecté à tous les neurones de la couche suivante et de la couche précédente mais pas aux neurones de la même couche.

Le PMC est un réseau passe-avant (feed-forward), c'est-à-dire que les informations ou activations ne vont circuler que dans un seul sens, des neurones de la couche d'entrée vers les neurones de la couche de sortie (Figure 3.8).

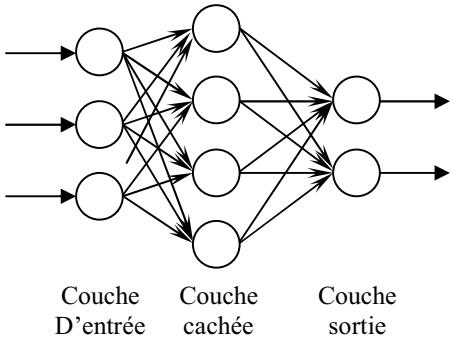


Figure 3.8 Architecture d'un perceptron Multi-Couches à une couche cachée.

Une couche cachée dans un PMC correspond à une couche qui n'est ni la couche d'entrée, ni celle de sortie. De plus, un PMC peut avoir autant de couches cachées que désirées mais il a été montré [59] que quelque soit le nombre de couches cachées dans un PMC, il existe un PMC équivalent avec une seule couche cachée. Cette couche cachée permet de modéliser des fonctions de décisions complexes et non linéaires entre n'importe quels espaces d'entrée et de sortie.
L'apprentissage des PMC se fait par rétropropagation du gradient de l'erreur [62]. Le principe est d'adapter les différents poids des connexions en propageant l'erreur commise en sortie du réseau.

3.5.3 Les Machines à Vecteurs Support (SVM)

Les SVM, introduites par Vapnik et ses collègues [51] comme une nouvelle classe d'algorithmes d'apprentissage, constituent une application directe du principe inductif de minimisation structurel du risque [63]. Elles sont utilisées dans les trois problèmes classiques en apprentissage (régression, estimation de densité et discrimination). Ces différents algorithmes se caractérisent par le choix de maximiser les capacités en généralisation d'une fonction de discrimination *f* en minimisant une borne supérieure sur le risque. Le risque est l'erreur en généralisation de la fonction de discrimination *f* et correspondant à la probabilité que le résultat de *f* soit erroné. La borne supérieure sur le risque est ce que l'on appelle le risque *garanti*.
Dans le cadre de la discrimination, la SVM, à l'instar d'un perceptron, tente de séparer linéairement les données. Cependant, dans l'espace où elles se trouvent, les données ne sont généralement pas linéairement séparables. Dans ce cas, il devient utile d'effectuer un pré-traitement sur les données avant de les séparer avec des hyperplans. Ainsi, dans l'exemple représenté sur la figure 3.9, on peut pré-traiter les points de R^2, en les projetant sur la surface d'un

paraboloïde bien choisi. D'une manière générale, on projette les données, à l'aide d'une fonction Φ dans un espace de plus grande dimension, appelé "espace de représentation", où l'on espère qu'elles seront linéairement séparables. On parle alors de SVM linéaire lorsque cette application Φ correspond à la fonction identité, i.e. lorsqu'elle ne renvoie pas les données dans un nouvel espace de représentation, et de SVM non linéaire dans le cas contraire.

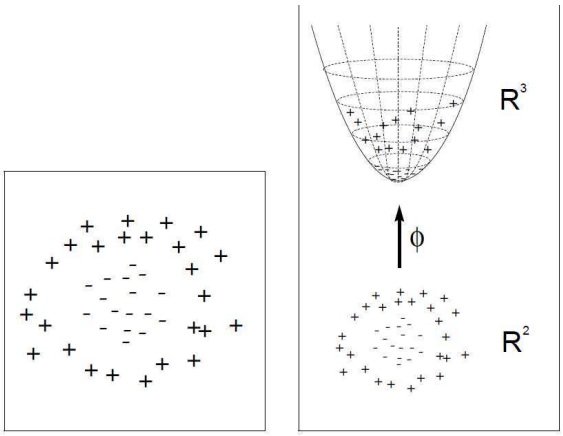


Figure. 3.9 (a) données non linéairement séparables. (b) Pré-traitement des données, choix d'une transformation Φ (projection sur un paraboloïde) rendant les données linéairement séparables.

Enfin, les SVM ont été développés initialement dans le cadre d'une classification bi-classes, mais des extensions multi-classes ont été proposés, comme la M-SVM [12] Les SVMs ont récemment été introduites en reconnaissance de la parole et ont donné des résultats prometteurs [13], [14].

3.5.4 Méthodes "hybrides"

Les HMMs sont largement utilisés dans le domaine de la parole, plus particulièrement en reconnaissance de la parole. Mais ils présentent aussi quelques limitations comme le besoin de faire des hypothèses simplificatrices pour leur fonctionnement qui entraîne une limitation de leur généralité. De plus, leur apprentissage n'est en général pas discriminant.

La combinaison des HMMs avec des méthodes discriminantes semble intéressante et a été utilisée avec succès en reconnaissance de la parole et en discrimination parole/bruit [22], [32]. Deux associations sont souvent utilisées: HMM-RN et HMM-SVM.

3.5.4.1 HMM et réseaux de neurones

Dans cette approche hybride, le réseau de neurones (la plupart du temps un PMC) se situe en aval

d'un HMM et est utilisé comme estimateur de probabilités a posteriori d'appartenance à une classe. En effet, il a été démontré [61], [56] qu'un PMC entraîné dans des conditions adéquates est équivalent à un estimateur de probabilités a posteriori l'appartenance à une classe.

Un perceptron peut ainsi apprendre les probabilités a posteriori des classes de phonèmes. Ces probabilités, grâce à la formule de Bayes, permettent d'obtenir les vraisemblances des observations qui vont être utilisées à la place de celles fou mies par un mélange de gaussiennes dans un HMM classique.

Une autre façon d'utiliser la sortie du PMC comme entrée d'un HMM est illustrée par la figure 3.10 Ce système est celui de [22].

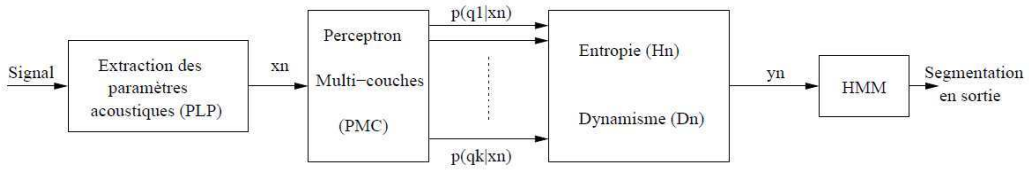


Figure 3.10 Système de segmentation parole/musique.

Des coefficients cepstreux (PLP) sont extraits tous les 16ms. Un PMC reçoit ces coefficients en entrée et donne en sortie des probabilités a posteriori pour les différentes classes de phonèmes. Les probabilités a posteriori des classes de phonèmes sont ensuite analysées selon leur "entropie" et "dynamisme" pour finalement arriver en entrée du classifieur HMM qui effectuera la segmentation (les probabilités d'émission du HMM ont été estimées en utilisant soit un GMM, soit un deuxième PMC).

3.5.4.2 HMM et SVM

L'hybridation HMM/PMC donnant de bons résultats, il est donc normal de vouloir coupler les HMMs avec d'autres méthodes discriminantes telles les SVM Contrairement aux PMC qui estiment des distributions de probabilité, les SVM estiment directement, à partir des données d'apprentissage, des surfaces de décision. Différentes méthodes ont été proposées pour convertir la distance d'une observation inconnue à une surface de décision fournie par une SVM en probabilités *a posteriori* exploitables par un HMM une de ces implantations a consisté à entraîner une SVM sur des données segmentales, en transformant les informations de distance fournie par une SVM en estimation de probabilités a posteriori pour les HMMs [32]. Utilisée notamment en

reconnaissance de la parole bruitée, cette hybridation donne déjà des résultats prometteurs.

Chapitre 4

Application des ondelettes au signal de la parole

4.1 Présentation des ondelettes

4.1.1 Introduction

Cette section présente rapidement la base de notre approche en paramétrisation du signal, à savoir la décomposition du signal en ondelettes. Cette présentation est un rapide aperçu des fondements théoriques des ondelettes. Pour aller plus loin sur cette théorie du traitement du signal à l'aide d'ondelettes, le lecteur pourra se porter au livre de Mallat [34], [29].

Au quotidien, notre attention (visuelle ou auditive) est attirée par le mouvement et les phénomènes transitoires, au contraire des stimuli stationnaires qui sont vite ignorés.

Cette stratégie qui donne la priorité aux phénomènes transitoires permet de sélectionner les informations importantes de notre environnement, information qui, en des temps anciens, nous ont permis de survivre. Pourtant le traitement du signal classique s'est surtout concentré sur l'étude d'opérateurs invariants dans le temps et dans l'espace, qui modifie les propriétés stationnaires des signaux. Cela conduit à l'hégémonie indiscutable de la transformée de fourier. La transformée de Fourier est un outil fondamental pour une grande variété d'applications, telles que la transmissions ou traitements des signaux stationnaires. Néanmoins, si nous nous intéressons à des phénomènes transitoires, la transformée de Fourier s'avère inadéquat. En effet, nous pouvons définir un morceau musical comme un ensemble de 'fréquences' sonores qui varient dans le temps.

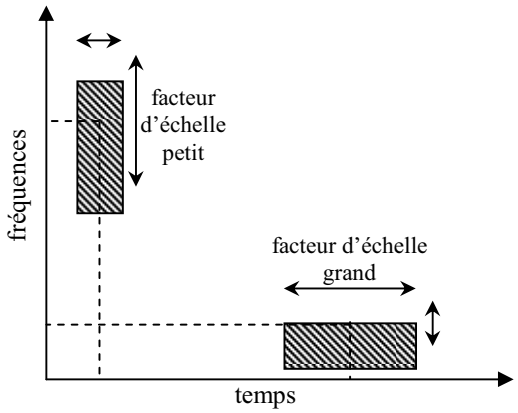


Figure 4.1 Boite de Heisenberg correspondant au pavage du plan temps/fréquence de la transformée en ondelettes à des échelles différentes. Une échelle plus petite réduit l'étalement en temps mais augmente la taille du support fréquentiel.

De telles évolutions temps/fréquence peuvent être mises en évidence en décomposant le signal en fonctions élémentaires bien concentrées en temps et en fréquence. La transformée de Fourier à

fenêtre et la transformée en ondelettes sont deux exemples importants de décomposition temps/fréquence. C'est en 1946 que le physicien Gabor [80] propose d'analyser les signaux sonores avec des atomes élémentaires qui sont des fonctions bien concentrées en temps et en fréquence. En montrant que de telles décompositions sont étroitement liées à notre perception des sons, et qu'elles isolent les structures importantes des signaux de parole et de bruit, les travaux de Gabor furent à la base de 'analyse temps/fréquence. Gabor introduit ainsi en 1946 les atomes de Fourier à fenêtre afin de mesurer les 'variations fréquentielles' des sons.

La résolution temps/fréquence de la transformée de Fourier à fenêtre dépend de l'étalement de la fenêtre en temps et en fréquence. Cet étalement correspond à la surface de la boite de Heisenberg. En effet, les concentrations en temps et en fréquence sont limitées par le principe d'incertitude d'Heisenberg. Ce principe, qui dit que l'énergie d'une fonction et de sa transformée de Fourier ne peuvent être simultanément concentrées sur des intervalles arbitrairement petits, a une interprétation importantes en mécanique quantique, en tant qu'incertitude sur la position et la quantité du mouvement d'une particule libre. En d'autres termes, le principe d'incertitude d'Heisenberg indique qu'un signal ne peut pas être simultanément connu avec des précisions en temps Δt et en fréquence Δf quelconques, le produit de ces deux quantités étant borné inférieurement [29]

$$\Delta t.\Delta f \geq \frac{1}{2} \qquad (4.1)$$

Un inconvénient de la transformée de Fourier à fenêtre est le réglage de taille de la fenêtre d'analyse. Ce réglage est un compromis entre résolution temporelle et résolution fréquentielle. On perd en localisation fréquentielle ce qu'on a gagné en localisation temporelle, ceci à cause du principe d'incertitude de Heisenberg (figure 4.1). Ainsi, une représentation satisfaisante de la structure temporelle fine du signal permettant par exemple de voir les transitions entre phonèmes se fera au détriment de la résolution fréquentielle (analyse large bande). Inversement, une analyse permettant de bien faire apparaître les composantes harmoniques du signal se fera au détriment de la résolution temporelle et ne rendra pas en compte des événements temporels brefs (analyse bande étroite). Une fois ce réglage effectué, la taille de la fenêtre sera fixée et la résolution de la transformée de Fourier à fenêtre restera la même sur tout le plan temps/fréquence (figure 4.2). Mais pour analyser des composantes transitoires de durées différentes comme c'est souvent le cas en parole. Il nécessaire d'utiliser des atomes dont les supports temporels ont des tailles variables.

La transformée en ondelettes en est la solution (figure 4.3)

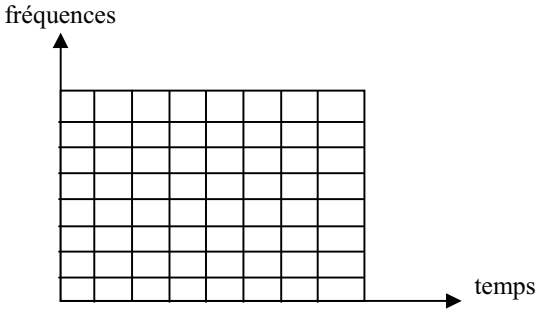


Figure 4.2 un exemple de couverture temps/fréquence avec la transformée de Fourier à fenêtre. Les résolutions temporelle et fréquentielle restent inchangées quelque soit le temps et la fréquence.

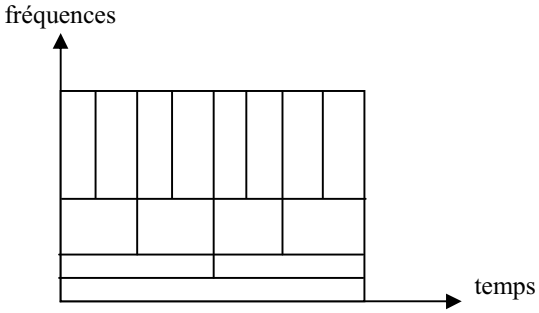


Figure 4.3 un exemple de couverture temps/fréquence avec la transformée en ondelettes.

4.1.2 Définitions

Nous avons vu qu'une alternative, pour dépasser les limitations de la transformée de Fourier à fenêtre, se trouve être l'utilisation de la transformée en ondelettes. Nous pouvons à présent définir ce qu'est une ondelettes [29] et comment une transformée en ondelettes du signal.

4.1.2.1 Les ondelettes

Une ondelette [29] est une fonction $\psi \in L^2(\Re)$ de moyenne nulle :

$$\int_{-\infty}^{+\infty} \psi(t)dt = 0 \qquad (4.2)$$

Et à énergie finie :

$$\int_{-\infty}^{+\infty} |\psi(t)|^2 dt \prec +\infty \qquad (4.3)$$

Elle est normalisée à $\|\psi(t)\| = 1$, et centrée au voisinage de $t = 0$. Une famille d'atomes temps/fréquence s'obtient en dilatant l'ondelette par facteur s, et en la translatant par u :

$$\psi_{u,s}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right) \quad avec \ \ s \in \Re_+^* \qquad (4.4)$$

Si on considère $\left\|\psi_{u,s}(t)\right\| = 1$, alors les ondelettes dilatées restent de norme unitaire. L'ondelette peut être réelle ou analytique complexe. Selon les applications, on peut choisir l'une ou l'autre. Pour notre part, nous avons opté pour une ondelette réelle. Nous allons maintenant définir la transformée en ondelettes.

4.1.2.2 La transformée en ondelettes

La transformée en ondelettes d'un signal $f(t)$ à l'échelle s et au temps u se calcule en corrélant $f(t)$ avec l'ondelette $\psi_{u,s}$ correspondante. Ceci nous donne la définition suivante de la transformée en ondelettes :

$$Wf(u,s) = \left\langle f, \psi_{u,s} \right\rangle = \int_{-\infty}^{+\infty} f(t) \frac{1}{\sqrt{s}} \psi^*\left(\frac{t-u}{s}\right) dt \qquad (4.5)$$

où

- W est l'initiale de *Wavelet* qui signifie ondelette en anglais,
- ψ^* est le complexe conjugué de ψ.

Nous utiliserons par la suite uniquement des transformée en ondelettes réelles car elles permettent de mesurer la variation de $f(t)$ dans un certain voisinage de u (dépendant de ψ) de taille proportionnelle à s. Il a été démonté que lorsque s tend vers 0, la décroissance des coefficients d'ondelettes caractérisent la régularité de $f(t)$ au voisinage de u. Cette propriété est très importante pour nous car elle permet de détecter des transitoires.

Enfin, une transformée en ondelettes réelles est complète et préserve l'énergie tant que l'ondelette ψ satisfait une condition d'admissibilité donnée par le théorème suivant :

Théorème 1 (Calderon, Grossmann, Morlet) soit $\psi \in L^2(\Re)$ une fonction réelle (ou un signal réel) vérifiant :

$$C_\psi = \int_0^{+\infty} \frac{\left|\hat{\psi}(\omega)\right|^2}{\omega} d\omega \prec +\infty \qquad (4.6)$$

Où $\hat{\psi}$ est la transformée de Fourier de ψ.

Toute fonction $x(t) \in L^2(\Re)$ vérifie :

$$x(t) = \frac{1}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} W_x(u,s) \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right) du \frac{ds}{s^2} \qquad (4.7)$$

et

$$\int_{-\infty}^{+\infty} |x(t)|^2 dt = \frac{1}{C_\psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} |W_x(u,s)|^2 du \frac{ds}{s^2} \qquad (4.8)$$

La condition

$$C_\psi = \int_0^{+\infty} \frac{|\hat{\psi}(\omega)|^2}{\omega} d\omega \prec +\infty \qquad (4.9)$$

Du théorème précèdent s'appelle la condition d'admissibilité de l'ondelette. Pour que l'intégrale soit finie, il faut s'assurer que $\hat{\psi}(0) = 0$, ce qui explique pourquoi les ondelettes doivent être de moyenne nulle. Cette condition est presque suffisante. Si $\hat{\psi}(0) = 0$ avec $\hat{\psi}(\omega)$ continûment différentiable, la condition d'admissibilité est alors satisfaite. On vérifie assez facilement que $\hat{\psi}(\omega)$ est continûment différentiable si ψ décroît assez vite à l'infini. C'est pourquoi on choisit aussi des ondelettes à décroissance rapide. Enfin, la dernière équation du théorème démontre la conservation de l'énergie entre le domaine temporel et le domaine des ondelettes.

Le signal de parole est continu mais nous travaillons sur un signal discret $f[n] = f(n)$ (de taille N). Nous utiliserons donc la version discrète de la transformée en ondelettes. La transformée en ondelettes discrète se calcule aux échelles $s = a^j$, avec $a = 2^{1/v}$ ce qui fournit v échelles intermédiaires pour chaque octave $\left[2^j, 2^{j+1}\right[$. De plus, la transformée en ondelettes de f ne pourra être calculée que pour les échelles :

$$\frac{1}{N} \prec s \prec 1 \qquad (4.10)$$

4.1.3 La transformée en ondelettes discrète utilisée pour le débruitage de la parole

Le traitement du signal basé sur les ondelettes a été utilisé avec succès pour des problèmes très variés, comme la reconnaissance de la parole [30], [17], le débruitage de la parole [25], la classification audio [24], [10] et la compression d'image [31],…etc.

L'utilisation des ondelettes permet de faire une analyse multi-résolution du signal. Nous verrons

l'intérêt de ce type d'analyse dans le cadre de débruitage de la parole pour la reconnaissance. Mais tout d'abord, définissons la transformée en ondelettes discrète.

Soit un signal $f(t)$ échantionné uniformément sur $[0,1]$ avec un pas d'échantillonnage de $\frac{1}{N}$. On obtient un signal discret $f[n]=f\left(\frac{n}{N}\right)$ composé de N échantillons.

Soit $\psi(t)$ une ondelette en temps continu dont le support est inclus dans $[-k/2,k/2]$. Pour $2 \le a^i \le \frac{N}{k}$, on définit une ondelette discrète dilatée par a^i :

$$\psi_j(n)=\frac{1}{\sqrt{a^j}}\psi\left(\frac{n}{a^j}\right) \qquad (4.11)$$

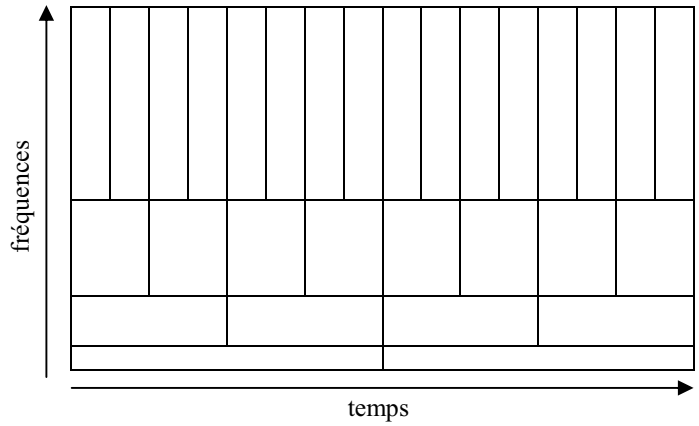


Figure 4.4 Décomposition temps/fréquence du signal. Une décomposition dyadique appliquée à la fois sur l'axe du temps et l'axe des fréquences.

Elle a KNa^j valeur non nulles sur $[-N/2,N/2]$. L'échelle a^j doit être supérieur à 2 pour que le pas d'échantillonnage soit plus petit que le support de l'ondelette. Afin d'éviter des problèmes de bords, $f[n]$ et $\psi[n]$ sont traité comme des signaux de période N.
La transformée en ondelettes discrète peut alors s'écrire comme une convolution circulaire avec

$$\overline{\psi}_j[n]=\psi_j^*[-n] \qquad (4.12)$$

$$Wf[n,a^j]=\sum_{m=0}^{N-1} f[m]\psi_j^*[m-n]=f\otimes\overline{\psi}_j[n] \qquad (4.13)$$

Où ψ^* est le conjugué complexe de ψ et $\otimes$ est l'opérateur de convolution circulaire.

Si nous prenons le cas ou l'échelle est découpée selon une suite dyadique $\{2^j\}_{j\in Z}$, c'est-à-dire lorsque le paramètre d'échelle est $a^j = 2^j$, alors la transformée en ondelettes discrète et dyadique s'écrit :

$$Wf[n,2^j] = \sum_{m=0}^{N-1} f[m]\psi^*_{2^j}[m-n] = f \otimes \overline{\psi}_{2^j}[n] \qquad (4.14)$$

avec

$$\psi_{2^j}(n) = \frac{1}{\sqrt{2^j}}\psi\left(\frac{n}{2^j}\right) \qquad (4.15)$$

La figure 4.5 montre la décomposition temps/fréquence du signal en utilisant la transformée en ondelettes dyadique. La transformée dyadique de f ne peut pas être calculée que pour des échelles $1 \succ 2^j \geq \frac{1}{N}$ la valeur absolue de j sera utilisée par la suite pour représenter les différentes échelles dans l'analyse multi-resolution ainsi que les différentes bandes de fréquence.

L'utilisation de la transformée en ondelettes dyadique nous permet d'obtenir une partition dyadique du plan temps/fréquence de telle sorte que les basses fréquences sont représentées avec une haute résolution fréquentielle et une faible résolution temporelle alors que les hautes fréquences sont représentées avec une haute résolution temporelle et une faible résolution fréquentielle (figure 4.4). La résolution temporelle est inversement proportionnelle à la résolution fréquentielle à cause du principe d'incertitude d'Heisenberg. Cette partition permet d'avoir une résolution fréquentielle qui se rapproche de celle de l'oreille humaine, analyse fine des basses fréquences et qui diminue de manière logarithmique lorsque l'on montre en fréquence (figure 4.5). C'est une approximation de l'échelle Mel, très utilisée en reconnaissance de la parole et notamment avec les MFCC.

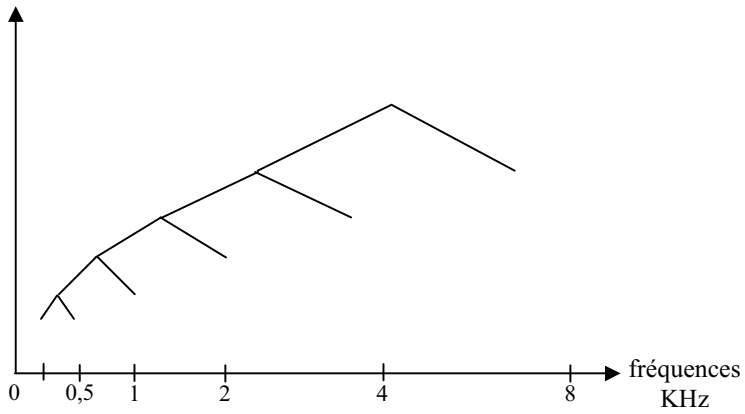


figure 4.5 Résolution fréquentielle obtenue à l'aide de la décomposition en ondelettes dyadique. (Arbre de décomposition dyadique avec 5 niveaux de décomposition).

4.1.4 Algorithme rapide pour la transformée en ondelettes

Mallat [34] a montré que les coefficients de la décomposition du signal sur une base orthonormée d'ondelettes se calculent par un algorithme rapide (algorithme pyramidal) qui cascade des convolutions discrète avec des filtres passe-bas (G) et passe haut (H) dont les sorties sont sous échantillonnées.

Dans notre cas, les coefficients de décomposition du signal par la transformée en ondelettes dyadique sont obtenus par filtrage successif passe-haut (H) et passe bas (G) de la sortie du filtre passe bas (G). Les sorties des filtres sont sous échantillonnées par un facteur de 2 l'algorithme est illustré à la figure 4.6.

Ces banc de filtres implémentent une transformée rapide en ondelettes orthogonales, qui ne nécessite que $O(N)$ calculs pour un signal de taille N.

Le symbole $'\downarrow 2'$ correspond au sous-échantillonnage par un facteur de 2. La figure montre qu'à chaque niveau de décomposition j, le signal est décomposé en coefficients d'approximation $a_j(k)$ (sortie du filtre passe bas) et en coefficients de détails $\omega_j(k)$ (la sortie du filtre passe haut (H)).

Les coefficients d'approximation correspondent à des moyennes locales du signal tandis que les coefficients de détails, aussi appelés coefficients d'ondelettes, dépeignent les différences entre deux moyennes locales successives, c'est-à-dire entre deux approximations successives du signal.

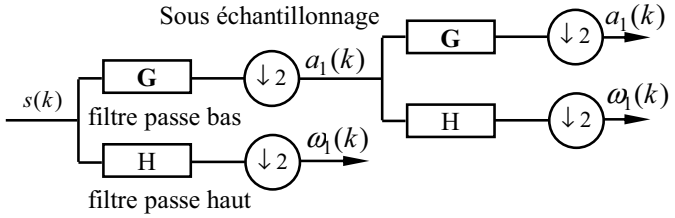


Figure 4.6 transformée en ondelettes Dyadique avec 2 niveaux de décomposition.

D'une manière plus imagée, les coefficients d'approximation donnent une représentation lissée du signal et les coefficients d'ondelettes (de détails) nous donnent les détails (le bruit) qui ont été supprimés lors du lissage. Il est tout à fait possible de reconstruire le signal de départ à partir de ces coefficients d'approximation et de détails.

4.2 Types d'ondelettes utilisées

Il existe un nombre très important de type d'ondelettes que l'on appelle aussi familles. Cette richesse dans le choix de la base d'ondelettes, c'est-à-dire le chois des fonctions analysantes, est aussi l'un des intérêts de la transformée en ondelettes. Parmi la multitude de familles d'ondelettes qui ont été proposées, nous pouvons citer, par exemple , les Coeflets, les Symlets, les ondelettes de Daubechies, les ondelettes bi-orthogonales, l'ondelettes de Haar ...etc.

Lors de notre étude, nous nous somme limités à trois familles d'ondelettes bien connues en traitement du signal : les ondelettes de Daubechies, les Symlets et les Coeflets. Ces ondelettes sont toutes admissibles, selon le théoréme1, car de moyenne nulle et à décroissance rapide.
De plus elles ont déjà été étudié en reconnaissance de la parole et ont donné de bons résultats [23], [26]. Enfin, elles ont toutes la propriété d'avoir un support minimum pour un nombre de moments nuls donné. Avant d'aller plus loin, définissons les deux caractéristiques que nous venons de citer : le nombre de moments nuls et la taille du support d'une ondelette. Ces deux caractéristiques importantes sont généralement prises en compte dans le choix d'une ondelette.

a) Les moments nuls

Le nombre de moments nuls d'une ondelette s'exprime de la manière suivante :

$$\int_{-\infty}^{+\infty} t^k \psi(t) dt = 0 \quad \text{pour} \quad 0 \le k \prec p \qquad (4.16)$$

Si une ondelette ψ vérifie cette équation alors on dit que l'ondelette ψ a p moments nuls. Cela signifie que ψ est orthogonale à tout polynôme de degré $p-1$. L'intérêt d'avoir p moments nuls est d'obtenir des coefficients d'ondelettes ω_j proches de 0 au échelles fines 2^j (lorsque 2^j tend vers 0). En effet, si $f(t)$ est localement de classe C^k alors $f(t)$ est localement bien 'approximé' par un polynôme de Taylor de degré k, et si $k \prec p$ alors les ondelettes seront orthogonales à ce polynôme. La transformée en ondelettes aura donc des valeur proches de 0.
A contrario, quand $f(t)$ ne pourra être approximé correctement que par des polynômes de degré supérieur à *p*, alors la transformée en ondelettes aura de fortes amplitudes. Cette propriété est très utile pour détecter les transitions brutales. En effet, les zones stationnaires d'un signal correspondront à de petits coefficients d'ondelettes, et les transitions brutales à de grands coefficients.

b) Taille du support

Si $f(t)$ a une singularité isolée en t_0, et si t_0 est dans le support de l'ondelette ψ_j, alors la transformée en ondelette aux fines échelles : lorsque l'échelle s tend vers 0, il y aura k ondelettes aura des coefficients d'ondelettes de fortes amplitude autour de t_0. Si l'ondelettes ψ a un support de taille k, alors à haute résolution, c'est-à-dire aux fines échelles : alors l'échelle s tend vers 0 il y aura k ondelettes ψ_j dont le support contiendra t_0. L'idée est de minimiser la taille du support de ψ dans le but de diminuer le nombre de coefficients d'ondelettes de grande amplitude. Cela permet ainsi de faire de la détection de singularités.

Ces deux caractéristiques ne sont pas indépendantes. En effet, la taille du support et le nombre de moments nuls d'une ondelettes orthogonale sont liées par le fait que si ψ a p moments nuls alors son support est au mois de taille $2p-1$. Lors du choix d'une ondelettes, on doit donc faire un compromis entre la taille du support et le nombre de moments nuls. Si $f(t)$ a peu de singularités isolées, et est très régulier entre ces singularités, il est plus approprié de choisir une ondelette ayant de nombreux moments nuls afin d'obtenir un grand nombre de coefficients d'ondelettes de petite amplitude. Lorsque la densité de singularités augmente, il vaut

mieux diminuer la taille du support, quitte à avoir moins de moments nuls. En effet, les ondelettes dont le support passe par une singularité donnent des coefficients de grande amplitude.

Pour le choix des ondelettes, il faut aussi noter qu'en utilisant la transformée en ondelettes discrète, nous nous restreignons à n'utiliser que des ondelettes à filtres. En effet, seules les ondelettes à filtres pouvant être utilisées avec la transformée discrète, alors que dans le cas continu n'importe quelle fonction d'intégrale nulle convient. Ainsi, les ondelettes utilisées sont définies directement par leurs filtres associés (filtre passe-bas et passe-haut). En fait l'ondelette n'est pas toujours directement accessible, c'est-à-dire qu'aucune formule analytique ne la définit, comme par exemple l'ondelette définie implicitement, en utilisant un algorithme déduit de l'algorithme de reconstruction de Mallat [29]. Les filtres correspondant aux ondelettes que nous utilisons ont été construits à l'aide du logiciel 'Matlab'. La figure 4.7 représente la réponse impulsionnelle des filtres associés à l'ondelette de Daubichies 'dB4'. Nous présentons maintenant plus en détails les trois familles d'ondelettes choisies.

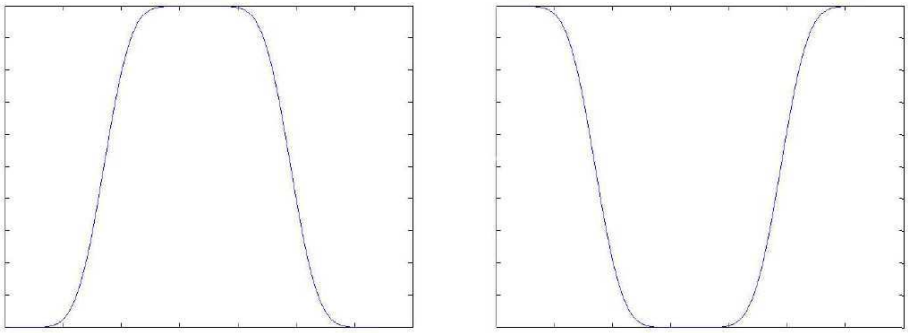


Figure 4.7 représentation en module dans le domaine des fréquences des effets des filtres d'analyse passe-haut (à gauche) et passe-bas (à droite) associé à l'ondelette 'db4'.

4.2.1 Les ondelettes de Daubechies

Cette famille d'ondelettes a été créée par Ingrid [49]. Nous noterons les ondelettes de cette famille *dbN* où *N* est l'ordre de l'ondelette. Nous retrouvons dans cette famille l'ondelette de Haar correspondant à *db1* et qui est la plus simple et certainement la plus ancienne des ondelettes. Excepté db1, les ondelettes de cette famille n'ont pas d'expression explicite. Cette famille possède certaines propriétés intéressantes. Le nombre de moments nuls de l'ondelette *dbN* et *N*. les ondelettes de Daubechies ont un support de taille minimale pour un nombre de moments nuls donné. Les ondelettes de Daubechies sont très asymétriques, en particulier pour les faibles valeurs

de *N*, sauf pour *db1*.

Figure 4.8 Exemple de Daubechies : de gauche à droite nous avons ‘db2’, ‘db4’ et ‘db8’

4.2.2 Les Symlets

Symlets, notées *symN*, ont été proposées par Daubechies en modifiant la coustruction des ondelettes *dbN* et constituent une famille d'ondelettes presque symétrique.

A part la symétrie, les propriétés de ces deux familles sont similaires. En regardant les figures des ondelettes de Daubechies et les Symlets, trous pouvons constater que la Symlet ressemble à une ondelette de Daubechies pour un nombre de moments nuls petit, et qu'elle est plus symétrique que sa consoeur.

Figure 4.9 Exemple de Symlets : de gauche à droite nous avons ‘sym2’, ‘sym4’ et ‘sym8’

4.2.3 Les Coiflet

Les Coitlets, comme les symlets, ont été construites par Daubechies. Elles ont été créées sur la demande de R. Coifman pour une application liée à l'analyse numérique. Nous prendrons comme notation de cette famille d'ondelettes: *coifN*.

Cette famille d'ondelettes est différente des deux précédentes, ici, l'ondelette *coifN* aura *2N* moments nuls. Toutefois, les Coiflets, comme nous pouvons le voir sur la figure 4.9, sont bien plus symétriques que les Symlets ou les ondelettes de Daubechies. L'intérêt principal des coiflets

réside dans le fait que si nous analysons une fonction *f* assez régulière, alors les coefficients d'approximation (pour un nombre de niveaux de décomposition assez grand) correspondent à l'échantillonnage de *f*.

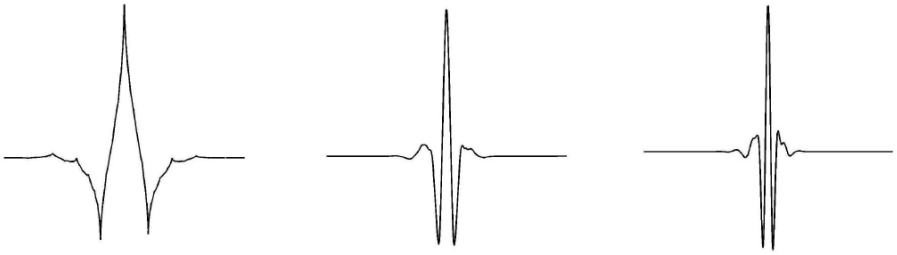


Figure 4.10 Exemple de Symlets : de gauche à droite nous avons 'coif2', 'coif3' et 'coif5'

4.3 Types d'énergies calculées sur les coefficients d'ondelettes

En reconnaissance de la parole l'énergie du signal est trop souvent utilisée en tant que paramètre et donne de plus de bons résultats. C'est pourquoi, nous avons décidé d'utiliser l'énergie, calculée sur les coefficients d'ondelettes obtenus à partir de la transformée en ondelettes dyadique, comme paramètre pour notre tache de reconnaissance. La nécessité d'utiliser L'énergie est aussi due au fait que les coefficients d'ondelettes sont trop nombreux dans chacune des bandes de fréquences (ou niveau de décomposition) pour être utilisé directement.

Nous avons choisi de présenter trois types d'énergies aux propriétés différentes.

Dans ce qui suit, ω_j^k dénote le coefficient d'ondelettes à la position temporelle k et à la bande de fréquence j. nous rappelons que les décomposition temporelles et en bande de fréquences suivent une échelle dyadique, c'est-à-dire que la résolution temporelle est divisée par deux alors que la résolution fréquentielle double à chaque niveau de décomposition. Le nombre de coefficients dans la bande j est noté N_j. Nous calculons finalement, à partir de l'ensemble des coefficients d'ondelettes ω_j^k pour la bande de fréquences j. Différents paramètres f_j pour cette bande de fréquences j en utilisant différents types d'énergie :

4.3.1 L'énergie instantanée

Ce type d'énergie, classiquement utilisé dans le domaine de la parole, nous donne la distribution de l'énergie dans chacune des bandes.

$$f_j = \log_{10}\left(\frac{1}{N_j}\sum_{k=0}^{N_J-1}\left(\omega_k^j\right)^2\right)$$

4.3.2 L'énergie de Teager

L'opérateur discret d'énergie de Teager (The discrete Teager Energy Operator ou *TEO*) introduit par Kaiser. Cet opérateur permet de calculer d'une façon simple l'énergie d'un signal et de pouvoir estimer son amplitude et sa fréquence instantanée (démodulation). Cet opérateur a été récemment utilisé en reconnaissance de la parole [11]. Il nous permis de suivre les modulation d'énergie et donne une meilleure représentation de l'information formantique du signal dans le vecteur paramètre, comparé aux MFCC [33]. Il permet aussi une réduction du bruit du signal en utilisant sa capacité de suivi de la modulation d'énergie.

$$f_j = \log_{10}\left(\frac{1}{N_j}\sum_{k=0}^{N_J-1}\left|\left(\omega_k^j\right) - \omega_{k-1}^j\omega_{k+1}^j\right|\right) \qquad (4.17)$$

4.3.3 L'énergie hiérarchique

Nous calculons ici des paramètres basés sur l'énergie mais avec une résolution temporelle hiérarchique. L'énergie hiérarchique correspond au calcul de l'énergie au centre de fenêtre d'analyse en prenant le même nombre de coefficients dans toutes les bandes :

$$f_j = \log_{10}\left(\frac{1}{N_j}\sum_{k=(N_j-N_J)/2}^{(N_j+N_J)/2}\left(\omega_k^j\right)^2\right) \qquad (4.18)$$

J correspond à la bande la plus basse.

Le choix de ce s'applique par le fait que les coefficients d'ondelettes ont une résolution temporelle plus fine dans les hautes fréquences. Ils recouvrent des intervalles de temps de plus en plus petits lorsque l'on monte en fréquence alors que lorsque l'on descend dans les basses fréquences les coefficients d'ondelettes vont recouvrir des zones temporelles de plus en plus grandes. Le nombre de coefficients donc différent d'une bande à l'autre, un grand nombre dans les hautes fréquences et un petit nombre dans les basses fréquences. La technique de résolution temporelle hiérarchique extrait les caractéristiques concentrées au centre de la fenêtre d'analyse, en prenant le même nombre de coefficients pour toutes les bandes.

Ce type d'énergie a été utilisé avec succès en reconnaissance automatique de la parole pour paramétriser le signal [26].

Chapitre 5

Nouveau paramètre acoustique pour la reconnaissance robuste

5.1 Introduction

La robustesse au bruit est un problème très difficile auquel sont confrontés les systèmes de reconnaissance de la parole dans les applications concrètes. Plusieurs techniques [36] ont été proposées pour améliorer les performances de la reconnaissance en présence de disparité entre les conditions d'apprentissage et celles de l'application. Ces techniques peuvent être classifiées en deux catégories : celles fondées sur le pré-traitement du signal de la parole (RASTA [53] ou amélioration de l'intelligibilité [37] par exemple) et les techniques de compensation. Dans ces dernières, des modèles acoustiques initiaux (généralement les modèles de parole propre) sont transformés pour représenter le nouvel environnement.

Dans notre approche nous considérons que les données de parole ont été enregistrées dans différentes conditions de bruit. Nous développons ensuite l'architecture de notre nouveau paramètre proposé (PNRF : Proposed Noise Robust Feature), qui est issu des études psycho-acoustique en relation avec la perception de l'oreille humaine, et qui dépend du module ainsi que de la phase du spectre du signal. Le principe de cette nouvelle paramétrisation acoustique est de proposer une phase préliminaire de pré-traitement dans le but d'améliorer le signal parole par un débruitage adaptatif pour une tache de reconnaissance, et cela par décomposition du signal parole en bandes critiques par paquet d'ondelettes perceptuelles suivi par un seuillage adaptatif proposé.

Ce chapitre est organisé de la façon suivante. Dans la prochaine section, nous commençons par développer l'architecture générale de notre nouveau paramètre acoustique. Et nous définissons les différentes étapes que nous utilisons dans ce développement. Dans la section 3, nous posons le cadre expérimental pour nos évaluations, nous décrivons la base de données vocale, la base de bruit noisex-92 et le système de reconnaissance markovien de référence. La quatrième section de ce chapitre est consacrée aux résultats et leur analyse.

5.2 Description de l'algorithme de paramétrisation

Les différentes phases de conception de notre nouveau paramètre proposé PNRF sont les suivantes (figure 5.1):

1. Segmentation du signal parole en fenêtre.
2. Décomposition fréquentielle du signal parole en bandes critiques par la transformation en paquet d'ondelettes perceptuelles TPWP.
3. Débruitage du signal parole par seuillage (seuillage doux et seuillage doux modifié) des coefficients d'ondelettes des différentes bandes critiques par sélection du seuil pénalisé.
4. Reconstruction du signal parole par transformation inverse par paquet d'ondelettes parceptuelles IPWP.
5. Calcul des coefficients MFPSCC.
6. Enfin, calcul des coefficients dynamiques (dérivé et accélération).

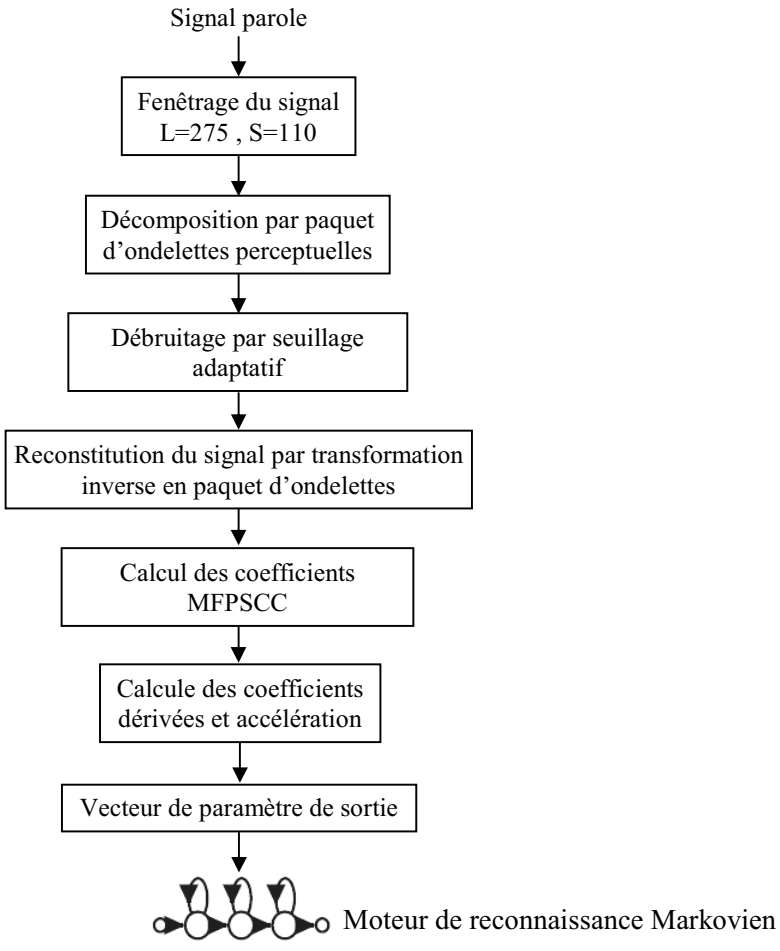


Figure 5.1. Bloc diagramme du paramètre robuste proposé

5.2.1 Segmentation en fenêtre

Dans la figure 5.1 du bloc diagramme de l'algorithme de paramétrisation robuste du signal parole. Le signal d'entrée est échantillonné à une fréquence $F_s = 11025Hz$. Une segmentation en trames est effectuée toutes les 10 ms pour permettre de découper le flot de parole continue en fenêtre de 25 ms soit de langueur $L = 275$ échantillons, dans lesquelles le signal est supposé quasi-stationnaire, le recouvrement entre deux fenêtre successives est de 10 ms ($L = 110$ échantillons). L'application d'une fenêtre classique (Hamming, Hanning,…..etc.) n'est pas nécessaire avant la phase de décomposition en d'ondelettes.

5.2.2 Décomposition du signal parole par paquet d'ondelettes

La décomposition du signal parole par paquet d'ondelettes a peu d'intérêt pour l'analyse du signal parole, parce que les coefficients d'ondelettes sont issus d'un banc de filtre de largeur de bande identique et à répartition linéaire. Cette répartition ne se rapproche pas de la répartition des fréquences au niveau de la membrane basilaire de l'oreille humaine. (Voir chapitre 2).

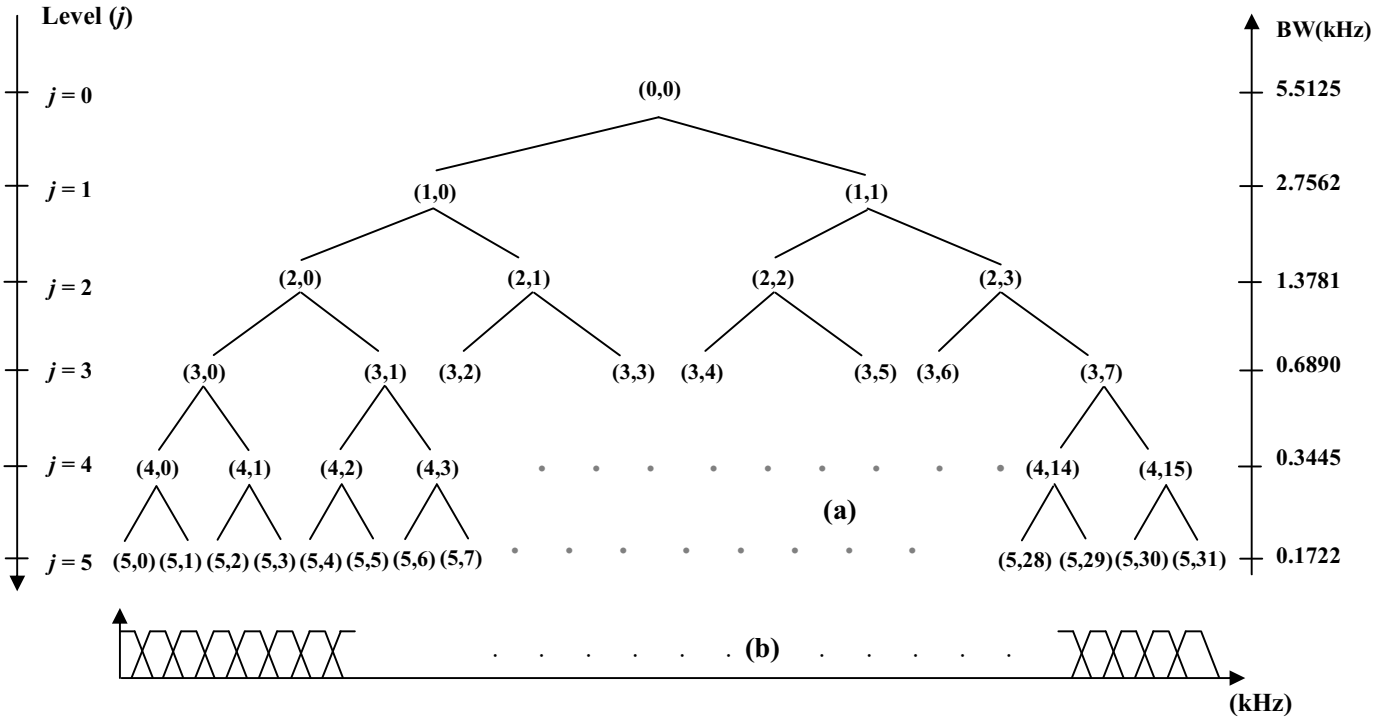


Figure 5.2 (a) structure de l'arbre WP. (b) correspondance de chaque bande (les bandes sont de largeurs identiques)

Index des Bandes fréquentielles	Centre de la bande (Hz)	Index des Bandes fréquentielles	Centre de la bande (Hz)
1	86	17	2838
2	258	18	3010
3	430	19	3182
4	602	20	3354
5	774	21	3526
6	946	22	3698
7	1118	23	3870
8	1290	24	4042
9	1462	25	4214
10	1634	26	4386
11	1806	27	4558
12	1978	28	4730
13	2150	29	4902
14	2322	30	5074
15	2494	31	5246
16	2666	32	5418

Tableau 5.1 : Description spectrale des sous bandes fréquentielles (toutes les bandes sont de largeurs identiques 172 Hz)

Pour cet effet nous avons proposé d'effectué une décomposition du signal parole en arbre perceptuel (PWPT : Perceptual Wavelet Packet Tree) comme il a été décrit dans ([19] par Pinter), (srinivasan et Jamieson dans [83]) et (carneno et drygajlo dans [82]), cette arbre est plus adaptée au système auditif, qui se comporte comme un banc de filtres passe-bande [43]. Les largeurs de bande de ces filtres, appelée bandes critiques se rapprochent d'échelles issues d'études sur la perception sonore (échelle Bark) et sur les bandes passantes critiques de l'oreille.

Une bande critique correspond à l'écart fréquentiel minimal pour que deux harmoniques d'un son soient discriminés perceptivement.

La décomposition en PWPT a été l'objet de plusieurs études récentes, elle est largement appliquée pour l'amélioration de la parole dans le milieu bruité [5], [6], [8], [15], [16], [35], ainsi que pour la reconnaissance robuste de la parole [1], [2], [7], [19], [20].

Les coefficients de la décomposition par PWP sont obtenus à partir de 17 sous bandes critiques, qui sont généralement considérés comme suffisantes pour les expériences de reconnaissance de la parole et du locuteur.

La décomposition par PWP a un intérêt majeur par rapport à la décomposition par WP, on peut citer quelques avantages :

- Les coefficients obtenus sont plus représentatifs et plus pertinents pour la tache de reconnaissance.
- Un nombre de nœud réduit par rapport à la décomposition en WP, ce qui nous rapporte un gain dans les calcules et dans le temps, et rend le système de reconnaissance plus souple ce qui permettra d'extraire les paramètres acoustiques en temps réel.
- Une représentation de l'information plus compacte, ce qui nous rapporte un gain considérable sur l'espace mémoire.

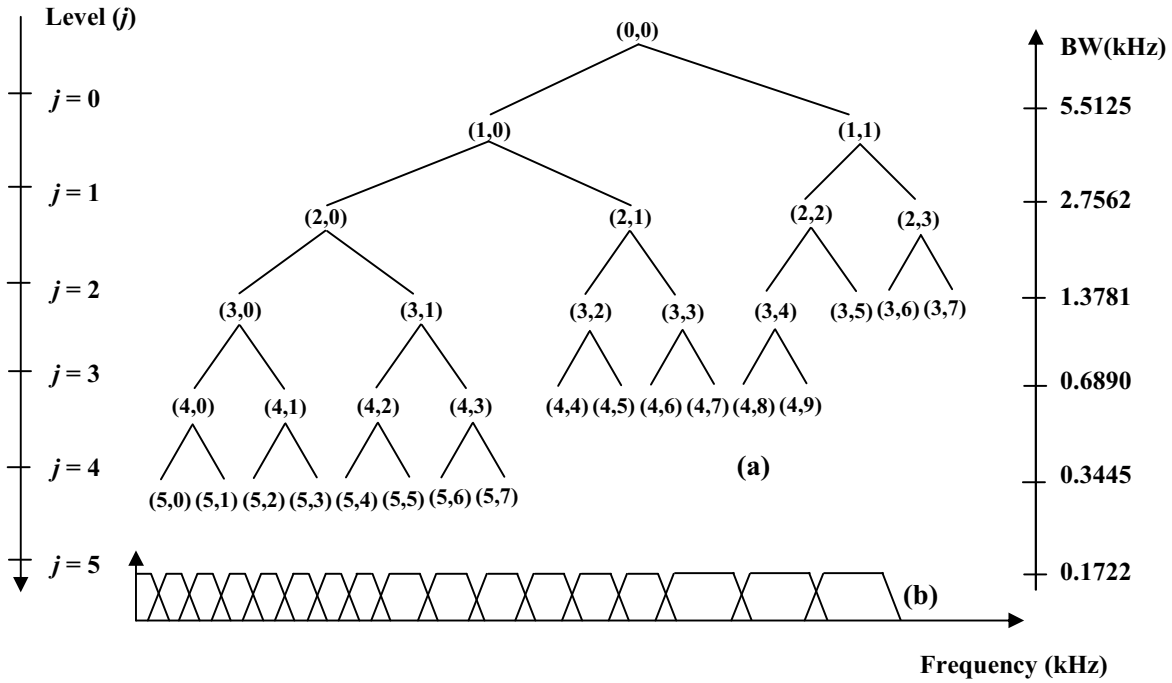


Figure 5.3 : (a) structure de l'arbre PWPT. (b) correspondance de chaque bande critique.

La figure 5.3 présente les détails de la décomposition en PWP, cinq niveaux de décomposition sont nécessaires pour la création de l'arbre, le dernier niveau de décomposition contient 17 sous bandes de largeur différentes.

Le tableau 5.2 montre que le nombre de coefficient dans une sous bande varie d'un niveau à un autre, il est plus élevé pour les niveaux ascendants qui sont moins énergétiques par rapport aux coefficients descendants.

Index des Bande critique	Largeur de la bande critique (Hz)	Centre de la bande (Hz)
1	172	86
2	172	258
3	172	430
4	172	602
5	172	774
6	172	946
7	172	1118
8	172	1290
9	344	1548
10	344	1892
11	344	2236
12	344	2580
13	344	2924
14	344	3268
15	689	3785
16	689	4473
17	689	5161

Tableau 5.2 : Description spectrale des bandes critiques.

Plusieurs études ont été faites pour déterminer quelles sont les ondelettes les plus adéquates pour une tâche de discrimination parole/bruit, c-à-d qui modélisent au mieux le signal parole en milieu précontraint. Ce choix ne peut être fait qu'empiriquement. Nous ne pouvons pas prédire, en regardant ses propriétés mathématiques, si une ondelette est meilleure pour telle ou telle tâche. Il existe de nombreuses familles d'ondelettes. Mais nous nous sommes limités aux ondelettes utilisables par l'algorithme rapide à base de bancs de filtres les ondelettes orthogonales Nous avons ainsi étudié trois familles d'ondelettes, les plus connues et les plus utilisées eu traitement du signal les ondelettes de Daubechies, les Symlets et les Coiflets que nous avons décrites au chapitre précédent.

Il a été montré dans plusieurs études récentes que les ondelettes de Daubechies sont les plus recommandés pour le débruitage du signal parole pour la RAP robuste [15], [12], [14] et [7], la détection de l'activité vocale en milieu bruité [3], [11] ainsi que pour l'amélioration du signal parole dans le milieu précontraint [5], [6], [35]. Apres plusieurs essais notre choix est porté sur l'ondelette Daubechies (Db8), quand vas le maintenir pour toutes les expériences.

5.2.3 Débruitage par les algorithmes de seuillage

Dans la littérature récente beaucoup de méthodes ont été développées dans le but de débruité les signaux contaminés par les bruits environnementaux [12], [16], [15], [13], [19]. Le débruitage par les ondelettes est réalisé par des algorithmes de seuillage, dont leurs coefficients sont inférieurs d'une certaine valeur spécifique, quand l'appelle seuil. Dans le domaine des ondelettes, ce terme signifie la rejection de bruit par un seuillage adéquat [84], [85].

Dans cette section nous allons présenter les techniques de seuillage les plus utilisées et nous introduisant une technique de seuillage plus adaptée pour traiter le signal parole.

5.2.3.1 Algorithme de seuillage dur (Hard thresholding)

Si un coefficient du signal de l'observation soit inférieur à un certain seuil, il est considéré comme étant du bruit pur et est remplacé par zéro, sinon il est gardé tel qu'il est. Il défini par : (voir figure 5.4. (b)).

$$\delta_{\lambda}^{H}(x) = \begin{cases} 0 & |x| \leq \lambda \\ x & |x| \succ \lambda \end{cases} \qquad (5.1)$$

λ : designe le seuil.

x : designe les coefficients d'ondelettes ($x \in w_{ij}$).

5.2.3.2 Algorithme de seuillage doux (Soft thresholding)

Si un coefficient du signal de l'observation soit inférieur à un certain seuil, il est considéré comme étant du bruit pur et est remplacé par zéro, sinon il est rétréci de la valeur du seuil (figure 5.4 (c))

$$\delta_{\lambda}^{S}(x) = \begin{cases} 0 & |x| \leq \lambda \\ sign(x)(|x| - \lambda) & |x| \succ \lambda \end{cases} \qquad (5.2)$$

5.2.3.3 Algorithme de seuillage doux modifié (Modified soft thresholding)

Chacun des algorithmes présentés ci-dessus a ses avantages et ses inconvénients. Le seuillage dur crée des discontinuités dans le signal de sortie, par contre le seuillage doux provoque l'apparition d'un biais qui est un inconvénient. Mais la technique de seuillage doux est plus optimale pour débruiter le signal parole corrompue par le bruit blanc gaussien additif.

Mais il y' a des considérations à prendre lors de l'application de la technique de seuillage (dur et doux) puisque les zones non voisées (consonnes) contiennent relativement beaucoup de plus

composantes de hautes fréquences qui peuvent être confondue avec le bruit et par conséquent éliminées durant la procédure de seuillage. Pour remédier a ces problèmes nous avons introduit l'algorithme de seuillage doux modifié, qui est défini comme suit :

$$y = \delta_{\lambda}^{Mst}(x) = \begin{cases} \theta x & |x| \prec \lambda \\ \mathrm{sgn}(x)(|x| + \lambda(\theta - 1)) & |x| \geq \lambda \end{cases} \qquad (5.3)$$

où $x \in w_{ij}$ et $y \in \overline{w}_{ij}$ qui représente la séquence des coefficients d'ondelettes résultantes.

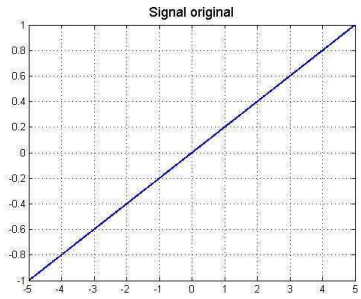
i désigne l'index de la sous bande de la WPD et j désigne le niveau de décomposition.

Le coefficient d'inclinaison θ introduit dans l'équation ci-dessus est définie par : (figure 5.4 (d))

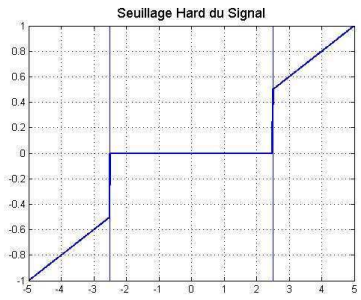
$$\theta = \beta \frac{\lambda}{\max(w_{ij})} \qquad (5.4)$$

β est la constante d'ajustement de l'inclinaison.

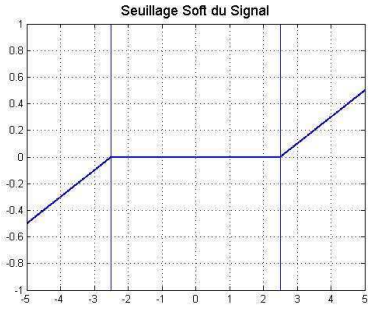
L'idée de base du seuillage doux modifié est l'introduction de coefficient d'inclinaison θ, pour ne pas forcer à zéro les coefficients dont leurs valeurs absolue est inférieurs au seuil λ. La technique de seuillage doux modifié est équivalente au seuillage doux lorsque $\beta = 0$. Dans notre cas $\beta = 0{,}5$.



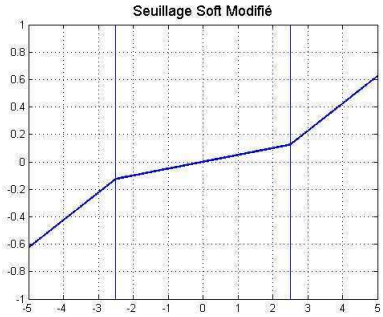
a) signal original



b) signal après un seuillage Hard (seuil=0.5)



c) signal après un seuillage Soft (seuil=0.5)



d) signal après un seuillage Soft Modifié (seuil=0.5)

Figure 5.4 Représentation graphique des différentes techniques de seuillage.

Une question naturelle se pose cependant : Comment est-elle choisie la valeur du seuil de décision ? C'est par la réponse à cette question que diffère une méthode d'une autre.

5.2.4 Sélection du seuil

Il y a de nombreuses formules pour obtenir la valeur du seuil, nous présentons dans le présent paragraphe deux méthodes qui sont les plus adapté au traitement du signal parole : la méthode du calcul du seuil universel et la méthode de seuil pénalisé.

5.2.4.1 Seuil obtenu par la méthode universelle

Donoho D. L. dans [84] a extrait un seuil optimal et général donnée par l'expression

$$\lambda = \sigma\sqrt{2\log_e(n)} \qquad (5.5)$$

n le nombre des échantillons dans une trame du signal parole, et σ l'écart type du bruit estimé par l'expression suivante

$$\sigma = \left| \frac{median(|w_{ij}|)}{0{,}6745} \right| \qquad (5.6)$$

où w_{ij} sont les coefficients de détail du 1er niveau de décomposition de la transformée en ondelettes du signal bruité.

5.2.4.2 Seuil obtenu par la méthode pénalisé

Soit la séquence des coefficients de paquet d'ondelettes $w_{j,i}$, où j représente le niveau de décomposition WPD et i est l'index des sous bande.

La variance est estimée de la même façon que dans [84]

$$\sigma = \frac{1}{\gamma_{mad}} median\left(\left|w_{1,1}\right|\right) \qquad (5.7)$$

$w_{1,1}$: est la séquence de WPC du premier nœud.

La constante $\gamma_{mad} = 0{,}6745$ estime la valeur médiane de la valeur absolue de l'écart type non biaisé de la distribution gaussienne.

nc : nombre de tous WPC du dernier niveau de décomposition.

cfs : contient tous les WPC du dernier niveau de décomposition $(W_{5,0}, W_{5,1} \ldots W_{3,7})$.

where $t = 1 \ldots ncd$

thres contient la valeur absolue des WPC sauvegardées dans l'ordre décroissant, *cd* contient le WPC du dernier niveau de la décomposition $(W_{5,1}, W_{5,2} \ldots W_{3,7})$ et *ncd* est le nombre de WPC dans *cd.*

$$A = cumsum(thres^2) \qquad (5.8)$$

cumsum : calcule la somme cumulative le long différente dimension.

$$valthr = index\_\min\left(2\sigma^2 t(\alpha + \log(nc/t)) - A\right) \qquad (5.9)$$

α : terme de pénalisation $(\alpha = 6.25)$

$$Maxthr = \max\left(\left|cfs\right|\right) \qquad (5.10)$$

$Valthr = \min(valthr, Maxthr)$ est la valeur du seuil. (5.11)

5.2.5 Comparaison entre les différents types de seuillage

Pour monter l'efficacité des techniques de seuillage, nous avons appliqué chacune d'entre elle sur un signal parole bruité à 10dB par un bruit blanc de SNR=5dB.

Pour le développement du graphique ci-dessous nous avons procédé de la manière suivante :

- Une décomposition du signal par paquet ondelettes perceptuelles PWP est effectuée.
- La décomposition est faite par les ondelettes de Daubechies (dB8).
- Les coefficients d'ondelettes sont issus de 17 bandes fréquentielles critiques.

- La restitution du signal est faite par la transformée inverse en paquet d'ondelettes ITWP.
- La fréquence d'échantillonnage est de 11025Hz.
- La fenêtre de traitement est de 25ms (275 échantillons) avec un pas de traitement de 10ms (110 échantillons).

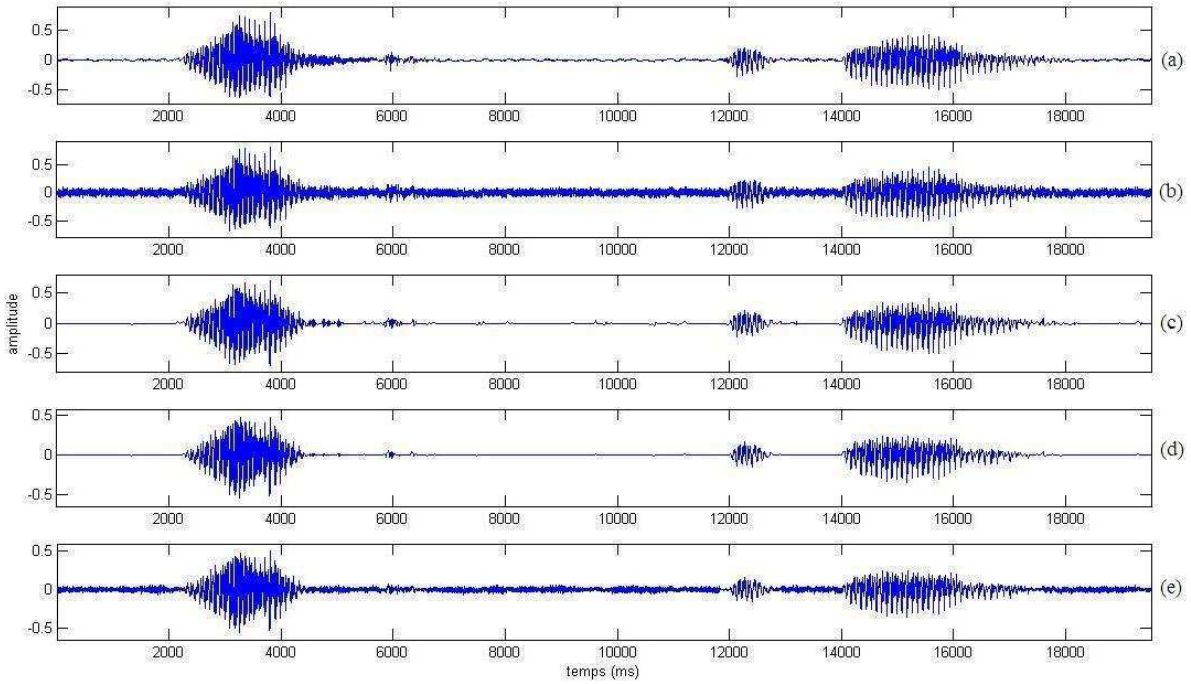


Figure 5.5 (a) signal original, (b) signal bruité par le bruit blanc à 5dB, (c) signal débruité par seuillage dur, (d) signal débruité par seuillage doux, (e) signal débruité par seuillage doux modifié. (dans l'ensemble de ces expérience nous avons utilisé le seuil pénalisé).

À première vue on constate que le seuillage dur et doux rejettent radicalement le bruit blanc, mais en réalité, ils éliminent aussi les composantes de hautes fréquences qui constituent la majeure partie du spectre des consonnes et qui sont l'essence même de la reconnaissance automatique de la parole. Par contre ils affectent moins les basses fréquences qui sont plus énergétiques que les hautes fréquences, et constituant les composantes principales des voyelles.

Le seuillage doux modifié (MST: modified soft thresholding) *ne rejettes pas les bruit mais les atténuent*, cette atténuation est due à l'introduction du coefficient d'inclinaison θ calculer par l'expression (5.4), ce type de seuillage est fait un bon compromis entre les composantes de hautes fréquences et les bruits.

5.2.6 Reconstitution du signal parole

En appliquant la transformée inverse par paquet d'ondelettes (IPWP : Inverse Perceptual Wavelet Packet Transform) nous obtenons le signal parole amélioré $n\tilde{x}(n)$

$$n\hat{x} = IPWPT\{\hat{w}_{j,i}\} \qquad (5.12)$$

$\hat{x}$: pour désigné le signal restitué.

$\hat{w}_{j,i}$: les coefficients d'ondelettes issus des sous-bandes critiques après seuillage.

5.2.7 Définition des coefficients de Mel cepstral du produit spectrale

A partir du signal restitué nous calculons les paramètres robustes proposés MFPSCC proposés par D. Zhu and K.K and Paliwal comme ils on décrit dans la réference [18].

Les coefficients Mfpscc ont été calculés à partir des quatre étapes suivantes :

1) Nous calculons le spectre du signal $\tilde{x}(n)$ et de $n\tilde{x}(n)$ par la FFT que l'on désigne respectivement par $X(k)$ et $Y(k)$.

2) Nous calculons ensuite le produit spectral donné par l'expression suivante:

$$Q(k) = \max\left(X_R(k)Y_R(k) + X_I(k)Y_I(k), \rho\right) \qquad (5.13)$$

où $\rho = 10^{\frac{\sigma}{10}} \cdot \max\left(X_R(k)Y_R(k) + X_I(k)Y_I(k)\right)$ (5.14)

σ est un seuil en dB (dans notre cas $\sigma = -60dB$).

3) Appliquer un banc de filter de Mel aux coefficients $Q(k)$ pour avoir les énergies issues des sous-bandes fréquentielles.

4) Enfin pour obtenir les coefficients Mfpscc, nous calculons la transformée en cosinus discrète.

Dans toutes nos expériences, nous ajoutons les coefficients dérivés du premier et du second ordre ainsi que le logarithme de l'énergie obtenu par chaque trame à tous les paramètres utilisés dans le reste de cette thèse.

$$E_j = \log_{10}\left(\frac{1}{N_j}\sum_{k=0}^{N_j-1}\left(\omega_k^j\right)^2\right) \qquad (5.15)$$

5.2.8 Coefficients différentiels

Les coefficients dérivés delta sont obtenus par la formule suivante :

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta (c_{\theta+1} - c_{\theta-1})}{2\sum_{\theta=1}^{\Theta} \theta^2} \qquad (5.16)$$

où d_t est le coefficient delta calculé à partir des coefficients $c_{t-\Theta}$ et $c_{t+\Theta}$. La même formule est utilisée pour le calcul des coefficients d'accélérations. Leurs utilisations améliorent les performances des systèmes markoviens de reconnaissance.

5.2.9 Comparaison graphique entre les différents types de paramètres acoustiques

Afin de déterminer quel est le paramètre acoustique le plus immunisé aux variations environnementales. Une comparaison graphique a été effectuée entre les différents paramètres soumis au bruit blanc de différentes intensités. Chaque vecteur de paramètre comporte 12 composantes fréquentielles. Les zones les plus sombres représentent les vecteurs de paramètres dont leurs coefficients sont moins énergétiques qui peuvent généralement être des segment de silence (au début et à la fin de chaque mot) ou des segments de signaux non voisés (consonnes occlusive, fricative …etc.).

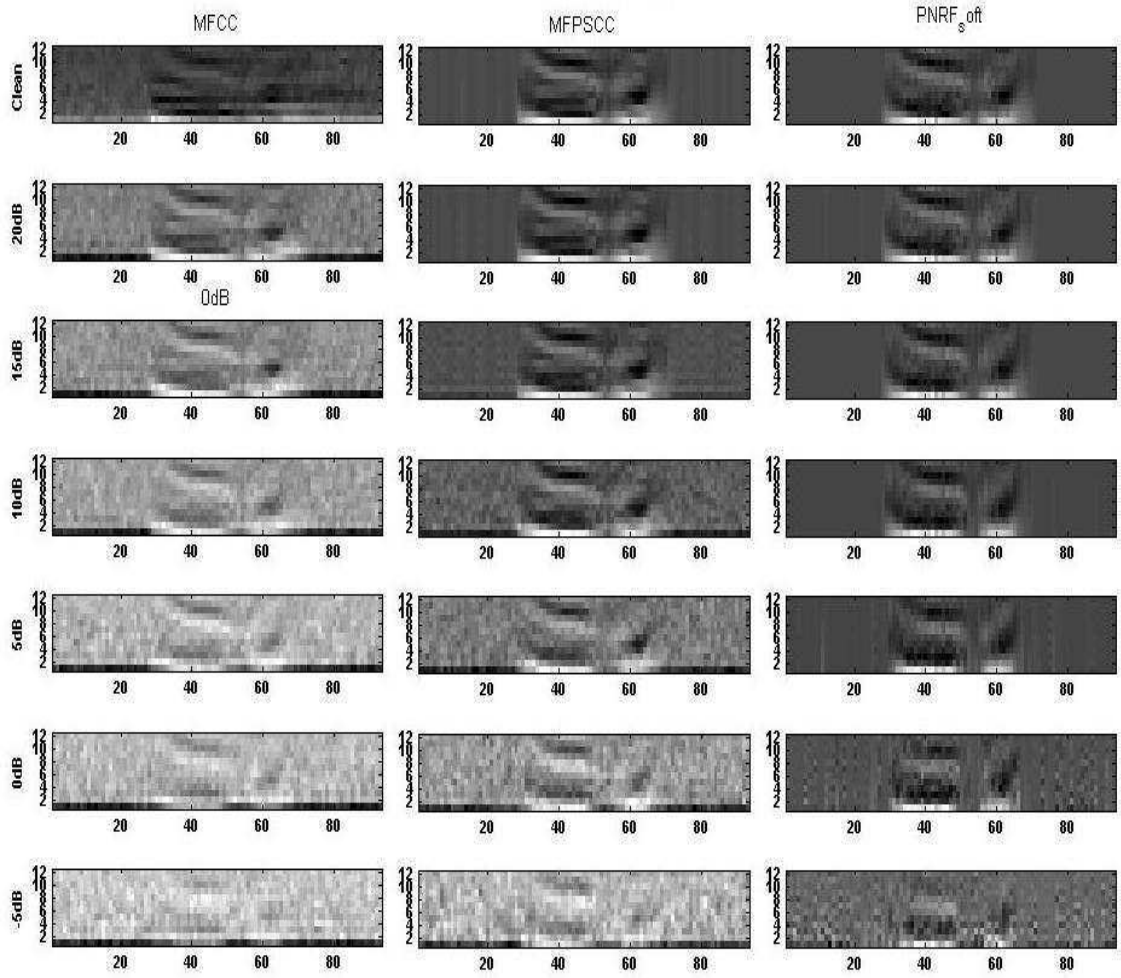


Figure 5.6 Représentation graphique des paramètres MFCC, MFPSCC et PNRF_Soft du mot un en arabe (واحد) corrompu par le bruit blanc sous 7 niveaux de SNR.

Comme le cas pour les coefficients MFCC, Mfpscc et PNRF_Soft sont calculés sur une fenêtre de 25 ms avec un pas de traitement de 10ms, les coefficients cepstraux sont obtenus a partir de la transformée en cosinus discrète du logarithme des énergies issues de 22 filtres répartis sur une échelle de Mel. A partir de la figure ci-dessus on peut constater que la dégradation des paramètres MFCC en présence de bruit blanc apparaît nettement pour des $SNR \leq 10dB$, alors que les coefficients MPSCC et PNRF_soft résistent mieux à des niveau de bruit plus élevés.

On peut voir clairement pour des $SNR \leq 5dB$ Les paramètres PNRF_soft sont les plus robustes et les plus adaptés a cette environnement par rapport aux paramètres MFPSCC.

5.3 Développement du système de reconnaissance

5.3.1 Description de la base de données

Toutes nos expériences ont été réalisées à l'aide d'une base de données vocale développée au niveau du laboratoire d'automatique et des signaux de Annaba (LASA), à l'université badji-Mokhtar, Annaba. Cette base a été acquise par un microphone mono-phonique relié à un ordinateur. La base contient 90 locuteurs, 46 locuteurs de sexe masculin et 44 autres de sexe féminins, qui appartiennent tous à la même tranche d'age, et dont la majorité sont de la même région (est d'Algérie). Chaque locuteur a prononcé 10 fois chaque chiffre arabe (0 à 9) d'une manière isolée (avec durée du silence importante entre deux locutions successives), cette base au totale contient 9 000 mots, les enregistrements ont été fait dans des condition moins bonnes (dans une salle fermée) . Le signal a été échantillonné avec une fréquence 11025 Hz et quantifié sur 16 bits.

Nous avons effectué un premier traitement qui consiste à enlever les bruits provoqués par le locuteur lors de la lecture, tel que les bruits d'inspirations et d'expirations de l'aire entre les élocutions, la toux et parfois des lapsus de prononciation.

Dans nos expériences, l'apprentissage du système de reconnaissance est fait dans des conditions non bruitées. Pour mieux évaluer nos paramètres acoustiques proposés plusieurs tests ont été faites lors de la phase d'évaluation du système et dans des conditions environnementales différentes de celle de l'apprentissage. Les signaux de parole obtenus sont corrompus par les bruits extraits du monde réel du corpus Noisex-92 développé par TNO. Quatre types de bruits ont été sélectionnés: le bruit blanc, le bruit rose, le bruit industriel (usinage de tôle) et le bruit du cockpit de l'avion de chasse F16.

Deux groupes de tests ont été envisagés, un groupe de test A, où les locuteurs ont contribués au deux phases, la phase d'apprentissage et la phase des tests. Nous avons pris de chaque 10 locutions de chaque chiffre prononcé (0 à 9) par chaque locuteur (90 locuteurs) 6 locutions pour servir à la phase de l'apprentissage et les 4 locutions restantes pour la phase des tests. Ce qui nous fait 5400 locutions pour l'apprentissage en clair et 3600 locutions restantes ont été utilisées pour l'évaluation du système de reconnaissance.

Un groupe de test B, dans ce groupe, les locuteurs qui ont servis à la phase d'apprentissage n'ont pas contribués à la phase des tests. L'apprentissage est fait avec les 10 locutions prononcées par

60 locuteurs (31 hommes et 29 femmes) ce qui ne donne un total de 6000 locutions. Les 30 locuteurs restant (15 hommes et 15 femmes) ont servis à la phase des tests avec un total de 3000 locutions.

5.3.2 Le corpus de bruits NOISEX-92

Nous allons présenter brièvement le corpus de bruit Noisex-92 avec lequel nous avons travaillé. Le but de ce corpus est de fournir un ensemble de bruits standard pouvant servir de base de comparaison pour les différentes méthodes de traitement et de reconnaissance de la parole dans le bruit.

Le corpus Noisex-92 a été conjointement mis au point, en 1992, à partir du corpus Noise-Rom-0 par l'Institut TNO pour l'étude de la perception et par l'équipe de recherche sur la parole de la 'Defense Research Agency' anglaise. Seuls certains bruits ont été sélectionnés par rapport à l'ensemble de ceux disponibles dans le corpus Noisex. En complément de ces bruits sont fournis des signaux de parole dans différentes conditions de bruits et, ce, pour tous les bruits du corpus : parole non bruitée et parole bruitée à des RSB de 18, 12, 6, 0 et -6 décibels.

Tous les fichiers des bruits de la base Noisex-92 [84] sont enregistrés sous format '.wav' avec une fréquence d'échantillonnage de 20 kHz et quantifié à 16 bits. La durée de chaque fichier est 255 secondes, obtenus à partir du corpus Noise-Rom-0. Le tableau 5.3 contient une brève description des différents types de bruit de la base Noisex-92.

Type de bruit	**Description**
White	bruit générer par générateur de bruit blanc analogique
Pink	bruit générer par générateur de bruit rose analogique
Babble	bruit de murmures de 100 personnes dans un restaurant
Volvo	bruit de voiture volovo340 à 120km/h en 4ème vitesse sur une route goudronnée
Factory1	bruit d'une usine de production de voitures : bruits de soudures électriques lors de l'assemblage du bas de caisse
Factory2	bruit d'une usine de production de voitures : bruits du hall d'assemblage
F16	bruit d'un chasseur F16 biplace à 500 noeuds et 300-600 pieds en place copilote
Destroyerops	Bruit de destroyer
Destroyerengine	Bruit de destroyer
hfchannel	bruit de canal radio hautes fréquences
Machinegun	bruit de mitrailleuse calibre 50mm
Buccaneer1	bruit de buccaner à 450 nœuds à 300 pieds
Buccaneer2	bruit de buccaner à 190 nœuds à 1000 pieds
M109	bruit de char de combat M 109 à 30 km/h
Leopard	bruit du Leopard 2 à 70 km/h

Tableau 5.3 Description des bruits de la base Noisex-92

5.3.3 Description du système de reconnaissance de référence à base des HMMc

Afin d'étudier l'apport du nouveau paramètre acoustique et pour montrer la pertinence des idées proposées, nous avons développé un système de référence à base de modèles de Markov cachés pour la reconnaissance des mots isolés.

Dans le système HMM de référence chaque mot est représenté par un HMM distinct. Dans l'étape d'apprentissage, chaque prononciation est convertie en une séquence de vecteur de paramètre acoustique (MFCC, MFPSCC, …etc.) qui constitue une séquence d'observation pour l'évaluation des paramètres HMM associés au mot respectif. L'évaluation est exécutée en optimisant la probabilité des données d'apprentissage correspondant à chaque mot dans le vocabulaire. Typiquement l'optimisation est exécutée en utilisant l'algorithme de Baum-Welch.

Dans l'étape de reconnaissance, la séquence d'observation représentant le mot à reconnaître est utilisée pour calculer les probabilités, pour tous les modèles possibles. Le mot reconnu correspond au modèle qui donne la probabilité la plus grande. Dans cette étape l'algorithme de Viterbi, est employé.

Les choix qui ont été fait pour le système de base sont les suivant :

- Nous avons modélisé chaque unité acoustique de notre vocabulaire par un modèle de Markov caché continu, 10 modèles sont nécessaires pour notre application (un modèle pour chaque chiffre arabe). Chaque modèle est représenté par 15 états, nous avons supposé que la majorité des chiffres arabes sont constitués de 5 phonèmes. Où chaque phonème est représenté par trois états, un état pour son début, un notre plus stable au milieu et le dernier pour sa fin.
- On a optée pour une topologie des modèles gauche droite (modèle de Bakis) proposé par Rabiner dans [42] pour sa bonne résolution des problèmes de reconnaissance des mots isolés.
- La Probabilité d'émission modélisée par une combinaison linéaire de 3 gaussiennes à matrice de covariance diagonale.

Tous les modèles ont la même topologie, et les probabilités d'émission de tous les états sont représentées par un nombre identique de gaussiennes. L'apprentissage et la reconnaissance des modèles isolés ont été réalisés avec les outils de la plate forme logicielle HTK.

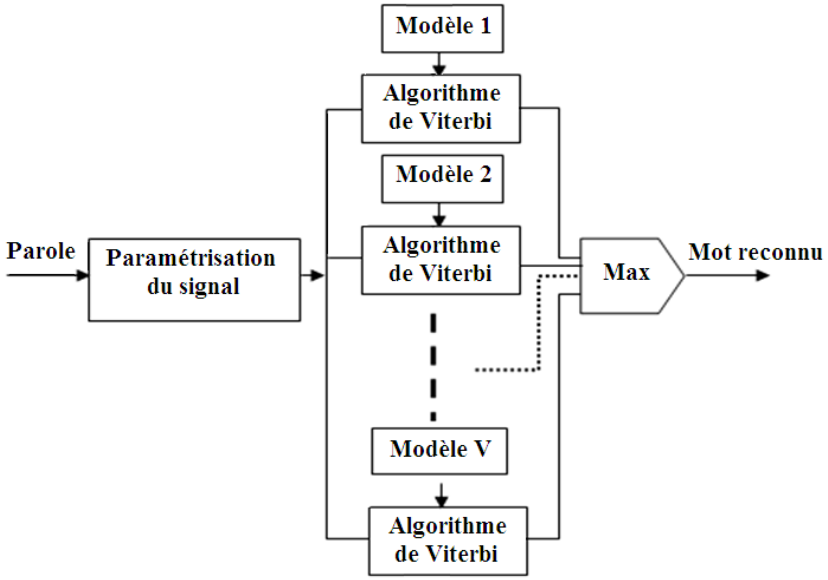


Figure 5.7 Structure du système de reconnaissance des mots isolé de référence

5.4 Expérimentation et résultats

Les tableaux ci-dessous montrent les taux de reconnaissance pour les séries d'expériences, nous rappelons que toutes nos expériences, l'apprentissage du système de reconnaissance est fait dans des conditions non bruitées..

5.4.1 Evaluation des performances du ASR en présence du bruit blanc pour les deux groupes de test (A et B)

SNR (dB)	MFCC	MFPSCC	PNRF_Mst	PNRF_Soft
Clean	98,55	**98,61**	97,78	97,08
20	97,55	**98,33**	97,72	96,50
15	96,03	**98,08**	97,50	95,94
10	90,78	96,44	**96,75**	95,03
5	76,69	92,47	**92,89**	92,69
0	48,04	75,85	80,11	**85,72**
-5	22,70	34,04	48,99	**65,05**
$\mathbf{V_{moyenne}}$	75,67	84,83	87,39	**89,71**

Tableau 5.3 Taux de reconnaissance (%) obtenus avec les différents paramètres en présence du bruit blanc (les locuteurs appartiennent au groupe de test A).

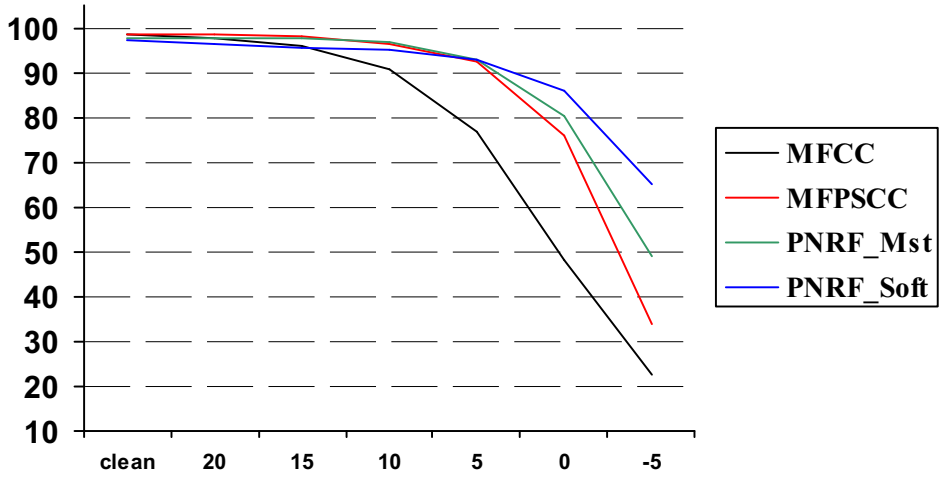


Figure 5.8 Représentation graphique des taux de reconnaissance (%) obtenus pour les différents paramètres en présence du bruit blanc (les locuteurs appartiennent au groupe de test A).

SNR (dB)	MFCC	MFPSCC	PNRF_Mst	PNRF_soft
Clean	**97,80**	97,60	97,00	96,27
20	96,77	**97,47**	96,87	95,67
15	95,03	**97,13**	96,67	95,07
10	88,93	**96,03**	95,47	93,73
5	74,49	92,13	**92,36**	91,00
0	43,91	77,33	80,83	**84,09**
-5	18,34	39,41	49,65	**63,05**
$V_{moyenne}$	73,61	85,30	86,97	**88,41**

Tableau 5.4 Taux de reconnaissance (%) obtenus avec les différents paramètres en présence du bruit blanc (les locuteurs appartiennent au groupe test B).

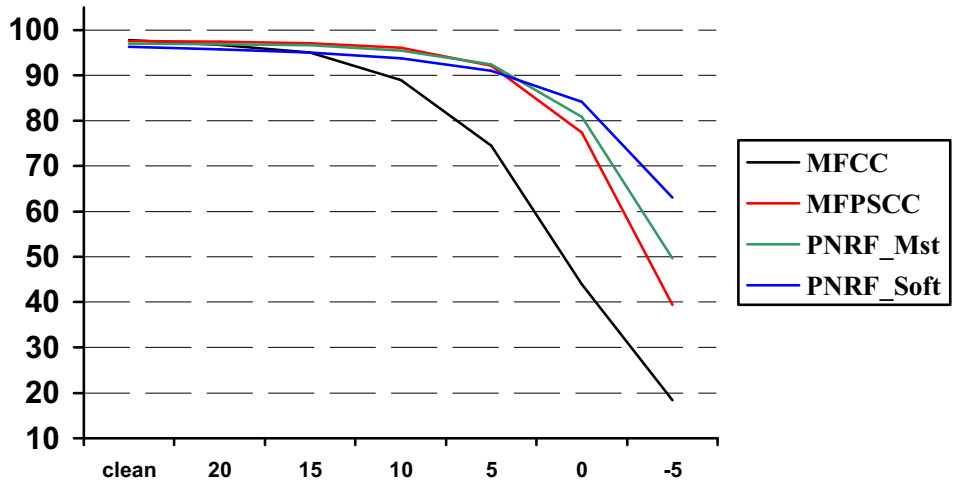


Figure 5.9 Représentation graphique des taux de reconnaissance (%) obtenu pour les différents paramètres en présence du bruit blanc (les locuteurs appartiennent au groupe de test B).

5.4.2 Discussion des résultats

A partir du tableau 5.3 et 5.4 on peut remarquer pour un bruit blanc de faible intensité SNR > 10 le paramètre Mfpscc a un apport meilleur sur le taux de reconnaissance, 0,6% par rapport au paramètre PNRF_soft et PNRF_Mst. Mais pour des niveaux de bruit élevés, pour des SNR ≤ 10db les meilleurs taux de reconnaissance sont obtenus avec nos paramètres proposés PNRF_Soft et PNRF_Mst, l'apport est de plus de 42% par rapport au paramètre Mfcc, et plus de 24% sur le taux de reconnaissance par rapport au paramètre Mfpscc. Le paramètre Mfcc se dégrade facilement en présence de bruit ce qui présente un handicap pour le système de reconnaissance qui opère dans ce genre de milieu.

On peut constaté aussi que le seuillage doux est mieux adapté que le seuillage doux modifié pour le traitement du signal parole corrompu par le bruit blanc et cela pour les deux séries de test effectués (avec les groupes de test A et B).

À partir des résultats obtenus on peut conclure, pour les 7 niveaux de SNR que le paramètre PNRF_soft est plus immunisé au bruit pour une tache de reconnaissance par les HMMc.

5.4.3 Evaluation des performances du ASR en présence du bruit rose pour les deux groupes de test (A et B)

SNR (dB)	MFCC	MFPSCC	PNRF_Mst	PNRF_soft
Clean	98,55	**98,61**	97,78	97,08
20	96,55	**89,60**	97,42	96,69
15	91,94	**97,75**	97,22	96,05
10	80,30	**96,33**	96,17	94,72
5	61,79	91,05	**92,14**	91,33
0	35,76	71,13	79,41	**81,24**
-5	16,00	42,01	49,37	**50,49**
$V_{moyenne}$	68,69	85,06	**87,07**	86,79

Tableau 5.5 Taux de reconnaissance (%) obtenus avec les différents paramètres en présence du bruit rose (les locuteurs appartiennent au groupe de test A).

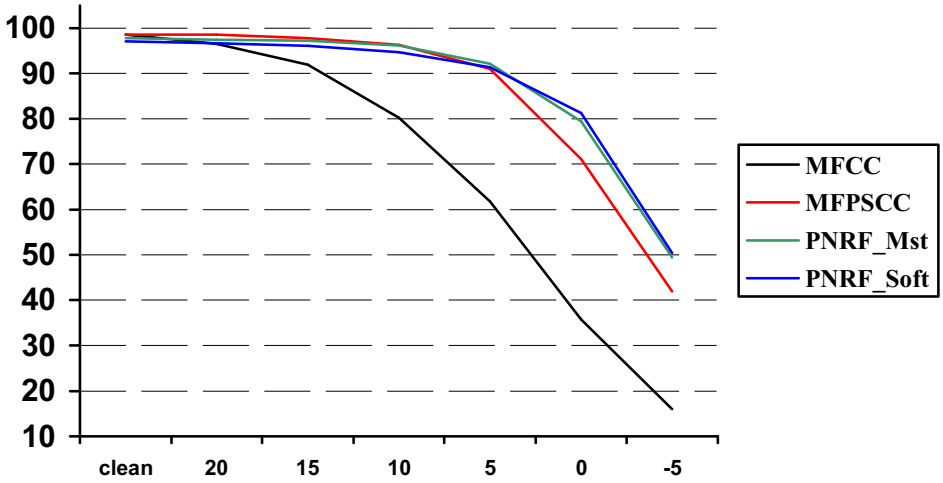


Figure 5.10 Représentation graphique des taux de reconnaissance (%) obtenu pour les différents paramètres en présence du bruit rose (les locuteurs appartiennent au groupe de test A).

SNR (dB)	MFCC	MFPSCC	PNRF_Mst	PNRF_soft
Clean	**97,80**	97,60	97,00	96,27
20	95,63	**97,59**	96,70	95,77
15	89,36	**97,07**	96,03	95,23
10	79,03	**95,50**	94,83	93,73
5	60,59	90,30	**90,43**	89,63
0	39,28	68,99	77,49	**78,09**
-5	22,44	39,48	46,28	**48,92**
$V_{moyenne}$	69,16	83,79	**85,53**	85,37

Tableau 5.6 Taux de reconnaissance (%) obtenus avec les différents paramètres en présence du bruit rose (les locuteurs appartiennent au groupe de test B).

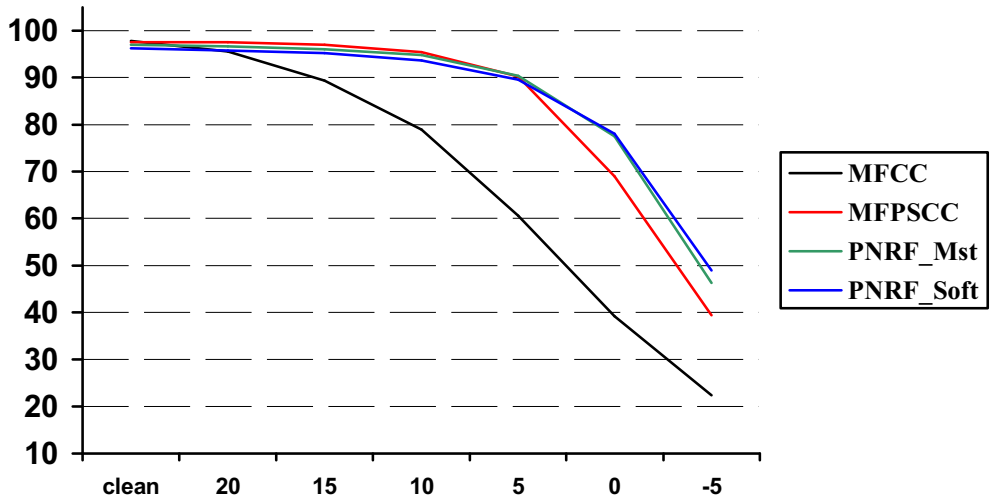


Figure 5.11 Représentation graphique des taux de reconnaissance (%) obtenus pour les différents paramètres en présence du bruit rose (les locuteurs appartiennent au groupe de test B).

5.4.4 Discussion des résultats

A partir des tableaux 5.5 et 5.6 pressentant les taux obtenus par le système de reconnaissance pour le signal parole corrompu par le bruit rose, on constate pour des SNR ≥ 10db le paramètre Mfpscc a un apport de 0,2 à 0,9% par rapport au paramètre PNRF_Mst et de 0,2 à 1,2% par

rapport au paramètre PNRF_Soft. Pour des SNR < 10db c'est-à-dire pour des bruits de niveau élevé, l'apport de paramètre PNRF_Mst est de 0,2 jusqu'à 7% par rapport au Mfpscc. avec le paramètre PNRF_soft l'apport sur le taux de reconnaissance est de 8 à 10% par rapport au Mfpscc et ce pour des SNR <5db. Les mêmes constatations peuvent être faites pour le paramètre Mfcc, où la dégradation est toujours importante même en présence de faible bruit.

En conclusion, le taux en valeur moyenne obtenu montre que paramètre PNRF_Mst offre plus de robustesse au système de reconnaissance par rapport aux autres paramètres. De plus on peut conclure que le seuillage doux modifié est mieux adapté pour le traitement du bruit rose.

5.4.5 Evaluation des performances du ASR en présence du bruit industriel pour les deux groupes de test (A et B)

SNR (dB)	MFCC	MFPSCC	PNRF_Mst	PNRF_Soft
Clean	98,55	**98,61**	97,78	97,08
20	95,11	**98,59**	97,22	96,64
15	88,77	**97,36**	96,92	95,75
10	75,44	**95,94**	95,42	93,61
5	57,57	**90,28**	90,08	89,08
0	35,59	71,69	**77,10**	74,91
-5	20,06	40,54	44,90	**45,23**
$\mathbf{V_{moyenne}}$	67,29	84,71	**85,63**	84,61

Tableau 5.7 Taux de reconnaissance (%) obtenus avec les différents paramètres en présence du bruit industriel (les locuteurs appartiennent au groupe de test A).

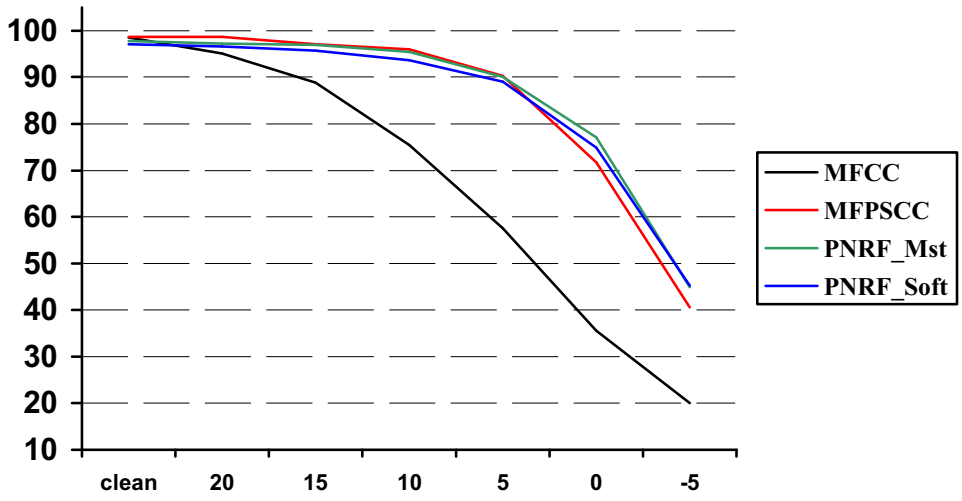


Figure 5.12 Représentation graphique des taux de reconnaissance (%) obtenus avec les différents paramètres en présence du bruit industriel (les locuteurs appartiennent au groupe de test A).

SNR (dB)	MFCC	MFPSCC	PNRF_Mst	PNRF_soft
Clean	**97,80**	97,60	97,00	96,27
20	93,93	**97,59**	96,37	95,33
15	87,16	**96,70**	95,47	94,43
10	73,12	**95,07**	93,70	92,23
5	54,52	**88,13**	87,86	86,76
0	35,88	69,66	**73,86**	73,02
-5	22,71	37,35	42,08	**42,61**
$V_{moyenne}$	66,44	83,15	**83,76**	82,95

Tableau 5.8 Taux de reconnaissance (%) obtenus avec les différents paramètres en présence du bruit industriel (les locuteurs appartiennent au groupe de test B).

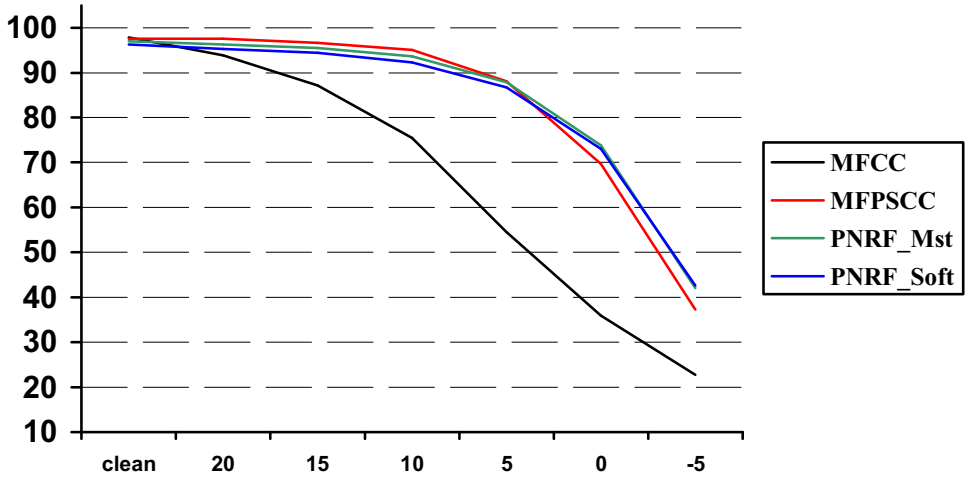


Figure 5.13 Représentation graphique des taux de reconnaissance (%) obtenus avec les différents paramètres en présence du bruit industriel (les locuteurs appartiennent au groupe de test B).

5.4.6 Discussion des résultats

A partir des tableaux 5.7 et 5.8 on peut constaté que nos paramètres proposés PNRF_Soft et PNRF_Mst ont un apport considérable de 4 à 6% par rapport aux paramètres Mfpscc sur le système de reconnaissance et ce pour un niveau de bruit industriel très élevé SNR < 5db, par contre le paramètre Mfpscc est meilleur pour des SNR ≥ 5db et l'apport varie de 0,6 à 1,2 par rapport a PNRF_Mst et de 1,4 à 3% pour le PNRF_Soft. Les mêmes constatations faites au dessus sont conservées pour le paramètre Mfcc. Pour ce type de bruit le seuillage doux modifié est mieux adapté que le seuillage doux pour la tache de reconnaissance par les HMMc.

5.4.7 Evaluation des performances du ASR en présence du bruit de cockpit de l'avion de chasse F16 pour les deux groupes de test (A et B)

SNR (dB)	MFCC	MFPSCC	PNRF_Mst	PNRF_Soft
Clean	98,55	**98,61**	97,78	97,08
20	94,28	**98,60**	97,19	96,47
15	85,94	**97,17**	96,80	95,61
10	72,55	**94,69**	94,64	93,22
5	54,29	85,79	**87,97**	87,44
0	34,04	63,02	**68,38**	67,88
-5	17,09	30,90	**37,09**	32,81
$\mathbf{V_{moyenne}}$	65,24	81,25	**82,83**	81,50

Tableau 5.9 Taux de reconnaissance (%) obtenus avec les différents paramètres en présence du bruit de cockpit de l'avion de chasse F16 (les locuteurs appartiennent au groupe A).

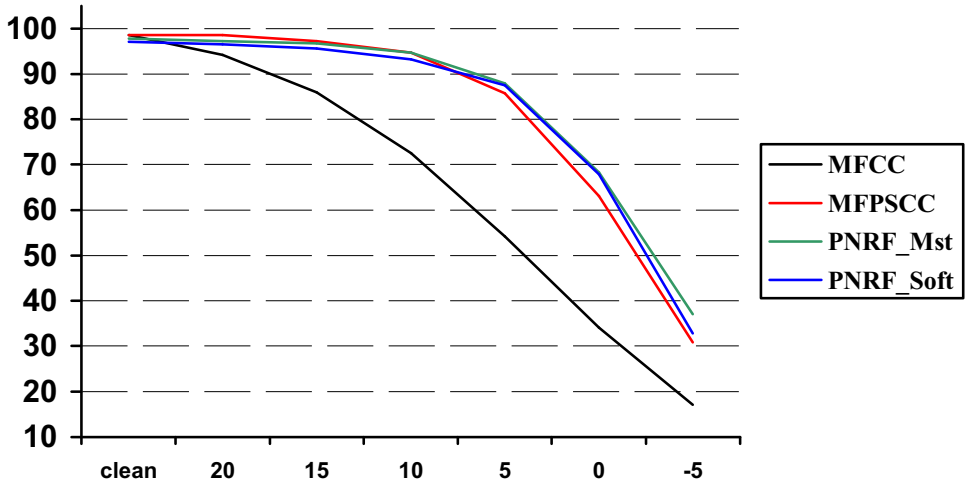


Figure 5.14 Représentation graphique des taux de reconnaissance (%) obtenus avec les différents paramètres en présence du bruit de cockpit de l'avion de chasse F16 (les locuteurs appartiennent au groupe de test A).

SNR (dB)	MFCC	MFPSCC	PNRF_Mst	PNRF_Soft
Clean	**97,80**	97,60	97,00	96,27
20	92,63	**97,59**	96,40	95,53
15	83,59	**95,93**	95,63	94,70
10	69,26	**93,30**	92,86	91,86
5	50,72	82,22	**86,26**	85,06
0	34,41	58,82	**66,32**	66,12
-5	19,87	29,68	**34,48**	31,78
$V_{moyenne}$	64,04	79,30	**81,26**	80,18

Tableau 5.10 Taux de reconnaissance (%) obtenus avec les différents paramètres en présence du bruit de cockpit de l'avion de chasse F16 (les locuteurs appartiennent au groupe de test B).

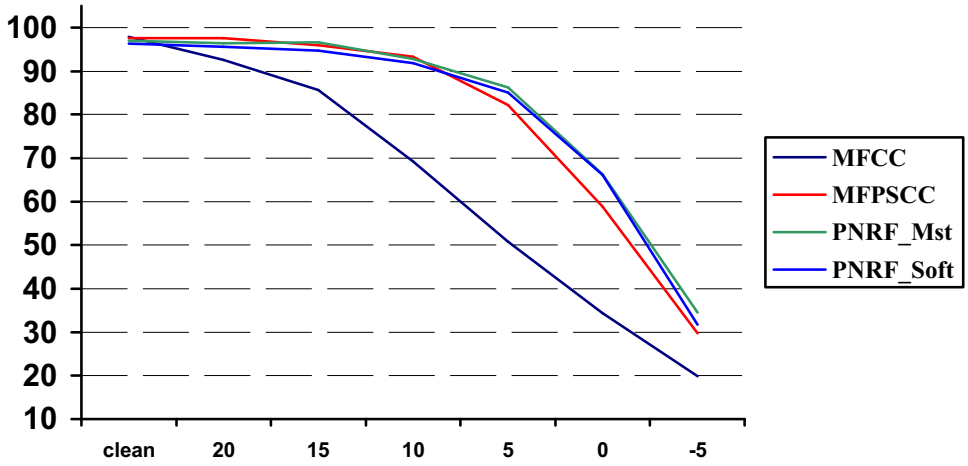


Figure 5.15 Représentation graphique des taux de reconnaissance (%) obtenus avec les différents paramètres en présence du bruit de cockpit de l'avion de chasse F16 (les locuteurs appartiennent au groupe de test B).

5.4.8 Discussion des résultats

A partir des tableaux 5.9 et 5.10 pressentant les taux obtenus par le système de reconnaissance pour le signal parole corrompu par le bruit de cockpit de l'avion de chasse F16, on peut constaté que le paramètres PNRF_Mst a un apport de 2 à 7% par rapport aux paramètres Mfpscc, et le paramètre PNRF_soft a un apport de 1,5 à 4% par rapport au Mfpscc sur le taux du système de reconnaissance et ce pour des SNR < 10db. Par contre le paramètre Mfpscc est meilleur pour des

SNR ≥ 10db et l'apport varie de 0,6 à 1,2 par rapport a PNRF_Mst et de 1,4 à 2,2% pour le PNRF_Soft. Les mêmes constatations pour le paramètre Mfcc. Pour ce type de bruit le seuillage doux modifié est mieux adapté que le seuillage doux.

Conclusion générale

Notre contribution consiste en la création d'un nouveau paramètre acoustique robuste et efficace, capable d'opérer dans des conditions environnementales aux caractéristiques acoustiques et sonores très différentes de l'environnement de l'apprentissage (en présence des différents bruits). Les majeurs contributions ont été fondue sur une décomposition du signal parole en paquet d'ondelettes perceptuel tout en respectant une repartions fréquentielles sur une échelle proche de la membrane basilaire, et par l'introduction la technique de seuillage doux modifié pour effectuer un débruitage tout en gardant l'ensemble des composantes spectrales. Le seuil utilisé est adaptatif, il a été calculé par un algorithme pénalisé.

Nos expériences ont été portées sur une base de donnés vocale contenant 9000 mots (chiffres arabes) prononcé par des hommes et des femmes ce qui est satisfaisant pour une tache de reconnaissance de mots isolés. Les différents résultats établis lors de cette thèse ont montré l'efficacité et l'apport important du paramètre acoustique proposé sur les performances du système de reconnaissance Markovien des mots isolés (chiffres arabes), et ce dans l'ensemble des environnements utilisés, mais plus particulièrement en présence de bruit blanc et du bruit rose qui sont des freins majeurs à l'emploi de la reconnaissance automatique de la parole.

Perspectives

Plusieurs tests ont montré que notre nouveau paramètre acoustique proposé permet une bonne modélisation acoustique du signal dans des conditions environnementales corrompues par différent type de bruit. Il convient dans un premiers temps de le tester sur d'autres bases de données vocales universelles (TIMIT, AURORA, TIDIGIT…etc.). Il convient également de le tester sur des moteurs de reconnaissance hybride (HMM/DTW, HMM/ANN, HMM/SVM) tel que les moteur développés au niveau du laboratoire LASA par les membres de notre équipe.

Dans le future, pour améliorer le système proposé de reconnaissance de la parole de mots isolés, plusieurs voies de recherche restent ouvertes. Plusieurs techniques peuvent être proposées : l'adaptation du moteur de reconnaissance aux nouvelles conditions environnementales, l'application des techniques d'adaptation au locuteur ou l'utilisation d'informations supplémentaires comme des informations visuelles sur la géométrie des lèvres.

Les techniques de traitement du signal appliquées au signal parole pour l'extraction des paramètres robustes ne suffisent pas pour rendre le système de reconnaissance insensible aux changements environnementaux (le type de microphone, l'écho de la salle, ou bien la distorsion de la transmission). Une adaptation du moteur de reconnaissance est nécessaire par l'utilisation des moteurs de reconnaissance hybride tel que les HMM/ANN, HMM/SVM….etc.

Pour généraliser l'application des systèmes de reconnaissance, il faut passer à la reconnaissance automatique de la parole continue, mais il y a plusieurs problèmes à résoudre.
En reconnaissance de parole continue, le décodage acoustique ne donne pas des résultats fiables à 100%. Nous avons donc besoin d'un modèle de langage. Le modèle de langage peut contenir une grammaire ou des modèles de langage stochastiques.
Pour cela, une étude de la syntaxe, ainsi que la construction d'une base de données de textes de la langue arabe sont nécessaires. Pour les données de textes, on peut envisager de les collecter à partir de plusieurs sources disponibles: les journaux, les livres ou bien les pages Web sur Internet.

Références bibliographiques

[1] M.C. Amara Korba, D. Messadeg, R. Djemili, H. Bourouba. "Robust Speech Recognition Using Perceptual Wavelet Denoising and Mel-frequency Product Spectrum Cepstral Coefficient Features", Informatica Journal, Vol. 32, No 3, pp. 283-288, 2008.

[2] N.Q. Trung; P.T. Nghia, "The perceptual wavelet feature for noise robust Vietnamese speech recognition", Communications and Electronics ICCE2008, Vol. 2, pp. 258-261, 2008.

[3] Shi-Huang Chen, Hsin-Te Wu, Yukon Chang, T. K. Truong, "Robust voice activity detection using perceptual wavelet-packet transform and Teager energy operator", Pattern Recognition Letters, vol. 28, pp. 1327-1332, 2007.

[4] M.T. Johnson, X. Yuan and Y. Ren, "Speech signal enhancement through adaptive wavelet thresholding", Speech Communication, Elsevier, Vol. 49, pp. 123-133, 2007.

[5] T. Haci, E. Ergun, "Speech Enhancement based on undecimated wavelet packet-perceptual filter-banks and MMSE-STSA estimation in various noise environments". Elsevier, Digital Signal Processing, 2007.

[6] A. Saeed, M. T. Manzuri, D. Roodhollah, "An improved wavelet-based speech enhancement by using speech signal features", Elsevier, Computer and Electrical Engineering Vol. 32, pp. 411-424. 2006.

[7] B. kotnik, Z. Kacic, "A noise robust feature extraction algorithm using joint wavelet packet subband decomposition and AR modelling of speech signals", Elsevier, Signal Processing, Vol. 87, pp. 1202-1223, 2006.

[8] Yu Shao, Chip-Hong Chang, "A versatile speech enhancement system based on perceptual wavelet denoising", ISCAS 2005, Vol. 2, pp. 864-867, 2005.

[9] Shao, Y. Chang, C.-H, "A versatile speech enhancement system based on perceptual wavelet denoising", IEEE international Symposium on circuits and systems, Vol. 2, pp. 864-867, 2005.

[10] C.C. lin, S.H. Chen, T.K. troung and Y. Chang, "Audio classification and categorization based on wavelets and support vector machine". IEEE transaction on speech and audio processing, vol.13, pages 644-651, 2005.

[11] D. Dimitriadis, P. Maragos and A. Potaminos, "Auditory teager energy cestrum coefficients for robust speech recognition". In European Conference On speech communication and Technology, pags 3013-3016, 2005.

[12] Y. Guermeur, A. Eliseef and D. Zelus, "A Comparative study of multi-class classifiers. Applied stochastic model in business and industry", Vol. 21, 2005.

[13] V. Wan & J. Carmichael, "Polynomial dynamic time Warping kernel support vector machines for dysarthric speech recognition with sparse training data". ln INTERSPEECH, 2005.

[14] V. Wan & S. Renais, "Speaker verification using sequence discriminant support vector machines". IEEE Transaction on Speech and Audio Processing, Vol. 13, 2005.

[15] M. Phothisonothai, P. Kumhom, and K. Chamnongthai, "Single-Channel Noise Reduction for Multiple Background Noises Using Perceptual Wavelet Packet Transform and Fuzzy Logic", Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 8, No. 6, pp. 613-620, 2004.

[16] S.H. Chen, J. Wang, "Speech Enhancement Using Perceptual Wavelet Packet Decomposition and Teager Energy Operator", Springer, The Journal of VLSI Signal Processing, Vol. 36, No. 2, pp. 125-139, 2004.

[17] M. Deviren, "Revising speech recognition systems: dynamic bayesian networks and new computational paradigms". Phd thesis, Université Henri poincaré, Nancy, france, 2004.

[18] Z. Donglai and K. K. Paliwal, "Product of power spectrum and group delay function for speech recognition", Proc. ICASSP, pp. 125-128, 2004.

[19] O. Farooq and S. Datta, "Wavelet-based Denoising for Robust Feature Extraction for Speech Recognition", electronics letters, Vol. 39, No 1, pp. 163-165, 2003.

[20] B. kotnik, Z. Kacic and B. Horvat, "The usage of wavelet packet transformation in automatic noisy speech recognition systems", Proceeding EROCON 2003, pp. 131-134, 2003.

[21] H.A. Murthy and V. Gadde, "The Modified Group Delay Function and Its Application to Phoneme Recognition", Proc. ICASSP, vol. 1, pp. 68-71, 2003.

[22] J. Ajmera, I. McCowan & H. Bourlard. "Speech/Music Discrimination using Entmpy and Dynamism Feaiures in a HMM Classification". Frameuiork; Speech Communication, vol. 40, pp. 351-363, 2003.

[23] M. Deviren and K.Daoudi, "Frequency filtering or wavelet filtering". In joint Intl. Conf. on Artificial Neural Networks and Neural Information Processing, ICANN/ICONIP, 2003.

[24] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals". IEEE transaction on speech and audio processing, Vol. 10, No. 5, pp. 293-302, 2002.

[25] I. J. Kim, S.I. Yang and Kwon, "Speech enhancement using adaptive wavelet shrinkage". In ISIE-2001, vol. 1, pp. 501-504, 2001.

[26] R. Gemello, D. Albesano, L. Moisa and R. De Mori, "Integration of fixed and Multiple resolution analysis in a speech recognition system". In ICASSP-01, 2001.

[27] O. Farooq and S. Datta, "Robust features for speech recognition based on admissible wavelet packets", Electronics letters, Vol. 37, No 5, pp. 1554-1556, 2001.

[28] N. Gowda and Z. Tufekci, "Mel-scaled discrete wavelet coefficients for speech recognition", Proc. Int. Conf. on acoustic, Speech and signal processing, Vol. 3, pp. 1351-1354, Istanbul, Turkey, 2000.

[29] S. Mallat. "Une exploration des signaux en ondelettes". Editions de l'Ecole polytechnique, 2000.

[30] 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, pp. 182-185, 2000.

[31] S. Saha, "Image compression from DCT to wavelets", ACM grossroads, Vol. 6, No. 3, pp. 644-651, 2000.

[32] A. Ganapathiraju & J. Picone, "Hybrid SVM/HMM Architectures for Speech Recognition". In Neural Information Processing Systems, 2000.

[33] F. Jabloun and A. Enis Cetin, "The teager energy based feature parameters for robust speech recognition in car noise", In ICASSP 99, 1999.

[34] S. Mallat, "A wavelet tour of signal of signal processing". Academic press, 1998.

[35] I. Pinter, "Perceptual wavelet-representation of speech signals and its application to speech enhancement", Computer speech & language, Vol. 10, No. 1, pp. 1-22, 1996.

[36] Yifan Gong, "Speech recognition in noisy environments: a survey". Speech communication, Vol. 16, pp. 261-291, 1995.

[37] Ephraim, "Gain-adapted Hidden Markov Models for Recognition of Clean and Noisy Speech". IEEE Trans. Signal Processing, Vol. 40, pp. 1303-1316, 1992.

[38] F. Jabloun and A. Enis Cetin, "The teager energy based feature parameters for robust speech recognition in car noise", In ICASSP 99, 1999.

[39] R. Vergin and D. O'Shaughnessy, "Generalized Mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous speech recognition", IEEE Trans. Speech, Audio Process., vol. 7, no. 5, pp. 525-532, Sept. 1999.

[40] S. Mallat, "A wavelet tour of signal of signal processing". Academic press, 1998.

[41] V.N. Vapnik, "Statistical learning theory", John Wiley ans Son, Inc N.Y. 1998.

[42] L.R Rabiner and B. juang, " A Tutorial on hidden Markovs Models and select application in speech recognition ", Proceedings of IEEE, Vol. 77, No. 2 , pp. 257-285, 1989.

[43] B. Moore, "An introduction to the psychology of hearing", Academic Press, 1997.

[44] J. S. Bridle, "Optimization and search in speech and language processing". Survey of the state of the art in human language technology, pp. 423-428, 1995.

[45] O. Cappé, J. Laroche et E. Moulines, "Regularized estimation of cepstrum envelope from discrete frequency points", IEEE ASSP Work-shop on application of signal processing to audio and acoustic (1995), p. 213-216. 1995.

[46] L. Rabiner, B. H. Juang, "Fundamentals of Speech Recognition". Prentice Hall Signal Processing Series, 1993.

[47] J.W. Picone, "Signal modeling techniques in speech recognition", Proc. IEEE, Vol. 81, No. 9, pp. 1215-1247, 1993.

[48] L. Rabiner and B.-H. Juang, "Fundamentals of speech recognition", Prentice-Hall, 1993.

[49] I. Daubechies, "Ten lectures on wavelets", Society for industrial and applied Mathematics, 1992.

[50] B. Yegnanarayana and H.A. Murthy, "Significance of Group Delay Functions in Spectrum Estimation", IEEE Trans. Signal Processing, Vol. 40, pp. 2281-2289, 1992.

[51] B. Boser, I. Guyon and V. Vapnik, "A training algorithm for optimal margin classifiers". In COLT'92, pp. 144-152, 1992.

[52] H. Hermansky, N. Morgan, A. Bayya et P. Kohn, "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)". Proceedings of the European Conference on Speech Communication and Technology, pp 1367-1370, 1991.

[53] H. Hermansky, N. Morgan, A. Bayya et P. Kohn. "RASTA-PLP speech analysis". Rapport technique TR-91-069, 6 pp, International Computer Science Institute, Berkeley (CA, États-Unis), 1991.

[54] J.-P. Haton, J.-M. Pierrel, G. Pérennou, J. Caelen et J.-L. Gauvain. "Reconnaissance automatique de la parole", 239 p, Collection AFCET - Dunod informatique, Dunod, 1991.

[55] T. Galas et X. Rodet, "Generalized functional approximation for source-filter system modeling", Proc. of Eurospeech 1991, pp. 1085-1088. 1991.

[56] H. Bourlard & C. Wellckcns. "Links between Markov models and multilayer perceptrons". In Trans. PAMI, Vol. 12, pp. 1167-1178, 1990.

[57] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech", Journal of Acoustical Society of America, Vol. 87, pp. 1738–1752, 1990.

[58] T. Galas et X. Rodet, "An improved cepstral method for deconvolution of source-filter systems with discret spectra : application to musical sound signals", In ICMC, 1990.

[59] K. Hornik, M.Stinchcombe and H. white. "Multilayer feedforward networks are universal approximators". Neural Networks, vol. 2, 1989.

[60] L.R. Rabiner, J.G. Wilpon & F.K. Soong, "High performance connected digit recognition using hidden Markov models", IEEE Trans. Acoust., Speech, Signal Processing, Vol. 37, No. 8, pp. 1214-1225, 1989.

[61] J.J. Hopfield. "Learning algorithms ans probability distributions in fee-forward networks". In Nat. Aead. Sei, pp. 8429-8433, 1987.

[62] Y. lecun, "Une procédure d'apprentissage pour réseaux à seuil asymétrique". In proc. Cognitiva, pp. 599-604, 1985.

[63] V.N. Vapnik, "Estimation of dependence based on empirical data", Springer-Verlag, N.Y, 1982.

[64] C. S. Myers et L. R. Rabiner, "Connected digit recognition using a level building DTW algorithm". IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 29, pp 351-363, 1981.

[65] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. Acoustics, Speech, Signal Processing, Vol. 28, No. 4, pp. 357-366, 1980.

[66] H. Sakoe, "Two level DP-matching a dynamic programming based pattern matching algorithm for connected word recognition". IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 27, pp. 588-595, 1979.

[67] J.-P. Zerling, "Articulation et coarticulation dans les groupes occlusive-voyelle en français". Thèse de doctorat de 3ème cycle, Université de Nancy 2, Nancy (France), 1979.

[68] L.R. Rabiner and R.W. Schafer, "Digital processing of speech signals", Prentice-Hall, 1978.

[69] H. Sakoe et S. Chiba, "Dynamic programming algorithms optimization for spoken word recognition". IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 26, No. 1, pp. 43-49, 1978.

[70] F. Itakura, "Minimum production residual principle applied to speech recognition". IEEE Transaction on Acoustics, Speech and Signal Processing, vol. 23, pp 67-72, 1975.

[71] J.M. Makhoul, "Linear prediction: a tutorial review", Proc. IEEE, Vol. 63, No. 4, pp. 561-579, 1975.

[72] A.V. Oppenheim and R.W. Schafer, "Digital Signal Processing", Englewood Cliffs, NJ: Prentice-Hall, 1975.

[73] A. V. Aho, J. E. Hopcroft et J. D. Ullman, "The design and analysis of computer algorithms, chapitre 7 : The fast Fourier transform and its application", pp. 251-276, 1974.

[74] H. Sakoe et S. Chiba, "A dynamic programming approach to continuous speech recognition", Proceedings of the 7th International Conference on Acoustics, article 20C-13, 6 pp, 1971.

[75] J.W. Cooley & J.W. Tukey, "An algorithm for machine calculation of complex Fourier series", Math.Comput, vol. 19, pp. 297-301, 1965.

[76] F. Rosenblatt, "Principles of neurodynamics", Spartan Books, 1962.

[77] R. E. Bellman, "On a routing problem", Quaterly Journal of Applied Mathematics, vol. 16, pp. 87-90, 1958.

[78] R. E. Bellman. "Dynamic Programming", Princeton University Press, 1957.

[79] W. Koenig, H.K. Duhn & L.Y. Lacy, "The sound spectrograph", J. Acoust. Soc. Am, Vol. 18, pp. 19- 49, 1946.

[80] D.Gabor, "Theory of communication", J. IEEE, Vol. 93, pp. 429-457, 1946.

[81] Calliope, "La parole et son traitement automatique", livre, Collection technique et scientifique des télécommunications, CNET - ENST, Masson.

[82] B. Carneno, A. Drygajlo, "Perceptual speech coding and enhancement using frame-synchronized fast wavelet-packet transform algorithms". IEEE Trans. Signal Process, Vol. 47, No 6, pp. 1622-1635, 1999.

[83] P. Srinivasan, L.H. Jamieson, "High quality audio compression using an adaptive wavelet decomposition and psychoacoustic modelling". IEEE Trans. Signal Process, Vol. 46, No 4, pp. 1085-1093, 1998.

[84] D. L. Donoho, "De-noising by Soft-thresholding", IEEE Trans. Inform Theory, Vol. 41, No. 3, pp. 613-627, 1995.

[85] D. L. Donoho, "Nonlinear Wavelet Methods for Recovering Signals, Images, and Densities from Indirect and Noisy Data", Proceedings of Symposia in Applies Mathematics, Vol. 47, pp. 173-205, 1993.

[86] A. Varga, H. Steeneken, M. Tomlinson, D. Jones, The NOISEX-92 study on the effect of additive noise on automatic speech recognition, Technical report, DRA Speech Research Unit, Malvern, England, 1992. Available from: http://spib.rice.edu/spib/select_noise