Sélection de paramètres acoustiques pertinents pour la reconnaissance de la parole

Abdenour Hacine-Gharbi

ÉCOLE DOCTORALE SCIENCES ET TECHNOLOGIES (ORLEANS)

FACULTÉ de TECHNOLOGIE (Sétif)

Laboratoire PRISME

THÈSE EN COTUTELLE INTERNATIONALE présentée par :

Abdenour Hacine-Gharbi

soutenue le : 09 décembre 2012

pour obtenir le grade de :

Docteur de l'Université d'Orléans
et de l'Université Ferhat Abbas-Sétif

Electronique

Sélection de paramètres acoustiques pertinents pour la reconnaissance de la parole

THÈSE dirigée par :

Rachid Harba	Professeur, Université d'Orléans (France)
Tayeb Mohamadi	Professeur, Université Ferhat Abbas, Sétif (Algérie)

RAPPORTEURS :

Olivier Alata	Professeur, Université Jean Monnet, Saint-Etienne (France)
Mohamed Debyeche	Professeur, Université USTHB, Alger (Algérie)

JURY:

Ameur Zegadi	Professeur, Université Ferhat Abbas, Sétif (Algérie)	Président du jury
Rachid Harba	Professeur, Université d'Orléans (France)	
Tayeb Mohamadi	Professeur, Université Ferhat Abbas, Sétif (Algérie)	
Olivier Alata	Professeur, Université Jean Monnet, Saint-Etienne (France)	
Mohamed Debyeche	Professeur, Université USTHB, Alger (Algérie)	
Philippe Ravier	Maître de Conférences (HDR), Université d'Orléans (France)	

Table des matières

INTRODUCTION GENERALE

La parole comme un moyen de dialogue homme-machine efficace, a donné naissance à plusieurs travaux de recherche dans le domaine de la Reconnaissance Automatique de la Parole (RAP). Un système de RAP est un système qui a la capacité de détecter à partir du signal vocal la parole et de l'analyser dans le but de transcrire ce signal en une chaîne de mots ou phonèmes représentant ce que la personne a prononcé. Ces propriétés offrent une grande variété d'applications comme l'aide aux handicapés, la messagerie vocale, les services vocaux dans les téléphones portables ("name dialing", numérotation automatique, reconnaissance de commandes vocales), la production de documents écrits par dictée, etc.

Cependant le signal de parole est l'un des signaux les plus complexes à caractériser ce qui rend difficile la tâche d'un système RAP. Cette complexité du signal de parole provient de la combinaison de plusieurs facteurs, la redondance du signal acoustique, la grande variabilité inter et intra-locuteur, les effets de la coarticulation en parole continue et les conditions d'enregistrement. Pour surmonter ces difficultés, de nombreuses méthodes et modèles mathématiques ont été développés, parmi lesquels on peut citer : la comparaison dynamique, les réseaux de neurones, les machines à vecteurs supports (Support Vector Machine SVM), les modèles de Markov stochastiques et en particulier les modèles de Markov cachés (Hidden Markov Models HMM).

Ces méthodes et modèles travaillent à partir d'une information extraite du signal de parole considérée comme pertinente. Cette extraction est effectuée par une analyse acoustique qui conduit à rassembler cette information sous le terme de vecteur de paramètres acoustiques dont la dimension et la nature sont déterminants pour accéder à de bonnes performances des systèmes de RAP. Les différents types de paramètres acoustiques couramment cités dans la littérature sont les coefficients : LPC, LSF, MSG, LPCC, LFCC, PLP, MFCC, etc. Généralement, les coefficients MFCC sont les paramètres acoustiques (caractéristiques) les plus utilisés dans les systèmes RAP [1].

Cependant, des travaux de recherche ont permis d'étudier l'amélioration des performances des systèmes de RAP en combinant les coefficients MFCC avec d'autres types de paramètres acoustiques tels que : LPCC, PLP, énergie, ondelettes [2] [3]. De plus d'autres travaux ont montré une amélioration de performances en intégrant les coefficients différentiels de premier ordre (appelés aussi delta Δ) et deuxième ordre (appelés delta-delta $\Delta\Delta$) issus des coefficients

MFCC initiaux (statiques) [4] [5]. Ces coefficients différentiels référés comme des paramètres dynamiques fournissent une information utile sur la trajectoire temporelle du signal de parole.

Cette démarche a cependant pour effet de doubler ou tripler la dimension des vecteurs acoustiques [6] [7]. On peut penser qu'accroître le nombre de paramètres pertinents pourrait améliorer la précision de la reconnaissance. Dans les faits, cette idée se heurte à un problème connu sous le terme de "malédiction de la dimensionnalité" [8]. En effet, l'augmentation du nombre de paramètres se fait au prix d'un accroissement exponentiel du nombre d'échantillons constituant la base de données utilisée pour l'apprentissage du système de reconnaissance. En conséquence, lorsque la base de données est de taille finie et figée, les performances viennent même à se détériorer avec l'accroissement du nombre de paramètres. De plus, cette augmentation exige une quantité de ressources importante qui n'est pas en adéquation avec celle disponible dans les systèmes de RAP embarqués sur les téléphones (notamment reconnaissance des mots isolés ou connectés) [7].

Il est donc nécessaire, si l'on veut concevoir un système acceptable en termes de précision, coût de calcul, et d'encombrement mémoire, de limiter le nombre de paramètres en sélectionnant les plus pertinents susceptibles de modéliser le mieux possible les données pour la tâche de reconnaissance. Il faut donc être capable, pour résoudre ce problème, de trouver les caractéristiques les plus informatives possibles pour la RAP tout en limitant leur nombre. Les travaux de recherche peuvent alors s'orienter selon deux axes :

- Un axe de recherche, dans un ensemble donné de paramètres, des paramètres les plus pertinents et de leur nombre optimal, par l'application d'un critère de sélection des paramètres les plus pertinents.
- Un axe de recherche d'une méthode de transformation d'un ensemble de paramètres acoustiques en un autre ensemble contenant un maximum d'information puisque l'algorithme développé dans le premier axe se charge ensuite de ne conserver qu'un nombre limité de paramètres acoustiques.

L'objectif de cette thèse est de proposer des solutions et améliorations de performance à certains problèmes de sélection des paramètres pertinents dans le cadre de la reconnaissance de la parole (premier point précédent). Les résultats obtenus permettent ainsi de travailler sur de nouvelles méthodes de transformation de paramètres (deuxième point précédent, cependant non traité dans cette thèse).

Sur le plan théorique, on s'est intéressé aux aspects théoriques de l'information mutuelle qui est un outil issu de la théorie de l'information. Cet outil permet de renseigner sur la pertinence qu'a une séquence de données à réaliser une tâche de classification spécifiée. L'idée est d'appliquer cet outil pour la sélection des paramètres pertinents du signal de parole.

Ainsi, notre première contribution consiste à proposer une nouvelle méthode de sélection de paramètres acoustiques pertinents fondée sur un développement exact de la redondance entre une caractéristique et les caractéristiques précédemment sélectionnées par un algorithme de recherche séquentielle ascendante (algorithme "greedy forward" dit glouton direct). Cette redondance se fonde sur l'estimation des densités de probabilités jointes dont le coût de calcul augmente avec la dimension. Ce problème est résolu par la troncature du développement théorique de cette redondance à des ordres acceptables. De plus, un critère d'arrêt de la procédure de sélection permet de fixer le nombre de caractéristiques sélectionnées en fonction de l'information mutuelle approximée à une itération j. Nous avons validé cette approche sur les paramètres MFCC extraits des données parole de la base Aurora2 utilisée dans la RAP. Cette contribution s'est traduite par un article de conférence [9].

Cependant l'estimation de l'information mutuelle est difficile puisque sa définition dépend des densités de probabilité des variables (paramètres) dans lesquelles le type de ces distributions est inconnu et leurs estimations sont effectuées sur un ensemble d'échantillons finis. Une approche pour l'estimation de ces distributions est fondée sur la méthode de l'histogramme. Cette méthode exige un bon choix du nombre de *bins* (cellules de l'histogramme). Ainsi, nous avons proposé une nouvelle formule de calcul du nombre de *bins* permettant de minimiser le biais de l'estimateur de l'entropie et de l'information mutuelle.

Ce nouvel estimateur a été validé sur des données simulées et des données de parole. Plus particulièrement cet estimateur a été appliqué dans la sélection des paramètres MFCC statiques et dynamiques les plus pertinents pour une tâche de reconnaissance des mots connectés de la base Aurora2. Ce travail a été concrétisé par la publication d'un article dans la revue « Pattern Recognition Letters » [10]. Ensuite cet estimateur a été appliqué dans la sélection des paramètres acoustiques de différents types (MFCC, PLP, LPCC).

Le manuscrit contient quatre chapitres :

- Le premier chapitre donne des généralités sur les systèmes de reconnaissance automatique de la parole et plus particulièrement ceux basés sur les modèles HMM. Les différentes étapes fonctionnelles de tels systèmes sont décrites comme : l'analyse acoustique, l'apprentissage des modèles HMM, le décodage acoustique.
- Dans le deuxième chapitre, les différentes méthodes de réduction de la dimensionnalité sont présentées. Les techniques de sélection des paramètres pertinents fondées sur le critère de maximisation de l'information mutuelle sont détaillées.
- Le chapitre 3 qui est le cœur de la thèse présente nos contributions théoriques sur la sélection des paramètres pertinents ainsi que sur l'estimation de l'information mutuelle. Des résultats de plusieurs expériences de simulations sont exposés à la fin du chapitre.

➢ Dans le chapitre 4, une application du nouvel estimateur de l'information mutuelle dans la sélection des paramètres MFCC statiques et dynamiques est présentée. Les performances des paramètres sélectionnés sont évaluées par la mesure de la précision d'un système de référence pour la reconnaissance des mots connectés de la base Aurora2. Ainsi une grande partie de ce chapitre est consacrée à la description du système de référence et la base de données parole Aurora2. Les résultats et leurs interprétations sont exposés à la fin du chapitre.

➢ Enfin une conclusion générale résume les différents travaux effectués ainsi que les perspectives qui permettront d'élargir et de poursuivre l'étude menée dans cette thèse.

CHAPITRE I

DESCRIPTION D'UN SYSTEME DE RECONNAISSANCE DE LA PAROLE

I.1 INTRODUCTION

La parole est l'un des moyens les plus directs d'échange de l'information, utilisés par l'homme. Cet avantage a donné naissance à plusieurs travaux de recherche dont l'objectif est la conception des systèmes permettant de reconnaître la séquence des mots parlés.

Un système de Reconnaissance Automatique de la Parole (RAP) est un système qui a la capacité de détecter la parole et de l'analyser dans le but de générer une chaîne de mots ou phonèmes représentant ce que la personne a prononcé. Cette analyse se fonde sur l'extraction des paramètres descriptifs de la parole. Cependant le signal parole ne contient pas seulement des informations sur le texte parlé mais aussi des informations sur le locuteur, la langue, les émotions dont leur extraction n'est pas l'objectif de la RAP. Cette thèse s'intéresse à une étape primordiale de la RAP, en l'occurrence la sélection des paramètres descriptifs pertinents pour la tâche de reconnaissance des mots parlés. Avant d'aborder cette étape, nous allons décrire dans ce chapitre les principes généraux et les problèmes de la RAP, ainsi que les différentes étapes constituant un tel système (le lecteur trouvera plus de détails dans [11] [12] [1] [13].

I.2 APPLICATION DE LA RECONNAISSANCE DE LA PAROLE

Les avantages que l'on attend de la reconnaissance de la parole sont multiples. Elle libère complètement l'usage de la vue et des mains (contrairement à l'écran et au clavier), et laisse l'utilisateur libre de ses mouvements. La vitesse de transmission des informations est supérieure, dans la RAP à celle que permet l'usage du clavier. Enfin tout le monde ou presque sait parler, alors que peu de gens sont à l'abri des fautes de frappe et d'orthographe.

Ces avantages sont à l'origine d'une grande variété d'applications comme :

- L'aide aux handicapés.
- La messagerie.
- L'avionique.
- La commande de machines ou de robots.
- Le contrôle de qualité et la saisie des données.

- L'accès à distance : téléphone, internet.
- La dictée vocale.

Toutes ces applications bénéficient de l'évolution technologique qui se traduit par l'apparition de composants intégrés spécialisés (en traitement du signal pour la programmation dynamique) et du développement des techniques et des méthodes algorithmiques de plus en plus performantes.

Enfin l'insertion d'un système RAP dans son environnement réel d'utilisation dépend de son contexte d'application et de ses conditions d'utilisation ce qui peut vite rendre le dispositif très complexe. Un système RAP peut être décrit selon quatre grands axes graduant cette complexité :

- La dépendance du locuteur (système optimisé pour un locuteur bien particulier) ou l'indépendance du locuteur (système pouvant reconnaître n'importe quel locuteur).
- Le mode d'élocution : mots isolés, mots connectés, mots-clés, parole continue lue ou parole continue spontanée.
- La complexité du langage autorisé : taille du vocabulaire et difficulté de la grammaire.
- La robustesse aux conditions d'enregistrement : systèmes nécessitant de la parole de bonne qualité ou fonctionnement en milieu bruité.

Tout système correspond à un compromis entre ces axes, choisi en fonction du but à atteindre. Les systèmes réalisés de la RAP sont conçus pour des applications spécifiques. Cela conduit à une restriction de l'univers du dialogue homme-machine. En effet, la conception de systèmes capables de comprendre la langue orale dans son intégralité est pour l'instant beaucoup trop complexe.

I.3 DIFFICULTES DE LA RECONNAISSANCE DE LA PAROLE

Le signal de parole est l'un des signaux les plus complexes à caractériser et analyser car sujet à une grande variabilité. Cette complexité est liée à la production du signal de parole, ainsi qu'à l'aspect technologique.

Le signal de parole varie non seulement avec les sons prononcés, mais également avec le locuteur, l'âge, les émotions, la santé, l'environnement.

De plus, la mesure du signal de parole est fortement influencée par la fonction de transfert du système de reconnaissance (les appareils d'acquisition et de transmission), ainsi que par le milieu ambiant.

Ainsi, l'obstacle majeur pour améliorer les performances d'un système RAP provient de la grande complexité du signal vocal due à la combinaison de plusieurs facteurs,

principalement la redondance du signal acoustique, la grande variabilité intra et inter-locuteurs, les effets de la coarticulation en parole continue, ainsi que les conditions d'enregistrement.

I.3.1 LA REDONDANCE

Le signal vocal présente un caractère redondant. Il contient plusieurs types d'information : les sons, la syntaxe et la sémantique de la phrase, l'identité du locuteur et son état émotionnel. Bien que cette redondance assure une certaine résistance du message au bruit, elle rend l'extraction des informations pertinentes pour la RAP plus délicate de part la multimodalité des sources d'information.

I.3.2 VARIABILITE

Le signal vocal de deux prononciations à contenu phonétique égal est différent pour un même locuteur (variabilité intralocuteur) ou pour des locuteurs différents (variabilité interlocuteur).

En effet, lorsque la même personne prononce deux fois le même énoncé, on constate des variations sensibles sur le signal vocal causées par :

- L'état physique, par exemple, la fatigue ou le rhume.
- Les conditions psychologiques, comme le stress.
- Les émotions du locuteur.
- Le rythme lié à la durée des phonèmes (façon dont s'exprime le locuteur) et l'amplitude (voix normale, voix chuchotée, voix criée).

Cependant la variabilité interlocuteur est a priori la plus importante. Elle s'explique par :

- Les différences physiologiques entre locuteurs.
- Les habitudes acquises en fonction du milieu social et géographique comme les accents régionaux.

Cette variabilité rend très difficile la définition d'invariants et complique la tâche de reconnaissance. Ainsi, il faut pouvoir séparer ce qui caractérise les phonèmes, de l'aspect particulier à chaque locuteur.

I.3.3 CONTINUITE ET COARTICULATION

La production d'un son est fortement influencée par le son qui le précède et qui le suit en raison de l'anticipation du geste articulatoire. La localisation correcte d'un segment de parole isolé de son contexte est parfois impossible. Évidement la reconnaissance des mots isolés bien séparés par un silence est plus facile que la reconnaissance des mots connectés. En effet, dans

ce dernier cas, non seulement la frontière entre mots n'est plus connue mais, de plus, les mots deviennent fortement articulés.

I.3.4 CONDITIONS D'ENREGISTREMENT

L'enregistrement du signal de parole dans de mauvaises conditions rend difficile l'extraction des informations pertinentes indispensables pour la reconnaissance des mots contenus dans ce signal. En effet, les perturbations apportées par le microphone (selon le type, la distance, l'orientation) et l'environnement (bruit, réverbération) compliquent beaucoup le problème de la reconnaissance.

Pour illustrer l'ensemble de ces difficultés, un système de RAP doit, en définitive, être capable de décider "qu'un [a] prononcé par un adulte masculin est plus proche d'un [a] prononcé par un enfant, dans un mot différent, dans un environnement différent et avec un autre microphone, que d'un [o] prononcé dans la même phrase par le même adulte masculin" [12].

I.4 APPROCHES DE LA RECONNAISSANCE DE LA PAROLE

Le principe général d'un système de RAP peut être décrit par la figure (I.1) :

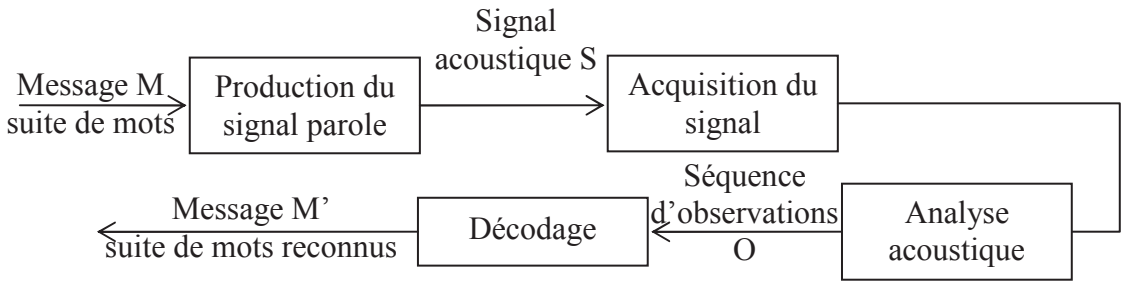


Figure.I.1 : Principe de la reconnaissance de la parole

La suite de mots prononcés M est convertie en un signal acoustique S par l'appareil phonatoire. Ensuite le signal acoustique est transformé en une séquence de vecteurs acoustiques ou d'observations O (chaque vecteur est un ensemble de paramètres acoustiques). Finalement le module de décodage consiste à associer à la séquence d'observations O une séquence de mots reconnus M'.

Un système RAP transcrit la séquence d'observations O en une séquence de mots M' en se basant sur le module d'analyse acoustique et celui de décodage.

Le problème de la RAP est généralement abordé selon deux approches que l'on peut opposer du point de vue de la démarche : l'approche globale et l'approche analytique [6]. La première considère un mot où une phrase en tant que forme globale à identifier en le comparant avec des références enregistrées. La deuxième, utilisée pour la parole continue,

cherche à analyser une phrase en une chaîne d'unités élémentaires en procédant à un décodage acoustico-phonétique exploité par des modules de niveau linguistique.

Cependant, les systèmes RAP en majorité utilisent des méthodes statistiques à base de modèles de Markov. Ces méthodes sont hybrides (globale et analytique).

I.4.1 APPROCHE GLOBALE

L'approche globale considère l'énoncé entier comme une seule unité indépendamment de la langue. Elle consiste ainsi à abstraire totalement les phénomènes linguistiques et ne retenir que l'aspect acoustique de la parole. Cette approche est destinée généralement pour la reconnaissance des mots isolés séparés par au moins 200 ms (voir figure I.2) ou enchaînés, appartenant à des vocabulaires réduits.

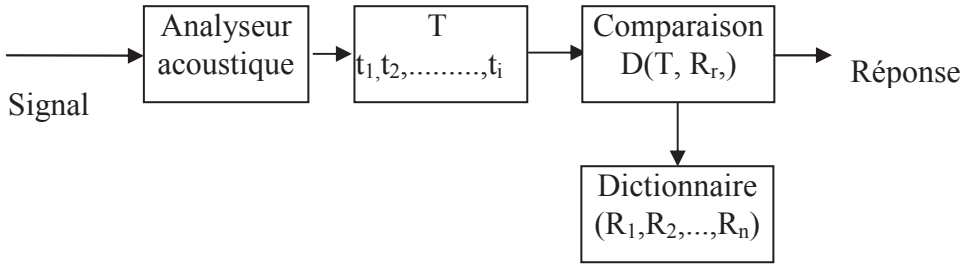


Figure.I.2 : Reconnaissance de mots isolés

Dans les systèmes de reconnaissance globale, une phase d'apprentissage est nécessaire, pendant laquelle l'utilisateur prononce la liste des mots du vocabulaire de son application. Pour chacun des mots prononcés, une analyse acoustique est effectuée permettant d'extraire les informations pertinentes sous forme de vecteurs de paramètres acoustiques. Le résultat est stocké ensuite en mémoire. Donc, les méthodes globales mettent en jeu une ou plusieurs images de références acoustiques (R_1,.....,R_n), a priori pour chaque mot.

Lors de la phase de reconnaissance, lorsque l'utilisateur prononce un mot T, la même analyse est effectuée : l'image acoustique du mot à reconnaître est alors comparée à toutes celles des mots de référence du vocabulaire au sens d'un indice de ressemblance D :

$$m = arg \max_{1 \le r \le N} [D(T, R_r)] \qquad (I.1)$$

Le mot ressemblant le plus au mot prononcé est alors reconnu.

Généralement, on rencontre deux problèmes : le premier est relatif à la durée d'un mot qui est variable d'une prononciation à l'autre, et le deuxième aux déformations qui ne sont pas linéaires en fonction du temps. Ces problèmes peuvent être résolus en appliquant un algorithme classique de la programmation dynamique appelé alignement temporel dynamique (*Dynamic Time Warping* DTW). Ce type d'algorithme permet sous certaines conditions d'obtenir une solution optimale à un problème de minimisation d'un certain critère d'erreur

sans devoir considérer toutes les solutions possibles. Dans le cas de la RAP, cet algorithme consiste à chercher le meilleur alignement temporel qui minimise la distance entre la représentation d'un mot de référence et la représentation d'un mot inconnu.

Dans le cas de grand vocabulaire ou de la parole naturelle continue, cette approche devient insuffisante et il est alors nécessaire d'adopter une nouvelle approche.

I.4.2 APPROCHE ANALYTIQUE

L'approche analytique cherche à trouver des solutions au problème de la reconnaissance de la parole continue ainsi qu'au problème du traitement de grands vocabulaires. Cette approche consiste à segmenter le signal vocal en constituants élémentaires (mot, phonème, biphone, triphone, syllabe), puis à identifier ces derniers, et enfin à reconstituer la phrase prononcée par étapes successives en exploitant des modules d'ordre linguistique (niveaux lexical, syntaxique ou sémantique). Ces constituants élémentaires peuvent être des phonèmes, des biphones, triphones ou des syllabes. Le processus de la reconnaissance de la parole dans une telle méthode peut être décomposé en deux opérations :

1. Représentation du message (signal vocal) sous la forme d'une suite de segments de parole, c'est la segmentation.
2. Interprétation des segments trouvés en termes d'unités phonétiques, c'est l'identification**.**

I.4.3 APPROCHE STATISTIQUE

L'approche statistique se fonde sur une formalisation statistique simple issue de la théorie de l'information permettant de décomposer le problème de la reconnaissance de la parole continue (figure I.3) [1].

Cette approche est construite sur le principe de fonctionnement des méthodes globales (avec phase d'apprentissage et de reconnaissance) mais avec l'exploitation des niveaux linguistiques. Ainsi une analyse acoustique est nécessaire pour convertir tout signal vocal en une suite de vecteurs acoustiques. Ces vecteurs sont considérés comme des observations dans la phase d'apprentissage des modèles statistiques et dans la phase de reconnaissance qui effectue une classification de chaque observation (par un index d'état dans le cas de la modélisation Markovienne).

On considère O une suite d'observations acoustiques résultant d'une analyse acoustique d'un signal de parole représentant une séquence de mots prononcés M. L'approche statistique consiste à chercher la séquence de mots $\bar{M}$ la plus probable parmi toutes les séquences de mots possibles E_M sachant les observations O. Ainsi, la séquence de mots optimale est celle qui maximise la probabilité a posteriori $P(M/O)$.

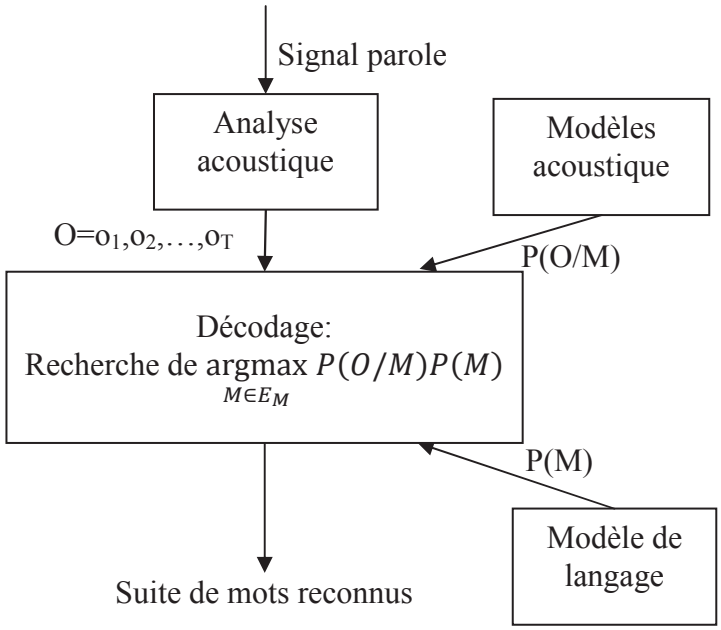


Figure.I.3 : Principe de la reconnaissance de formes bayésienne

$$\overline{M} = \text{argmax}_{M \in E_M} P(M/O) \qquad (\text{I. 2})$$

Selon la règle de Bayes, l'équation (I.2) peut être écrite comme suit :

$$\overline{M} = \underset{M \in E_M}{\text{argmax}} \frac{P(O/M)P(M)}{P(O)} \qquad (\text{I. 3})$$

Puisque P(O) ne dépend pas de M, alors l'équation (I.3) est équivalente à :

$$\overline{M} = \underset{M \in E_M}{\text{argmax}} \; P(O/M)P(M) \qquad (\text{I. 4})$$

où:

- $P(O/M)$ est la probabilité d'observer la séquence des vecteurs acoustiques O étant donnée la suite de mots M. Cette probabilité appelée vraisemblance est donnée par un **modèle acoustique** (figure I.3).
- $P(M)$, la probabilité *a priori* d'observer la séquence de mots M indépendamment de la séquence d'observations O, est estimée par un **modèle de langage** (figure I.3). Ce modèle exige des contraintes sur la syntaxe de la séquence des mots.

Les systèmes actuels de RAP continue exploitent la modélisation statistique du signal de parole par des Modèles de Markov Cachés HMM. La modélisation Markovienne tient compte non seulement de la non linéarité temporelle du signal de parole mais aussi de sa variabilité acoustique.

D'autres techniques ont été également développées pour la réalisation des systèmes RAP, parmi lesquelles, on peut citer : les réseaux de neurones, les machines à vecteurs supports (Support Vector Machine SVM), les réseaux bayésiens, les modèles hybrides obtenus par combinaison des réseaux de neurones et des modèles HMM [1].

Dans notre travail, nous nous sommes intéressés aux systèmes RAP à base des modèles HMM. Les différents éléments constituant ces systèmes sont décrits dans la section suivante.

I.5 SYSTEME DE RAP FONDE SUR LES MODELES HMM

L'hypothèse fondamentale des modèles HMM est que le signal vocal peut être caractérisé par un processus aléatoire paramétrique dont ses paramètres peuvent être déterminés avec précision par une méthode bien définie. La méthode HMM fournit une manière de reconnaître la parole, naturelle et très fiable pour une large gamme d'applications et intègre facilement les niveaux lexical et syntaxique [13].

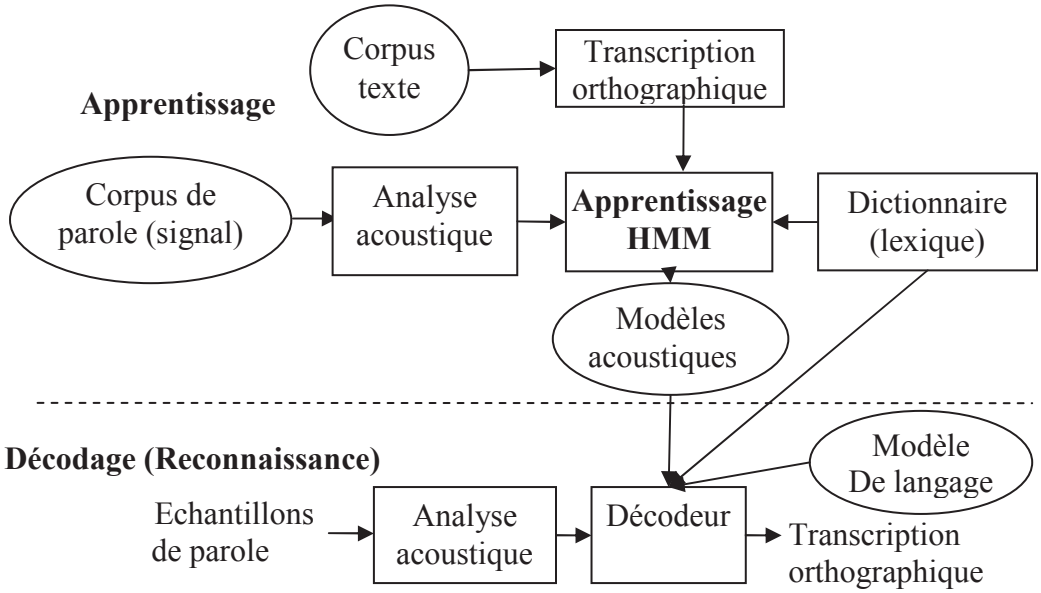


Figure.I.4 : Synoptique du système de reconnaissance de la parole incluant la procédure d'apprentissage et le décodage

Les différentes étapes d'un système de reconnaissance de la parole fondé sur les HMM sont représentées sur la figure (I.4). La ligne pointillée marque une séparation entre le processus d'apprentissage et le processus de reconnaissance. Les principaux composants utilisés pour le développement d'un tel système de reconnaissance sont les principales sources de connaissances (corpus de parole, corpus de texte, et lexiques de prononciations), le dispositif de paramétrisation acoustique (analyse acoustique), les modèles acoustiques et de langage dont les paramètres sont estimés durant la phase d'apprentissage, et le décodeur qui utilise ces modèles pour reconnaître la séquence de mots prononcés.

Les modèles acoustiques représentent les éléments à reconnaître : mots, ou unités phonétiques. Ces modèles sont usuellement développés à partir de grands corpus de données acoustiques et de textes. Ainsi, l'entraînement de ces modèles exige une définition des unités lexicales de base utilisées et un dictionnaire de prononciation décrivant la liste des mots qui pourront être reconnus.

Le modèle de langage fournit les informations syntaxiques pour la reconnaissance de la séquence de mots la plus probable.

Au centre de ce synoptique se trouve l'apprentissage par HMM qui est l'une des approches les plus utilisées dans les systèmes de RAP.

Lors de la reconnaissance, après l'analyse acoustique, un décodage est effectué et le système de reconnaissance fournit en sortie la séquence de mots la plus probable étant donné le modèle de langage et les modèles HMM.

Dans notre travail, nous avons proposé le système de reconnaissance des chiffres présenté dans [14] pour comparer les performances des paramètres acoustiques utilisés originalement dans ce système avec celles des paramètres acoustiques fournis par un algorithme de sélection. Ce système de référence est implémenté sous plate-forme HTK (Hidden Markov Model Toolkit, ou "boîte à outils de modèles de Markov cachés") [15] et évalué sur la base de données Aurora2 distribuée par ELRA [16]. Cette base conçue pour évaluer les performances des systèmes RAP dans différentes conditions de bruit, est utilisée pour sélectionner des paramètres acoustiques pertinents dans deux environnements: bruité et non bruité. La description détaillée de ce système de référence ainsi que ses performances sur la base Aurora2 seront abordés dans le chapitre IV.

Dans les sous-sections suivantes, nous détaillons brièvement l'étape d'analyse acoustique ainsi que les modèles HMM et leur mise en œuvre dans les systèmes RAP en utilisant la boîte à outils HTK. Au cours de ces étapes on introduit les grandes lignes de nos contributions, ainsi que le schéma synoptique des différentes étapes de notre système de reconnaissance.

I.5.1 ANALYSE ACOUSTIQUE

Le signal vocal transporte plusieurs informations comme le message linguistique, l'identité du locuteur, ainsi que ses émotions, la langue adoptée, etc. Un système RAP consiste à récupérer seulement le message linguistique indépendamment des autres informations. Dans un tel système, l'analyse acoustique consiste à extraire du signal vocal un ensemble de paramètres pertinents dans le but de réduire la redondance du signal vocal pour une tâche de reconnaissance de la parole. Le nombre de ces paramètres doit rester raisonnable, afin d'éviter la nécessité d'un grand espace mémoire ce qui accroît le coût de calcul dans le module de décodage. Ces paramètres doivent être discriminants en rendant les sons de base facilement séparables. Ils doivent être robustes au bruit.

Le calcul des paramètres acoustiques est réalisé par une chaîne de prétraitement selon les étapes suivantes (figure I.5) :

1. Filtrage et échantillonnage : le signal vocal est filtré puis échantillonné à une fréquence donnée. Cette fréquence est typiquement de 8 kHz pour la parole de qualité téléphonique et de 16 à 20 kHz pour la parole de bonne qualité [6].

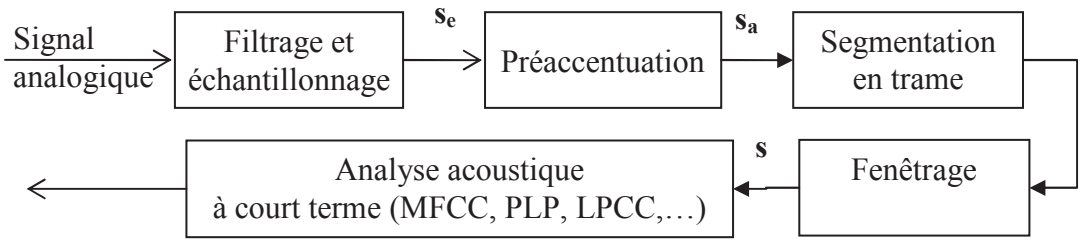


Figure.I.5 : Prétraitement acoustique du signal vocal

2. Préaccentuation : le signal échantillonné **s** est ensuite pré-accentué afin de relever les hautes fréquences qui sont moins énergétiques que les basses fréquences. Cette étape consiste à faire passer le signal s_n dans un filtre numérique à réponse impulsionnelle finie de premier ordre donné comme suit [13] :

$$H(z) = 1 - \alpha\, z^{-1} \;\; avec \qquad 0.9 \leq \alpha \leq 1 \tag{I.5}$$

Ainsi, le signal préaccentué s_a est lié au signal s_e par la formule suivante :

$$\boldsymbol{s_a}(n) = \boldsymbol{s_e}(n) - \alpha\, \boldsymbol{s_e}(n-1) \tag{I.6}$$

3. Segmentation : les méthodes du traitement de signal utilisées dans l'analyse du signal vocal opèrent sur des signaux stationnaires, alors que le signal vocal est un signal non stationnaire. Afin de remédier à ce problème, l'analyse de ce signal est effectuée sur des trames successives de parole, de durée relativement courte sur lesquelles le signal peut en général être considéré comme quasi stationnaire [1]. Dans cette étape de segmentation, le signal préaccentué est ainsi découpé en trames de N échantillons de parole. En général N est fixé de telle manière à ce que chaque trame corresponde à environ 20 à 30 ms de parole. Deux trames successives sont séparées de M échantillons correspondant à une période de l'ordre de la centi-seconde.

4. Fenêtrage : la segmentation du signal en trames produit des discontinuités aux frontières des trames. Dans le domaine spectral, ces discontinuités se manifestent par des lobes secondaires. Ces effets sont réduits en multipliant les échantillons $\{\boldsymbol{s_a}(n)\}_{n=0\ldots N-1}$ de la trame par une fenêtre de pondération $\{w(n)\}_{n=0\ldots N-1}$ telle que la fenêtre de Hamming [17].

$$\boldsymbol{s}(n) = w(n).\, \boldsymbol{s_a}(n) \tag{I.7}$$

$$\text{avec} \quad w(n) = 0.54 - 0.46\cos\left(2\pi \frac{n}{N-1}\right) \qquad O \leq n \leq N-1$$

5. Analyse à court terme : chaque trame fenêtrée du signal est ensuite convertie en un vecteur acoustique constitué d'un ensemble réduit de paramètres.

Différentes méthodes coexistent pour la transformation d'une trame fenêtrée de signal en un vecteur acoustique [18] :

- Les méthodes paramétriques qui se basent sur un modèle de production tel que le codage par prédiction linéaire LPC (*Linear Prediction Coding*), LPCC (*Linear Prediction Cepstral Coefficients*).
- Les méthodes non paramétriques telles que le taux de passage par zéro, la fréquence fondamental (*pitch*), la transformée de Fourier discrète, l'énergie du signal, les sorties d'un banc de filtres numériques et la transformée en ondelettes.
- Les méthodes fondées sur un modèle de perception tel que MFCC (*Mel Frequency Cepstral Coefficients)* et PLP *(Perceptual Linear Prediction*) [19].

Des études comparatives entre différents types de paramètres acoustiques ont été effectuées, pour déterminer ceux qui représentent mieux le signal vocal.

Dans [20], une étude comparative classique a été effectuée entre plusieurs représentations du signal vocal: cepstre en sortie d'un banc de filtres en échelle Mel (MFCC) ou en échelle linéaire (LFCC), coefficients de prédiction linéaire (LPC) ou de réflexion (RC), cepstre calculé à partir des coefficients auto-régressifs (LPCC). Ces représentations ont été appliquées dans un système de reconnaissance fondé sur un alignement DTW entre un mot de test et des mots de référence. Dans cette étude, les MFCC donnent les meilleurs résultats, ce qui montre plus généralement l'intérêt d'un pré-traitement par banc de filtres, d'une échelle fréquentielle non linéaire, et de la représentation cepstrale [6].

Dans [21], une analyse acoustique par LPC appliquée dans un système de reconnaissance fondé sur les modèles HMM discrets, a donné de meilleurs résultats qu'une analyse par un banc de filtre ou par transformée de Fourier. En revanche, dans [22], le prétraitement perceptif des coefficients PLP a amélioré les résultats de l'analyse par prédiction linéaire.

Dans [23], les paramètres MFCC, LPCC et pseudo-coefficients cepstraux obtenus à partir de deux modèles, ont été comparés sur un système de reconnaissance des mots isolés par les HMM pour diverses conditions d'enregistrements. Dans les conditions normales, les MFCC sont plus performants par rapport aux LPCC et présentent un écart très faible par rapport aux coefficients basés sur des modèles d'audition, alors que, selon [24], les PLP et MFCC donnent des résultats comparables.

Dans [25], des expériences ont montré que les coefficients MSG (*Modulation SpectroGram*) sont plutôt plus performants avec un classificateur réseau de neurones qu'avec un système standard HTK basé sur les GMM (*Gaussian Mixture Model*) ; mais dans ce dernier système, les performances des coefficients MSG sont inférieures à celles des paramètres MFCC.

Dans une étude de Furui [26] [4]; il est montré que la prise en compte de l'évolution est possible par l'introduction d'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. Dans [5] les auteurs ont montré que les coefficients différentiels du second ordre peuvent contribuer à l'amélioration des systèmes d'identification phonétique, ainsi que leur intérêt pour de la parole bruitée et soumise à l'effet Lombard.

En se basant sur les résultats de recherche décrits précédemment, nous avons retenu dans notre travail les coefficients MFCC, PLP, LPCC ainsi que leurs paramètres différentiels de premier et deuxième ordre pour sélectionner parmi ces paramètres, ceux les plus pertinents. Cette pertinence est validée à partir du système de référence qui est fondé originalement sur les paramètres MFCC, ainsi que leurs paramètres différentiels. Ces coefficients ont des particularités que nous nous proposons de décrire dans les paragraphes suivants. En revanche, les coefficients PLP et LPCC sont décrits dans l'annexe A.

I.5.1.1 Les coefficients cepstraux

Le signal vocal résulte de la convolution de la source par le conduit vocal. Dans le domaine spectral, cette convolution devient un produit qui rend difficile la séparation de la contribution de la source et celle du conduit. Ce problème peut être surmonté par l'analyse cepstrale par passage dans le domaine log-spectral [27]. En pratique, le cepstre réel d'un signal numérique *s*(n) estimé sur une fenêtre d'analyse de N échantillons, est obtenu comme suit [1]:

$$s(n) \xrightarrow{\text{FFT}} S(f) \xrightarrow{Log|\ \ |} Log|S(f)| \xrightarrow{\text{FFT}^{-1}} \text{cepstre}$$

Les coefficients cepstraux sont donnés par :

$$c(n) = \frac{1}{N}\sum_{j=0}^{N-1} Log(\,|S(j)|)\, e^{\frac{2ijn}{N}} \qquad pour\ n = 0, 1, \dots, N-1 \tag{I.8}$$

Ce type de calcul des coefficients cepstraux n'est pas utilisé en reconnaissance de la parole [1] du fait du calcul important de la FFT et de la FFT inverse [28]. En revanche, les coefficients cepstraux utilisés peuvent être obtenus à partir des coefficients de la prédiction linéaire ou des énergies d'un banc de filtres. Ainsi, les paramètres LPCC (*Linear Prediction Cepstral Coefficients*) sont calculés à partir d'une analyse par prédiction linéaire (voir annexe A). Si a_0=1, $\{a_i\}_{i=1:p}$ sont les coefficients de cette analyse, estimés sur une trame du signal, les d premiers coefficients cepstraux C_k sont calculés récursivement par :

$$C_k = -a_k - \sum_{i=1}^{k-1} \frac{(k-i)}{k}\, C_{k-i} a_i \qquad 1 \le k \le d \tag{I.9}$$

Un liftrage est effectué pour augmenter la robustesse des coefficients cepstraux [29]. Ce liftrage consiste à multiplier des coefficients cepstraux par une fenêtre de poids $W(k)$ pour être moins sensible au canal de transmission et au locuteur [5]:

$$\forall\, k \in [1, L] \quad W(k) = 1 + \frac{L}{2}.\sin\left(\frac{\pi.k}{L}\right) \tag{I.10}$$

où L est le nombre de coefficients.

Les coefficients MFCC (*Mel Frequency Cepstral Coefficients*) sont les paramètres les plus utilisés dans les systèmes de la reconnaissance de la parole. L'analyse MFCC consiste à exploiter les propriétés du système auditif humain par la transformation de l'échelle linéaire des fréquences en échelle Mel (voir figure I.6). Cette dernière échelle est codée au travers d'un banc de 15 à 24 filtres triangulaires espacés linéairement jusqu'à 1 KHz, puis espacés logarithmiquement jusqu'aux fréquences maximales. La conversion de l'échelle linéaire en échelle Mel est donnée par:

$$mel = 2595\, log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \tag{I.11}$$

Sur une trame d'analyse du signal, les coefficients MFCC sont calculés à partir des énergies issues d'un banc de filtres triangulaires en échelle de fréquence Mel [1]. Les d premiers coefficients cepstraux (en général d est choisi entre 10 et 15) C_k peuvent être calculés directement en appliquant la transformée en cosinus discrète sur le logarithme des énergies E_i sortant d'un banc de M filtres:

$$C_k = \sum_{i=1}^{M} \log(E_i)\,.\,cos\left[\frac{\pi k}{M}(i - \frac{1}{2})\right] \qquad k = 0, \dots, d \leq M \tag{I.12}$$

La transformée en cosinus discrète permet de fournir des coefficients peu corrélés [30]. Le coefficient C_0 représente la somme des énergies. Généralement, ce coefficient n'est pas utilisé. Il est remplacé par le logarithme de l'énergie totale E_0 calculée et normalisée sur la trame d'analyse dans le domaine temporel [6].

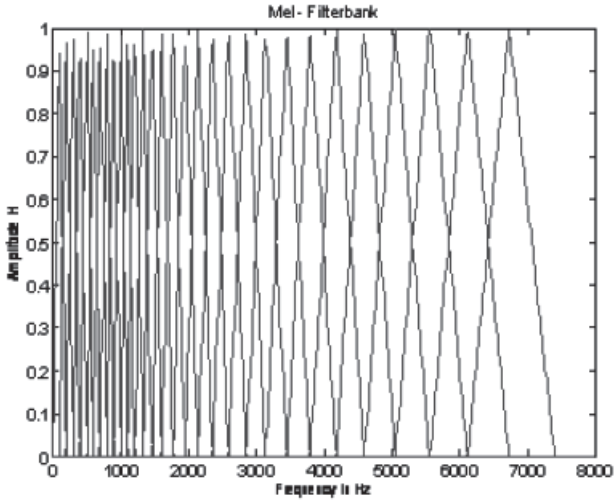


Figure.I.6 : Banc de filtres en échelle Mel

I.5.1.2 Coefficients différentiels

Généralement les coefficients MFCC sont désignés comme des paramètres statiques, puisqu'ils contiennent seulement l'information sur une trame donnée. Afin d'améliorer la représentation de la trame, il est souvent proposé d'introduire de nouveaux paramètres dans le vecteur des paramètres. Furui [26] [4] a proposé l'utilisation des paramètres dynamiques qui présentent l'information de transition spectrale dans le signal vocal. En particulier, il a proposé des coefficients différentiels du premier ordre appelés aussi coefficients delta, 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 $2n_\Delta$ trames d'analyse par l'estimation de la pente de la régression linéaire du coefficient C_k à l'instant t [15]:

$$\Delta C_k(t) = \frac{\sum_{i=-n_\Delta}^{i=n_\Delta} i\, C_k(t+i)}{2.\sum_{i=-n_\Delta}^{i=n_\Delta} i^2} \qquad (\text{I.13})$$

Le coefficient delta de l'énergie ΔE_0 est calculé de la même façon.

Les coefficients différentiels du second ordre $\Delta\Delta$ (delta-delta ou d'accélération) sont calculés de la même manière à partir des coefficients du premier ordre. Ces coefficients ont contribué eux aussi à l'amélioration des performances des systèmes RAP. Néanmoins, cette amélioration peut être négligeable comparée à celle des coefficients delta [31] [32].

Dans un système fondé sur l'analyse cepstrale, il est souvent utile de ne garder que les douze premiers paramètres MFCC auxquels on adjoint généralement le logarithme de l'énergie normalisée, ainsi que leurs coefficients différentiels du premier et deuxième ordre. L'ensemble constitue un vecteur de 39 paramètres.

Cependant l'ajout des paramètres différentiels demande plus de temps de calcul et plus d'espace mémoire ainsi que plus de nombre d'échantillons constituant la base de données utilisée pour l'apprentissage, ce qui devient critique pour les systèmes embarqués [7]. Une solution à ces problèmes est de limiter le nombre de paramètres en sélectionnant les plus pertinents susceptibles de modéliser le mieux possible les données pour la tâche de reconnaissance. Dans notre travail de thèse, nous nous sommes intéressés à cette solution en utilisant un outil de mesure de pertinence fondé sur la théorie d'information.

I.5.2 LES MODELES ACOUSTIQUES HMM

Les modèles de Markov cachés (*Hidden Markov Models* ou HMM) ont connu une grande importance depuis leur introduction en traitement de la parole [33] [34]. La plupart des systèmes de RAP utilisent les HMM pour modéliser les mots ou les unités élémentaires de la parole tels que les phonèmes, les syllabes. De nombreux travaux de recherches ont montré l'efficacité de ces modèles en reconnaissance de la parole continue ou isolée, indépendamment du locuteur pour de petits et grands vocabulaires. Ils sont utilisés avec des modèles de phonèmes indépendants où dépendant du contexte [28]. Les HMM supposent que le phénomène modélisé est un processus aléatoire et inobservable qui génère des émissions elles-mêmes aléatoires [6]. Ainsi, un HMM résulte de l'association de deux processus stochastiques : un processus interne, non observable $Q(t)$ et un processus observable $O(t)$, d'où le nom de modèle caché.

Dans le cas de la parole, la chaîne interne $Q(t)$ est une chaîne de Markov qui est supposée être à chaque instant t dans un état q_t où la fonction aléatoire correspondante émet un segment élémentaire de l'onde acoustique observée représenté par un vecteur de paramètres o_t (exemple des paramètres MFCC) extraits par une analyse acoustique [1]. Idéalement, il faudrait pouvoir associer à chaque phrase possible un modèle. Ceci est irréalisable en pratique car le nombre de modèles serait beaucoup trop élevé. Des sous-unités lexicales comme le mot, la syllabe, ou le phonème sont utilisées afin de réduire le nombre de paramètres à entraîner durant une phase d'apprentissage. En particulier, dans les systèmes de reconnaissance de parole continue à grand vocabulaire, il n'est pas raisonnable d'associer un modèle pour chaque mot, et l'utilisation d'unités acoustiques sous-lexicales devient ainsi indispensable. Par contre dans les systèmes de reconnaissance à petit vocabulaire, la modélisation par mot est très efficace [35].

A chacune de ces unités lexicales, est associé un modèle de Markov caché constitué d'un nombre fini d'états prédéterminés [28]. La reconnaissance de la parole revient à choisir le

HMM capable après avoir été entraîné, d'avoir la plus grande probabilité d'émettre la séquence d'observations (vecteurs acoustiques) correspondante au signal d'entrée.

I.5.2.1 Fonctionnement d'un HMM

Un modèle de Markov est un automate probabiliste d'états finis décrit par un ensemble de nœuds (ou états) reliés entre eux par des arcs de transitions. Cet automate est contrôlé par deux processus stochastiques. Le premier commence sur l'état initial du HMM et se déplace ensuite d'état en état à chaque instant t $(1 \leq t \leq T)$, en respectant les transitions autorisées par la topologie de l'automate. Le deuxième génère après chaque changement d'état à l'instant t une observation o_t. À chaque état i $(1 \leq i \leq N)$ est associée une distribution de probabilité $b_i(o)$ et à chaque transition de l'état i à l'état j est associée une probabilité de transition a_{ij} [36].

La distribution de probabilité $b_i(o)$ représente la probabilité d'émission sur l'état i de l'observation o. Ainsi, si l'ensemble des observations possibles est un alphabet fini, alors la classe de distribution de probabilité est discrète et le HMM est dit discret. Si les observations sont définies dans un espace continu R^d, alors la classe de distribution de probabilité est continue et le HMM est dit continu (voir figure I.7).

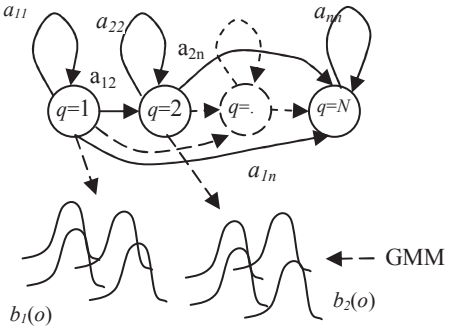


Figure.I.7 : Exemple d'un Modèle de Markov gauche-droite

La mise en œuvre d'un HMM en reconnaissance de la parole nécessite de résoudre certains problèmes et de poser un ensemble d'hypothèses simplificatrices. Les trois problèmes principaux sont [1]]:

1. Le choix des paramètres du modèle : quelle topologie utiliser pour définir un modèle (nombre d'états, transitions, loi de probabilité d'émission) ?
2. L'apprentissage : étant donné un ensemble de J séquences d'observations O_j associées à chacun des Modèles de Markov M_j, comment estimer les paramètres λ_j de ces modèles afin de maximiser la vraisemblance de la suite d'observations O_j.

$$arg \max_{\lambda} \prod_{j=1}^{J} P(O_j \backslash M_j, \lambda_j) \qquad (I.14)$$

où λ représente l'ensemble des paramètres de tous les modèles λ_j.

3. La reconnaissance : étant donnée une séquence d'observation *X*, et un ensemble de HMM, quelle est la séquence de ces modèles qui maximise la probabilité de générer *X* ?

Les différentes hypothèses simplificatrices pour résoudre ces problèmes peuvent être résumées comme suit [1]:

- Le signal de parole est généré par une suite d'états, chaque état est géré par une loi de probabilité. Chaque unité lexicale élémentaire (phone, diphone, triphone, mot) est associée à un modèle de Markov et la concaténation de tels modèles permet d'obtenir des mots ou des phrases.
- La probabilité que le modèle de Markov soit dans l'état *i* au temps *t* ne dépend que de l'état du modèle au temps *t*-1. Cette condition conduit à un modèle de Markov de premier ordre appelé chaîne simple de Markov:

$$P(q_t \backslash q_{t-1} q_{t-2} \dots q_0) \cong P(q_t \backslash q_{t-1}) \qquad (\text{I. } 15)$$

- La chaîne e de Markov est stationnaire:

$$P(q_t = j \backslash q_{t-1} = i) \cong P(q_{t+\tau} = j \backslash q_{t+\tau-1} = i) \qquad \forall\, \tau \qquad (\text{I. } 16)$$
$$= a_{ij}$$

où a_{ij} est la probabilité de transition de l'état *i* à l'état *j*.

- La probabilité qu'un vecteur soit émis au temps *t* dépend uniquement de l'état au temps *t*:

$$P(o_t \backslash q_0 q_1 \dots q_t, o_1 \dots o_{t-1}) \cong P(o_t \backslash q_t) \qquad (\text{I. } 17)$$

- Les HMM utilisés pour représenter la parole sont, la plupart du temps, des modèles "gauche-droit" qui ne permettent pas de "retour en arrière". L'automate probabiliste correspondant ne contient pas de transition entre les états *i* et *j* de telle sorte que:

$$i > j \Rightarrow a_{ij} = 0 \qquad (\text{I. } 18)$$

- Dans le cas d'un HMM continu, la distribution de probabilité d'émettre l'observation o sachant que le processus markovien est dans l'état j, est représentée par un modèle de mélange de k gaussiennes (GMM Gaussian Mixture Model) ayant la forme suivante:

$$b_j(o_t) = \sum_{i=1}^{k} \frac{c_i}{\sqrt{(2\pi)^d \left|\sum_{\mathrm{i}}\right|}} \exp\left(-\frac{1}{2}(o_t - \mu_i)'.\sum_{\mathrm{i}}^{-1}.(o_t - \mu_i)\right) \qquad (\text{I. } 19)$$

μ_i et $\sum_{\mathrm{i}}$ représentent la moyenne et la matrice de covariance de la i$^{\text{ième}}$ gaussienne, c_i est le poids ou la probabilité a priori de la $i^{\text{ième}}$ gaussienne vérifiant la condition : $\sum_{i=1}^{k} c_i$=1.

$|\sum_i|$ est le déterminant de la matrice $\sum_i$.

Ainsi, compte tenu des détails présentés ci-dessus, un modèle HMM du premier ordre à N états est défini par la connaissance des paramètres $\lambda = \{\Pi, A, B\}$:

- L'ensemble $\Pi=\{\pi_i, 1 \leq i \leq N\}$ des probabilités initiales π_i , probabilité d'être dans l'état i à l'instant initial.
- La matrice de transition A de taille N×N d'éléments a_{ij}.
- L'ensemble de distributions de probabilités d'émission $B=\{b_i(o), 1 \leq i \leq N\}$: $b_i(o)$ est la probabilité d'émettre l'observation o sachant que le processus markovien est dans l'état i.

Différent critères sont proposés pour l'estimation de ces paramètres, comme le critère du maximum de vraisemblance (Maximum Likelihood Estimation ou MLE), le critère MAP (Maximum A Posteriori) et MMI (Maximum Mutual Information). Cependant la mise en œuvre de ces deux derniers est généralement plus difficile [6]. Par contre le critère du maximum de vraisemblance MLE est souvent utilisé pour l'apprentissage car il est moins coûteux en temps de calcul. Ainsi dans les paragraphes suivants, on décrit brièvement l'algorithme de Baum-Welch réalisant le critère MLE.

I.5.2.2 Phase d'apprentissage par critère MLE

Le critère MLE consiste à chercher le meilleur ensemble de paramètre λ qui maximise la probabilité d'émission de la séquence d'observations O_j par le modèle M_j :

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

Cependant cette maximisation n'a pas une solution analytique, mais pratiquement les formules de Baum-Welch permettent une réestimation itérative des paramètres a_{ij} et $b_i(o)$ en appliquant ce critère [37]. En partant d'une estimation initiale λ_0, on réestime les paramètres du nouvel ensemble des paramètres λ_1. Ensuite on effectue des itérations pour obtenir de meilleurs ré-estimations :

$$P(O/\lambda_{n+1}) \geq P(O/\lambda_n) \geq \cdots \geq P(O/\lambda_2) \geq P(\lambda_1) \geq P(O/\lambda_0) \qquad (I.20)$$

Pratiquement la convergence de cet algorithme nécessite une bonne initialisation des paramètres des modèles et un nombre élevé de données d'apprentissage. Cette initialisation peut être effectuée par l'algorithme de Viterbi utilisé dans le décodage.

Les formules de Baum-Welch pour la réestimation des paramètres du modèle sont données directement sans démonstration (le lecteur trouve plus de détails sur ces formules dans [1] [13] [6]). Les formules suivantes concernent les paramètres d'un modèle HMM continu dont

chaque distribution de probabilités d'émission b_k d'un état k est une Gaussienne multivariée de moyenne μ_k et de covariance $\sum_k$. Ces formules peuvent être généralisées au cas d'une distribution d'un mélange de lois de Gauss.

La moyenne μ_k est donnée par [1]:

$$\mu_k = \frac{\sum_{t=1}^{T} \gamma_t(k).\, o_t}{\sum_{t=1}^{T} \gamma_t(k)} \qquad (I.21)$$

$\gamma_t(i) = P(q_t=i\backslash O,\lambda)$ est la probabilité que l'état à l'instant *t* soit à l'état *i* sachant l'observation O et le modèle λ.

La covariance $\sum_k$ est donnée par :

$$\sum_k = \frac{\sum_{t=1}^{T} \gamma_t(k).\,(o_t - \mu_k)\,.\,(o_t - \mu_k)'}{\sum_{t=1}^{T} \gamma_t(k)} \qquad (I.22)$$

La probabilité de transition a_{ij} est donnée comme suit :

$$a_{ij} = \frac{\sum_{t=1}^{T} \xi_t(i,j)}{\sum_{t=1}^{T} \gamma_t(i)} \qquad (I.23)$$

avec $\xi_t(i,j) = P(q_t = i, q_{t+1} = j \backslash O, \lambda).$

Les probabilités γ et ξ sont données comme suit :

$$\gamma_t(i) = \frac{P(o_1 \dots o_t, q_t = i\backslash\, \lambda).\, P(o_{t+1} \dots o_T \backslash q_t = i, \lambda)}{P(O\backslash\lambda)} = \frac{\alpha_t(i)\ \beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\ \beta_t(i)} \qquad (I.24)$$

$$\xi_t(i,j) = \frac{P(q_t = i, q_{t+1} = j\backslash O, \lambda)}{P(O\backslash\lambda)} = \frac{\alpha_t(i)\ a_{ij}\ b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i) a_{ij}\ b_j(o_{t+1})\beta_{t+1}(j)} \qquad (I.25)$$

où $\alpha_t(i) = P(o_1 \dots o_t, q_t = i\backslash\, \lambda)$ et $\beta_t(i) = P(o_{t+1} \dots o_T\ q_t = i, \lambda).$

α et β s'obtiennent par récurrence par la méthode appelée forward-backward :

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i)\ \ a_{ij}\right] b_j(o_{t+1}) \qquad (I.26)$$

$$\beta_t(i) = \left[\sum_{i=1}^{N} a_{ij}\ \ b_j(o_{t+1})\ \ \beta_{t+1}(j)\right] \qquad (I.27)$$

Ces paramètres sont recalculés itérativement en utilisant l'algorithme EM.

En pratique, l'apprentissage des modèles s'effectue comme suit [1][3]:

1. Création de modèles dont les probabilités de transition sont équiprobables. La moyenne et l'écart-type de l'ensemble des observations sont calculés et affectés à chaque état.
2. En utilisant les formules de réestimation ci-dessus, on obtient une première approximation pour les probabilités de transition et pour les paramètres des lois de Gauss.

Après un certain nombre d'itérations, les paramètres des gaussiennes pour chaque état sont estimés.

Afin d'obtenir un modèle plus précis, il est nécessaire d'augmenter le nombre de gaussiennes pour chaque état à condition de disposer de suffisamment d'exemples d'apprentissage [1].

I.5.2.3 Phase de reconnaissance

A). <u>Reconnaissance de mots isolés</u>

Considérant V mots d'un vocabulaire dont chaque mot W_i est modélisé par un modèle M_i, la reconnaissance d'un mot inconnu appartenant au vocabulaire revient à chercher le meilleur modèle permettant de générer la meilleure séquence d'états $Q=(q_1,q_2,\ldots,q_T)$ qui peut générer la séquence d'observations $O=(o_1,o_2,\ldots,o_T)$ correspondant à la prononciation du mot inconnu.

Une solution consiste à évaluer le maximum $P(O\backslash M,Q)$ pour chaque séquence d'états possibles Q, avec :

$$P(O\backslash M,Q) = \pi_0 \prod_{t=1}^{T} a_{q_{t-1}q_t}.\ b_{q_t}(o_t) \qquad (\text{I.28})$$

Le nombre de séquences d'états possible est très grand, de l'ordre de $T.N^T$. Ainsi, cette solution est inapplicable en général. Une autre solution basée sur une variante stochastique de la programmation dynamique appelée l'algorithme de Viterbi [1] [13], permet de trouver la séquence optimale d'états avec une complexité de calcul limitée à $T.N^2$.

Cet algorithme consiste à construire de façon itérative la meilleure séquence d'états à partir d'un tableau T×N contenant les valeurs $\delta_t(i)$ définissant la vraisemblance du meilleur chemin fini à l'état i au temps t. La valeur $\delta_t(i)$ peut être calculée par récurrence :

- ✓ Initialisation : $\delta_0(i)=\pi_i$, probabilité d'être dans l'état i à l'instant initial
- ✓ Récursion :

$$\delta_t(i) = \arg\max_j(\delta_{t-1}(i)\ a_{ij})\ b_i(o_t) \qquad (\text{I.29})$$

- ✓ Terminaison : $P = \arg\max_i \delta_T(i)$, cette valeur détermine la vraisemblance donnée par la formule (I.29).

B) <u>Reconnaissance de la parole continue</u>

Dans la reconnaissance de la parole continue, on ne connait pas le nombre de mots qui composent une phrase, ni les frontières de chaque mot. Cette difficulté peut être résolue en effectuant une transition du dernier état d'un mot vers le premier état d'un des mots du

vocabulaire. Ainsi, une modification de l'algorithme de reconnaissance des mots isolés est nécessaire.

L'algorithme de Viterbi devient [1][3]:

- ✓ Initialisation : $\delta_0^k = \pi_i^k$, k désignant l'un des mots du vocabulaire.
- ✓ Récursion:

Etat non initial du mot k : $\delta_t^k(i) = \max_j(\delta_{t-1}^k(i).\, a_{ij}^k)\ .\ b_i(o_t)$,

Etat initial du mot k:

$$\delta_t\ (i) = \max\ (\max_j(\delta_{t-1}^k(i).\, a_{ij}^k)\ ,$$

$$\max_l(\delta_{t-1}^l(\text{étatfinal}(l)).\, P(M_l \backslash M_k)))\ .\ b_i(o_t)$$

- ✓ Terminaison : $P = \arg\max_k(\delta_T^k(i))$.

La terminaison indique seulement le dernier état qui maximise la séquence ; afin de retrouver la séquence de modèles, deux solutions peuvent être envisagées. La première consiste à mémoriser dans le tableau des δ l'état qui avait contribué à calculer le maximum. Ainsi, à la dernière trame, il suffit de revenir en arrière à partir de l'état qui maximise $\delta_T^k(i)$ pour retrouver la séquence optimale d'états et la séquence de mots. La seconde consiste à mémoriser dans $\delta_T^k(i)$ la séquence de mots qui a permis d'arriver à l'état i du modèle k à l'instant t.

A la fin de la reconnaissance, on obtient non seulement la séquence de mots prononcés mais aussi les frontières de chaque mot [1].

Dans l'annexe B, nous décrivons la mise en œuvre pratique d'un système de reconnaissance de mots connectés sous la plateforme HTK. Plus particulièrement ce système se base sur les modèles HMM associées aux modèles GMM.

I.6 CONCLUSION

Le signal acoustique de la parole présente une grande variabilité qui complique la tâche des systèmes RAP. Cette complexité provient de la combinaison de plusieurs facteurs, comme la redondance du signal acoustique, la grande variabilité intra et inter-locuteurs, les effets de la coarticulation en parole continue, ainsi que les conditions d'enregistrement. Pour surmonter ces problèmes, différentes approches sont envisagées pour la reconnaissance de la parole telles que les méthodes analytiques, globales et les méthodes statistiques. Actuellement la majorité des systèmes RAP sont construits selon la méthode statistique en utilisant les modèles de Markov cachés HMM. Ainsi, dans ce chapitre, nous avons décrit brièvement le principe de fonctionnement des systèmes RAP basés sur les modèles HMM ainsi que leur mise en œuvre pratique.

Ces systèmes s'articulent autour de deux phases principales (apprentissage et reconnaissance) dont chacune réalise une analyse acoustique qui transforme tout signal vocal en une suite de vecteurs constitués des paramètres acoustiques afin d'éliminer la redondance présente dans le signal vocal. Ces vecteurs acoustiques sont utilisés comme des observations dans la modélisation markovienne des données parole.

Les coefficients MFCC sont les paramètres les plus utilisés dans les systèmes de RAP. Ces coefficients sont généralement utilisés avec leurs paramètres dynamiques Δ et $\Delta\Delta$ afin d'améliorer les performances de ces systèmes. Néanmoins l'ajout de paramètres exige dans certaines applications de RAP de réduire le nombre de paramètres acoustiques en sélectionnant les plus pertinents. Ainsi, notre travail de thèse a consisté principalement à intégrer une étape de sélection des paramètres acoustiques les plus pertinents en s'appuyant sur une mesure de pertinence issue de la théorie de l'information. Dans le chapitre suivant, nous ferons un état de l'art des méthodes de la réduction de la dimensionnalité ainsi qu'une description sur la sélection des variables aléatoires les plus pertinentes.

CHAPITRE II

REDUCTION DE LA DIMENSIONNALITE

II.1 POSITION DU PROBLEME

L'objectif primaire de la reconnaissance de forme est la classification, supervisée ou non supervisée. L'évaluation des performances d'un classifieur dépend du nombre d'échantillons utilisés lors de la conception du classifieur, du nombre de caractéristiques (variables, paramètres, ou *features*) retenues pour la tâche de classification et de la complexité du classifieur [8]. Augmenter le nombre de caractéristiques nécessite l'accès à un nombre d'échantillons d'entraînement de plus en plus important, si l'on veut conserver au moins le même niveau de performance [38]. Ce phénomène nommé la malédiction de la dimensionnalité (« *curse of dimensionnality* ») conduit au phénomène de « *peaking* » [8]. Ainsi, il est observé en pratique que l'ajout de caractéristiques peut être la cause d'une dégradation des performances d'un classifieur si le nombre d'échantillons des données d'entraînement utilisées pour la conception du classifieur reste inférieur relativement au nombre des caractéristiques [39] [40].

Face à ce problème où le nombre de caractéristiques peut être important, il est souvent essentiel d'adopter une stratégie efficace de réduction du nombre de caractéristiques pertinentes pour cette tâche de classification afin :

- d'apporter une réponse au problème de la malédiction de la dimensionnalité ;
- d'accéder à une modélisation la plus fine et la plus compacte possible des formes pour la tâche de classification ;
- de réduire les coûts de calcul et d'encombrement mémoire qui deviennent prohibitifs avec la dimension.

Dans la reconnaissance de la parole, de nombreuses approches proposées pour améliorer la robustesse et les performances des systèmes de RAP consistent à travailler sur les méthodes de transformation du signal de parole en une séquence de vecteurs de paramètres acoustiques. Plusieurs transformations produisent plusieurs types de paramètres qu'il est possible de combiner comme par exemple :

- Paramètres MFCC, PLP et auditifs [2].
- Paramètres spectraux et discriminants [41].
- Paramètres acoustiques et articulatoires [42] [43].
- Paramètres LPCC, MFCC, PLP, énergies et durée moyenne [44] [45] [46].

- Paramètres PLP, MFCC et ondelettes [3].

La combinaison optimale n'est cependant pas connue. Une solution consisterait à prendre en compte tous les coefficients issus de toutes les transformations puis de réduire de façon adéquate la dimension de l'espace acoustique créé, en ne retenant qu'un nombre limité de ces coefficients.

La réduction de la dimension de l'espace acoustique peut se faire principalement selon les techniques suivantes:

- Techniques d'extraction des caractéristiques (*Feature extraction*).
- Techniques de sélection des caractéristiques (*Feature selection*)

Les techniques d'extraction permettent de créer de nouveaux ensembles de caractéristiques, en utilisant une transformation ou une combinaison d'un espace de départ. Les principales techniques sont l'analyse en composantes principales (ACP) obtenues selon un développement de Karhunen-Loève [47], l'analyse en composantes indépendantes [48] (plus appropriée que l'ACP pour les distributions non-gaussiennes), l'analyse linéaire discriminante (ALD) [47]. L'ACP est une transformation non supervisée, c.-à-d. qu'aucune information sur les classes présentes dans les données n'est utilisée, même si cette information est disponible. L'ACP a la propriété de décorréler les coefficients. Il y a ainsi peu à attendre *a priori* d'une ACP sur des coefficients cepstraux, puisque la transformation cepstrale produit déjà des coefficients faiblement corrélés [23] [49]. Par opposition à l'ACP, l'ALD est supervisée, car elle utilise les étiquettes des classes afin de trouver la transformée qui sépare le mieux les classes. Le critère maximisé par l'ALD est le rapport de la variance intraclasse sur la variance interclasses. Le résultat de la transformation est une représentation de basse dimension, comme l'ACP, mais avec des classes mieux séparées et compactes. L'ALD est cependant limitée à des probabilités conditionnelles normales de classes. De surcroît, en conditions bruitées, la nature du bruit conditionne très fortement la robustesse des paramètres issus de l'ALD [50].

Pour les raisons évoquées ci-dessus, les techniques d'extraction ne seront pas abordées dans ce document.

Les techniques de sélection correspondent aux algorithmes permettant de sélectionner un sous-ensemble de caractéristiques les plus pertinentes ou informantes parmi un ensemble de départ, pour la réalisation d'une tâche pour laquelle un système a été conçu et ceci utilisant divers critères et différentes méthodes.

Les domaines d'application des techniques de sélection de caractéristiques sont variés tels que la modélisation, la classification, l'apprentissage automatique (*Machine Learning*) et

l'analyse exploratoire de données (*Data Mining*) [51]. Dans cette thèse nous nous intéressons plus particulièrement à la sélection de caractéristiques pour la classification.

II.2 SELECTION DE CARACTERISTIQUES

II.2.1 ETAT DE L'ART

Plusieurs définitions de la sélection de caractéristiques sont proposées dans la littérature. Dans [52], l'énoncé de la sélection de caractéristiques proposée est la suivante : étant donné un ensemble de dimension *N*, il faut sélectionner le sous-ensemble de dimension *M* tel que $M < N$, conduisant au taux d'erreur le plus faible en classification.

Dans [53], Dash a proposé de regrouper et identifier les techniques de sélection de caractéristiques en quatre classes distinctes selon la fonction objectif visée :

1. idéalisée "*Idealized*" : trouver le sous-ensemble de taille minimale qui est nécessaire et suffisant pour atteindre l'objectif fixé [54].
2. classique "*Classic*" : sélectionner un sous-ensemble de *M* caractéristiques à partir d'un ensemble de *N* caractéristiques $M < N$, tel que une fonction de critère soit optimisée sur tous les sous-ensembles de taille *M* [55].
3. améliorer la précision de prédiction "*Improving prediction accuracy*" : l'objectif de la sélection des caractéristiques est de sélectionner un sous-ensemble de ces caractéristiques pour améliorer la précision de la prédiction ou diminuer la taille de la structure sans diminution significative de la précision de prédiction du classificateur, construit en utilisant seulement les variables sélectionnées.
4. approximer la distribution originale de classe "*Approximating original class distribution*": le but de la sélection des caractéristiques est de sélectionner le plus petit sous-ensemble tel que la distribution des classes résultante, étant donné seulement les valeurs des caractéristiques sélectionnées, soit aussi proche que possible de la distribution des classes originale, étant donné l'ensemble complet des caractéristiques.

Ces deux dernières approches sont reprises par Koller et Sahami dans [56]. L'avantage de cette dernière est qu'elle est générale et permet ainsi de regrouper l'ensemble des algorithmes de sélection [51].

Selon [53], le processus typique de sélection des caractéristiques se base sur les quatre étapes présentées dans la figure (II.1) :

1. une procédure de recherche (ou génération) permet d'explorer différentes combinaisons de caractéristiques.
2. un critère d'évaluation des caractéristiques permet de comparer les différents sous ensembles et de choisir le meilleur.

3. un critère d'arrêt permet de stopper la procédure de recherche ou de relancer une nouvelle procédure de recherche.
4. une procédure de validation vérifie si le sous-ensemble sélectionné est valide (§II.2.5).

La mise en œuvre d'un algorithme de sélection de caractéristiques se base généralement sur les trois premières étapes ci-dessus [57] [51].

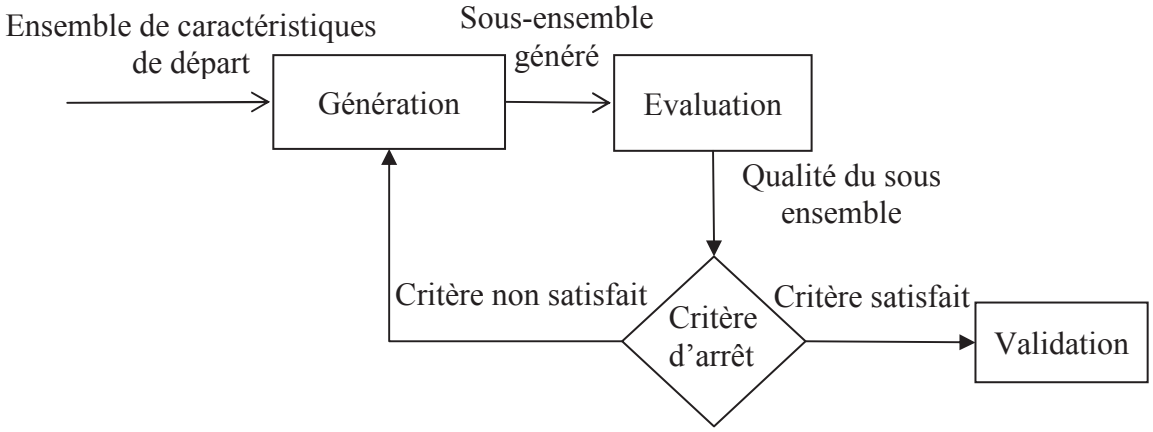


Figure.II.1 : représentation graphique du processus de sélection de caractéristiques d'après Dash

II.2.2 PROCEDURE DE RECHERCHE

Idéalement, la procédure de recherche consiste, à partir d'un ensemble de N caractéristiques, à générer 2^N-1 sous-ensembles de caractéristiques candidats à évaluer. Cette procédure est exhaustive et devient rapidement prohibitive et très coûteuse, même pour une taille moyenne de l'ensemble de caractéristiques (N). Des méthodes heuristiques ou aléatoires sont introduites pour réduire l'espace de recherche et la complexité de calcul. Dans [58] les techniques de génération des sous-ensembles sont regroupées en trois catégories : génération complète, heuristique et aléatoire. Nous allons les passer en revue dans la suite de cette section.

II.2.2.1 Génération complète

Les procédures de génération complètes effectuent une recherche complète pour trouver le sous-ensemble de variables optimal au sens de la mesure d'évaluation choisie. Une méthode est dite complète ou exacte si elle assure de retourner toujours le sous-ensemble optimal. Il est important de distinguer une recherche complète d'une recherche exhaustive. En effet, une recherche exhaustive est toujours complète puisque qu'elle consiste à parcourir tous les sous-ensembles possibles. Ainsi, le meilleur sous-ensemble est toujours évalué et donc choisi. En revanche, la réciproque est fausse : dans certains cas, une recherche complète n'est pas exhaustive. Par exemple, si la mesure d'évaluation est monotone, nous n'aurons pas besoin de

regarder tous les sous-ensembles (donc d'être exhaustif) pour retourner le sous-ensemble optimal. La recherche complète est donc encore coûteuse [53], mais elle évalue toujours moins de sous-ensembles que la recherche exhaustive [59].

II.2.2.2 Génération heuristique

Également appelée séquentielle (*greedy*), ce type de génération regroupe les algorithmes itératifs pour lesquels chaque itération permet de sélectionner ou rejeter une ou plusieurs variables. Ces algorithmes sont simples et rapides. Cependant ils ne permettent de parcourir généralement qu'un petit sous-espace de l'espace total des possibilités. Trois catégories d'algorithmes peuvent être envisagées [51]:

1. *Forward* : cette approche part d'un ensemble de variables vide auquel, à chaque itération sont ajoutées une ou plusieurs variables. Elle est également appelée approche ascendante.
2. *Backward* : c'est l'approche inverse ; l'ensemble total des variables est considéré au départ de la procédure itérative, chaque itération permet d'en supprimer. Une autre appellation est approche descendante.
3. *Stepwise* : cette dernière approche est une version hybride des deux précédentes car elle consiste à ajouter ou retirer successivement des variables à l'ensemble déjà sélectionné.

II.2.2.3 Génération aléatoire

Au lieu d'effectuer des recherches exhaustives, ces techniques ne parcourent qu'une partie de l'espace des solutions. Chaque procédure de génération aléatoire nécessite le choix de différents paramètres. Un choix judicieux de ces paramètres est nécessaire pour l'obtention de bons résultats à l'aide de ces techniques. Les algorithmes génétiques font partie de cette catégorie.

En conclusion, la procédure de recherche associée à la sélection de variables tient une place importante. En effet, il est quasiment impossible d'effectuer une recherche exhaustive de l'espace des solutions. La technique de recherche choisie conditionnera donc le sous-espace exploré et l'optimalité de la solution finale [53].

II.2.3 ÉVALUATION DES CARACTERISTIQUES

A tout moment, dans la procédure de recherche choisie, il est nécessaire de mesurer quantitativement la qualité d'un sous-ensemble généré. La fonction (ou critère) d'évaluation doit permettre de mesurer la qualité d'une variable ou d'un sous-ensemble de variables pour

expliquer la variable que l'on cherche à comprendre. Il est important de mentionner qu'un sous-ensemble de caractéristiques est optimal uniquement par rapport au critère d'évaluation. De ce fait, le choix de ce critère est très important. Les mesures généralement utilisées sont les mesures de consistance, les mesures de précision et les mesures basées sur l'information mutuelle [59]. Mais il existe d'autres possibilités de mesures comme les distances ou mesures de similarités, de dépendance ou d'erreur de classification.

II.2.3.1 Distances et mesures de similarités

Ces mesures sont également appelées mesures de séparabilité ou de discrimination. Dans le cas d'un problème à deux classes, une variable X est préférée par rapport à Y, si X introduit une plus grande différence entre les probabilités conditionnelles des deux classes que Y. Si la différence est zéro, alors X et Y sont indistinguables. Un exemple de ce critère est la distance euclidienne.

II.2.3.2 Mesures d'information

Ces mesures déterminent le gain d'information que peut apporter une variable. Ce gain est défini pour une variable X comme la différence entre l'incertitude *a priori* et celle *a posteriori*, c'est-à-dire avant et après sélection de la variable X. La variable sélectionnée parmi deux variables alternatives, correspond à celle qui apporte le plus d'information. L'information mutuelle est un critère appartenant à cette catégorie.

L'information mutuelle est une mesure issue de la théorie de l'information. Elle mesure a la fois l'information qu'apporte une variable aléatoire sur une autre et la réduction d'incertitude sur une variable aléatoire grâce à la connaissance d'une autre [59]. Elle se note généralement I et sera détaillée dans le chapitre III.

II.2.3.3 Mesures de dépendance

Les mesures de dépendance (ou mesures de corrélation) permettent de qualifier la capacité de prédire la valeur d'une variable à partir d'une autre. Ce critère peut être utilisé en classification pour mesurer la corrélation entre une caractéristique X et la variable classe C. Pae exemple, si la corrélation entre X et C est plus importante que la corrélation entre une caractéristique Y et C alors X sera sélectionnée [53].

II.2.3.4 Mesures de précision

Les mesures de précision sont utilisées lorsque l'on définit *a priori* un modèle des données. Dans ce cas là, la sélection de variables sert à optimiser le processus en simplifiant les calculs

par une diminution du nombre de variables à prendre en compte dans le modèle. En théorie, un algorithme d'apprentissage doit être doté d'une mesure de précision permettant d'évaluer la qualité du modèle construit. Les mesures de précision sont généralement utilisées avec un algorithme de type recherche séquentielle descendante. A chaque étape, une variable est enlevée et l'on vérifie que le modèle est toujours suffisamment précis. Ce type d'algorithme s'arrête soit quand il n'y a plus de variables à enlever soit quand la précision est jugée trop mauvaise. Une mesure de précision classique est le taux d'erreur des données reconstruites par le modèle [60], c'est-a-dire le nombre d'individus (observations, formes) mal classés [59]. Les algorithmes de sélection de caractéristiques utilisant cette mesure sont appelés *Wrapper*.

Ainsi, les algorithmes de sélection de caractéristiques sont classés en deux groupes, en fonction du critère d'évaluation utilisé [61] [62]: *Filter Model* et *Wrapper Model*. Cette séparation est basée sur la dépendance ou non de l'algorithme de sélection de variables avec le modèle choisit pour les données. Les méthodes filtres sont indépendantes du modèle. Elles retournent un ensemble de variables qui peut être utilisé pour construire n'importe quel modèle de données. Par opposition, les méthodes *Wrapper* utilisent le modèle de données que l'on cherche à construire pour évaluer la qualité d'un sous-ensemble. Elles retournent donc le sous-ensemble optimal pour un modèle donnée, comme par exemple un réseau de neurones [63] ou un modèle de Markov caché HMM.

Les résultats obtenus par les algorithmes *Wrappers* sont fiables mais au prix d'un temps de calcul important [51]. Les algorithmes Filtres n'ont pas les défauts des *Wrappers*. Ils sont beaucoup plus rapides, ils permettent de mieux comprendre les relations de dépendance entre variables. Mais, comme ils ne prennent pas en compte les algorithmes de classification, les sous-ensembles de variables générés donnent un taux de reconnaissance plus faible.

II.2.4 CRITERE D'ARRET

Les techniques de sélection fondées sur des méthodes de recherche heuristique ou aléatoire ont besoin d'un critère d'arrêt pour éviter une recherche exhaustive des sous-ensembles.

Le critère d'arrêt peut être lié à la procédure de recherche ou bien à la mesure d'évaluation [55]. Dans le premier cas, le critère d'arrêt est soit la taille prédéfinie du sous-ensemble à sélectionner, soit un nombre fixe d'itérations de l'algorithme de sélection de variables. Dans le deuxième cas, un critère d'arrêt lié à la mesure d'évaluation est soit une différence de qualité entre deux ensembles non significative (l'ajout ou la suppression d'une variable n'améliore pas la qualité du sous-ensemble), soit un seuil pour la fonction d'évaluation à atteindre [61].

Dans le cas des méthodes de sélection séquentielle, le critère d'arrêt est fortement conditionné par la mesure de pertinence des variables. La recherche est arrêtée lorsqu'aucune des variables restantes n'est jugée pertinente au sens du critère d'évaluation.

II.2.5 PROCEDURE DE VALIDATION

Dans [53], les auteurs proposent d'ajouter une quatrième composante à un algorithme de sélection de caractéristiques : une procédure de validation. Cette procédure est fonction de la nature des données utilisées, synthétiques ou réelles.

Généralement, une base de données synthétique est construite dans le but de tester un concept ou une application particulière. De ce fait les variables pertinentes sont connues et identifiées. La validation d'un algorithme sera alors directe puisqu'il suffit de vérifier si le sous-ensemble retenu correspond bien aux variables pertinentes.

Dans le cas de données réelles, les variables pertinentes ne sont généralement pas connues. La procédure consiste alors à évaluer la précision de la classification obtenue avec le sous-ensemble de variables sélectionnées par l'intermédiaire d'un classifieur (classifieur de Bayes). Cette dernière peut alors être comparée à d'autres approches ou à celle obtenue par des techniques classiques.

II.3 LES BASES DE LA THEORIE DE L'INFORMATION

Nous présentons dans cette section quelques notions de base de théorie de l'information nécessaires pour comprendre le principe de fonctionnement des méthodes de sélection des paramètres pertinents fondées sur l'information mutuelle. Les théorèmes ou les lemmes sont extraits de [64] et ne seront pas démontrés.

II.3.1 L'ENTROPIE DE SHANNON

En (1948) Shannon avait proposé initialement le concept d'entropie qui est une mesure de l'incertitude d'une variable aléatoire [65].

Soit X une variable aléatoire discrète définie sur un alphabet $\mathcal{L}x$ et décrite par la mesure de probabilité $p(x)=\Pr\{X=x\}$, $x \in \mathcal{L}x$. On dénote $p(x)$ plutôt que $p_X(x)$ pour la commodité. Donc $p(x)$ et $p(y)$ se réfèrent à deux variables différentes et elles représentent deux mesures de probabilité différentes $p_X(x)$ et $p_Y(y)$, respectivement.

Définition II.1 : l'entropie $H(X)$ d'une variable aléatoire discrète X est définie par:

$$H(X) = -\sum_{x \in \mathcal{L}x} p(x) \log\big(p(x)\big) \qquad \text{(II. 1)}$$

On peut également écrire la quantité ci-dessus sous la forme $H(p)$. Le log est à la base 2 (base par défaut) et l'entropie est exprimée en bit. On utilise la convention $0 \log 0 = 0$. Donc, ajouter des termes de probabilité égale à 0 ne change pas l'entropie. Si la base du logarithme est b, on note l'entropie par $H_b(X)$. Si le logarithme népérien est utilisé, alors l'entropie est exprimée en Nat. Notons que l'entropie dépend seulement de la fonction de probabilité $p(x)$ de la variable aléatoire X.

L'entropie de Shannon $H(X)$ s'interprète aussi comme l'espérance mathématique de $\log\left(\frac{1}{p(x)}\right)$:

$$H(X) = E_p\left(\log\left(\frac{1}{p(X)}\right)\right) \quad \text{(II. 2)}$$

Théorème II.1 (*Entropie*): L'entropie de Shannon est bornée :

$$H(X) \leq \log(|\mathcal{L}x|) \quad \text{(II. 3)}$$

avec une égalité si et seulement la variable aléatoire p_X suit une distribution uniforme sur $\mathcal{L}x$.

$|\mathcal{L}x|$ désigne le nombre d'éléments de l'alphabet $\mathcal{L}x$.

La définition de l'entropie d'une seule variable aléatoire s'étend naturellement à une paire de variables aléatoires puisqu'un couple de variables aléatoires (X,Y) peut être considéré comme une variable aléatoire évaluée par un vecteur simple [64].

Définition II.2 (*entropie conjointe*): l'entropie conjointe $H(X,Y)$ d'un couple de variables aléatoires discrètes (X,Y) ayant une distribution de probabilité $p(x,y)$ est définie par :

$$H(X,Y) = -\sum_{x\in\mathcal{L}x}\sum_{y\in\mathcal{L}y} p(x,y)\log\big(p(x,y)\big) \quad \text{(II. 4)}$$

Ce qui donne sous la forme d'une espérance :

$$H(X,Y) = -E_{p(x,y)}\big(\log\big(p(X,Y)\big)\big) \quad \text{(II. 5)}$$

Définition II.3 (*entropie conditionnelle*): On définit également l'entropie conditionnelle $H(X\backslash Y)$ d'une variable aléatoire X étant donnée une autre variable aléatoire Y comme :

$$H(X\backslash Y) = -\sum_{y\in\mathcal{L}y} p(y)\, H(X\backslash Y = y) \quad \text{(II. 6)}$$

$$H(X\backslash Y) = -E_{p(x,y)}\big(\log\big(p(X\backslash Y)\big)\big) \quad \text{(II. 7)}$$

Remarque : $H(X\backslash Y) \neq H(Y\backslash X)$

Théorème II.2 : (*la règle de chaînage pour l'entropie*) : soient $X_1, X_2, \ldots, X_n$ n variables aléatoires discrètes décrites par une fonction de probabilité conjointe $p(x_1, x_2, \ldots, x_n)$. Alors :

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i \setminus X_{i-1}, \ldots, X_1) \quad \text{(II. 8)}$$

Une application directe de cette règle de chaînage est le théorème suivant :

Théorème II.3 (*Majoration de l'entropie*) : soient $X_1,X_2,\ldots,X_n$ n variables aléatoires discrètes, alors :

$$H(X_1,\ldots,X_n) \leq \sum_{i=1}^{n} H(X_i) \qquad \text{(II. 9)}$$

Avec égalité si et seulement si les variables X_i sont indépendantes.

II.3.1.1 Propriétés de l'entropie:

1. L'entropie de Shannon est toujours positive :

$$H(X) \geq 0 \qquad \text{(II. 10)}$$

2. L'entropie conditionnelle est majorée :

$$H(X\backslash Y) \leq H(X) \qquad \text{(II. 11)}$$

Avec égalité si et seulement si X et Y sont indépendantes.

3. L'entropie conditionnelle se généralise et en considérant trois variables aléatoires X,Y et Z, on a :

$$H(X,Y\backslash Z) = H(X\backslash Z) + H(Y\backslash X,Z) \qquad \text{(II. 12)}$$

4. $H_b(X) = log_b(a)\, H_a(X)$ (II. 13)

Cette propriété permet de changer la valeur de l'entropie d'une base en une autre.

II.3.2 L'ENTROPIE RELATIVE

L'entropie d'une variable aléatoire est une mesure de l'incertitude que l'on a sur cette variable aléatoire. Elle permet aussi de mesurer la quantité d'information nécessaire pour décrire la variable aléatoire.

L'entropie relative est un concept directement liée à cette notion d'information. En effet, elle mesure la distance entre deux distributions de probabilité. En statistiques, l'entropie relative $D(p \parallel q)$ permet de mesurer l'inefficacité causée par l'hypothèse que la distribution de la variable aléatoire X est q_X alors que la vraie distribution est p_X.

Définition II.4 (*Entropie relative***):** L'entropie relative ou divergence de Kullback-Leibler entre deux distributions de probabilités p_X et q_X est définie par :

$$D(p \parallel q) = \sum_{x \in \mathcal{L}x} p(x) \log\left(\frac{p(x)}{q(x)}\right) \qquad \text{(II. 14)}$$

$$= E_p\left[\log\left(\frac{p(X)}{q(X)}\right)\right] \qquad \text{(II. 15)}$$

Avec la convention $0\ log\frac{0}{q} = 0\ et\ 0\ log\frac{p}{0} = \infty$

Théorème II.4 (*Inégalité de l'information***):** l'entropie relative est toujours positive :

$$D(p \parallel q) \geq 0 \tag{II.16}$$

Et elle est nulle si et seulement si $p=q \ \ \forall x \in \mathcal{L}x$

Théorème II.5 : (*Règle de chaînage pour l'entropie relative*)

$$D(p(x,y) \parallel q(x,y)) = D(p(x) \parallel q(x)) + D(p(y\backslash x) \parallel q(y\backslash x)) \tag{II.17}$$

Définition II.5 (*entropie relative conditionnelle***):** L'entropie relative conditionnelle est la moyenne des entropies relatives entre les distributions de probabilités conditionnelles pondérée par la distribution de probabilité *p(y)* :

$$D(p(x\backslash y) \parallel q(x\backslash y)) = \sum_{y \in \mathcal{L}y} p(y) \sum_{x \in \mathcal{L}x} p(x\backslash y) \log\left(\frac{p(x\backslash y)}{q(x\backslash y)}\right) \tag{II.18}$$

$$= E_{p(x,y)}\left[\log\left(\frac{p(X\backslash Y)}{q(X\backslash Y)}\right)\right] \tag{II.19}$$

Corollaire II.1: L'entropie relative conditionnelle est toujours positive :

$$D(p(x\backslash y) \parallel q(x\backslash y)) \geq 0 \tag{II.20}$$

Et elle est nulle si et seulement si :

$$p(x\backslash y) = q(x\backslash y) \qquad \forall x \in \mathcal{L}x \text{ et } \ \forall y \in \mathcal{L}y \ avec \ \ p(x) > 0$$

II.3.3 L'INFORMATION MUTUELLE

L'information mutuelle est une mesure de la quantité d'information qu'apporte une variable aléatoire sur une autre. Elle représente la réduction de l'incertitude d'une variable aléatoire due à la connaissance d'une autre. Généralement elle est utilisée comme une mesure de dépendance statistique entre deux variables aléatoires.

Définition II.6 (*Information mutuelle*) : Soient deux variables aléatoires discrètes *X* et *Y* ayant une fonction de probabilité conjointe *p*(*x,y*), et des fonctions de probabilité marginales *p*(*x*) et *p*(*y*). L'information mutuelle entre *X* et *Y*, *I*(*X; Y*) est l'entropie relative entre la distribution conjointe *p(x,y)* et le produit des distributions marginales *p*(*x*) *p*(*y*):

$$I(X;Y) = \sum_{x \in \mathcal{L}x} \sum_{y \in \mathcal{L}y} p(x,y) \log\left(\frac{p(x,y)}{p(x).p(y)}\right) \tag{II.21}$$

$$= D\big(p(x,y) \parallel p(x).p(y)\big) \tag{II.22}$$

$$= E_{p(x,y)}\left[\log\left(\frac{p(X,Y)}{p(X).p(Y)}\right)\right] \tag{II.23}$$

Théorème II.6 : (*Information mutuelle et entropie*)

$$I(X;Y) = H(X) - H(X\backslash Y) \tag{II.24}$$

$$I(X;Y) = H(Y) - H(Y\backslash X) \tag{II.25}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{II.26}$$

$$I(X;Y) = I(Y;X) \tag{II.27}$$

$$I(X;X) = H(X) \tag{II.28}$$

L'équation (II.28) nous indique que l'information mutuelle d'une variable aléatoire avec elle même est égale à l'entropie de cette variable. C'est pour cette raison que l'entropie est parfois désignée comme la ″self-information″ d'une variable aléatoire.

Les relations entre $H(X)$, $H(Y)$, $H(X,Y)$, $H(X\backslash Y)$, $H(Y\backslash X)$ et $I(X;Y)$ peuvent être illustrées par le diagramme de Venn représenté dans la figure (II.2) :

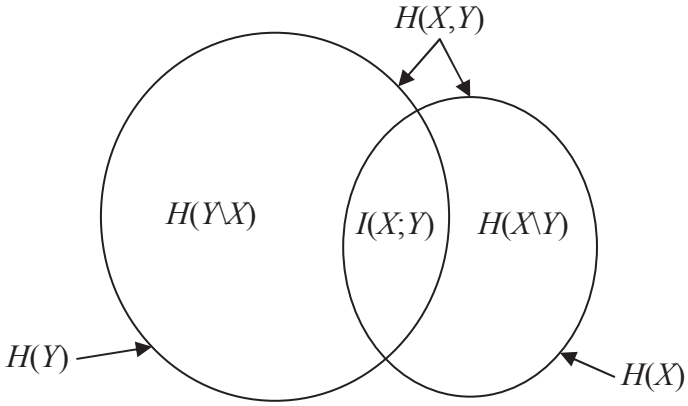


Figure.II.2 : relations entre l'entropie et l'information mutuelle

On note que l'information mutuelle correspond à l'intersection de l'information de X avec celle de Y.

Définition 2.7 (*Information mutuelle conditionnelle*) : L'information mutuelle conditionnelle de deux variables X,Y compte tenu de la variable Z est donnée par :

$$I(X;Y\backslash Z) = H(X\backslash Z) - H(X\backslash Y,Z) \tag{II.29}$$

$$= E_{p(x,y,z)}\left[\log\left(\frac{p(X,Y\backslash Z)}{p(X\backslash Z).p(Y\backslash Z)}\right)\right] \tag{II.30}$$

L'information mutuelle satisfait la règle de chainage.

Théorème II.7 : (*Règle de chaînage pour l'information mutuelle*)

Soient $X_1,X_2,\ldots,X_n$ n variables aléatoires discrètes décrites par une fonction de probabilité conjointe $p(x_1,x_2,\ldots,x_n)$, alors :

$$I(X_1,\ldots,X_n;Y) = \sum_{i=1}^{n} I(X_i\,;Y\backslash\, X_{i-1},\ldots,X_1) \tag{II.31}$$

Propriétés de l'information mutuelle :

1. L'information mutuelle entre deux variables aléatoires X et Y est positive:

$$I(X;Y) \geq 0 \tag{II.32}$$

   Elle est nulle si et seulement si X et Y sont indépendantes.

2. Considérons trois variables aléatoires, X, Y et Z, alors :

$$I(X;Y\backslash Z) \geq 0 \tag{II.33}$$

avec l'égalité si et seulement si X et Y sont indépendantes conditionnellement à Z.

II.3.4 L'INFORMATION MUTUELLE MULTIVARIEE (IMV)

L'information mutuelle est un bon indicateur de dépendance entres deux variables [64]. Cependant, plusieurs problèmes dans la théorie de l'information exigent la connaissance de l'interaction entre plus de deux variables. En particulier, les méthodes employées en analyse multivariée nécessitent d'évaluer de telles interactions. L'évaluation de l'information mutuelle entre deux variables est à ce jour bien établie. En revanche, cette évaluation dans le cas multivarié l'est beaucoup moins car délicate et bien plus lourde car on a besoin d'estimer des densités de grande dimension.

Les premiers travaux sur l'analyse d'information théorique de l'interaction entre plus de deux variables ou en d'autres termes, l'Information Mutuelle Multivariée (IMV), ont été présentés par McGill [66]. La définition de l'IMV ou K-information [67] a été étendue au cas général (au-dessus de trois variables) par Fano [68] et reformulée en une structure en treillis par Han [69].

Dans cette section, nous introduisons une nouvelle notation qui aide à généraliser l'information mutuelle à un ensemble de variables. L'information mutuelle est indicée par le nombre de variables aléatoires qu'elle prend en paramètres. Par exemple, l'information mutuelle classique I entre deux variables aléatoires s'écrira I_2. La généralisation pour K ≥ 2 variables aléatoires est notée I_K.

Définition 2.8 (*Information mutuelle entre trois variables aléatoires*). L'information mutuelle entre trois variables aléatoires X, Y et Z est définie par :

$$I_3(X;Y;Z) = I_2(X;Y) - I_2(X;Y\backslash Z) \tag{II.34}$$

$$= H(X) + H(Y) + H(Z) - H(X,Y) - H(X,Z) - H(Y,Z) + H(X,Y,Z) \tag{II.35}$$

L'IMV de trois variables peut être représentée par le diagramme de Venn (figure II.3).

Définition 2.9 (*Information mutuelle multivariée*)

Soient K ≥ 2 et $X_1, X_2, \ldots, X_r$, K variables, l'information mutuelle multivariée IMV d'ordre K entre ces variables est définie par [69]:

$$I_K(X_1;X_2;\ldots;X_K) = \sum_{k=1}^{\mathrm{K}} (-1)^{k+1} \sum_{\substack{S\subseteq\{\mathrm{X}_1,\mathrm{X}_2,\ldots,\mathrm{X}_\mathrm{K}\} \\ |S|=k}} H(S) \tag{II.36}$$

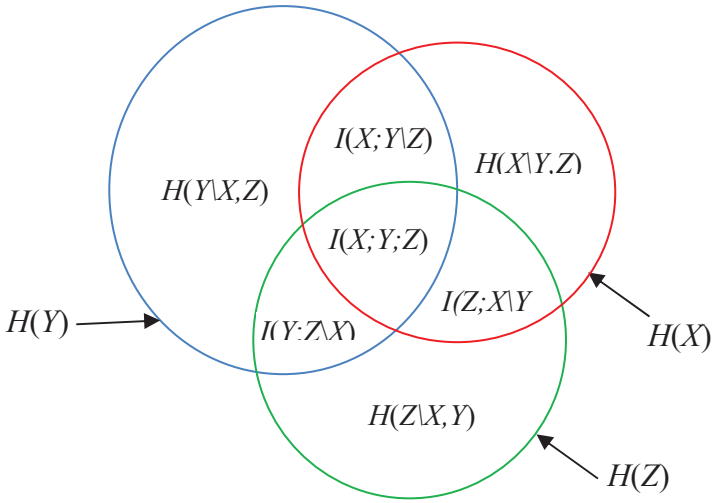


Figure.II.3 : Information mutuelle multivariée de trois variables

De même, l'IMV conditionnelle est définie par :

Définition 2.10 (*Information mutuelle multivariée conditionnelle*)

$$I_K(X_1;X_2;\dots;X_K/Y)=\sum_{k=1}^{K}(-1)^{k-1}\sum_{\substack{S\subset\{X_1,X_2,\dots,X_K\}\\|S|=k}}H(S/Y) \qquad (II.37)$$

Han [69] a montré que l'information mutuelle multivariée n'est pas toujours positive. Une IMV positive entre deux variables explicatives et une variable à expliquer, signifie que ces variables sont redondantes alors qu'une IMV négative indique que ces variables sont synergiques.

Propriétés de l'information mutuelle multivariée:

1. **La récursivité :**

Fano [68] définit l'IMV à K variables comme une extension de l'information mutuelle de deux variables. Il aboutit à la propriété de récursivité suivante :

$$I_K(X_1;X_2;\dots;X_K)=I_{K-1}(X_1;X_2;\dots;X_{K-1})-I_{K-1}(X_1;X_2;\dots;X_{K-1}\backslash X_K) \qquad (II.38)$$

Cette dernière expression s'écrit, pour deux et trois variables :

$$I_2(X_1;X_2)=I_1(X_1)-I_1(X_1\backslash X_2)$$
$$=H(X_1)-H(X_1\backslash X_2)$$
$$I_3(X_1;X_2,X_3)=I_2(X_1;X_2)-I_2(X_1;X_2\backslash X_3)$$

où I_1 est l'entropie H (la "self-information").

2. **La règle de chaînage :**

Une propriété importante de l' information mutuelle multivariée est la règle de chaînage, définie par [67] :

$$I_{K+1}(Y_1; Y_2; \dots; Y_K; X_1, X_2, \dots, X_K) = \sum_{i=1}^{K} I_{K+1}(Y_1; Y_2; \dots; Y_K; X_i \backslash X_{i-1}, \dots, X_1) \quad \text{(II. 39)}$$

II.4 METHODES DE SELECTION FONDEES SUR L'INFORMATION MUTUELLE

Dans la mise en place d'une stratégie de sélection des caractéristiques, il est nécessaire d'accéder à une mesure d'évaluation ou de pertinence des caractéristiques qui permet de chiffrer leur importance dans la tâche de classification. Dans le cas d'une modélisation non-linéaire, l'information mutuelle (IM) est souvent utilisée comme critère de pertinence de variables. En effet, l'IM $I(C;Y)$ entre une variable (caractéristique) Y et la variable index de classes C représente la réduction de l'incertitude de C apportée par la connaissance de Y. Elle peut facilement être étendue à des groupes de variables, ce qui est essentiel dans des procédures de sélection de type *greedy* (procédures itératives *forward*, *forward-backward*, etc.).

Dans cette section, nous présentons les algorithmes les plus connus de la sélection des caractéristiques [60] fondés sur l'IM. Le but est de sélectionner, d'un ensemble F de n caractéristiques $\{Y_1, Y_2, \dots, Y_n\}$, un sous ensemble S de k caractéristiques $\{Y_{P_1}, Y_{P_2}, \dots, Y_{P_k}\}$, tel que S retient tout où la plupart de l'information dans F pour une tâche de classification [70].

Puisque le nombre des sous-ensembles possibles est très grand (combinaison C_k^n) pour permettre le traitement de chaque candidat, ceci conduit aux algorithmes itératifs « greedy » qui sélectionnent les caractéristiques une par une selon une mesure.

Utilisant l'inégalité de Fano [68], la probabilité minimale P_E d'une estimation incorrecte de l'index de classe C par le sous ensemble de caractéristiques S est bornée inférieurement par :

$$P_E \geq \frac{H(C|\boldsymbol{S}) - 1}{log(N)} = \frac{H(C) - I(C;\boldsymbol{S}) - 1}{log(N)} \quad \text{(II. 40)}$$

Selon l'inégalité de Hellman-Raviv [71], elle est bornée supérieurement par:

$$P_E \leq \frac{1}{2} H(C\backslash\boldsymbol{S}) = \frac{H(C) - I_2(C;\boldsymbol{S})}{2} \quad \text{(II. 41)}$$

où $\boldsymbol{S}$ est le vecteur aléatoire dont les composantes sont les éléments de S et N est le nombre de classes.

Puisque l'entropie de l'index de classe $H(C)$ et le nombre de classes N sont fixés, la limite inférieure de P_E est minimisée quand l'information mutuelle $I(C;\boldsymbol{S})$ entre l'index C et l'ensemble des caractéristiques de S, devient maximale. La recherche d'une bonne méthode de sélection des caractéristiques correspond donc à chercher celle qui maximise $I(C;\boldsymbol{S})$ soit :

$$\boldsymbol{S_{opt}} = arg \max_{\boldsymbol{S} \in F} I_2(C;\boldsymbol{S})$$

Dans [72], Battiti formalise ce concept de sélection des k caractéristiques les plus pertinentes de l'ensemble F des n caractéristiques, en problème FRn-k et adopte le mode opératoire de sélection *greedy* pour résoudre ce problème.

La procédure adoptée par Battiti commence par considérer un sous-ensemble de caractéristiques vide. A chaque itération, on ajoute la meilleure caractéristique restante au sous-ensemble des caractéristiques déjà sélectionnées, jusqu'à ce que la taille du sous-ensemble atteigne k.

L'algorithme idéal *greedy* de sélection utilisant l'IM est réalisé comme suit :

1. (initialisation), ensemble $F \leftarrow$ l'ensemble des n caractéristiques, $S \leftarrow$ le sous ensemble vide.
2. (Calcul de l'IM), $\forall\ Y_i \in F$, calculer $I(C;Y_i)$.
3. (sélection de la première caractéristique Y_{P_1}), trouver la caractéristique qui maximise $I(C;Y_i)$,

   affecter $F \leftarrow F\text{-}\{Y_{P_1}\}$, $S \leftarrow \{Y_{P_1}\}$.
4. (sélection greedy), répéter jusqu'au nombre désiré des caractéristiques :
   a. (calculer les IM conjointes entre la variable index C et les caractéristiques) : $\forall\ Y_i \in F$, calculer $I(C;S,Y_i)$.
   b. (sélection de la caractéristique suivante Y_{P_j}) choisir la caractéristique $Y_i \in F$, qui maximise $I(C;S,Y_i)$ à l'étape j et effectuer : $F \leftarrow F\text{-}\{Y_{P_j}\}$, $S \leftarrow S \cup \{Y_{P_j}\}$.
5. Faire sortir le sous-ensemble des caractéristiques sélectionnées S.

Le calcul de l'IM à partir des données nécessite l'estimation de la densité de probabilité, qui ne peut pas être précise pour des grandes dimensions. De ce fait, la majorité des algorithmes utilise des mesures basées au maximum sur trois variables (deux caractéristiques plus l'index des classes).

Nous présentons les méthodes les plus connues de la sélection des caractéristiques fondées sur des critères d'information mutuelle pour lesquels l'ordre est limité au maximum à 3 [70] [73] [74].

Le critère le plus simple de sélection d'une caractéristique à l'étape j+1 est [75] [60] :

$$Y_{P_{j+1}} = arg \max_{Y_i \in F - S_j} I(C;Y_i) \qquad (II.42)$$

où $S_j = S_{j-1} \cup \{Y_{P_j}\}$ est le sous ensemble des caractéristiques sélectionnées à l'étape j.

Cependant, cette méthode ordonne la pertinence de chaque caractéristique individuellement, sans se soucier des choix antérieurs, car on ne tient pas compte de toutes les caractéristiques déjà sélectionnées. Cette procédure peut conduire à sélectionner des

caractéristiques redondantes (qui partagent la même information avec l'index de classe C). Cette redondance doit cependant être éliminée. Ainsi les algorithmes proposés justifient leur utilisation par différents arguments, tous avec l'idée d' «augmenter la pertinence et diminuer la redondance ».

Battiti (1994) [72] propose le critère MIFS (Mutual Information-Based Feature Selection) :

$$Y_{P_{j+1}} = arg \max_{Y_i \in F - S_j} \left[I(C; Y_i) - \beta \sum_{k=1}^{j} I\left(Y_i; Y_{P_k}\right) \right] \quad (II.43)$$

Cette formule inclut le terme $I(C; Y_i)$ qui assure la pertinence du paramètre, mais introduit une approximation de la redondance sous forme de pénalité $\beta\, I\left(Y_i; Y_{P_j}\right)$. β est un paramètre configurable, qui doit être fixé expérimentalement [73].

Peng et al. (2005) [76] proposent le critère MRMR (Maximum-Relevance Minimum-Redundancy) :

$$Y_{P_{j+1}} = arg \max_{Y_i \in F - S_j} \left[I(C; Y_i) - \frac{1}{j} \sum_{k=1}^{j} I\left(Y_i; Y_{P_k}\right) \right] \quad (II.44)$$

Il est clair que le critère MRMR est équivalent au critère MIFS avec $\beta = \frac{1}{j}$.

Yang and Moody (1999) [77] proposent l'utilisation de l'information mutuelle conjointe JMI (Joint Mutual Information).

$$Y_{P_{j+1}} = ag \max_{Y_i \in F - S_j} \left[\sum_{k=1}^{j} I(C; Y_i, Y_{P_k}) \right] \quad (II.45)$$

$$= arg \max_{Y_i \in F - S_j} \left[\sum_{k=1}^{j} \left[I\left(C; Y_{P_k}\right) + I(C; Y_i \backslash Y_{P_k}) \right] \right]$$

Le terme $\sum_{k=1}^{j} I\left(C; Y_{P_k}\right)$ dans ce qui précède est constant en tenant compte de l'argument Y_i. Ainsi ce terme peut être ignoré. Donc, le critère peut être réduit à :

$$Y_{P_{j+1}} = arg \max_{Y_i \in F - S_j} \left[\sum_{k=1}^{j} I(C; Y_i \backslash Y_{P_k}) \right]$$

$$= arg \max_{Y_i \in F - S_j} \left[\sum_{k=1}^{j} \left[I(C; Y_i) - I\left(C; Y_i; Y_{P_k}\right) \right] \right]$$

$$= arg \max_{Y_i \in F - S_j} \left[j\,.\, I(C; Y_i) - \sum_{k=1}^{j} I\left(C; Y_i; Y_{P_k}\right) \right]$$

En divisant les termes de la dernière formule par j, le critère peut être réduit à :

$$Y_{P_{j+1}} = arg \max_{Y_i \in F - S_j} \left[I(C;Y_i) - \frac{1}{j}\sum_{k=1}^{j} I(Y_i;Y_{P_k};C) \right]$$

$$= arg \max_{Y_i \in F - S_j} \left[I(C;Y_i) - \frac{1}{j}\sum_{k=1}^{j} \left[I\left(Y_i;Y_{P_k}\right) - I(Y_i;Y_{P_k}\backslash C) \right] \right] \qquad \text{(II. 46)}$$

La stratégie JMI, d'une autre manière, ajoute un terme de pénalité qui prend en compte la redondance moyenne $\frac{1}{j}\sum_{k=1}^{j} I(Y_i;Y_{P_k};C)$ évaluée entre un paramètre candidat Y_i et chaque paramètre déjà sélectionné Y_{P_k}.

On peut voir la relation entre MRMR et JMI. Le critère JMI équivaut au critère MRMR en ajoutant le terme $-\frac{1}{j}\sum_{k=1}^{j} I(Y_i;Y_{P_k}\backslash C)$.

Kwak et Choi (2002) [78] proposent une amélioration à MIFS, appelé MIFS-U, appropriée aux problèmes où l'information est distribuée uniformément dans l'espace, selon :

$$Y_{P_{j+1}} = arg \max_{Y_i \in F - S_j} \left[I(C;Y_i) - \beta \sum_{k=1}^{j} \frac{I\left(C;Y_{P_k}\right)}{H\left(Y_{P_k}\right)} I\left(Y_i;Y_{P_k}\right) \right] \qquad \text{(II. 47)}$$

En pratique, les auteurs trouvent $\beta = 1$ optimal.

Vidal-Naquet et Ullman (2003) [79] proposent un critère utilisé en vision par ordinateur, appelé IF (*Informative Fragments*) décrit comme suit :

$$Y_{P_{j+1}} = arg \max_{Y_i \in F - S_j} \left[\min_{Y_{P_k} \in S_j} \left[I\left(C;Y_i,Y_{P_k}\right) - I\left(C;Y_{P_k}\right) \right] \right] \qquad \text{(II. 48)}$$

Fleuret (2004) [80] propose le critère CMI (*Conditional Mutual Information*), le plus connu, basé sur la maximisation de l'information mutuelle conditionnelle :

$$Y_{P_{j+1}} = arg \max_{Y_i \in F - S_j} \left[\min_{Y_{P_k} \in S_j} \left[I\left(C;Y_i \backslash Y_{P_k}\right) \right] \right] \qquad \text{(II. 49)}$$

En décomposant $$I\left(C;Y_i\backslash Y_{P_k}\right) = I(C;Y_i) - I\left(C;Y_i;Y_{P_k}\right)$$

$$= I(C;Y_i) - \left[I\left(Y_i;Y_{P_k}\right) - I\left(Y_i;Y_{P_k}\backslash C\right) \right]$$

La formule (II.49) devient :

$$Y_{P_{j+1}} = arg \max_{Y_i \in F - S_j} \left[I(C;Y_i) - \max_{Y_{P_k} \in S_j} \left[I\left(Y_i;Y_{P_k}\right) - I\left(Y_i;Y_{P_k}\backslash C\right) \right] \right] \qquad \text{(II. 50)}$$

En utilisant la règle de chaînage :

$$I\left(C;Y_i,Y_{P_k}\right) = I\left(C;Y_{P_k}\right) + I\left(C;Y_i\backslash Y_{P_k}\right)$$

La formule (II.50) devient :

$$Y_{P_{j+1}} = arg \max_{Y_i \in F - S_j} \left[\min_{Y_{P_k} \in S_j} \left[I\left(C;Y_{P_k}\right) + I\left(C;Y_i\backslash Y_{P_k}\right) - I\left(C;Y_{P_k}\right) \right] \right]$$

$$Y_{P_{j+1}} = arg \max_{Y_i \in F - S_j} \left[\min_{Y_{P_k} \in S_j} \left[I\left(C; Y_i \backslash Y_{P_k}\right) \right] \right]$$

Donc la formule (II.48) est équivalente à la formule (II.49). Ainsi on peut dire que le critère IF est équivalent au critère CMI.

Lin et Tang (2006) [81] proposent un autre critère également utilisé en vision par ordinateur, appelé CIFE (*Conditional Infomax Feature Extraction*) et décrit comme suit:

$$Y_{P_{j+1}} = arg \max_{Y_i \in F - S_j} \left[I(C; Y_i) - \sum_{k=1}^{j} \left[I\left(Y_i; Y_{P_k}\right) - I(Y_i; Y_{P_k} \backslash C) \right] \right] \qquad (II.51)$$

Ce critère est déjà étudié en 2005 par Kojadinovic [82] pour la sélection des variables pour une tâche de régression. Kojadinovic a pu démontrer que le calcul de l'information mutuelle $I(X; Y_1, ..., Y_n)$ entre un ensemble de n variables explicatives et une variable à expliquer X peut être décomposé en termes de signe opposé successivement dont chaque i^{ieme} terme est une somme de différentes informations mutuelles multivariées (IMV) de i-1 variables et la variable X. Une démonstration est faite par Brown [73] dans le cas de sélection des paramètres pour une tâche de classification. Les démonstrations de Kojadinovic et Brown sont basées sur la théorie des fonctions de Mobius [83]. Dans la section suivante, nous présentons notre démonstration du calcul de l'information mutuelle $I(C;Y_1,\ldots,Y_n)$ entre un ensemble de caractéristiques déjà sélectionnées et la variable index de classe en se basant sur les propriétés de l'information mutuelle multivariée [9]. Pour cela, nous rappelons un théorème de décomposition de l'IM entre un ensemble de variables et une autre variable.

Théorème [73]**:**

Etant donné un ensemble de paramètres d'entrée $S=\{X_1,\ldots,X_N\}$ et une variable à expliquer Y, leur information mutuelle de Shannon peut être formulée comme suit :

$$I(X_1, \ldots, X_N; Y) = \sum_{T \subset S} I(\{T \cup Y\}), \qquad |T| \geq 1 \qquad (II.52)$$

En d'autres termes, l'information mutuelle de Shannon entre l'ensemble $\{X_1,...,X_N\}$ et la variable Y se décompose en une somme de termes d'Informations Mutuelle Multivariée IMV. $\sum_{T \subset S}$ devrait être lu, " sommer sur tous les sous-ensembles possibles T tirés de S ". Le terme $I(\{T \cup Y\})$ désigne l'information mutuelle multivariée entre tous les paramètres de T ainsi que Y.

Exemple : comme exemple illustratif pour le cas de 4 variables, l'information mutuelle entre la variable Y et l'ensemble de 3 variables $S=\{X_1, X_2, X_3\}$ peut être écrite comme suit:

$I(X_1, X_2, X_3; Y) = \sum_{T \subset S} I(\{T \cup Y\})$

$$= I(X_1;Y) + I(X_2;Y) + I(X_3;Y) - I_3(X_1;X_2;Y) - I_3(X_1;X_3;Y) \ldots$$
$$-I_3(X_2;X_3;Y) + I_4(X_1,X_2,X_3;Y) \quad \text{(II.53)}$$

Le critère (II.51) peut être déduit en tronquant les termes d'ordre supérieur de (II.52) sous l'hypothèse que les IMV de plus de 3 variables sont négligeables. Ainsi, Kojadinovic a appelé ce critère *K-additive*, alors que Brown a lui donné le nom FOU (*first-order utility*).

Au final, le but de tous ces algorithmes est de maximiser l'IM entre le sous-ensemble S des caractéristiques et la variable index de classes, ce qui peut être étendu comme suit :

$$I(C;\boldsymbol{S}) = I\left(C;Y_{P_1},Y_{P_2},,\ldots,Y_{P_k}\right) \quad \text{(II.54)}$$

$$= \sum_{j=1}^{k} I(C;Y_{P_j}\backslash Y_{P_1},,\ldots,Y_{P_{j-1}}) \quad \text{(II.55)}$$

$$= \sum_{j=1}^{k} \left[I(C;Y_{P_j}) - I(C;Y_{P_j};Y_{P_1},,\ldots,Y_{P_{j-1}})\right] \quad \text{(II.56)}$$

Puisque le Y_{P_j}est le Y_i particulier qui maximise le j[ième] terme de cette somme, tous les critères mentionnés précédemment peuvent être interprétés comme une approximation de l'optimisation générale. Tous maximisent la différence entre $I(C;Y_i)$ et une approximation de la redondance $I(C;Y_i;Y_{P_1},\ldots,Y_{P_{j-1}})$ entre *Yi, S_{j-1}*, et l'index de classe *C*.

II.4.1 METHODE PROPOSEE (TMI)

Notre méthode de sélection diffère des autres méthodes par le développement théorique de la redondance entre Y_i, S_{j-1}, et l'index de classe C (deuxième terme de (II.56)), exprimée par:

$$R_j = I\left(C;Y_i;,\ldots,Y_{P_{j-1}}\right) \quad \text{(II.57)}$$

A l' étape j, on a:

$$Y_{P_j} = arg\,max_{Y_i \in F - S_{j-1}} \left[I_2\left(C;Y_i,Y_{P_1},\ldots,Y_{P_{j-1}}\right)\right] \quad \text{(II.58)}$$

où $I_2\left(C;Y_i,Y_{P_1},\ldots,Y_{P_{j-1}}\right) = I_2(C;Y_i,S_{j-1})$

$$= I_2(C;S_{j-1}) + I_2(C;Y_i\backslash S_{j-1}) \quad \text{(II.59)}$$

Donc, $$Y_{P_j} = arg\,max_{Y_i \in F - S_{j-1}} \left[I_2\left(C;Y_i\backslash Y_{P_1},\ldots,Y_{P_{j-1}}\right)\right] \quad \text{(II.60)}$$

$$= arg\,max_{Y_i \in F - S_{j-1}} \left[I_2(C;Y_i) - I_3\left(C;Y_i;Y_{P_1},\ldots,Y_{P_{j-1}}\right)\right] \quad \text{(II.61)}$$

$$= arg\,max_{Y_i \in F - S_{j-1}} \left[I_2(C;Y_i) - R_j\right]$$

En appliquant la propriété de chaînage (II.39), la redondance R_j s'écrit :

$R_j = \sum_{m=1}^{j-1} I_3\left(C;Y_i;Y_{P_m}\backslash Y_{P_1},\ldots,Y_{P_{m-1}}\right)$

En appliquant la propriété de récursivité (II.38), on a :

$R_j = \sum_{m=1}^{j-1}\left[I_3\left(C; Y_i; Y_{P_m}\right) - I_4\left(C; Y_i; Y_{P_m}; Y_{P_1}, \dots, Y_{P_{m-1}}\right)\right]$

En continuant le développement par application des propriétés de chaînage et de récursivité, on aboutit à la formule suivante :

$$R_j = \sum_{m=1}^{j-1} I_3\left(C; Y_i; Y_{P_m}\right) + \dots$$

$$+(-1)^{r+1} \sum_{j_1=r}^{j-1} \dots \sum_{j_q=r-q+1}^{j_{q-1}-1} \dots \sum_{j_r=1}^{j_{r-1}-1} I_{r+2}\left(C; Y_i; Y_{P_{j_1}}; \dots; Y_{P_{j_q}}; \dots; Y_{P_{j_r}}\right)$$

$$+ \dots + (-1)^j I_{j+1}\left(C; Y_i; Y_{P_{j-1}}; \dots; Y_{P_1}\right) \qquad \text{(II.62)}$$

avec $j \geq 2$ et $r = 1,2, \dots, j-1$

Ainsi, la redondance à l'étape j peut être décomposée en une somme de j-1 termes de signe opposé successivement dont chaque p$^{\text{ieme}}$ terme est une somme de différentes informations mutuelles multivariées (IMV) des (p-2) variables déjà sélectionnées, ainsi que la variable Y_i et la variable index de classe *C*.

De l'équation (II.59), on peut calculer l'information mutuelle $I_2\left(C; Y_i, Y_{P_1}, , \dots, Y_{P_{j-1}}\right)$ à l'étape j par:

$$I_2\left(C; Y_i, Y_{P_1}, \dots, Y_{P_{j-1}}\right) = I_2\left(C; S_{j-1}\right) + I_2\left(C; Y_i \backslash S_{j-1}\right)$$

$$= I_2\left(C; S_{j-1}\right) + I_2(C; Y_i) - R_j \qquad \text{(II.63)}$$

Le calcul de R_j à partir des données nécessite l'estimation des densités de probabilité, qui ne peuvent pas être précises pour des grandes dimensions. Une approximée $\widehat{R}_j$ de R_j est obtenue par la troncature des termes d'ordre supérieur de l'Eq (II.62) sous l'hypothèse que les IMV de plus de L variables sont négligeables.

$$\widehat{R}_j = \sum_{m=1}^{j-1} I_3\left(C; Y_i; Y_{P_m}\right) + \dots +$$

$$(-1)^{r+1} \sum_{j_1=r}^{j-1} \dots \sum_{j_q=r-q+1}^{j_{q-1}-1} \dots \sum_{j_r=1}^{j_{r-1}-1} I_{r+2}\left(C; Y_i; Y_{P_{j_1}}; \dots; Y_{P_{j_q}}; \dots; Y_{P_{j_r}}\right) \qquad \text{(II.64)}$$

Avec $j \geq 2$ et $r = 1,2, \dots, L-2 < j-1$

Des équations (II.57) , (II.61) et (II.64) on peut énoncer un nouveau critère de sélection basé sur la troncature des IMV d'ordre supérieur :

A l'étape j: $Y_{P_j} = arg\, max_{Y_i \in F - S_{j-1}}\left[I_2(C; Y_i) - \widehat{R}_j\right]$ (II.65)

Notre algorithme peut se formaliser comme suit:

1. (initialisation), ensemble F←l'ensemble des n caractéristiques, $S \leftarrow$ le sous ensemble vide.
   (Calcul de l'IM), $\forall\ Y_i \in F$, calculer $I_2(C;Y_i)$.
2. (sélection de la première caractéristique Y_{P_1}), trouver la caractéristique qui maximise $I(C;Y_i)$,
   affecter $F \leftarrow F\text{-}\{Y_{P_1}\}$, $S \leftarrow \{Y_{P_1}\}$, $IMP \leftarrow Max(I(C;Y_i))$
3. (selection à l'étape j):
   a. $\forall\ Y_i \in F$, *calculer* $IMT = I_2(C;Y_i) - \widehat{R}_j$ (IMT est l'approximée de $I(C;Y_i \backslash S)$.
   b. (sélection de la caractéristique Y_{P_j}) choisir la caractéristique $Y_i \in F$, qui maximise IMT et effectuer : $F \leftarrow F\text{-}\{Y_{P_j}\}$, S ← S∪{$Y_{P_j}$}, IMP ← IMP + Max(IMT).
   IMP est l'approximée de $I(C; Y_{P_j}, S)$,
4. Répéter l'étape 3 jusqu'à $IMP \geq \alpha\ H(C)$ ou jusqu'à atteindre un nombre de paramètres désiré.
5. Faire sortir le sous-ensemble des caractéristiques sélectionnées S.

Le critère d'arrêt utilisé dans cet algorithme est soit un nombre désiré de caractéristiques, soit un pourcentage atteint d'entropie par l'IMP.

Remarques :

- IMT est l'approximée de l'information mutuelle conditionnelle entre la variable index de classe C et la caractéristique Y_i connaissant l'ensemble des caractéristique sélectionnées ($\{Y_{P_1}, Y_{P_2},...,Y_{P_{j-1}}\}$) à l'itération j-1.
- IMP est l'approximée de l'information mutuelle conjointe entre la variable index de classe C et l'ensemble des caractéristiques sélectionnées ($\{Y_{P_1}, Y_{P_2},...,Y_{P_j}\}$) à l'itération j.
- Le critère d'arrêt ainsi que le choix de la valeur de α seront détaillés dans le chapitre IV.

II.5 CONCLUSION

Dans la reconnaissance de la parole, plusieurs approches proposées pour améliorer la robustesse et les performances des systèmes de Reconnaissance Automatique de la Parole, consistent à enrichir les vecteurs acoustiques en plus des coefficients initiaux (statiques) par leurs coefficients différentiels (dynamiques). Cependant l'augmentation de la dimension de l'espace acoustique a pour effet d'augmenter les coûts de calcul et d'encombrement mémoire et de dégrader probablement les performances des systèmes RAP (phénomène de la malédiction de la dimensionnalité). Il est donc nécessaire, si l'on veut concevoir un système acceptable en termes de coût de calcul et d'encombrement mémoire, de limiter le nombre de paramètres constituant les vecteurs acoustiques. Cette réduction peut se faire principalement par les techniques d'extraction (ACP, ALD) ou par les techniques de sélection. Les premières

permettent de créer de nouveaux ensembles de caractéristiques, en utilisant une transformation ou une combinaison d'un espace de départ, alors que les techniques de sélection permettent de sélectionner un sous-ensemble de caractéristiques les plus pertinentes parmi un ensemble de départ.

Dans notre travail de thèse on s'est intéressé à la technique de sélection basée sur la théorie de l'information et plus particulièrement celle basée sur le critère de l'information mutuelle. Une nouvelle méthode de sélection fondée sur la théorie de l'information mutuelle est proposée. Cependant, l'estimation de l'information mutuelle à partir des données reste un problème délicat. Ce problème et l'application de l'IM estimée dans la sélection des paramètres pertinents font l'objet du chapitre III.

CHAPITRE III

ESTIMATION DE L'ENTROPIE ET DE L'INFORMATION MUTUELLE

III.1 INTRODUCTION

L'information mutuelle introduite par Shannon [65] est une mesure quantifiant la dépendance statistique entre deux variables malgré la possibilité d'existence d'une relation non linéaire entre ces variables. Elle est largement utilisée, soit comme une mesure de similarité entre des variables aléatoires (paramètres), soit comme une mesure de pertinence pour la sélection des variables dans un contexte de régression [84] ou de classification [85] [86] [87]. Elle est appliquée dans plusieurs domaines scientifiques tels que : le traitement du langage [88], le traitement du signal parole [74] [89] [90], le traitement d'images [91] [92], et le biomédical [93].

Cependant, le calcul de l'information mutuelle à partir des données exige l'estimation des Fonctions de Densités de Probabilités (fdp) conjointes et marginales [64] qui ne sont pas connues en pratique et qui sont estimées à partir d'un nombre d'échantillons fini. L'estimation s'effectue par des méthodes non paramétriques telles que la méthode d'histogramme [94] [95] [96], la méthode à noyau (*Kernel Density Estimation* KDE) [97] [98] [99], ou les méthodes paramétriques dont le modèle de mélange gaussien (*Gaussian mixture model* GMM) [100].

Dans notre travail, on retient l'estimateur de l'information mutuelle fondé sur la méthode d'histogramme pour ses avantages indéniables en termes de simplicité et de complexité de calcul [101] [102] [103]. Cependant, les statistiques de cet estimateur souffrent de la variance et du bais [94] [104] dont les causes sont dépendantes du nombre de données, du choix du nombre de *bins* (cellules) de l'histogramme et du lissage de la fdp. Les deux dernières causes ont une importance dans le cas des variables continues. Le calcul du biais et la variance est bien détaillé dans [94].

Le nombre de *bins* k d'un histogramme est un paramètre important qui doit être choisi soigneusement. La partition de l'histogramme en *bins* peut être adaptative ou uniforme (largeur de *bins* constante) [95] [96]. La partition adaptative de l'espace d'observations pour l'estimation de l'IM est prometteuse mais elle souffre d'un coût de calcul important [95]. Dans le cas de la partition uniforme, il existe différents critères heuristiques pour le choix du

nombre de *bins* qui prennent en considération le nombre d'observations et la fdp de référence gaussienne : Sturges [105], Scott [106], Freedman [107].

Ainsi, notre contribution consiste à estimer le nombre de *bins* permettant d'avoir un estimateur d'information mutuelle moins biaisé en se basant sur l'approximation théorique du biais donné dans [94]. D'autres critères de choix du nombre de *bins* seront présentés également dans la section suivante.

Dans la deuxième partie de ce chapitre, nous proposons d'appliquer ces critères dans la sélection des caractéristiques gaussiennes pour une tâche de classification de données simulées. L'application de ces critères dans la sélection des paramètres acoustiques pour une tâche de reconnaissance de digits sera l'objet du chapitre IV.

III.2 ESTIMATION DE L'ENTROPIE ET DE L'INFORMATION MUTUELLE DES VARIABLES CONTINUES

Dans le chapitre deux ont été présentées des notions de base sur la théorie d'information, notamment le concept de l'entropie et celui de l'information mutuelle pour des variables aléatoires discrètes. Lorsqu'on passe des variables aléatoires discrètes aux variables aléatoires continues, les sommes dans les formules de l'entropie et de l'information mutuelle se remplacent par des intégrales, et les mesures de probabilité se remplacent par des densités de probabilités.

Soit X une variable aléatoire (ou vecteur aléatoire) continue ayant la fdp $f_X(x)$ et la fonction de répartition $F_X(x)$. Rappelons la définition de l'entropie $H(X)$ (appelée aussi entropie différentielle) [64] :

$$H(X) = -\int_{-\infty}^{+\infty} f_X(x).log\big(f_X(x)\big)\,dx \quad nat \qquad \text{(III. 1)}$$

L'information mutuelle $I(X;Y)$ entre deux variables aléatoires (vecteurs aléatoires) continues X, Y ayant respectivement les fdps marginales$f_X(x)$, $f_Y(y)$ ainsi que la fdp conjointe $f_{XY}(x,y)$ est définie théoriquement de la manière suivante [64]:

$$I(X;Y) = I_{XY} = \int_{-\infty}^{-\infty}\int_{-\infty}^{+\infty} f_{XY}(x,y).log\left(\frac{f_{XY}(x,y)}{f_X(x).f_Y(y)}\right)dx\,dy \quad nat \qquad \text{(III. 2)}$$

L'estimation de l'entropie et de l'information mutuelle est un processus en deux étapes : premièrement, la fdp est estimée, ensuite l'entropie ou l'IM est calculée.

En pratique, il est difficile de connaitre exactement les fonctions $f_X(x)$, $f_Y(y), f_{XY}(x,y)$. Ces quantités doivent être estimées. L'approche par histogramme peut être utilisée pour estimer ces fonctions. Quand cette approche est utilisée avec un nombre d'échantillons de données fini, différents types d'erreurs apparaissent. Le calcul du biais, de la variance et de

l'Erreur Quadratique Moyenne (EQM ou *Mean Squared Error MSE*) des estimateurs de l'entropie et de la IM basés sur la méthode d'histogramme sont détaillés dans la section suivante.

En pratique, l'estimateur de l'entropie d'une variable (ou vecteur aléatoire) X représentée par la fdp $\hat{f}_X(x)$ estimée à partir des échantillons $x_1, x_2, \ldots, x_i, \ldots, x_N$ peut être obtenu par [83]-:

$$\widehat{H}(X) = -\frac{1}{N}\sum_{i=1}^{N} log\left(\hat{f}_X(x_i)\right) \qquad \text{(III. 3)}$$

Cet estimateur a été étudié par Joe [97] et Hall et Morton [108] dans le cas où $f_X(x)$ est estimée par une approche à noyau (KDE). Cependant dans notre travail, nous proposons d'estimer la fdp par un GMM afin de réaliser une comparaison avec la méthode par histogramme. Ainsi pour une densité approchée par GMM, l'estimateur de l'entropie est donné comme suit [109] :

$$\widehat{H}_x = -\frac{1}{N}\sum_{i=1}^{N} log\left(\sum_{m=1}^{M} \alpha_m G_m(x_i)\right) \qquad \text{(III. 4)}$$

où $G_m$est la m^{ième} densité gaussienne et α_m est son poids.

Pratiquement l'estimateur de l'IM est obtenu par :

$$\hat{\mathrm{I}}_{\mathrm{XY}} = \frac{1}{N}\sum_{\mathrm{i=1}}^{N} \log\left(\frac{\hat{f}_{XY}(x_i, y_i)}{\hat{f}_X(x_i).\hat{f}_Y(y_i)}\right) \qquad \text{(III. 5)}$$

Dans le cas d'une estimation par les GMM, l'estimateur est donné par :

$$\hat{I}_{XY} = \frac{1}{N}\sum_{i=1}^{N} log\left(\frac{\sum_{m=1}^{M_{xy}} \alpha_m G_m(x_i, y_i)}{\sum_{m=1}^{M_x} \alpha_m G_m(x_i) \,.\, \sum_{m=1}^{M_y} \alpha_m G_m(y_i)}\right) \qquad \text{(III. 6)}$$

Nous présentons dans la section suivante l'estimation de l'entropie et de l'IM par la méthode d'histogramme et une étude plus détaillée est consacrée au choix du nombre de *bins* de l'histogramme ainsi que son influence sur ces estimateurs. De plus, de nouveaux estimateurs de l'IM et de l'entropie à faible biais en s'appuyant sur le travail de Moddemeijer [94] sont proposés.

III.3 ESTIMATION DE L'ENTROPIE ET DE L'IM PAR LA METHODE D'HISTOGRAMME

Dans cette section, les estimations de l'entropie et de l'IM basées sur la méthode par histogramme sont dérivées. Finalement, le biais, la variance et l'EQM de ces estimateurs sont exprimés. Le lecteur intéressé pourra trouver plus de détails dans [101].

III.3.1 ESTIMATION DE L'ENTROPIE PAR LA METHODE D'HISTOGRAMME

L'estimation de l'entropie $H(X)$ basée sur la méthode d'histogramme divise l'axe x en k_x *bins* (segment) de largeur constante Δx et de position i, sur l'intervalle de définition des données (typiquement égale à 6 fois l'écart type de données σ_x autour de la moyenne). L'approximation discrète de (III.1) s'écrit comme suit :

$$H(X) \approx -\sum_{i=1}^{k_x} f_X(x_i) log\big(f_X(x_i)\big)\ \Delta x \qquad \text{(III.7)}$$

où x_i est le centre du i^{ième} *bin*.

Utilisant le fait que $p_i \approx f_x(x_i)\ \Delta x$, la probabilité p_i d'observer un échantillon dans le bin i est estimée par $\hat{p}_i \approx \frac{k_i}{N}$ où k_i est le nombre d'échantillons contenus dans le bin i parmi un nombre total de N données. L'Eq. (III.7) conduit à l'estimateur de H(X) suivant :

$$\widehat{H}(X) = -\sum_{i=1}^{k_x} \frac{k_i}{N} log\left(\frac{k_i}{N}\right) + log\ \Delta x \qquad \text{(III. 8)}$$

Par définition, le biais de $\widehat{H}(X)$ est $E\{\widehat{H}(X) - H(X)\}$. Ce biais peut être décomposé en deux parties : le R_{biais} causé par la représentation insuffisante de la fdp par l'histogramme et le N_{biais} causé par le nombre fini de données.

Concernant le R_{biais} , l' Eq. (III.1) peut être formulée comme suit:

$$H(X) = -\sum_{i=-\infty}^{+\infty} \int_{x_i-\frac{\Delta x}{2}}^{x_i+\frac{\Delta x}{2}} f_X(x_i) log\big(f_X(x_i)\big)\, dx$$

Par ailleurs, l'Eq. (III.8) peut être vue comme une approximation de l'expression suivante:

$$\widetilde{H}(X) = -\sum_{i=-\infty}^{\infty} \int_{x_i-\frac{\Delta x}{2}}^{x_i+\frac{\Delta x}{2}} f_X(x_i)\ dx\ log\left(\int_{x_i-\frac{\Delta x}{2}}^{x_i+\frac{\Delta x}{2}} f_X(x_i) dx\ \right) + log\ \Delta x$$

La différence entre $\widetilde{H}(X)$ et $H(X)$ correspond au R_{biais} défini par :

$$R_{biais} = \widetilde{H}(X) - H(X) \qquad \text{(III. 9)}$$

Le développement de Taylor du deuxième ordre de $f_X(x)$ autour de x_i dans $\widetilde{H}(X)$ et $H(X)$ conduit à [94] :

$$R_{biais} = \int_{-\infty}^{+\infty} \frac{1}{24\ f_X(x)} \left(\frac{df_X(x)}{dx}\right)^2 {\Delta_x}^2\ dx \qquad \text{(III. 10)}$$

Dans le cas où $f_X(x)$ est une gaussienne avec une moyenne arbitraire, le R_{biais} peut s'écrire comme :

$$R_{biais} = \frac{1}{24} \left(\frac{\Delta_x}{\sigma_x}\right)^2 \qquad \text{(III. 11)}$$

Le N_{biais} est défini comme suit :

$$N_{biais} = E\left\{\widehat{H}(X) - \widetilde{H}(X)\right\} \quad \text{(III. 12)}$$

Le développement de Taylor limité aux quatre premiers termes de la variable k_i autour de sa valeur moyenne dans l'Eq. (III.8) est effectué et sa version stochastique donne :

$$N_{biais} = -\frac{k_x - 1}{2\,N} \quad \text{(III.13)}$$

Il est supposé dans ce calcul que les échantillons sont indépendants.

Le biais total de premier ordre pour $\widehat{H}(X)$ est la somme de R_{biais} (III.11) et N_{biais} (III.13) :

$$Biais\{\widehat{H}(X)\} = B_{H_x} = -\frac{k_x - 1}{2\,N} + \frac{1}{24}\left(\frac{\Delta_x}{\sigma_x}\right)^2 \quad \text{(III. 14)}$$

Selon [94], la variance de cet estimateur est approximée par :

$$var(\widehat{H}(X)) \approx \frac{1}{2\,N} \quad \text{(III. 15)}$$

Généralement les performances d'un estimateur $\hat{\theta}$ se basent sur le critère de l'erreur quadratique moyenne EQM qui est définie par :

$$EQM(\hat{\theta}) = var(\hat{\theta}) + (B_{\hat{\theta}})^2 \quad \text{(III. 16)}$$

Ainsi, l'erreur EQM de l'entropie $\widehat{H}(X)$ est donnée par :

$$EQM\left(\widehat{H}(X)\right) = EQM_{H_x} = \frac{1}{2\,N} + \left(-\frac{k_x - 1}{2\,N} + \frac{1}{24}\left(\frac{\Delta_x}{\sigma_x}\right)^2\right)^2 \quad \text{(III. 17)}$$

III.3.2 ESTIMATION DE L'INFORMATION MUTUELLE PAR LA METHODE D'HISTOGRAMME

L'estimation de l'IM basée sur la méthode d'histogramme divise le plan euclidien $\{(x, y)/\ x \in R \text{ et } y \in R\}$ en (k_x x k_y) cellules (*bins*) de taille constante $(\Delta_X \times \Delta_Y)$ et de coordonnées (*i,j*).

L'approximation discrète de l'IM (III.2) s'écrit comme suit :

$$I(X;Y) \approx \sum_{j=1}^{kx}\sum_{j=2}^{ky} f_{XY}(x_i, y_j).\, log\left(\frac{f_{XY}(x_i, y_j)}{f_X(x_i).\, f_Y(y_j)}\right) \Delta_X\, \Delta_Y \quad \text{(III. 18)}$$

où (x_i, y_j) représente le centre du *bin* (*i,j*).

Utilisant le fait que $p_{ij} \approx f_{XY}(x_i, y_j)\, \Delta_x\, \Delta_y$, la probabilité p_{ij} d'observer un échantillon dans le *bin* (*i,j)* est estimée par $\hat{p}_{ij} \approx \frac{k_{ij}}{N}$ (k_{ij} est le nombre d'échantillons apparus dans le *bin* (*i,j)* parmi un nombre total de N données), l'Eq. (III.19) conduit à l'estimateur de *I*(*X ;Y*):

$$\hat{I}(X;Y) = \sum_{j=1}^{ky}\sum_{i=1}^{kx}\left(\frac{k_{ij}}{N}\right) log\left(\frac{k_{ij}N}{k_i\, k_j}\right) \quad \text{(III. 19)}$$

Le biais de cet estimateur peut être obtenu en suivant la même procédure de dérivation du biais de l'estimateur d'entropie. Pour deux variables gaussiennes *X* et *Y* ayant respectivement les écarts types σ_x, $\sigma_y$et une corrélation ρ, le biais s'exprime comme suit :

$$Biais\{\hat{I}(X;Y)\} = B_{I_{xy}} = \frac{(k_x-1)(k_y-1)}{2\,N} - \frac{\rho^2}{24(1-\rho^2)}\left(\left(\frac{\Delta_x}{\sigma_x}\right)^2 + \left(\frac{\Delta_y}{\sigma_y}\right)^2\right) \qquad \text{(III. 20)}$$

Selon [94], la variance est approximée par:

$$var(\hat{I}_{xy}) \approx \frac{\rho^2}{N} \qquad \text{(III. 21)}$$

D'après l'Eq.(III.16), l'EQM de $\hat{I}(X;Y)$ peut s'exprimer comme suit :

$$EQM\left(\hat{I}(X;Y)\right) = EQM_{I_{xy}} = \frac{\rho^2}{N} + \left(\frac{(k_x-1)(k_y-1)}{2\,N} - \frac{\rho^2}{24(1-\rho^2)}\left(\left(\frac{\Delta_x}{\sigma_x}\right)^2 + \left(\frac{\Delta_y}{\sigma_y}\right)^2\right)\right)^2 \qquad \text{(III. 22)}$$

L'originalité de notre travail, présentée dans [10], consiste à trouver des nombres de *bins* k_x et k_y permettant d'avoir des estimateurs $\widehat{H}(X)$ et $\hat{I}(X;Y)$ à faible biais. Le calcul de ces nombres de *bins* sera présenté dans la section (III.4). On montrera également que ces nombres de *bins* permettent aussi d'avoir une Erreur Quadratique Moyenne Minimale (EQMM) pour ces estimateurs. Pour comparer les performances de ces nouveaux estimateurs, plusieurs choix du nombre de *bins* utilisés dans la littérature sont présentés dans la section suivante.

III.3.3 FORMULES DU CHOIX DE NOMBRE DE BINS

Le choix du nombre de *bins* ou la largeur de *bins* est un paramètre très important dans la construction d'un histogramme décrivant une densité de probabilité d'une variable aléatoire.

Dans la littérature, plusieurs formules existent pour ce choix.

Soit *X* une variable aléatoire, *k* le nombre de *bin* et *h* la largeur de *bin*. Le nombre k peut être déduit de la largeur *h* par :

$$k = \frac{A_x}{h} \qquad \text{(III. 23)}$$

avec A_x l'étendue de la variable *x*,

Considérant *N* le nombre de données et σ_x l'ecart type de la variable *X,* la largeur *h* ou le nombre de bin *k* peuvent être donnés par les formules suivantes :

1. Formule de Sturges [105] :

$$k = 1 + log_2(N) \qquad \text{(III. 24)}$$

2. Formule de Scott [106] :

$$h = \frac{3.5\,\sigma}{N^{(\frac{1}{3})}} \qquad \text{(III. 25)}$$

3. Formule de Freedman-Diaconis [107] :

$$h = 2.\frac{IQR(x)}{N^{(\frac{1}{3})}} \quad \text{(III. 26)}$$

où IQR est l'intervalle interquartile.

Récemment, Shimazaki et Shnomoto [110] ont proposé un choix basé sur la minimisation d'une fonction de risque L^2 conduisant à une largeur de bin h :

$$h = \ arg\min_{\Delta} \frac{2\,\bar{m} - v}{\Delta^2} \quad \text{(III. 27)}$$

où $\bar{m}$ and v sont respectivement la moyenne et la variance biaisées de la largeur Δ.

Bien que les trois premières formules soient anciennes, elles sont encore utilisées [111] [112].

III.4 METHODES

III.4.1 NOUVELLE ESTIMATION DE L'ENTROPIE

Dans le problème de sélection des variables pertinentes (paramètres, caractéristiques) dans une tâche de classification, on a besoin généralement d'estimer l'information mutuelle $I_2(C;X)$ entre la variable discrète index de classe C et une variable continue X. Dans ce cas, l'IM $I_2(C;X)$ peut être exprimée en fonction de l'entropie $H(X)$ qui est estimée sur les données de l'ensemble des classes, et les entropies conditionnelles H(X\C=c), chacune étant estimée sur les données de la classe correspondante c :

$$I_2(X;C) \ = H(X) - \sum_{c=1}^{N_C} p_C(c)\mathrm{H}(X \backslash C = c) \quad \text{(III. 28)}$$

où p_C est la probabilité pour que la variable C prenne la valeur c.

L'estimation de $I_2(X;C)$ exige ainsi de bonnes estimations de l'entropie $H(X)$ et des entropies $H(X\backslash C=c)$. En prenant l'hypothèse que la fdp de la variable X est une gaussienne, une bonne estimation de l'entropie $\widehat{H}(X)$ de l'Eq. (III.8) peut être réalisée en cherchant le nombre de *bins* k_x permettant de minimiser l'erreur EQM de l'Eq. (III.17) :

$$k_{x\,opt} = arg\min_{k_x} EQM_{H_x} \quad \text{(III. 29)}$$

$$= arg\min_{k_x}\left(var\left(\widehat{H}_x\right) + \left(B_{\widehat{H}_x}\right)^2\right)$$

Puisque la variance de $\widehat{H}_x$ est indépendante du nombre k_x, le nombre de *bins* optimal est dépendant seulement du biais de $\widehat{H}_x$:

$$k_{x\,opt} = arg\min_{k_x}\left[\frac{1}{2\,N} + \left(B_{\widehat{H}_x}\right)^2\right]$$

$$= arg\min_{k_x}\left[{B_{H_x}}^2\right] \quad \text{(III. 30)}$$

Ainsi, les équations (III.29) et (III.30) nous montrent que le nombre de *bins* permettant de minimiser l'erreur EQM_{H_x} est équivalent à celui qui minimise le biais B_{H_x}. Cette minimisation peut être effectuée en annulant le biais B_{H_x} [10]:

$$-\frac{k_x-1}{2\,N}+\frac{1}{24}\left(\frac{\Delta_x}{\sigma_x}\right)^2=0 \qquad \text{(III.31)}$$

L'histogramme de la variable X peut être défini sur une étendue pratique $A_x=\alpha_x.\sigma_x$ conduisant à: $k_x\,\Delta_x=\alpha_x.\sigma_x$, avec α_x est une constante. Les échantillons à l'extérieur de cette étendue sont négligés.

Ainsi l'éq. (III.31) s'écrit :

$$\frac{1}{24}\left(\frac{\alpha_x}{k_x}\right)^2-\frac{k_x-1}{2\,N}=0$$

Cette dernière peut être écrite comme une équation de troisième ordre de la variable k_x:

$${k_x}^3-{k_x}^2-G=0 \qquad \text{(III.32)}$$

Avec la constante $G=\frac{N\,\alpha_x^2}{12}=\frac{N}{12}\,\frac{A_x^2}{\sigma_x^2}$.

Résolvant l'Eq. (III.32), on dérive le nombre de *bins* comme la plus proche valeur entière de la solution réelle positive (la solution de cette équation est détaillée dans [10] :

$$k_{xopt}=\ round\left\{\frac{\zeta}{6}+\frac{2}{3\zeta}+\frac{1}{3}\right\} \qquad \text{(III.33)}$$

Avec :

$$\zeta=\sqrt[3]{\left(8+108\,G+12\,\sqrt[2]{12\,G+81\,G^{\,2}}\right)} \qquad \text{(III.34)}$$

Pratiquement, G peut être déterminé par : $G=\frac{N}{12}\,\frac{A_x^2}{\hat{\sigma}_x^2}$, avec $\hat{\sigma}_x$ est l'estimée pratique de l'écart type σ_x et l'étendue $A_x$égale à la différence entre la valeur maximale et la valeur minimale de x : $A_x=x_{max}-x_{min}$. Ainsi, l'estimation du nombre de bin optimal demande l'estimation de la variance σ_x qui ne peut pas être précis pour un nombre de données réduit. Cependant, dans le cas d'une variable gaussienne, l'étendue pratique de l'histogramme peut être prise égale à $6\sigma_x$ ($\alpha_x=6$) permettant ainsi d'avoir un biais et une erreur EQM_{H_x}donnés par:

$$B_{\widehat{H}_x}=-\frac{k_{xopt}-1}{2\,N}+\frac{3}{2}\,\frac{1}{{k_{xopt}}^2} \qquad \text{(III.35)}$$

$$\mathrm{EQM}_{H_x}=\frac{1}{2\,N}+\left(-\frac{k_{xopt}-1}{2\,N}+\frac{3}{2}\,\frac{1}{{k_{xopt}}^2}\right)^2 \qquad \text{(III.36)}$$

Dans ce cas, le nombre de *bins* $k_{x\,opt}$ donné par (III.33) pour l'estimateur $\widehat{H}(X)$ devient indépendant de la variance σ_x puisque ζ ne dépend que du nombre de données N :

$$\zeta = \sqrt[3]{8 + 324N + 12\sqrt{36N + 729N^2}} \qquad \text{(III. 37)}$$

III.4.2 NOUVELLE ESTIMATION DE L'INFORMATION MUTUELLE

Le calcul de l'Information Mutuelle Multivariée (IMV) $I_3(C;X;Y)$ entre la variable index de classe C et deux variables continues X et Y est indispensable dans certains algorithmes de mesure de pertinence pour une tâche de classification tels que la JMI, CMI et TMI (voir chapitre II). L'IMV $I_3(C;X;Y)$ peut être exprimée en fonction de l'IM $I_2(X;Y)$ et les IM conditionnelles $I_2(X;Y\backslash C = c)$:

$$I_3(C;X;Y) = I_2(X;Y) - \sum_{c=1}^{N_C} p_C(c)\, I_2(X;Y\backslash C = c) \qquad \text{(III. 38)}$$

L'estimation de $I_3(C;X;Y)$ peut être obtenue par une estimation de l'IM $I_2(X;Y)$ et celles des IMs $I_2(X;Y\backslash C = c)$.

Supposant que la fdp conjointe des deux variables X,Y est une gaussienne, et admettant que $k = k_x = k_y$, une bonne estimation de l'IM $\hat{I}(X;Y)$ de l'Eq. (III.19) peut être obtenue en cherchant le nombre de *bins* k conduisant à une erreur $\mathrm{EQM}_{I_{xy}}$ minimale:

$$k_{opt} = arg\min_k EQM_{\hat{I}_{xy}} \qquad \text{(III. 39)}$$

$$= \arg\min_k\left(var\left(\hat{I}_{xy}\right) + \left(B_{\hat{I}_{xy}}\right)^2\right)$$

Puisque la variance de $\hat{I}_{xy}$ est indépendante du nombre k, le nombre de bin optimal est dépendante seulement du biais de $\hat{I}_{xy}$.

$$k_{opt} = \arg\min_k\left[\frac{\rho^2}{N} + \left(B_{\hat{I}_{xy}}\right)^2\right]$$

$$= \arg\min_k\left[\left(B_{\hat{I}_{xy}}\right)^2\right] \qquad \text{(III. 40)}$$

Les équations (III.39) et (III.40) nous montrent que le nombre de *bins* permettant de minimiser l'EQM de l'estimateur $\hat{I}(X;Y)$ est équivalent au nombre de *bins* qui minimise le biais de cet estimateur [10]. Cette minimisation peut être effectuée en posant le biais $B_{\hat{I}_{xy}}$ égal à zéro.

$$\frac{(k-1)^2}{2\,N} - \frac{\rho^2}{24(1-\rho^2)}\left(\left(\frac{\Delta_x}{\sigma_x}\right)^2 + \left(\frac{\Delta_y}{\sigma_y}\right)^2\right) = 0 \qquad \text{(III. 41)}$$

L'histogramme bidimensionnel conjoint des variables X, Y peut être défini pratiquement sur une étendue $A_x = \alpha_x.\sigma_x$ sur l'axe X et sur une étendue $A_y = \alpha_y.\sigma_y$ sur l'axe Y conduisant à : $k_x\,\Delta_x = \alpha_x.\sigma_x$, $k_y\,\Delta_y = \alpha_y.\sigma_y$ avec α_x et α_y des constantes. Les échantillons à

l'extérieur de ces étendues sont négligés. Ainsi l'eq (III.41) peut se mettre sous la forme suivante :

$$k^2(k-1)^2 = L \qquad \text{(III. 42)}$$

Avec une constante $L = \frac{N\,\rho^2}{12\,(1-\rho^2)}\left(\alpha_x^2 + \alpha_y^2\right)$

L'équation (III.42) peut être réduite à deux équations polynomiales d'ordre 2 après la prise de la racine carré de ses membres :

$$k(k-1) = \mp\sqrt{L} \qquad \text{(III. 43)}$$

La résolution de l'Eq. (III.43) permet d'accéder au nombre de *bins* optimal comme la plus proche valeur entière de la solution réelle positive :

$$k = round\left\{\frac{1}{2} + \frac{1}{2}\sqrt{1+4\sqrt{L}}\right\} \qquad \text{(III. 44)}$$

L'histogramme d'une distribution gaussienne peut être défini sur une étendue $6\,\sigma$, ainsi $A_x = 6\sigma_x$ et $A_y = 6\sigma_y$. Donc, le nombre de bins pour l'estimateur de l'IM basé sur l'histogramme est donné par :

$$k = round\left\{\frac{1}{2} + \frac{1}{2}\sqrt{1+4\sqrt{\frac{6N\,\hat{\rho}^2}{1-\hat{\rho}^2}}}\right\} \qquad \text{(III. 45)}$$

où le coefficient de corrélation inconnu ρ est remplacé par son estimateur classique $\hat{\rho}$.

Le biais et l'erreur quadratique moyenne correspondant à ce nombre de *bins* sont donnés par :

$$B_{\hat{I}_{xy}} = \frac{(k-1)^2}{2.N} - \frac{3\,\rho^2}{(1-\rho^2)k^2} \qquad \text{(III. 46)}$$

$$EQM_{\hat{I}_{xy}} = \frac{\rho^2}{N} + \left(\frac{(k-1)^2}{2.N} - \frac{3\,\rho^2}{(1-\rho^2)k^2}\right)^2 \qquad \text{(III. 47)}$$

Selon l'équation (III.45), le nombre de *bins* est fonction du nombre d'échantillons et du coefficient de corrélation ρ. Les figures (III.1) et (III.2) donnent respectivement la courbe du nombre de *bins* k en fonction du ρ et en fonction du nombre d'échantillons.

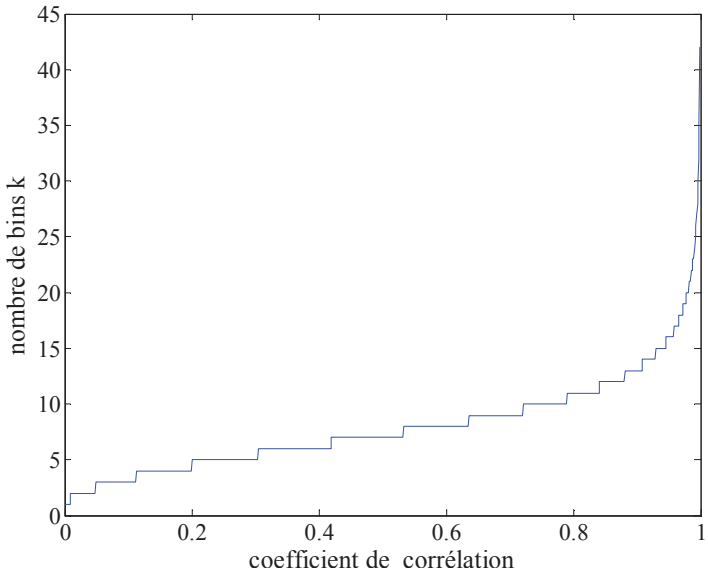
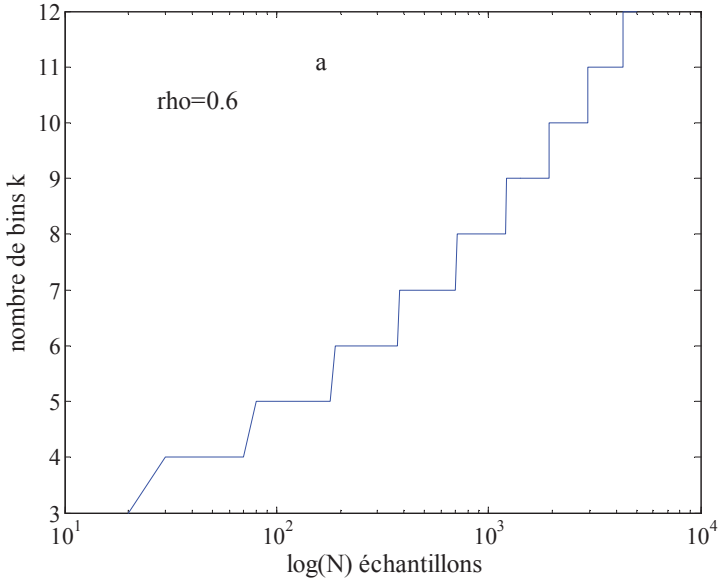


Figure.III.1 : nombre de *bins* en fonction du coefficient de corrélation ρ pour N=1000



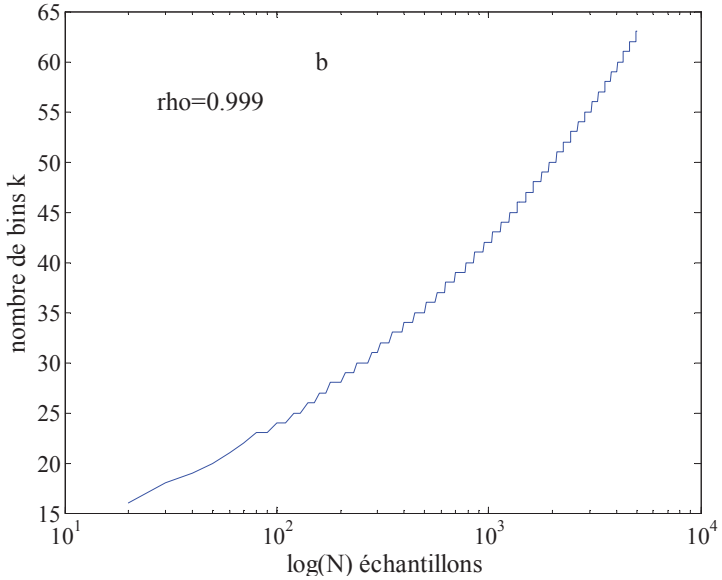


Figure.III.2 : nombre de *bins* en fonction du nombre d'échantillons N pour ρ=0.6 et ρ=0.999

Les figures (III.1) et (III.2) montrent que l'estimation de l'IM entre deux variables décorrélées exige un petit nombre de *bins* alors que celle entre deux variables très corrélées exige un très grand nombre de *bins*. Selon l'Eq (III.19), cette dernière estimation demande un coût de calcul très grand par rapport à la première estimation.

III.5 SIMULATIONS ET RESULTATS

On discute dans cette section plusieurs expériences pour évaluer les performances de la nouvelle approche pour le choix du nombre de *bins* pour l'estimation de l'entropie et de l'information mutuelle ainsi que leur application dans la sélection des variables pertinentes. La section est structurée en deux parties. Premièrement, on compare les performances de l'approche proposée aux autres approches couramment utilisées dans la littérature pour l'estimation de l'entropie et de l'information mutuelle en utilisant des données gaussiennes simulées. Deuxièmement, on montre l'intérêt de l'approche proposée pour trouver le nombre exact des paramètres pertinents (features) dans un problème de sélection des variables gaussiennes, simulées, pour une tâche de classification.

Bien que l'approche se base sur la fdp de référence gaussienne, nous montrons qu'en pratique, qu'elle reste robuste pour d'autres lois de probabilités telles que la loi uniforme et la loi Lognormal dans une tâche de sélection des paramètres pertinents.

Les différents calculs sont effectués sur un PC doté d'un CPU I5 d'horloge 3 Ghz et d'une mémoire RAM de 8 Go avec le langage de calcul Matlab.

III.5.1 ESTIMATION DE L'ENTROPIE ET DE L'IM DES VARIABLES SIMULEES

Dans cette section, on mesure les performances des nouveaux choix du nombre de *bins* donnés dans les équations (III.33) et (III.45) au sens de l'EQM et du temps de calcul pour l'estimation de l'entropie et de l'information mutuelle basées sur l'approche d'histogramme. Les performances de l'approche proposée sont comparées avec celles des approches déjà décrites dans la section (III.3.3) à savoir : Scott [SC], Sturges [ST], Freedman [FD] et Shimazaki [SH], ainsi que la méthode GMM.

III.5.1.1 Expérience d'estimation de l'Entropie d'une variable gaussienne

L'objectif de cette expérience est d'évaluer les performances de différents estimateurs de l'entropie d'une variable aléatoire gaussienne. La valeur théorique de l'entropie d'une variable aléatoire gaussienne de moyenne μ et d'écart-type σ est donnée par :

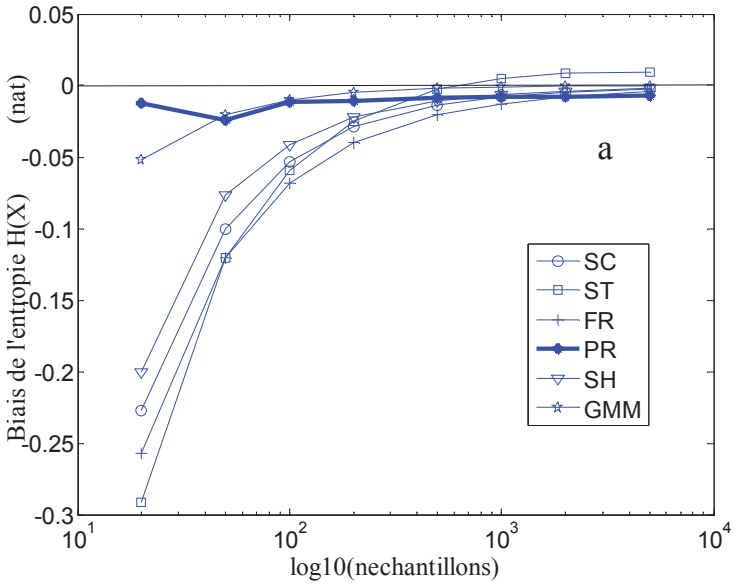
$$Ht_x = \frac{1}{2}\log(2\pi e\sigma^2) \qquad (nat) \qquad \text{(III. 48)}$$

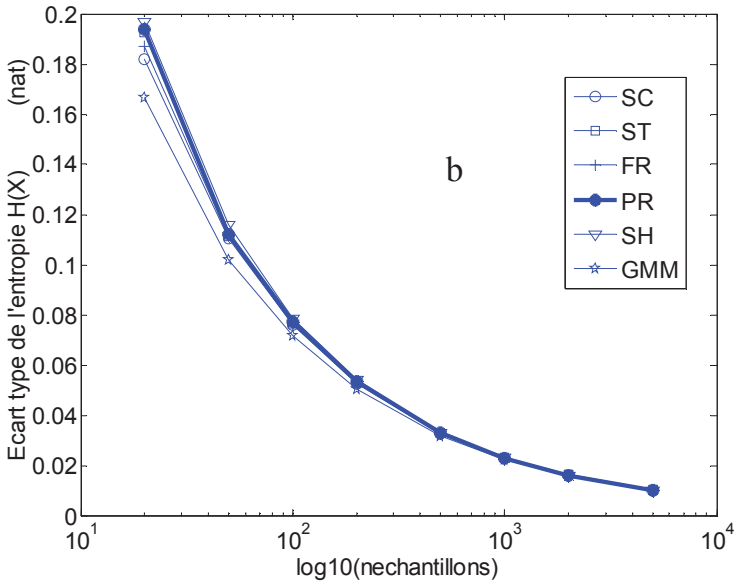
Théoriquement et selon les équations (III.15), (III.35) et (III.36), la variance de l'entropie dépend seulement du nombre d'échantillons alors que le biais et l'EQM dépendent du nombre d'échantillons N ainsi que du nombre de *bins* k qui prend sa valeur selon l'approche à utiliser.

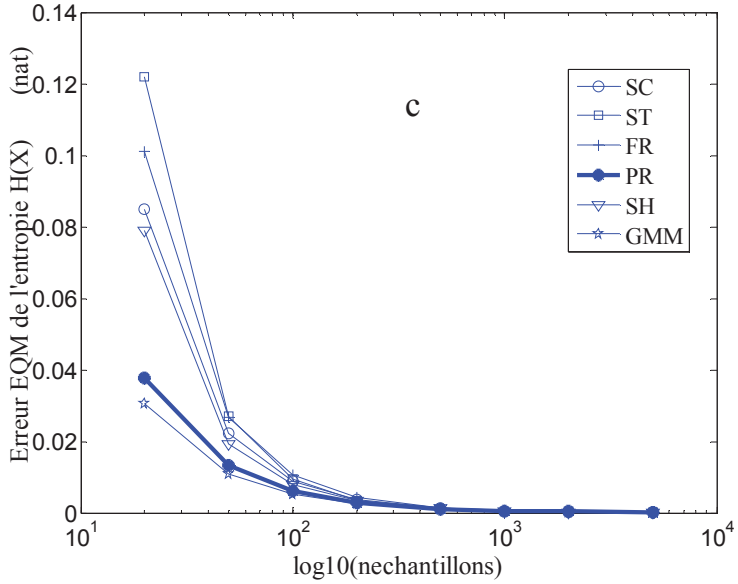
Ainsi, dans cette expérience, nous discutons au point de vue pratique à l'aide de données simulées issues d'une loi de gauss, l'influence du nombre d'échantillons N sur les performances des différentes approches d'estimation de l'entropie au sens de l'EQM ainsi que du temps de calcul.

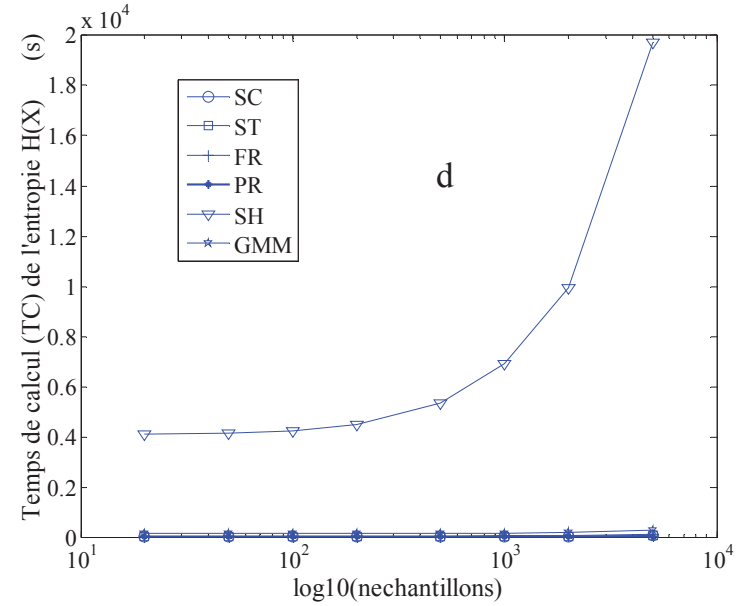
Soit X une variable aléatoire gaussienne de moyenne μ=0 et d'écart-type σ=1, générée sur N échantillons. La génération de cette variable, ainsi que l'estimation de son entropie par 5 approches basées histogramme (Scott [SC], Sturges [ST], Freedman [FD], Proposée [PR], Shimazaki [SH],) et 1'approche basée modèle (GMM) sont effectuées 100000 fois ce qui permet de calculer empiriquement le biais, l'écart-type et l'EQM de l'estimation de $H(X)$ pour chacune de ces approches. Le biais et l'EQM sont déterminés par rapport à la valeur théorique $Ht_x(X)$ donnée ci-dessus.

La figure (III.3) montre, le biais, l'écart-type et l'EQM en fonction du nombre d'échantillons N (20, 50, 100, 200, 500, 1000, 2000, 5000), ainsi que le temps de calcul (TC) de 100000 estimations de l'entropie. Ces trois premières grandeurs sont mesurées en *nat*, alors que le temps de calcul est mesuré en seconde (s). Dans l'estimation de $H(X)$ par GMM donnée dans l'équation (III.4), on a proposé de fixer le nombre de gaussiennes à 1 afin de considérer la fonction pdf comme une gaussienne.




b
SC
ST
FR
PR
SH
GMM
Ecart type de l'entropie H(X) (nat)
log10(nechantillons)


c
SC
ST
FR
PR
SH
GMM
Erreur EQM de l'entropie H(X) (nat)
log10(nechantillons)


d
x 10^4
SC
ST
FR
PR
SH
GMM
Temps de calcul (TC) de l'entropie H(X) (s)
log10(nechantillons)

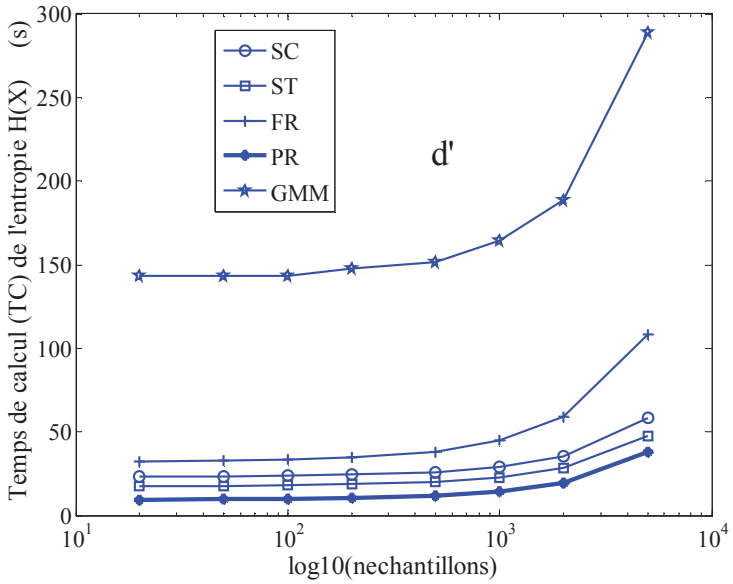


Figure.III.3 : Biais (a), Ecart type (b), EQM (c) d'estimation de $H(X)$ pour une distribution Gaussienne*;* temps de calcul *TC* (d,d') ; Méthodes d'estimation: Scott [SC], Sturges [ST], Freedman-Diaconis [FD], , Proposée [PR], Shemazaki [SH] et GMM.

Le tableau (III.1) donne les valeurs de ces grandeurs pour N={20,100,1000} afin de donner plus de détails par rapport à la figure (III.3).

TABLEAU III.1:

Estimation de l'entropie $H(X)$ pour une distribution Gaussienne : Méthodes d'estimation de l'IM: Scott [SC], Sturges [ST], Freedman-Diaconis [FD], Méthode proposée [PR], Shimazaki [SH] et GMM ; $\widehat{\boldsymbol{H}}(\boldsymbol{X})$: entropie estimée; Biais ; Ecart type (std); EQM; *TC*: temps de calcul

$Ht(X)$	1.4189																	
N	**20**						**100**						**1000**					
Method	**SC**	**ST**	**FD**	**PR**	**SH**	**GMM**	**SC**	**ST**	**FD**	**PR**	**SH**	**GMM**	**SC**	**ST**	**FD**	**PR**	**SH**	**GMM**
$\widehat{H}(X)$	1.1915	1.1277	1.1618	1.4066	1.2186	1.3668	1.3187	1.2981	1.2987	1.3943	1.3423	1.3984	1.4108	1.4237	1.4062	1.4110	1.4125	1.4178
Biais	-0.2275	-0.2913	-0.2572	**-0.0123**	-0.2003	**-0.0522**	-0.0537	-0.0594	-0.0686	**-0.0117**	-0.0413	**-0.0104**	-0.0081	0.0048	-0.0127	-0.0080	**-0.0064**	**-0.0011**
Std	0.1666	0.1970	0.1940	0.1868	0.1925	0.1820	0.0715	0.0787	0.0772	0.0761	0.0764	0.0760	0.0228	0.0233	0.0228	**0.0230**	0.0231	**0.0224**
EQM	0.0848	0.1219	0.1011	**0.0378**	0.0789	**0.0305**	0.0087	0.0094	0.0105	*0.0061*	0.0079	**0.0052**	0.0006	0.0006	0.0007	0.0006	0.0006	**0.0005**
***TC*(ks)**	0.0234	0.0177	0.0325	**0.0095**	4.0868	0.1435	0.0236	0.0180	0.0333	**0.0099**	4.2422	0.1432	0.0293	0.0229	0.0449	**0.0146**	6.9248	0.1643

A partir de la figure III.4 et du tableau III.1, nous pouvons exprimer les commentaires suivants:

1. La méthode proposée PR présente un biais minimal de $H(X)$ dans le cas d'un nombre d'échantillons réduit ($N \leq 200$) et un biais proche de celui des autres méthodes dans le cas d'un N supérieur à 200.
2. Selon la figure (III.3:b), l'écart type de $H(X)$ est presque le même pour toutes les méthodes, ce qui justifie l'équation (III.15) qui montre que l'écart type est indépendant du nombre de *bins* k et dépend inversement seulement du nombre d'échantillons.

3. La méthode PR présente une EQM minimale par rapport aux méthodes : SC, ST, FR, SH pour un nombre d'échantillons *N* inférieur à 200 et une EQM proche de celle des autres méthodes dans le cas d'un *N* supérieur à 200 (Fig.III.4:c). Dans la première gamme d'échantillons, la méthode de Sturges présente l'erreur maximale (exemple : erreur de 0.1219 *nat* pour ST et 0.0378 *nat* pour PR dans le cas de 20 échantillons).
4. La méthode PR présente également un temps de calcul (*TC*) minimal par rapport aux autres méthodes (Fig.III.4:d). De plus la méthode de Shimazaki [SH] exige un temps de calcul très grand (plus grand de 474 fois par rapport à PR dans le cas de 1000 échantillons) conduisant cette méthode à être inutilisable dans des applications en temps réel.
5. La méthode GMM à une composante présente une EQM minimale mais exige un temps de calcul qui peut atteindre 15 fois celui de la méthode PR (Tableau.III.1).

Ainsi, la méthode proposée réalise un bon compromis entre la précision et le temps de calcul pour un nombre d'échantillons soit limité (N=20) soit très grand (N=1000).

III.5.1.2 Expérience d'estimation de l'information mutuelle entre deux variables gaussiennes

L'objectif de cette expérience est d'évaluer les performances de différents estimateurs de l'information mutuelle IM entre deux variables de type gaussien. La valeur théorique de la IM entre deux variables gaussiennes *X, Y* ayant un coefficient de corrélation ρ est donnée par :

$$I_2(X;Y) = -\frac{1}{2}\log(1-\rho^2) \qquad (nat) \qquad \text{(III. 49)}$$

Théoriquement et selon les équations (III.21), (III.46) et (III.47), la variance dépend du nombre d'échantillons *N* et du coefficient de corrélation ρ alors que le biais et l'erreur EQM dépendent en plus du nombre de *bins k* qui prend sa valeur selon l'approche à utiliser.

Ainsi, dans cette expérience, nous discutons sur un plan pratique, à l'aide de données simulées représentant deux variables gaussiennes, l'influence du nombre d'échantillons *N* et de la valeur du coefficient de corrélation ρ sur les performances des différentes approches d'estimation de l'IM, au sens de l'EQM et du temps de calcul (*TC*).

Soient deux variables *X,Y* ayant respectivement les moyennes $\mu_x = 0, \mu_y = 0$ et les écarts type $\sigma_x = 1, \sigma_y = 1$, corrélées avec un coefficient ρ. La génération de ces variables sur *N* échantillons, ainsi que l'estimation de leur IM par les approches (Scott [SC], Sturges [ST], Freedman [FD], Proposée [PR], Shimazaki [SH], GMM) sont effectuées 10000 fois afin d'estimer l'écart-type, le biais et l'EQM de ces estimateurs. Le biais et l'EQM sont déterminés par rapport à la valeur théorique $IMt_x(X)$ donnée en équation (III.49).

Dans l'estimation de $I_2(X;Y)$ par les méthodes de Scott, de Fredman et de Shimazaki, le nombre de *bins* pour X est déterminé indépendamment du nombre de *bins* de Y.

Le nombre de gaussiennes pour la méthode GMM est fixé à 1 afin de considérer la fonction de référence binormale.

Le tableau (III.2.a) donne les valeurs du biais, l'écart-type (std) et l'EQM en fonction du coefficient de corrélation $\rho = \{0,0.6,0.999\}$, pour le cas N=1000, alors que le tableau (III.2.b) donne les valeurs de ces grandeurs pour le cas N=50. Ces deux cas sont considérés pour voir l'influence du nombre limité de données sur l'estimation de l'IM.

La figure (III,4) nous montre l'EQM en fonction du nombre d'échantillons N (20,50, 100, 200, 500, 1000, 2000, 5000) pour le cas du $\rho = \{0,0.6,0.999\}$ ainsi que le temps de calcul (*TC*) de 100000 estimations de l'IM. Dans cette figure, l'erreur EQM, de l'estimation de l'IM est mesurée en *nat* alors que le temps de calcul (*TC*) est mesuré en seconde (*s*). Dans cette figure l'estimation par la méthode de Shimazaki n'est pas considérée car elle exige un temps de calcul très élevée par rapport aux autres méthodes (voir tableau (III,2)).

A partir de la figure (III.4) et du tableau (III.2), nous remarquons les points suivants:

1. La méthode PR présente une EQM minimale par rapport aux méthodes SC, ST, FD, SH, pour un petit ou un grand nombre d'échantillons N. Plus particulièrement, dans le cas de nombre limité des données, ces dernières méthodes commettent plus d'erreur.
2. La méthode SH commet la plus grande erreur pour le cas de grande corrélation et de nombre limité de données (tableau III.2.b). De plus elle exige un temps de calcul qui peut atteindre 608 fois celui de la méthode proposée (Tableau III.2.a). Ainsi cette méthode devient impraticable dans le cas des applications exigeant un très grand nombre d'estimations de l'IM).
3. Pour des basses et moyennes corrélations (Fig.III.4:a,b) la méthode PR a approximativement la même EQM que la méthode GMM à une composante mais pour de grandes corrélations, cette dernière est meilleure (Fig.III.4:c).
4. La méthode PR est plus rapide que les autres méthodes: SC, ST, FD, GMM, SH.
5. La méthode GMM présente approximativement un temps de calcul (*TC*) plus élevé que celui de la méthode proposée (Tableau III.2.b).

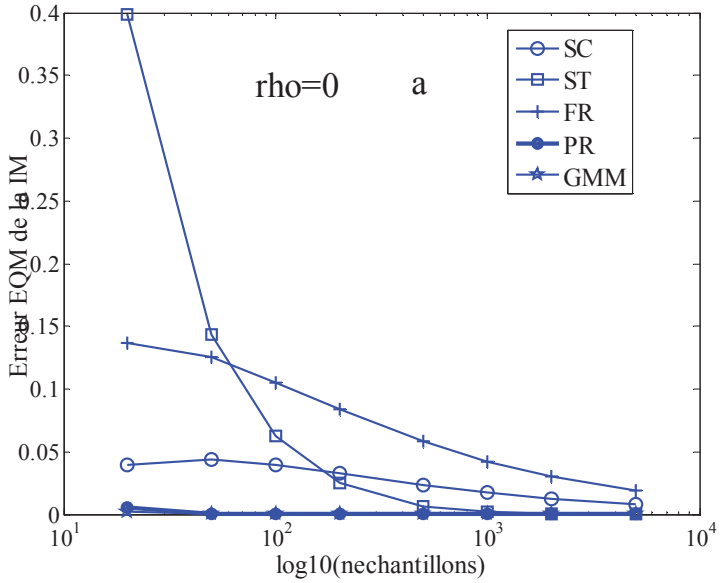
Tableau III.2 : Estimation de l'IM pour une distribution Gaussienne : Méthodes d'estimation : Scott [SC], Sturges [ST], Freedman-Diaconis [FD], Méthode proposée [PR], Shimazaki [SH] et GMM ;
I_{XY}: IM théorique; $\hat{I}_{XY}$: IM estimée; Biais:Ecart type (std); EQM; TC: temps de calcul

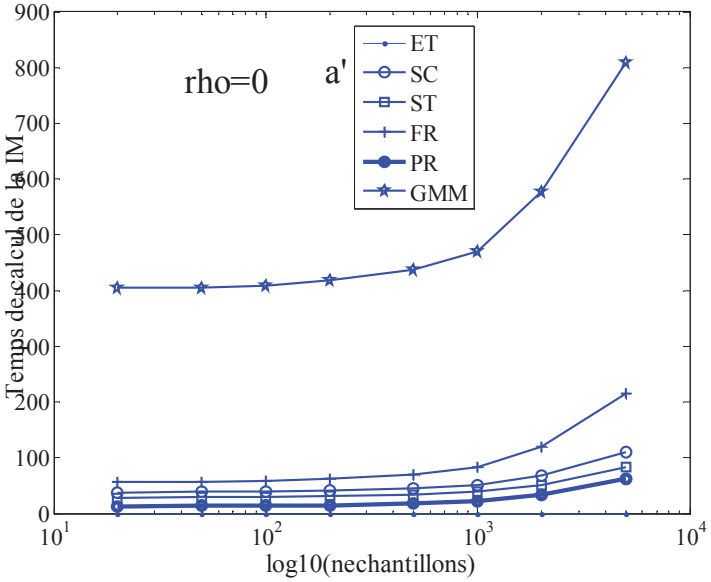
Tableau III.2.a (N=1000 échantillons par réalisation)

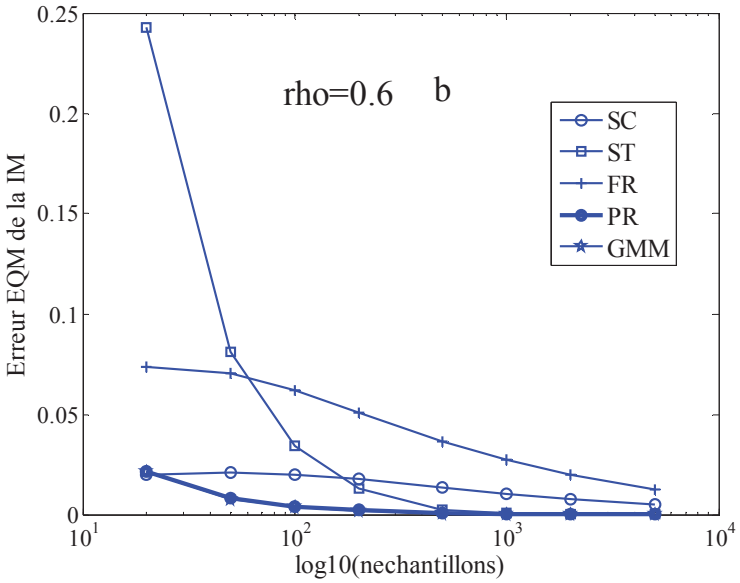
ρ	0						0.6						0.999					
I_{XY}	0						0.2231						3.1076					
Meth	SC	ST	FD	PR	SH	GMM	SC	ST	FD	PR	SH	GMM	SC	ST	FD	PR	SH	GMM
$\hat{I}_{XY}$	0.1315	0.0467	0.2058	0.0007	0.1266	0.0005	0.3232	0.2443	0.3872	0.2188	0.3190	0.2237	2.0405	1.6731	2.2067	2.6480	1.9332	3.1076
Biais	0.1315	0.0467	0.2058	**0.0007**	0.1266	**0.0005**	0.1000	0.0212	0.1640	**-0.0044**	0.0959	**0.0005**	-1.0671	-1.4344	-0.9009	**-0.4595**	-1.1743	**0.0001**
Std	0.0109	0.0066	0.0160	**0.0012**	0.0284	**0.0007**	0.0194	0.0187	0.0214	**0.0188**	0.0296	**0.0189**	0.0859	0.0656	0.0757	**0.0328**	0.1377	**0.0317**
EQM	0.0174	0.0022	0.0426	**0.0000**	0.0168	**0.0000**	0.0104	0.0008	0.0274	**0.0004**	0.0101	**0.0004**	1.1460	2.0619	0.8173	**0.2122**	1.3980	**0.0010**
***TC* (s)**	05.50	03.60	08.60	**02.30**	1398.3	61.8	05.00	03.60	08.40	**02.40**	1396.6	61.90	04.90	03.60	08.20	**03.20**	1395.6	61.8

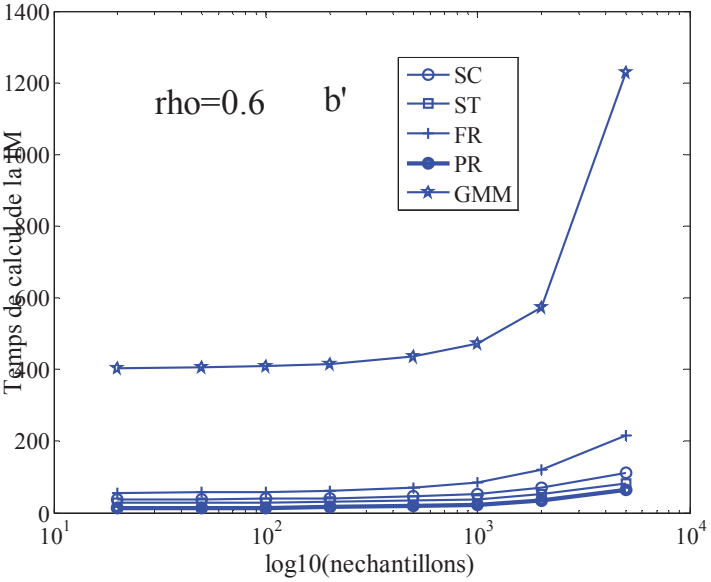
Tableau III.2.b (N=50 échantillons par réalisation)

ρ	0						0.6						0.999					
I_{XY}	0						0.2231						3.1076					
Meth	SC	ST	FD	PR	SH	GMM	SC	ST	FD	PR	SH	GMM	SC	ST	FD	PR	SH	GMM
$\hat{I}_{XY}$	0.1997	0.3711	0.3396	0.0147	0.1077	0.0106	0.3447	0.4959	0.4684	0.2309	0.2383	0.2344	1.3173	1.5707	1.4649	2.2032	0.9681	3.1188
Biais	0.1997	0.3711	0.3396	**0.0147**	0.1077	**0.0106**	0.1216	0.2728	0.2452	**0.0077**	0.0152	**0.0113**	-1.7903	-1.5369	-1.6426	**-0.9043**	-2.1395	**0.0113**
Std	0.0604	0.0712	0.0995	**0.0251**	0.0962	**0.0149**	**0.0810**	**0.0851**	0.1042	0.0897	0.1163	0.0873	**0.1291**	**0.1240**	0.1813	0.1294	0.2995	0.1433
EQM	0.0435	0.1428	0.1252	**0.0008**	0.0208	**0.0003**	0.0213	0.0816	0.0710	**0.0081**	0.0138	**0.0077**	3.2217	2.3773	2.7311	**0.8346**	4.6671	**0.0207**
***TC* (s)**	3.732	2.604	5.575	**1.443**	835.5	54.02	3.699	2.599	5.592	**1.444**	835.4	53.87	3.646	2.582	5.547	**1.620**	834.4	53.42




rho=0
a'
ET
SC
ST
FR
PR
GMM
Temps de calcul de la IM
log10(nechantillons)


rho=0.6
b
SC
ST
FR
PR
GMM
Erreur EQM de la IM
log10(nechantillons)


rho=0.6
b'
SC
ST
FR
PR
GMM
Temps de calcul de la IM
log10(nechantillons)

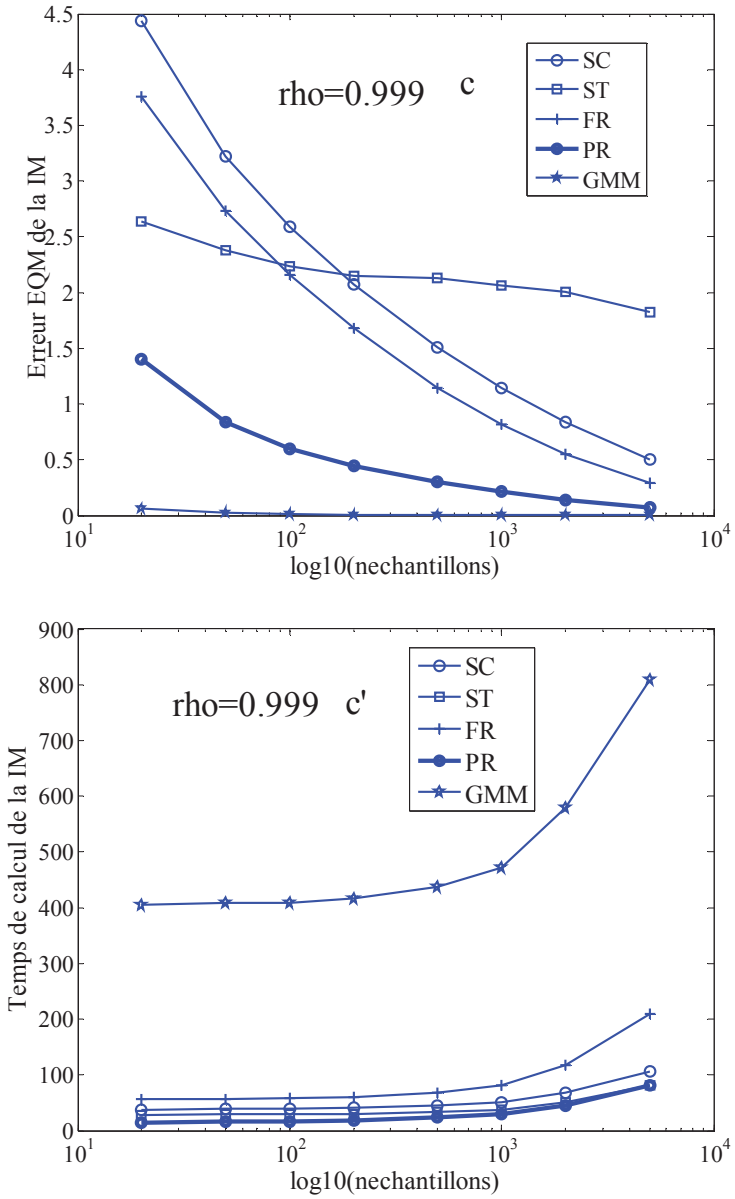


Figure.III.4 : Erreur d'estimation de l'information mutuelle (EQM) et temps de calcul (*TC*). Méthodes d'estimation: Scott [SC], Sturges [ST], Freedman-Diaconis [FD], Proposée [PR] et GMM

III.5.1.3 Expérience d'estimation de l'information mutuelle de deux variables uniformes

On discute dans cette section l'estimation de l'information mutuelle entre deux variables aléatoires uniformes par les différentes approches citées précédemment. Néanmoins, à cause de la difficulté et de la diversité de représentation de la fonction de densité de probabilité conjointe (fdpc) bivariée de deux variables uniformes X et Y [113], on ne discute que le cas de la fdp de deux variables indépendantes dans lequel l'information mutuelle théorique est $I(X;Y)=0$ [64].

Soient X, Y deux variables aléatoires uniformes continues sur le segment [0,10], générées sur N échantillons. La génération de ces variables, ainsi que l'estimation de leur information

mutuelle par les 7 approches (Scott [SC], Sturges [ST], Freedman [FD], Proposée [PR], Shimazaki [SH], GMM) sont effectuées 10000 fois afin de calculer l'EQM de l'estimation de $I(X;Y)$ par ces approches. Dans l'estimation de $I(X;Y)$ par GMM donnée dans l'équation (III.5), nous avons proposé de déterminer le nombre de gaussiennes par le critère d'information de Akaike [114].

Le tableau (III,3) nous montre l'EQM ainsi que le temps de calcul (*TC* en *s*) correspond à 10000 estimations de l'information mutuelle dans le cas où la variable *X* est indépendante de *Y*.

TABLEAU III.3:

Estimation de l'information mutuelle entre deux variables uniformes indépendantes. Méthodes d'estimation de l'IM: Scott [SC], Sturges [ST], Freedman-Diaconis [FD], proposée [PR], Shimazaki [SH]], [GMM]; EQM ; *TC* : temps de calcul.

		Méthode											
		SC		**ST**		**FD**		**PR**		**SH**		**GMM**	
		EQM	*TC (s)*	EQM	*TC*	EQM	TC	EQM	TC	EQM	TC	EQM	TC
Nombre d'échantillons	20	0.0230	3.57	0.4370	2.50	0.0368	5.29	**0.0057**	**1.36**	**0.0051**	774	**0.0026**	60
	50	0.0134	4.62	0.1909	2.80	0.0179	6.53	**0.0009**	**1.63**	**9.4E-4**	812	0.0148	224
	100	0.0088	4.78	0.0894	2.99	0.0105	6.76	**1.9E-4**	**1.78**	**1.9E-4**	832	0.0061	750
	200	0.0050	5.07	0.0345	3.18	0.0063	7.18	**4.7 E-5**	**1.81**	**3.7E-5**	889	0.0020	1303
	500	0.0029	5.09	0.0075	3.28	0.0034	7.22	**7.5E-6**	**1.96**	**4.5E-6**	1105	0.0011	2083
	1000	0.0019	5.76	0.0027	3.80	0.0022	8.67	**1.7E-6**	**2.47**	**8.5E-7**	1521	0.0009	3329
	2000	0.0014	7.32	0.0010	4.96	0.0014	11.90	**4E-8**	**3.50**	**1.9E-7**	2331	0.0008	6781
	5000	0.0007	10.91	0.0003	7.82	0.0008	20.14	**6.7E-8**	**6.23**	**3E-8**	31695	0.0007	15285

A partir de ce tableau on peut remarquer les points suivants :

- La méthode proposée PR et celle de Shimazaki sont les plus précises. Néanmoins la méthode PR est la plus rapide aux autres méthodes et celle de Shimazaki est la plus lente.
- La méthode GMM est la moins précise et très lente. Ainsi la méthode GMM est un mauvais choix pour représenter une variable uniforme.
- Dans le cas d'un nombre de données inférieur à 2000, la méthode Sturges est moins précise par rapport aux méthodes SC, FD, PR mais elle est plus rapide par rapport à SC et FD.
- La méthode PR est plus précise et plus rapide par rapport les méthodes SC, ST, FD, GMM.

Ainsi, on peut conclure que les estimateurs proposés pour le calcul de l'information mutuelle, fondés sur le bon choix du nombre de *bins* de l'histogramme présentent une erreur quadratique moyenne minimale et un temps de calcul très réduit par rapport aux autres

estimateurs qui se basent sur le choix du nombre de *bins* de Scott, de Sturges, et de Freedman. De plus, nos estimateurs présentent également un bon compromis entre la précision et le temps de calcul par rapport à la méthode GMM. Bien que la méthode de Shimazaki est précise dans certains cas, elle est très lente et reste une méthode inutilisable dans des applications en temps réel.

Comme déjà discuté dans le chapitre précédent, l'estimation de l'entropie et celle de l'information mutuelle peuvent être utilisées dans le problème de sélection des variables pertinentes. Ainsi pour valider notre méthode, nous discutons dans la section suivante ce problème pour différents types de variables aléatoires : gaussiennes, uniformes, lognormales.

III.5.2 SELECTION DES VARIABLES PERTINENTES APPLIQUEES SUR DES DONNEES SIMULEES

Dans cette expérience on montre la capacité de notre approche dans la détection du naombre de paramètres (variables) pertinents dans une tâche de classification de données simulées. Les différents algorithmes de sélection appliqués dans cette expérience diffèrent selon le degré de redondance qu'ils prennent en compte entre un sous-ensemble de variables déjà sélectionné et une autre à sélectionner (voir chapitre II). Les algorithmes considérés sont : MMI (II.42), JMI (II.46), CMI (II.49), TMI (II.65). Ces algorithmes sont appliqués premièrement sur des données simulées gaussiennes afin de prendre en considération l'hypothèse de la fdp de référence gaussienne. Ensuite on montre la robustesse de notre approche pour des données non gaussiennes générées à l'aide de la loi uniforme et la loi lognormale. Les différentes méthodes considérées pour l'estimation de l'information mutuelle et de l'entropie sont : Scott [SC], Sturges [ST], Freedman-Diaconis [FD], proposée [PR], Shimazaki [SH], mélange de Gaussiennes [GMM].

La simulation considère 20 classes de données caractérisées par 15 variables. Ces variables sont synthétisées comme suit : les 5 premières (X_1 à X_5) considérées comme pertinentes sont indépendantes et elles expliquent les 20 classes, les 5 variables suivantes X_6 à X_{10} sont respectivement redondantes par rapport aux variables X_1 à X_5, alors que les 5 dernières variables sont indépendantes de toutes les variables et de la variable index de classe C.

Dans le cas des données gaussiennes, X_1 à X_5 sont des variables gaussiennes générées dans chaque classe avec des moyennes μ_i et des écarts types σ_i $(i = 1{:}5)$ définis par : $\mu_i = \sigma_i^2$. Afin de fixer ces moyennes pour les différentes classes, chaque μ_i $(i = 1{:}5)$ est générée aléatoirement 20 fois par une loi uniforme sur un intervalle $[a_i, b_i]$ tel que $a_{i+1} = 5\,a_i$ et $b_{i+1} = 5b_i$ commençant par $a_1 = 1$ et $b_1 = 2$. Les variables X_6 à X_{10} sont définies comme suit : $X_{i+5} = X_i + Y_i$ pour i allant de 1 à 5 où Y_i sont des variables gaussiennes de moyennes et

écart types définis par : $\mu_i = \sigma_i^2 = 10b_i$. Les variables X_{11} à X_{15} sont des variables gaussiennes indépendantes avec des moyennes $\mu_i = \sigma_i^2 = 1$ (*i=11 :15*).

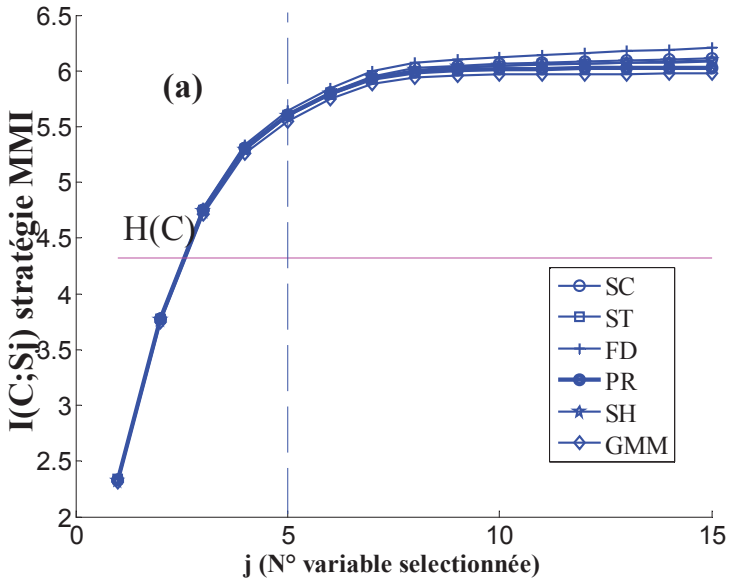
Deux simulations sont considérées, la première considère 1000 réalisations multivariées par classe, l'autre prend 50 réalisations par classe. Les deux cas sont considérés afin d'analyser l'effet du nombre limité de réalisations ou d'observations sur l'estimation de l'information mutuelle $I(C;S)$ entre la variable index classe C et un sous-ensemble de variables sélectionnées S. Un simple GMM basé sur le sous-ensemble S est utilisé comme classifieur afin de valider la pertinence de ce sous-ensemble.

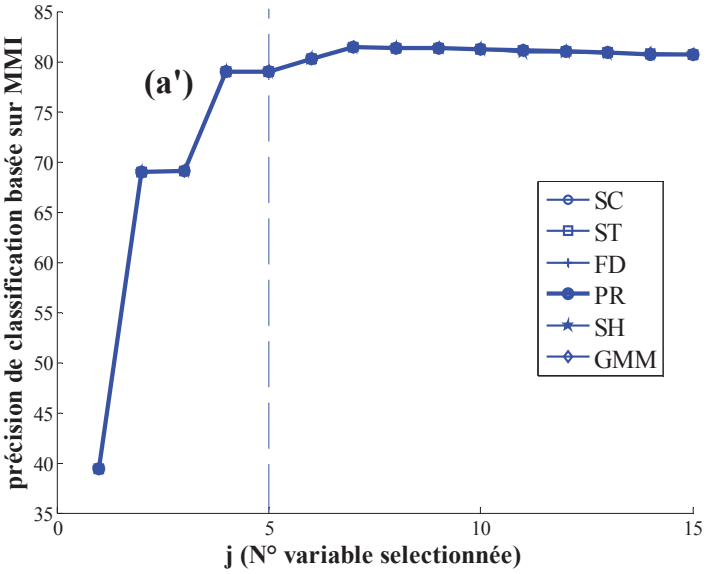
Théoriquement, si S comprend des variables redondantes, alors $I(C;S)$ doit croître à chaque itération de la procédure de sélection jusqu'à ce qu'elle atteigne un plateau correspondant au nombre des variables pertinentes. La valeur maximale de $I(C;S)$ ainsi que ce plateau ne doit pas dépasser l'entropie $H(C)$ qui est égale à 4.3 bits. Dans notre cas le nombre optimal des variables pertinentes est égal à 5, donc théoriquement $I(C;S)$ atteint ce plateau pour ce nombre de variables. Cependant, pratiquement $I(C;S)$ peut avoir une valeur supérieure à $H(C)$. Ceci est probablement causé par les stratégies de sélection des paramètres (MMI, JMI, CMI, TMI) qui ne tiennent pas en compte la redondance exacte où par l'estimation biaisée des termes des informations mutuelle incluses dans le calcul de $I(C;S)$.

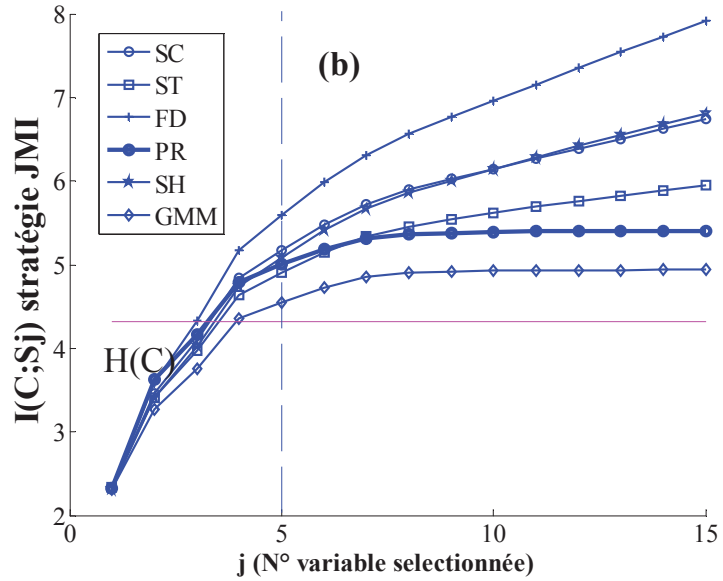
La figure III.5.a montre les résultats pour la stratégie MMI dans le cas de 1000 échantillons par classe. Toutes les méthodes SC, ST, FD, PR, SH, GMM atteignent approximativement leur plateau à partir d'un nombre de variables égal à 10. Ce résultat montre la surestimation du nombre de variables qui doit être égal à 5. De plus ce plateau atteint une valeur supérieure à l'entropie $H(C)$. Cette estimation erronée est due principalement au fait que la stratégie MMI ne tient pas en compte la redondance entre une variable à sélectionner à l'itération j et un sous-ensemble de variables déjà sélectionné à l'itération j-1. Cet argument est justifié par les résultats qui montrent l'existence des variables redondantes dans le sous-ensemble des 5 premières variables sélectionnées: 5,4,10,3,9. La figure III.5 (a') montre la précision de classification en fonction des variables sélectionnées. Pour toutes les méthodes d'estimation de l'IM, la précision atteint approximativement un plateau à partir d'un nombre de variables égal à 10. En plus la précision confirme la redondance de la variable 10 et la variable 9 puisqu'elles n'ont aucune contribution sur la classification. Ceci est justifié par la redondance de la variable 10 avec la variable 5 et la redondance de la variable 4 avec la variable 9.

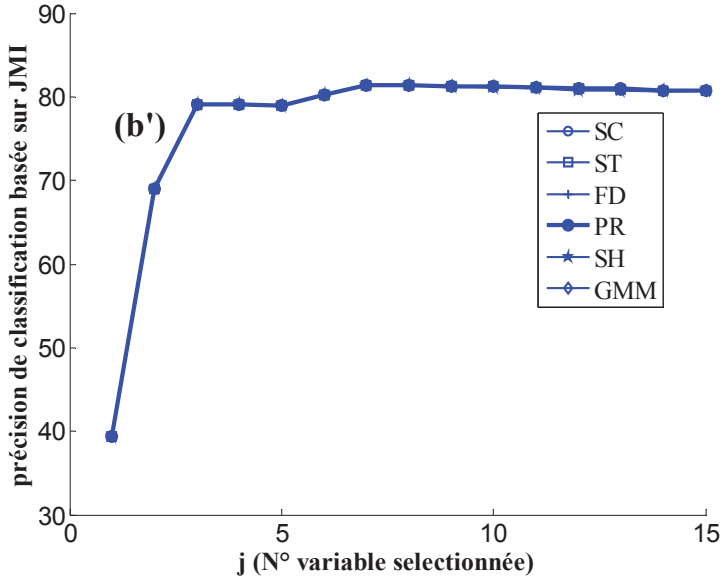
Dans le cas de 50 réalisations par classe (voir figure III.6.a), seulement les méthodes PR et GMM permettent d'avoir approximativement un plateau à partir d'un nombre de variables égal à 10. Ainsi, ces deux méthodes sont les moins erronées. Parmi les autres méthodes d'estimation, la méthode SC est la moins erronée, comparée aux méthodes ST, FD et SH,

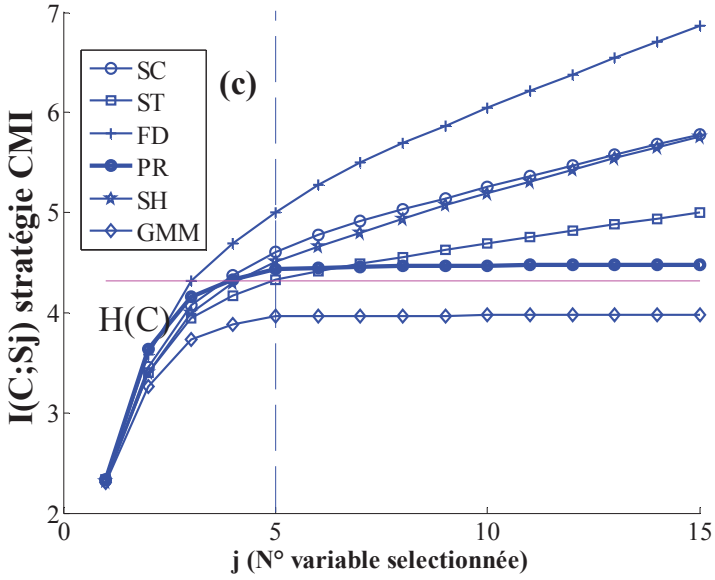
alors que dans le cas de 1000 échantillons par classe, la méthode ST est la meilleure comparée aux méthodes SC, FD et SH. Ces résultats peuvent être confirmés par les tableaux III.2.a et III.2.b. De plus, la figure III.6.a' montre clairement, dans le cas de 50 observations par classe, l'effet du phénomène de la malédiction de la dimensionnalité des données (la décroissance de la précision malgré l'augmentation de la dimension des observations). Ce phénomène causé par le manque de données pour la modélisation des classes exige ainsi une réduction de la dimensionnalité des observations.






(b)
SC
ST
FD
PR
SH
GMM
H(C)
I(C;Sj) stratégie JMI
j (N° variable selectionnée)


(b')
SC
ST
FD
PR
SH
GMM
précision de classification basée sur JMI
j (N° variable selectionnée)


(c)
SC
ST
FD
PR
SH
GMM
H(C)
I(C;Sj) stratégie CMI
j (N° variable selectionnée)

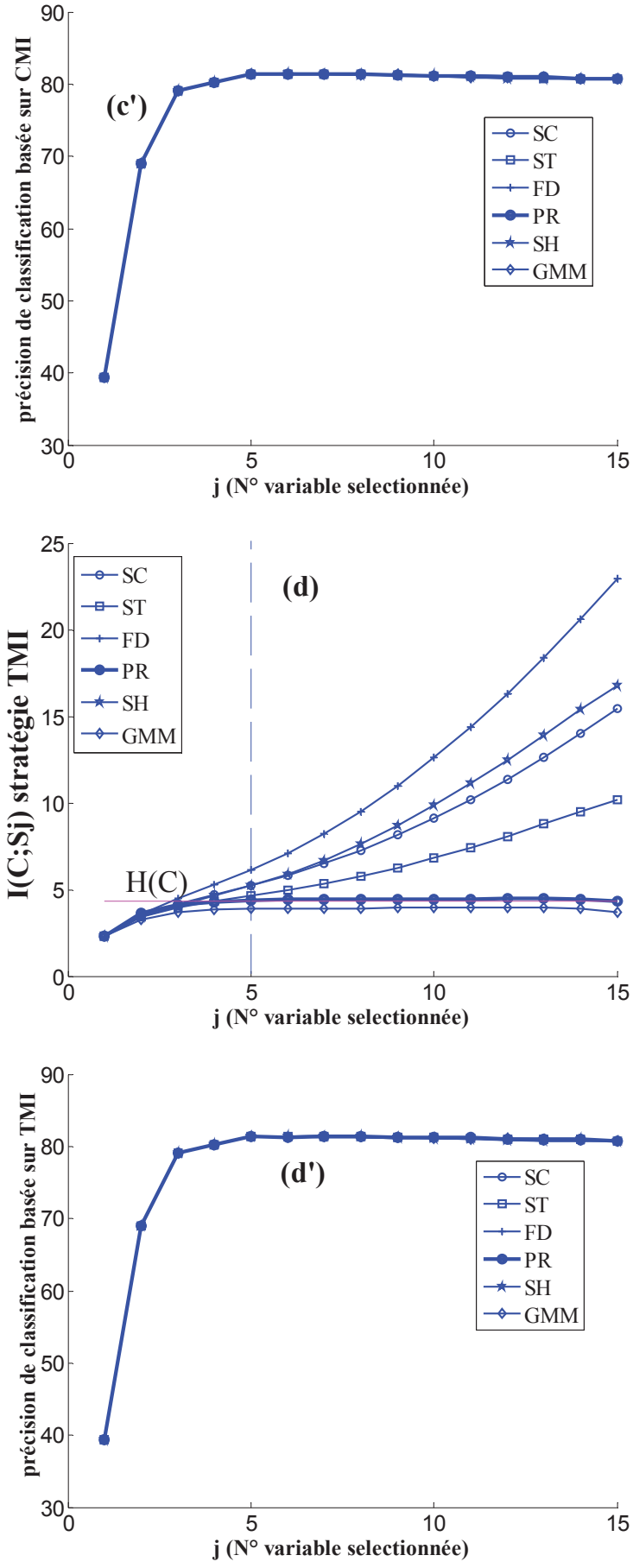
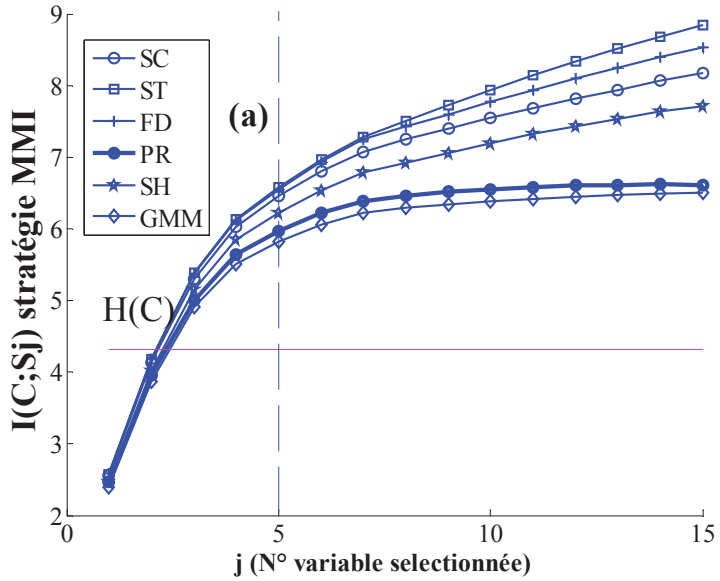
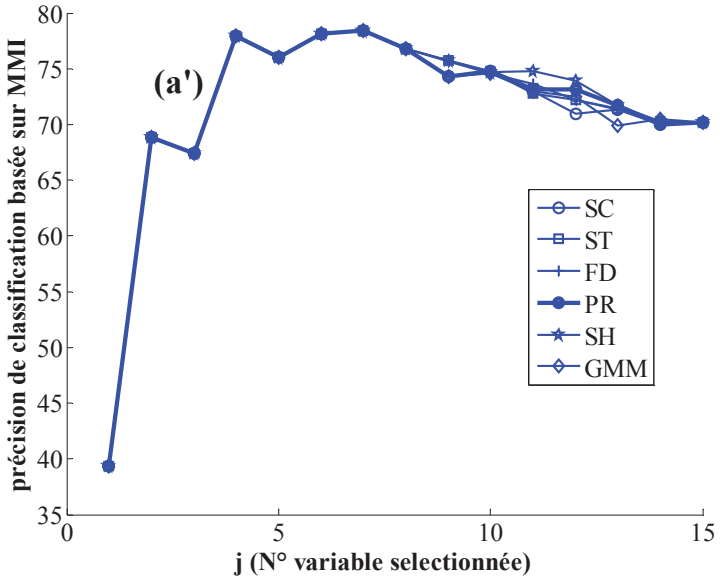
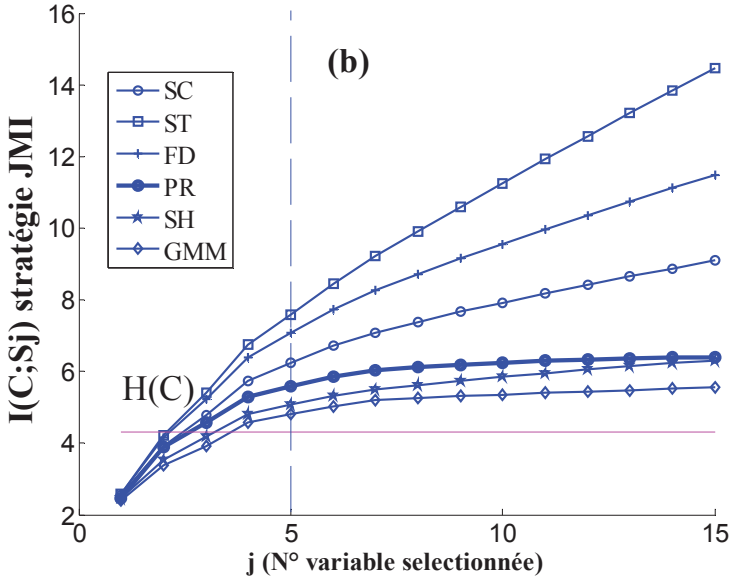
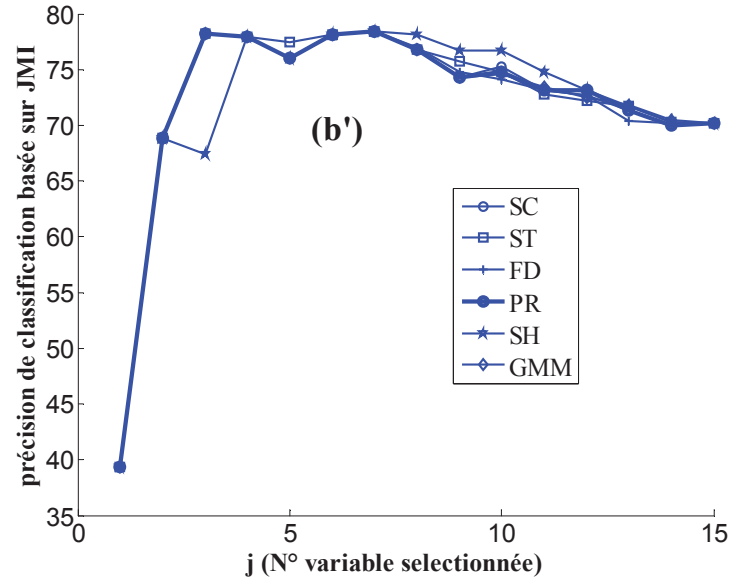


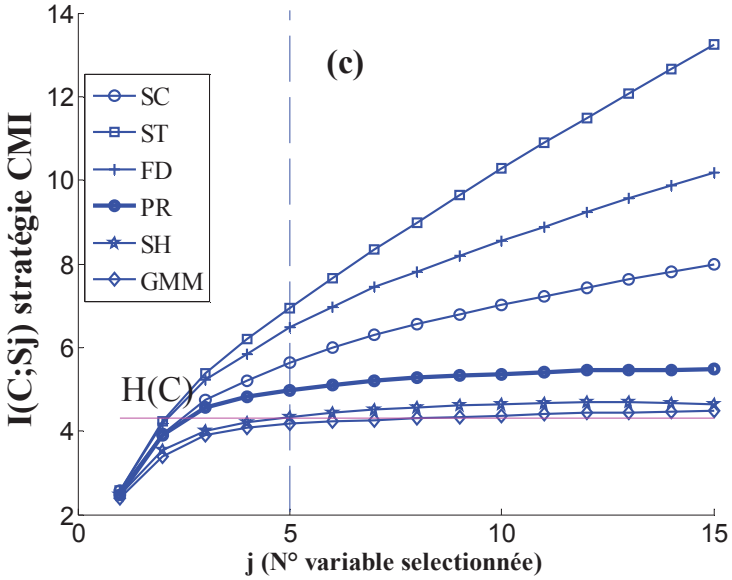
Figure.III.5 : Sélection des paramètres pertinents de type gaussien pour les quatre stratégies: (a) MMI, (b) JMI , (c) CMI and (d) TMI avec 1000 observations gaussiennes par classe.

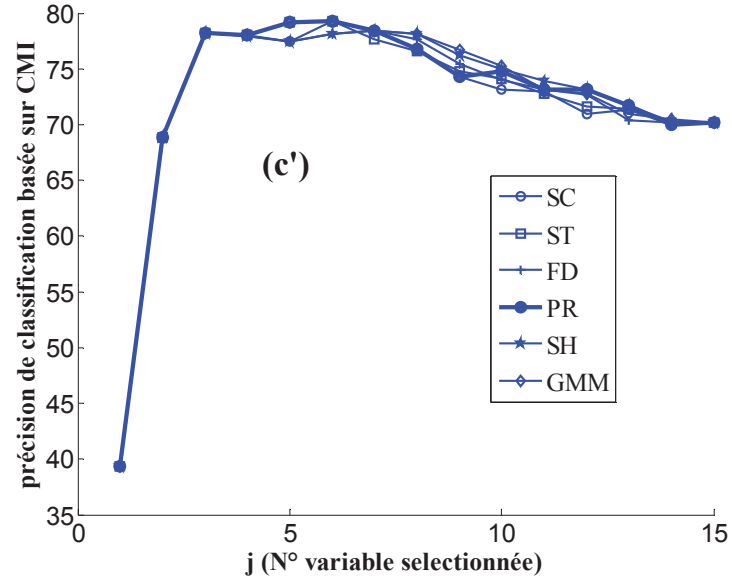

I(C;Sj) stratégie MMI
(a)
SC
ST
FD
PR
SH
GMM
H(C)
j (N° variable selectionnée)


précision de classification basée sur MMI
(a')
SC
ST
FD
PR
SH
GMM
j (N° variable selectionnée)


I(C;Sj) stratégie JMI
(b)
SC
ST
FD
PR
SH
GMM
H(C)
j (N° variable selectionnée)


(b')
précision de classification basée sur JMI
j (N° variable selectionnée)
SC
ST
FD
PR
SH
GMM


(c)
I(C;Sj) stratégie CMI
H(C)
j (N° variable selectionnée)
SC
ST
FD
PR
SH
GMM


(c')
précision de classification basée sur CMI
j (N° variable selectionnée)
SC
ST
FD
PR
SH
GMM

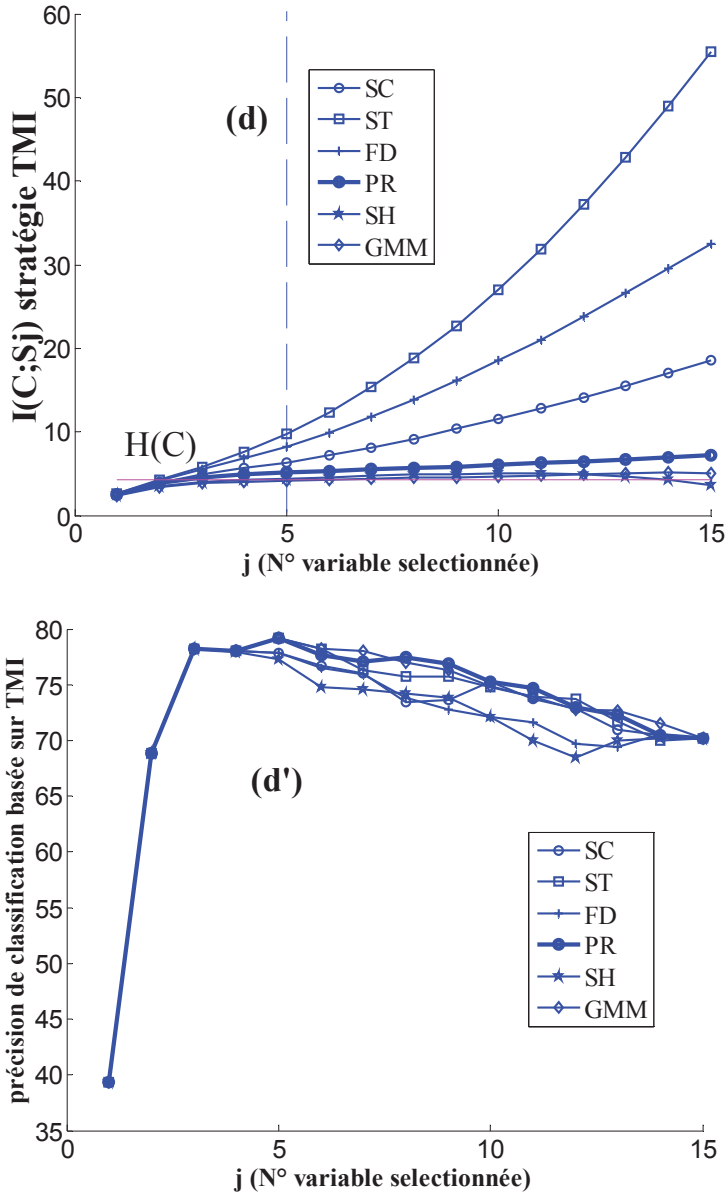


Figure.III.6 : Sélection des paramètres pertinents de type gaussien sous les quatre stratégies: (a) MMI, (b) JMI , (c) CMI and (d) TMI avec 50 observations gaussienne par classe

La figure III.5.b montre les résultats de sélection sous la stratégie JMI. Bien que toutes les méthodes permettent de sélectionner des sous-ensembles de variables conduisant approximativement à la même précision de classification (voir la figure III.5.b'), seulement les méthodes PR et GMM atteignent approximativement un plateau confirmant ainsi l'erreur minimale d'estimation de l'IM commise par ces dernières méthodes. Néanmoins ce plateau atteint une valeur supérieure à $H(C)$ à partir d'un nombre de variables (8) plus grand que le nombre optimal (5), ceci est dû principalement à la non suppression complète de la redondance qui peut être justifiée par l'existence des variables redondantes dans les sous ensemble des 5 premières variables sélectionnées : 5, 4, 3, 10, 9.

La figure III.5.b montre que la méthode ST est la moins biaisée par rapport aux méthodes SC, FD, SH. Ce peut être justifié par le tableau III.2.a qui montre que cette méthode est la moins biaisée dans le cas d'un grand nombre d'observations. Dans le cas de 50 observations par classe, la figure III.6.b montre que les méthodes PR, SH et GMM sont les moins biaisées mais seulement PR et GMM atteignent approximativement un plateau. Cette figure montre aussi que la méthode ST commet une grande erreur qui peut être expliquée par le tableau III.2.b. Du point de vue de classification, la figure III.6.b' montre une croissance de la précision ; ensuite une décroissance à partir d'un nombre de variables égal à 7. Ce comportement justifie clairement le phénomène de la malédiction de la dimensionnalité.

La figure III.5.c résume les résultats dans le cas de 1000 observations par classe sous la stratégie CMI. La figure montre que pour les méthodes GMM et PR, la IM I(C ;S) atteint approximativement un plateau d'un niveau proche de l'entropie à partir d'un nombre de variables égal au nombre optimal, alors que pour les autres méthodes, l'IM continue à croître. Ce résultat est dû principalement à l'erreur minimale d'estimation de l'IM par GMM et PR ainsi que de la suppression presque complète de la redondance. Cette suppression peut être justifiée par l'utilisation de la stratégie CMI qui tient compte de la redondance maximale avec une variable déjà sélectionnée. En plus selon la figure III.5.c', la précision atteint également un plateau à partir du même nombre de variables obtenu ci-dessus.

La figure III.6.c montre, malgré le nombre limité de données, que seulement les méthodes PR, SH et GMM atteignent approximativement un plateau, alors que la figure III.6.c' montre l'accroissement de la précision, ensuite sa décroissance expliquant ainsi le phénomène de la malédiction de la dimensionnalité. Remarquant dans le cas d'un grand nombre de données, la grande erreur d'estimation de l'IM est commise par la méthode FD alors que dans le cas d'un nombre limité de données la grande erreur est commise par la méthode ST. Ce résultat peut être justifié par le tableau III.2.

Sous la stratégie TMI, la figure III.5.d montre que seulement les méthodes PR et GMM atteignent approximativement un plateau d'un niveau proche de l'entropie *H(C)* à partir d'un nombre de variables égal au nombre optimal. Ce résultat est dû principalement à l'erreur minimale d'estimation de l'IM par GMM et PR ainsi que de la suppression complète de la redondance. Cette suppression peut être justifiée par la vérification des hypothèses de la stratégie TMI qui tient compte de toute redondance d'une variable déjà sélectionnée avec celle à sélectionner. En outre la figure III.5.d montre qu'une large erreur d'estimation de l'IM est commise par la méthode FD tandis que la figure III.6.d montre qu'une très large erreur est commise par la méthode ST. Cette dernière figure montre également que les seules méthodes qui convergent vers un plateau sont SH, GMM, PR. En plus selon la figure III.5.d', la

précision de classification atteint également un plateau à partir du même nombre déjà obtenu par l'application de la stratégie TMI, alors que la figure III.6.d' montre une croissance de la précision jusqu'à la 5^{ième} variable ensuite une décroissance justifiant ainsi le phénomène de la malédiction.

Bien que toutes les méthodes ont approximativement la même courbe de précision de classification, seules les méthodes GMM et PR peuvent avoir un plateau et estimer le nombre de variables pertinentes sous les stratégies CMI, TMI soit dans les cas de grand ou petit nombre de données. Néanmoins la méthode GMM exige de fixer plusieurs paramètres comme le nombre maximale d'itérations de l'algorithme EM et le nombre maximal de composantes gaussiennes. Pratiquement on a fixé le nombre d'itérations à 100 et on a utilisé le critère d' Akaike pour le choix du nombre de composantes parmi 20 dans le cas d'un grand nombre de données et 10 dans le cas d'un nombre limité de données. Ainsi, du point de vue du temps de calcul et de la capacité de prédire le nombre de variables pertinentes, la méthode PR s'avère la plus performante selon les résultats présentés ci-dessus.

En plus des résultats de sélection des variables gaussiennes, nous avons appliqué les méthodes SC, ST, FD, PR, SH, GMM sur des variables de type Uniforme et de type Lognormal sous la stratégie TMI pour examiner la robustesse de ces méthodes. En conservant les couples (μ_i, σ_i) du cas Gaussien, nous déduisons les limites $[a_i; b_i]$ de la loi uniforme selon :

$$a_i = \mu_i - \sqrt{3}\sigma_i$$

$$b_i = \mu_i + \sqrt{3}\sigma_i$$

Rappelons que la fdp d'une loi lognormale est définie comme suit :

$$f(x\backslash m, v) = \frac{1}{xv\sqrt{2\pi}} e^{\frac{-(\ln x - m)^2}{2v^2}}$$

Dans ce cas, les paramètres m_i et v_i de la loi Lognormale sont donnés par les formules suivantes :

$$m_i = \log\left(\frac{{\mu_i}^2}{\sqrt{{\sigma_i}^2 + {\mu_i}^2}}\right)$$

$$v_i = \sqrt{\log\left(\frac{{\sigma_i}^2}{{\mu_i}^2} + 1\right)}$$

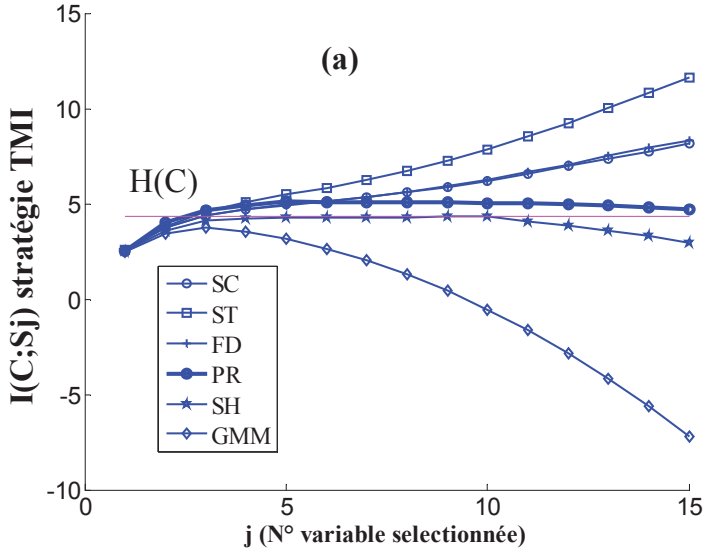
La figure III.7.a montre les résultats de sélection des variables uniformes sous la stratégie TMI dans le cas de 1000 observations par classe, tandis que la figure III.7.a' montre les résultats dans le cas de 50 observations. Plusieurs remarques peuvent être effectuées:

- la méthode PR converge vers un plateau ce qui justifie sa robustesse, alors que la méthode GMM commet une large erreur.
- la méthode SH atteint un plateau à partir de la 5^{ième} variable mais une décroissance apparait après la 10^{ième} variable.
- La méthode ST commet une grande erreur pour un nombre de données limité ou très grand.

La figure III.8 montre les résultats de sélection des variables de type lognormal. Plusieurs remarques peuvent être tirées :

- dans le cas d'un grand nombre de données, seulement la méthode PR atteint approximativement un plateau à partir de la 5ème variable.
- la méthode SH commet une large erreur dans le cas d'un grand nombre de données, alors que la méthode ST est la plus biaisée dans le cas d'un nombre limité de données.
- dans le cas d'un grand nombre de données, la méthode GMM atteint approximativement un plateau à partir de la 5ème variable, ensuite une apparition d'une décroissance montre la mauvaise robustesse de cette méthode.

De ces résultats sur la sélection, on peut conclure que seulement notre méthode est performante dans toutes les conditions (quelque soit le nombre de données par classe, quelque soit le type de variables).



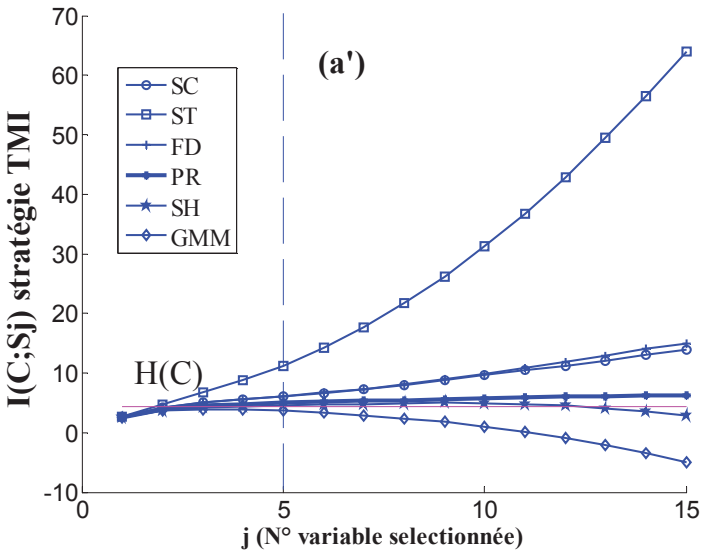


Figure.III.7 : Sélection des paramètres pertinents de type uniforme sous la stratégie TMI avec 1000 observations (a) et 50 observations (a') par classe.

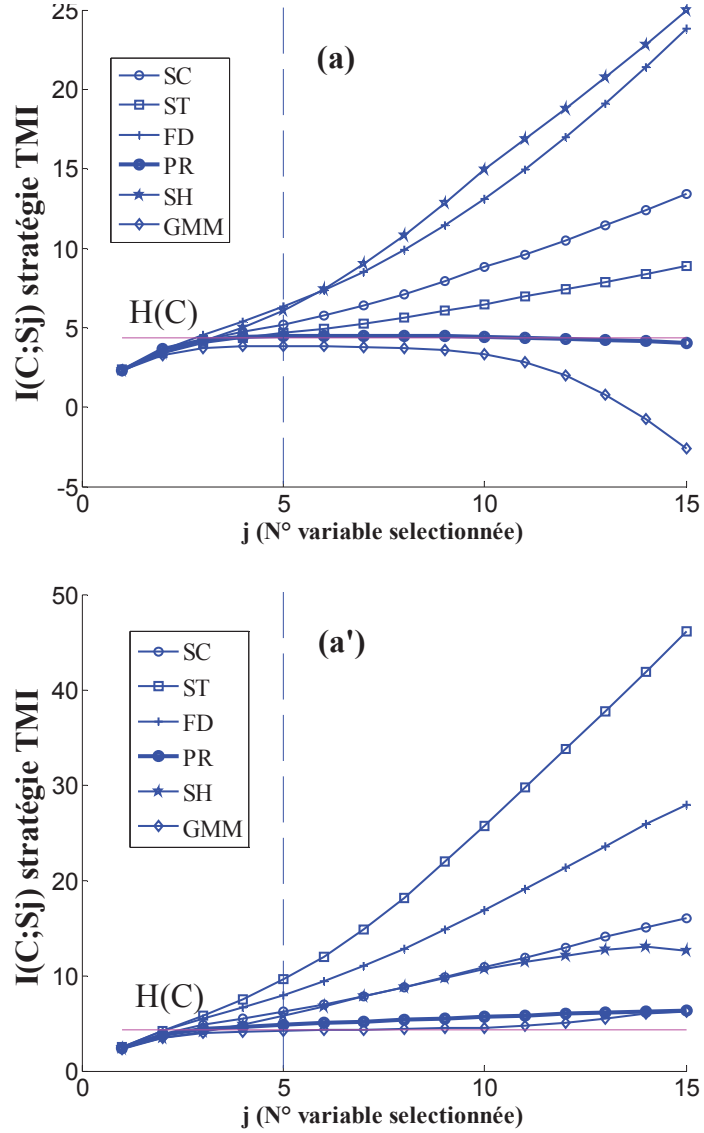


Figure.III.8 : Sélection des paramètres pertinents de type lognormal sous la stratégie TMI avec 1000 observations (a) et 50 observations (a') par classe.

III.6 CONCLUSION

Le calcul de l'information mutuelle à partir des données exige l'estimation des Fonctions de Densités de Probabilités (fdp) conjointes et marginales. En pratique, l'estimation de ces fonctions se heurte à la non connaissance du type de ces fonctions ainsi qu'à un nombre d'échantillons limité. Plusieurs méthodes d'estimation de ces fonctions existent telles que la méthode par histogramme, GMM, KDE,…

Dans notre travail de thèse, nous nous sommes intéressés à la méthode histogramme pour ses avantages en termes de simplicité et de faible complexité de calcul. Cependant cette méthode souffre du choix de nombre de *bins* constituant l'histogramme, ainsi nous avons proposé un nouvel algorithme pour le choix du nombre de *bins* permettant de minimiser le biais et l'erreur quadratique moyenne de l'estimateur de l'information mutuelle.

Ce nouvel estimateur de l'IM a été appliqué pour la sélection des variables pertinentes. Les résultats de simulations sur des variables gaussiennes ont montré que cet estimateur est plus précis que les méthodes de Sturges, de Scott, Fredman, Shimazaki, et comparable à la méthode GMM, qui demeure cependant très lente par rapport à notre méthode. La simulation sur d'autres types de variables a également prouvé la robustesse de notre méthode malgré une hypothèse restrictive de fdp gaussienne.

Dans le chapitre IV, nous appliquons ce nouvel estimateur dans la sélection des paramètres acoustiques pertinents pour une tâche de reconnaissance de la parole.

CHAPITRE IV

SELECTION DES PARAMETRES ACOUSTIQUES POUR LA RECONNAISSANCE DE LA PAROLE

IV.1 INTRODUCTION

Dans le chapitre III, on a présenté une nouvelle méthode de calcul du nombre de *bins* d'un histogramme pour l'estimation de l'entropie et de l'information mutuelle dans le cas des variables continues. Une étude comparative effectuée sur des données simulées montre que la méthode proposée (PR) est la plus performante par rapport aux approches suivantes : mélange de gaussiennes (GMM), Scott (SC), Sturges (ST), Freedman (FD) et Shimazaki (SH), du point de vue du temps de calcul et de l'erreur quadratique moyenne. Dans le domaine de sélection des variables pertinentes, les résultats d'application de ces méthodes sur des données synthétiques ont montré que seulement la méthode proposée a pu avoir approximativement un plateau à partir du nombre optimal des variables pertinentes pour différents types de variables (gaussien, uniforme et lognormal) dans le cas d'un nombre important de données.

L'objectif de ce chapitre est d'appliquer ces méthodes sur des données de la parole. Plus particulièrement on essaye d'évaluer les performances de ces méthodes dans l'étape de sélection des paramètres acoustiques les plus pertinents pour une tâche de reconnaissance des mots connectés. Cette étape permet à partir d'un ensemble initial de paramètres acoustiques de réduire la dimension des vecteurs acoustiques en ne retenant que les paramètres jugés utiles et non redondants pour la discrimination entre les mots constituant le vocabulaire considéré.

Les coefficients MFCC à court terme sont généralement les paramètres acoustiques les plus utilisés en RAP et ils sont désignés par les paramètres statiques. Dans des environnements en laboratoire, les MFCC associés à une modélisation statistique par le moyen des Modèles HMM et des modèles GMM ont atteint de bonnes performances en RAP [115].

Cependant, dans le cas de la reconnaissance de la parole bruitée, les performances se dégradent rapidement et il est devenu pratiquement indispensable d'inclure les coefficients différentiels du premier (delta) et second ordre (delta-delta) en tant que partie du vecteur des paramètres acoustiques. Ces coefficients nommés par les paramètres dynamiques, fournissent une information sur la trajectoire temporelle de la parole [116]. Ainsi, les coefficients MFCC statiques et dynamiques constituent un vecteur acoustique.

Dans ce chapitre, nous présentons tout d'abord le problème de la sélection des paramètres acoustiques MFCC statiques et dynamiques dans différents environnements pour la reconnaissance vocale. Puis, l'effet de l'augmentation de la dimension des vecteurs acoustiques sur les performances du système RAP est abordé ainsi que la minimisation de cet effet par la sélection des paramètres acoustiques les plus pertinents. Cette augmentation de la dimension est effectuée en combinant les paramètres MFCC avec les paramètres statiques et dynamiques des coefficients PLP, LPCC.

Afin d'évaluer les performances d'un système RAP basé sur les paramètres acoustiques sélectionnés, il est nécessaire de les comparer à celles d'un système existant. Ainsi, on a proposé le système de référence pour la reconnaissance des chiffres présenté dans [14]. Ce système est évalué sur la base de données Aurora2 distribuée par ELRA [16]. Dans la section suivante, nous décrivons brièvement ce système et les résultats de la reconnaissance sur cette base de données.

IV.2 RESULTATS EXPERIMENTAUX DU SYSTEME DE REFERENCE

Le système RAP de référence permet la reconnaissance des séquences de chiffres en mode indépendant du locuteur fonctionnant sous différentes conditions de bruit [14]. Ce système est basé sur la modélisation de chaque chiffre par un HMM.

Toutes les données de parole utilisées dans l'apprentissage des modèles HMM et l'évaluation des performances du système appartiennent à la base Aurora2.

IV.2.1 BASE DE DONNEES AURORA2

La source de parole de la base Aurora2 est la base TIdigits [117]. Cette dernière contient des enregistrements d'adultes Américains masculins et féminins qui prononcent des séquences au maximum de 7 chiffres connectés. Les signaux originellement échantillonnés sur 20 Khz sont sous échantillonnés à 8 Khz. Ces signaux sont considérés comme des données sans bruit. Des distorsions sont artificiellement ajoutées à ces données pour définir les sous-ensembles de la base Aurora2 soit dans l'apprentissage soit dans le test.

Deux modes d'apprentissage sont utilisés [14]:

- Apprentissage avec des données non bruitées.
- Apprentissage multiconditions utilisant des données bruitées et des données sans bruit.

Dans le cas du premier mode, 8440 séquences sont sélectionnées de la partie d'apprentissage de la base TIDigits contenant des enregistrements de 55 hommes et 55 femmes adultes. Afin de considérer des caractéristiques fréquentielles réelles des terminaux et des équipements dans le domaine des télécommunications, les signaux de ces séquences ont

subit un filtrage appelé G.712 selon [14] sans ajouter aucun bruit. Les mêmes séquences sont prises pour le deuxième mode. Elles sont divisées en 20 sous-ensembles avec 422 séquences dans chacun. Chaque sous-ensemble contient quelques séquences de tous les locuteurs. Les 20 sous-ensembles représentent 4 scénarios de bruits différents sous 5 niveaux du Rapport Signal sur Bruit (RSB). Les 4 bruits sont: *suburban train, babble*, *car* et *exhibition hall*. Les RSBs sont 20 dB, 15 dB, 10 dB, 5 dB, et la condition sans-bruit. La parole et le bruit sont subit un filtrage G.712 avant d'être ajoutés. Les ensembles de ces deux modes sont utilisés premièrement dans l'apprentissage des modèles HMM mots, deuxièmement dans la sélection des paramètres pertinents.

Concernant le test, trois ensembles différents appelés respectivement A, B et C sont définis dans la base Aurora2 pour l'évaluation du système de la reconnaissance. 4004 séquences de 52 hommes et 52 femmes de la partie test de la base TIdigits, sont divisées en 4 sous-ensembles avec 1001 séquences dans chacune. Des enregistrements de tous les locuteurs sont présents dans chaque sous-ensemble. Un seul type de bruit est ajouté pour chaque sous-ensemble de 1001 séquences sous différents RSBs : 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, -5 dB. En plus, le cas sans-bruit est pris comme la $7^{ième}$ condition. La parole et le bruit sont subis un filtrage G.712 avant d'être ajoutés.

Dans l'ensemble de test A, les 4 bruits: *suburban train, babble, car* et *exhibition hall* sont ajoutés à ces quatre sous-ensembles. Le total cet ensemble est constitué de 4 fois de 7 fois de 1001 = 28028 séquences. Il contient les mêmes bruits utilisés dans l'apprentissage multiconditions qui justifient probablement les bonnes performances du système dans le cas d'évaluation sur cet ensemble.

Les ensembles de test B et C sont crées de la même façon que l'ensemble A, mais l'ensemble B est construit sur des bruits différents de ceux utilisés dans l'apprentissage muticonditions, alors que la génération de l'ensemble C est effectuée principalement sur un filtrage différent de celui utilisé dans la génération des ensembles des modes d'apprentissage. Le tableau (IV.1) résume les différents ensembles d'apprentissage et de test qui composent la base Aurora2.

La pertinence des paramètres acoustiques est validée en mesurant la précision de reconnaissance des séquences enregistrées dans les mêmes conditions de celles utilisées dans la procédure de sélection. Ainsi, seulement les séquences de l'ensemble A créées sous les 20 conditions du deuxième mode d'apprentissage sont considérées pour l'évaluation des performances du système de reconnaissance. Dans le cas du premier mode, seulement 4 sous-ensembles de l'ensemble A correspondant à la condition sans-bruit sont considérés. Donc 4004 séquences de test sont utilisées dans le cas d'apprentissage sans-bruit et 20020

séquences sont utilisées dans le cas d'apprentissage multiconditions. Dans la suite on décrit brièvement le système de référence ainsi que les résultats de reconnaissance de l'ensemble de test A sous la plate forme HTK.

Tableau IV.1 : Définitions des ensembles d'apprentissage et de test de la base Aurora2

	Données d'apprentissage		Données de test		
	Condition clean	Multi-condition	Ensemble A	Ensemble B	Ensemble C
Filtre	G.712				MIRS
RSB(dB)	Clean	Clean,20,15,10,5	Clean,20,15,10,5,0,-5		
Bruit		Subway Babble Car exhibition	Subway Babble Car Exhibition hall	Restaurant Street Airoport Station	Subway Street

IV.2.2 SYSTEME DE REFERENCE SOUS PLATE FORME HTK [14]

Le système de référence basé sur les modèles HMM est implémenté à partir de la plate-forme HTK. Toutes les étapes du système comme l'analyse acoustique, l'apprentissage des modèles HMM, la reconnaissance des séquences des chiffres inconnus ainsi que le résultat d'évaluation sur un ensemble de séquences de test sont effectuées en utilisant la version 4.3 de la librairie HTK. Ainsi, un certain nombre de choix sont faits sur ce système, comme le nombre d'états des modèles, le type de densités de probabilité d'émission associées aux états et l'espace de représentation du signal par des coefficients acoustiques.

Dans la phase d'apprentissage, tout signal d'une séquence de mots de l'ensemble d'apprentissage est transformée en une séquence de vecteurs acoustiques pour qu'elle soit utilisée en tant que séquence d'observations d'entrée dans la modélisation markovienne des mots. L'ensemble des paramètres constituant chaque vecteur acoustique est constitué de 12 coefficients cepstraux en échelle de fréquence Mel MFCC (excepté le coefficient 0), le logarithme de l'énergie, leurs coefficients delta et delta-delta. Cet ensemble de 39 coefficients MFCC est calculé sur chaque trame d'analyse de 25ms avec un chevauchement de 10ms (outil *HCopy* de la librairie HTK). Les opérations appliquées sur chaque trame d'analyse sont:

- Préaccentuation du signal avec un facteur de 0.97.
- Application d'une fenêtre de Hamming.
- Application de la transformée cosinus DCT sur les logarithmes d'énergies des bancs de filtres en échelle Mel ayant 23 bandes de fréquence dans la gamme de 64 hertz jusqu'à la moitié de la fréquence d'échantillonnage.

Après l'analyse acoustique, chaque chiffre est modélisé par un HMM représentant entièrement un mot avec les paramètres suivants :

- 16 états par mot (plus 1 état d'entrée et 1'état de sortie).
- Modèle gauche-droite sans saut sur les transitions.
- GMM de 3 Gaussiennes par état.
- Seulement les variances des paramètres sont prises en compte (matrice de covariance diagonale).

Deux modèles de pause sont définis. Le premier, appelé 'sil', est constitué de 3 états dont chacun est modélisé par un GMM de 6 Gaussiennes. Le premier modèle illustré dans la figure (IV,1) modélise le silence du début et de la fin d'une séquence de mots. Le deuxième, appelé 'sp', est constitué d'un état partagé avec l'état milieu de 'sil'; il est utilisé pour modéliser les pauses entre les mots. Ainsi, le nombre d'états de tous les modèles HMM créés est égal à 180 états.

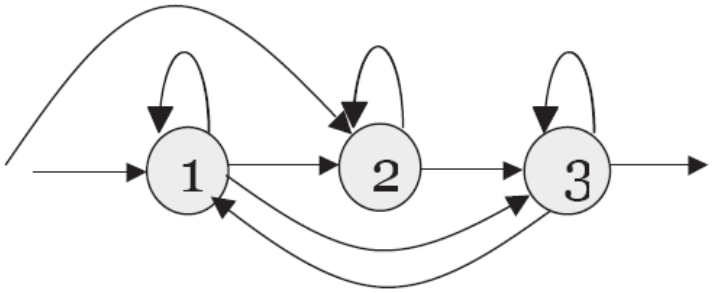


Figure.IV.1 : Transitions possibles dans le modèle 'sil'

L'apprentissage est effectué en plusieurs étapes en appliquant l'algorithme de réestimation de Baum-Welch (outil *HERest* de HTK):

- Initialisation de tous les modèles HMM mots et du modèle 'sil' de 3 états avec des moyennes et des variances globales (*HcompV* de HTK). Les modèles mot et 'sil' contiennent seulement une gaussienne par état dans cette étape d'initialisation.
- Trois itérations de réestimation Baum-Welch (option d'élagage - t de l'outil *HERest* à 250 150 1000).
- Introduction du modèle pause 'SP' entre les mots, augmentation du nombre de gaussiennes à 2 pour les états des modèles 'sil' et pause, application encore de trois itérations de réestimation Baum-Welch.
- Augmentation du nombre de gaussiennes à 2 pour tous les états des modèles mots, augmentation du nombre de gaussiennes à 3 pour tous les états des modèles 'sil' et 'SP' et application de trois autres itérations de réestimation Baum-Welch.

- Augmentation du nombre de gaussiennes à 3 pour tous les états des modèles mots, augmentation du nombre de gaussiennes à 6 pour tous les états des modèles 'sil' et 'SP' et application de sept itérations de réestimation Baum-Welch.

Durant la reconnaissance, tout signal d'une séquence de mots de l'ensemble de test est transformé en une séquence de vecteurs acoustiques lequel sera transcrit en une séquence de mots (outil *HVite* de HTK). La grammaire accepte toute séquence constituée de n'importe quelle combinaison de chiffres, débutée et terminée par un silence avec la possibilité de pause entre les chiffres.

Les performances du système sont basées sur le calcul de la précision de la reconnaissance. Cette précision est définie comme le nombre de mots correctement reconnus moins les mots insérés, le tout divisé par le nombre total de mots testés :

$$\text{précision} = \frac{H-I}{N} \qquad \text{(IV.1)}$$

où N est le nombre total des labels (mots) dans la transcription de référence, H est le nombre de mots correctement reconnus et I est le nombre de mots insérés. Cette précision est basée sur une comparaison entre les mots de chaque transcription de référence d'un signal et ceux de sa transcription obtenue par l'outil *HVite*. Le calcul de cette précision est effectué en utilisant l'outil HTK *HResult*.

Le tableau (IV.2) présente la précision du système de reconnaissance en mode d'apprentissage sans-bruit utilisant une analyse acoustique qui produit des vecteurs acoustiques de 39 coefficients MFCC. Cette précision est calculée sur les séquences de l'ensemble A sans les RSB de 0 et -5 dB.

Tableau.IV.2 Précision du système pour l'ensemble de test A en mode d'apprentissage sans-bruit

Bruit / RSB/dB	Subway	Babble	Car	Exhibition	Moyenne
Sans-bruit	98.83	98.97	98.81	99.14	98.94
20	96.96	89.96	96.84	96.20	94.99
15	92.91	73.43	89.53	91.85	86.93
10	78.72	49.06	66.24	75.10	67.28
5	53.39	27.03	33.49	43.51	39.35
Moyenne	84.16	67.69	76.98	81.16	77.50

Le tableau (IV.3) présente la précision du système de reconnaissance en mode d'apprentissage multiconditions utilisant une analyse acoustique qui produit des vecteurs

acoustiques de 39 coefficients MFCC. Cette précision est calculée sur les séquences de mots de l'ensemble de test A sans les RSB de 0 et -5 dB.

Tableau IV.3 Précision du système pour l'ensemble de test A en mode d'apprentissage multi-conditions

Bruit / RSB/dB	Subway	Babble	Car	Exhibition	Moyenne
Sans-bruit	98.59	98.52	98.48	98.55	98.54
20	97.64	97.61	97.85	96.98	97.52
15	96.75	96.80	97.64	96.58	96.94
10	94.38	95.22	95.65	93.12	94.59
5	88.42	87.67	87.24	86.95	87.57
Moyenne	95.16	95.16	95.37	94.46	95.03

L'avantage du mode d'apprentissage sur des données non bruitées est seulement la modélisation de la parole sans distorsion par n'importe quel type de bruit. Les modèles obtenus sont les meilleurs pour représenter toute l'information disponible de la parole. Ainsi selon les tableaux (IV.2) et (IV.3), les meilleures performances sont obtenues avec ce type d'apprentissage dans le cas du test sur des données non bruitées. Néanmoins ces modèles ne contiennent aucune information sur les distorsions possibles qui influent sur les performances dans le cas du test sur des données bruitées. L'avantage de l'apprentissage multiconditions est d'avoir cette information dans ces modèles. Ce type d'apprentissage conduit aux bonnes performances du système de reconnaissance quand les données de test et d'apprentissage sont prises dans les mêmes conditions de bruit. Ceci est justifié pratiquement par les résultats du tableau (IV,3).

Dans la section suivante, le problème de la sélection des paramètres MFCC parmi les 39 paramètres décrits précédemment est discuté.

IV.3 SELECTION DES PARAMETRES ACOUSTIQUES

Dans cette section nous abordons le problème de la sélection des paramètres acoustiques les plus pertinents pour la tâche reconnaissance des chiffres de la base Aurora2. Dans notre travail, les paramètres à sélectionner sont jugés pertinents s'ils expliquent mieux les données utilisées dans l'apprentissage des modèles HMM. Ainsi, tous les vecteurs acoustiques obtenus par l'analyse acoustique dans la phase d'apprentissage sont utilisés dans la procédure de sélection des paramètres. Chaque séquence de vecteurs acoustiques de l'ensemble

d'apprentissage est segmentée automatiquement en une séquence d'états HMM (outil *HVite* de HTK). Ainsi chaque vecteur est classifié par un label représentant un état parmi 180 états des modèles HMM [70]. Les étapes nécessaires pour la sélection des paramètres acoustiques sont résumées dans la figure (IV.2).

Trois modes d'apprentissage sont considérés pour cette tâche de sélection:

1. Mode ″clean″ basé sur l'ensemble de données d'apprentissage sans bruit de la base Aurora2 (8440 séquences de mots correspondent aux 1468957 vecteurs acoustiques).
2. Mode ″clean réduit″ basé sur 15.64 % de l'ensemble de données d'apprentissage sans bruit de la base Aurora2 (1320 séquences de mots correspondant aux 260522 vecteurs acoustiques).
3. Mode ″multiconditions″ basé sur l'ensemble de données d'apprentissage multiconditions de la base Aurora2 (8440 séquences de mots correspondent aux 1468957 vecteurs acoustiques).

Ces deux derniers modes sont considérés pour étudier respectivement l'influence du nombre de données réduit et l'environnement bruité sur la sélection des paramètres acoustiques.

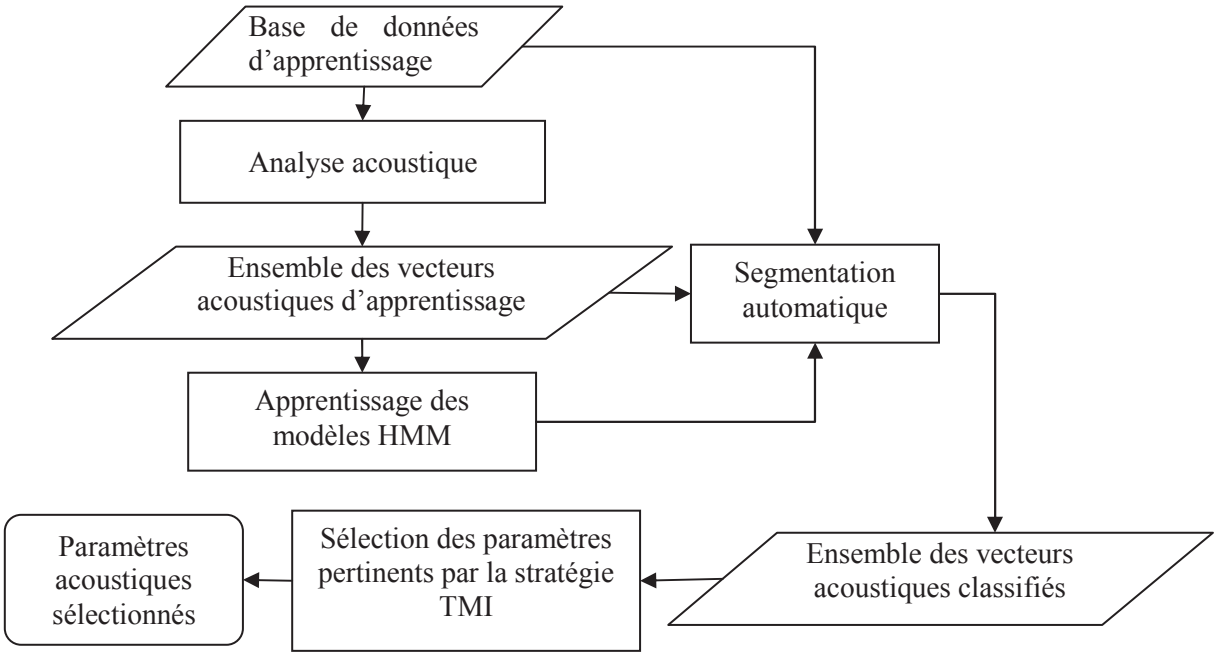


Figure.IV.2 : les étapes nécessaires pour la sélection des paramètres acoustiques pertinents

On propose dans ce travail d'appliquer la stratégie TMI pour la sélection des paramètres pertinents puisque, premièrement, elle se base sur un fondement théorique de l'estimation de l'information mutuelle conjointe entre la variable index de classe C et l'ensemble des variables candidates à sélectionner et deuxièmement, cette stratégie a donné pratiquement de

bons résultats dans le cas des données simulées vérifiant les hypothèses de cette stratégie. L'objectif est de comparer les performances des méthodes d'estimation de l'information mutuelle : PR, GMM, SC, ST, FD et SH dans la sélection des paramètres acoustiques en appliquant cette stratégie.

Dans la section suivante, on applique la procédure de sélection à partir de l'ensemble des 39 coefficients MFCC statiques et dynamiques en considérant les données des trois modes d'apprentissage.

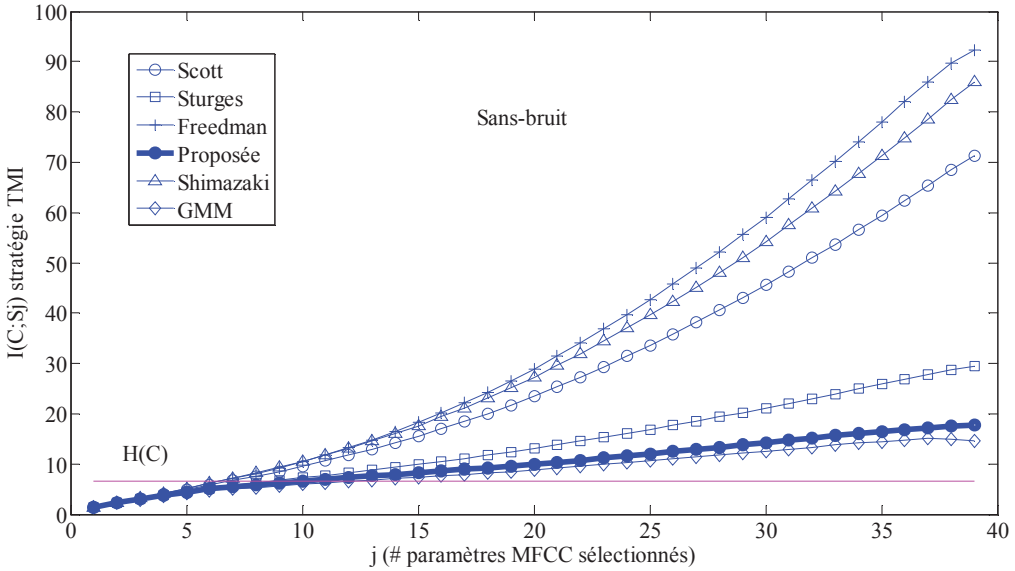
IV.3.1 SELECTION DES PARAMETRES MFCC

Considérant la variable index de classe C qui prend sa valeur dans l'ensemble des états HMM {1,…,180} et considérant également l'ensemble initial des variables S constitué des 39 paramètres acoustiques MFCC ordonnés comme suit:

$S = \{MFCC_1,\ldots,MFCC_{12},E, \Delta MFCC_1,\ldots, \Delta MFCC_{12}, \Delta E, \Delta\Delta MFCC_1,\ldots, \Delta\Delta MFCC_{12}, \Delta\Delta E\}$ où ΔMFCC représente le delta MFCC et ΔΔMFCC représente le delta-delta MFCC.

Les différents calculs pour cette sélection sont effectués sur un PC doté d'un microprocesseur I5 d'horloge 3 Ghz et d'une mémoire RAM de 8 Go.

Dans le cas de la méthode GMM, le nombre d'itérations est fixé à 10, et le nombre de gaussiennes est fixé à 20 pour l'estimation des fdp sur l'ensemble de données de toutes les classes, alors que ce nombre de composantes est fixé au maximum à 3 pour l'estimation sur l'ensemble de données d'une seule classe. Contrairement à [10], le choix de *bins* d'un histogramme bidimensionnel est choisi différent d'une variable à l'autre pour les méthodes de Shimazaki, de Freedman et de Scott.



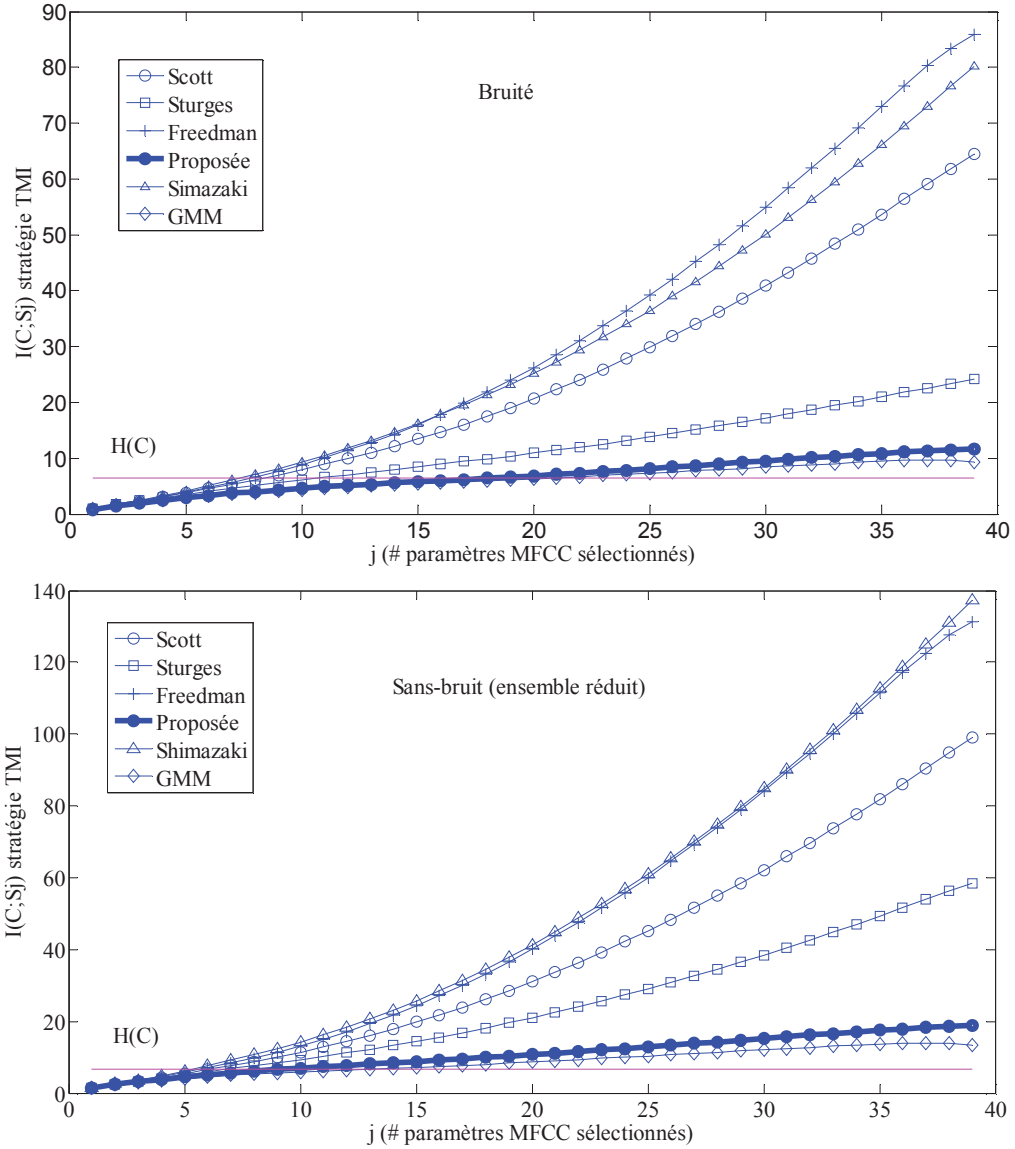


Figure. IV.3: Estimation de l'IM conjointe basée sur la méthode d'histogramme utilisant la stratégie TMI: Scott (○), Sturges (□), Freedman (+), Shimazaki (Δ), GMM (◊) et la méthode proposée (●). La ligne horizontale indique la valeur de l'entropie H(C)

La figure (IV.3) montre les résultats d'application de la stratégie TMI dans les trois cas d'apprentissage. Il est possible d'estimer l'entropie de l'index de classe $H(C)$ qui est égale à 6,69 bits dans le premier cas, 6,72 bits dans le deuxième cas et à 6,46 bits dans le troisième cas. Il est prévu que le nombre de paramètres pertinent atteigne sa valeur optimale quand l'IM atteint un plateau dont sa valeur maximale égale $H(C)$. Différemment à l'expérience de sélection des variables pertinentes présentée dans le chapitre 3, aucune méthode n'a montré un plateau. Les raisons possibles peuvent venir du fait que les données sont non gaussiennes (les paramètres d'énergie sont non gaussiens) [90], où que les termes des IM multivariées d'ordre supérieur à 3 aient été négligés dans la stratégie TMI sans connaître leur impact réel. Cette hypothèse de données non gaussiennes peut être confirmée pratiquement par la figure (IV.4) qui illustre les histogrammes des paramètres énergétiques : énergie E, ΔE et ΔΔE de la

classe de données c=51 correspondant à l'état actif 3 du modèle HMM « nine ». Ces histogrammes sont construits sur un nombre de bins égal à 30.

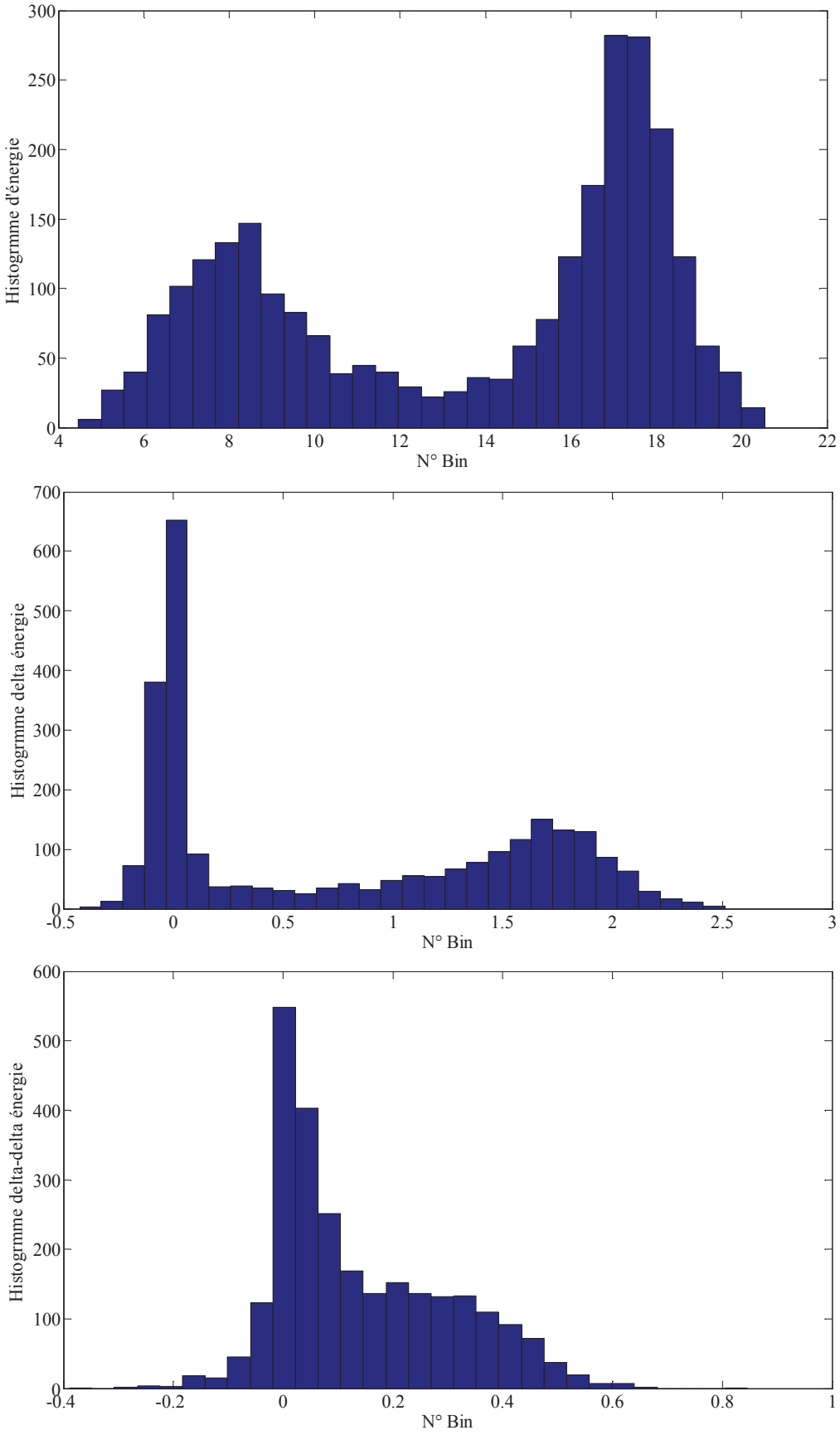


Figure.IV.4: Histogramme des paramètres énergétiques des données de la classe 51 du modèle HMM « nine ».

Cependant, on constate que pour un nombre donné de paramètres, la méthode proposée et celle de GMM présentent une meilleure estimation de l'IM qui atteint des valeurs maximales proches de l'entropie *H(C)*, alors que les autres méthodes peuvent avoir des valeurs très supérieures à cette entropie.

IV.3.2 ESTIMATION DU NOMBRE OPTIMAL DE PARAMETRES PERTINENTS

Afin d'estimer le nombre optimal des paramètres pertinents, on propose un simple critère d'arrêt de la procédure TMI. Ce critère s'exprime de la manière suivante : si la courbe de l'IM excède le niveau de l'entropie $H(C)$ alors le nombre optimal de paramètres correspond au point de croisement entre ce niveau d'entropie et la courbe de l'IM ; si l'IM est au-dessous de ce niveau et si aucun plateau ne se présente, alors le nombre optimal de paramètres correspond à la valeur maximale de la courbe IM. En utilisant ce critère, on obtient dans le cas d'apprentissage sans-bruit pour les méthodes de Freedman, de Shimazaki, de Scott, de Sturges, GMM et la méthode proposée, respectivement un nombre de paramètres pertinents de 7, 7, 7, 9, 12 et 10. Dans le cas d'apprentissage sans-bruit avec un nombre de données réduit, on obtient respectivement 6, 5, 6, 7, 13 et 9 paramètres. Dans le cas d'apprentissage multiconditions, on obtient respectivement 8, 7, 8, 11, 19 et 21 paramètres pertinents.

Dans le but d'analyser ces résultats, on a utilisé le système de reconnaissance de la parole décrit ci-dessus pour estimer la précision du système en fonction du nombre des paramètres MFCC sélectionnés. Ainsi la phase d'apprentissage et la phase d'évaluation de la précision se répètent pour chaque sous-ensemble de paramètres acoustiques sélectionnés dans une itération de la procédure TMI. La figure (IV.5) illustre cette précision dans le cas d'application de la TMI combinée avec la méthode proposée. La précision est évaluée en moyennant les précisions obtenues sur les différents sous-ensembles de test considérés. Dans n'importe quel type d'apprentissage, les résultats de la figure (IV.5) montrent que la précision augmente jusqu'à ce qu'elle atteigne un plateau correspondant au nombre optimal des paramètres pertinents. Ce nombre a approximativement la même valeur que celui obtenu par l'application de la TMI utilisant la méthode proposée et celle de GMM.

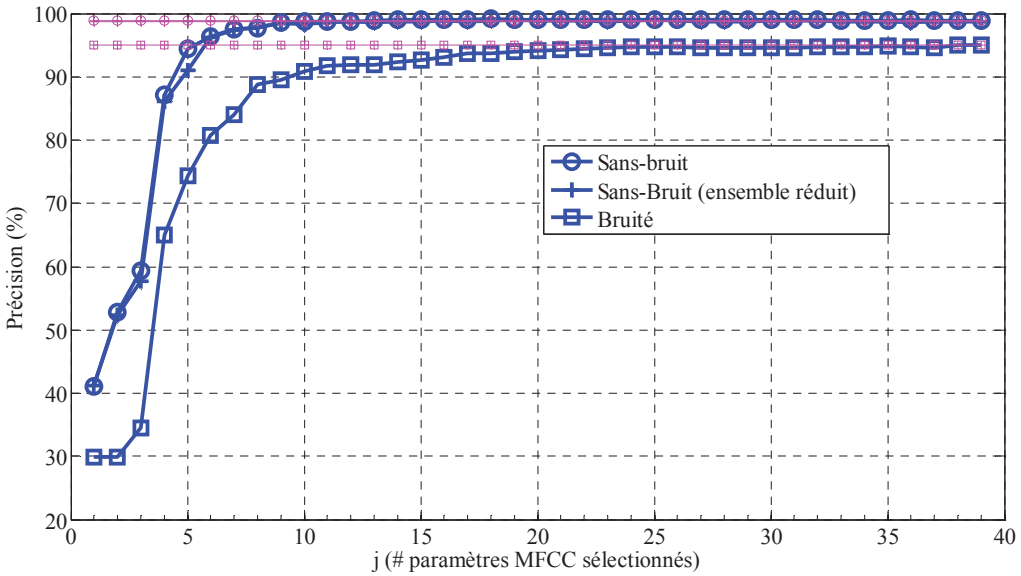


Figure. IV.5: Précision (%) du système RAP en fonction du nombre des paramètres sélectionnés par la stratégie TMI combinée avec la méthode proposée. Les modèles entrainés et testés utilisant l'ensemble clean (○), l'ensemble réduit clean (+) où l'ensemble bruité (□). Les lignes horizontales correspondent à la précision obtenue avec un large vecteur acoustique (39 paramètres).

Le tableau IV.4 donne plus de détails comparatifs sur la sélection des paramètres acoustiques ainsi que sur la précision de reconnaissance. Ces détails sont : le temps de calcul *TC* pour la sélection (jusqu'à 39 paramètres), la valeur maximale de la IM (Vmax) sur les différentes itérations de la procédure de sélection, nombre de paramètres pertinents, le sous-ensemble des paramètres pertinents, ainsi que la précision correspondant à ce sous-ensemble. La meilleure méthode est celle qui propose le meilleur compromis entre une grande précision et un petit nombre de paramètres sélectionnés avec un temps de calcul réduit.

Tableau IV.4: Résultats de sélection des paramètres acoustiques pertinents utilisant la stratégie TMI

Tableau IV.4.a : Données non bruitées

Méthode	Scott	Sturges	Freedman	Proposée	Shimazaki	GMM
Nombre paramètres pertinent	7	9	7	10	7	12
Valeur max de IM	71.37	29.45	92.28	17.71	86.02	15.06
Temps calcul (s)	1.48e+3	1.402e+3	1.88e+3	1.37e+3	7.38e+4	6.44e+4
Sous ensemble des paramètres pertinents	13, 26,2,1,15, 14, 39	13, 26,1,2,14 ,15,39,16,3	13,26,2,1,15, 14,4	13,26,39,2,1, 15,14, 28,16,3	13,26,2,1,15, ,14,39	13,26, 2,15,1, 14,16,4,3,5,28 ,18
Précision de reconnaissance (%)	97.57	98.59	97.46	98.67	97.57	98.68
Précision (39 paramètres) (%)	98.93					

Tableau IV.4.b : Données non bruitées (ensemble réduit)

Methode	Scott	Sturges	Freedman	Proposée	Shimazaki	GMM
Nombre paramètres pertinent	6	7	6	9	5	13
Vmax de IM(bit)	99.01	58.50	131.30	18.94	137.38	13.98
Temps calcul (s)	349.88	329.07	446.85	298.26	3.36e+4	1.74e+4
Sous ensemble des paramètres pertinents	13, 26,2,15,1, 14	13, 26,1,2,15, 14,39	13, 26,1,2,15, 16	13, 26,39,2,15 ,1,14,29,16	13,26,39,15, 2	13,26,2,15,1, 14,16,17,4,5,2 9,6,19
Précision de reconnaissance (%)	96.90	97.31	97.65	98.43	90.95	98.37
Précision (39 paramètres) (%)	98.71					

Tableau IV.4.c : Données bruitées

Methode	Scott	Sturges	Freedman	Proposée	Shimazaki	GMM
Nombre paramètres pertinent	8	11	8	18	7	21
Vmax de IM	64.52	24.17	85.90	11.55	80.25	9.62
Temps calcul (s)	1.55e+3	1.44e+3	1.87e+3	1.37+3	1.21e+5	6.34e+4
Sous ensemble des paramètres pertinents	26,13, 2,1,15, 4, 39,3	26,13,39,2,1, 15,14,4,3,6,16	26,13,2,1,15,4 ,6,3	26,13,39,2,1, 15, 4,3,6,16,5, 28, 29,7,17,8, 18,19	26,13,2,1,15, 39,4	26,13,2,15,39, 4,1,5,16, 6,17, 29,19,18,8,7,3 ,36,9,38,20
Précision de reconnaissance (%)	88.78	91.78	88.65	93.71	84.04	94.23
Précision (39 paramètres) (%)	95.03					

Dans le cas des données non bruitées, une sélection aux alentour de 10 paramètres est suffisante pour obtenir une précision proche de celle obtenue avec les 39 paramètres MFCC selon la figure (IV.5). Les résultats montrent que le nombre de paramètres est sous-estimé en utilisant les méthodes de Freedman, Shimazaki et Scott alors que ce nombre est bien estimé en utilisant les méthodes : Sturges, proposée et GMM. En plus la méthode Shimazaki présente le temps de calcul maximal (7.38e+4 s), alors que la méthode proposée présente le temps de calcul minimal (1.37e+3 s). Bien que la méthode GMM atteigne presque la même précision de la méthode proposée, elle reste moins pratique à cause du temps de calcul qu'elle exige pour l'estimation de l'IM (6.44e+4 s). Ainsi, dans le cas des données non bruitées, la méthode proposée permet d'avoir un nombre minimal de paramètres pertinents avec un bon compromis entre la précision de la reconnaissance et le temps de calcul exigé pour cette estimation.

Dans le cas du nombre limité de données non bruitées, notre méthode atteint de bonnes performances avec une meilleure précision de reconnaissance avec un nombre réduit de paramètres acoustiques et un temps de calcul minimal. En plus elle commet une erreur faible

dans l'estimation de l'MI par rapport aux autres méthodes (Vmax de IM proche de $H(C)$). La méthode GMM atteint aussi une bonne précision et une faible erreur sur IM mais exige plus de paramètres acoustiques et un temps de calcul très grand. La méthode de Shimazaki s'avère la moins performante avec une grande erreur de l'IM, le plus grand temps de calcul, une mauvaise précision de reconnaissance. Ces résultats montrent la robustesse de la méthode proposée et sa meilleure estimation du nombre de paramètres pertinents dans le cas des petites bases de données.

Dans le cas des données bruitées, la figure (IV.5) montre qu'au moins 20 paramètres sont nécessaires pour obtenir une bonne précision de reconnaissance. Utilisant les résultats ci-dessus, seulement la méthode proposée et celle de GMM peuvent estimer approximativement le nombre correct des paramètres pertinents alors que les autres méthodes donnent des valeurs trop sous-estimées. Néanmoins la méthode GMM reste très lente (6.34e+4s) par rapport à la méthode proposée (1.37+3s). Là encore la méthode de Shimazaki est la plus lente dans la sélection et la moins précise pour la reconnaissance.

Ainsi, notre méthode peut être jugée la plus performante du fait qu'elle a pu déterminer un nombre de paramètres réduit, réalisant un bon compromis entre le temps de calcul et la précision d'estimation de la IM et celle de reconnaissance.

On peut constater également quelques remarques générales sur l'effet du nombre réduit des données ainsi que le bruit sur la sélection des paramètres et la précision du système de reconnaissance:

- En comparant la valeur maximale de IM obtenue dans le cas d'apprentissage sans-bruit et celui sans-bruit avec ensemble réduit, on remarque que la réduction du nombre de données conduit à une augmentation des valeurs maximales de la IM pour toutes les méthodes (sauf la méthode GMM). Néanmoins la méthode proposée commet la plus faible erreur, proche de $H(C)$. En outre cette réduction des données a diminué le nombre de paramètres sélectionnés ainsi que la précision de la reconnaissance.
- En comparant la valeur maximale de IM obtenue dans le cas d'apprentissage sans-bruit et celle de données bruitées, on remarque que cette valeur est diminuée dans le deuxième cas. Ceci peut être justifié probablement par la diminution de la dépendance entre les paramètres acoustiques et la variable index de classe C, causée par l'ajout du bruit au signal original. Cette diminution peut être expliquée par une diminution de la précision de reconnaissance pour un même nombre de paramètres (39 paramètres). Ainsi ces classes (états des modèles HMM) exigent plus de paramètres acoustiques pour mieux être expliquées. Effectivement, dans le cas d'apprentissage avec des données bruitées, un nombre proche de 20 paramètres est

indispensable pour avoir approximativement la même précision des 39 paramètres, comparé seulement à 10 paramètres dans le cas d'un apprentissage sans-bruit.

- Les paramètres dynamiques constituent une bonne partie du sous-ensemble des paramètres pertinents selon la stratégie TMI pour les différents modes d'apprentissage et pour la majorité des méthodes d'estimation de l'IM. En plus les paramètres énergétiques : log E (paramètre N°13) et delta-énergie (paramètre N°26) sont les plus pertinents pour toutes les méthodes, alors que le delta-delta-énergie est jugé pertinent dans certaines méthodes et non pour d'autres. En particulier, le paramètre statique N°13 (énergie) est plus pertinent que celui N°26 (delta-énergie) dans le cas d'apprentissage sans-bruit, et inversement dans le cas d'apprentissage sur des données bruitées.

La dernière expérience que nous présentons consiste à répondre sur le dernier problème exposé dans (VI.1), c.-à-d : l'effet de l'augmentation de la dimension des vecteurs acoustiques sur la précision du système RAP basé sur une base de données limité. On présente ainsi la solution en sélectionnant les paramètres les plus pertinents par l'application de l'algorithme TMI.

Dans cette expérience, le système de reconnaissance de chiffres est basé sur la représentation de chaque trame d'analyse par 111 paramètres acoustiques ainsi que la base de données sans-bruit réduit. L'ensemble des paramètres considérés sont :

- 39 paramètres MFCC présenté précédemment.
- 36 paramètres LPCC dont 12 paramètres statiques et 24 paramètres dynamiques.
- 36 paramètres PLP dont 12 paramètres statiques et 24 paramètres dynamiques.

Le résultat de reconnaissance donne une précision de 97.81% pour ces 111 paramètres, alors que la précision de système basé sur les 39 paramètres a donné 98.71% (Tableau IV.4.b).

Ce résultat montre qu'augmenter la dimension des vecteurs acoustiques n'est pas toujours une solution pour améliorer les performances des systèmes RAP. Ceci est dû principalement par le phénomène de la malédiction de la dimensionnalité.

Ainsi, pour résoudre ce problème, il faut réduire la dimension des vecteurs en cherchant les paramètres les plus pertinents. Cette réduction peut être effectuée en sélectionnant ces paramètres en appliquant l'algorithme TMI.

La figure (IV.6) donne les résultats de cette sélection pour différentes méthodes d'estimation d'IM : Scott, Sturges, Freedman , Shimazaki , GMM et proposée. Les méthodes de Freedman, Shimazaki commettent la plus grande erreur. Alors que la méthode proposée et GMM sont les plus proches de $H(C)$ justifiant l'erreur minimale commise.

En utilisant le critère d'arrêt proposé, on obtient pour les méthodes Freedman, Shimazaki, Scott, Sturges, GMM et la méthode proposée, respectivement un nombre de paramètres pertinents de 6, 5, 6, 7, 8 et 9.

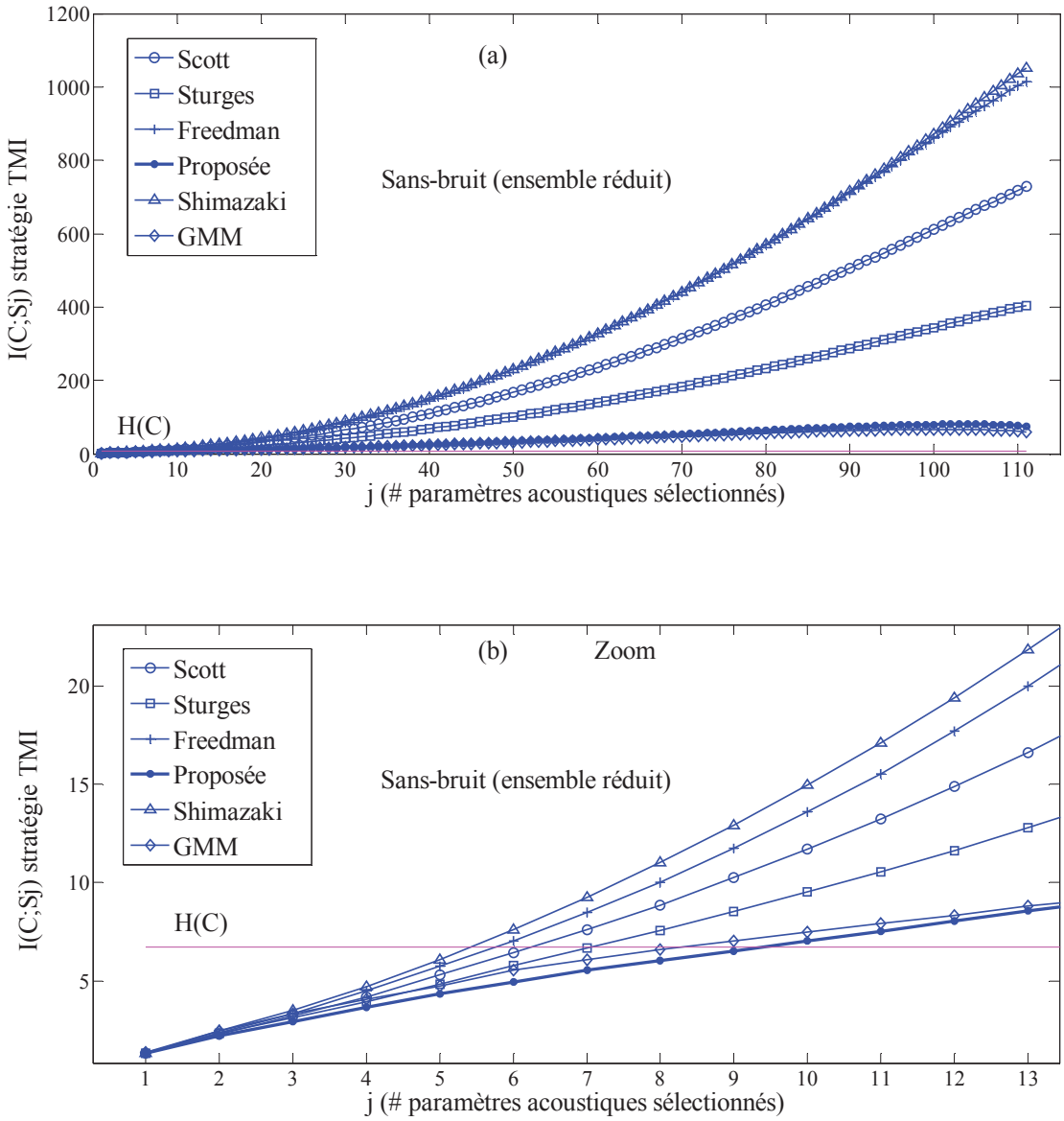


Figure. IV.6: Estimation de l'IM conjointe basée sur la méthode d'histogramme utilisant la stratégie TMI appliquée sur l'ensemble des paramètres combinés: Scott (○), Sturges (□), Freedman (+), Shimazaki (Δ), GMM (◊),et la méthode proposée (●). La ligne horizontale indique la valeur de l'entropie H(C)

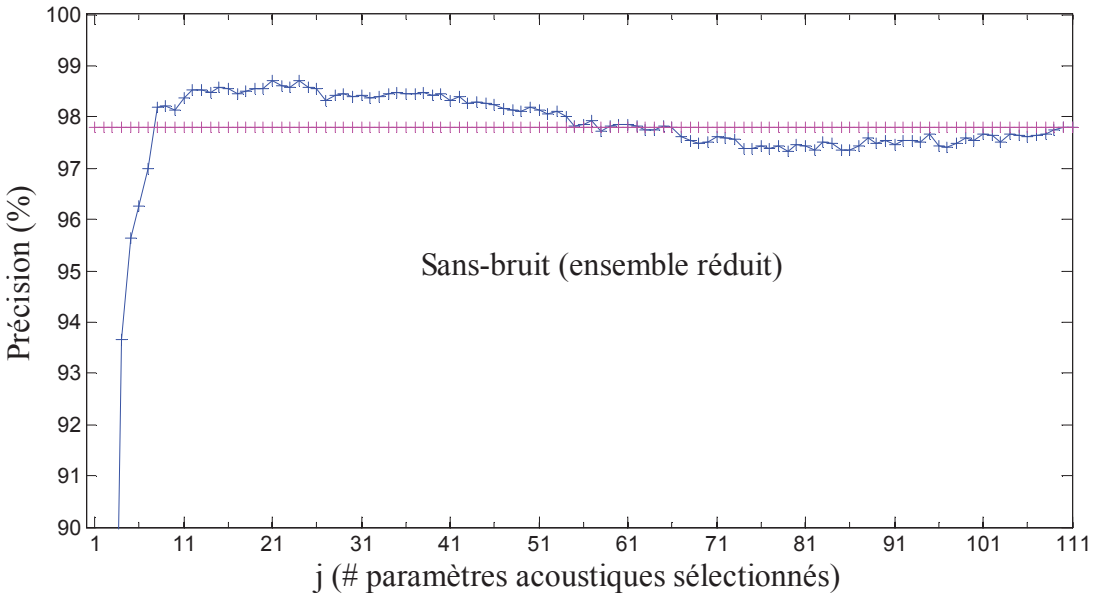


Figure.IV.7: Précision (%) du système RAP en fonction du nombre des paramètres sélectionnés. La stratégie TMI est utilisée. Les modèles entrainés et testés utilisant l'ensemble réduit clean (+). La ligne horizontale correspond à la précision obtenue avec un large vecteur acoustique (111 paramètres combinés des coefficients MFCC, PLP, LPCC et leur dérivés delta et delta-delta).

La figure (IV.7) montre la précision du système RAP. La courbe de précision augmente puis décroit. Le résultat montre que les 9 paramètres acoustiques sélectionnés donnent une précision de reconnaissance de 98.22% qui est plus grande que celle des 111 paramètres initiaux. Ces 9 paramètres sont: E, ΔE, $MFCC_2$, $MFCC_1$, $\Delta MFCC_2$, $\Delta\Delta$E, $\Delta LPCC_1$, $\Delta MFCC_3$, $\Delta\Delta MFCC_3$. Ce résultat montre que les paramètres MFCC sont plus pertinents par rapport aux paramètres PLP et LPCC.

IV.4 CONCLUSION

Dans ce chapitre nous avons appliqué le nouvel estimateur de l'information mutuelle dans la sélection des paramètres acoustiques les plus pertinents parmi l'ensemble des paramètres statiques et dynamiques des coefficients MFCC. On a proposé également d'appliquer l'algorithme de sélection TMI qui n'a jamais été appliqué dans le domaine de la RAP.

Les expériences sont effectuées sur un système de référence basé sur les modèles HMM. Ce système permet la reconnaissance des séquences de mots connectés en mode indépendant du locuteur fonctionnant sous différentes conditions de bruit. Toutes les données de parole utilisées dans l'apprentissage des modèles HMM et l'évaluation des performances du système appartiennent à la base Aurora2.

Les résultats nous ont montré que la méthode proposée et GMM sont les seules qui ont pu aboutir un nombre réduit de paramètres permettant d'obtenir une précision de reconnaissance

proche de celle de l'ensemble initial des paramètres. Par ailleurs, la méthode de Shimazaki ainsi que la méthode GMM exigent un temps de calcul très grand par rapport aux autres méthodes. De fait, la méthode proposée réalise un bon compromis entre la précision et le temps de calcul. Les résultats montrent aussi que les paramètres statiques sont plus pertinents dans le cas des données non bruitées alors que les paramètres dynamiques sont plus pertinents dans le cas des données bruitées. Ce résultat avait déjà été observé sans avoir été démontré dans la thèse de C. Barras [6].

Les résultats nous ont montré aussi qu'une augmentation de la dimension des vecteurs acoustiques n'implique pas une amélioration des performances du système de reconnaissance. Une expérience de sélection des paramètres acoustiques est effectuée pour réduire le nombre de paramètres en utilisant l'algorithme TMI. Cet algorithme associé à la méthode proposée PR sélectionne seulement 9 paramètres. Ces paramètres donnent une précision plus grande que celle de 111 paramètres.

Ainsi dans le cas de la reconnaissance de la parole, la méthode proposée a montré son avantage dans la sélection des paramètres acoustiques pertinents en termes de rapidité et précision.

CONCLUSION ET PERSPECTIVES

L'objectif initial de cette thèse était d'appliquer l'idée de la sélection des paramètres pertinents dans la reconnaissance de la parole. Plus particulièrement, nous avons étudié la sélection des paramètres acoustiques MFCC (*Mel Frequency Cepstrum Coefficients*) statiques et dynamiques pour une tâche de reconnaissance des chiffres de la base Aurora2. Ensuite nous avons présenté la sélection des paramètres à partir d'un ensemble de grande dimension constitué des paramètres MFCC, LPCC, LPP et leurs paramètres dynamiques.

Dans la mise en place d'une stratégie de sélection des paramètres, il est nécessaire d'accéder à une mesure de pertinence des paramètres qui permet de chiffrer leur importance dans la tâche de classification. Dans le cas d'une modélisation non-linéaire, l'information mutuelle (IM) I(C;Y**)** entre une variable (paramètre) Y et la variable index de classes C est souvent utilisée comme telle mesure. L'information mutuelle peut facilement être étendue à des groupes de variables, ce qui est essentiel dans des procédures de sélection de type "greedy" (procédures itératives "*forward*", "*forward-backward*", etc.).

Cependant, l'estimation de l'information mutuelle sur des données de taille finie est difficile, surtout lorsque le nombre de variables augmente, puisque sa définition exige l'estimation des Fonctions de Densités de Probabilités (fdp) qui ne sont pas connues en pratique et qui sont difficiles à estimer pour des variables de grandes dimensions.

Afin de remédier à ce problème, nous avons montré théoriquement en se basant sur la théorie de l'information que le calcul de l'information mutuelle entre un ensemble de k variables (caractéristiques) et la variable index des classes C, peut être tronqué, sous hypothèse à des ordres acceptables. Ainsi, cette troncature nous permet d'approximer l'IM, au travers des seules densités de probabilités d'ordre inférieur évaluées à partir de données de taille finie. Sur ce principe, nous avons proposé une nouvelle méthode de sélection des paramètres, utilisant la procédure "*forward*". De plus, un critère d'arrêt permet de fixer le nombre de caractéristiques sélectionnées en fonction de l'information mutuelle approximée à l'itération j. Cette approche appelée TMI (*Truncated Mutual Information*) a été appliquée dans la sélection des paramètres MFCC et validée sur des données de la base Aurora2 utilisée dans la reconnaissance de la parole. Les résultats présentés dans la conférence TAIMA 2009, ont montré que notre approche est plus performante que l'approche CMI (conditional Mutual Information).

Dans cette application, nous avons retenu l'estimation de la IM basée sur la méthode d'histogramme pour ses avantages indéniables en termes de simplicité et de complexité de

calcul. Néanmoins cette estimation souffre des problèmes de biais et de variance causés par le nombre fini de données, la quantification, ainsi que la représentation insuffisante de l'histogramme. Ce problème influe fortement sur la sélection des paramètres pertinents par la possibilité de commettre de grande erreur d'estimation de l'information mutuelle entre un sous ensemble de paramètres à sélectionner et la variable index de classe. Par ailleurs, ceci discrédite tout critère d'arrêt permettant d'estimer le nombre optimal des paramètres pertinents. Afin de réduire ces erreurs d'estimation d'IM et d'entropie, nous avons essayé plusieurs choix du nombre de *bins* proposés dans la littérature comme ceux de Sturges, Freedman, Scott, Shimazaki. Malheureusement aucun de ces choix n'est satisfaisant. Ainsi, nous avons proposé de nouvelles formules mathématiques du choix de nombre de *bins* des histogrammes permettant de minimiser l'Erreur Quadratique Moyenne (EQM) et le biais de la IM en se basant sur l'hypothèse de la fdp gaussienne. Les résultats de simulation ont montré que cette nouvelle méthode d'estimation du nombre de *bins* améliorait l'erreur quadratique moyenne de l'IM et de l'entropie, ce qui a permis, par voie de conséquence, de mieux estimer le nombre optimal de variables pertinentes. De plus, cette approche affiche un temps de calcul réduit par rapport aux méthodes citées précédemment.

Cet estimateur de l'IM a également été appliqué dans la sélection des paramètres MFCC et des paramètres de différents types (MFCC, PLP, LPCC) extraits des données de la base Aurora2. Les résultats de cette application nous ont confirmé la capacité de cet estimateur à obtenir un nombre de paramètres pertinents à l'inverse des autres méthodes qui ont sous-estimé ce nombre. Bien que le nouveau choix du nombre de *bins* soit basé sur l'hypothèse d'une fdp gaussienne, il reste le choix le plus convenable dans le cas des variables de type uniforme et lognormal mais demeure plus biaisé vis-à-vis des variables de type gaussien. Ainsi, un travail futur consiste à développer de nouveaux choix de *bins* pour d'autre type de fdp.

Finalement, nous envisagerons l'application de la TMI dans la sélection des paramètres les plus pertinents pour une tâche de classification parole/silence. Cette application aura pour rôle la segmentation de la base de données non étiquetée ou la suppression du silence pour vérification ou identification de locuteurs.

Annexe A

A.1. Analyse par prédiction linéaire (LPC)

Le signal vocal résulte de l'excitation du conduit vocal par un train d'impulsions ou un bruit produisant respectivement des sons voisés et non voisés. Ainsi, ce signal peut être modélisé par la convolution de la fonction de transfert du conduit vocal (filtre) avec le signal d'excitation (source).

Dans l'analyse par prédiction linéaire (LPC : Linear Prediction Coding), la fonction de transfert du conduit vocal peut être modélisée par un filtre linéaire tout-pôles qui produit un signal auto-régressif (AR). La fonction de transfert est donnée par :

$$H(z) = \frac{1}{\sum_{i=0}^{p} a_i z^{-i}} \quad \text{(A.1)}$$

Où p est le nombre de pôles (l'ordre de prédiction) , a_0=1, $\{a_i\}_{i=1:p}$ sont les coefficients du filtre. Chaque échantillon $s(n)$ est prédit comme une combinaison linéaire des p échantillons précédents, à laquelle s'ajoute un bruit blanc gaussien **e** de variance σ^2:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + e(n) \quad \text{(A.2)}$$

Les coefficients $\{a_i\}$ sont choisis de telle façon à minimiser l'erreur de prédiction estimée sur la fenêtre d'analyse par la méthode des moindres carrés. Cette minimisation conduit aux équations de Yule-Walker qui expriment le vecteur des coefficients $A= (1,a_1 \ldots a_p)^t$ comme:

$$R.A= (\sigma^2,0 \ldots 0)^t \quad \text{(A.3)}$$

Où R est une matrice de Toëplitz constituée des p+1 premiers coefficients d'autocorrélation.

Levinson en 1947, a développé un algorithme rapide pour résoudre l'équation (A.2) et calculer les coefficients auto-régressifs $\{ai\}i=1 \ldots p$. Cet algorithme a été ensuite modifié par Durbin en 1960 [118]. D'autres paramètres peuvent être extraits à partir d'une analyse LPC comme les coefficients de réflexion (ou PARCOR) et les coefficients cepstraux LPCC.

Un des inconvénients de la méthode de prédiction linéaire est que le filtre identifie uniformément le spectre sur toutes les fréquences de la bande d'analyse, alors que l'oreille humaine est plus sensible aux fréquences situées au milieu de la bande d'analyse du spectre (100 à 3000 Hz) [119]. Ainsi il est possible que l'analyse LP ne prenne pas en compte ces détails du spectre [28]. L'introduction des connaissances issues de la psycho-acoustique dans l'estimation d'un modèle AR [120] a permet de résoudre ce problème et a conduit à l'analyse par PLP. Cette analyse développée dans [19], modélise un spectre auditif par un modèle tout

pôle. Elle diffère de l'analyse standard LPC par une intégration en bandes critiques du spectre de puissance sur une échelle Bark [121], suivie d'une préaccentuation du signal non linéaire selon une courbe isotonique [122], d'une compression en racine cubique du spectre résultant pour simuler la loi de la perception humaine en puissance sonore [123], et finalement d'une modélisation tout pole [124].

Le diagramme fonctionnel de l'analyse PLP est illustré dans la figure (A,1).

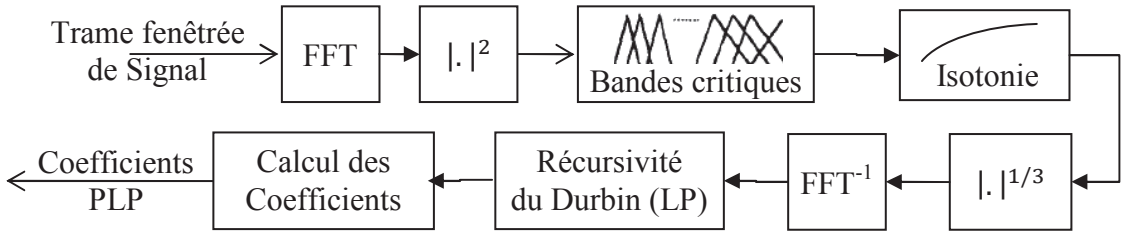


Figure. A.1 : Diagramme fonctionnel de la technique d'analyse PLP

Les coefficients cepstraux peuvent être obtenus à partir des coefficients de la prédiction linéaire ou des énergies d'un banc de filtres. Ainsi, les paramètres LPCC (*Linear Prediction Cepstral Coefficients*) sont calculés à partir d'une analyse par prédiction linéaire décrite en dessus. Si a_0=1, $\{a_i\}_{i=1:p}$ sont les coefficients de cette analyse, estimés sur une trame du signal, les d premiers coefficients cepstraux C_k sont calculés récursivement par:

$$C_k = -a_k - \sum_{i=1}^{k-1} \frac{(k-i)}{k} C_{k-i} a_i \qquad 1 \leq k \leq d \tag{A.4}$$

Un liftrage est effectué pour augmenter la robustesse des coefficients cepstraux [29]. Ce liftrage consiste à multiplier des coefficients cepstraux par une fenêtre de poids W(k) pour être moins sensible au canal de transmission et au locuteur :

$$\forall\, k \in [1, L] \quad W(k) = 1 + \frac{L}{2}.\sin\left(\frac{\pi.k}{L}\right) \tag{A.5}$$

où L est le nombre de coefficients.

Annexe B

MISE EN ŒUVRE D'UN SYSTEME DE RECONNAISSANCE DES MOTS CONNECTES SOUS HTK [15]

B.1 INTRODUCTION

HTK est une boîte à outils de modèles de Markov cachés HMM, conçue pour la construction et la manipulation de ces modèles. Cette boîte est constituée d'un ensemble de modules bibliothèque et d'outils disponibles en codes sources C. Ces outils HTK sont conçus pour fonctionner en ligne de commande, généralement sous l'environnement linux avec le Shell C. Chaque outil a un nombre d'arguments obligatoires en plus d'arguments optionnels préfixés par le signe "–". Le chapitre "Reference section" de l'ouvrage htkbook [15][7] décrit en détail tous les outils de la boîte HTK ainsi que leurs arguments.

Principalement, la boîte à outils HTK est utilisée pour la construction des systèmes RAP basés sur les modèles HMM dans un but de recherche scientifique. Généralement les deux processus indispensables pour le fonctionnement d'un RAP sont le processus d'apprentissage et celui de reconnaissance (ou décodage). La figure (B.1) illustre l'enchaînement de ces processus. Premièrement, les outils d'apprentissage HTK sont utilisés pour estimer les paramètres de l'ensemble des modèles HMM en utilisant des signaux de parole ainsi que leurs transcriptions associées. Ensuite, les signaux de parole inconnue sont transcrits en utilisant les outils de reconnaissance. Le lecteur peut consulter le livre htkbook pour plus de détails sur l'implémentation des systèmes RAP sous la plateforme HTK [15].

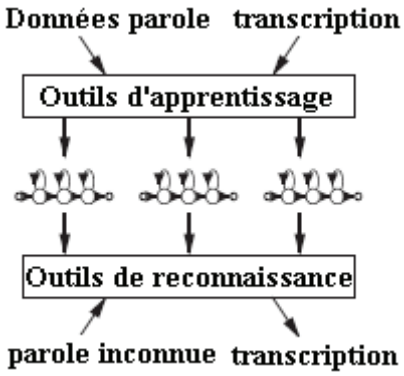


Figure.B.1 : Processus d'un système RAP

Pratiquement, la construction d'un système RAP se base sur 4 phases principales: préparation des données, apprentissage, test, analyse. La figure (B.2) illustre les différents outils HTK de chaque phase d'un système de reconnaissance de la parole continue.

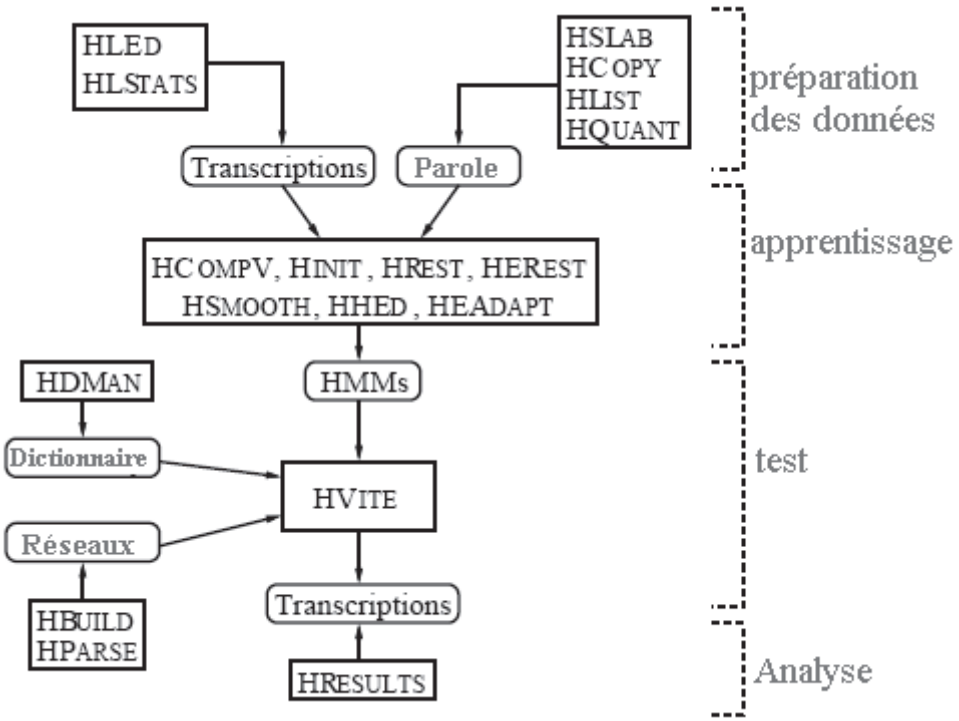


Figure.B.2 : Différentes phases du système RAP sous HTK et outils associés

Dans les paragraphes suivants, on décrit brièvement ces outils pour le cas d'un système de reconnaissance de mots connectés (les chiffres anglais). Ce type de système se base sur la modélisation entière de chaque mot par un HMM.

B.2 OUTILS DE PREPARATION DES DONNEES

La construction d'un ensemble de modèles HMM exige un ensemble de fichiers de données de parole (signaux), ainsi que leurs transcriptions correspondantes. Souvent les données de parole sont récupérées à partir d'une base de données, fournie généralement sur un CD-ROM. Cette base doit être répartie en un corpus d'apprentissage et un corpus de test. Chacun de ces corpus contient un ensemble de fichiers texte contenant la transcription orthographique des phrases et un ensemble de fichiers de données contenant les échantillons des signaux correspondant aux fichiers texte. Avant d'être utilisées dans l'apprentissage, ces données doivent être converties en un format paramétrique approprié et ses transcriptions associées doivent être converties en format correct (étiquetées en label de mot dans notre travail). Si les données de parole ne sont pas disponibles, alors l'outil HSLab peut être utilisé pour enregistrer la parole et l'étiqueter manuellement par n'importe quelle transcription (par phonème ou mot). Ainsi pour chaque phrase prononcée, on lui correspond un fichier signal (exemple d'extensions : wav, sig,…) et un fichier de transcription (extension lab).

Cependant, avant d'effectuer ces transcriptions, un dictionnaire des mots doit être défini afin d'être utilisé dans la phase d'apprentissage et celle de test. Dans le cas d'un système basé

sur des modèles HMM représentant des phonèmes, la construction du dictionnaire s'effectue par l'outil HDMan. De plus la grammaire de la tâche considérée doit être définie en utilisant l'outil HParse. Cet outil génère un réseau de mots définissant la grammaire considérée (figure B.2).

La dernière étape dans la phase de préparation des données est la conversion du signal de chaque phrase en une séquence de vecteurs acoustiques (voir figure B.3). Cette conversion est effectuée par une analyse acoustique en utilisant l'outil HCopie. Différents types de paramètres acoustiques sont supportés par cet outil comme : LPC, LPCC, MFCC, PLP, FBANK (log mel-filter bank), MELSPEC (linear mel-filter bank), LPCEPSTRA (LPC cepstral coefficients), LPREFC (linear prediction reflection coefficients), USER (type défini par l'utilisateur).

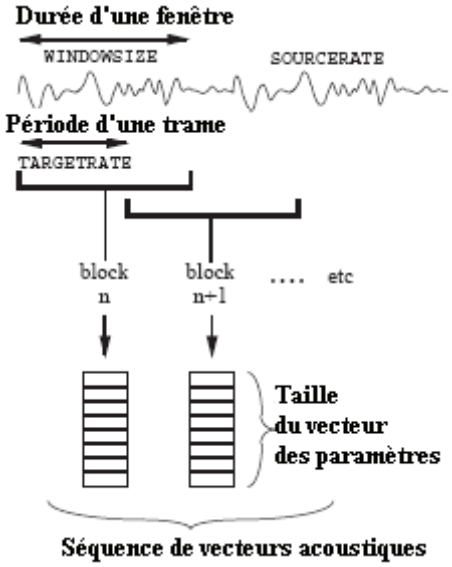


Figure.B.3: Processus de l'analyse acoustique

La ligne de commande pour l'exécution de HCopy s'écrit comme suit :

HCopy -T 1 -C config -S codetr.scp

La figure (B.4) montre le principe de fonctionnement de cet outil pour la conversion d'un ensemble de fichiers parole d'extension wav en un ensemble de fichiers d'extension mfc contenant des vecteurs de paramètres acoustiques MFCC. La liste de l'ensemble de ces fichiers est donnée dans un fichier appelé codetr.dcp dont un extrait est fourni :

root/training/corpus/sig/S0001.wav root/training/corpus/mfcc/S0001.mfc
root/training/corpus/sig/S0002.wav root/training/corpus/mfcc/S0002.mfc
root/training/corpus/sig/S0003.wav root/training/corpus/mfcc/S0003.mfc
etc.

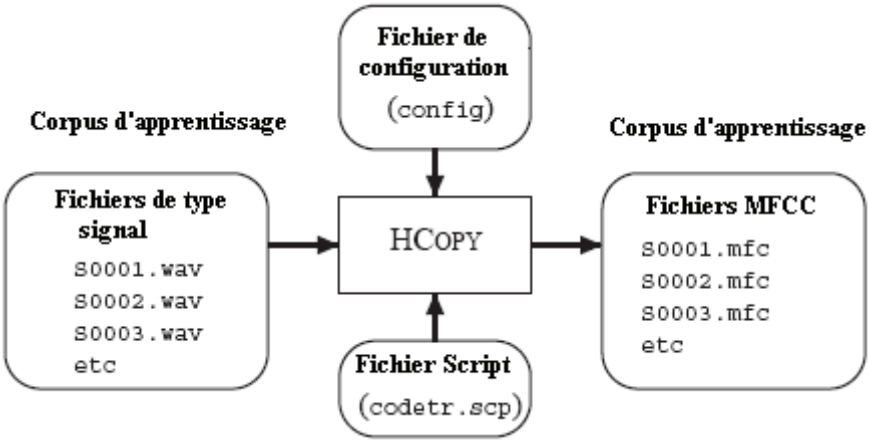


Figure.B.4: Principe de fonctionnement de l'outil HCopy

Cependant l'exécution de l'outil HCopy exige un fichier de configuration (config) pour définir les différents paramètres de l'analyse acoustique considérée. Voici un exemple de ce type de fichier associé à une analyse acoustique MFCC :

```
# Exemple d'un fichier de configuration pour une analyse acoustique MFCC
SOURCEFORMAT = HTK # donne le format des fichiers des signaux
TARGETKIND = MFCC_0_D_A # identificateur des coefficients à utiliser
# Unit = 0.1 micro-second :
WINDOWSIZE = 250000.0 # = 25 ms = longueur de la durée d'une trame
TARGETRATE = 100000.0 # = 10 ms = période des trames
NUMCEPS = 12 # nombre des coefficients MFCC(ici de c1 to c12)
USEHAMMING = T # utilisation de la fonction de Hamming pour le fenêtrage
des trames
PREEMCOEF = 0.97 # coefficient de prè-accentuation
NUMCHANS = 26 # nombre de canaux des bancs de filtres
CEPLIFTER = 22 # longueur de liftrage cepstral
# la fin
```

B.3 OUTILS D'APPRENTISSAGE

La deuxième phase consiste à construire les modèles HMM des mots appartenant au dictionnaire de la tâche considérée. Premièrement, pour chaque mot, il faut définir un modèle prototype contenant la topologie choisie à savoir le nombre d'états du modèle, la disposition de transitions entre les états, le type de la loi de probabilité associée à chaque état. L'état initial et final de chaque modèle n'émettent pas des observations mais servent seulement à la connexion des modèles dans la parole continue]. Les probabilités d'émissions associées aux états sont des mélanges de gaussiennes multivariées dont les composantes sont les probabilités a priori définies chacune par une matrice de covariance et un vecteur de moyennes dans l'espace des paramètres acoustiques. La matrice de covariance peut être choisie diagonale si l'on suppose l'indépendance entre les composantes des vecteurs acoustiques [6].

Ces modèles prototypes sont générés dans le but de définir la topologie globale des modèles HMM. Ainsi, l'estimation de l'ensemble des paramètres de chaque modèle HMM est

le rôle du processus d'apprentissage. Les différents outils d'apprentissage sont illustrés dans la figure (B.5).

Selon cette figure, deux chaines de traitement peuvent être envisagés pour l'initialisation des modèles HMM. La première chaîne tient en compte des signaux étiquetés en label de mot. Dans ce cas, l'outil HInit extrait tous les segments correspondant au mot modélisé et initialise les probabilités d'émission des états du modèle au moyen de la procédure itérative des "k moyennes segmentales". Ensuite l'estimation des paramètres d'un modèle est affinée avec HRest, qui applique l'algorithme optimal de Baum-Welch jusqu'à la convergence et réestime les probabilités d'émission et de transition.

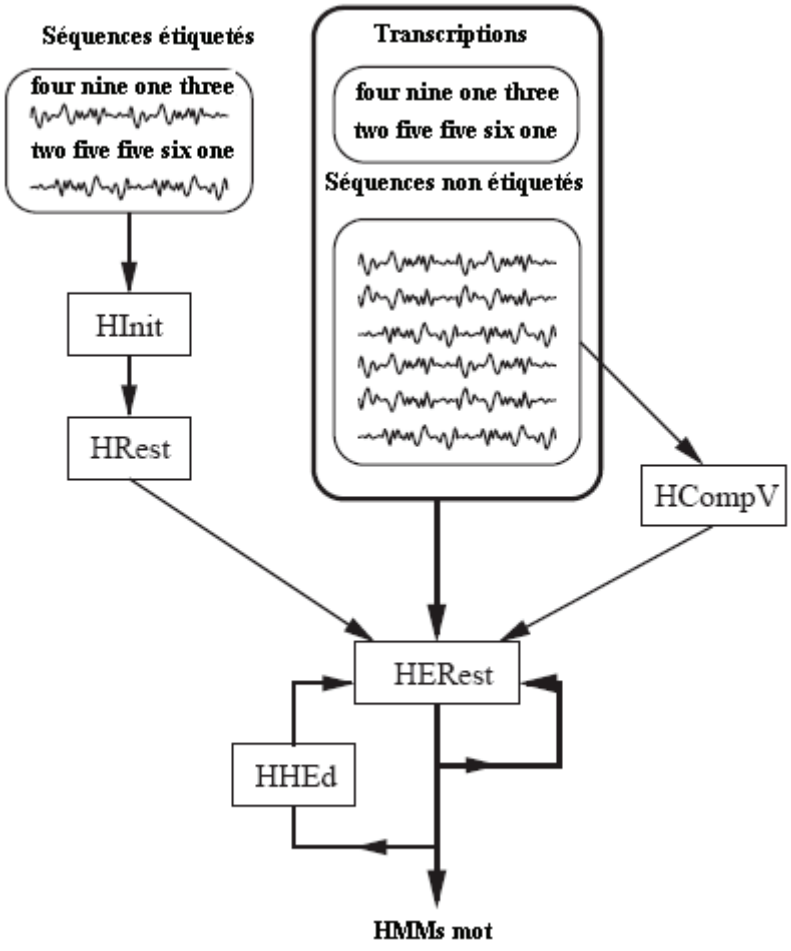


Figure.B.5: Outils d'apprentissage HTK

Dans la deuxième chaîne, les signaux ne sont pas étiquetés. Dans ce cas, tous les modèles HMM sont initialisés avec le même modèle dont les moyennes et les variances sont égales respectivement à la moyenne et la variance globales de tous les vecteurs acoustiques du corpus d'apprentissage. Cette opération est effectuée par l'outil HCompV.

Après l'initialisation des modèles, l'outil HERest est appliqué en plusieurs itérations pour réestimer simultanément l'ensemble des modèles sur l'ensemble de toutes les séquences de vecteurs acoustiques non étiquetés. Les modèles obtenus peuvent être améliorés, en

augmentant par exemple le nombre de gaussiennes servant à estimer la probabilité d'émission d'une observation dans un état. Cette augmentation est effectuée par l'outil HHEd. Les modèles doivent être ensuite réestimés par HRest ou HERest.

B.4 OUTILS DE RECONNAISSANCE

La boîte HTK fournit un outil de reconnaissance appelé HVite qui permet la transcription d'une séquence de vecteurs acoustiques en une séquence de mots. Le processus de reconnaissance est illustré dans la figure (B.6).

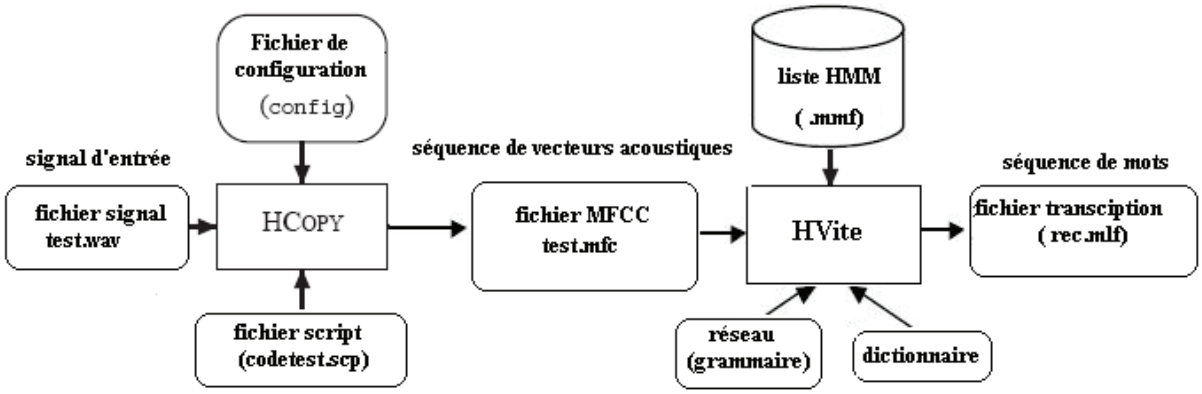


Figure.B.6: Processus de reconnaissance sous HTK

HVite utilise l'algorithme de Viterbi (§ I.5.2.3) pour trouver la séquence d'états la plus probable qui génère la séquence d'observations (vecteurs acoustiques) selon un modèle HMM composite, ceci afin d'en déduire les mots correspondants. Le modèle composite permet la succession des modèles acoustiques en fonction du réseau de mots qui définit la grammaire de la tâche considérée.

Le résultat de décodage par l'outil HVite est enregistré dans un fichier d'extension (.mlf) contenant l'étiquetage en mots du signal d'entrée.

B.5 OUTILS D'EVALUATION

Généralement les performances des systèmes RAP sont évaluées sur un corpus de test contenant un ensemble de fichiers d'échantillons parole ainsi que leurs fichiers d'étiquetage associés. Les résultats de reconnaissance des signaux du corpus de test sont comparés aux étiquettes de référence par un alignement dynamique réalisé par HResults, afin de compter les étiquettes identifiées, omises, substituées par une autre, et insérées. Ces statistiques permettent de calculer le taux ou la précision de reconnaissance définis dans le chapitre IV.

Liste des figures

Liste des tableaux

Liste des acronymes

ACP	Analyse en Composante Principal
ALD	Analyse Linéaire Discriminante
CIFE	Conditional Infomax Feature Extraction
CMI	Conditional Mutual Information
CPU	Central Processing Unit
DTW	Dynamic Time Warping
ELRA	European Language Resources Association
EQM	Erreur Quadratique Moyenne
Fdp	Fonctions de Densités de Probabilités
FFT	Fast Fourier Transform
FOU	first-order utility
GMM	Gaussian mixture model
HMM	Hidden Markov Model
IF	Informative Fragments
IM	Information Mutuelle
IMV	Information Mutuelle Multivariée
JMI	Joint Mutual Information
KDE	Kernel Density Estimation
LAR	Logarithm Area Ratios
LDC	Linguistic Data Consortium
LFCC	Linear Frequency Cepstral Coefficients
LPC	Linear Predictive Coefficients
LPCC	Linear Predictive Cepstral Coefficients
LSF	Line Spectral Frequencies
MAP	Maximum A Posteriori
MFCC	Mel Frequency Cepstral Coefficients
MIFS	Mutual Information-Based Feature Selection
MLE	Maximum Likelihood Estimation
MMI	Maximum Mutual Information
MRMR	Maximum-Relevance Minimum-Redundancy
MSG	Modulation SpectroGram
PLP	Perceptual Linear Predictive
RAM	Random Access Memory
RAP	Reconnaissance Automatique de la Parole
RSB	Rapport Signal sur Bruit
SVM	Support Vector Machine
TMI	Truncated Mutual Information

Bibliographie

[1] J.P. Haton, C. cerisara, D. Fohr, Y. Laprie, and K. Smaili, *Reconnaissance automatique de la parole: du signal à son interprétation*. Paris: Dunod, 2006.

[2] L. Jiang and X. Huang, "Acoustic feature selection using speech recognizers," in *Proc of IEEE ASRU Workshop*, Keystone, Colorado, 1999.

[3] R. Gemello, F. Mana, D. Albesano, and R.D. Mori, "Multiple resolution analysis for robust automatic speech recognition," *computer, Speech and Language*, vol. 20, no. 1, pp. 2–21, January 2006.

[4] S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 1, pp. 52–59, February 1986.

[5] B.A. Hanson and T.H. Applebaum, "Robust speaker independent word recognition using static, dynamic and acceleration features: experiments with Lombard and noisy speech," in *Proc. ICASSP*, vol. 2, Albuquerque, NM, USA, april 1990, pp. 857-860.

[6] C. Barras, "Reconnaissance de la parole continue : Adaptation au locuteur et contrôle temporel dans les modèles de Markov Cachés," Université de Paris VI, Paris, Thèse de Doctorat 1996.

[7] C. Levy, "Modèles acoustiques compacts pour les systèmes embarqués," Université d'Avignon, Avignon, Thèse de Doctorat 2006.

[8] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4-37, January 2000.

[9] A. Hacine-Gharbi, P. Ravier, and T. Mohamadi, "Une nouvelle méthode de sélection des paramètres pertinents : application en reconnaissance de la parole," in *Proc. conférence TAIMA*, Hammamet, Tunisie, Mai 2009, pp. 399-407.

[10] A. Hacine-gharbi, P. Ravier, R. Harba, and T. Mohamadi, "Low bias Histogram-based Estimation of Mutual Information for Feature Selection," *Pattern Recognition Letters*, vol. 33, no. 10, pp. 1302–1308, July 2012.

[11] Calliope, *La parole et son traitement automatique*. Paris: Masson et CNET-ENST, 1989.

[12] J. Mariani, "Reconnaissance automatique de la parole: progrès et tendances," *Traitement du signal*, vol. 7, no. 4, pp. 239-266, 1990.

[13] L.R. Rabiner and B.H. Juan, *Fundamentals of speech recognition*. Englewood Cliffs, N.J., USA: Prentice Hall, 1993.

[14] H.G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under Noisy conditions," in *Proc. ISCA ITRW ASR2000*, Paris, France, September 2000, pp. 181-188.

[15] S. Young et al., *The HTK book (HTK version 3.4)*. Cambridge : Cambridge University Press, 2006.

[16] (2008) ELRA - European Language Resources Association. [Online]. www.elra.info

[17] F.J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," in *Proc. IEEE*, vol. 66 (1), Jan 1978, pp. 51-83.

[18] L.R. Rabiner and R.W. Schaffer, *Digital processing of speech signals*. Englewood Cliffs, N.J., USA: Prentice-Hal, 1978.

[19] H. Hermansky, "Perceptual Linear Predictive (PLP) analysis of speech," *Acoustic Society Am*, vol. 87, no. 4, pp. 1738-1752, April 1990.

[20] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for

monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, no. 4, pp. 357-366, August 1980.

[21] V. Neelakantan and J.N. Gowdy, "A comparative study of using different speech parameters in the design of discrete hidden Markov model," in *Proc. IEEE*, vol. 2, Southeaton. USA, April 1992, pp. 475-478.

[22] J.C. Junqua, H. Wakita, and H. Hermansky, "Evaluation and optimization of perceptually-based ASR front-end," *IEEE Trans. on Speech and Audio Processing*, vol. 1, no. 1, pp. 39-48, January 1993.

[23] C.R. Jankovski, H.D. Vo, and R.P. Lipmann, "A comparison of signal processing front ends for automatic word recognition," *IEEE trans. on Speech and Audio Processing*, vol. 3, no. 4, pp. 286-293, July 1995.

[24] J. Holmes and W. Holmes, *Speech Synthesis and Recognition*. Bristol, PA, USA: Francis & Taylor, 2002.

[25] D. Ellis. (1999, Jully) Feature Statistics Comparison Page. [Online]. http://www.icsi.berkeley.edu/~dpwe/respite/multistream/msgmfcc.html

[26] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, April 1981.

[27] L.R. Rabiner and R.W. Schafer, "Introduction to digital speech processing," vol. 1, no. 1, pp. 1-194, January 2007.

[28] O. Deroo, "Modèles dépendants du contexte et méthodes de fusion de données appliqués à la reconnaissance de la parole par modèles hybrides HMM_MLP," Faculté Polytechniques de Mons, Mons, Belgique, Thèse de Doctorat 1998.

[29] B.H. Juang, L.R. Rabiner, and J.G. Wilpon, "On the use of band pass liftering in speech recognition," in *Proc. ICASSP*, vol. 11, Tokyo, Japan, April 1986, pp. 765-768.

[30] F.K. Soong and A.E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing,* , vol. 36, no. 6, pp. 871-879, Jun 1988.

[31] C.H. Lee, E. Giachin, L.R. Rabiner, R. Pieraccini, and Rosenberg. A.E., "Improved acoustic modeling for large vocabulary continuous speech recognition," *Computer Speech and Language*, vol. 6, no. 2, pp. 103–127, April 1992.

[32] S. Nefti, "Segmentation automatique de parole en phones. Correction d'étiquetage par l'introduction de mesures de confiance," Université de Rennes 1, Rennes, Thèse de Doctorat 2004.

[33] J.K. Baker, "The DRAGON system - An overview," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, no. 1, pp. 24-29, February 1975.

[34] F. Jelinek, "Continuous speech recognition by statistical methods," in *Proc. of the IEEE*, vol. 64(4), April 1976, pp. 532-556.

[35] L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proc. of the IEEE*, vol. 77(2), February 1989, pp. 257-286.

[36] S. Igounet, "Eléments pour un système de reconnaissance automatique de la parole continue du Français," Université d'Avignon, Avignon, Thèse de Doctorat 1998.

[37] L. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic function of a Markov process," *Inequality*, vol. 3, pp. 1-8, 1972.

[38] C.M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, UK: Oxford University Press, 1995.

[39] A.K. Jain and B. Chandrasekaran, "Dimensionality and Sample Size Considerations in Pattern Recognition Practice," , vol. 2 of Handbook of Statistics, pages 835–855,

Amsterdam, North-Holland, 1982.

[40] S.J. Raudys and V. Pikelis, "On Dimensionality, Sample Size, Classification Error, and Complexity of Classification Algorithms in Pattern Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 2, no. 3, pp. 243-251, March 1980.

[41] M.C. Benitez et al., "Robust ASR front-end using spectral based and discriminant features: experiments on the aurora task," in *Proc. Interspeech*, Aalborg, Denmark, September 2001, pp. 429–432.

[42] K. Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noise and reverberant environments," in *Proc. ICSLP*, Sydney, Australia, 1998, pp. 891–894.

[43] H. Tolba, S.A. Selouani, and D. Shaughnessy, "Auditory-based acoustic distinctive features and spectral cues for automatic speech recognition using a multi-stream paradigm," in *Proc. ICASSP*, vol. 1, Orlando, FL, USA, May 2002, pp. 837–840.

[44] M.K. Omar, K. Chen, M. Hasegawa-Johnson, and Y. Bradman, "An evaluation of using mutual information for selection of acoustic features representation of phonemes for speech recognition," in *Proc. ICSLP*, Denver, Colorado, USA, September 2002, pp. 2129–2132.

[45] M.K. Omar and M.A Hasegawa-Johnson, "Maximum mutual information based acoustic features representation of phonological features for speech recognition," in *Proc. ICASSP*, vol. 1, Canada, May 2002, pp. 81–84.

[46] A. Zolnay, R. Schlüter, and H.J. Ney, "Acoustic feature combination for robust speech recognition," in *Proc. ICASSP*, Philadelphia, March 2005, pp. 457–460.

[47] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Second Edition ed. New York, USA: Academic Press, 1990.

[48] A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Survey*, vol. 2, pp. 94–128, 1999.

[49] B. Chigier and H.C. Leung, "The effects of signal representations, phonetic classification techniques, and the telephone network," in *Proc. ICSLP*, Banff, Alberta, Canada, October 1992, pp. 97-100.

[50] O. Siohan, Y. Gong, and J.P. Haton, "A comparison of three noisy speech recognition approaches," in *Proc. ICSLP*, vol. 3, Yokohama, Japan, September 1994, pp. 1031-1034.

[51] F. Grandidier, "un nouvel algorithme de sélection de caractéristiques Application à la lecture automatique de l'écriture manuscrite," Ecole de technologies supérieure, université du Québec, Québec, Thèse de Doctorat 2003.

[52] A. Jain and D. Zongker, "Feature selection: Evaluation, application and small sample performance," *IEEE Trans. on Pattern Analysis and Machine Recognition*, vol. 19, no. 2, pp. 153–158, February 1997.

[53] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 3, pp. 131–156, 1997.

[54] K. Kira and L.A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proc. Conference on Artificial Intelligence*, 1992, pp. 129–134.

[55] P.M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature selection," *IEEE Transactions on Computers*, vol. 26, no. 9, pp. 917–922, September 1977.

[56] D. Koller and M. Sahami, "Toward optimal feature selection," in *Proc.of the International Conference on Machine Learning*, Bari, Italy, July 1996, pp. 284–292.

[57] P. Leray, "Apprentissage et Diagnostic de Systèmes Complexe: Réseaux de Neurones et Réseaux Bayésiens," Université de Paris 6, Paris, Thèse de Doctorat 1998.

[58] D.W. Aha and R.L. Bankert, "A comparative evaluation of sequential feature selection algorithms," in *Proc. of the 5th International Workshop on Artificial Intelligence and Statistics*, Ft. Lauderdale, FL, USA, 1995, pp. 1–7.

[59] H. Daviet, "ClassAdd, une procédure de sélection de variables basée sur une troncature k additive de l'information mutuelle et sur une Classification Ascendante Hiérarchique en prétraitement," Laboratoire d'Informatique de Nantes Atlantique, Université de Nantes, Nantes, Thèse de Doctorat 2009.

[60] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. MA, USA: Academic Kluwer Publishers, Norwell, 1998.

[61] G.H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and subset selection problem," in *Proc. of the Conference on Machine Learning*, San Francisco, CA, 1994, pp. 121–129.

[62] P. Langley, "Selection of relevant features in machine learning," in *Proc. AAAI Fall Symposium on Relevance*, New Orleans, 1994, pp. 140–144.

[63] Leray, P.; Gallinari, P., "Feature selection with neural networks," *Behaviormetrika*, vol. 26, no. 1, pp. 145-166, 1999.

[64] T.M Cover and J.A Thomas, *Elements of information theory*, 2nd ed.: Wiley Series in telecommunications and Signal Processing, 2006.

[65] C.E. Shannon, "A mathematical theory of communication," *Bell Systems Technical Journal*, vol. 27, pp. 379-423, July 1948.

[66] W.J. McGill, "Multivariate Information Transmission," *IEEE Trans. Information Theory*, vol. 4, no. 4, pp. 93-111, September 1954.

[67] A.P. Hekstra and F.M.J. Willems, "Dependence Balance Bounds for Single-output Two-way Channels," *IEEE Trans. Information Theory*, vol. 35, no. 1, pp. 44-53, January 1989.

[68] R. Fano, *Transmission of Information: Statistical,Theory of Communications*. New York, USA: Wiley, 1961.

[69] T.S. Han, "Multiple Mutual Informations and Multiple Interactions in Frequency Data," *Information and Control*, vol. 46, no. 1, pp. 26-45, July 1980.

[70] T. Drugman, M. Gurban, and J.P. Thiran, "Relevant Feature Selection for Audio-visual Speech recognition," in *Proc. of the International Workshop on Multimedia Signal Processing*, Crete, Greece, October 2007, pp. 179 - 182.

[71] M.E. Hellman and J. Raviv, "Probability of error, equivocation, and the Chernoffv bound," *IEEE Transactions on Information Theory*, vol. 16, no. 4, pp. 368–372, July 1970.

[72] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," *IEEE Trans. Neural Networks*, vol. 5, no. 4, pp. 537- 550, July 1994.

[73] G. Brown, "A New Perspective for Information Theoretic Feature Selection," in *Proc. of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 5 of JMLR, Clearwater Beach, Florida, USA, 2009, pp. 49-56.

[74] M. Gurban and J.P Thiran, "Information Theoretic Feature Extraction for Audio-Visual Speech Recognition," *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4765-4776, December 2009.

[75] P. Scanlon and R. Reilly, "Feature analysis for automatic speech reading," in *Proc. Workshop on Multimedia Signal Processing,* cannes, October 2001, pp. 625–630.

[76] H.C Peng, F. Long, and C.H.Q Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–

1238, august 2005.

[77] H.H Yang and J. Moody, "Feature selection based on joint mutual information," in *Advances in Intelligent Data Analysis (AIDA) and Computational Intelligent Methods and Application (CIMA)*, Rochester, New York, 1999.

[78] N Kwak and C.H Choi, "Input Feature Selection for Classification Problems," *IEEE Transactions on Neural Networks*, vol. 13, no. 1, pp. 143–159, January 2002.

[79] M. Vidal-Naquet and S. Ullman, "Object recognition with informative features and linear classification," in *Proc. IEEE on Computer Vision and Pattern Recognition*, vol. 1, Nice, France, October 2003, pp. 281-288.

[80] F. Fleuret, "Fast Binary Feature Selection with Conditional Mutual Information," *Machine Learning Research*, vol. 5, pp. 1531–1555, November 2004.

[81] D. Lin and X. Tang, "Conditional Infomax Learning: An Integrated Framework for Feature Extraction and Fusion," in *Proc. European Conference on Computer Vision*, vol. Part I, Graz , Autriche, May 2006, pp. 68-82.

[82] I. Kojadinovic, "Relevance measures for subset variable selection in regression problems based on k-additive mutual information," *Computational Statistics & Data Analysis*, vol. 49, no. 4, pp. 1205-1227, June 2005.

[83] G.C Rota, "On the Foundations of Combinatorial Theory I. Theory of Mobius Functions," *Probability Theory and Related Fields*, vol. 2, no. 4, pp. 340–368, 1964.

[84] B.V. Bonnlander and A.W. Weigend, "Selecting input variables using mutual information and nonparametric density estimation," in *Proc. of International Symposium on Artificial Neural Networks (ISANN'94)*, Tainan, Taiwan, 1994, pp. 42–50.

[85] H. Wang, D. Bell, and F. Murtagh, "Relevance approach to feature subset selection," *Feature Extraction, Construction and Selection*, vol. 453, pp. 85–99, 1998.

[86] M. Hutter and M. Zaffalon, "Distribution of mutual information from complete and incomplete data," *Comput. Statist. Data Anal*, vol. 48, no. 3, pp. 633--657, March 2005.

[87] V. Gómez-Verdejo, M. Verleysen, and J. Fleury, "Information-theoretic feature selection for functional data classification," *Neurocomputing*, vol. 72, no. 16-18, pp. 3580–3589, October 2009.

[88] Y. Deng and J. Liu, "Feature Selection Based on Mutual Information for Language Recognition," in *Proc. of the 2nd International Congress on Image and Signal Processing*, Oct 2009, pp. 1-4.

[89] P. Scanlon, D.P.W Ellis, and R. Reilly, "Using mutual information to design, class-specific phone recognizers," in *Proc. Eurospeech*, Geneva, Switzerland, eptember S2003.

[90] H.H. Yang, S.V. vuuren, S. Sharma, and H. Hermansky, "Relevance of time-frequency feature for phonetic and speaker-channel classification," *Speech communication*, vol. 31, no. 1, pp. 35-50, May 2000.

[91] J.P.W. Pluim, J.B.A. Maintz, and M.A. Viergever, "Mutual Information Based Registration of Medical Images: A Survey," *IEEE Trans on Medical Imaging*, vol. 22, no. 8, pp. 986 - 1004, August 2003.

[92] G.D. Tourassi, E.D. Frederick, M.K. Markey, and C.E. Floyd, "Application of the Mutual Information Criterion for Feature Selection in Computer-aided Diagnosis," *Medical Physics*, vol. 28, no. 12, pp. 2394-2402, December 2001.

[93] S. Aoyagi, A. Azusa, A. Takesawa, A. yamashita, and M. kudo, "Mutual information theory for biomedical applications : estimation of three protein-adsorbed dialysis membranes," *Applied surface Science*, vol. 252, no. 19, pp. 6697-6701, July 2006.

[94] R. Moddemeijer, "On estimation entropy and mutual information of continuous distributions," *Signal Processing*, vol. 16, no. 3, pp. 233-248, 1989.

[95] G.A. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observations space," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1315–1321, May 1999.

[96] G. Coq, "Utilisation d'approches probabilistes basées sur les critères entropiques pour la recherche d'informations sur supports multimédia," Université de Poitiers, Poitiers, Thèse de Doctorat 2008.

[97] H. Joe, "Estimation of entropy and other functionals of a multivariate density," *Ann. Inst. Statist. Math*, vol. 41, no. 4, pp. 683–697, 1989.

[98] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall, 1986.

[99] Y.I. Moon, B. Rajagopalan, and U. Lall, "Estimation of mutual information using kernel density estimators," *Phys. Rev*, vol. 52, no. 3, pp. 2318-2321, 1995.

[100] M. Ait Kerroum, A. Hammouch, and D. Aboutajdine, "Textural feature selection by joint mutual information based on Gaussian mixture model for multispectral image classification," *Pattern Recognition Letters*, vol. 31, no. 10, pp. 1168-1174, July 2010.

[101] R. Moddemeijer, "A statistic to estimate the variance of the histogram-based mutual information estimator based on dependent pairs of Observations," *Signal Processing*, vol. 75, no. 1, pp. 51-63, 1999.

[102] N.J.I. Mars and G.W. van-Arragon, "Time delay estimation in non-linear systems using average amount of mutual information analysis," *Signal Processing*, vol. 4, no. 2, pp. 139-153, 1982.

[103] P. Legg, "Improving Accuracy and Efficiency of Registration by Mutual Information using Sturges Histogram Rule," in *Proc. Medical Image Understanding and Analysis*, Aberystwyth, July 2007, pp. 26-30.

[104] S. Panzeri, R. Senatore, M. Montemurro, and R. Petersen, "Correcting for the Sampling Bias Problem in Spike Train Information Measures," *Neurophysiology*, vol. 98, no. 3, pp. 1064-1072, September 2007.

[105] H. Sturges, "The choice of a class-interval," *J. Amer. Statist. Assoc*, vol. 21, pp. 65–66, 1926.

[106] D.W. Scott, "On optimal and data-based histograms," *Biometrika*, vol. 66, no. 3, pp. 605–610, 1979.

[107] D. Freedman and P. Diaconis, "On the Maximum Deviation Between the Histogram and the Underlying Density," *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 58, no. 2, pp. 139-167, 1981.

[108] P. Hall and S. Morton, "On the estimation of entropy," *Ann. Inst. Statist. Math*, vol. 45, no. 1, pp. 69–88, 1993.

[109] T. Lan, D. Erdogmus, U. Özertem, and Y. Huang, "Estimating Mutual Information Using Gaussian Mixture Model for Feature Ranking and Selection," in *Proc. International Joint Conference on Neural Networks*, Vancouver, BC, canada, 2006, pp. 5034-5039.

[110] H. Shimazaki and S. Shinomoto, "A method for selecting the bin size of a time histogram," *Neural Computation*, vol. 19, no. 6, pp. 1503-1527, June 2007.

[111] H. Nies, O.U. Loffeld, B. Dömnez, A. Ben Hammadi, and R. Wang, "Image Registration of TerraSAR-X Data using Different Information Measures," in *Proc. IEEE Geoscience and Remote Sensing Symposium*, Boston, MA, July 2008, pp. 419-422.

[112] E. Schaffernicht, R. Kaltenhaeuser, S. Verma, and H.M Gross, "On Estimating Mutual Information for Feature Selection," in *Proc. ICANN*, 2010, pp. 362-367.

[113] T.T Soong, *Fundamentals of Probability and Statistics for Engineers*.: John Wiley & Sons, April 2004.

[114] G. McLachlan and D. Peel, *Finite Mixture Models*. USA: John Wiley & Sons, 2000.

[115] X. Zhou, Y. Fu, M. Liu, M.A. Hasegawa-Johnson, and T.S. Huang, "Robust analysis and weighting on MFCC components for speech recognition and speaker identification," in *Proc IEEE on Multimedia and Expo*, Beijing, 2007, pp. 188–191.

[116] C. Yang, F. Soong, and T.M Lee, "Static and dynamic spectral features: Their noise robustness and optimal weights for ASR," *IEEE Trans. Audio Speech Language Process*, vol. 15, no. 3, pp. 1087–1097, March 2007.

[117] (2010) LDC - Linguistic Data Consortium. [Online]. http://ldc.upenn.edu/

[118] J.I Makhoul, "Linear prediction: a tutorial review," in *Proc. of the IEEE*, vol. 63(4), April 1975, pp. 561-580.

[119] A. Amehraye, "Débruitage perceptuel de la parole," Ecole Nationale Supérieure des Télécommunications de Bretagne, Thèse de Doctorat 2009.

[120] H. Hermansky, "An efficient speaker-independent automatic speech recognition by simulation of properties of humain auditory perception," in *Proc. of the IEEE Conference on ICASSP*, vol. 12, Dallas, TX, April 1987, pp. 1159-1162.

[121] H. Fletcher, "Auditory Patterns," *Review of Modern Physics*, vol. 12, no. 1, pp. 47-65, January 1940.

[122] D.W. Robinson and R.S. Dadson, "A Redetermination of the Equal-Loudness relations for Pure Tones," *British Journal of Applied Physics*, vol. 7, pp. 166-181, 1956.

[123] S.S. Stevens, "On the Psychophysical Law," *Phychological Review*, vol. 64, no. 3, pp. 153-181, 1957.

[124] J.C Junqua, "Utilisation d'un modèle d'audition et de connaissances phonétiques en reconnaissance automatique de la parole," *Traitement du signal*, vol. 7, no. 4, pp. 275–284, 1990.

Abdenour Hacine-Gharbi

Sélection de paramètres acoustiques pertinents pour la reconnaissance de la parole

Résumé :

L'objectif de cette thèse est de proposer des solutions et améliorations de performance à certains problèmes de sélection des paramètres acoustiques pertinents dans le cadre de la reconnaissance de la parole. Ainsi, notre première contribution consiste à proposer une nouvelle méthode de sélection de paramètres pertinents fondée sur un développement exact de la redondance entre une caractéristique et les caractéristiques précédemment sélectionnées par un algorithme de recherche séquentielle ascendante. Le problème de l'estimation des densités de probabilités d'ordre supérieur est résolu par la troncature du développement théorique de cette redondance à des ordres acceptables. En outre, nous avons proposé un critère d'arrêt qui permet de fixer le nombre de caractéristiques sélectionnées en fonction de l'information mutuelle approximée à l'itération j de l'algorithme de recherche.

Cependant l'estimation de l'information mutuelle est difficile puisque sa définition dépend des densités de probabilités des variables (paramètres) dans lesquelles le type de ces distributions est inconnu et leurs estimations sont effectuées sur un ensemble d'échantillons finis. Une approche pour l'estimation de ces distributions est basée sur la méthode de l'histogramme. Cette méthode exige un bon choix du nombre de bins (cellules de l'histogramme). Ainsi, on a proposé également une nouvelle formule de calcul du nombre de bins permettant de minimiser le biais de l'estimateur de l'entropie et de l'information mutuelle. Ce nouvel estimateur a été validé sur des données simulées et des données de parole. Plus particulièrement cet estimateur a été appliqué dans la sélection des paramètres MFCC statiques et dynamiques les plus pertinents pour une tâche de reconnaissance des mots connectés de la base Aurora2.

Mots clés : reconnaissance de la parole, paramètres acoustiques, coefficients MFCC, modèles de Markov cachés (MMC), entropie, information mutuelle, histogramme, nombre de bins, sélection des paramètres, pertinence, redondance, biais.

Relevant acoustic feature selection for speech recognition

Summary :

The objective of this thesis is to propose solutions and performance improvements to certain problems of relevant acoustic features selection in the framework of the speech recognition. Thus, our first contribution consists in proposing a new method of relevant feature selection based on an exact development of the redundancy between a feature and the feature previously selected using Forward search algorithm. The estimation problem of the higher order probability densities is solved by the truncation of the theoretical development of this redundancy up to acceptable orders. Moreover, we proposed a stopping criterion which allows fixing the number of features selected according to the mutual information approximated at the iteration J of the search algorithm.

However, the mutual information estimation is difficult since its definition depends on the probability densities of the variables (features) in which the type of these distributions is unknown and their estimates are carried out on a finite sample set. An approach for the estimate of these distributions is based on the histogram method. This method requires a good choice of the bin number (cells of the histogram). Thus, we also proposed a new formula of computation of bin number that allows minimizing the estimator bias of the entropy and mutual information. This new estimator was validated on simulated data and speech data. More particularly, this estimator was applied in the selection of the static and dynamic MFCC parameters that were the most relevant for a recognition task of the connected words of the Aurora2 base.

Keywords : speech recognition, acoustic feature, MFCC coefficient, hidden Markov models (HMM), entropy, mutual information, histogram, bins number, feature selection, relevance, redundancy, bias.

UNIVERSITE D'ORLEANS

Laboratoire PRISME
Université d'Orléans, France

Université Ferhat Abbas - Sétif