

République Algérienne Démocratique et populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université des Sciences et de la Technologie
Houari Boumediène
Faculté d'Electronique et d'Informatique



MEMOIRE
Présenté pour l'obtention du diplôme de MAGISTER
En : ELECTRONIQUE
Spécialité : Communication Parlée

Par
BOUBAKEUR Khadidja Nesrine

Thème

**Modélisation Prosodique pour la Reconnaissance
Automatique du Locuteur**

Soutenu publiquement le 19 Juin 2012, devant le jury composé de :

<i>M^{me} R. Touhami</i>	<i>Professeur</i>	<i>à l'USTHB</i>	<i>Présidente</i>
<i>M^r M. Debyeche</i>	<i>Maître de Conférences /A</i>	<i>à l'USTHB</i>	<i>Directeur de Mémoire</i>
<i>M^r B. Fergani</i>	<i>Maître de Conférences /A</i>	<i>à l'USTHB</i>	<i>Examineur</i>
<i>M^{me} L. Falek</i>	<i>Maître de Conférences /A</i>	<i>à l'USTHB</i>	<i>Examinatrice</i>

Remerciements

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

Je tiens à remercier Mme R. TOUHAMI, Professeur à la Faculté d'Electronique et d'Informatique, USTHB, pour l'honneur qu'elle me fait de présider le Jury de soutenance.

Mes remerciements s'adressent également à Mr B. FERGANI, Maître de Conférences à la Faculté d'Electronique et d'Informatique de l'USTHB, et Mme L. FALEK, Maître de Conférences à la Faculté d'Electronique et d'Informatique de l'USTHB, membres examinateurs du jury, pour l'intérêt qu'ils ont porté à notre travail et pour leur disponibilité.

Grand merci à mes chers parents, qui m'ont inculqué la soif du savoir. Je remercie mon fiancé qui m'a toujours soutenu, encouragé et qui a toujours été présent quand j'en avais besoin.

Je remercie également tous ceux qui m'ont aidé et encouragé de près ou de loin.

Abstract

The Automatic Speaker Recognition (ASR) is the process that automatically determines the identity of the speaker based on his voice characteristics. Currently, this type of systems are widely used in many fields, such as securing access to protected sites, to do banking, or to meet a court for a crime or a civil dispute.

Systems of automatic speaker recognition typically use cepstral-based features (standard features) as acoustic modeling of speech signal. The most powerful features currently used are the MFCC (Mel-Frequency Cepstral Coefficients), the LPCC (Linear Predictive Cepstrum Coefficients) and the PLP (Perceptual Linear Predictive). However, these features are very sensitive to speech signal variability under real-life conditions. The speech signal variability is mostly due to environmental factor (presence of noise), it leads to different kinds of mismatch between acoustic features and acoustic models. This causes a reduction on the recognition rate under real-life conditions. The sensitivity of standard features to noise motivates many authors to look for new parameters to make the acoustic models more robust.

The main focus of this work is to improve the speaker recognition in real environment (noisy), by integrating prosodic parameters into standard ASR systems so as to make the acoustic models more robust to the variability in the speech signal under real-life conditions. Thus, two systems of Automatic Speaker Identification (ASI) in independent mode of text have been implemented. The first one is based on Hidden Markov Models (HMM) and the second uses the Support Vector Machines (SVM). The standard parameters used in the two systems are the MFCC (Mel Frequency Cepstral Coefficients).

The prosodic parameters proposed in this work are the pitch frequency and energy. The introduction of these parameters in both systems was implemented by a strategy called "Direct Integration strategy". The fusion process is based on the concatenation of the prosodic parameters with the standard cepstral features in the same acoustic vector.

Very interesting improvements are obtained in noisy environments for different SNR (Signal to Noise Ratio). This improvement is more pronounced for the system based on SVM. This motivates the use of prosodic parameters for speaker recognition in real environment.

Résumé

La Reconnaissance Automatique du Locuteur (RAL) est le processus qui détermine automatiquement l'identité de celui qui parle en se basant sur ses caractéristiques vocales. Actuellement, ce type de système est largement utilisé dans plusieurs domaines, tels que la sécurisation d'accès à des sites protégés, pour faire des transactions bancaires, ou pour répondre aux besoins d'un tribunal pour un crime ou un litige civil.

Les systèmes de Reconnaissance Automatique du Locuteur utilisent généralement des paramètres cepstraux dits paramètres standards comme modélisation acoustique du signal de parole. Les paramètres cepstraux les plus performants actuellement sont les coefficients MFCCs (Mel Frequency Cepstral Coefficients), les coefficients LPCC (Linear Predictive Cepstrum Coding) et les coefficients PLP (Perceptual Linear Predictive). Cependant, ces paramètres restent très sensibles aux variations du signal vocal. Cette variabilité est causée par l'environnement (présence de bruit) et le locuteur (l'âge, la fatigue, le sexe). Elle conduit souvent à différents types de disparités entre observations et modèles acoustiques, ce qui engendre des faibles taux de reconnaissance dans les conditions réelles (bruitées). La sensibilité des paramètres standards à la variabilité du signal a motivé plusieurs chercheurs à utiliser de nouveaux paramètres et de nouveaux paradigmes pour rendre les modèles acoustiques plus robustes.

L'objet de ce mémoire est de chercher à améliorer la reconnaissance du locuteur en environnement réel, en intégrant des paramètres prosodiques. Ainsi deux systèmes d'Identification Automatique du Locuteur (IAL) en mode indépendant du texte ont été mis en œuvre. Le premier basé sur les modèles de Markov cachés (HMM) réalisé sous la plateforme HTK (Hidden Markov Toolkit), le deuxième utilisant les Support Vector Machines (SVM). Les paramètres standards utilisés pour les deux systèmes sont les coefficients MFCCs.

Les paramètres prosodiques proposés dans ce travail sont la fréquence fondamentale (pitch) et l'énergie. L'introduction de ces paramètres dans les deux systèmes d'IAL a été réalisée par une stratégie dite à Identification Directe (ID) ou fusion de paramètres, les deux types de paramètres (standards et prosodiques) sont concaténés dans le même vecteur pour former une seule observation à l'entrée du système d'IAL.

Les expériences de validation réalisées sur la base de données ARADIGITS en milieu clean et bruité à différents niveaux RSB (Rapport Signal/Bruit) ont montré que l'intégration des paramètres prosodiques améliore les performances des deux systèmes. Cette amélioration est plus marquée pour le système basé sur les SVM. Ceci motive l'utilisation des paramètres prosodiques pour la reconnaissance du locuteur en environnement réel.

Table des matières

Table des matières.....	i
Liste des figures.....	iv
Liste des tableaux.....	vi
Liste des abréviations.....	vii
Introduction générale.....	01
Chapitre 1: Généralités sur la Reconnaissance Automatique du Locuteur.....	05
1.1 Introduction.....	06
1.2 Tâches de la RAL.....	06
1.2.1 Vérification Automatique du Locuteur.....	06
1.2.2 Identification Automatique du Locuteur.....	07
1.3 Structure d'un système de RAL.....	08
1.4 Paramétrisation acoustique.....	09
1.5 Approches de reconnaissance.....	10
1.5.1 L'approche DTW.....	10
1.5.2 La Quantification Vectorielle.....	11
1.5.3 Les Modèles de Markov cachés.....	11
1.5.4 Approche basée sur les SVM.....	12
1.6 Structure d'un système de RAL.....	12
1.7 Domaines d'applications.....	13
1.7.1 Applications sur l'accès restreint sécurisé à des sites sensibles.....	14
1.7.2 Applications dans les systèmes de communication.....	14
1.7.3 Applications juridiques.....	14
1.8 Conclusion.....	15
Chapitre 2 : Approches de reconnaissance mises en œuvre et Fusion.....	16
2.1 Introduction.....	17
2.2 Modèles de Markov Cachés.....	17
2.2.1 Définition.....	17
2.2.2 Les paramètres d'un Modèle de Markov Caché.....	18
2.2.3 Principaux types de HMM.....	19

2.2.3.1 Contraintes liées à la matrice de transition A	19
2.2.3.2 Type de densité de probabilité d'observations $b_j(O_t)$	20
2.2.4 Topologie des modèles HMMs utilisés pour la reconnaissance du locuteur	22
2.2.5 Trois problèmes fondamentaux d'un HMM	23
2.2.6 Reconnaissance du Locuteur par HMMs	23
2.2.6.1 Evaluation de la probabilité d'émission d'une séquence	24
2.2.6.2 Apprentissage des modèles	25
2.2.6.3 La reconnaissance	25
2.3 Théorie des Machines à Vecteurs de Support	27
2.3.1 Minimisation du risque structurel	27
2.3.2 Classification binaire par hyperplan	29
2.3.2.1 SVM Linéaire	29
2.3.2.2 SVM non linéaires	33
2.3.2.3 Séparation imparfaite (Cas de données inséparables)	38
2.4 SVM à plusieurs classes	41
2.4.1 Approches "un contre tous"	41
2.4.2 Approches "un contre un"	42
2.5 Stratégies de fusion de données	42
2.5.1 Fusion des paramètres	43
2.5.2 Fusion des scores	44
2.5.2.1 Méthodes de fusion	45
2.6 Conclusion	46
Chapitre 3 : Paramétrisation du signal de parole	47
3.1 Introduction	48
3.2 Signal de parole	48
3.2.1 Production et perception de la parole	48
3.2.2 Variabilité du signal de parole	50
3.2.2.1 Variabilité inter-locuteurs	50
3.2.2.2 Variabilité intra-locuteur	51
3.2.2.3 Variabilité contextuelle	52
3.2.2.3 facteurs de distorsion	52
3.3 Paramètres de l'analyse spectrale	53
3.3.1 Coefficients cepstraux	55

3.3.2 Les coefficients MFCC.....	56
3.4 Paramètres prosodiques.....	60
3.4.1 La fréquence fondamentale (pitch).....	60
3.4.1.1 Méthodes de détection de la F0.....	60
3.4.1.2 Méthode basée sur la fonction autocorrélation.....	62
3.4.1.3 La fonction d'AMDF basée sur le Clippage	64
3.4.1.4 Le Produit Harmonique Spectral	64
3.4.2 L'énergie.....	65
3.5 Paramètres dynamiques.....	65
3.6 Conclusion	66
Chapitre 4 : Résultats Expérimentaux.....	67
4.1 Introduction	68
4.2 Protocole expérimental.....	68
4.2.1 La base de données d'évaluation.....	68
4.2.2 Évaluation des résultats.....	69
4.2.3 Paramétrisation du signal vocal.....	70
4.2.1.1 Extraction des coefficients MFCC.....	70
4.2.1.2 Paramètres prosodiques.....	70
4.3 Construction d'un système de référence.....	70
4.3.1 Caractéristiques du système de référence	71
4.3.2 Validation du système de référence.....	72
4.3.2.1 Apport des coefficients différentiels.....	72
4.3.2.2 Apport du nombre d'états.....	74
4.4 Construction d'un système d'IAL basé sur les SVM.....	76
4.5 Fusion des paramètres prosodiques dans les deux systèmes.....	77
4.6 Conclusion.....	83
Conclusion et Perspectives.....	84
Bibliographie.....	87
Annexe A : Algorithmes Forward, Backward et Forward-Backward.....	95
Annexe B : Construction d'un système de RAL sous HTK.....	101
Annexe C : Le corpus Noisex-92.....	108

Table des Figures

Figure 1.1 : Schéma modulaire d'un système de VAL.....	7
Figure 1.2 : Schéma modulaire d'un système de IAL.....	8
Figure 1.3 : Schéma synoptique d'un système de RAL.....	9
Figure 1.4 : Calcul de la distance dynamique par DTW.....	11
Figure 2.1 : Exemple de modèle ergodique à 4 états.....	19
Figure 2.2 : Exemple de modèle gauche-droite à 4 états.....	
Figure 2.3 : Exemple de modèle de Bakis à trois états.....	22
Figure 2.4 : Hyperplans séparateurs.....	30
Figure 2.5 : Transposition du problème vers un espace de dimension élevé.....	34
Figure 2.6 : Frontières de décision obtenues par trois SVM à noyau linéaire, polynomial et RBF.....	36
Figure 2.7 : Application du noyau RBF à deux données x et x'	37
Figure 2.8 : La similarité des données vue en utilisant un noyau RBF à σ tendant vers 0.....	38
Figure 2.9 : Hyperplans séparateurs dans le cas de données non-linéairement séparables (VS sont les Vecteurs Supports).....	38
Figure 2.10 : Problème à trois classes : frontières de décision linéaires dans la stratégie Un-contre-Tous.....	42
Figure 2.11 : Fusion des paramètres standard et prosodique dans un système de RAL.....	42
Figure 2.12 : Fusion des scores des systèmes standard et prosodique dans un système RAL par identification séparée.....	43
Figure 3.1 : L'appareil phonatoire humain.....	49
Figure 3.2 : Système auditif périphérique humain.....	50
Figure 3.3 : Variabilité inter-locuteurs pour le chiffre arabe « siffer ».....	51
Figure 3.4 : Variabilité intra-locuteur pour le chiffre arabe « siffer ».....	52
Figure 3.5 : Variabilité due à un environnement bruité avec un niveau RSB (Rapport Signal/Bruit) = 0 dB d'un bruit d'usine pour le chiffre arabe « siffer ».....	53
Figure 3.6 : étapes de mise en forme du signal de parole.....	54
Figure 3.7 : Etapes de calcul du cepstre.....	55
Figure 3.8 : Processus de calcul des coefficients MFCC.....	57
Figure 3.9 : Banc de filtres espacés (a) sur l'échelle Mel en Hertz (b) en Mel-Fréquence (B(f)).....	58

Figure 3.10 : Graphe de la fonction 3-level center clipping.....	63
Figure 4.1 : Création de la base de test bruitée.....	69
Figure 4.2 : Principe d'identification du locuteur par HMM.....	71
Figure 4.3 : Taux de reconnaissance comparatifs coefficients statiques/coefficients.....	74
Figure 4.4 : Modèle de Bakis à 3 états émetteurs.....	74
Figure 4.5 : Influence du nombre d'état sur le taux d'identification.....	75
Figure 4.6 : Construction de vecteurs composites de dimension $3n$	77
Figure 4.7 : Taux comparatifs des systèmes de fusion/système de référence dans un milieu calme et bruité par le d'usine à différents RSB.....	79
Figure 4.8 : Taux comparatifs système basé sur les SVM / systèmes de fusion pour le bruit. de la foule.....	80
Figure 4.9 : Influence du bruit sur les systèmes à paramètres standards.....	82
Figure 4.10 : Influence de l'ajout des paramètres prosodiques sur les deux systèmes.....	82

Liste des tableaux

Tableau 4.1: Taux comparatifs coefficients statiques/coefficients dynamiques en fonction du nombre de gaussiennes.....	73
Tableau 4.2: Influence du nombre d'état sur le taux d'identification.....	75
Tableau 4.3 : Taux comparatifs des systèmes de fusion à base de HMM /système de référence pour trois types de bruits.....	78
Tableau 4.4 : Taux comparatifs système basé sur les SVM / systèmes de fusion pour trois types de bruits.....	80
Tableau 4.5 : Taux comparatifs des différents systèmes mis en œuvre.....	81

Liste des abréviations

ACP	Analyse en Composantes Principales.
DTW	Dynamic Time Warping (alignement temporel dynamique).
EM	Expectation-Maximisation
GMM	Gaussian Mixture Model.
HMM	Hidden Markov Models (modèles de Markov cachés).
HO	Hyperplan Optimal.
HTK	HMM Toolkit (boîte à outils pour modèles de Markov cachés).
IAL	Identification Automatique du Locuteur.
LPC	Linear Predictive Coding.
MLE	Maximum Likelihood Estimation.
MFCC	Mel Frequency Cepstral Coefficient (coefficient cepstral en échelle Mel).
PLP	Perceptual Linear Prediction (prédiction linéaire perceptive).
QV	Quantification Vectorielle.
RAL	Reconnaissance Automatique du Locuteur.
RBF	Radial Basis Function.
RSB	Rapport Signal sur Bruit.
SVM	Support Vector Machine.
TCDI	Transformée en Cosinus Discrète Inverse.
TFD	Transformée de Fourier Discrète.
VAL	Vérification Automatique du Locuteur.

Introduction générale

Introduction générale

Le domaine de l'authentification biométrique est depuis quelques années en pleine expansion. En effet, un grand nombre d'applications nécessitent une phase d'authentification de l'utilisateur. Ceci requiert des moyens fiables pour reconnaître l'identité de la personne qui se présente au système. Or les moyens classiques reposant sur des mots de passe ou des cartes magnétiques associées à un code personnel présentent un certain nombre d'inconvénients. Un mot de passe peut être oublié ou volé par un autre individu de même pour les cartes d'accès. C'est ainsi que l'exploitation de caractéristiques ou mesures liées à la physiologie de l'individu (sa voix, son visage, sa signature, ses empreintes digitales, la forme de sa main,...) est apparue naturellement comme la solution la plus fiable.

La voix est une caractéristique biométrique porteuse d'informations variées telles que la langue parlée par le locuteur, son identité ou même des indications sur son âge ou son état émotionnel.

La Reconnaissance Automatique du Locuteur (RAL), s'intéresse à l'extraction de l'information caractéristique de l'identité du locuteur. Cette information, appelée empreinte vocale, servira comme indice biométrique d'identification. L'intérêt essentiel de cette discipline est de garantir la confidentialité et l'authentification lors des transactions numérisées.

Les recherches en RAL réalisées progressivement depuis une vingtaine d'années, ont fait des progrès considérables et ont obtenu des taux de reconnaissance quasiment parfait. Cependant, cette performance a été atteinte avec de la parole enregistrée dans des conditions propres, tandis que pour la parole bruitée (cas réel), les performances obtenues ne sont pas pour autant satisfaisantes.

Les paramètres cepstraux dits paramètres standards sont les plus utilisés comme modélisation acoustique du signal de parole. Cependant, ces paramètres ont montré une certaine sensibilité à la variabilité du signal vocal. Cette sensibilité affecte négativement les taux de reconnaissance dans les conditions réelles (bruitées). Une des solutions envisagée

par les recherches menées actuellement dans ce domaine est l'intégration des paramètres prosodiques dans le système de RAL.

Dans ce cadre, ce mémoire s'intéresse au problème de la robustesse des systèmes de RAL en environnement réel. L'étude a porté sur l'influence de l'approche de reconnaissance et de l'intégration des paramètres prosodiques dans le front-end. Ainsi deux systèmes d'Identification Automatique du Locuteur (IAL) en mode indépendant du texte ont été mis en œuvre. Le premier basé sur les modèles de Markov cachés (HMM) réalisé sous la plateforme HTK (Hidden Markov Toolkit), le deuxième utilisant les Support Vector Machines (SVM).

Les paramètres prosodiques proposés dans ce travail sont la fréquence fondamentale (pitch) et l'énergie. Les deux types de paramètres (standards et prosodiques) ont été concaténés dans le même vecteur pour former une seule observation à l'entrée du système d'IAL.

Organisation du mémoire

Le premier chapitre présente une introduction à la reconnaissance du locuteur. Les différentes caractéristiques et la structure générale d'un système de RAL, ainsi que les paramètres les plus efficaces pour représenter le signal de parole et les méthodes de reconnaissance les plus utilisées actuellement.

Dans le deuxième chapitre nous détaillons les deux approches de reconnaissance utilisées dans notre système et les méthodes de fusion. Les modèles de Markov cachés, leur théorie, leurs principes pour la modélisation du locuteur ainsi que les algorithmes d'apprentissage et de décodage permettant la construction d'un système de RA sont d'abord présentés. Ensuite, nous présentons la formulation mathématique des SVM et le principe de classification par hyperplan. En dernier, nous présentons quelques méthodes de fusion de données dans un système de RAL.

Le troisième chapitre est le chapitre noyau de ce mémoire, il présente les paramètres standards et les paramètres prosodiques que nous proposons comme vecteur d'information additionnelle et leurs méthodes d'extraction.

Dans le chapitre quatre nous allons présenter le contexte expérimental et les résultats obtenus avec les différents systèmes mis au point ainsi que l'évaluation objective de ces systèmes.

Nous terminons ce manuscrit par une conclusion générale en mettant en évidence le travail réalisé dans ce mémoire. A la lumière des résultats obtenus nous donnons aussi les perspectives futures envisagées.

Chapitre 1

Généralités sur la Reconnaissance Automatique du Locuteur

1.1 Introduction

La caractérisation automatique du locuteur est un vaste domaine dans lequel la "machine" a pour tâche d'extraire du signal de parole les informations de nature à renseigner sur les spécificités d'un individu : identité, caractéristiques physiques, émotivité, état pathologique, particularités régionales, etc. Elle s'applique à différents thèmes de recherche traitant des informations extralinguistiques véhiculées par la voix tels que la classification d'individus, ou l'étude psychique ou physiologique d'une personne.

La Reconnaissance Automatique du Locuteur (RAL) est un sous-problème de la caractérisation automatique du locuteur. Son objectif est de reconnaître l'identité d'une personne à partir de sa voix [Dodd-85] [Rose-91]. La variabilité de la parole entre locuteurs (variabilité inter-locuteur) est l'essence même de la RAL. Sans cette variabilité, il serait impossible de reconnaître une voix parmi plusieurs voix possibles.

On peut distinguer deux modes de RAL :

- **RAL en mode dépendant du texte** : dans ce genre de système, la reconnaissance du d'une personne est réalisée sur la base d'un message dont le contenu linguistique (mot de passe, code PIN, phrase,...) est connu à priori.
- **RAL en mode indépendant du texte** : dans ce genre de système, il n'existe aucune contrainte sur le message prononcé par la personne.

1.2 Tâches de la RAL

On distingue deux tâches principales en reconnaissance du locuteur : l'identification et la vérification du locuteur.

1.2.1 Vérification Automatique du Locuteur

La Vérification Automatique du Locuteur (VAL) consiste à déterminer si un locuteur est bien celui qu'il prétend être. Dans ce type d'applications, il s'agit donc de trancher entre les deux hypothèses : soit le locuteur est bien le locuteur autorisé, c'est à dire celui dont l'identité est revendiquée, soit nous avons affaire à un imposteur qui cherche à se faire passer pour un locuteur autorisé. Le système dispose en entrée d'un échantillon de parole et d'une identité proclamée. Une mesure de ressemblance est calculée entre l'échantillon et la référence du locuteur correspondant à l'identité proclamée. Si cette mesure est en dessous d'un certain

seuil, le système accepte le locuteur ; dans le cas contraire, le locuteur est considéré comme un imposteur et rejeté.

La figure 1.1 représente un schéma modulaire d'un système de VAL.

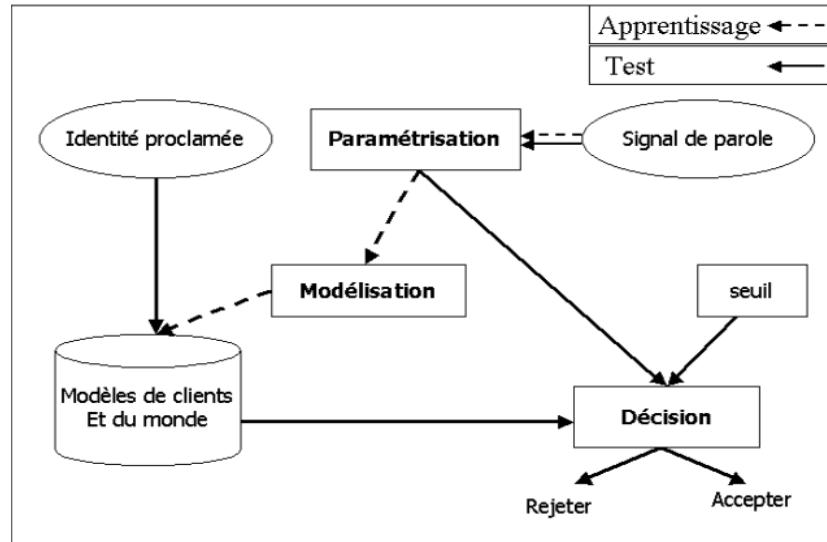


Figure 1.1 : Schéma modulaire d'un système de VAL.

1.2.2 Identification Automatique du Locuteur

L'Identification Automatique du Locuteur (IAL) consiste à reconnaître la voix d'un locuteur parmi une population (une base de données) composée de N locuteurs connus.

Deux conditions d'identification sont connues : milieu ouvert ou fermé.

- **IAL en milieu fermé :** Le locuteur est supposé être l'un des N locuteurs du système. L'entrée du système est l'échantillon de parole d'un locuteur inconnu. La sortie du système correspond à l'identité du locuteur de la base de référence qui est le plus "proche" du signal de parole inconnu.
- **IAL en milieu ouvert :** Le système peut décider qu'aucune des N identités connues n'est celle du locuteur. C'est une combinaison des deux tâches précédentes :
 - a) Identification du locuteur le plus probable parmi les locuteurs de la base.
 - b) Vérification que l'échantillon inconnu a bien été prononcé par le locuteur choisi dans l'étape d'identification.

La figure 1.2 représente un schéma modulaire d'un système d'IAL.

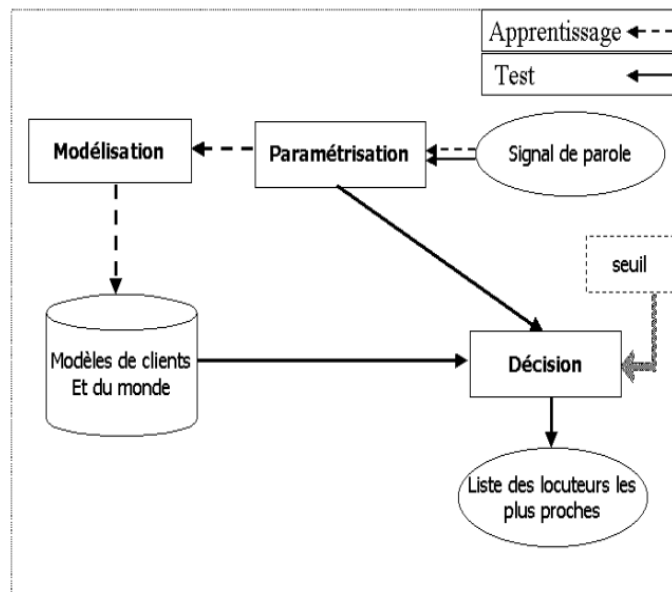


Figure 1.2 : Schéma modulaire d'un système d'IAL.

1.3 Structure d'un système de RAL

Nous distinguons deux phases dans le processus de la reconnaissance du locuteur: l'apprentissage et le test tel qu'illustré par la figure 1.3.

Lors de la phase d'apprentissage, un modèle associé à chaque locuteur de la base d'apprentissage. Cette phase se déroule en deux étapes :

- a) La paramétrisation qui consiste essentiellement à réduire l'information du signal vocal en quantité et en redondance tout en augmentant la discrimination. A la sortie, le signal est représenté par un ensemble de vecteurs de coefficients.
- b) La modélisation qui consiste à estimer les paramètres d'un modèle mathématique approprié pour les vecteurs de coefficients.

Lors de la phase de test, une mesure de similarité est établie entre les vecteurs testés et les données stockées comme références pour chaque locuteur. En sortie de cette phase, le système émet une réponse (décision): une identité pour la tâche d'identification ou une décision accès/rejet pour la vérification. Notons que seule l'étape de paramétrisation est nécessaire pour la phase de test.

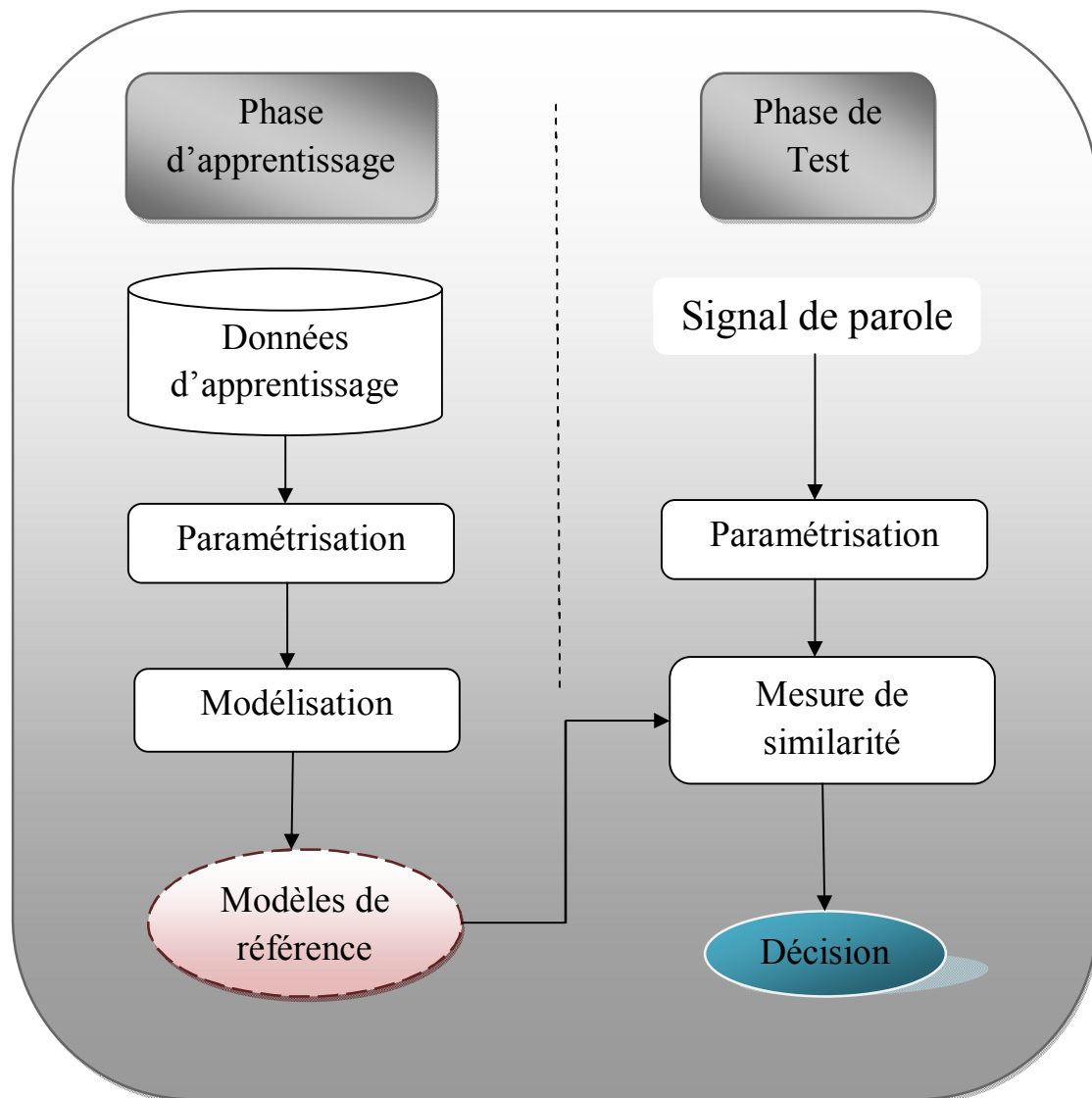


Figure 1.3 : Schéma synoptique d'un système de RAL.

1.4 Paramétrisation acoustique

Le processus de paramétrisation consiste à extraire du signal de parole les informations pertinentes en vue de la reconnaissance. Le signal de parole, de part sa complexité (multitudes d'informations et redondance), ne peut être exploité directement. Une représentation simplifiée du signal de parole est par conséquent nécessaire. Cette représentation repose généralement sur des vecteurs de paramètres acoustiques, calculés périodiquement sur le signal de parole. Dans son article sur le choix de paramètres efficaces pour la reconnaissance

du locuteur, Wolf [Wolf, 72] décrit les attributs nécessaires des "bons paramètres" pour la RAL.

Idéalement, les paramètres (ou traits acoustiques) :

- doivent être fréquents (i.e. ne pas correspondre à des événements ne survenant que très rarement dans le signal de parole d'un locuteur),
- doivent être facilement mesurables, ne pas être trop sensibles à la variabilité intra-locuteur,
- ne doivent pas être affectés par le bruit ambiant ou par les variations dûes au canal de transmission,
- doivent être robustes face aux imitateurs.

En pratique, il est très difficile de réunir tous ces attributs en même temps. La sélection de traits acoustiques pertinents pour la RAL est donc un sujet largement traité dans la littérature. Il ressort que les seuls types de paramètres vraiment pertinents et utilisables efficacement sont les paramètres de l'analyse spectrale et éventuellement les paramètres prosodiques ainsi que les paramètres dynamiques. L'extraction de ces paramètres sera détaillée au chapitre 3.

1.5 Approches de reconnaissance

Plusieurs approches de reconnaissance ont été utilisées. Nous pouvons citer ;

- l'approche basée sur la programmation dynamique connue sous l'acronyme DTW (Dynamic Time Warping) [Bell-57],
- l'approche basée sur la quantification vectorielle,
- l'approche statistique à base de modèles de Markov cachés (Hidden Markov Models : HMM en anglais),
- l'approche basée sur les Support Vector Machine (SVM) [Wan-00].

Les deux dernières approches qui seront utilisées dans le cadre de notre travail.

1.5.1 L'approche DTW

Utilisée en mode dépendant du texte [Mami-03], cette technique effectue la comparaison entre la forme d'entrée à reconnaître et la forme de référence en calculant la distance entre les paramètres des deux formes. Elle détermine le meilleur chemin reliant le début et la fin des deux blocs de paramètres. Ainsi cet algorithme permet de trouver un alignement temporel optimal entre la forme d'entrée et la forme de référence (figure 1.4). Cet alignement est réalisé par une technique de programmation dynamique. Malgré les bonnes performances obtenues

par cette technique, elle reste très sensible à la qualité de l'alignement et notamment le choix du point de départ des deux formes à comparer.

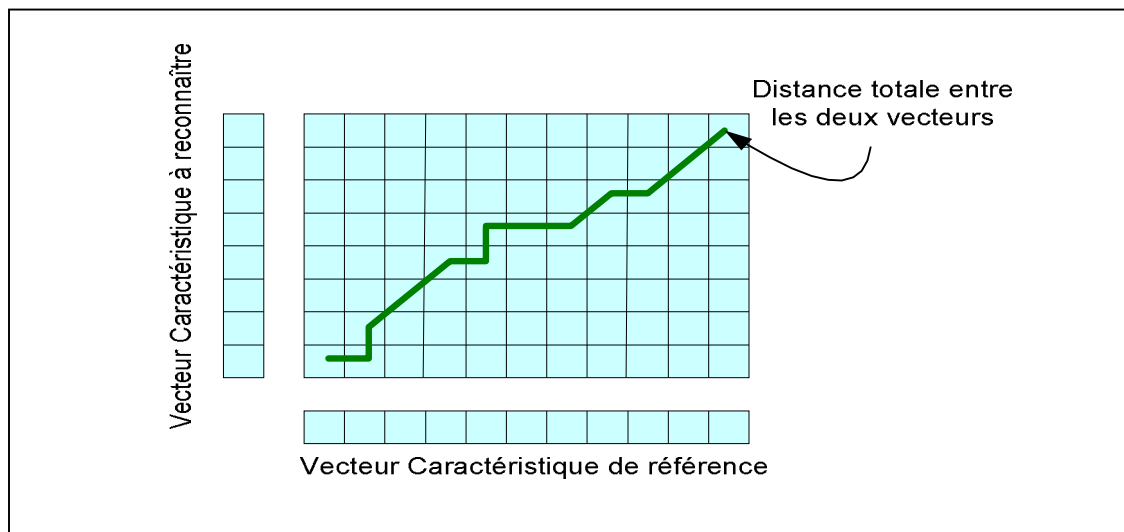


Figure 1.4: Calcul de la distance dynamique par DTW.

1.5.2 La Quantification Vectorielle

La quantification vectorielle ou (VQ : Vector Quantization) est une méthode non paramétrique qui permet de décrire un ensemble de données par un faible nombre de vecteurs formant un dictionnaire (code book) associé aux données. Le dictionnaire est en général calculé de façon à ce que la distance moyenne entre un vecteur issu des données et son plus proche voisin dans le dictionnaire soit la plus petite possible. En reconnaissance de locuteur ce dictionnaire est réalisé à partir des vecteurs spectraux provenant de l'analyse du signal de parole de chaque locuteur. Les performances et la rapidité de cette technique dépendent fortement de la taille du dictionnaire. En effet, plus la taille du dictionnaire est grande meilleures sont les performances, mais le processus de test devient trop lent [Kharr-03].

1.5.3 Les Modèles de Markov cachés

Les modèles de Markov (ou HMM pour Hidden Markov Models) ont été initialement introduits en reconnaissance de la parole. Puis leur utilisation s'est étendue peu à peu au domaine de la reconnaissance du locuteur. Les HMM permettent de caractériser les variations temporelles du signal vocal. Ils reposent sur une machine à états, i.e. une succession d'états associés à des probabilités de transition d'un état à l'autre. Une ou plusieurs distributions de probabilité associées à chaque état caractérisent les probabilités d'émission des vecteurs

acoustiques par un état. Lors de la reconnaissance, la vraisemblance pour qu'une séquence de vecteurs de test soit issue de la chaîne de Markov est calculée. Cette approche est détaillée dans le chapitre 2.

1.5.4 Approche basée sur les SVM

Les Support Vector Machines (SVM) sont des nouvelles techniques discriminantes dans la théorie de l'apprentissage statistique. Elles ont été proposées en 1995 par V. Vapnik dans son livre « The nature of statistical learning theory » [Vap-95]. La première application des SVM en reconnaissance du locuteur, notamment pour l'identification du locuteur a été en 1996 [Sch-96]. Elles ont été utilisées aussi pour la vérification du locuteur [Kha-02],

L'idée essentielle des SVM consiste à projeter les données de l'espace d'entrée (appartenant à deux classes différentes) non-linéairement séparables dans un espace de plus grande dimension appelé espace de caractéristiques de façon à ce que les données deviennent linéairement séparables. Dans cet espace, la technique de construction de l'hyperplan optimal est utilisée pour calculer la fonction de classement séparant les deux classes.

Pour appliquer les SVM à la tâche de reconnaissance, une idée simple consisterait à apprendre pour chaque locuteur une frontière de décision (non linéaire) dans l'espace des vecteurs d'observation et à moyenner les sorties sur chaque vecteur en phase de test. Cette approche est détaillée dans le chapitre 2.

1.6 Décision et mesure de performance

Dans tous les systèmes de reconnaissance du locuteur, il faut, à un moment ou à un autre, prendre la décision d'accepter ou de rejeter un locuteur. Nous allons présenter le module de décision ci-dessous pour les tâches de vérification et d'identification.

a) Identification du locuteur

En identification, un signal de test est comparé à toutes les références de locuteurs connus du système, résultant en un ensemble de mesures de similarité (ou un ensemble de mesure de distance) en entrée du processus de décision. Aussi, ce processus a pour tâche de rechercher la mesure de similarité maximale (ou distance minimale dans le cas de mesures de distance) et de désigner l'identité du locuteur correspondant.

Dans ce contexte, la mesure des performances d'un système d'IAL est généralement donnée en termes de Taux d'Identification Correcte (TIC). Ce taux s'obtient par la formulation suivante :

$$\text{TIC} = \frac{\text{nombre de tests correctement identifiés}}{\text{nombre total de tentatives}} \quad (1.5)$$

b) Vérification du locuteur

En vérification, le processus de décision consiste à comparer la mesure de similarité entre le signal de test et le modèle client à un seuil de décision. Si la mesure est supérieure au seuil, le locuteur est considéré comme un client et il est accepté. Dans le cas contraire, le locuteur est considéré comme un imposteur et il est rejeté.

Plusieurs solutions sont proposées dans la littérature pour mesurer les performances d'un système de VAL. Toutes s'appuient sur deux erreurs caractéristiques des systèmes de VAL qui sont :

- **l'erreur de fausse acceptation** : le système reconnaît à tort l'identité revendiquée (acceptation d'un imposteur).
- **l'erreur de faux rejet** : le système rejette à tort l'identité revendiquée (rejet d'un client).

Ces erreurs conduisent aux taux d'erreurs de fausse acceptation f_a et de faux rejet f_r calculés pour un seuil de décision donné et définis par :

$$f_a = \frac{\text{nombre de tentatives d'imposteurs acceptés}}{\text{nombre total de tentatives d'imposteurs}} \quad (1.6)$$

$$f_r = \frac{\text{nombre de tentatives d'abonnés rejetés}}{\text{nombre total de tentatives rejetés}} \quad (1.7)$$

1.7 Domaines d'applications

Dans ce paragraphe on donne quelques exemples d'applications en RAL et que l'on peut regrouper en trois catégories principales : applications en contrôle d'accès sur sites sensibles, application dans le domaine sécuritaire et juridiques (notamment en sciences forensiques) et les applications dans les systèmes de communication.

1.7.1 Applications sur l'accès restreint sécurisé à des sites sensibles

Cette catégorie concerne les applications qui se trouvent sur un site géographique particulier, elles sont utilisées principalement pour limiter l'accès à des lieux privés. Voici quelques exemples de ce type d'applications :

- **Le Verrouillage automatique:** ces applications sont utilisées comme une sorte de verrous électroniques comme par exemple la protection de domicile, garage, bâtiment, etc.
- **Validation des transactions sur site** (comme contrôle supplémentaire au niveau des distributeurs bancaires).
- **Accès aux lieux de production des usines :** qui sont en général réservés aux employés, ouvriers et inspecteurs afin de protéger le secret de la production et du matériel.

1.7.2 Applications dans les systèmes de communication

Ce type d'applications utilise par exemple le téléphone comme un moyen matériel de communication entre l'homme et la machine. C'est la catégorie la plus importante parce qu'elle permet de vérifier ou identifier le locuteur à longue distance. Il existe plusieurs applications dans cette catégorie et parmi elles :

- Validation de transactions bancaires par téléphone (pour améliorer le service bancaire, ainsi que pour valider légalement la transaction effectuée).
- Accès à des bases de données pour plus de sécurité et pour plus de protection (ex : consultation d'email, consultation de répondeur, etc.).
- Accès à des services téléphoniques (ex : téléphoner sur son compte de facturation personnelle de n'importe quelle ligne téléphonique).

1.7.3 Applications juridiques

Enfin on trouve le domaine d'applications qui pose actuellement le plus de problèmes, c'est le domaine juridique. La reconnaissance de locuteur est utilisée par exemple pour :

- L'orientation des enquêtes.
- La constitution des éléments de preuves au cours d'un procès.

Dans ces applications on trouve beaucoup plus d'inconvénients que d'avantages :

- La quantité de la parole à disposition est en général très limitée.
- Les conditions d'environnement sont très mauvaises.
- Les locuteurs impliqués sont très rarement coopératifs.

1.8 Conclusion

Dans ce chapitre, nous avons introduit le principe de la reconnaissance automatique du locuteur ainsi que les différentes étapes d'un système de RAL. Nous avons présenté également les techniques de paramétrisation les mieux adaptées pour représenter un locuteur ainsi que quelques approches de reconnaissance. L'approche basée sur les modèles de Markov cachés et celle basée sur les SVM sont utilisées dans notre travail. Elles seront présentées plus en détails dans le chapitre suivant.

Chapitre 2

**Approches de reconnaissance
mises en œuvre et Fusion**

2.1. Introduction

Plusieurs approches de reconnaissance du locuteur existent dans la littérature. Nous pouvons citer l'approche basée sur la programmation dynamique (DTW), l'approche basée sur la quantification vectorielle, l'approche neuronale, l'approche basée sur les modèles de Markov cachés (HMM), l'approche basée sur les Machines à Vecteurs Support (SVM), etc. Les deux dernières approches ont été utilisées dans notre travail. Dans ce qui suit, nous allons les détailler. Nous présenterons leurs principes et nous verrons quels sont leurs paramètres et comment ils sont obtenus.

2.2. Modèles de Markov Cachés

Les Modèles de Markov Cachés (en anglais : Hidden Markov Model ou HMM) ont été introduits par Baum et ses collaborateurs dans les années 1960-70 [Rabi-89]. Les HMMs ont été initialement introduits en reconnaissance de la parole. Puis leur utilisation s'est étendue peu à peu au domaine de la reconnaissance du locuteur.

Les HMM permettent de modéliser des processus stochastiques variant dans le temps. Le modèle d'un locuteur est constitué de l'association d'une chaîne de Markov, une succession d'états avec des probabilités de transition d'un état à un autre et de lois de probabilités (probabilités d'observation d'un vecteur acoustique dans un état).

2.2.1 Définition

Un modèle de Markov λ est un automate stochastique capable, après apprentissage, d'estimer la probabilité qu'une séquence d'observations ait été générée par ce modèle. Idéalement, il faudrait pouvoir associer à chaque locuteur un modèle constitué d'un nombre fini d'état prédéterminé.

Deux processus stochastiques concurrents sont observés :

- Un premier processus qui est la séquence d'observations produites $(O_t)_{1 \leq t \leq T}$, qui vérifie :

$$P(O_t = o_t / Q_t = q_i, \dots, Q_1 = q_1, O_{t-1} = o_{t-1}, \dots, Q_1 = o_1) = P(O_t = o_t / Q_t = q_i) \\ = b_i(o_t) \quad (2.1)$$

- Un second processus : la séquence d'états parcourus (Q_t) qui est une chaîne de Markov d'ordre 1 (la connaissance du passé se résume à celle du dernier état occupé), il doit vérifier:

$$P(Q_{t+1} = q_j / Q_t = q_j, \dots, Q_0 = q_0) = P(Q_{t+1} = q_j / Q_t = q_j) \\ = a_{ij} \quad \text{pour tout } t \geq 1 \quad (2.2)$$

Les modèles de Markov sont dits 'cachés' parce que la suite d'états parcourus Q_t pour générer la séquence O_t , n'est pas directement observable.

2.2.2 Les paramètres d'un Modèle de Markov Caché

A chaque locuteur est associé un HMM constitué d'un nombre fini d'états prédéterminés. Formellement, un modèle de Markov caché λ peut être défini par les paramètres suivants :

$$\lambda = (N, A, B, \pi)$$

1. un ensemble d'états cachés représentés par $S = \{s_1, s_2, \dots, s_N\}$ avec N est le nombre d'état du modèle.
2. La matrice de transition de probabilité, pour la chaîne de Markov : $A_{[N \times N]} = (a_{ij})$, a_{ij} représente la probabilité de transiter de l'état i vers l'état j .

$$a_{ij} = P(q_{t+1} = s_j | q_t = s_i), \quad 1 \leq i \leq N, \quad 1 \leq j \leq N, \quad 1 \leq t \leq T \quad (2.3)$$

$$\sum_{i=1}^N a_{ij} = 1 \quad 1 \leq j \leq N$$

3. $B = \{b_i(o_t)\}$ la matrice (ou la fonction de densité) de probabilité d'émission de l'observation o_t dans l'état s_i selon le type de modélisation d'observations :

$$b_i(o_t) = P(o_t / q_t = s_i) \quad (2.4)$$

- Dans le cas d'entrées discrètes (après quantification vectorielle des observations), les probabilités d'émission peuvent être décrites comme des fonctions de densité de probabilité discrète.
- Dans le cas d'entrées continues $b_i(o_t)$ est :
- Supposée être de la forme d'une distribution multivariable gaussienne, entièrement définie par le vecteur moyenne et la matrice de covariance ou d'une distribution de type multigaussienne (somme pondérée de gaussienne).

4. Un ensemble de densités de probabilités initiales $\pi = [\pi_i]$ où π_i représente la probabilité initiale du modèle à l'état s_i , soit :

$$\pi_i = P(q_1 = s_i) , \quad 1 \leq i \leq N \quad (2.5)$$

2.2.3 Principaux types de HMM

Le type d'un HMM peut être spécifié selon :

- Les contraintes liées à la matrice de transition A .
- Le type de densité de probabilité d'observations $b_j(O_t)$.

2.2.3.1 Contraintes liées à la matrice de transition A

Deux types de modèles peuvent être obtenus selon les contraintes imposées sur les éléments de la matrice A :

a. Modèle ergodique

Il est sans contrainte, toutes les transitions d'un état vers un autre sont possibles, c'est-à-dire, que tous les états peuvent être atteints de n'importe où en un seul pas. On peut remarquer que la connaissance de π est primordiale pour le choix de l'état de départ (figure 2.1).

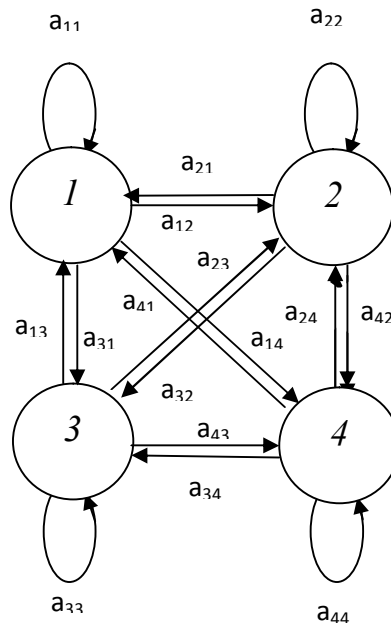


Figure 2.1: Exemple de modèle ergodique à 4 états.

b. Modèle gauche-droite

Il n'autorise les transitions que d'un état vers lui-même et vers un état suivant. Ce type de modèle peuvent être utilisés pour modéliser des processus possédant des propriétés variantes dans le temps, tels que les signaux de la parole et d'écriture manuscrite [Augu-01]. Ils traduisent la causalité du processus de production de ces signaux. Leurs probabilités de transition vérifient: $i > j \Rightarrow a_{ij} = 0$ (figure 2.2).

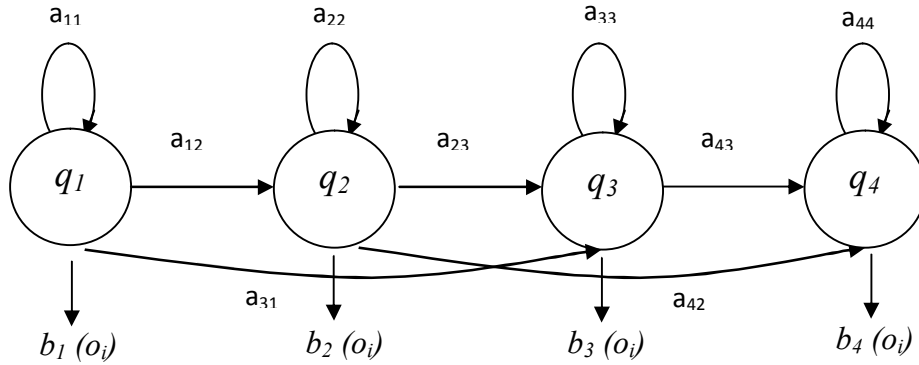


Figure 2.2: Exemple de modèle gauche-droite à 4 états.

2.2.3.2 Type de densité de probabilité d'observations $b_j(O_t)$

Selon le type de densité de probabilité d'observations générer par les symboles, discrète ou continue, nous pouvons construire deux types de modèles HMM : soit un HMM discret soit un HMM continu.

a. HMM discret "Discret Hidden Markov Models (DHMM)"

L'espace de représentation du signal acoustique représentatif du locuteur est généralement un espace multidimensionnel continu E et les vecteurs de coefficients calculés sur une trame de signal sont des points de cet espace. Il est possible de modéliser ces observations par des modèles de Markov à émissions discrètes au moyen de la quantification vectorielle (QV).

La QV permet le passage de l'espace E vers un espace discret, en partitionnant cet espace et en choisissant un représentant pour chaque classe. L'ensemble des représentants constitue un dictionnaire de M prototypes noté $V = \{v_k\}_{1 \leq k \leq M}$. Chaque vecteur $O \in E$ est quantifié par le prototype du dictionnaire dont il est le plus proche au sens d'une distance $d(O, \hat{O})$ définie dans l'espace E :

$$O \xrightarrow{QV} \hat{O} = v_{r'}, \quad \text{avec} \quad r' = \arg \min_{1 \leq r \leq M} d(O, v_r) \quad (2.6)$$

Divers algorithmes ont été proposés pour la réalisation du dictionnaire des prototypes, basés le plus souvent sur une classification hiérarchique descendante ou sur les nuées dynamiques [Gray-84]. Cependant, la QV introduit des distorsions, et il est souvent préférable de travailler directement dans l'espace continu E .

b. HMM continu "Continuous Hidden Markov Models (CHMM)"

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

$$b_j(O) = \sum_{k=1}^M c_{jk} N(O, \mu_{jk}, \Sigma_{jk}) = \sum_{k=1}^M c_{jk} b_{jk}(O) \quad 1 \leq j \leq N \quad (2.7)$$

M est le nombre de gaussiennes à l'état j . $N(O, \mu_{jk}, \Sigma_{jk})$ où $b_{jk}(O)$ est la distribution gaussienne du $k^{\text{ème}}$ mélange de l'état j . Cette distribution est définie par le vecteur moyen μ_{jk} et la matrice de covariance Σ_{jk} à l'état j . c_{jk} est le coefficient de pondération du $k^{\text{ème}}$ mélange qui satisfait la contrainte stochastique :

$$\sum_{k=1}^M c_{jk} = 1 \quad c_{jk} \geq 0, \quad 1 \leq j \leq N \quad (2.8)$$

Nous rappelons que la densité de probabilité d'une loi normale de moyenne μ et de matrice de covariance Σ en dimension d est:

$$N(O, \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (O - \mu)^t \Sigma^{-1} (O - \mu) \right\} \quad (2.9)$$

L'hypothèse qui est souvent faite d'une indépendance entre les d dimensions de l'espace autorise l'utilisation de matrices de covariance diagonales, cela limite le nombre de paramètres à estimer et simplifie les calculs. D'autres types de densités de probabilité sont

possibles, comme la sortie d'un réseau de neurones dans le cas de systèmes de reconnaissance hybride [Bour-90].

2.2.4 Topologie des modèles HMMs utilisés pour la reconnaissance du locuteur

Les HMMs supposent que la séquence de vecteurs acoustiques représentatifs du locuteur est une succession de segments stationnaires. Ainsi le locuteur est modélisé par une succession d'états, avec des transitions instantanées possibles entre ces états. Chaque observation est supposée être une fonction probabiliste de l'état.

La topologie principalement employée dans la littérature est le modèle gauche-droit d'ordre 1 qui ne permet pas de "retour en arrière" et qui démarre toujours depuis l'état initial ($i=1$).

C'est-à-dire que leurs probabilités vérifient :

$$i > j \Rightarrow a_{ij} = 0 \quad 2 \leq i \leq N, \quad 1 \leq j \leq N-1 \quad (2.10)$$

$$\pi_i = P(q_1 = i) = \begin{cases} 1 & \text{pour } i = 1 \\ 0 & \text{pour } 1 < i \leq N \end{cases} \quad (2.11)$$

Dans ce cadre, R. Bakis a proposé un modèle type pour représenter un locuteur qui permet le bouclage sur l'état courant (progression acoustique stationnaire) et le passage à l'état suivant (progression acoustique "standard") (figure 2.3). Le nombre d'états du modèle est normalement proportionnel à la durée moyenne du signal acoustique représentatif du locuteur. La plupart des systèmes de reconnaissance utilisent des modèles à trois états. Le but initial étant d'avoir l'état central modélisant la partie stable de ce signal alors que les deux états extrêmes modélisent la partie transitoire.

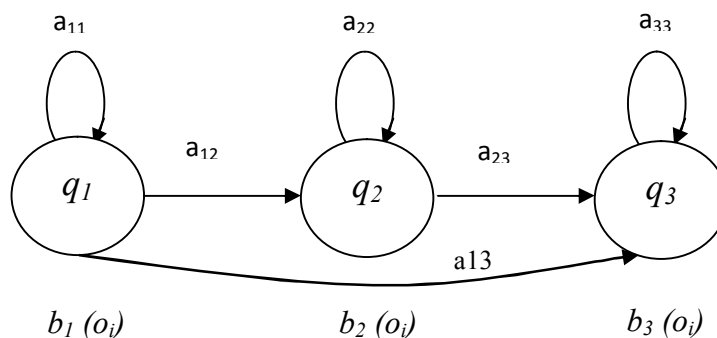


Figure 2.3: Exemple de modèle de Bakis à trois états.

2.2.5 Trois problèmes fondamentaux d'un HMM

L'utilisation des HMM donne lieu à trois problèmes fondamentaux :

➤ **Problème 1 : Evaluation de la probabilité d'émission d'une séquence**

Etant donné une suite d'observations O et un modèle λ . Comment peut-on calculer efficacement $P(O/\lambda)$, la probabilité d'observer la séquence sachant le modèle λ ?

➤ **Problème 2 : Apprentissage des modèles**

Comment ajuster le modèle λ pour maximiser la vraisemblance de la séquence d'observations $P(O/\lambda)$?

➤ **Problème 3 : Reconnaissance**

Etant donné une suite d'observations visibles O et un modèle λ . Comment peut-on trouver la séquence d'états cachés la plus probable pour produire cette séquence d'observables ?

La réponse à ces problèmes va nous permettre d'utiliser les HMM comme outil de modélisation des locuteurs.

2.2.6 Reconnaissance du Locuteur par HMMs

Dans le but de réaliser un système de RAL, on suppose que chaque locuteur est modélisé par une machine de Markov.

Soit λ un modèle de Markov caché et O une séquence d'observations acoustiques représentant un locuteur. La reconnaissance de cette séquence s'effectue en trouvant le modèle λ qui maximise la probabilité $P(\lambda/O)$ (probabilité qu'un modèle λ génère une séquence de vecteurs acoustiques O). Cette probabilité est aussi appelée probabilité à posteriori.

Malheureusement, il n'est pas possible d'accéder directement à cette probabilité. Mais on peut la décomposer en trois probabilités :

$$P(\lambda/O) = \frac{P(O/\lambda) \times P(\lambda)}{P(O)} \quad (2.12)$$

- $P(O/\lambda)$ est la vraisemblance de la séquence d'observations O étant donné le modèle λ .
- $P(\lambda)$ est la probabilité à priori du modèle. Cette probabilité peut être calculée indépendamment des observations.
- $P(O)$ est la probabilité à priori de la séquence des vecteurs acoustique. Cette probabilité peut être considérée constante, puisqu'elle est indépendante du modèle λ . Ainsi, maximiser $P(\lambda/O)$ revient à maximiser $P(O/\lambda) \times P(\lambda)$.

2.2.6.1 Evaluation de la probabilité d'émission d'une séquence

Soit le modèle $\lambda = (N, A, B, \pi)$, $O = \{o_1, o_2, \dots, o_T\}$ une séquence d'observations, $Q = \{q_1, q_2, \dots, q_T\}$ une séquence d'états. La probabilité que la suite d'observations O soit émise par le modèle λ peut être calculée comme la somme des probabilités conjointes de l'observation O et du chemin Q pour l'ensemble de tous les chemins possibles de longueur T :

$$P(O/\lambda) = \sum_Q P(O, Q/\lambda) \quad (2.13)$$

La probabilité d'observer la séquence O pour une séquence d'états Q est :

$$P(O/Q, \lambda) = b_{q_1}(O_1) \cdot b_{q_2}(O_2) \cdot \dots \cdot b_{q_T}(O_T). \quad (2.14)$$

Or, la probabilité de la séquence Q peut s'écrire sous la forme suivante :

$$P(Q/\lambda) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \dots a_{q_{T-1} q_T} \quad (2.15)$$

La probabilité conjointe du chemin Q et des observations O est :

$$P(O, Q/\lambda) = P(O/Q, \lambda) P(Q/\lambda) \quad (2.16)$$

La probabilité de la séquence d'observations O sachant le modèle λ est obtenue par la sommation de $P(O, Q/\lambda)$ sur toutes les séquences d'états Q possibles.

Ainsi la probabilité d'émission des observations est :

$$P(O/\lambda) = \sum_Q P(O, Q/\lambda) = \sum_Q P(O/Q, \lambda) P(Q/\lambda) = \sum_Q \prod_{t=1}^T a_{q_{t-1} q_t} b_{q_t}(o_t) \quad (2.17)$$

Pour un modèle HMM à N états, le calcul direct nécessite $(2T-1) \times N^T$ multiplications et $N^T - 1$ additions, ce qui le rend trop complexe et impossible à implémenter. Il existe une autre procédure pour le calcul de cette probabilité qui consomme moins d'opérations, c'est la méthode 'Forward-Backward' [Rabi-89]. Cet algorithme est détaillé dans l'annexe A.

2.2.6.2 Apprentissage des modèles

L'un des principaux problèmes de l'utilisation des HMMs réside dans la phase d'apprentissage. Il s'agit de déterminer les paramètres du modèle maximisant la probabilité de la suite d'observations O générée par le modèle.

Ces paramètres sont les probabilités de transition entre états et les probabilités d'émission associées aux états, car la topologie du modèle, à savoir le nombre d'états des modèles et les transitions autorisées entre ces états est supposée fixée à priori. Ainsi, connaissant une suite d'observations émises par un modèle, il est possible de modifier les paramètres du modèle de manière à rendre plus probable l'émission des observations par le modèle. Il s'agit d'une estimation par le critère du maximum de vraisemblance (Maximum Likelihood Estimation ou MLE).

$$\tilde{\lambda} = \arg \max_{\lambda} P(O / \lambda) \quad (2.18)$$

Une résolution analytique directe n'est pas possible, mais les formules Baum-Welch [Baum-72] [Rabi-89] permettent une ré-estimation itérative des paramètres du modèle. L'algorithme de Baum-Welch est décrit dans l'annexe A.

2.2.6.3 La reconnaissance

Après l'entraînement des paramètres des modèles HMM, la reconnaissance ou le décodage d'une séquence acoustique O correspondant à un locuteur consiste à déterminer le 'meilleur' chemin des états qui maximise la probabilité de cette séquence. Le problème de reconnaissance est généralement résolu par l'intermédiaire de l'algorithme de Viterbi.

➤ Algorithme de Viterbi

L'algorithme Viterbi [Vit-67] est un algorithme issu de la programmation dynamique permettant de déterminer la meilleure séquence d'états, $Q = \{q_1, q_2, \dots, q_T\}$, connaissant la séquence d'observations $O = (o_1 \dots o_T)$. Nous définissons alors $\delta_t(i)$, comme la probabilité maximale que les observations observées jusqu'à l'instant t aient été émises par le modèle λ en suivant un chemin qui arrive à l'état s_i :

$$\delta_t(i) = \max_{q_1, q_2, \dots, q_{t-1}} P(o_1, o_2, \dots, o_t, q_1, q_2, \dots, q_t = i \mid \lambda) \quad (2.19)$$

Par induction, nous calculons :

$$\delta_{t+1}(j) = [\max_i \delta_t(i) a_{ij}] b_j(o_{t+1}) \quad (2.20)$$

Tout en gardant la trace, lors du calcul, de la suite d'états qui donne le meilleur chemin amenant à l'état s_i à l'instant t dans un tableau ψ . Formellement, l'algorithme se déroule de la façon suivante :

Initialisation : (début de l'algorithme)

$$\begin{cases} \delta_1(i) = \pi_i b_j(o_1) & 1 \leq i \leq N \\ \psi_1(i) = 0 \end{cases} \quad (2.21)$$

Induction : (calcul intermédiaire)

Pour t allant de 2 à T ,

Pour j allant de 1 à N ,

$$\delta_t(j) = \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}] b_j(o_t) \quad (2.22)$$

On conserve la mémoire du meilleur état précédent par:

$$\psi_t(j) = \arg \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}] \quad (2.23)$$

Fin

Fin

Terminaison : (fin de l'algorithme) $P(O, Q \setminus \lambda) = \max_{1 \leq i \leq N} [\delta_T(i)] \quad (2.24)$

Construction du meilleur chemin $\tilde{Q} = (\tilde{q}_1, \tilde{q}_2, \dots, \tilde{q}_T)$ avec:

$$\tilde{q}_T = \arg \max_{1 \leq i \leq N} [\delta_T(i)] \quad (2.25)$$

Puis, pour t allant de $T-1$ à 1 (en back-traking):

$$\tilde{q}_t = \psi_{t+1}(\tilde{q}_{t+1}) \quad (2.26)$$

2.3 Théorie des Machines à Vecteurs de Support

Depuis quelques années, de nouvelles méthodes d'apprentissage se développent sur la base de la théorie de l'apprentissage statistique (Statistical Learning Theory) de Vapnik et Chervonenkis. L'une de ces méthodes, appelée machines à vecteurs de support (SVM de l'anglais Support Vector Machine) qui fut introduite par Vapnik en 1995, permet de réaliser des estimations en classification (à deux classes ou plus) ou en régression. Le grand succès de cette méthode dans beaucoup de domaines se justifie essentiellement par des solides bases théoriques qui la soutiennent [Call-03].

La première application des SVM en reconnaissance du locuteur, notamment pour l'identification du locuteur, été en 1996 [Sch-96]. Elles ont été utilisées aussi pour la vérification du locuteur [Kha-02], la détection des mots clés [Ben-03], la reconnaissance de langue [Cam-04], etc.

Les SVM présentent notamment deux avantages majeurs :

1. Les méthodes à vecteurs de support sont issues de la théorie de l'apprentissage statistique, qui définit un cadre théorique rigoureux;
2. Les SVM affichent pour des applications pratiques de très bonnes performances, et leur succès est dû à l'utilisation des noyaux pour transformer le SVM en algorithme non linéaire pouvant être appliqué sur des données variées. Les noyaux et les SVM ont alors permis d'utiliser l'apprentissage numérique dans des problèmes traitant des données complexes.

Les SVM sont une classe d'algorithmes basés sur le principe de minimisation du « risque structurel » décrit par la théorie de l'apprentissage statistique de Vapnik et Chervonenkis.

2.3.1 Minimisation du risque structurel

Lorsque nous utilisons des méthodes d'apprentissage, ce sont généralement deux jeux de données principaux : le jeu d'entraînement et le jeu de test. Le jeu d'entraînement représente la part des données d'origine utilisée pour calculer le modèle. Le jeu de test est l'autre partie, inconnue de l'algorithme d'apprentissage et est utilisée pour évaluer les performances de généralisation du modèle. La qualité de ce modèle est alors jugée à sa capacité à réduire l'erreur de test ou de « généralisation ». Cependant, comme le modèle n'est pas construit en

utilisant le jeu de test, l'erreur de généralisation ne peut pas être évaluée exactement car elle dépend de la distribution de probabilité des données :

$$R[f] = \int \frac{1}{2} Q(x) dP(x, y) \quad (2.27)$$

Q est la fonction d'erreur (erreur absolue dans le cas des SVM), x est le vecteur d'entrée, et $P(x, y)$ est la distribution des données (qui nous est inconnue).

La seule information dont nous disposons comme évaluation de l'erreur est l'erreur d'entraînement :

$$R_{emp}[f] = \frac{1}{l} \sum_{i=1}^l \frac{1}{2} Q(x_i) \quad (2.28)$$

l est le nombre de points d'entraînement.

Ce n'est pas suffisant. La raison en est que l'on peut facilement trouver un modèle minimisant l'erreur d'entraînement mais pour lequel l'erreur de généralisation sera très grande. Un exemple simple est la régression de données linéaires bruitées au moyen d'une fonction polynomiale : plus le degré du polynôme sera grand, plus l'erreur d'entraînement sera faible, mais plus l'erreur de généralisation sera élevée. Nous pouvons donc comprendre que cette dernière est aussi liée à la famille de fonction utilisée comme modèle. Cette dépendance est nommée « risque structurel ».

Dans leur Théorie de l'Apprentissage Statistique, Vapnik et Chervonenkis ont prouvé qu'il est possible de définir une majoration du risque structurel en fonction de la famille de fonction utilisée pour le modèle. L'une de ces majorations peut être calculée en utilisant la dimension de Vapnik-Chervonenkis (VC_{dim}) qui représente le plus grand nombre de points pouvant être séparés de toutes les façons possibles par un membre de l'ensemble de fonction. La borne VC est alors définie ainsi : si la dimension $VC(h)$ de la famille de fonction utilisée est inférieure au nombre de points d'entraînement l , alors avec une probabilité d'au moins $1-\eta$, on a :

$$R[f] \leq R_{emp}[f] + \sqrt{\frac{h \left(\log\left(\frac{2l}{h}\right) + 1 \right) - \log\left(\frac{\eta}{4}\right)}{l}} \quad (2.29)$$

2.3.2 Classification binaire par hyperplan

La méthode des SVM est destinée à trouver les hyperplans qui séparent le mieux entre les classes. Il s'agit de trouver l'hyperplan « optimal » séparant entre les classes.

Le SVM cherche le signe de la fonction $f(x) = \text{sgn}(w \cdot x + b)$ pour prendre une décision à propos de la classification d'un locuteur.

Les vecteurs d'apprentissage ne sont autres que les vecteurs de caractéristiques x_i de dimension N avec les étiquettes des classes correspondantes y_i .

2.3.2.1 SVM Linéaire

Dans un premier temps nous allons nous limiter au cas où les SVM sont linéaires, c'est-à-dire que nous sommes dans le cas où les données sont linéairement séparables. L'ensemble d'apprentissage est $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$; où $x_i \in \mathbb{R}^N$ et $y_i \in \{-1, +1\}$, le but est de trouver un hyperplan de séparation pour classer un nouveau locuteur.

La fonction de décision à déterminer est :

$$f(x) = \text{sgn}(w \cdot x + b) \quad (2.30)$$

La majorité des problèmes de classification comportent plus de deux classes. Toutefois, un problème multi-classes peut être réduit à un ensemble de deux classes (binaire) suivant une certaine stratégie de subdivision.

Pour le cas binaire, où les données sont linéairement séparable on cherche à maximiser la marge qui sépare les deux classes.

L'Hyperplan Optimal (HO) possède comme fonction $y = \text{sgn}(w \cdot x + b) = 0$, la fonction de décision est :

$$f(x) = \text{sgn}(w \cdot x + b) \cdot \begin{cases} f(x) > 0 & \text{décision à propos de } C_1 \\ f(x) < 0 & \text{décision à propos de } C_2 \end{cases} \quad (2.31)$$

Pour décider quel hyperplan il faut prendre, nous définissons deux hyperplans H_1 et H_2 qui sont parallèle à H_0 et ayant la même distance par rapport à ce dernier.

$$\begin{aligned} H_1 : y &= w \cdot x + b = +1 \\ H_2 : y &= w \cdot x + b = -1 \end{aligned} \quad (2.32)$$

Avec la condition qu'il n'y ait aucun point à l'intérieur de la marge qui appartient à l'une des deux classes délimité par les deux hyperplans (figure 2.4).

➤ Maximisation de la marge

Le problème à résoudre donc est de trouver les deux hyperplans qui maximisent la marge sans erreur, et qui séparent bien les deux classes l'une de l'autre.

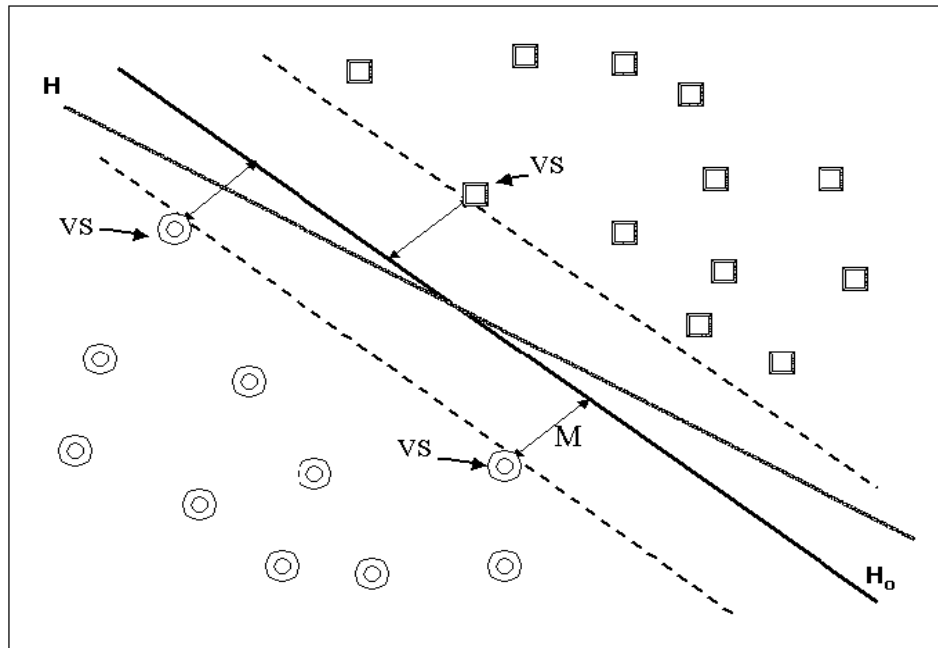


Figure 2.4: Hyperplans séparateurs.

H est un hyperplan quelconque, H_o est l'hyperplan optimal et M est la marge qui représente la distance entre les différentes classes et H_o (VS sont les Vecteurs Supports).

Pour assurer une séparation optimale des deux classes l'une par rapport à l'autre, il nous faut donc une maximisation de la marge séparant les deux hyperplans H_1 et H_2 . L'algorithme SVM conduit au choix de l'hyperplan séparateur dont la marge est maximal.

Les vecteurs d'apprentissage situés sur les deux hyperplans séparateurs sont appelés 'les Vecteurs de Support'. L'hyperplan optimal est défini par la somme pondérée de ces vecteurs.

➤ Calcul de la marge

La distance entre un point situé sur H_1 et l'hyperplan H_0 est donnée par :

$$\frac{|w \cdot x + b|}{\|w\|} = \frac{1}{\|w\|} \quad (2.33)$$

Cela donne nécessairement une distance entre H_1 et H_2 égale à $\frac{2}{\|w\|}$. Les vecteurs w et b sont les deux vecteurs qui définissent l'hyperplan séparateur.

Maximiser la quantité $\frac{2}{\|w\|}$ revient à minimiser la quantité inverse $\frac{\|w\|^2}{2}$, avec la condition qu'il n'y ait aucun point entre H_1 et H_2 .

$$\begin{cases} w \cdot x + b \geq +1 & , \quad y_i = +1 \\ w \cdot x + b \leq -1 & , \quad y_i = -1 \end{cases} \quad (2.34)$$

Ces deux conditions peuvent être combinées dans une seule condition donnée par :

$$y_i \cdot (w \cdot x_i + b) \geq 1 \quad (2.35)$$

Le problème peut être reformulé de la manière suivante :

$$\begin{cases} \min \left(\frac{1}{2} \|w\|^2 \right) \\ \text{sous les contraintes : } y_i \cdot (w \cdot x_i + b) \geq 1 \end{cases} \quad (2.36)$$

Il s'agit d'un problème d'optimisation quadratique sous contraintes linéaires dont la fonction objective (2.36) est à minimiser. Cette fonction objective est le carré de l'inverse du double de la marge. L'unique contrainte stipule que les données doivent être bien classées et qu'ils ne dépassent pas les hyperplans, c'est-à-dire aucune donnée d'apprentissage dans la marge n'est permise.

Dans cette formulation, les variables à fixer sont les composantes w_i et b . Ceci implique le réglage de $n+1$ paramètres (n étant la dimension de l'espace d'entrée plus la dimension du scalaire b). En gardant cette formulation telle quelle, nous tombons dans le même problème critique des méthodes classiques d'apprentissage. Pour éviter cela, il est nécessaire d'introduire une formulation dite duale du problème. Un problème dual est un problème fournissant la même solution que le primal (initial) mais dont la formulation est différente.

Nous appellerons ‘variables primales’, les variables du problème primal, et ‘variables duales’, les variables du problème dual qui n’interviennent pas dans le primal. Nous utiliserons souvent les termes primal et dual pour se référer aux problèmes.

Pour résoudre ce problème, nous devons former ce que nous appelons le Lagrangien. Il s’agit de faire rentrer les contraintes dans la fonction objective et de pondérer chacune d’entre elles par une variable duale α :

$$L(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^N \alpha_i y_i (w \cdot x_i + b) + \sum_{i=1}^N \alpha_i \quad (2.37)$$

Les variables duales α_i intervenant dans le Lagrangien sont appelées multiplicateurs de Lagrange, une pour chaque point de données. Selon le type de contraintes qu’ils représentent, et sachant que L doit être minimisé par rapport aux variables primaires w et b (maximisation de la marge) et maximisé par rapport aux variables duales α_i (pour s’assurer qu’un maximum de contraintes soient satisfaites), les multiplicateurs doivent respecter la condition suivante :

$$\alpha_i \geq 0 \quad \text{Avec :} \quad 1 \leq i \leq N \quad (2.38)$$

A présent, nous introduisons les deux conditions de Karush Kuhn et Tucker (KKT), statuant sur l’optimalité d’une solution, elles peuvent être obtenues en exploitant directement les propriétés du point de selle de lagrangien. Au point de selle, la dérivée du lagrangien par rapport aux variables primaires doit s’annuler, ceci s’écrit :

$$\frac{\partial}{\partial w} L(w, b, \alpha) = 0 \quad \text{Et,} \quad \frac{\partial}{\partial b} L(w, b, \alpha) = 0 \quad (2.39)$$

$$\left\{ \begin{array}{l} \frac{\partial L}{\partial w} = 0 \\ \frac{\partial L}{\partial b} = 0 \end{array} \right\} \Rightarrow \left\{ \begin{array}{l} w = \sum_{i=1}^N \alpha_i y_i x_i \\ \sum_{i=1}^N \alpha_i y_i = 0 \end{array} \right. \quad (2.40)$$

Nous remplaçons (2.40) dans le Lagrangien :

$$L(w, b, \alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j \quad (2.41)$$

A partir de quoi nous pouvons formuler le problème dual :

$$\text{Maximiser : } L = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j \quad (2.42)$$

$$\text{Tel que : } \begin{cases} \sum_{i=1}^N \alpha_i y_i = 0 \\ \alpha_i \geq 0 \quad \forall i = 1 \dots N \end{cases}$$

C'est une forme d'optimisation quadratique de dimension N bien étudiée, pour laquelle il existe de bons algorithmes d'optimisation. La résolution du dual permet donc de calculer la fonction de l'hyperplan. Nous avons à présent tous les éléments nécessaires pour exprimer la fonction décision de notre classifieur :

$$f(x) = \text{sgn}(w.x + b) = \text{sgn}\left(\sum_{i=1}^N \alpha_i y_i [x_i x] + b\right) \quad (2.43)$$

Le calcul de la fonction de décision requiert la résolution d'un gros problème d'optimisation quadratique. Il existe un vaste choix d'algorithmes performants pour le résoudre. Un algorithme spécifique, l'Optimisation Séquentielle Minimale (SMO), a été utilisé afin de pouvoir manipuler des ensembles de très grande dimension. Il optimise la fonction objective duale du problème global en opérant à chaque itération sur un ensemble réduit à deux multiplicateurs de Lagrange.

2.3.2.2 SVM non linéaires

Le classifieur à marge maximale que nous venons de présenter, permet d'obtenir de très bons résultats lorsque les données sont linéairement séparables. L'intérêt principal d'un classifieur de ce type réside dans le fait que l'on contrôle facilement le pouvoir de généralisation. Naturellement, un grand nombre de jeux de données est non-linéairement séparable. Pour classer ce genre de données, nous pourrions utiliser une fonction de décision non linéaire. Géométriquement, cela reviendrait à avoir une (hyper)courbe qui marquerait la frontière entre les données positives et négatives. Le danger de ce genre de méthode vient du fait que le nombre de paramètres à calculer est très élevé et provoque souvent du sur-apprentissage.

Boser et al., ont su produire des frontières de décision non-linéaire avec les SVM. Dans ce cas, le problème est transposé dans un espace de dimension plus élevé pour rendre les nuages de points des deux classes linéairement séparables.

Il s'agit d'introduire une transformation mathématique pour ne pas sortir du domaine des données disponibles, le but est toujours de chercher à maximiser la marge de séparation (figure 2.5). Pour effectuer cela, le problème est transposé dans un espace d'ordre plus élevé. (De l'espace $\mathcal{R}^N$ à $\mathcal{R}^M$, $M > N$) selon la transformation ϕ . Le problème dual d'optimisation devient:

$$W(\alpha) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \phi(x_i) \phi(x_j) \quad \text{Avec :} \quad \phi(x_i) \phi(x_j) = K(x_i, x_j) \quad (2.44)$$

$$\text{Tel que :} \quad \begin{cases} \sum_{i=1}^N \alpha_i y_i = 0 \\ \alpha_i \geq 0 \quad \forall i = 1 \dots N \end{cases}$$

La quantité $\phi(x_i) \phi(x_j) = K(x_i, x_j)$ définit un produit scalaire dans le nouvel espace. Il est équivalent à une 'fonction noyau' de l'espace d'entrée. Tous les noyaux doivent satisfaire les conditions de Mercer suivantes :

Quelque soit la fonction $g(x)$ telle que : $\int g(x)^2 dx < \infty$, le noyau K doit satisfaire l'inégalité : $\int K(x, y) g(x) g(y) dx dy \geq 0$ et l'équation de l'hyperplan séparateur dans le nouvel espace devient:

$$f(x) = \sum_i \alpha_i y_i K(x_i, x) + b \quad (2.45)$$

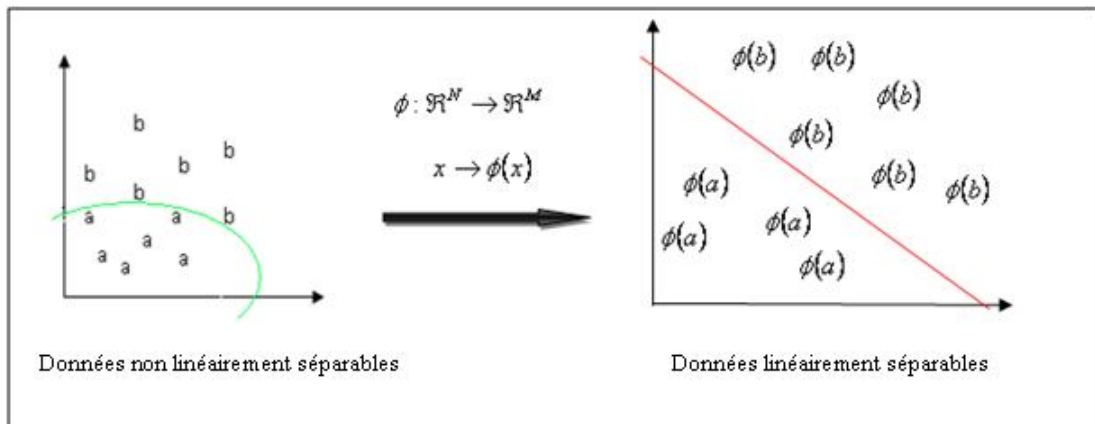


Figure 2.5: Transposition du problème vers un espace de dimension élevé.

a) Noyaux d'usage courant

Il est généralement difficile de vérifier analytiquement la condition de Mercer. Néanmoins, un certain nombre de noyaux sont connus comme étant définis positifs et largement exploités par la communauté :

- **Linéaire** : $K(x, y) = x^T y$

Celui-ci correspond au produit scalaire sans transformation. Il traduit donc la forme traditionnelle des algorithmes, sans l'usage du kernel trick.

- **Polynômial homogène** : $K(x, y) = (x^T y)^\delta$

Celui-ci permet indirectement d'appliquer le principe de maximisation de la marge aux classifieurs polynômiaux.

- **Polynômial inhomogène** : $K(x, y) = \left(1 + \frac{c}{d} x^T y\right)^\delta$

L'ajout d'une constante au produit scalaire permet d'inclure dans la transformation ϕ tous les monômes d'ordre inférieur ou égal à δ . Le noyau polynômial inhomogène implique donc un espace transformé de dimension supérieure au noyau homogène.

- **Gaussien RBF** : $K(x, y) = \exp\left(-\frac{\|x - y\|^2}{d\sigma^2}\right)$

Les fonctions à base radiale (RBF Radial Basis Functions) sont définies par le fait qu'elles ne dépendent que de la distance entre leurs arguments.

Le noyau Gaussien RBF applique ainsi une gaussienne sur la distance entre les exemples.

- **Sigmoïdal** : $K(x, y) = \tanh\left(\frac{c}{d} x^T y + \theta\right)$ La fonction de décision construite avec un

noyau sigmoïdal est égale à celle d'un réseau de neurones à deux couches. Le noyau sigmoïdal ne respecte pas la condition de Mercer pour toutes les valeurs de c et θ . Il demeure cependant couramment utilisé et reste généralement exploitable, malgré sa possible inadéquation théorique.

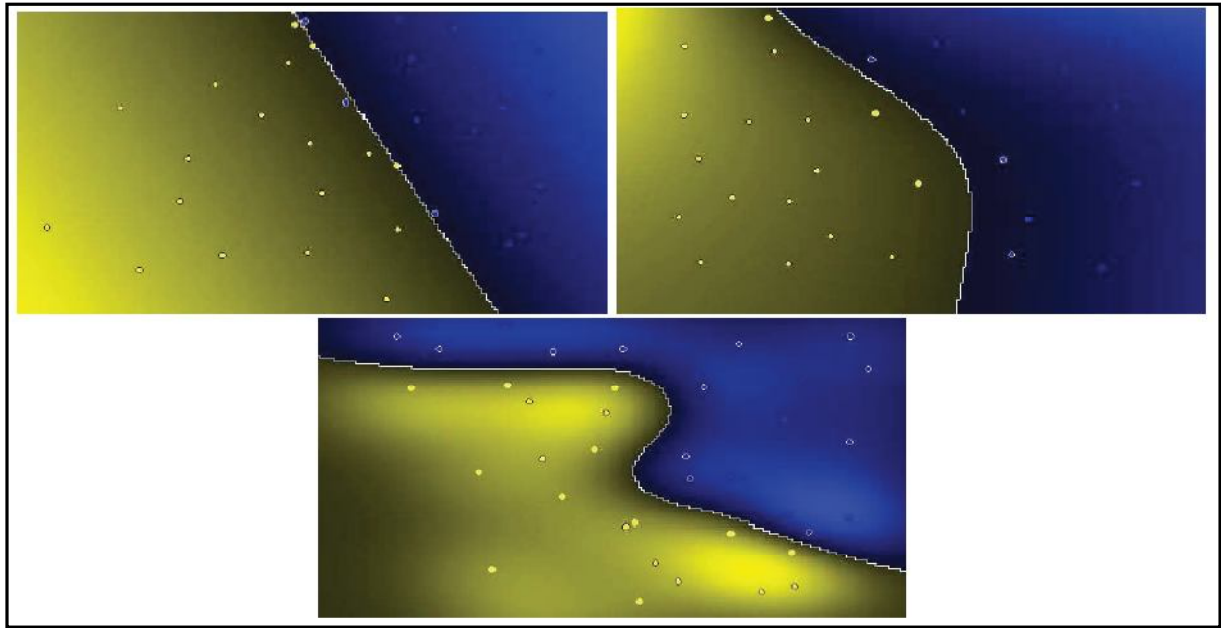


Figure 2.6: Frontières de décision obtenues par trois SVM à noyau linéaire, polynomial et RBF.

Les noyaux les plus utilisés sont les noyaux linéaires et RBF. Le choix du noyau RBF est souvent recommandé. En effet, il correspond à un meilleur “éclatement” des données; alors que le noyau linéaire ne permet pas d’augmenter la dimension de l’espace des descripteurs. Il permet aussi de traiter des cas non-linéaires. Sathiya et ses co-auteurs ont montré dans [Sath-03] que la classification linéaire était une simple restriction de la classification RBF. Une autre raison est le faible nombre de paramètres à régler pour le RBF, qui influe dans la complexité du modèle, un atout que n’ont pas les autres noyaux.

Enfin, le RBF, de module ≤ 1 , présente moins de problèmes numériques que les noyaux linéaires et polynomiaux et donc l’avantage principal de ce type de noyau est la rapidité de la phase d’apprentissage.

b) Le noyau RBF (Radial Basis Function)

La forme générique de ce noyau est :

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \quad (2.46)$$

Cette fonction noyau est une exponentielle du carré de la distance euclidienne entre les vecteurs, divisé par sigma au carré. C'est la formule d'une gaussienne dans l'espace des caractéristiques, où sigma est la l'écart type de la gaussienne. C'est un paramètre de l'apprentissage qui détermine les propriétés de la solution.

La transformation ϕ induite par ce type de noyau est un peu spéciale. En fait, une donnée va être transformée en une fonction gaussienne représentant la similarité de la donnée avec toutes les données de X. La figure 2.7 représente l'application du noyau RBF à deux données x et x'.

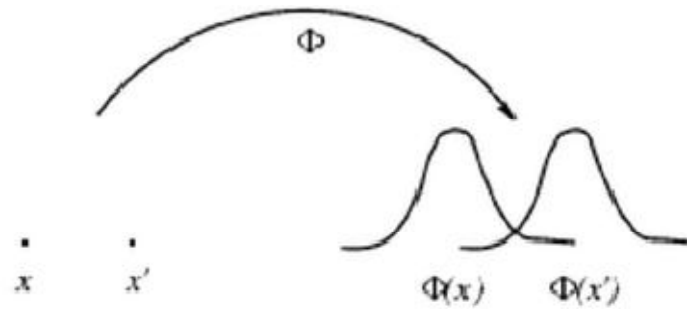


Figure 2.7: Application du noyau RBF à deux données x et x'.

L'espace d'arrivée F (connu comme espace de Hilbert) de la fonction ϕ est de dimension infinie étant donné que ϕ fait correspondre une fonction continue à chaque donnée, le noyau RBF permet donc de calculer des similarités dans un espace de dimension infinie !

Avec le noyau RBF, toutes les données sont placées sur une hyper-sphère de rayon 1 :

$$k(x, x) = \exp\left(-\frac{\|x - x\|^2}{2\sigma^2}\right) = 1 \quad (2.47)$$

Le paramètre σ permet de régler la largeur de la gaussienne. En prenant un σ grand, la similarité d'une donnée par rapport à ceux qui l'entoure sera assez élevée, alors qu'en prenant un σ tendant vers 0, la donnée ne sera similaire à aucune autre (voir figure 2.8). En resserrant fortement la gaussienne, un classifieur (faisant usage de ce noyau) peut arriver à apprendre n'importe quel ensemble d'entraînement sans commettre d'erreur. En fait, σ est un paramètre permettant de contrôler le pouvoir de généralisation d'un classificateur. La valeur du paramètre noyau affecte directement la complexité de la frontière de décision du classifieur.

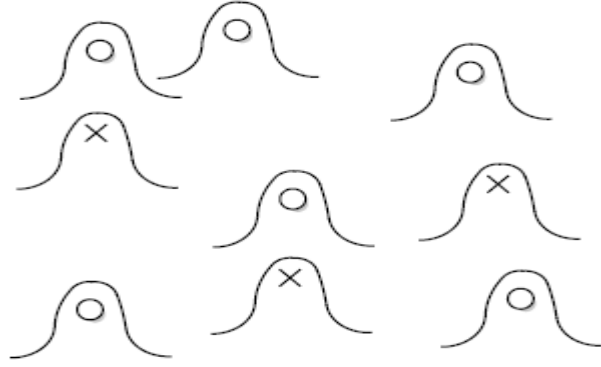


Figure 2.8: La similarité des données vue en utilisant un noyau RBF à σ tendant vers 0.

Pour connaître le domaine de variation de σ (avec : $\sigma \in [0, \sigma_{\max}]$), nous utilisons la distance maximale $d_{\max}$ euclidienne qui sépare deux vecteurs $\phi(x_i)$ dans l'espace transformé de rayon unité, qui est égale à :

$$d_{\max} = \|\phi(x_i) - \phi(x_j)\|_{\max}^2 = 2\sigma_{\max}^2 \quad (2.48)$$

$$\text{Sachant que : } \|\phi(x_i) - \phi(x_j)\|_{\max} = \sqrt{2} \quad \text{Alors : } \sigma_{\max} = \sqrt{2} = 1.4$$

2.3.2.3 Séparation imparfaite (Cas de données inséparables)

La plupart du temps, les données d'entraînement comportent du bruit. Par conséquent, même s'il existe une relation linéaire entre les données et leur catégorie, un classifieur linéaire pourrait commettre des erreurs et la formulation précédente n'est plus valable. Nous devons donc introduire au classifieur la possibilité de tenir compte des erreurs et les pénaliser en conséquence.

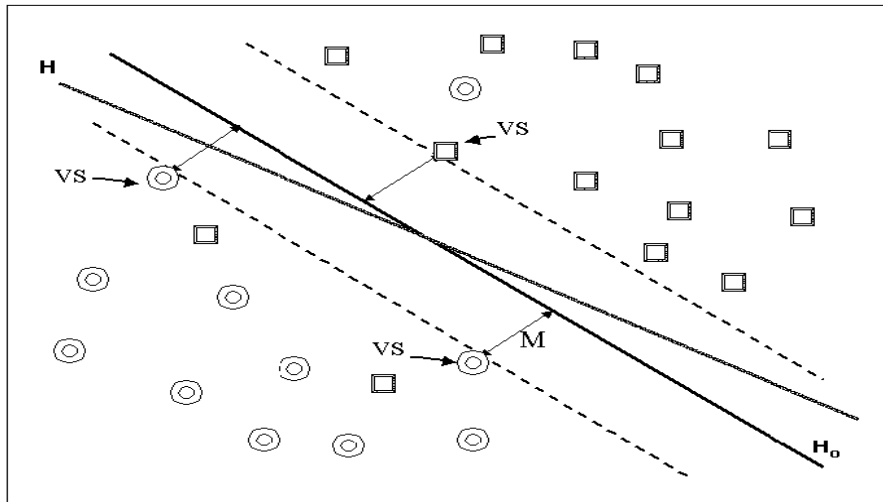


Figure 2.9: Hyperplans séparateurs dans le cas de données non-linéairement séparables (VS sont les Vecteurs Supports).

Pour effectuer cela, des ‘variables de pénalité’ (ou variables d’écart) ζ_i sont introduites aux inégalités pour les décontracter de manière à imposer que toutes les données soient bien classées.

$$\begin{cases} y_i(w.x_i + b) \geq 1 - \zeta_i & \forall i = 1 \dots N \\ \zeta_i \geq 0 & \forall i = 1 \dots N \end{cases} \quad (2.49)$$

L’objectif est de minimiser la fonction suivante :

$$\frac{1}{2} K(w, w) + C \sum_{i=1}^N \zeta_i \quad (2.50)$$

Où C est un paramètre ajustable par l’utilisateur et q est une fonction coût qui pénalise les erreurs commises. Le second terme dans l’équation (2.70) correspond au risque empirique.

➤ Choix de la fonction coût

Le choix de la fonction coût associée aux variables d’écart est important et affecte la solution du SVM. Il y a deux types de fonctions coût : linéaire et quadratique.

a) Fonction coût linéaire

La fonction coût linéaire $q(\zeta) = \zeta$ est la plus utilisée dans les SVM. Le premier objectif est de minimiser la quantité suivante :

$$L = \frac{1}{2} \|w\|^2 + C \sum_i \zeta_i \quad (2.51)$$

Avec les contraintes suivantes :

$$\begin{cases} x_i w + b \geq 1 - \zeta_i & \text{pour } y_i = 1 \\ x_i w + b \leq 1 - \zeta_i & \text{pour } y_i = -1 \\ \zeta_i \geq 0 & \forall i \end{cases} \quad (2.52)$$

De la même manière que le cas séparable, nous introduisons les multiplicateurs de Lagrange et le Lagrangien.

L'expression duale à maximiser s'écrit :

$$L = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j' \quad (2.53)$$

En ajustant les multiplicateurs de Lagrange soumis aux conditions :

$$\begin{aligned} 0 &\leq \alpha_i \leq C \\ \sum_i \alpha_i y_i &= 0 \end{aligned} \quad (2.54)$$

La solution w est donnée par :

$$w = \sum_i \alpha_i y_i x_i \quad (2.55)$$

Nous constatons que l'expression de l'optimisation duale avec une fonction coût linéaire est la même que l'expression dans le cas séparable sauf qu'il y a une borne supérieure imposée aux multiplicateurs de Lagrange.

b) Fonction coût quadratique

La fonction coût quadratique peut aussi être utilisée dans le SVM de sorte que l'étape d'optimisation soit un problème d'optimisation quadratique. En appliquant la fonction coût quadratique sur les variables, le problème d'optimisation est de minimiser la quantité :

$$L = \frac{1}{2} \|w\|^2 + C \sum_i \zeta_i^2 \quad (2.56)$$

Sous les contraintes :

$$\left\{ \begin{array}{ll} \text{Min } \frac{1}{2} \|w\|^2 + C \sum_i \zeta_i & \\ x_i \cdot w + b \geq 1 - \zeta_i & \text{pour } y_i = 1 \\ x_i \cdot w + b \leq 1 - \zeta_i & \text{pour } y_i = -1 \\ \zeta_i \geq 0 & \forall i \end{array} \right. \quad (2.57)$$

L'expression duale équivalente à maximiser est :

$$L = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j (x_i x_j) + \frac{1}{C} \sum_i \alpha_i^2 \quad (2.58)$$

En ajustant des coefficients α soumis aux conditions :

$$\begin{aligned} 0 &\leq \alpha_i \leq C \\ \sum_i \alpha_i y_i &= 0 \end{aligned} \quad (2.59)$$

Encore une fois, la solution est :

$$w = \sum_i \alpha_i y_i x_i \quad (2.60)$$

Lorsque les données d'apprentissage sont complètement séparables et la solution du SVM est peu dense (i.e. que le nombre de vecteurs de support est faible), la différence entre la fonction coût linéaire et quadratique est négligeable parce que la plupart des ζ_i seront nuls.

C'est seulement lorsque les données ne sont pas linéairement séparables que la fonction coût affecte la solution.

2.4 SVM à plusieurs classes

Les SVM multi-classes sont des modèles d'apprentissage de conception relativement récente, dont l'étude est actuellement en plein essor. Ceci résulte en premier lieu du fait que la communauté des théoriciens, qui avait jusque dans un passé récent consacré l'essentiel de ses forces au développement de la théorie statistique du calcul des dichotomies, exprime à présent un intérêt de plus en plus marqué pour le cas multi-classes, dont elle perçoit mieux les spécificités. Cette situation nouvelle fait naître un besoin, celui de disposer d'une étude synthétique sur les SVM multi-classes, ou plus généralement l'utilisation de SVM pour la discrimination à catégories multiples.

2.4.1 Approches "un contre tous"

L'approche "un contre tous" est la plus simple et la plus ancienne des méthodes de décomposition. Elle consiste à utiliser un classifieur binaire (à valeurs réelles) par catégorie. Le k -ième classifieur est destiné à distinguer la catégorie d'indice k de toutes les autres. Pour affecter un exemple, nous le présentons donc à Q classifieurs, et la décision s'obtient en application du principe "winner-takes-all" : l'étiquette retenue est celle associée au classifieur ayant renvoyé la valeur la plus élevée. Les auteurs soutiennent la thèse selon laquelle cette approche, aussi simple soit-elle, lorsqu'elle est mise en œuvre avec des SVM correctement paramétrées, obtient des performances qui ne sont pas significativement inférieures à celles des autres méthodes de décomposition et des SVM multi-classes actuelles.

Il convient cependant de souligner qu'elle implique d'effectuer des apprentissages aux répartitions entre catégories très déséquilibrées, ce qui soulève souvent des difficultés pratiques.

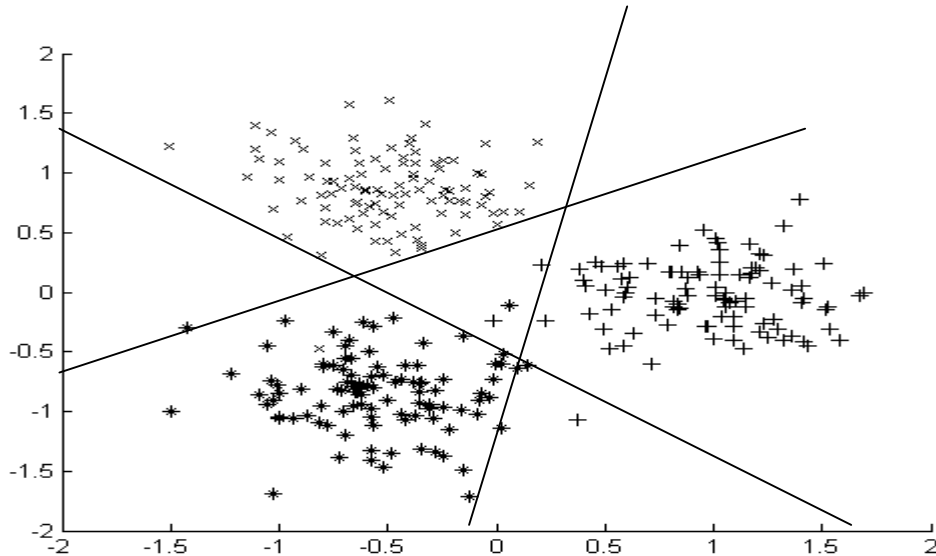


Figure 2.10: Problème à trois classes : frontières de décision linéaires dans la stratégie Un-contre-Tous

2.4.2 Approches "un contre un"

Une autre approche appelée "un contre un", ordinairement attribuée à Knerr et ses co-auteurs, consiste à utiliser un classifieur par couple de catégories. Le classifieur indicé par le couple (k, l) (avec $1 < k < l < Q$), est destiné à distinguer la catégorie d'indice k de celle d'indice l . Pour affecter un exemple, nous le présentons donc à C_Q^2 classifieurs, et la décision s'obtient habituellement en effectuant un vote majoritaire ("max-wins voting").

Sous l'hypothèse que la frontière séparant une catégorie d'une autre peut être moins complexe que celle séparant cette même catégorie de toutes les autres, il y voit un moyen d'obtenir des estimateurs présentant un biais plus faible qu'avec l'approche un contre tous. Naturellement, le prix à payer est un possible accroissement de la variance de ces estimateurs, compte tenu du fait que les bases d'apprentissage de chacun des classifieurs sont plus petites que l'échantillon initial.

2.5 Stratégies de fusion de données

La fusion des paramètres peut prendre différentes formes selon le niveau où elle est effectuée. En s'inspirant des méthodes de fusion appliquées entre autre dans le domaine de la

reconnaissance audio-visuelle [Delé-96] [Rogo-99] [Baki-08], nous avons étudié deux principales stratégies d'intégration des paramètres prosodiques dans un système de la RAL.

- Dans la première stratégie dite « à identification directe ou fusion de paramètres », les deux types de paramètres (standards et prosodiques) sont concaténés dans le même vecteur acoustique pour former une seule observation acoustique à l'entrée d'un système de RAL.
- Une deuxième stratégie dite « à identification séparée ou fusion des scores » consiste à modéliser chaque type de paramètres par un sous système de reconnaissance indépendant, et de fusionner les scores (les probabilités) à la sortie des deux sous systèmes.

2.5.1 Fusion des paramètres

Le processus de fusion de paramètres consiste à concaténer les paramètres standards et prosodiques dans le même vecteur acoustique à l'entrée du système de reconnaissance, comme le montre la figure 2.11. Les deux types de paramètres sont intégrés directement dans le système de reconnaissance sans aucun prétraitement et avant tout processus de classification (reconnaissance). Dans cette stratégie les paramètres prosodiques sont considérés comme s'il s'agissait de paramètres cepstraux supplémentaires, sans tenir compte de leurs natures hétérogènes (deux prétraitements différents pour le même signal acoustique). C'est ce type de fusion que nous avons utilisé dans notre travail.

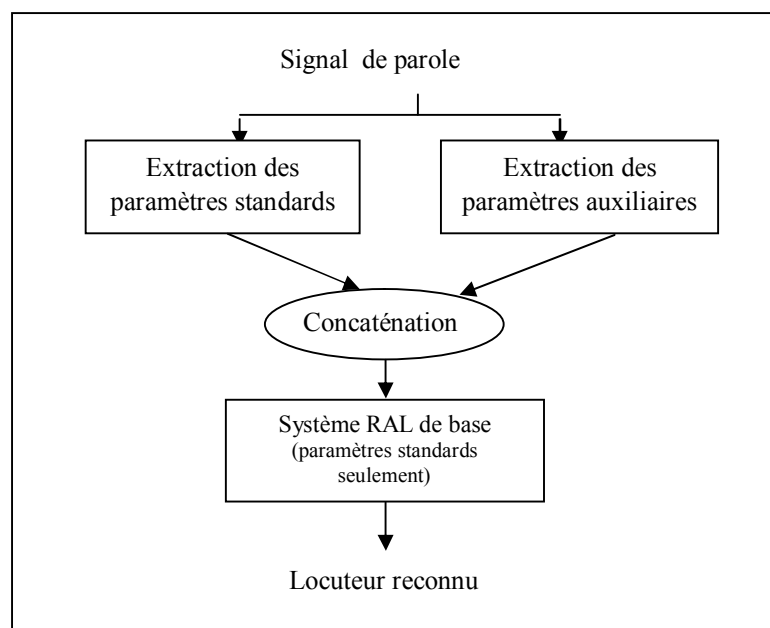


Figure 2.11 : Fusion des paramètres standard et prosodique dans un système de RAL.

2.5.2 Fusion des scores

Dans cette stratégie, chaque type de paramètres (standard et prosodiques) est modélisé par un sous système de reconnaissance séparé. Les résultats (scores probabilistes) en sortie de chacun de ces sous systèmes de reconnaissance sont fusionnés dans un module de décision qui produit le résultat final.

Les paramètres standards et prosodiques sont traités séparément dans chacun des sous systèmes de RAL.

Les deux sous systèmes, standard et prosodique, sont pratiquement identiques à un système de la RAL standard. La seule différence réside dans le module de décision. La décision de reconnaissance consiste alors à choisir parmi les N meilleures solutions proposées celle qui représente le mieux le locuteur à reconnaître.

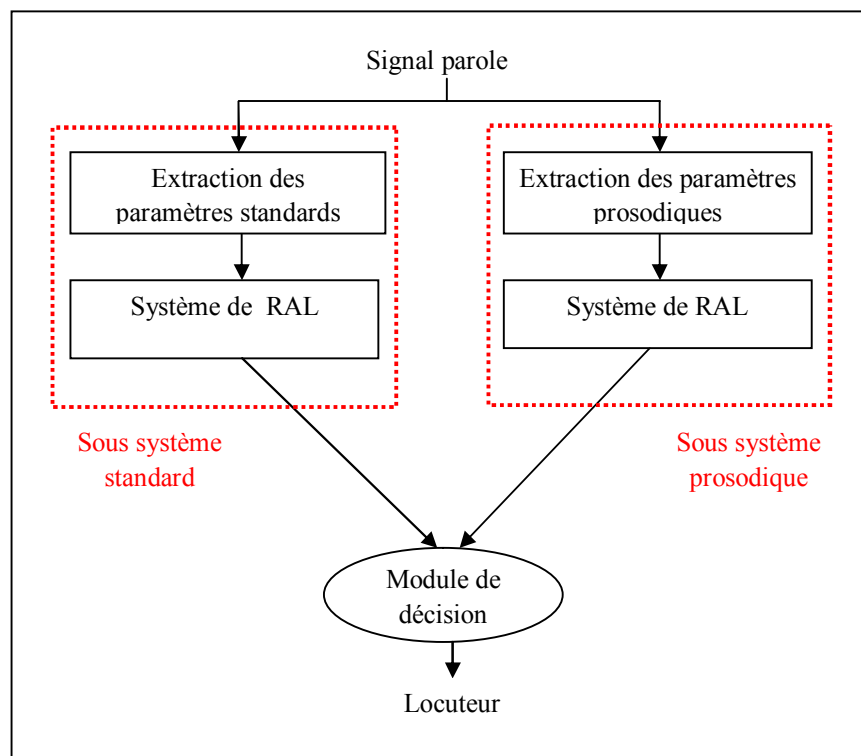


Figure 2.12: Fusion des scores des systèmes standard et prosodique dans un système RAL par identification séparée.

2.5.2.1 Méthodes de fusion

Plusieurs méthodes peuvent être utilisées pour réaliser la fusion des données dans notre cas les scores à l'intérieur d'un module de décision. Parmi ces méthodes, nous pouvons citer :

➤ **La fusion dans un cadre stochastique**

Il s'agit de la méthode de fusion la plus simple, où la sortie du module de décision n'est que l'entité correspondante au maximum de l'addition vectorielle des vecteurs probabilistes *standard* ($p_1, \dots, p_N$) et *auxiliaire* ($p'_1, \dots, p'_N$) obtenues respectivement à la sortie des deux sous systèmes pour une même entité à reconnaître. Cette fusion peut être formulée par l'équation 2.61 suivante :

$$Sortie = \arg \text{Max}_N \left(\frac{1}{2} \sum_{i=1}^N \text{standard}(p_i) + \text{auxiliaire}(p'_i) \right) \quad (2.61)$$

Sortie : représente la décision obtenue à la sortie du module de décision.

N : représente les N meilleurs scores obtenues à la sortie des deux sous systèmes standard et auxiliaire.

Cette méthode de fusion est fortement liée à la fiabilité des scores standards et auxiliaires qui change d'un contexte à un autre (présence ou absence du bruit). Ceci crée la nécessité d'adapter la simple somme de l'équation (2.61) à la fiabilité relative des deux types de paramètres, en utilisant par exemple une combinaison linéaire des scores. L'équation (2.62) peut être reformulée comme suit :

$$Sortie = \arg \text{Max}_N \left(\sum_{i=1}^N \alpha \text{standard}(p_i) + \beta \text{auxiliaire}(p'_i) \right) \quad (2.62)$$

Tel que :

$\alpha + \beta = 1$: α et β représentent le degré de fiabilité supposé pour les paramètres standards et prosodiques respectivement. Ceci pose le problème du choix des pondérations, qui change d'un contexte à un autre.

➤ La fusion par la technique du vote

Cette méthode de fusion est appelée méthode du vote et est couramment utilisée pour la fusion de données provenant de différentes modalités. La décision repose sur l'accord de la majorité des candidats (les différents sous systèmes ou modalités) à l'entrée du module de décision. Pour chaque locuteur modélisé par le système de reconnaissance, nous calculons le nombre de candidats mis en accord pour qu'elle corresponde à l'entité à reconnaître. Le module de décision rend le locuteur dont le nombre de mise en accord est le plus grand. La fiabilité de cette méthode de fusion exige la présence d'un nombre assez grand de candidats (au moins trois).

➤ La fusion neuronale

Un autre moyen de fusion des données est le cadre neuronal ; la fusion des scores est dans ce cas réalisée au moyen d'un réseau de neurones qui choisit le meilleur score parmi les $2N$ scores fournis en entrée ; son choix est basé bien entendu sur une phase d'apprentissage préalable.

2.6 Conclusion

Dans ce chapitre, nous avons présenté les principes des HMM et des SVM qui se basent sur des théories complètement différentes. Nous avons commencé par présenter le concept des modèles de Markov cachés, leur formalisme ainsi que leur emploi dans le domaine de la reconnaissance du locuteur. L'utilisation des HMM pour la reconnaissance du locuteur pose généralement trois problèmes à résoudre : l'initialisation, l'apprentissage et le décodage. C'est ainsi que nous avons étudié les différents algorithmes utilisés pour pallier à ces problèmes. Ensuite, nous avons présenté la formulation mathématique des SVM. Cette méthode de classification est basée sur la recherche d'un hyperplan qui permet de séparer au mieux des ensembles de données. Nous avons exposé les cas linéairement séparables et les cas non linéairement séparables qui nécessitent l'utilisation de fonction noyau (Kernel) pour changer d'espace. Cette méthode est applicable pour des tâches de classification à deux classes, mais il existe des extensions pour la classification multi classes. En dernier, nous avons présenté quelques techniques de fusion des données dans un système de RAL.

Chapitre 3

Paramétrisation du signal de parole

3.1 Introduction

Les systèmes de reconnaissance du locuteur utilisent des représentations acoustiques du signal de parole dans lesquelles le bruit et la redondance sont réduits afin de ne conserver que les informations considérées comme utiles à la tâche spécifiée. Ce traitement est appelé paramétrisation.

Les paramètres de l'enveloppe spectrale dits paramètres standards tels que les coefficients MFCC (Mel Frequency Cepstral Coefficients) sont actuellement les plus utilisés en reconnaissance du locuteur. Ces coefficients captent assez bien les résonances du conduit vocal mais ils sont très sensibles à la variabilité du signal de parole causée par l'environnement ou la personne [Baud-93] [Mary-08] [Ezza-02] et ne prennent pas en compte l'information de nature prosodique. Il semble donc intéressant de combiner les deux types d'analyse, d'enveloppe spectrale et prosodique [Mary-08] [Ada-03] [Carey-96] [Dehak-07] [Xie-05] [Shri-05] dans un système de RAL. Dans ce chapitre nous allons présenter les paramètres standards et prosodiques utilisés dans notre travail.

3.2 Signal de parole

3.2.1 Production et perception de la parole

La parole est le résultat de deux fonctions mécaniques de bases : la phonation et l'articulation. La phonation est la production du signal acoustique par le mouvement du larynx. L'articulation est la modulation du signal acoustique par les articulateurs (les lèvres et la langue) et la résonance de ce signal dans les cavités (la bouche et le nez) [Meu-00].

L'appareil vocal humain (figure 3.1) est constitué d'un excitateur, le complexe glotte-cordes vocales, et d'un ensemble de résonateurs de l'appareil phonatoire : le pharynx, la cavité labiale, les fosses nasales. Lorsqu'un excitateur entre en vibration, il fournit un signal, dont le résonateur va amplifier certaines composantes. La présence ou l'absence d'obstacles sur le parcours de la colonne d'air modifie la nature du son produit.

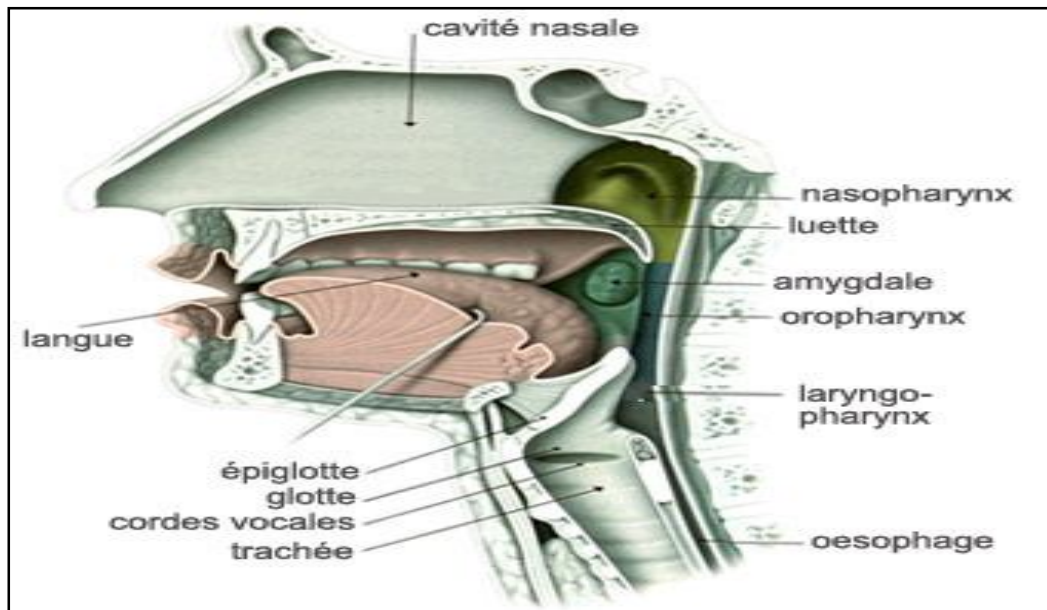


Figure 3.1 : L'appareil phonatoire humain.

L'utilité du signal de la parole produit par l'appareil phonatoire est liée à l'existence d'un récepteur qui peut capter et analyser ce signal. L'appareil auditif (figure 3.2) se divise en deux parties : le système auditif périphérique correspondant à ce que nous nommons communément l'oreille et le système auditif central.

- a) **Le système auditif périphérique** comporte trois parties : l'oreille externe, moyenne et interne.
 - **L'oreille externe** (pavillon, conduit auditif) : Elle transmet les vibrations acoustiques aériennes reçues à l'oreille moyenne.
 - **L'oreille moyenne** (le marteau, l'enclume et l'étrier) : Elle assure la transmission de la vibration sonore d'un milieu aérien à un milieu liquidien.
 - **L'oreille interne** : Elle transforme le signal acoustique en message nerveux.
- b) **Le système auditif central** est composé de la partie nerveuse, au niveau de cette partie l'information auditive périphérique sera traitée.

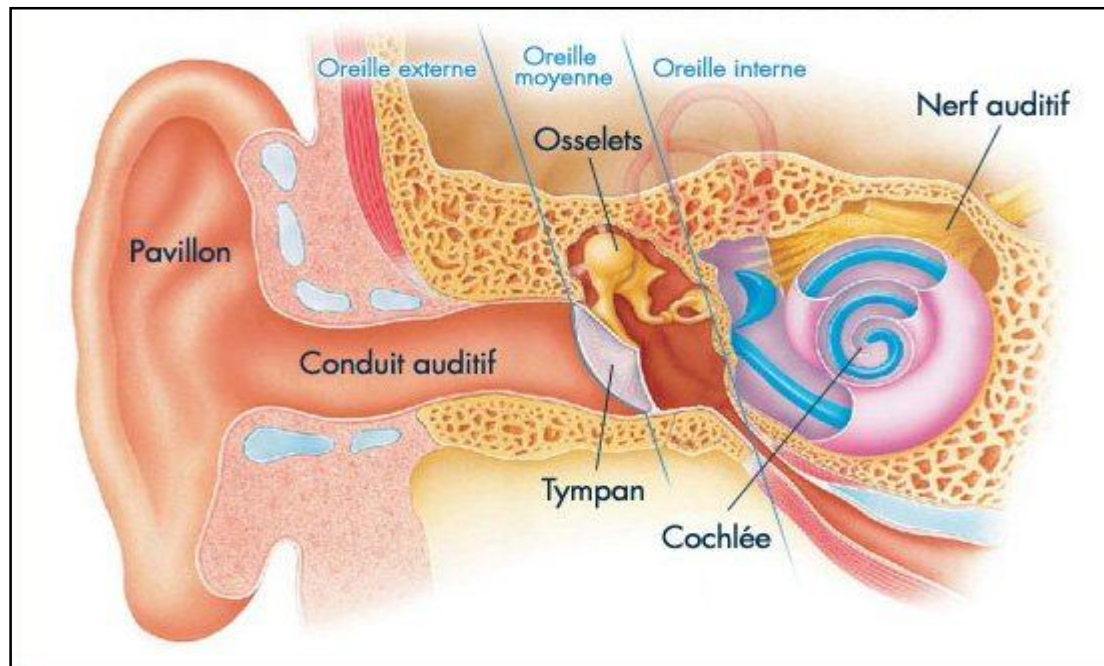


Figure 3.2 : Système auditif périphérique humain.

3.2.2 Variabilité du signal de parole

La capacité des systèmes de RAL à différencier plusieurs individus repose essentiellement sur la variabilité inter-locuteur i.e. la disposition du signal de parole à varier entre différents individus. Néanmoins, le signal de parole renferme d'autres types de variabilité qui rendent problématique la tâche de reconnaissance, telles que la variabilité intra-locuteur ou la variabilité contextuelle. Par ailleurs, les systèmes de RAL doivent faire face à d'autres difficultés liées davantage au domaine applicatif, comme l'utilisation des systèmes dans des conditions difficiles, les tentatives d'imposture, etc.

3.2.2.1 Variabilités inter-locuteurs

Elles proviennent des différences physiologiques (différences dimensionnelles du conduit vocal, fréquence d'oscillation des cordes vocales) et de différences de style de prononciation (accent, niveau social, etc. . .).

La figure suivante présente la prononciation du chiffre *siffer* par deux locuteurs différents.

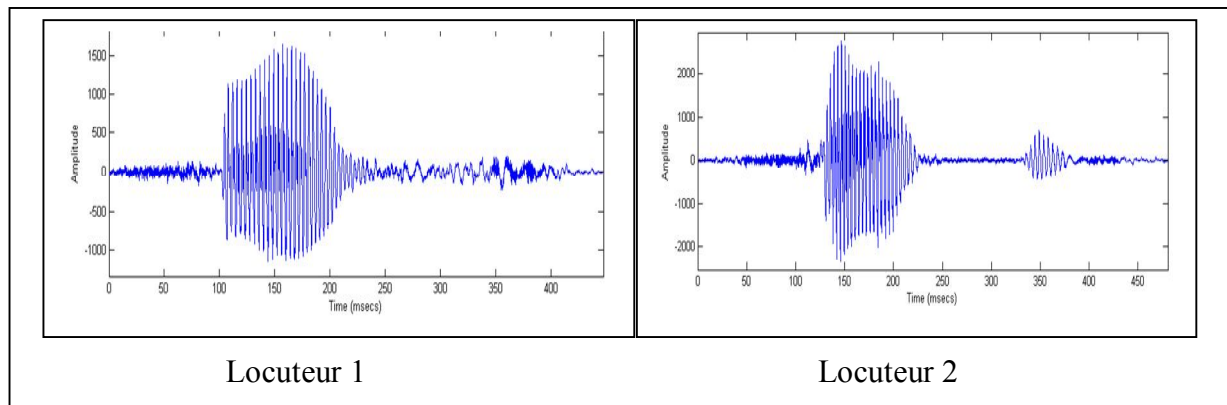


Figure 3.3: Variabilité inter-locuteurs pour le chiffre arabe « siffer ».

3.2.2.2 Variabilités intra-locuteur

Les systèmes de RAL se basent sur le principe de la variation du signal de parole entre les individus. Il est cependant clair qu'une variation du signal de parole est aussi présente pour un même individu. Cette variation peut être :

- **Ponctuelle ou à très court terme :** L'état pathologique (fatigue, rhume, etc.) ou émotionnel (stress) [Homa-95], d'une personne provoquent des altérations momentanées dans sa voix. Dans ce sens, la voix d'une personne peut évoluer entre le début et la fin de la journée (fatigue, irritation due à la pollution). D'autre part, il est impossible pour un individu de répéter consécutivement deux phrases identiques et de produire un même signal de parole pour ces deux phrases. Une légère variation est toujours observée. Finalement, une personne a la possibilité de modifier volontairement sa voix.
- **A moyen terme :** En RAL, le comportement d'un individu se modifie au fur et à mesure de son utilisation du système. L'individu devient de plus en plus confiant et sa voix évolue dans ce sens.
- **A long terme :** Une altération de la voix due à l'âge est présente chez tous les individus.

La figure suivante présente deux prononciations différentes du chiffre *siffer* par un même locuteur.

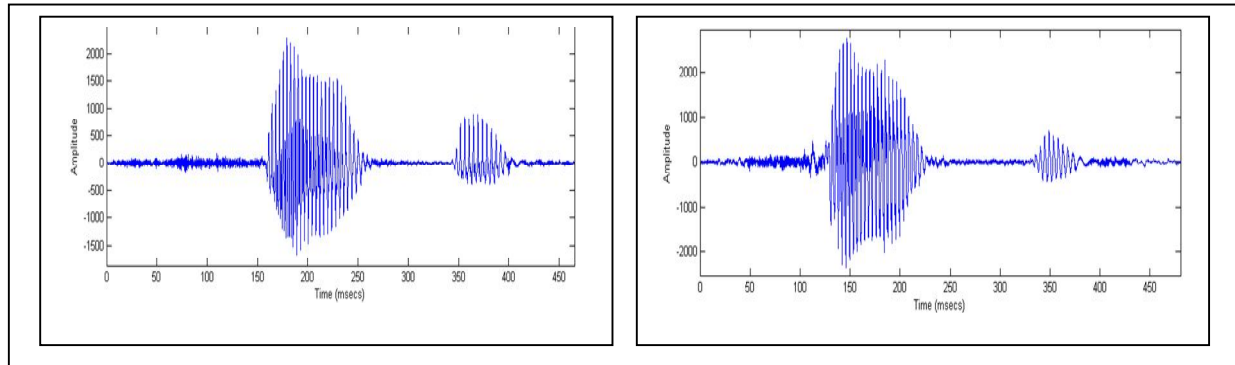


Figure 3.4 : Variabilité intra-locuteur pour le chiffre arabe « siffer ».

La variabilité intra-locuteur pose le problème de la représentativité des signaux de parole collectés lors des sessions d'enrôlement (et des modèles des locuteurs correspondant) au sein des systèmes de RAL. Des travaux ont montré que les performances d'un système sont très fortement corrélées au temps qui sépare les sessions d'enrôlement et les tests [Fur-77], [Rose-76], [Set-94]. Plus ce temps augmente, plus les performances se dégradent. Néanmoins, même les variations à court terme (émotion, état pathologique) peuvent être très préjudiciables aux systèmes de RAL.

3.2.2.3 Variabilité contextuelle

Une variabilité aussi importante est la variabilité contextuelle attribuée aux phénomènes dits de coarticulation. Nous désignons par coarticulation le fait que les mouvements accomplis par les articulateurs dans la production de la parole se chevauchent sur l'axe temporel. Le déplacement de ces organes est limité par une certaine inertie mécanique. Les sons émis subissent alors l'influence des sons qui les précèdent ou les suivent. Ces effets de coarticulation sont des interférences sur le signal de parole. Ils entraînent l'altération des formes sonores en fonction du contexte. Nous devons noter d'ailleurs que la prise en compte de cette variabilité améliore la performance des systèmes de reconnaissance.

3.2.2.4 Facteurs de distorsion

Les facteurs de distorsion dans le signal de parole peuvent être de deux natures, la première provient des facteurs environnementaux, la seconde provient des changements des conditions

d'acquisition et de transmission entre deux enregistrements. Lorsque la chaîne de transmission est maîtrisée et stable d'un enregistrement à un autre, les variations de celle-ci ne posent pas de problèmes particuliers et toute normalisation par rapport à ces variations est peu utile. Dans un cadre applicatif, ce n'est souvent pas le cas et cette variation doit être modélisée. Le protocole opératoire peut apporter, de par sa définition, des variabilités entre deux enregistrements. Celles-ci apparaissent par la quantité de données disponible, ou par l'utilisation d'un microphone différent entre les deux enregistrements. Par exemple, pour une application embarquée dans un véhicule, l'enregistrement peut être fait chez l'utilisateur dans un environnement calme à l'aide d'un microphone de qualité alors que l'accès aux données se fait dans la voiture, la fenêtre potentiellement ouverte (environnement) avec un téléphone de qualité réduite (acquisition et transmission).

La figure suivante présente la variabilité due à un environnement bruité avec un niveau RSB (Rapport Signal/Bruit) = 0 dB d'un bruit d'usine pour le chiffre arabe « siffer ».

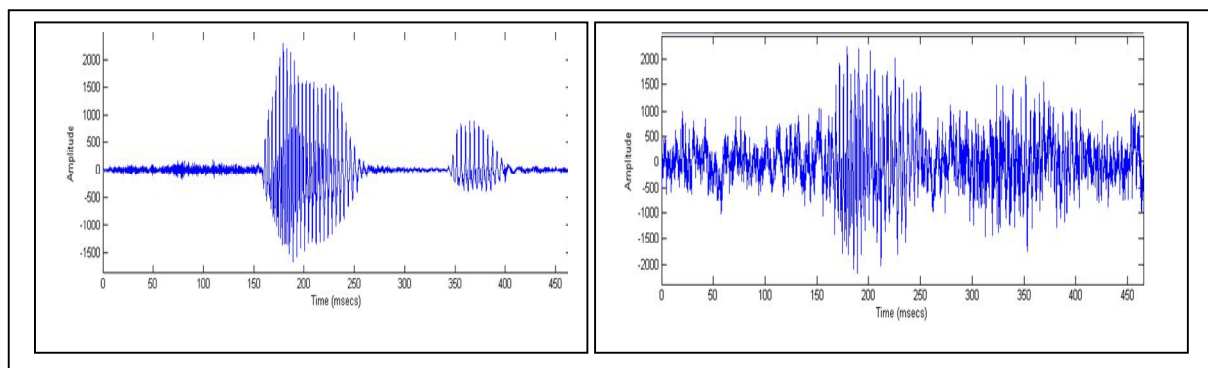


Figure 3.5: Variabilité due à un environnement bruité avec un niveau RSB (Rapport Signal/Bruit) = 0 dB d'un bruit d'usine pour le chiffre arabe « siffer ».

3.3 Paramètres de l'analyse spectrale

L'analyse spectrale est l'analyse la plus employée en RAL. Les paramètres qui en découlent sont généralement représentatifs des caractéristiques physiques de l'appareil phonatoire (forme du conduit vocal) de chaque individu.

Les principaux paramètres de l'analyse spectrale utilisés en reconnaissance du locuteur sont les coefficients issus d'une analyse par prédiction linéaire, les coefficients issus de l'analyse en banc de filtres, ainsi que les paramètres cepstraux [Fur-81] [Herm-90]. A de rares exceptions près, les caractéristiques utilisées dans les systèmes actuels de reconnaissance sont donc soit les paramètres cepstraux standards soit des paramètres de type MFCC (Mel Frequency Cepstral Coefficients).

Il faut noter ici qu'avant toute analyse (extraction de paramètres), le signal de parole subit des opérations de mise en forme dont le but est de l'adapter au calcul de coefficients acoustiques. La figure 3.6 illustre ces différentes opérations.

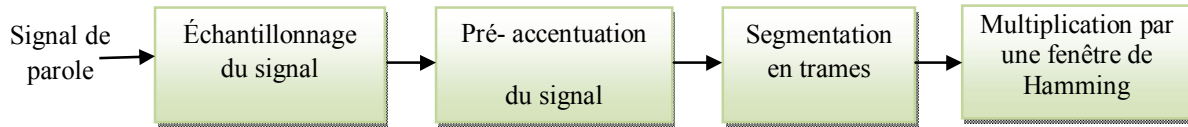


Figure 3.6 : étapes de mise en forme du signal de parole.

- **Echantillonnage** : Le signal de parole est tout d'abord filtré puis échantillonné à une fréquence donnée f_e compatible avec les exigences du théorème de Shannon :

$$f_e \geq 2.f_{\max} \quad (3.1)$$

Avec :

f_e : La fréquence d'échantillonnage.

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

- **Pré- accentuation** : elle permet d'amplifier les hautes fréquences qui sont moins énergétiques que les basses fréquences. La pré-accentuation consiste à faire passer le signal de parole dans un filtre passe-haut, dit filtre de Pré-accentuation. Ce filtre est du premier ordre et a pour fonction de transfert :

$$H(Z) = 1 - \alpha.Z^{-1} \quad (3.2)$$

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

- **Segmentation en trame** : Cette opération consiste à découper le flot de parole continue en trames pendant lesquelles le signal est supposé quasi-stationnaire. Chaque trame a habituellement une durée identique d'environ 20 à 30 ms, et le décalage entre deux trames consécutives est d'environ 15 à 10 ms.
- **Multiplication par une fenêtre de Hamming** : pour réduire les effets de bord produits par la segmentation, les trames sont alors multipliées par une fenêtre de

pondération (équation 3.4). Un exemple de fenêtres de pondération utilisées en reconnaissance, est la fenêtre de Hamming dont la formule est donnée par l'équation 3.5.

$$s''_n = w_n s'_n \quad (3.3)$$

Avec :

$$w_n = 0,54 - 0,46 \cdot \cos\left(2\pi \frac{n}{N-1}\right), \quad 0 \leq n \leq N-1 \quad (3.4)$$

3.3.1 Coefficients cepstraux

La représentation cepstrale est basée sur une connaissance du modèle de production de la parole. C'est une méthode qui vise à séparer la contribution du conduit et de la source d'excitation par déconvolution. Pour cela on fait l'hypothèse que le signal de parole $s(t)$ est produit par un signal excitateur $e(t)$ (source glottique) traversant un système linéaire passif de réponse impulsionnelle $h(t)$ (conduit). Avec ces hypothèses nous pouvons écrire :

$$s(t) = e(t) * h(t) \quad (3.5)$$

Pour déconvoluer $s(t)$ il suffit de transposer le problème par homomorphisme dans un espace où la convolution correspond à une addition. En pratique cette transposition par homomorphisme est réalisée par les étapes schématisées sur la figure suivante :

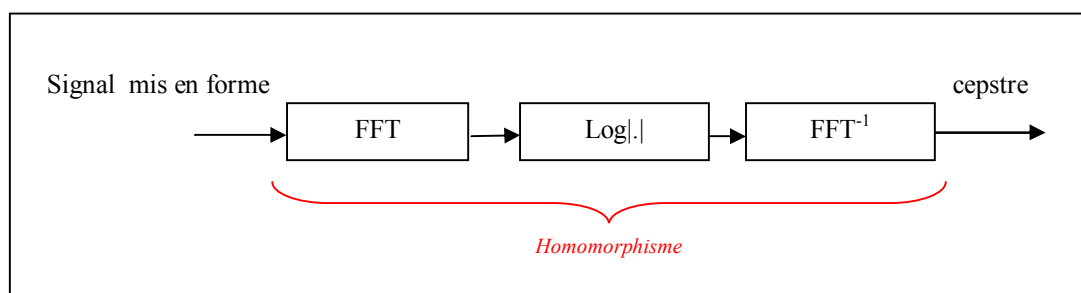


Figure 3.7 : Etapes de calcul du cepstre.

L'application du logarithme du module de la transformée de fourier à l'équation 3.5 donne :

$$\log|S(f)| = \log|E(f)| + \log|H(f)| \quad (3.6)$$

Par une transformée de Fourier inverse, nous obtenons un signal représenté dans le domaine Cepstral ou Quefrentiel (équation 3.7) :

$$s'(q) = e'(q) + h'(q) \quad (3.7)$$

$s'(q)$ sont les coefficients cepstraux approchés prenant leurs valeurs dans un domaine pseudo-temporel réel appelé domaine quéfrentiel. Cette méthode de calcul des cepstres est élémentaire, il existe d'autres méthodes itératives effectuant un lissage, ce qui permet d'obtenir des cepstres de meilleure qualité [Deb-07].

Les coefficients cepstraux ont la propriété d'être faiblement corrélés entre eux, ce qui permet d'éviter l'analyse de corrélation inter-coefficients lors de la modélisation.

3.3.2 Les coefficients MFCC

Les coefficients MFCC sont une extension des coefficients cepstraux par le passage de l'échelle fréquentielle linéaire à une échelle fréquentielle non linéaire proche de l'audition humaine. Cette échelle non linéaire est l'échelle perceptive Mel. Ces coefficients sont très utilisés en RAL du fait des bons résultats qu'ils ont permis d'obtenir. La majorité des systèmes de reconnaissance actuels utilisent ces coefficients pour représenter le signal vocal. D'ailleurs nous avons nous aussi utilisé ces coefficients dans notre travail. L'utilisation des MFCC est motivée par les deux propriétés suivantes [Rose-06] :

- 1) **Déconvolution** : les MFCC découplent les caractéristiques du conduit vocal (qui véhicule la plus grande partie de l'information disponible sur les traits distinctifs de la parole) des caractéristiques générées par l'excitation (l'information dépendante du locuteur).
- 2) **Décorrélation** : La transformée en cosinus discrète possède un effet de décorrélation entre les éléments du vecteur de traits.

➤ Echelle Mel

L'échelle Mel redistribue les fréquences selon une échelle non linéaire qui simule la perception humaine des sons. Cette échelle est linéaire pour les basses fréquences (inférieures à 1000Hz) et logarithmique pour les hautes fréquences.

La transformation de la fréquence linéaire f par cette échelle est donnée par :

$$B(f) = x \log_{10} \left(1 + \frac{f_{Hz}}{y} \right) \quad (3.8)$$

Plusieurs valeurs sont utilisées pour x et y . De nos jours, les valeurs les plus couramment utilisées sont $x = 2595$ et $y = 700$.

Le calcul des paramètres MFCC se réalise de la façon suivante (figure 3.8):

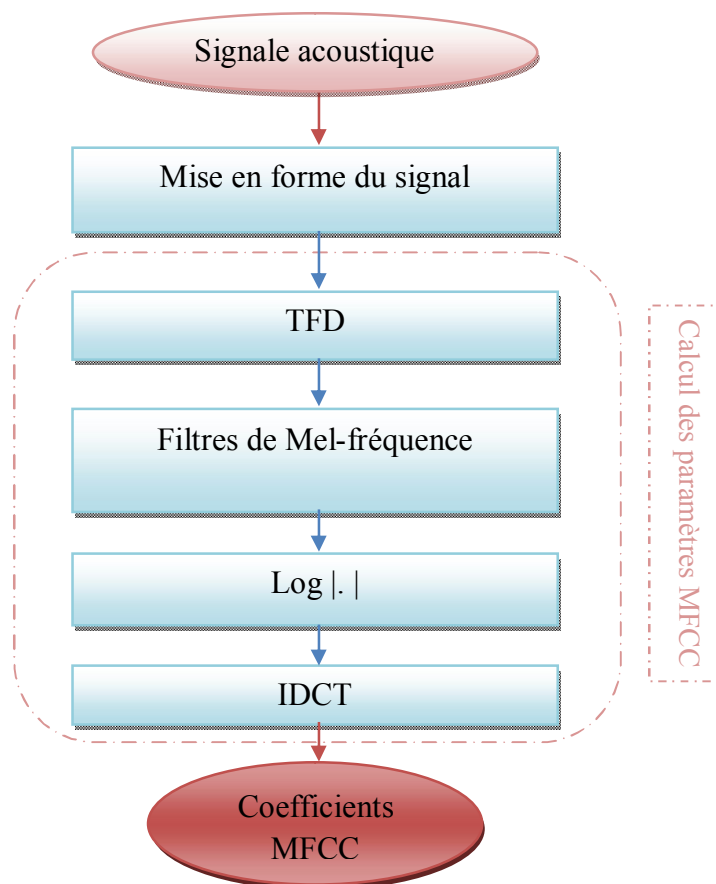


Figure 3.8: Processus de calcul des coefficients MFCC.

Après la mise en forme du signal, une TFD (Transformée de Fourier Discrète) est calculée pour faire passer le signal de parole dans le domaine spectral :

Pour un signal discret $\{s[n]\}$ avec $0 < n < N$, où N est le nombre d'échantillons d'une fenêtre analysée, F_s est la fréquence d'échantillonnage, la transformée de Fourier discrète $s[k]$ est obtenue par :

$$s[k] = \sum_{n=0}^{N-1} s[n] e^{-j 2 \pi n k / N} \quad (3.9)$$

Le spectre du signal est multiplié par des filtres triangulaires (figure 3.9) dont les bandes passantes sont équivalentes en domaine mel-fréquence. Ces filtres possèdent la caractéristique suivante : plus la fréquence est élevée, plus la bande passante est large, et donc une meilleure résolution temporelle des hautes fréquences. Ce qui permet de mieux approximer le système auditif humain que dans le cas d'un positionnement linéaire. Ce rôle perceptif est bien approprié pour les tâches de reconnaissance.

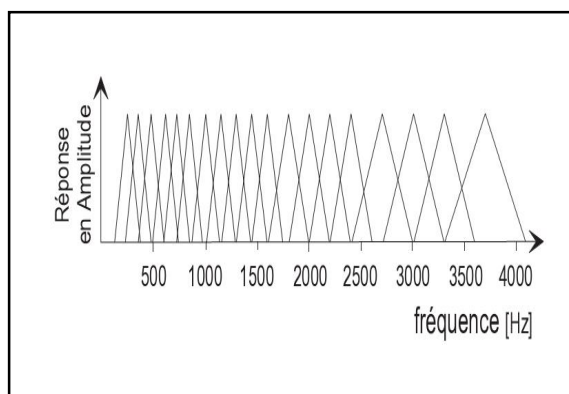
Les points frontières $B[m]$ des filtres en mel-fréquence sont calculés ainsi :

$$B[m] = B(f_l) + m \frac{B(f_h) - B(f_l)}{M + 1} \quad 0 \leq m \leq M + 1 \quad (3.10)$$

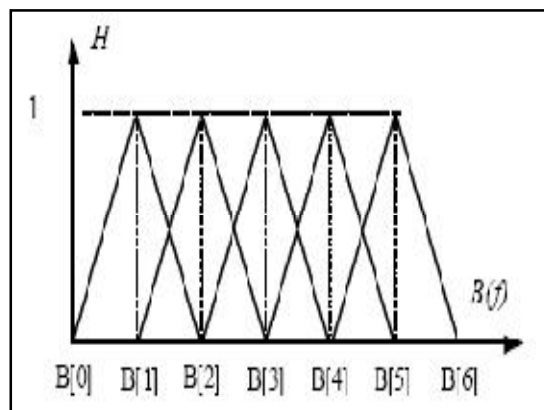
M : le nombre de filtres.

f_h : la fréquence la plus haute du signal.

f_l : la fréquence la plus basse du signal.



(a)



(b)

Figure 3.9: Banc de filtres espacés (a) sur l'échelle Mel en Hertz
(b) en Mel-Fréquence ($B(f)$).

Dans le domaine fréquentiel, les points $f[m]$ discrets correspondants sont calculés par l'équation :

$$f[m] = \left(\frac{N}{F_s} \right) B^{-1} \left(B(f_l) + m \frac{B(f_h) - B(f_l)}{M+1} \right) \quad (3.11)$$

Où B^{-1} est la transformée de mel-fréquence en fréquence.

$$B^{-1}(m) = 700 * (10^{m/2595} - 1). \quad (3.12)$$

Le coefficient $H_m[k]$ de chaque filtre est déterminé par le système suivant :

$$H_m[k] = \begin{cases} 0 & si \quad k \leq f[m-1] \\ \frac{k - f[m-1]}{f[m] - f[m-1]} & si \quad f[m-1] \leq k \leq f[m] \\ \frac{f[m+1] - k}{f[m+1] - f[m]} & si \quad f[m] \leq k \leq f[m+1] \\ 0 & si \quad k \geq f[m+1] \end{cases} \quad (3.13)$$

Pour un spectre lissé et stable à la sortie des filtres, le logarithme de spectre d'amplitude est calculé :

$$E[m] = \log \left[\sum_{k=0}^{N-1} |S[k]|^2 H_m[k] \right] \quad 0 \leq m \leq M \quad (3.14)$$

Les coefficients cepstraux de mel-fréquence (MFCCs) seront obtenus par une TCDI (Transformée en Cosinus Discrète Inverse) qui permet d'obtenir des coefficients peu corrélés à partir des coefficients aux sorties des filtres :

$$c[n] = \sum_{m=0}^{M-1} E[m] \cos \left(\frac{\pi n (m + \frac{1}{2})}{M} \right) \quad 0 \leq n \leq M \quad (3.15)$$

Une dizaine de coefficient MFCCs sont généralement considérés comme suffisants pour les expériences de reconnaissance du locuteur.

3.4 Paramètres prosodiques

La prosodie désigne la mélodie, le rythme et l'intensité de la voix. C'est un canal parallèle au contenu sémantique du message parlé dans les conversations quotidiennes à travers lequel l'auditeur peut percevoir les intentions et l'état émotionnel de l'orateur. C'est à travers la prosodie également que le locuteur peut donner à l'énoncé un ton de déclaration, d'une question, ou d'une commande [Huang-01]. De plus, les coefficients prosodiques sont plus robustes aux variations du signal acoustique.

La fréquence fondamentale (F_0) et l'énergie, mesures instantanées de la mélodie et de l'intensité, sont les paramètres les plus couramment utilisés et font l'objet de nos expériences.

3.4.1 La fréquence fondamentale (pitch)

Le pitch est le phénomène prosodique le plus expressif. Les systèmes de traitement de la parole utilisent la fréquence fondamentale, appelée encore F_0 , pour estimer le pitch. La fréquence fondamentale représente la cadence du cycle d'ouverture et de fermeture des cordes vocales de larynx durant la phonation des sons voisés. La plage de variation moyenne de cette fréquence varie d'un locuteur à l'autre en fonction principalement de son âge et de son sexe (de 80 à 200 Hz pour une voix masculine, de 150 à 450 Hz pour une voix féminine et de 200 à 600 Hz pour une voix d'enfant) [Roua-05] [Lang-95]. En parlant, nous varions systématiquement notre fréquence fondamentale pour exprimer nos sentiments ou pour diriger l'attention de l'auditeur vers un aspect important de notre message parlé. Un paragraphe prononcé avec un pitch constant et uniforme paraîtra peu naturel [Huang-01].

L'évaluation de la fréquence fondamentale est une tâche difficile pour de nombreuses raisons telles que la non-stationnarité du signal vocal, une certaines irrégularités dans l'excitation glottique ou encore une interaction avec le premier formant, la décision du voisement, la distinction entre les segments non voisée et les segments voisée à énergie réduite, la difficulté inhérente en définissant le début et la fin exacts de chaque période de F_0 durant les segments de la parole Voisée et enfin le doublement de période local [Bim-95]. C'est un type d'erreur qui affecte pratiquement toutes les méthodes d'estimation de la F_0 .

3.4.1.1 Méthodes de détection de la F_0

D'après les travaux de Hess [Hess-83], les algorithmes de détection de pitch sont classés en trois groupes principaux : temporelles, spectrales et Hybrides.

- Les méthodes temporelles permettent l'estimation de la F0 avec des calculs très simples. Elles sont relativement peu coûteuses en temps de calcul car elles nécessitent peu d'opérations arithmétiques de multiplications et d'additions [Dub-76]. Toutefois, elles manquent de précision. De nombreuses techniques temporelles sont décrites dans la littérature. Parmi les techniques de base nous pouvons citer : la Fonction d'AutoCorrélation (FAC) et ses versions modifiées [Dub-76], la Fonction de différence d'AMDF (Average Magnitude Difference Function) et ses variantes [Yu-03], la Fonction de réduction de donnée, DARD (DATA ReDUCTION method) [Mill-75] et la Fonction du calcul parallèle, PPROC (Parallel PROcessing method) [Rose-75].
- Les méthodes spectrales sont définies comme étant celles qui permettent d'obtenir un F0 en traitant le spectre de la parole directement. Parmi ces techniques, nous pouvons citer : la technique Cepstrale (CEP) [Sha-70], le Produit Harmonique Spectral (HPS), et l'intercorrélacion avec le Peigne Spectrale (PS) [Fle-06].
- Les méthodes hybrides, visent à combiner différentes approches pour augmenter les performances globales du système d'extraction. Elles appliquent différents analyseurs simultanément sur le signal et combinent les différents estimateurs [Hess-83].

Dans la plupart des algorithmes d'extraction de la F0, trois phases essentielles durant le traitement s'appliquent : le prétraitement, le traitement et le post-traitement.

- a) La phase de prétraitement est réservée à la préparation du signal issue d'un microphone. Elle consiste à choisir la durée des trames d'analyse et du recouvrement afin de moins compromettre la condition de stationnarité exigée par les algorithmes de traitement et l'effet de bord lié aux fenêtres de pondération appliquées. La durée de la trame est généralement choisie entre 20 et 50ms avec un recouvrement de 30 à 50%, pour assurer la présence d'au moins une période du Fondamental [Boi-89]. Nous trouvons souvent d'autres techniques permettant d'améliorer la rapidité d'extraction tels que le filtrage, la décimation et les techniques de transformation non linéaire dites Clipping.
- b) La phase de traitement est réservée à l'extraction de la F0 et dépend donc de l'algorithme utilisé.
- c) La phase de post traitement a pour but de diminuer les erreurs qui peuvent être de plusieurs types. On présente dans ce paragraphe les techniques choisies pour une éventuelle évaluation des performances.

3.4.1.2 Méthode basée sur la fonction autocorrélation

a) Définition de la fonction d'autocorrélation :

D'après Rabiner [Rabi-77], la fonction d'autocorrélation r_s du signal temporel $s(n)$ est donnée par la relation:

$$r_s(m) = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{n=-N}^N s(n)s(n+m) \quad (3.17)$$

Si le signal $s(n)$ est périodique, sa fonction d'autocorrélation est aussi périodique. Sa période est égale à celle du signal $s(n)$:

$$\forall n : s(n) = s(n+p) \Rightarrow r_s(m) = r_s(m+p) \quad (3.18)$$

Pour le signal de la parole qui est non stationnaire à long terme, la fonction d'autocorrélation est redéfinie pour une trame w de longueur N :

$$r_l(m) = \frac{1}{N} \sum_{n=0}^{N-m-1} [s(n+l)w(n)][s(n+l+m)w(n+m)] \quad 0 \leq m \leq M_0 - 1 \quad (3.19)$$

M_0 : est le nombre de points de la fonction d'autocorrélation.

l : est l'indice du début de la trame.

b) Application de la fonction d'autocorrélation pour l'estimation du pitch

Nous résumons la méthode d'extraction de pitch basée sur la méthode d'autocorrélation introduite par L. Rabiner [Rabi-77] en trois étapes importantes :

1) Filtrage passe bas

L'utilisation d'un filtre passe-bas (aux alentours de 900 Hertz) permet l'élimination partielle des effets des formants d'ordre supérieure à 2 sur la fonction d'autocorrélation.

2) Dispositif non linéaire

Cette étape consiste à introduire le signal dans un dispositif non linéaire appelé 3-level center. Ce dispositif non linéaire rend le calcul de la fonction d'autocorrélation plus simple et le spectre de puissance du signal plus plat pour mieux détecter le pic correspondant au pitch. La relation entre l'entrée et la sortie du dispositif est donnée par l'équation suivante :

$$Y(n) = \begin{cases} 1 & \text{si } s(n) \geq C_L \\ 0 & \text{si } |s(n)| < C_L \\ -1 & \text{si } s(n) \leq -C_L \end{cases} \quad (3.20)$$

C_L est le seuil d'écroulement (center clipped signal). Il est défini par la relation suivante :

$$C_L = 0.64 \times M \quad (3.21)$$

M représente le plus petit des maximums absolus Max1 et Max3 calculés respectivement sur le premier et le dernier tiers de la fenêtre d'analyse.

La figure suivante montre le graphe correspondant au dispositif non linéaire :

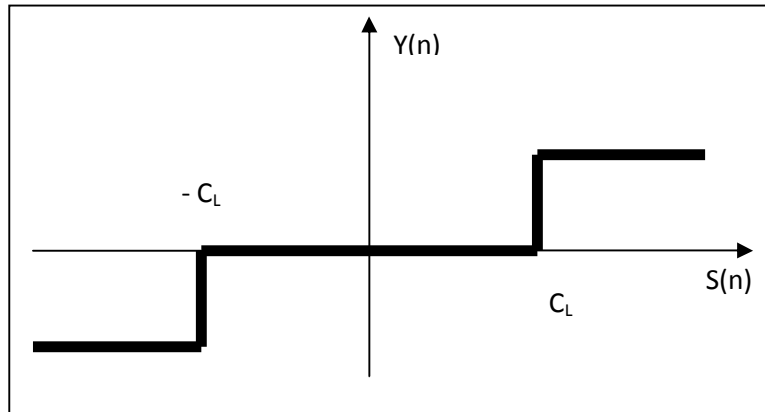


Figure 3.10 : graphe de la fonction 3-level center clipping.

Introduire le signal dans ce dispositif revient à coder chaque échantillon $s(n)$ sur trois niveaux $(-1, 0, 1)$. De ce fait la fonction d'autocorrélation du signal codé est particulièrement simple à calculer.

c) Calcul de la fonction d'autocorrélation

Une fois le signal S est codé, nous procédons au calcul de sa fonction d'autocorrélation r_s puis à la recherche du maximum de cette fonction. La décision de voisement est déterminée en comparant $R(0)$ au plus grand pic $R_{\max}(k)$:

$$\begin{cases} \frac{R_{\max}(k)}{R(0)} \leq S & , \text{ la séquence est voisée.} \\ \frac{R_{\max}(k)}{R(0)} > S & , \text{ la séquence est non voisée.} \end{cases} \quad (3.22)$$

S est un seuil à déterminer expérimentalement (exemple 30%).

La fréquence fondamentale est calculée pour les trames voisées comme suit :

$$F_0 = \frac{f_e}{K_{\max}} \quad (3.23)$$

f_e : est la fréquence d'échantillonnage du signal. $K_{\max}$: est l'abscisse de $R_{\max}(k)$

3.4.1.3 La fonction d'AMDF basée sur le Clippage

Plusieurs versions d'AMDF existent pour la détection de la F_0 [Yu-75]. Nous présentons ici la méthode basée sur le Clippage (C-AMDF) (Clipping -Average Magnitude Diffrence Function). En premier lieu, le signal vocal est filtré par un filtre passe bas de type Butterworth puis segmenté en trames. L'opération de clippage consiste à appliquer sur les fenêtres à court terme résultantes une transformation non linéaire définie par le clippage central donné par l'équation (3.20). Un seuil de clippage CL doit être calculé pour chaque trame. L'AMDF est calculée sur le signal clippé pour chaque trame d'analyse.

3.4.1.4 Le Produit Harmonique Spectral

La méthode HPS, pour Harmonic Product Spectrum (HPS), a été publiée pour la première fois par R. Noll en 1970 [Noll-70]. Cette méthode est basée sur le principe de la compression des raies spectrales. Pour chaque trame stationnaire du signal vocal $x(n)$, le logarithme de sa densité spectrale de puissance est calculé le long de l'axe des fréquences sur des facteurs entiers. La valeur logarithmique de l'HPS est obtenue par l'addition du spectre original et ses versions compressées (décimées) [Fle-06]:

$$HPS(m) = \sum_{r=1}^R \log|x(rm)|^2 \quad (3.24)$$

R : représente le nombre total des spectres impliqués dans le calcul et $X(k)$ la Transformée de Fourier Discrète (TFD) de $x(n)$. Pour obtenir le Produit Spectral d'Harmonique, l'exponentiel de la fonction doit être pris. Le choix de la constante R joue un rôle principal sur la précision du détecteur. De nombreux travaux de recherche étaient basés

fondamentalement sur un facteur de décimation $R=5$. C'est un choix qui offre une meilleure estimation de la F_0 pour une Fréquence d'échantillonnage (F_e) de 16kHz [Yu-03][Huin-06].

Quand les spectres compressés sont ajoutés, les harmoniques présentes dans le signal de parole s'ajoutent de manière constructive, puisqu'ils sont multiples de la F_0 . Les composantes fréquentielles du bruit et des sons non Voisés, ne montrent pas le même rapport, par conséquent, elles seront noyées par l'opération de la somme [Fle-06].

3.4.2 L'énergie

L'énergie ou l'intensité est une sensation auditive basée sur la perception de la force du signal acoustique. Elle correspond à la variation de l'amplitude de signal de parole causée par une force plus ou moins forte provenant du pharynx et provoquant une variation de la pression de l'air sous la glotte. Ce descripteur permet de fournir une mesure de la force sonore de la voix (faible ou forte).

L'énergie à court terme d'un signal échantillonné sur une fenêtre de longueur T , $(s_t)_{t=1,T}$, est définie par :

$$E = \frac{1}{T} \sum_{t=1}^T s_t^2 \quad (3.24)$$

Pour respecter l'échelle perceptive, elle est généralement exprimée en décibels :

$$E_{db} = 10 \times \log_{10} \left(\frac{1}{T} \sum_{t=1}^T s_t^2 \right) \quad (3.25)$$

Pour éliminer la variabilité de ce paramètre liée aux conditions d'enregistrement, cette énergie est normalisée par rapport au maximum d'énergie observé sur le signal global.

3.5 Paramètres dynamiques

L'information dynamique véhiculée par le signal de parole est une source potentielle d'informations pour les systèmes de RAL.

Les paramètres dynamiques les plus répandus demeurent les coefficients dérivés des vecteurs de paramètres instantanés, appelés coefficients Delta (première dérivée) et Delta-Delta (seconde dérivée) [Furui-81] [Soong-88], [Bern-90].

Soit le vecteur acoustique à N composantes $C_t = \{c_t^1, c_t^2, \dots, c_t^N\}$. Les coefficients delta de premier ordre sont alors estimés par :

$$\Delta C_t = \frac{\sum_{K=-L}^L K C_t}{\sum_{K=-L}^L K^2} \quad (3.26)$$

Les coefficients du second ordre sont calculés en itérant deux fois l'expression (3.26).

3.6 Conclusion

Dans ce chapitre, nous avons présenté les paramètres standards (MFCC) et prosodiques (fréquence fondamentale et énergie) utilisés dans notre système de RAL ainsi que leurs méthodes d'extraction.

Chapitre 4

Résultats Expérimentaux

4.1 Introduction

Ce chapitre présente le contexte expérimental et l'évaluation de notre travail. Notre objectif est d'améliorer la reconnaissance du locuteur en environnement réel, en intégrant des paramètres prosodiques. Ainsi deux systèmes d'identification automatique du locuteur en mode indépendant du texte ont été mis en œuvre.

- Le premier système basé sur les modèles de Markov cachés réalisé sous la plateforme HTK (Hidden Markov Toolkit) de l'Université de Cambridge. Il est construit avec une représentation du signal par les coefficients MFCC et leurs coefficients différentiels. Ce système est pris comme système de référence.
- Le deuxième système basé sur les SVM (Support Vector Machines), réalisé sous MATLAB et utilisant la même paramétrisation que le premier système.
- A partir de ces deux systèmes, deux autres systèmes basés sur la fusion des paramètres sont réalisés en fusionnant les paramètres standards avec les paramètres prosodiques.

4.2 Protocole expérimental

4.2.1 La base de données d'évaluation

La base de données utilisée dans ce travail est la base de données ARADIGITS. Cette base a été conçue au laboratoire LCPTS de la faculté d'Electronique et d'Informatique de l'USTHB [Amro-07]. Elle est constituée de prononciations de 10 chiffres de la langue arabe de zéro jusqu'à neuf prononcés par 60 locuteurs algériens de différentes régions âgés entre 18 et 50 ans. L'enregistrement a été effectué dans un environnement calme avec un niveau de bruit ambiant inférieur à 35 dB, sous le format WAV, avec une fréquence d'échantillonnage égale à 16 KHz.

Pour développer nos systèmes de reconnaissance, cette base de données est divisée en deux bases :

- **La base d'apprentissage** : Cette base comporte 1440 fichiers vocaux prononcés par 60 locuteurs comprenant les deux sexes (30 masculins et 30 féminins). Chaque locuteur répète les chiffres (0, 1, 2, 3, 4, 5, 6, 7) trois fois.
- **La base de test** : Elle comporte 360 fichiers vocaux prononcés par 60 locuteurs, comprenant les deux sexes (30 masculins et 30 féminins). Chaque locuteur répète les chiffres (8, 9) trois fois.

A partir de cette base de test, nous avons réalisé une nouvelle base bruitée :

- **La base de test bruitée:** La base de test est bruitée additivement (figure 4.1) par trois types de bruits (bruit de voiture, bruit de la foule et le bruit d'usine) pris de la base de donnée NOISEX-92 (Annexe C) avec des rapports signal sur bruit variant de 0 dB à 10 dB. Le Rapport Signal sur Bruit (RSB) d'un signal parole est défini par la formule suivante:

$$RSB = 20 \log_{10} \left(\frac{A_{signal}}{A_{bruit}} \right) \quad (4.1)$$

Tel que :

A_{signal} et A_{bruit} représentent l'amplitude du signal et du bruit.

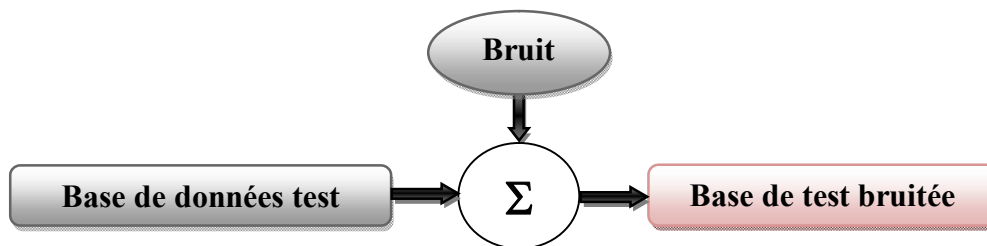


Figure 4.1 Création de la base de test bruitée.

4.2.2 Évaluation des résultats

La mesure des performances des systèmes d'identification du locuteur se base sur le calcul du Taux d'Identification Correcte (TIC) défini par :

$$TIC(\%) = \frac{\text{nombre de tests correctement identifiés}}{\text{nombre total de tests}} * 100 \quad (4.1)$$

4.2.3 Paramétrisation du signal vocal

Cette étape consiste à extraire du signal de parole les caractéristiques du locuteur qui vont permettre de le reconnaître. Deux types de paramètres ont été extraits, les paramètres standards représentés par les coefficients MFCC et les paramètres prosodiques portés par le pitch et l'énergie.

4.2.3.1 Extraction des coefficients MFCC

L'extraction des paramètres se fait selon les étapes suivantes :

- Un filtre de préaccentuation de fonction de transfert $H(z) = 1 - 0,97z^{-1}$, est appliqué sur le signal numérisé (sur 16 bits, avec $F_e = 16$ kHz).
- Une fenêtre de Hamming est appliquée sur des trames de signal de durées 25 ms avec un déplacement de 10 ms.
- Une transformée de Fourier rapide est appliquée sur chaque trame, pour obtenir le spectre.
- Un banc de 22 filtres triangulaires est utilisé pour passer à l'échelle Mel.
- Une transformée discrète en cosinus (DCT) est appliquée sur le spectre logarithmique du Mel, pour obtenir le cepstre.
- Chaque trame du signal est représentée par un vecteur de 12 premiers coefficients MFCC et leurs coefficients différentiels pour prendre en compte la dynamique du signal.
- Les vecteurs d'un même locuteur sont concaténés pour donner une matrice de taille $(nb_coef) \times (nb_vect)$ représentant le cepstre du locuteur.

4.2.3.2 Paramètres prosodiques

Les paramètres prosodiques utilisés dans notre travail sont l'Energie (E) et la Fréquence fondamentale (F0). L'outil HTK a été utilisé pour l'extraction de l'énergie et le l'outil Praat pour la fréquence fondamentale.

4.3 Construction d'un système de référence

Lors du développement d'un système de RAL, il est nécessaire d'évaluer les performances de ce système pour le comparer aux systèmes existants. De plus, les améliorations apportées à ce

système au cours de son développement doivent être quantifiées, afin de juger objectivement leur pertinence. Deux approches sont envisageables:

- tester le système sur une base de données de référence admise par la communauté,
- comparer dans les mêmes conditions de test les performances du système développé à celles d'un système de référence.

4.3.1 Caractéristiques du système de référence

Notre système de référence est un système d'Identification Automatique du Locuteur (IAL) basé sur les HMM. Ce système est mis au point à partir de la plate-forme HTK de l'Université de Cambridge et tester sur la base de données ARADIGIT. 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 la représentation du signal par des coefficients MFCC et des coefficients différentiels. Le schéma suivant représente le principe d'IAL par HMM :

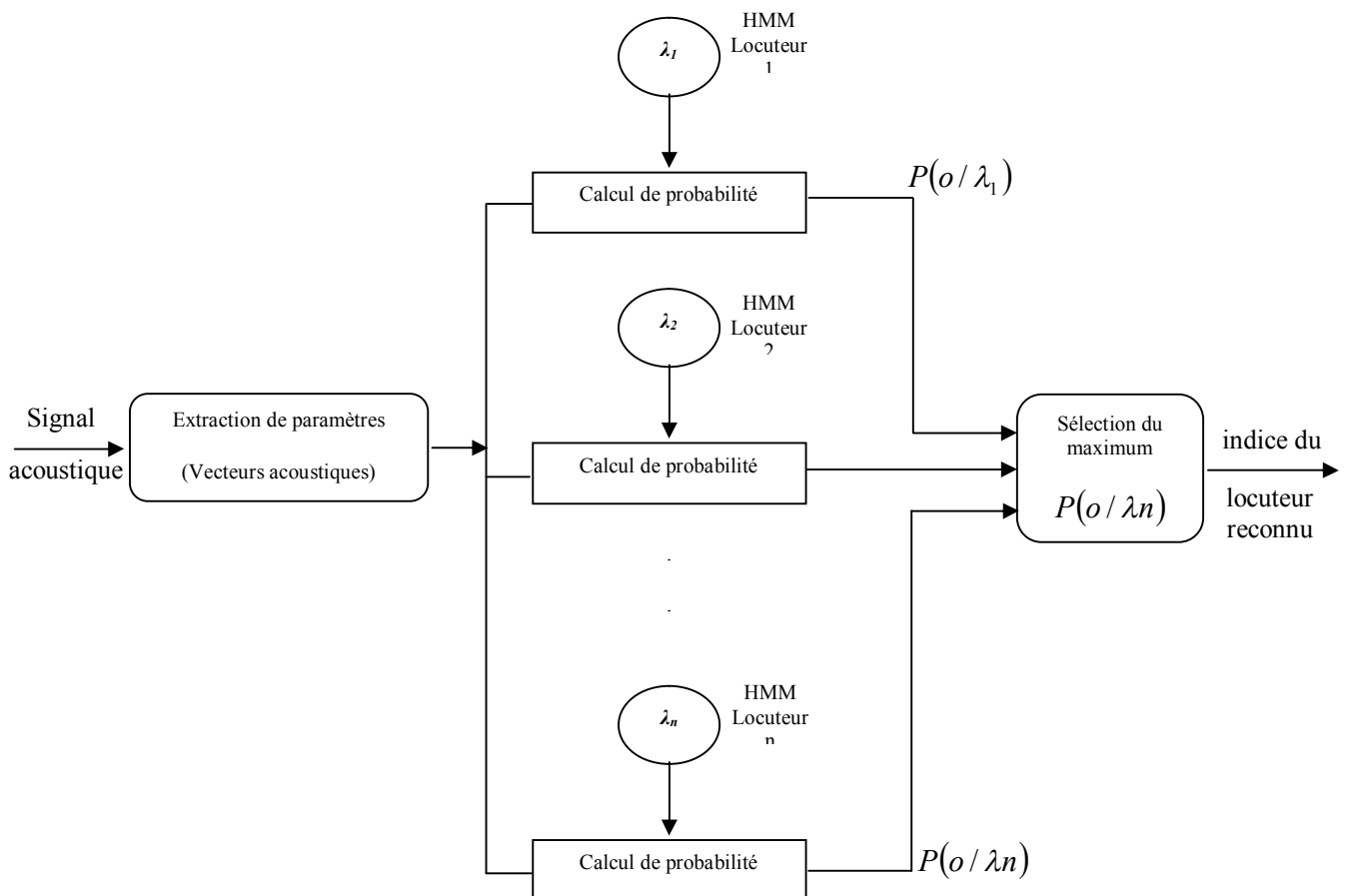


Figure 4.2: Principe d'identification du locuteur par HMM.

Chaque locuteur est modélisé par une machine de Bakis à 3 états. La probabilité d'émission de chacun des états des modèles est représentée par une combinaison linéaire de mélanges de gaussiennes à matrice de covariance diagonale. Des modèles avec une seule gaussienne par état sont initialisés par l'algorithme de Viterbi et ré-estimés par les équations de Baum-Welch. Le nombre de gaussiennes est ensuite augmenté de manière itérative: des modèles à G gaussiennes par état servent à l'initialisation de modèles à 2G gaussiennes par état, en suivant une procédure standard de HTK décrite dans l'Annexe B.

Les choix qui ont été faits pour ce système sont les suivants:

- Représentation du signal par 12 MFCC et 12 coefficients différentiels du premier ordre, soit 24 coefficients;
- Chacun des 60 locuteurs est modélisé par un HMM.
- Tous les modèles acoustiques (HMM) ont une topologie de type Bakis à 3 états émetteurs.
- Pour chaque état émetteur du modèle, la probabilité d'émission est modélisée par une combinaison linéaire de 8 gaussiennes à matrice de covariance diagonale.

4.3.2 Validation du système de référence

La robustesse du système de référence relativement à certains des choix réalisés a été étudiée. En particulier, nous testons :

- L'apport des coefficients différentiels du premier et du second ordre par rapport aux coefficients statiques seuls.
- L'apport du nombre d'état par modèle.
- L'apport du nombre de gaussiennes par état.

4.3.2.1 Apport des coefficients différentiels

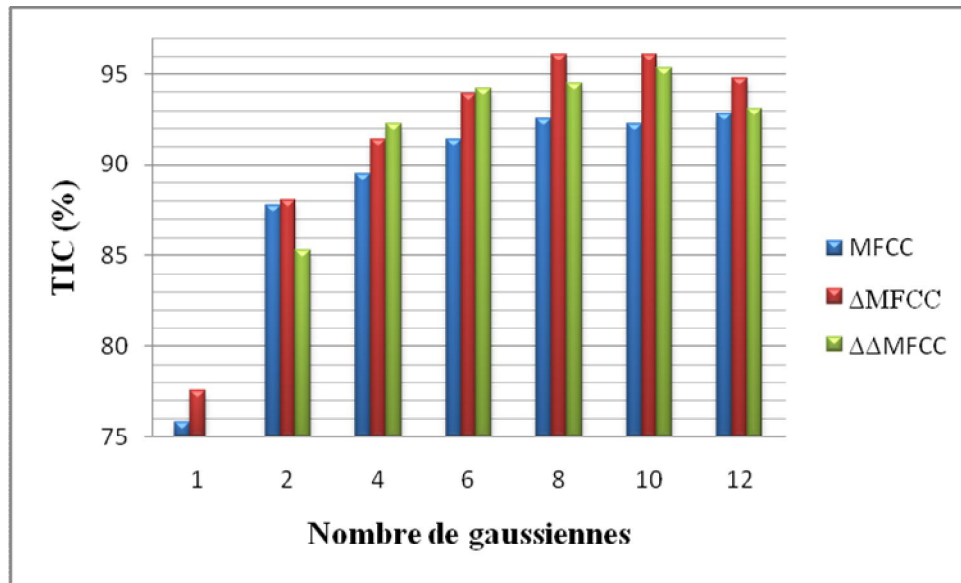
Le système de référence utilise 12 coefficients MFCC, ainsi que leurs coefficients différentiels, calculés automatiquement par les outils de HTK. Nous avons testé l'apport de coefficients différentiels du premier puis du deuxième ordre aux 12 coefficients initiaux.

Le Tableau 4.1 présente les taux d'identification de locuteurs en fonction du nombre de gaussiennes par densité de probabilité. Les meilleurs résultats sont obtenus pour 24 coefficients (12MFCC + 12 Δ MFCC) et 8 gaussiennes par état. Au delà de 8 gaussiennes, les performances du système ne sont pas significativement supérieures, tandis que la convergence de l'apprentissage est plus longue à atteindre. L'apport des coefficients différentiels du premier ordre est majeur, de 2 à 4 % sur le taux d'identification suivant les expériences. Pour un nombre identique de gaussiennes, la dérivée seconde n'apporte pas d'amélioration significative dans le cadre de ces expériences.

Pour un même nombre de paramètres d'apprentissage du système de reconnaissance, les 24 coefficients statiques et delta sont les plus performants (Figure 4.3). Ils ont donc été choisis pour le système de référence.

Tableau 4.1: Taux comparatifs coefficients statiques/coefficients dynamiques en fonction du nombre de gaussiennes.

Nombre de gaussiennes	TIC (%)		
	12 MFCC	12MFCC+12 Δ MFCC	12MFCC+12 Δ MFCC+ 12 $\Delta\Delta$ MFCC
1	75.83	77.50	73.06
2	87.78	88.06	85.28
4	89.44	91.39	92.22
6	91.39	93.89	94.17
8	92.50	96.11	94.44
10	92.22	96.11	95.28
12	92.78	94.72	93.06



Figures 4.3: Taux de reconnaissance comparatifs coefficients statiques/coefficients dynamiques en fonction du nombre de gaussiennes .

4.3.2.2 Apport du nombre d'états

Tous les locuteurs sont modélisés par une machine de Bakis à trois états (Figure 4.4). Il s'agit d'un choix classique pour la RAL. Nous pouvons interpréter l'état initial et l'état final comme des états transitoires qui encadrent un état central plus stationnaire. L'état central peut être éventuellement sauté si la réalisation acoustique est très courte.

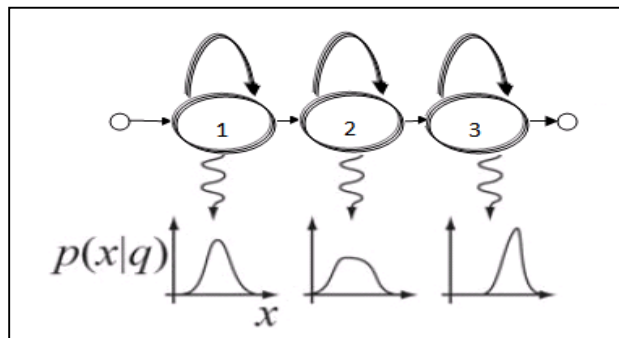


Figure 4.4 : Modèle de Bakis à 3 états émetteurs.

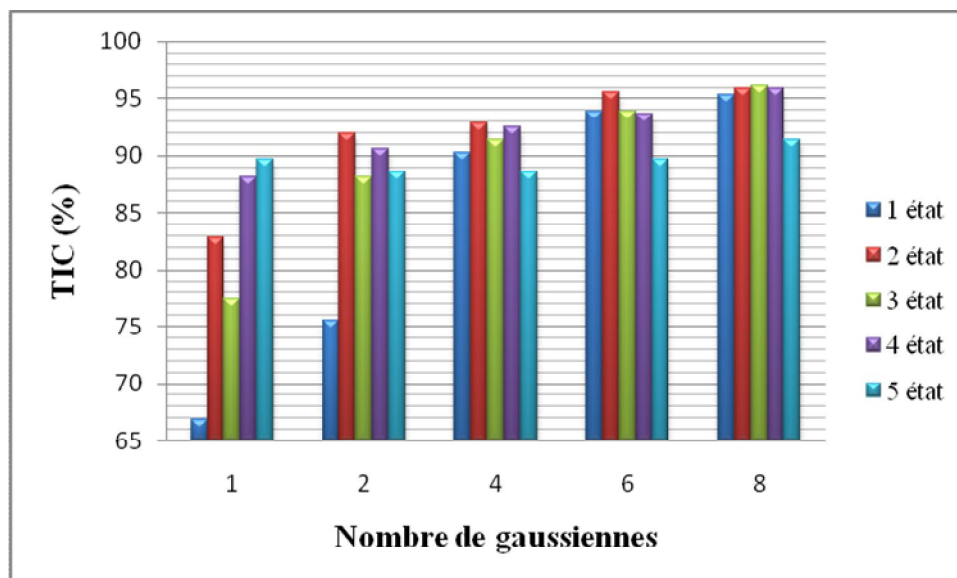
Pour valider le choix du nombre d'états du modèle, nous avons réalisé l'apprentissage de modèles à un, deux, trois, quatre et cinq états. Différents choix du nombre de gaussiennes par

densité de probabilité ont été testés afin d'obtenir des systèmes de complexité comparable (Tableau 4.32).

Dans la figure 4.5, Nous remarquons que pour 8 gaussiennes, un modèle à 3 états donne le meilleur taux d'identification (96.11%). Ceci valide notre choix.

Tableau 4.2: Influence du nombre d'état sur le taux d'identification.

		TIC (%)				
Nombre de gaussiennes	Nombre d'états	1	2	3	4	5
1		66.94	82.78	77.50	88.06	89.72
2		75.56	91.94	88.06	90.56	88.61
4		90.28	92.78	91.39	92.50	88.61
6		93.89	95.56	93.89	93.61	89.72
8		95.28	95.83	96.11	95.83	91.39



Figures 4.5: Influence du nombre d'état sur le taux d'identification.

4.4 Construction d'un système d'IAL basé sur les SVM

Pour appliquer les SVM à la tâche de reconnaissance, une idée simple consisterait à apprendre pour chaque locuteur une frontière de décision (non linéaire) dans l'espace des vecteurs d'observation. Identifier un locuteur (λ) revient à lui calculer un score S_λ^X donné par l'équation suivante :

$$S_\lambda^X = \sum_{t \in X} f_\lambda(t) \quad (4.3)$$

Où X est le segment de test, t : une trame du segment X , λ est le locuteur à identifier et f_λ est le classifieur SVM de ce dernier donné par l'équation (2.65).

Dans notre problématique, une séquence est définie comme un ensemble de vecteurs produits par un même locuteur. Ces vecteurs doivent être de taille fixe afin de pouvoir les utiliser comme paramètres d'entrée pour les SVM. Dans ce cas, une normalisation de la taille de ces derniers est nécessaire.

➤ Construction des vecteurs d'entrée des SVM

Une approche simple mais efficace utilisée dans la plupart des systèmes de reconnaissance vocale est de considérer que les signaux de parole sont composés d'un nombre fixe de trames [Chang-97] [Halb-99] [Gana-98]. Après l'étape d'extraction des coefficients MFCC et leurs coefficients différentiels, T vecteurs acoustiques représentatifs de chaque locuteur (T dépend de la durée du signal de parole) sont concaténés dans un même vecteur z_i , sachant que chaque trame est constituée de $n=36$ coefficients. Ensuite, les trames sont divisées en trois parties (P_1, P_2, P_3) de longueurs respectives (3-4-3). Un vecteur composite de dimension $3 \times n$ est obtenu par le calcul du vecteur moyen de chaque partie.

La figure suivante représente la démarche suivie pour la construction des vecteurs d'entrée de notre système à base de SVM.

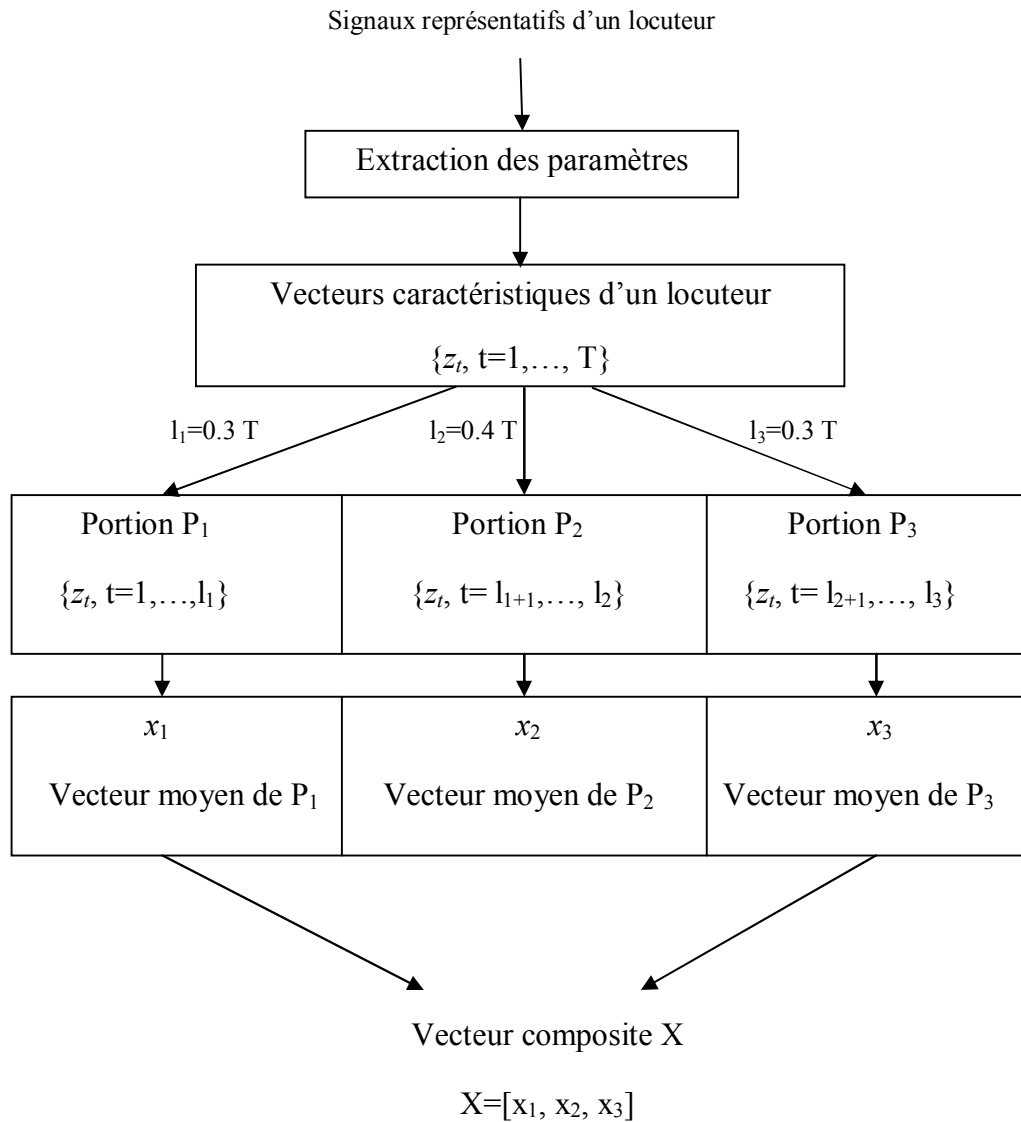


Figure 4.6 : Construction de vecteurs composites de dimension $3n$.

4.5 Fusion des paramètres prosodiques dans les deux systèmes

Nous avons étudié la robustesse des différents systèmes réalisés dans un environnement calme et bruité. L'introduction des paramètres prosodiques dans les deux systèmes d'IAL a été réalisée par une concaténation des deux types de paramètres (standards et prosodiques) ainsi que leurs dérivées premières dans un même vecteur pour former une seule observation à l'entrée du système d'IAL.

Trois nouveaux systèmes ont été obtenus :

- le premier utilise les paramètres MFCC et l'énergie ainsi que leurs dérivées comme paramétrisation, soit 26 coefficients.
- le second utilise les paramètres MFCC et le Pitch ainsi que leurs dérivées comme paramétrisation, soit 26 coefficients.
- le troisième utilise les paramètres MFCC, l'énergie et le Pitch ainsi que leurs dérivées comme paramétrisation, soit 28 coefficients.

Le tableau 4.3 présente les taux comparatifs entre le système de référence et les systèmes de fusion à base de HMM dans différents environnements pour un nombre de gaussiennes pris égal à 8.

Tableau 4.3 : Taux comparatifs des systèmes de fusion à base de HMM /système de référence pour trois types de bruits.

		TIC(%)			
Environnement		Système de référence	MFCC+E	MFCC+F0	MFCC+E +F0
calme		96.11	96.39	75.00	80.83
Bruit d'usine	10 dB	21.39	31.67	22.78	27.22
	5 dB	6.94	8.33	12.22	12.78
	0 dB	1.39	3.06	8.33	5.83
Bruit de voiture	10 dB	85.00	88.06	66.67	73.06
	5 dB	73.33	78.61	60.83	64.72
	0 dB	60.28	58.61	56.94	56.94
Bruit de la foule	10 dB	53.30	64.84	44.78	60.79
	5 dB	29.40	36.26	28.02	33.79
	0 dB	14.44	16.11	12.50	16.94

Dans la figure 4.7 nous remarquons que les performances des différents systèmes mis en œuvre se dégradent dans un milieu bruité.

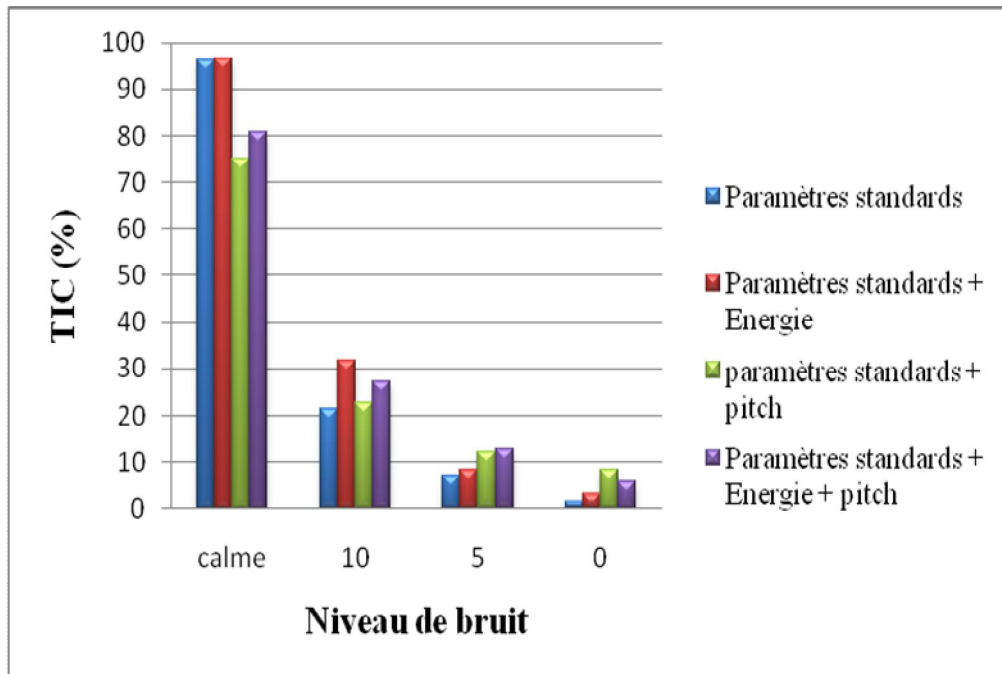


Figure 4.7 : Taux comparatifs des systèmes de fusion/système de référence dans un milieu calme et bruité par le d'usine à différents RSB.

Le tableau 4.4 présente les taux d'identification de locuteurs obtenus pour les systèmes à base de SVM dans un milieu calme et bruité par trois types de bruits (bruit d'usine, de voiture et de la foule) avec différents SNR (10, 5, 0 dB).

Dans nos expériences, Le noyau RBF (Radial Basis Function) a été utilisé avec $\sigma = 5$.

Tableau 4.4 : Taux comparatifs système basé sur les SVM / systèmes de fusion pour trois types de bruits.

		TIC(%)			
Environnement		Système SVM	MFCC+E	MFCC+F0	MFCC+E +F0
Calme ($\delta = 5$)		100	97.78	99.44	97.78
Bruit d'usine	10 dB	96.67	97.22	97.22	96.67
	5 dB	87.78	91.11	96.11	86.11
	0 dB	65	71.67	77.22	60.56
Bruit de voiture	10 dB	82.22	78.89	82.22	80
	5 dB	58.89	55.56	61.67	56.67
	0 dB	50	50	50	50
Bruit de la foule	10 dB	98.89	97.22	97.22	96.67
	5 dB	95.56	91.11	96.11	86.11
	0 dB	92.22	71.67	77.22	60.56

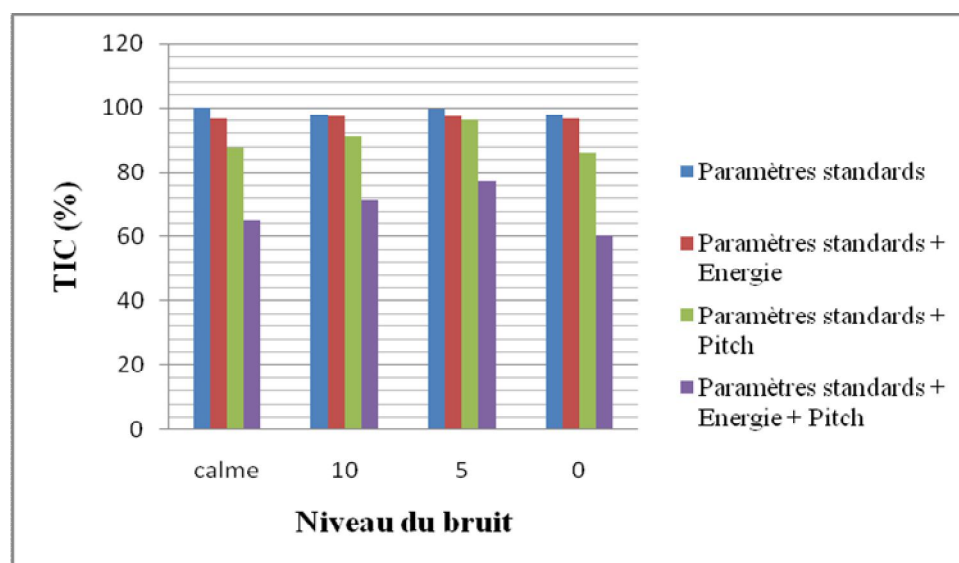


Figure 4.8: Taux comparatifs système basé sur les SVM / systèmes de fusion pour le bruit de la foule.

➤ Interprétation des résultats

L'interprétation des résultats est basée sur la comparaison des TIC des différents systèmes implémentés : le système de référence, le système à base de SVM et les six systèmes de fusion.

Le tableau 4.5 présente les taux d'identifications obtenus par ces différents systèmes, pour un environnement calme et bruité par différents bruits.

Tableau 4.5 : Taux comparatifs des différents systèmes mis en œuvre.

		TIC(%)								
Environnement Systèmes	Calme	Bruit de la foule			Bruit de voiture			Bruit d'usine		
		10 dB	5 dB	0 dB	10 dB	5 dB	0 dB	10 dB	5 dB	0 dB
Système de référence (paramètres standards)	96.11	53.30	29.40	14.44	85	73.33	60.28	21.39	6.94	1.39
SVM (Paramètres standard)	100	98.33	95.56	92.22	82.22	58.89	50	96.67	87.78	65
HMM (Paramètres standard +Energie)	96.39	64.84	36.26	16.11	88.06	78.61	58.61	36.67	8.33	3.06
SVM (Paramètres standard +Energie)	97.98	98.33	96.11	91.67	78.89	55.56	50	97.22	91.11	71.67
HMM (Paramètres standard +pitch)	75.00	44.78	28.02	12.50	66.67	60.83	56.94	22.78	12.22	8.33
SVM (Paramètres standard +pitch)	99.44	98.33	97.78	97.78	82.22	61.67	50	97.22	96.11	77.22
HMM (Paramètres standard +Energie+pitch)	80.83	60.79	33.79	16.94	73.06	64.72	56.94	27.22	12.78	5.83
SVM (Paramètres standard +Energie+pitch)	97.78	98.33	97.22	91.66	80	56.67	50	96.67	86.11	60.56

Les deux systèmes basés sur les paramètres cepstraux de type MFCC donnent de très bonnes performances dans un environnement calme avec un avantage pour les SVM (TIC = 100%). Nous avons constaté que les performances de ces deux systèmes se dégradent dans des conditions réelles (figure 4.9). Cette dégradation est très marquée pour le système à base de HMM, il atteint un taux de 1.39% pour un milieu fortement bruité (bruit d'usine à 0dB).

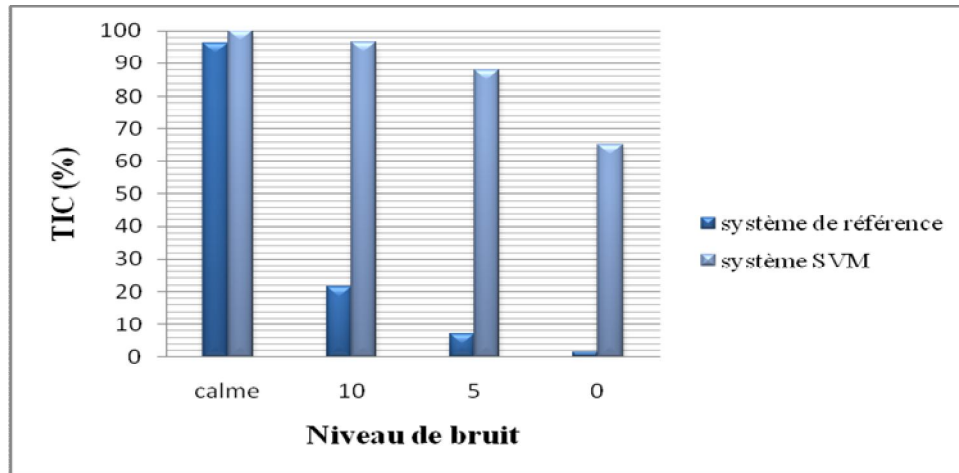
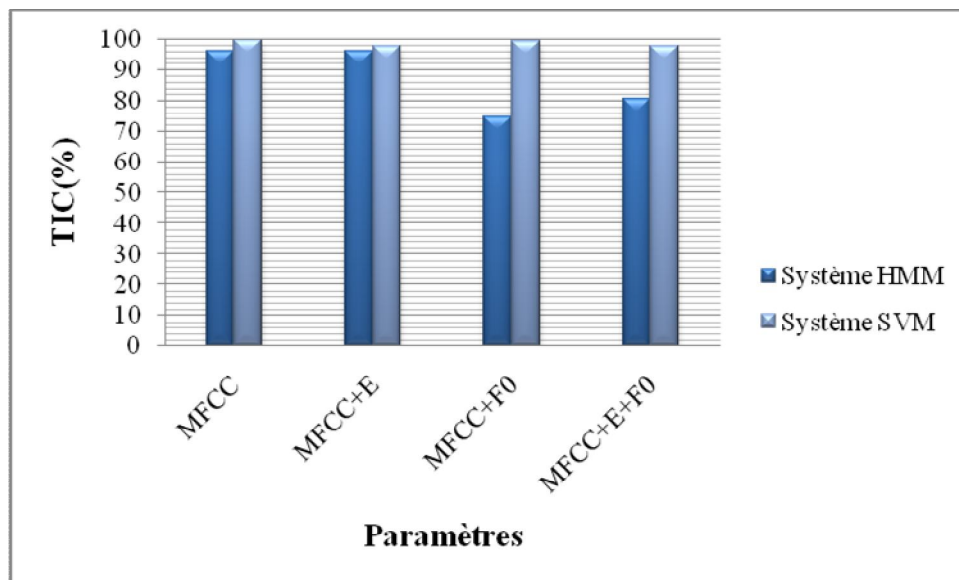


Figure 4.9: Influence du bruit sur les systèmes à paramètres standards.

Nous remarquons que dans un environnement calme, l'ajout de l'énergie améliore le taux d'identification du système à base de HMM (96.11% vs. 96.39%) et dégrade celui du système à base de SVM (100% vs. 97.98%) (Figure 4.10). Par contre, l'ajout du Pitch provoque une dégradation des taux d'identification des deux systèmes (96.11% vs. 75% pour le système à base de HMM et 100% vs. 97.78% pour le système à base de SVM). Cette dégradation des taux est due au fait que les paramètres prosodiques additifs perturbent les paramètres standards.



Figures 4.10: Influence de l'ajout des paramètres prosodiques sur les deux systèmes.

Nous constatons que les performances du système à base de HMM sont plus touchées que celles du système à base de SVM. Ceci peut s'expliquer par le fait que la modélisation des

paramètres du système à fusion de paramètres par une distribution multi-gaussiennes peut être non adéquate pour modéliser le vecteur allongé par les paramètres prosodiques.

Cependant en présence du bruit, l'énergie apporte une nette amélioration pour les deux systèmes (53.30% vs. 64.84% pour un système à base de HMM c-à-d un gain d'environ 11% et 65% vs. 71.67% c-à-d un gain d'environ 7%).

L'ajout du Pitch améliore le taux d'identification du système à base de SVM dans un milieu bruité quelque soit le bruit utilisé et son degré (92.22% vs. 97.78% pour le bruit de la foule à 0 dB, 58.89% vs. 67.67% pour le bruit de voiture à 5dB et 65% vs. 77.22% pour le bruit d'usine à 0 dB). Par contre dans le cas d'un système à base de HMM l'amélioration n'est remarquée que pour un milieu fortement bruité (1.39% vs. 8.33% pour le bruit d'usine à 0 dB).

Les performances des systèmes à base de SVM sont toujours meilleures que celles obtenues avec les systèmes à base de HMM (dans les conditions calmes et bruitées). L'ajout des paramètres prosodiques améliore la reconnaissance surtout en milieu fortement bruité.

4.6 Conclusion

Dans ce chapitre, nous avons réalisé deux systèmes d'identification du locuteur. Le premier basé sur les HMM, pris comme système de référence, et le deuxième sur les SVM utilisant les paramètres MFCC comme modélisation acoustique du signal de parole représentatif du locuteur. Par la suite, nous avons introduit un autre type de paramètres dits paramètres prosodiques, qui sont dans notre cas : le pitch et l'énergie, dans les deux systèmes. Nous avons évalué, dans les mêmes conditions, les performances des systèmes développés au regard de celles du système de référence. D'après les résultats, l'intégration des paramètres prosodiques pour les deux types de systèmes apporte une amélioration remarquable des taux d'identification dans un milieu réel avec un avantage pour les SVM. Ceci montre la sensibilité réduite des paramètres prosodiques au bruit par rapport aux paramètres MFCC et motive l'utilisation des paramètres prosodiques pour la reconnaissance du locuteur en environnement réel.

Conclusion et perspectives

Conclusion et perspectives

Dans le cadre de ce mémoire nous nous sommes intéressés à la robustesse des systèmes de RAL en environnement réel donc bruité. Pour cela, nous avons procédé à l'étude de l'influence de l'approche de reconnaissance et de l'intégration des paramètres prosodiques dans le front-end. Deux systèmes d'Identification Automatique du Locuteur (IAL) en mode indépendant du texte basés sur deux approches de reconnaissance différentes et utilisant les paramètres MFCC comme paramétrisation acoustique ont été mis en œuvre.

Le premier basé sur les modèles de Markov cachés (HMM) a été réalisé avec la plate-forme HTK (Hidden Markov Toolkit), pris comme système de référence afin d'évaluer les performances des systèmes réalisés au cours de ce travail.

Le deuxième basé sur les Support Vector Machines (SVM) a été réalisé avec le logiciel MATLAB. Un problème important se pose lors de l'utilisation des SVM pour la RAL. Les vecteurs de paramètres acoustiques représentatifs des locuteurs ont des tailles différentes. Or, les vecteurs d'entrée des SVM doivent avoir une taille fixe. Pour palier à ce problème, nous avons effectué une normalisation de la taille de ces vecteurs.

Par la suite, nous avons introduit les paramètres prosodiques, qui sont dans notre cas : le pitch et l'énergie, dans les deux systèmes. L'intégration de ces derniers dans le système de RAL a été réalisée par une concaténation des deux types de paramètres dans le même vecteur acoustique.

Nous avons évalué, dans les mêmes conditions, les performances des systèmes développés au regard de celles du système de référence.

En environnement calme, les performances des différents systèmes sont presque similaires, avec un léger avantage pour les SVM. Cependant, en milieu réel les performances de ces deux systèmes se dégradent.

Nous avons constaté à travers nos expériences que les SVM sont plus robustes au bruit que les HMM. C'est ce qui explique l'intérêt grandissant pour les SVM dans les systèmes de reconnaissances de locuteur.

L'intégration des paramètres prosodiques pour les deux types de systèmes apporte une amélioration remarquable des taux d'identification dans un milieu réel avec un avantage pour les SVM. Ceci montre la sensibilité réduite des paramètres prosodiques au bruit par rapport aux paramètres MFCC et motive l'utilisation des paramètres prosodiques pour la reconnaissance du locuteur en environnement réel.

Malgré les améliorations remarquables des taux de reconnaissance obtenus dans les environnements bruités, le travail proposé dans ce mémoire reste à notre avis perfectible.

Comme perspectives à ce travail nous proposons :

- L'introduction d'autres paramètres prosodiques tels que la durée, le rythme et la vitesse d'élocution.
- L'utilisation d'un module de rehaussement de la parole à l'entrée du système de reconnaissance afin de le rendre moins sensible au bruit.
- L'Utilisation d'une approche hybride de type SVM/HMM pour allier le pouvoir discriminant des SVM au pouvoir de modélisation des HMM.
- l'utilisation des techniques de prétraitement de données (ex : ACP Analyse en Composantes Principales) peut être très efficace pour améliorer la modélisation de la distribution acoustique par des lois gaussiennes.

Bibliographie

- [Ada-03] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey. Modeling prosodic dynamics for speaker recognition. In Proceedings of ICASSP, volume IV, pages 788–791, 2003.
- [Amro-07] A. Amrouche “Reconnaissance automatique de la parole par les modèles connexionnistes“. Thèse de doctorat, faculté d’électronique et d’informatique, USTHB. 2007.
- [Augu-01] E. Augustin “Reconnaissance de mots manuscrits par systèmes hybrides Réseaux de Neurones et Modèles de Markov Cachés”, Thèse de Doctorat, Paris V, 188 pages, 2001
- [Barr-96] C. Barras, "Reconnaissance de la parole continue : adaptation au locuteur et contrôle temporel dans les modèles de Markov cachés", Thèse de doctorat en informatique, Université de Paris VI, mai 1996.
- [Baud-93] G. Baudoin, P. Jardin et d’autres « Comparaison de techniques paramétrisation spectrale pour la reconnaissance vocale en milieu bruité » quatorzième colloque gretsi. Septembre 1993.
- [Baum-72] L. Baum « An inequality and association maximization technique in statistical estimation for probabilistic functions of Markov processes » Inequality, vol. 3, 1972.
- [Bell-57] R. Bellman, «Dynamic programming » Princeton University Press, 1957.
- [Ben-03] Y. Ben Ayed, “Détection de mots clés dans un flux de parole”, thèse de l’Ecole Nationale Supérieure des Télécommunications, 2003.
- [Bern-90] C. Bernasconi, “On instantaneous and transitional spectral information for text-dependent speaker verification”. Speech Communication, volume 9(2), pages 129-139, 1990.
- [Boser-92] B. Boser, I. Guyon, et V. Vapnik, “An training algorithm for optimal margin classifiers,” in Proc. Fifth Annual Workshop on Computational Learning Theory, pp. 144-152, 1992.
- [Bour-90] H. Bourlard & C.J. Wellekens, « Links between Markov models and multilayer perceptrons» IEEE Trans. on Pattern Anal. and Machine Intell., vol. 12, no. 12, pp. 1-12, 1990.

- [Bon-03] J. Bonastre, F. Bimbot, L. Boë, J. Campbell, D. Reynolds, et I. Magrin-Chagnolleau, "Person authentication by voice : a need for caution. Dans Proceedings of Interspeech", European Conference on Speech Communication and Technology, Geneva, Switzerland, pp. 33–36, 2003.
- [Burg-98] C. J.C. Burges, "A tutorial on support vector machines for pattern recognition", *Data Mining and Knowledge Discovery*, 2(2):121–167, juin 1998.
- [Chang-97] J. Chang and J. Glass, "Segmentation and modeling in segment-based recognition," in *Eurospeech*, Rhodes, Greece, 1997.
- [Call-03] J. Callut, "Implémentation efficace des Support Vector Machines pour la Classification", Université Libre de Bruxelles, Faculté des sciences, Département d'informatique, Année académique 2002-2003.
- [Cam-04] W. M. Campbell, E. Singer, P. A. Torres-Carrasquillo et D. A. Reynolds, "Language Recognition with Support Vector Machines", In *Proc. Odyssey: The Speaker and Language Recognition Workshop in Toledo, Spain*, ISCA, pages 41–44, 2004.
- [Carey-96] M. J. Carey, E. S. Parris, H. Lloyd-Thomas, and S. Bennet. Robust prosodic features for speaker identification. In *Proceedings of ICSLP*, pages 1800–1803, 1996.
- [Deby-07] M. Debyeche « Reconnaissance automatique de la parole appliqué à la langue arabe » Thèse de Doctorat d'Etat, USTHB, 2007.
- [Dehak-07] N. Dehak, P. Kenny, and P. Dumouchel. Continuous prosodic features and formant modeling with joint factor analysis for speaker verification. In *Proc. of Interspeech*, pages 1234–1237, 2007.
- [Demp-77] A. P. Dempster, N. M. Laird, et D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm", *Journal of Acoustical Society of America JASA*, Volume 39, pp. 1–38, 1977.
- [Cort-95] C. Cortes, and V. Vapnik, "Support Vector Networks", *Machine Learning* 20, 273 297, 1995.

[Dodd-85] G. R. Doddington, "Speaker recognition. identifying people by their voices", IEEE transactions, Volume 7311, pp. 1651–1664, 1985.

[Ezza-02] H. Ezzaidi « Discrimination Parole/Musique et étude de nouveaux paramètres et modèles pour un système d'identification du locuteur dans le contexte de conférences téléphoniques » Thèse de doctorat. L'Université du Québec à Chicoutimi. Département des Sciences Appliquées. Octobre 2002.

[Forn-73] G.D. Forney, "The Viterbi algorithm", Proc. of the IEEE, vol. 61, no. 3, pp. 268-278, 1973.

[Furui-81] Furui S. "Cepstral analysis technique for automatic speaker verification". IEEE Transactions Acoustics, Speech, and Signal Processing (ASSP), volume 29(2), pages 254-272, Avril 1981.

[Gana-98] A. Ganapathiraju, J. Hamaker, and J. Picone, "Support vector machines for speech recognition," in International Conference on Spoken Language Processing (ICSLP), Sydney, Australia, Nov. 1998.

[Gray-84] R.M. Gray « Vector quantization » IEEE ASSP Mag., vol. 1(2):4-29, 1984.

[Halb-99] A. Halberstadt, "Heterogeneous acoustic measurements and multiple classifiers for speech recognition," Ph.D. dissertation, Massachusetts Institute of Technology, Dept. of Electrical Engineering and Computer Science, 1999.

[Herm-90] P. Hermansky, «Perceptual Linear Predictive (PLP) analysis of speech » Journal of the Acoustical Society of America, 87-4 :1738–1752, 1990.

[Hsu-10] C. Hsu, C. Chang, et C. Lin, A practical guide to Support Vector Classification.

[Igou-98] S. Igounet "Eléments pour un système de reconnaissance automatique de la parole continue du français " Thèse de doctorat en Informatique, Université d'Avignon et des Pays de Vaucluse, juillet 1998.

[Kharr-03] J. Kharroubi, "Etude de Techniques de Classement 'Machines à Vecteurs Supports' pour la Vérification Automatique du Locuteur", Thèse de doctorat en signal et image, Ecole nationale supérieure de télécommunications, Paris, Juillet 2002.

[Li-02] Z. Li, S. Tang et S. Yan, “Multi-class SVM classifier based on pairwise coupling”, In Proc. of the First International Workshop on Pattern Recognition with Support Vector Machines, pages 321–333, 10 août 2002.

[Lar-09] A. Larcher, “Modèles acoustiques à structure temporelle renforcée pour la vérification du locuteur embarquée”, thèse de doctorat en informatique, l’Université d’Avignon et des Pays de Vaucluse, septembre 2009.

[Mami-03] Y. Mami, “Reconnaissance de locuteurs par localisation dans un espace de référence”, Thèse de doctorat en signal et image, Ecole nationale supérieure de télécommunications, Paris, Octobre 2003.

[Mary-08] L. Mary, B. Yegnanarayana « Extraction and representation of prosodic features for language and speaker recognition » Speech Communication 50, 782–796. April 2008.

[Pogg-90] T. Poggio et F. Girosi, “Regularization algorithms for learning that are equivalent to multilayer networks”, Science, 247(4945):978–982, février 1990.

[Rabi-89] L.R. Rabiner, «A tutorial on hidden Markov models and selected applications in speech recognition» Proc. of the IEEE, vol. 77, no. 2, pp. 257-286, 1989.

[Rose-91] A. E. Rosenberg et F. K. Soong, “Recent research in automatic speaker recognition”, Advances in speech signal processing, pp.701–737, 1991.

[Rossi-89] M. Rossi, “De la quiddité des variables. Actes du séminaire Variabilité et spécificité du locuteur : Etudes et Applications, 11–31, 1989.

[Sath-03] S. Sathiya Keerthi and Chih-Jen Lin. Asymptotic behaviors of support vector Machines with Gaussian kernel. Neural Computation, 15(7):1667-1689, 2003.

[Sch-96] M. Schmidt et H. Gish, “Speaker identification via support vector classifiers ”, Acoustics, Speech and Signal Processing, USA, page 105-108, vol. 1, 1996.

[Shri-05] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke. Modeling prosodic feature sequences for speaker recognition. Speech Communication, 46(3- 4) :455–472, July 2005.

[Smo-98] A. J. Smola et B. Scholkopt, "Atutorial on support vector regression", Neuro COLT2 Technical Report Series NC2-TR-1998-030, Octobre 1998.

[Son-99] K. Sonmez, L. P. Heck et M. Weintraub, "Speaker tracking and detection with multiple speakers", Proceedings of European Conference on Speech Communication and Technology, 1999.

[Soong-88] F. K. Soong et A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition". IEEE Acoustics Transactions, Speech, and Signal Processing (ASSP), volume 36(6), pages 871-879, Juin 1988.

[Tou-08] B. Tounsi, "Inférence d'identité dans le domaine forensique en utilisant un système de reconnaissance automatique du locuteur adapté au dialecte Algérien", Mémoire de magister en Informatique, Institut National de Formation en Informatique (I.N.I), Oued-Smar Alger, Année Universitaire 2007-2008.

[Vap-74] V. N. Vapnik et A. Y. Chervonenkis, "Theory of pattern Recogniton", Nauka, Moscow, 1974.

[Vap-95] V. Vapnik, "The nature of statistical learning theory", N-Y: Springer-Verlag, 1995.

[Vit-67] A.J. Viterbi. Error bound for convnutional codes and an asymptotically optimum algorithm. Transactions on Information Theory, 13(2) :260–269, 1967.

[Wan-00] V. Wan et W. M. Campbell, 2000. "Support Vector Machines for Speaker Verification and Identification". IEEE Signal Processing Society Workshop Neural Networks for Signal Processing, Volume 2,

[Wolf-72] J. J. Wolf, "Efficient Acoustic Parameters for Speaker Recognition", the journal of the Acoustical Society of America, 1972.

[Xie-05] Y. L. Xie, X. Zhou, Z. Q. Yao, J. X. Chen, and M. H. Liu. University of science and technology of china SSIP laboratory NIST SRE 2005 system. In NIST SRE Workshop, June 2005.

[Youn-05] S. Young, J. Odell et d'autres, "The HTK Book Version 3.3.", Speech group, Engineering Department, Cambridge University, Sydney (Australia), 775–784. 58, Avril 2005.

Annexes

Annexe A

Algorithmes Forward, Backward et Forward-Backward

A.1 Algorithme Forward (Estimation directe)

L'algorithme Forward est utilisé pour estimer la probabilité d'émission d'une séquence d'observations O par un HMM de paramètres λ . Cet algorithme de programmation dynamique permet grâce à une factorisation des chemins de réduire considérablement le nombre d'opérations et de réaliser efficacement le calcul de la probabilité d'émission.

Soit $\alpha_t(i) = P(O_1, O_2, \dots, O_t, q_t = s_i / \lambda)$ la probabilité d'avoir généré la séquence $O = o_1, o_2, \dots, o_t$ et d'être arrivé sur l'état s_i à l'instant t . Cette probabilité est calculée d'une manière récursive.

$$\textbf{Initialisation:} \quad \alpha_1(i) = \pi_i b_i(O_1) \quad 1 \leq i \leq N \quad (\text{A.1})$$

Induction : Pour t allant de 1 à T ,

$$\begin{array}{l} \vdots \\ \text{Pour } j \text{ allant de } 1 \text{ à } N \\ \vdots \\ \alpha_{t+1}(j) = \left[\sum_{i=1}^N \alpha_t(i) a_{ij} \right] b_j(o_{t+1}) \\ \vdots \\ \text{Fin} \\ \vdots \\ \text{Fin} \end{array} \quad (\text{A.2})$$

$$\textbf{Terminaison:} \quad P(O / \lambda) = \sum_{i=1}^N \alpha_T(i) \quad (\text{A.3})$$

Cet algorithme est appelé Forward car l'induction est réalisée en avant (figure A.1) : nous calculons tout d'abord la probabilité de générer le premier symbole de la séquence, puis à chaque étape de l'induction nous rajoutons un symbole et nous réitérons la procédure jusqu'à ce que l'on soit calculée la probabilité de génération de la séquence entière.

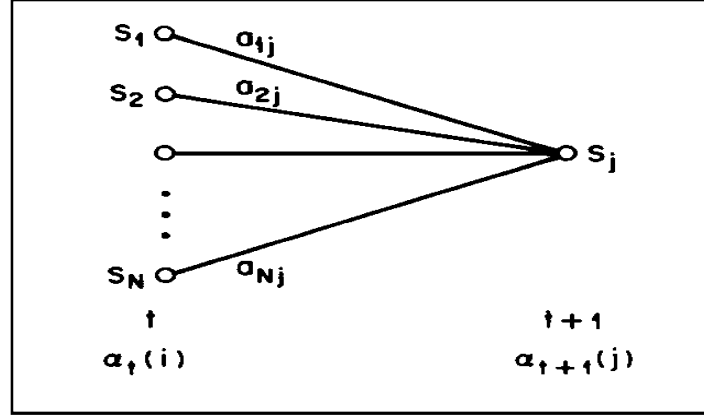


Figure A.1: Progression de la procédure Forward.

A.2 Algorithme Backward (Estimation rétrograde)

Un algorithme similaire, l'algorithme backward (figure A.2), peut être utilisé pour réaliser ce calcul à l'envers. Nous utilisons alors la variable backward $\beta_t(i) = P(o_{t+1} \dots o_T / q_t = i, \lambda)$ qui exprime la probabilité de générer la séquence $O = o_1, o_2 \dots o_t$ en partant de l'état s_i . L'induction suit alors le schéma :

Initialisation: $\beta_T(i) = 1 \quad 1 \leq i \leq N$ (A.4)

Induction :

Pour t allant de T à 1 ,

 Fin

Pour i allant de 1 à N

$$\beta_t(i) = \sum_{j=1}^N a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$$

 Fin

.....
 Fin

(A.5)

Terminaison : $P(O / \lambda) = \sum_{i=1}^N \pi_i b_i(o_1) \beta_1(i)$ (A.6)

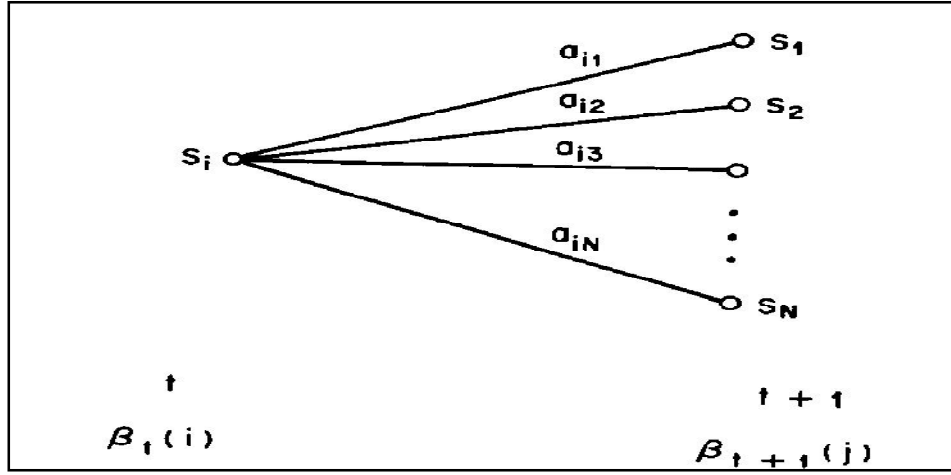


Figure A.2: Progression de la procédure Backward.

A.3 Algorithme de Baum-Welch (forward-backward)

L'algorithme de Baum-Welch [Baum-70] est un algorithme d'entraînement itératif qui permet de maximiser la probabilité de génération d'une séquence $O = \{o_1, o_2, \dots, o_T\}$ par un modèle de Markov caché de paramètres $\lambda = (N, A, B, \pi)$. C'est une forme généralisée de l'algorithme Expectation-Maximisation (EM) qui a été introduit par Dempster et al [Demp-77]. Il s'agit d'une procédure de ré-estimation itérative permettant l'augmentation de la vraisemblance des paramètres a_{ij} et $b_j(k)$ qui affinent le modèle $\tilde{\lambda}_n$ en suivant les étapes suivantes :

- Initialiser le modèle $\lambda_0(A_0, B_0, \pi_0)$.
- Calculer $\tilde{\lambda}_{n+1}$ à partir de $\tilde{\lambda}_n$.
- Répéter ce processus jusqu'à un critère de fin en vérifiant que :

$$P(O \mid \tilde{\lambda}_{n+1}) \geq P(O \mid \tilde{\lambda}_n) \quad (\text{A.7})$$

Le meilleur modèle trouvé, est enregistré pour former une base d'apprentissage (figure A.3).

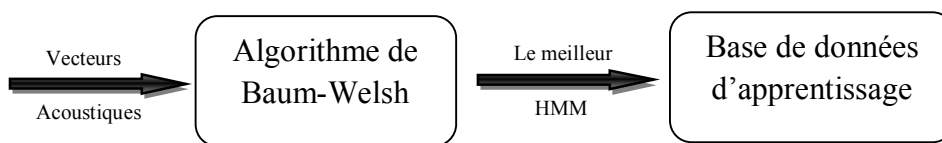


Figure A.3: Schéma représentatif de données d'apprentissage

Pour estimer les nouveaux paramètres du modèle, l'algorithme de Baum-Welch utilise deux nouvelles variables :

- $\gamma_t(i)$ représente la probabilité d'être dans l'état s_i au temps t , sachant le modèle λ et la séquence d'observation O :

$$\gamma_t(i) = P(q_t = s_i \mid O, \lambda) \quad (\text{A.8})$$

- $\xi_t(i, j)$ la probabilité d'être dans l'état i au temps t et dans l'état s_j au temps $t+1$, sachant le modèle λ et la séquence d'observation O :

$$\xi_t(i, j) = P(q_t = s_i, q_{t+1} = s_j \mid O, \lambda) \quad (\text{A.9})$$

Le calcul de ces probabilités est rapidement réalisé grâce aux variables α et β délivrées par les algorithmes Forward et Backward (figure A.4).

$$\xi_t(i, j) = \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O \mid \lambda)} \quad (\text{A.10})$$

$$\gamma_t(i) = \frac{\alpha_t(i) \beta_t(j)}{P(O \mid \lambda)} = \frac{\sum_{j=1}^N \alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O \mid \lambda)} = \sum_{j=1}^N \frac{\alpha_t(i) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{P(O \mid \lambda)} = \sum_{j=1}^N \xi_t(i, j) \quad (\text{A.11})$$

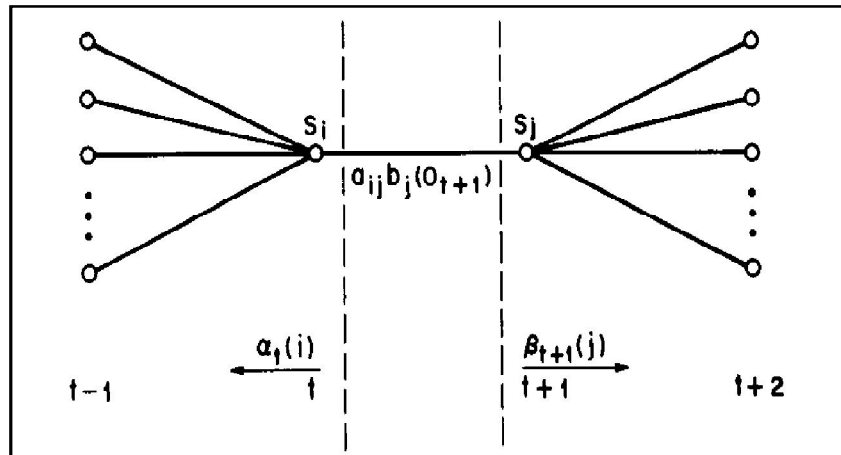


Figure A.4 : Progression de la procédure de Baum-Welch

La somme de $\gamma_t(i)$ et $\xi_t(i, j)$ au cours de temps t , peut être interprétée comme :

- Le nombre de transitions depuis l'état i :

$$\sum_{t=1}^{T-1} \gamma_t(i) = \sum_{t=1}^{T-1} P(q_t = s_i \mid O, \lambda) \quad (\text{A.12})$$

- Le nombre de transitions depuis l'état i vers l'état j .

$$\sum_{t=1}^{T-1} \xi_t(i, j) = \sum_{t=1}^{T-1} P(q_t = s_i, q_{t+1} = s_j \mid O, \lambda) \quad (\text{A.13})$$

Ainsi, la ré-estimation des nouveaux paramètres $\bar{\lambda} = (\bar{A}, \bar{B}, \bar{\pi})$ du modèle Markovien $\lambda = (N, A, B, \pi)$ est telle que :

$$\bar{\pi}_i = \text{probabilité d'être à l'état } s_i \text{ à l'instant } (t=1) = \gamma_1(i) \quad (\text{A.14})$$

$$\bar{a}_{ij} = \frac{\text{nombre de transitions de l'état } s_i \text{ vers l'état } s_j}{\text{nombre de transitions depuis l'état } s_i} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (\text{A.15})$$

- Dans le cas d'une modélisation discrète des observations :

$$\bar{b}_j(k) = \frac{\text{nombre d'observations du symbole } v_k \text{ dans l'état } s_j}{\text{nombre de passages par l'état } s_j} = \frac{\sum_{t=1}^T \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)} \quad (\text{A.16})$$

- Dans le cas d'une modélisation continue des observations, le coefficient de pondération c_{jk} , le vecteur de moyenne μ_{jk} et la matrice de covariance Σ_{jk} de la densité de probabilité associée au $k^{\text{ème}}$ mélange de l'état j sont recalculés comme suit:

$$\bar{c}_{jk} = \frac{\sum_{t=1}^T \gamma_t(j, k)}{\sum_{t=1}^T \sum_{k=1}^M \gamma_t(j, k)} \quad (\text{A.17})$$

$$\bar{\mu}_{jk} = \frac{\sum_{t=1}^T \gamma_t(j, k) \cdot o_t}{\sum_{t=1}^T \gamma_t(j, k)} \quad (\text{A.18})$$

$$\bar{\Sigma}_{jk} = \frac{\sum_{t=1}^T \gamma_t(j, k) \cdot (o_t - \bar{\mu}_{jk})(o_t - \bar{\mu}_{jk})'}{\sum_{t=1}^T \gamma_t(j, k)} \quad (\text{A.19})$$

Où $\gamma_t(j, k)$ est une généralisation de terme $\gamma_t(j)$, il est défini comme la probabilité d'être à l'état j au temps t pour le $k^{\text{ème}}$ mélange de gaussienne :

$$\gamma_t(j, k) = \left[\frac{\alpha_t(j) \beta_t(j)}{\sum_{j=1}^N \alpha_t(j) \beta_t(j)} \right] \left[\frac{c_{jk} N(o_t, \mu_{jk}, \Sigma_{jk})}{\sum_{k=1}^M c_{jk} N(o_t, \mu_{jk}, \Sigma_{jk})} \right] \quad (\text{A.20})$$

Il a été démontré par Baum et ses collègues [Baum-68] que $P(O|\bar{\lambda}) \geq P(O|\lambda)$.

Annexe B

Construction d'un système de RAL sous HTK

Dans cette annexe nous présentons les étapes de conception d'un système de reconnaissance automatique du locuteur en utilisant l'outil HTK (Hidden Markov model Toolkit). L'implémentation de ce système est réalisée par un programme écrit en langage script *Perl* sous le système d'exploitation *Windows*. Les étapes de la reconnaissance d'un locuteur par HTK sont :

- Préparation des données acoustiques.
- Description des modèles de Markov.
- Apprentissage des modèles HMMs avec le corpus d'entraînement.
- Reconnaissance : définir le réseau lexico-syntaxique à suivre, puis effectuer la reconnaissance et l'évaluation des performances sur un corpus de test.

B.1 Préparation des données

L'étape de préparation de données consiste à fournir les données acoustiques et lexicales (dictionnaire) nécessaires pour l'étape d'entraînement et l'étape de test.

B.1.1 Paramètres acoustiques

Les paramètres acoustiques retenus pour modéliser le signal parole sont 12 coefficients MFCC plus leurs dérivées premières. Les fichiers de paramètres sont calculés par l'outil **HCop**y qui prend comme entrée deux fichiers (Figure B.1).

- Le premier fichier indique la liste des fichiers parole représentatifs des locuteurs à paramétrer et l'emplacement souhaité pour les stocker (Figure B.2).

- Le deuxième fichier sert à mentionner la configuration acoustique souhaitée. L'outil Hcopy rend les fichiers de paramètres calculés dans l'emplacement indiqué (Figure B.3).

Le calcul des paramètres MFCC peut être effectué avec la commande perl 'system' comme suit :

```
system("HCOPY -C $fich_config_mfcc $fich_liste_chemins ")
```

(B.1)

Avec :

\$ fich_config_mfcc : est le fichier de configuration des paramètres acoustiques.

\$ fich_liste_chemins : est le fichier qui contient la liste des fichiers parole représentatifs des locuteurs et des coefficients MFCC.

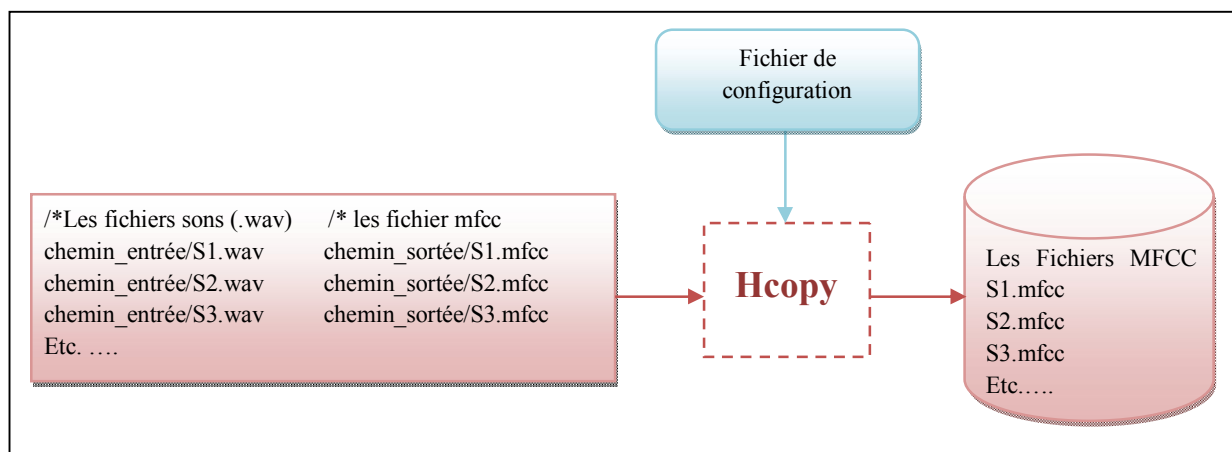


Figure B.1 : Le calcul des paramètres acoustiques par l'outil Hcopy.

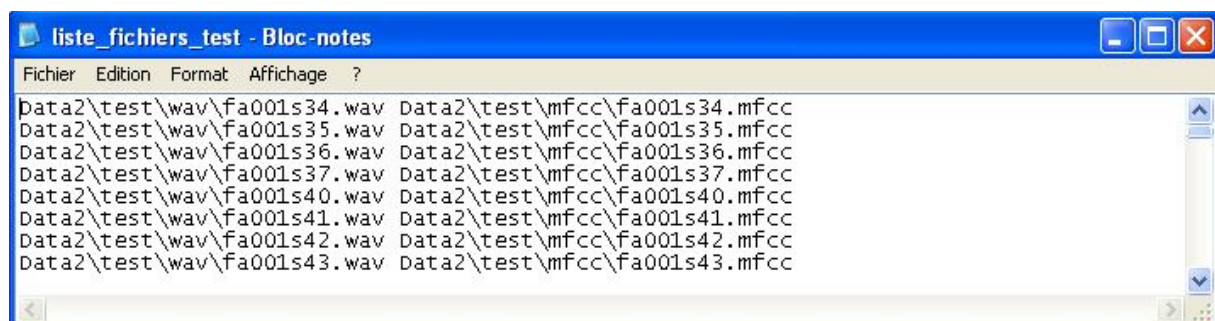


Figure B.2 : Exemple d'un fichier de liste de fichiers à paramétrer et leur emplacement de stockage.

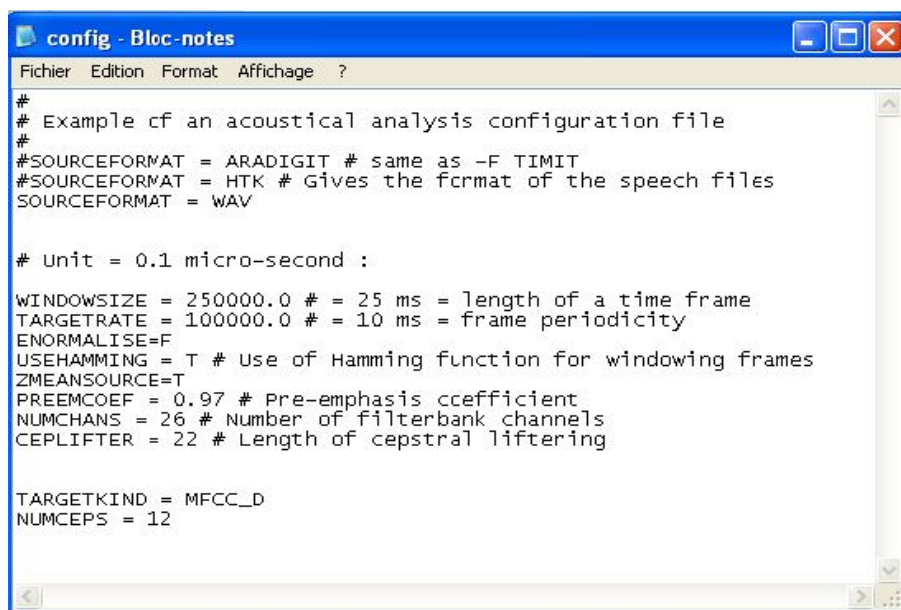


Figure B.3 : Exemple d'un fichier de configuration pour l'outil HTK Hcopy.

La figure B.4 représente le dossier contenant les fichiers .mfcc issus de l'étape de paramétrisation.

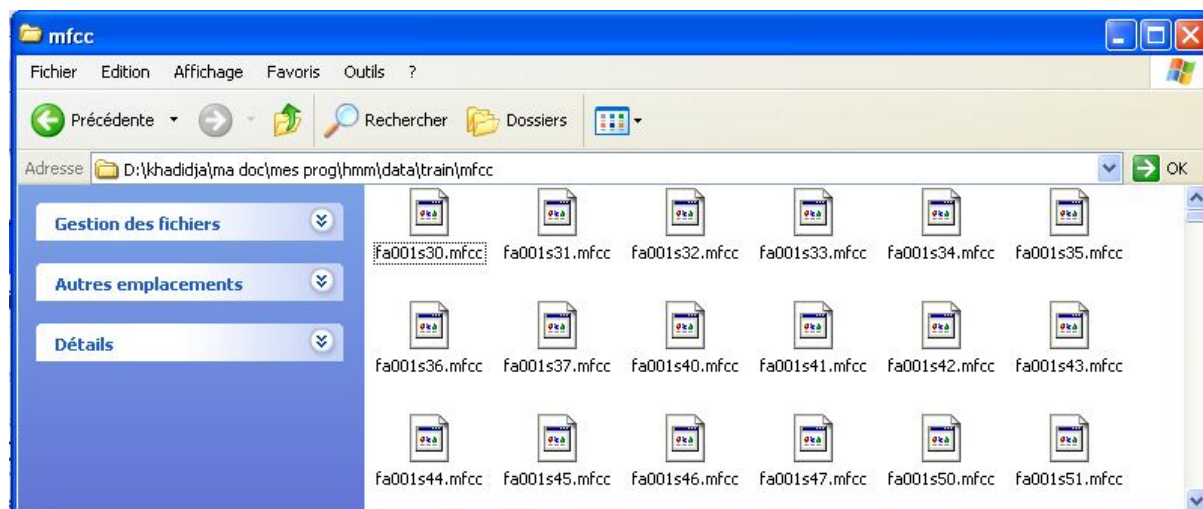


Figure B.4 : Dossier qui contient les coefficients .mfcc.

B.1.2 Création du dictionnaire

La première étape dans la création du dictionnaire (Figure B.6) est de créer une liste ordonnée de tous des noms de locuteurs appartenant à la base de données appelée grammaire (Figure B.5).

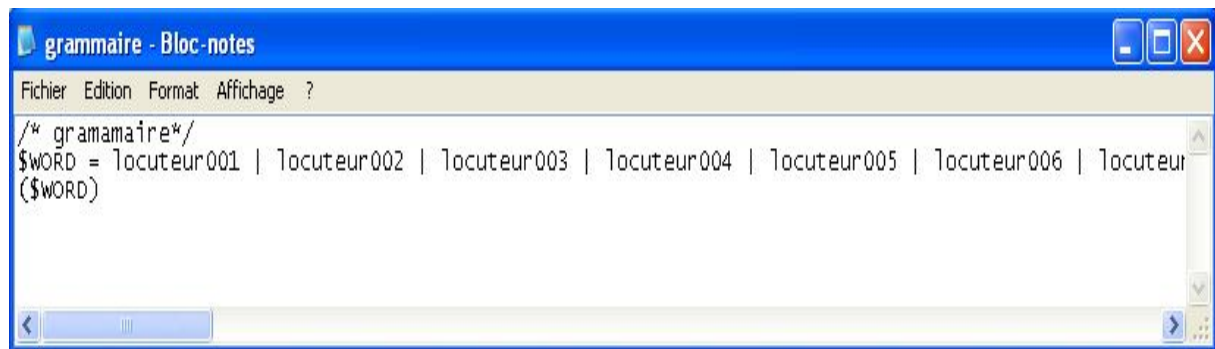


Figure B.5 : Liste des noms de locuteurs de la base de données.

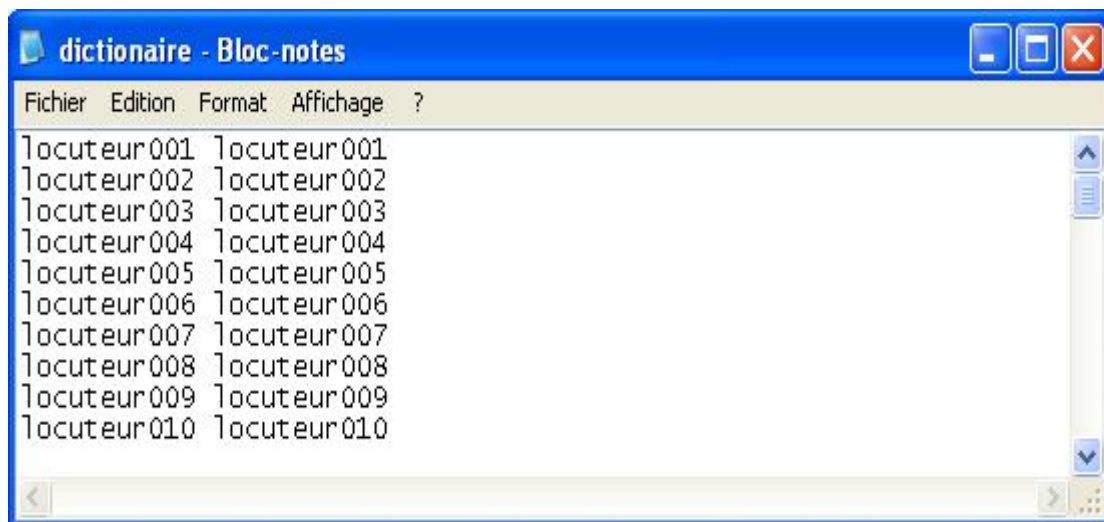


Figure B.6 : Dictionnaire de locuteurs de la base de données.

B.2 Création des modèles HMM prototypes

Une fois le dictionnaire et la grammaire définis, nous passons à la description des modèles de Markov cachés. Pour chaque locuteur, un modèle associé est défini (Figure B.7). Pour cela, nous donnerons la topologie de chaque modèle, le nombre d'états et les probabilités de transition entre les états. La topologie HMM choisie est de type gauche-droit à trois états émetteurs avec une gaussienne à matrice de covariance diagonale de dimension 24 (12 coefficients MFCC plus leurs dérivées premières). La moyenne est initialisée à 0 et la variance à 1.

```

~o <VecSize> 24 <MFCC_D>
~h "locuteur001"
<BeginHMM>
<NumStates> 5
<State> 2 <NumMixes> 1
<Mixture> 1 1.0000
<Mean> 24
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
<Variance> 24
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
<State> 3 <NumMixes> 1
<Mixture> 1 1.0000
<Mean> 24
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
<Variance> 24
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
<State> 4 <NumMixes> 1
<Mixture> 1 1.0000
<Mean> 24
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
<Variance> 36
1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
<TransP> 5
0.0 1.0 0.0 0.0 0.0
0.0 0.6 0.4 0.0 0.0
0.0 0.0 0.6 0.4 0.0
0.0 0.0 0.0 0.6 0.4
0.0 0.0 0.0 0.0 0.0
<EndHMM>

```

Figure B.7 : Modèle HMM prototype d'un locuteur.

B.3. Apprentissage

Lors de la phase d'apprentissage les paramètres du modèle HMM seront ré-estimés. L'apprentissage est réalisé en deux étapes majeures: l'initialisation et la ré-estimation. La phase d'initialisation permet de mettre à jour les moyennes, les variances et les probabilités de transition entre états jusqu'à ce qu'un seuil de convergence ou qu'un nombre maximum d'itération soient atteint. Ceci est fait par l'algorithme de Viterbi. Pour initialiser les modèles HMM, nous appliquons l'outil **HInit** (ligne de commande B.2), la figure (B.8) représente le fichier résultat de cette commande.

HInit (\$liste_hmm,\$rep_hmms_inities,\$rep_prototypes,\$rep_mfcc) (B.2)

```

~o
<STREAMINFO> 1 24
<VECSIZE> 12<NULLD><MFCC_D><DIAGC>
~h "Loc001"
<BEGINHMM>
<NUMSTATES> 5
<STATE> 2
<MEAN> 12
-1.904085e+001 4.508417e+000 -6.964932e-001 7.493671e-001 -1.102431e+000 -4.810043e-001
<VARIANCE> 12
4.535616e+001 3.705126e+001 4.547272e+001 1.828408e+001 6.717820e+001 3.860729e+000
<GCONST> 6.338995e+001
<STATE> 3
<MEAN> 12
-6.836818e+000 -2.443465e+000 -4.635944e+000 -2.743177e+000 -6.129978e+000 -7.54301e-001
<VARIANCE> 12
5.042480e+001 7.384899e+001 5.077778e+001 3.927658e+001 5.391684e+001 4.188755e+000
<GCONST> 6.910456e+001
<STATE> 4
<MEAN> 12
-2.097432e+000 -1.470220e+000 -1.101946e+001 -1.354867e+001 9.577645e-001 3.478635e-001
<VARIANCE> 12
9.810580e+000 2.045344e+001 1.663062e+001 3.005595e+001 1.789572e+001 3.906615e+000
<GCONST> 5.848419e+001
<TRANSP> 5
0.000000e+000 1.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000
0.000000e+000 9.163637e-001 8.363637e-002 0.000000e+000 0.000000e+000
0.000000e+000 0.000000e+000 9.698163e-001 3.018373e-002 0.000000e+000
0.000000e+000 0.000000e+000 0.000000e+000 9.148148e-001 8.518518e-002
0.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000
<ENDHMM>

```

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

Par la suite, les modèles HMM sont ré estimés de façon indépendante avec l’algorithme de Baum Welch en utilisant la commande **HRest** (ligne de commande B.3), la figure (B.9) représente le fichier résultat de cette commande.

HRest(\$liste_hmm,\$rep_hmms_entraines,\$rep_hmms_inities,\$rep_mfcc) (B.3)

```

~0
<STREAMINFO> 1 24
<VECSIZE> 12<NULLD><MFCC_D><DIAGC>
~h "Loc001"
<BEGINHMM>
<NUMSTATES> 5
<STATE> 2
<MEAN> 12
-1.906903e+001 4.529743e+000 -6.849962e-001 8.035937e-001 -1.077268e+000 -4.779068e-001
<VARIANCE> 12
4.506742e+001 3.680485e+001 4.563195e+001 1.803404e+001 6.725626e+001 3.855447e+000
<GCONST> 6.341635e+001
<STATE> 3
<MEAN> 12
-6.825397e+000 -2.427993e+000 -4.663682e+000 -2.795325e+000 -6.121409e+000 -7.55591e-001
<VARIANCE> 12
5.033836e+001 7.381094e+001 5.077445e+001 3.953537e+001 5.371869e+001 4.178919e+000
<GCONST> 6.909997e+001
<STATE> 4
<MEAN> 12
-2.088565e+000 -1.508272e+000 -1.100261e+001 -1.354951e+001 9.891163e-001 3.597015e-001
<VARIANCE> 12
9.776853e+000 2.010918e+001 1.668732e+001 3.016451e+001 1.791739e+001 3.775498e+000
<GCONST> 5.840616e+001
<TRANSP> 5
0.000000e+000 1.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000
0.000000e+000 9.161438e-001 8.385620e-002 0.000000e+000 0.000000e+000
0.000000e+000 0.000000e+000 9.699461e-001 3.005391e-002 0.000000e+000
0.000000e+000 0.000000e+000 0.000000e+000 9.139960e-001 8.600402e-002
0.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000 0.000000e+000
<ENDHMM>

```

Figure B.9 : Fichier résultat de la commande HRest.

B.4 Reconnaissance

La reconnaissance consiste à comparer les paramètres acoustiques représentatifs du locuteur à identifier avec ceux de la base de référence en utilisant successivement les modèles issus de **HInit** et ceux issus de **HRest**. Pour cela nous allons utiliser la commande **HVite** (ligne de commande B.5)

```
HVite -d$rep_hmms_entraines-i $reco-w $reseau_syntaxique-S$liste_test_mfcc$dictionnaire
$liste_hmm
```

(B.4)

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

```
HRESULTS -p -I $fich_ref$liste_hmm $reco> $resultats
```

(B.5)

Annexe C

Le corpus Noisex-92

L'objectif de cette annexe est de présenter brièvement le corpus de bruit Noisex-92 avec lequel nous avons travaillé lors de ce mémoire. Ce corpus a été mis au point, en 1992, à partir du corpus Noise-Rom-0 par l'Institut TNO pour l'étude de la perception et par l'équipe de recherche sur la parole de la *Defense Research Agency* anglaise.

Dans cette annexe, nous avons choisi de classer les différents bruits en fonction de leurs caractéristiques de stationnarité plutôt que de les classer selon leurs origines (synthétique, civile ou militaire). Nous avons ainsi défini trois classes : celle des bruits stationnaires, celle des bruits rythmiques et celle des bruits aléatoires.

C.1 Bruits stationnaires ou quasi stationnaires

Les bruits stationnaires de Noisex-92 sont au nombre de quatre. Ils présentent des difficultés diverses et plus ou moins importantes.

- Le bruit de voiture (bruit d'une voiture Volvo 340 à 120 kilomètres à l'heure sur une route goudronnée et en 4ème vitesse) n'est pas fondamentalement problématique à de forts Rapports Signal sur Bruit (RSB). Il peut cependant devenir très inconfortable pour un système de RAL à de faibles RSB puisque l'énergie de ce bruit est concentrée dans les basses fréquences, fréquences où se trouvent les formants des voyelles (figure C.1).

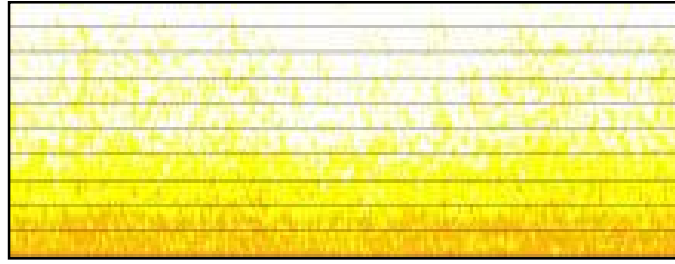


Figure C.1 : Spectrogramme du bruit de voiture.

- Le bruit synthétique de parole (*speech noise*) présente la même répartition spectrale que le bruit de voiture mais est cependant plus énergétique (figure C.2). Ce bruit, plus que le bruit de la voiture Volvo, peut présenter de réels inconvénients pour un système de RAL puisque l'énergie de ce bruit est concentrée en basses fréquences, en dessous de 4 MHz. Il est cependant loin d'avoir le caractère problématique du *voice babble*, qui est un véritable bruit de parole (figure C.7).

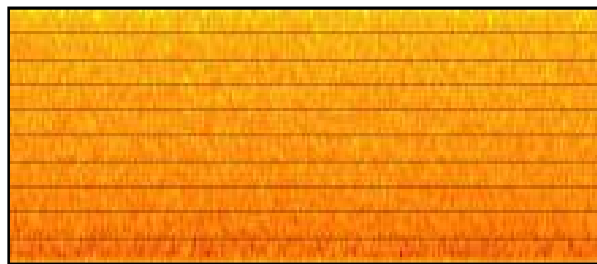


Figure C.2 : Spectrogramme du bruit synthétique de parole.

Les deux bruits que nous venons de présenter peuvent nettement amoindrir les capacités des systèmes de RAL mais possèdent des répartitions spectrales assez différentes de celles de la parole véritable. D'autres bruits stationnaires possèdent des caractéristiques qui les font plus ressembler à de la parole, tout au moins à des voyelles : ces bruits possèdent des traces formantiques dans leur spectre.

- Le premier de ces deux bruits qui rentrent dans la catégorie des bruits stationnaires de nature formantique est le bruit de l'avion F16. Ce bruit a été enregistré en place arrière d'un F16 biplace volant à une vitesse de 500 noeuds et à une altitude comprise entre 300 et 600 pieds. Ce bruit est, comme le montre son spectrogramme (figure C.3), de nature formantique. Ces formants correspondent à la vitesse de fonctionnement des organes moteur de l'avion, en l'occurrence du réacteur. Ces formants, en se

superposant à la parole, peuvent créer l'illusion d'une présence de voyelles pour certains systèmes. Leur stationnarité, dans un régime de vol donné, permet cependant de les isoler assez facilement.

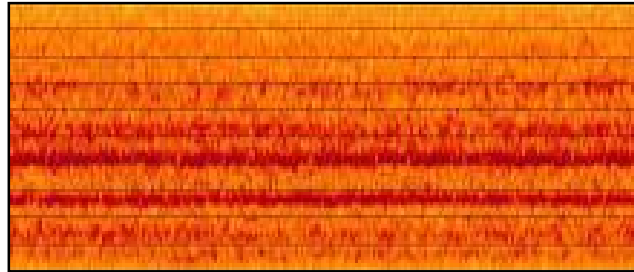


Figure C.3 : Spectrogramme du bruit de l'avion F16.

- Le deuxième bruit stationnaire de nature formantique est le bruit de l'hélicoptère Lynx enregistré sur une plate-forme. Les caractéristiques générales de ce bruit (figure C.4) sont similaires aux caractéristiques du bruit du F16 (figure C.3). Les organes moteurs fonctionnant cependant moins vite, les traces formantiques sont de fréquences moins élevées. Cette dernière caractéristique fait que le bruit de l'hélicoptère Lynx dégradera plus facilement les voyelles que ne le fait le bruit de l'avion F16.

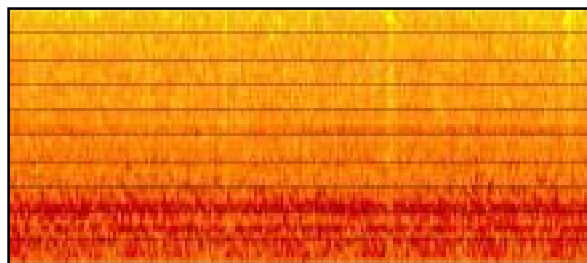


Figure C.4 : Spectrogramme du bruit de l'hélicoptère Lynx.

C.2 Bruits rythmiques

- Le premier bruit rythmique de Noisex-92 correspond au bruit enregistré dans une salle d'opérations d'un contre-torpilleur. Ce bruit, dont le spectrogramme est présenté dans la figure B.5, présente un rythme simple qui est visible dans la succession des raies spectrales. Ce bruit est assez intense mais l'énergie des raies n'est pas beaucoup plus importante que l'énergie du bruit de fond. Il est, en outre, possible d'observer la présence de bruits aléatoires supplémentaires en basses fréquences, entre 0 et 3000 Hz. Ce bruit est cependant très diffus.

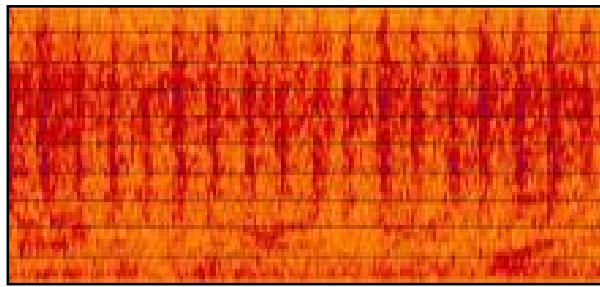


Figure C.5 : Spectrogramme du bruit d'une salle d'opérations d'un contre-torpilleur.

- Un deuxième bruit rythmique présent dans la base Noisex-92 est le bruit de test du bateau STITEL (figure C.6). Le rythme est ici établi sur quatre bandes de fréquences différentes. Ces bandes s'échelonnent respectivement de 1000 à 1500 Hz, de 1500 à 3000 Hz, de 3000 à 5500 Hz et de 5500 à 10000 Hz. Elles constituent un bruit qu'il serait possible de traiter aisément avec un système ayant appris les différents intervalles de temps de présence et d'absence d'énergie dans ces quatre bandes de fréquences.

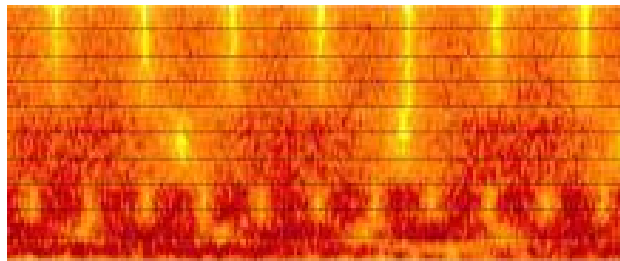


Figure C.6 : Spectrogramme du bruit STITEL ship test signal.

C.3 Bruits aléatoires

Rappelons tout d'abord que nous qualifions ici des bruits aléatoires, les bruits générés par des processus aléatoires de haut niveau, ces processus pouvant correspondre à une discussion entre plusieurs personnes, à une séance de travail en atelier ou dans un bureau. Ces bruits ne correspondent donc pas à des signaux temporels générés aléatoirement, cette génération aléatoire étant un processus de bas niveau.

- Le bruit de parole voice babble, appelé également bruit de cocktail party, est un bruit qui génère de grandes difficultés dans les systèmes de RAL. Il a été enregistré dans un réfectoire où se trouvait une centaine de personnes. Le bruit est donc composé de

différents signaux de parole se superposant plus ou moins les uns aux autres en fonction de la distance du locuteur au microphone à un instant donné. Ce bruit devient très problématique lorsque le rapport signal sur bruit devient négatif. Il présente en effet strictement les mêmes caractéristiques que le signal de parole à analyser et peut donc conduire à de très nombreuses erreurs. Ces bruits semblent difficiles à traiter avec les techniques actuelles de la RAL puisque l'homme, en tant qu'auditeur, semble utiliser toutes ses facultés pour surmonter ce type de problème avec l'utilisation de la binauralité et de la connaissance du timbre de la voix du locuteur, connaissance qui permet d'effectuer un "suivi" de la voix.

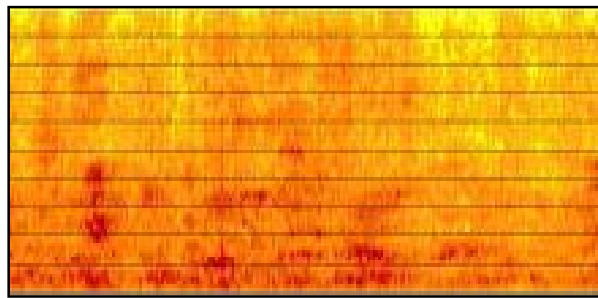


Figure C.7 : Spectrogramme du bruit de cocktail party.

- Les bruits associés à des ateliers de fabrication ou à des usines peuvent également générer de nombreuses erreurs dans les systèmes de RAL. L'univers de la production est généralement très bruyant et le bruit, qui en constitue l'environnement sonore, est généralement très élevé. Au niveau de la caractérisation de ces bruits, il faut tout d'abord considérer le cas des bruits rythmiques correspondant aux automates mis en place dans une unité de production. Mais ces bruits ne sont pas les seuls.

L'ouvrier, qui est avant tout un être humain, ne travaille pas en suivant le rythme d'une horloge à la manière d'un robot et produit donc des bruits à intervalles irréguliers. Ce fait peut conduire à la génération de bruit tel que celui dont le spectrogramme est donné en figure C.8. Ce bruit correspond à l'enregistrement des bruits de soudures électriques dans un atelier de montage de bas de caisse de voitures. Des bruits de friction (raies verticales) correspondant aux instants de soudure y sont visibles, à intervalles irréguliers.

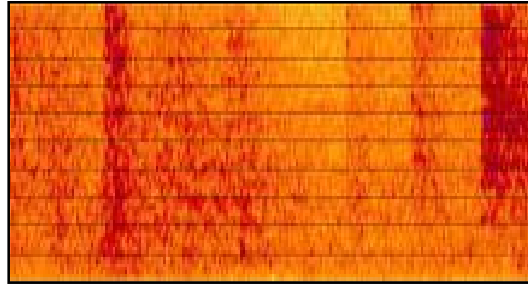


Figure C.8 : Spectrogramme de bruits de soudures électriques dans un atelier de montage de bas de caisse.

- Le dernier bruit que nous avons classé dans les bruits aléatoires est le bruit de rafales de mitrailleuse. Ce bruit est aléatoire car, comme nous pouvons le voir sur le spectrogramme de la figure C.9, l'intervalle de temps entre deux détonations successives peut, approximativement, varier du simple au triple. Ce bruit est également très problématique car il peut créer pendant de courts instants de très fortes perturbations dans le signal de parole.

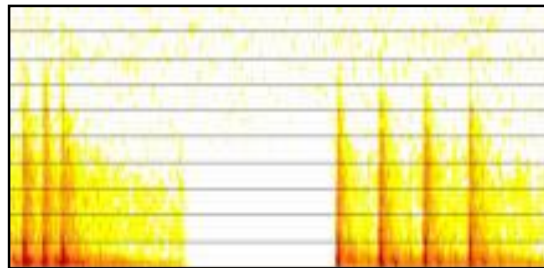


Figure C.9 : Spectrogramme du bruit de rafales de mitrailleuse.