REPUPLIQUE ALGERIENNE DEMOCRATIQUE ET POLPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

ECOLE NATIONALE POLYTECHNIQUE

Département d'Electronique

Spécialité Electronique

Projet de fin d'études

En vue de l'obtention du diplôme d'Ingénieur d'Etat en Electronique

THEME

Identification du locuteur en mode indépendant du texte

Proposé et dirigé par :

M. B.BOUSSEKSOU

Etudié par :

M. Samer MECHHOUD

Juin 2010

E.N.P. 10, Avenue Hassen Badi, El Harrach, Alger

ملخص

يندرج هذا العمل ضمن التشخيص الأوتوماتيكي للمتكلم، هذا الميدان الغني بالتطبيقات البالغة الأهمية بدءا من تامين المعابر و التطبيقات القضائية إلى تنظيم الملفات الصوتية. حتى نترك المجال مفتوح، اهتممنا بالتشخيص الأوتوماتيكي للمتكلم المستقل عن النص المنطوق. اهتممنا بكيفية تمثيل المتكلمين و استخراج الشارات الصوتية للمتكلمين حتى نتمكن من تطبيقها في نظام تشخيص المتكلم.

كلمات مفتاحية : تشخيص المتكلم، الفضاء الصوتي، وسائط مال سابستر، الوسائط الطيفية الخطية، التكميم الشعاعي، النموذج متعدد الغوصيات، شبكة العصبونات الاصطناعية.

Résumé

Ce travail s'inscrit dans le domaine de la reconnaissance automatique du locuteur, domaine riche d'applications potentielles allant de la sécurisation d'accès et les applications d'ordre juridique à l'indexation de documents audio. Afin de laisser le champ à un large éventail d'applications, nous nous intéressons à l'identification du locuteur en mode indépendant du texte. Nous nous intéressons, plus particulièrement, à la modélisation et à la représentation des locuteurs. Il s'agit d'extraire, à partir des signaux de parole, les informations relatives à l'identité du locuteur et d'estimer un modèle du locuteur permettant son identification.

Mots clés : identification du locuteur, espace acoustique, MFCC, LSP, quantification vectorielle QV, modèles de mélanges des gaussiennes GMM, réseaux de neurones artificiels RNA.

Abstract

This work relates to the automatic speaker recognition which has many potential applications ranging from access security to audio indexing. In this thesis, the text-independent speaker identification is studied with a specific focus on speaker modeling and representation. We are especially interested to extract, from speech signals, the relative information of the speaker identity and estimate a sufficiently robust speaker's model.

Keywords: speaker identification, acoustic space, Mel frequency cepstral coefficients MFCC, line spectral pair LSP, vector quantization VQ, Gaussian mixture models GMM, artificial neural networks ANN.

Remerciements

Je tiens à remercier en premier lieu « ALLAH » le tout puissant, qui m'a donné la force, le courage et la volonté pour mener à bien ce modeste travail.

J'exprime ma profonde gratitude, mon grand respect et ma sincère reconnaissance à mon promoteur monsieur B. BOUSSEKSOU pour avoir assumé la lourde responsabilité de m'encadrer, de m'avoir orienté et conseiller tout au long de ce travail ainsi pour la confiance qu'il m'a accordée.

Je remercie l'ensemble de mes enseignants d'Electronique de l'Ecole Nationale Polytechnique.

Je remercie vivement tous mes enseignants et encadreurs de l'Ecole Nationale Préparatoire aux Etudes d'Ingéniorat.

Finalement, je remercie toute personne qui m'a soutenu de près ou de loin tout au long de mon parcours pour la réalisation de ce travail.

Dédicaces

A mes très chers parents

A mes sœurs

A mes frères

A ma tante

A tous mes amis

Table des matières

Chapitre 4 Evaluation expérimentale

Liste des Figures

Liste des Tableaux

Liste des acronymes

AR : **A**uto **R**égressif.

DCT: **D**iscrete **C**osine **T**ransform.

DTW: **D**ynamic **T**ime **W**arping.

EM: **E**xpectation **M**aximisation.

FFT: **F**ast **F**ourier **T**ransform.

GMM: **G**aussian **M**ixture **M**odels.

HMM: **H**idden **M**arkov **M**odels.

IAL: **I**dentification **A**utomatique du **L**ocuteur.

LBG: **L**inde **B**uzo **G**ray.

LFCC: **L**inear **F**requency **C**epstral **C**oefficients.

LPC: **L**inear **P**rediction **C**oefficients.

LPCC: **L**inear **P**rediction **C**epstral **C**oefficients.

LSP: **L**ine **S**pectral **P**air (**L**ine **S**pectral **F**requencies)

MFCC: **M**el **F**requency **C**epstral **C**oefficients.

MLP : **M**ulti-**L**ayer **P**erceptron.

MV : **M**aximum de **V**raisemblance.

PNN : **P**robabilistic **N**eural **N**etworks.

QV : **Q**uantification **V**ectorielle.

RAL : **R**econnaissance **A**utomatique du **L**ocuteur.

RNA : **R**éseau de **N**eurones **A**rtificiel.

RTC : **R**éseau **T**éléphonique **C**ommuté.

SNR: **S**ignal to **N**oise **R**atio.

TFD: **T**ransformée de **F**ourier **D**iscrète.

VAL : **V**érification **A**utomatique du **L**ocuteur.

Introduction générale

Le traitement de la parole est une composante fondamentale des sciences de l'ingénieur. Située au croisement du traitement numérique du signal et du traitement du langage, cette discipline scientifique a connu depuis les années 1960 une expansion fulgurante, liée au développement des moyens et des techniques de télécommunications.

Ce travail s'inscrit dans le domaine de la reconnaissance automatique du locuteur, domaine riche d'applications potentielles allant de la sécurisation d'accès à l'indexation des documents audio. Afin de laisser le champ à un large éventail d'applications, nous nous intéressons à l'identification du locuteur en mode indépendant du texte. Nous nous intéressons plus particulièrement à l'extraction des paramètres distinctifs et la modélisation des locuteurs.

Nous avons commencé par rappeler le principe de la reconnaissance automatique du locuteur. Cette introduction présente les différents traits distinctifs entre locuteurs (variabilité interlocuteur) qui constituent l'essence même de la reconnaissance et les conditions qu'ils doivent remplir afin de permettre une bonne discrimination.

Dans le deuxième chapitre, nous avons présenté l'extraction des vecteurs acoustiques représentant les locuteurs et les différentes étapes nécessaires pour l'analyse du signal vocal. Nous avons cités les différents paramètres acoustiques utilisés dans la majorité des systèmes d'identification du locuteur. Ce chapitre donne une idée générale sur le choix des paramètres acoustiques convenables.

Au troisième chapitre, nous nous sommes intéressés aux différentes modélisations des locuteurs, plus particulièrement à l'approche connexionniste où les locuteurs sont modélisés

par des réseaux de neurones. Cette approche a été récemment et largement utilisée dans ce domaine et fournit de très bonnes performances.

Le quatrième et dernier chapitre décrit le contexte expérimental et expose les résultats des différents tests effectués sur un échantillon de 38 locuteurs extrait de la base de données de l'Ecole Nationale Polytechnique. Nous avons essayé d'examiner et de voir l'influence d'un certain nombre de paramètres (la qualité des données d'apprentissage et de test, le nombre de coefficients acoustiques, le nombre de classes, la quantité des données d'apprentissage et les conditions de transmission à travers le réseau téléphonique commuté RTC) sur le taux d'identification correcte et sélectionner, par la suite, l'ensemble des paramètres qui donne les meilleures performances pour une éventuelle conception d'un système d'identification du locuteur.

Chapitre 1

Introduction à la reconnaissance automatique du locuteur (RAL)

La reconnaissance automatique du locuteur s'inscrit dans le domaine du traitement de la parole. Elle est interprétée comme une tâche particulière de la reconnaissance des formes. Elle exploite la variabilité interlocuteur et s'intéresse aux informations extralinguistiques du signal vocal. Ce domaine regroupe les problèmes relatifs à l'identification ou à la vérification du locuteur sur la base de l'information contenue dans le signal acoustique : il s'agit d'extraire du signal de la parole la part relative à l'identité du locuteur.

Les variations individuelles entre locuteurs ont deux origines essentielles. D'abord, les caractéristiques morphologiques de l'appareil de phonation qui diffèrent d'un locuteur à un autre, et ensuite les différences dans les débits d'élocution, l'étendue des variations du pitch ou encore les différences liées au milieu socioculturel. Cette variabilité interlocuteur est l'essence même de la reconnaissance automatique du locuteur.

L'utilité en est la possibilité de vérifier automatiquement l'identité d'une personne demandant d'accéder à des informations protégées.

1.1 Système de reconnaissance automatique du locuteur

Un système de reconnaissance automatique du locuteur comporte plusieurs modules. Tout d'abord, un module d'acquisition qui capte le signal vocal et le convertie en un signal numérique. Ensuite le module d'analyse acoustique servant à extraire des vecteurs de

coefficients pertinents pour modéliser les locuteurs. Dans l'étape d'apprentissage, un modèle est créé pour chaque locuteur. Dans l'étape de reconnaissance, un module va mesurer la similarité entre les données de test et tous les modèles de locuteurs présents dans la base. En dernier lieu, un module de décision, basé sur une stratégie de décision donnée, fournit la réponse du système.

1.1.1 Identification automatique du locuteur (IAL)

L'identification automatique du locuteur consiste à reconnaître une personne parmi un ensemble de locuteurs en comparant ses paramètres de test aux différents modèles de locuteurs présents dans la base. Dans le cas où le locuteur doit appartenir à l'ensemble des locuteurs de la base de données, on parle d'une identification en ensemble fermé (qui constitue le cadre de notre travail). Dans le cas où le système peut être amené à fournir un ensemble vide comme réponse, on parle d'une identification en ensemble ouvert. La structure d'un système d'identification automatique du locuteur est représentée sur la figure 1.1.

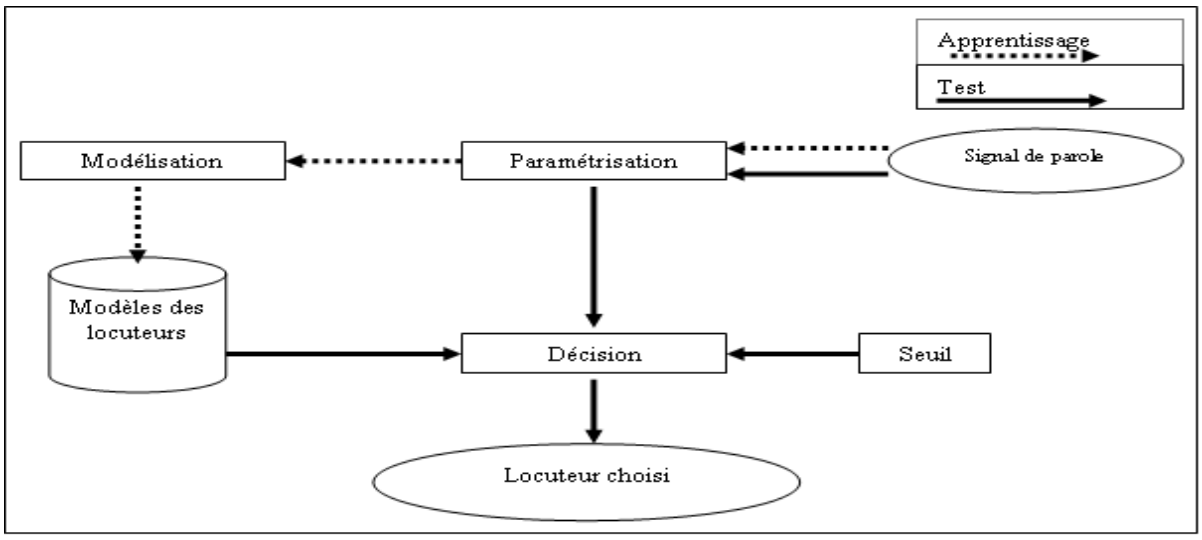

Fig. 1.1 Schéma modulaire d'un système d'IAL [2].

1.1.2 Vérification automatique du locuteur (VAL)

La vérification du locuteur consiste, après que le locuteur ait décliné son identité, à vérifier l'adéquation de son message vocal avec la référence acoustique du locuteur qu'il prétend être. La structure d'un système de vérification automatique du locuteur est représentée sur la figure 1.2.

Dans la pratique ces deux tâches, identification et vérification, s'effectuent de la même manière, seul le test final diffère .Comparaison du résultat du traitement de la donnée test avec la référence en fonction d'un seuil dans le cas de la vérification. Comparaison du test avec plusieurs références et sélection du plus proche dans le cas de l'identification.

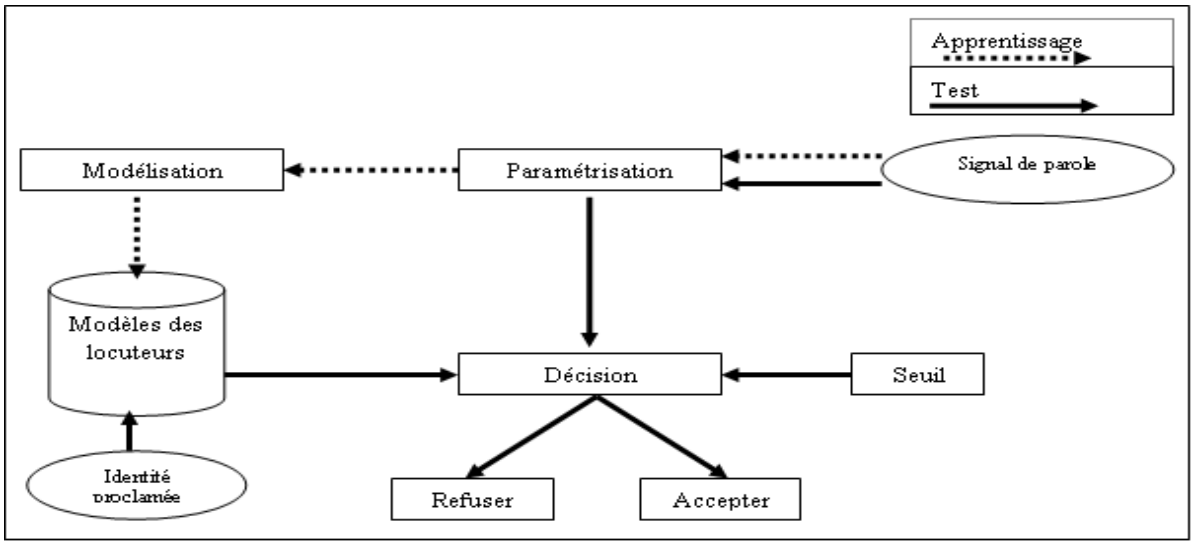

Fig. 1.2 Schéma modulaire d'un système de VAL [2].

1.2 Modes dépendant et indépendant du texte

On peut classer les systèmes de reconnaissance automatique du locuteur en deux catégories, qui correspondent aux deux modes dépendant et indépendant du texte. Les niveaux de dépendance au texte sont classés suivant les applications :

- Systèmes à texte libre : le locuteur est libre de prononcer ce qu'il veut, et les phrases d'apprentissage et de test sont différentes.
- Systèmes à texte suggéré : un texte, différent à chaque session et pour chaque locuteur, est imposé par la machine. Les phrases d'apprentissage et de test peuvent être différentes.
- Systèmes dépendants du vocabulaire : le locuteur prononce une séquence de mots issus d'un vocabulaire limité. Dans ce cas, l'apprentissage et le test sont réalisés sur des textes constitués à partir du même vocabulaire.
- Systèmes personnalisés dépendants du texte : chaque locuteur a son propre mot de passe. Dans ce mode, l'apprentissage et le test sont réalisés sur le même texte.

1.3 Variabilité intra locuteur

Le signal acoustique de la parole présente des caractéristiques qui rendent complexe son interprétation. L'information portée par ce signal peut être analysée de bien des façons et à plusieurs niveaux (acoustique, phonétique, phonologique, morphologique, syntaxique, sémantique et pragmatique), ce qui rend la tâche de traitement de la parole complexe. La variabilité intra locuteur exprime les différences dans le signal produit par une même

personne. Il existe plusieurs facteurs qui peuvent augmenter cette variabilité comme par exemple :

- L'état pathologique du locuteur (maladie, émotions,...).
- Vieillissement (la voix d'une personne change avec son âge).
- Facteurs socioculturels (le locuteur peut changer d'accent).
- Locuteurs non coopératifs (notamment dans des applications judiciaires).
- Conditions de prise de son : bruit ambiant ...etc.

1.4 Traits distinctifs du locuteur

Le premier problème qui se pose lors de la réalisation d'un système de reconnaissance du locuteur est évidemment le choix des paramètres. Il s'agit de pouvoir extraire du signal vocal les caractéristiques de la voix de chaque locuteur en sélectionnant les traits acoustiques signifiants (chargés d'une quantité suffisante d'informations sur l'identité du locuteur).

La mise en œuvre d'une tâche de reconnaissance de locuteur (ou de parole) est loin d'être facile, et ce pour deux raisons majeures. La première tient au fait que l'on ne maîtrise pas l'espace acoustique et en particulier la fonction de production d'un signal de parole. Aucune méthode analytique ne permet de prédire quelle va être la forme du signal de parole correspondant à l'émission d'un symbole donné par un locuteur particulier. La seconde est que la réalisation acoustique d'un symbole donné n'est pas unique.

1.4.1 Production de la parole

L'appareil phonatoire humain est constitué d'un organe respiratoire (source d'énergie), des cordes vocales (qui jouent le rôle d'oscillateurs) et des cavités buccales et nasales qui tiennent lieu de résonateur et anti résonateur (figure 1.3).

Dans la voix parlée, le signal est produit par une excitation acoustique du canal vocal. Ce canal peut être considéré comme un tube acoustique partant des cordes vocales (région du larynx) et allant jusqu'aux lèvres. Sur ce tube, peut se brancher en dérivation le tube constitué par les cavités nasales qui est normalement occulté et n'agit que lors de la production des lettres nasalisées, par l'ouverture du velum (palais mou).

Deux modes d'excitation du canal vocal existent :

- Par les vibrations des cordes vocales.
- Par des bruits.

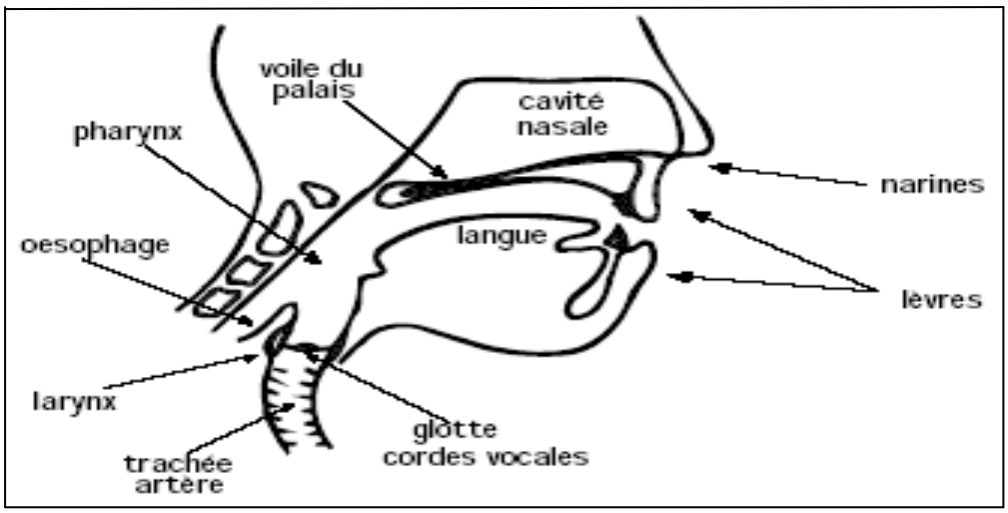

Fig. 1.3 L'appareil phonatoire humain [1].

1.4.1.1 Sons voisés

Les sons voisés ont une structure quasi-périodique (figure 1.4), ils résultent de l'excitation du conduit vocal par un train périodique d'impulsions de pression liées aux oscillations des cordes vocales. L'ouverture brusque de la glotte libère la pression accumulée en amont, elle se referme ensuite plus graduellement.

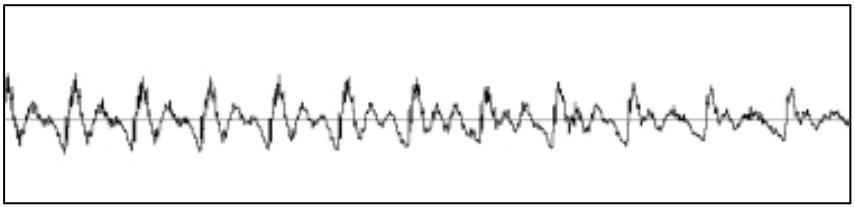

Fig. 1.4 Son voisé [1].

1.4.1.2 Sons non voisés

Le second mode d'excitation est obtenu par divers bruits produits par le passage de l'air en un point de resserrement du canal vocal ou par des bruits d'occlusion provoqués par la fermeture ou l'ouverture des lèvres, ou des chocs de la langue contre le palais. Dans cette catégorie de sons les cordes vocales ne vibrent pas. Un son non voisé ne présente pas de structure périodique (figure 1.5), il peut être approximé par la réponse du conduit vocal à un bruit blanc gaussien.

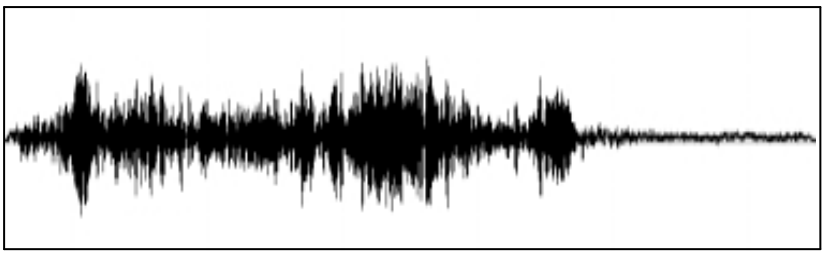

Fig. 1.5 Son non voisé [1].

1.4.2 Fréquence fondamentale

La fréquence fondamentale, appelée aussi pitch, est la fréquence de vibration des cordes vocales. La gamme de variation moyenne de la fréquence fondamentale dépend, essentiellement, de l'âge et du sexe du locuteur. Elle peut varier [1] :

- de 70 à 250 Hz chez l'homme.
- de 150 à 400 Hz chez la femme.
- de 200 à 600 Hz chez l'enfant.

Elle peut présenter des variations considérables chez un même locuteur selon le type de la phrase prononcée, son attitude et son état émotionnel.

1.4.3 Timbre

La caractéristique première de la voix d'un locuteur est son timbre, qui n'est perceptible que pour les sons voisés, et surtout les voyelles et semi-voyelles. Le timbre est déterminé par les amplitudes relatives des harmoniques du pitch. Le timbre est totalement exprimé par le spectre du signal. Or, le signal vocal est constitué de la convolution du signal glottique (excitation) et de la réponse impulsionnelle du canal vocal, son spectre est le produit du spectre de la source et de celui du canal.

1.4.3.1 Spectre de la source

Le spectre de la source est composé par une fréquence fondamentale, et un grand nombre d'harmoniques, ayant une enveloppe dont la forme est proche d'une exponentielle. Une part de l'identité du locuteur est sans doute associée à cette forme [3].

1.4.3.2 Spectre du conduit vocal

Il s'agit en fait de la fonction de transfert du canal vocal en tant que tube acoustique. L'examen des spectres du signal vocal (éventuellement lissés pour éliminer l'influence de la source) montre la présence d'un certain nombre de pics dans le spectre des voyelles. Ce sont les zones correspondant aux fréquences renforcées par les différents résonateurs couplés. Ces zones sont appelées zones formantiques ou formants [3]. Les paramètres

fréquentiels des voyelles sont liés à l'anatomie du sujet (donc à son identité), mais aussi à son état physique et émotionnel, ce qui peut gêner leur utilisation en tant que traits caractéristiques du locuteur.

1.4.4 Mélodie

La source (impulsion de glotte) est caractérisée non seulement par son spectre, mais aussi par la période de vibration. La fréquence de la source est la fréquence fondamentale, qui est le pitch. Cette fréquence n'est pas stable. Elle varie très rapidement en fonction du temps (mélodie), et porte une information sémantique au moyen des patrons intonatifs ou de la micro mélodie (évolution du fondamental d'un phonème à un autre, ou même au sein d'un même phonème). La mélodie porte également une information sur l'identité du locuteur qui apparaît dans la distribution statistique de la fréquence (pitch moyen, variance de pitch, …) et dans l'évolution temporelle de l'élocution.

1.4.5 Articulation

Les phénomènes d'articulation sont à l'origine des traits distinctifs entre locuteurs, et concernent l'activité musculaire du locuteur.

1.4.5.1 Coarticulation

La coarticulation est l'influence d'un son sur un autre son contiguë ou voisin. Le locuteur prononçant une phrase produit une suite de phonèmes qui sont enchaînés les uns des autres de façon continue, en reliant les parties stables du signal par des zones de transition. La dynamique du canal vocal représentée par les variations de la fonction de transfert est donc un ensemble de traits distinctifs du locuteur. Néanmoins elle est très variable suivant l'état physique ou émotionnel du locuteur.

1.4.5.2 Occlusives

C'est la durée du silence qui précède l'explosion dans les plosives [p], [t], [k]. Il s'agit donc d'un paramètre temporel qui est très difficile à imiter, comme la coarticulation.

1.4.5.3 Enveloppe énergétique

L'enveloppe énergétique qui est par définition la distribution de l'énergie du signal dans le domaine fréquentiel, est liée à l'identité du locuteur. On conçoit qu'il s'agisse aussi d'une donnée assez facile à imiter [3], ce qui explique qu'elle ne soit utilisée que conjointement à d'autres paramètres, moins sensible à l'imitation.

1.5 Qualité des traits distinctifs

Les conditions que doivent remplir les paramètres pour l'identification du locuteur sont [4]:

- Etre aptes à représenter l'information utile sur l'identité du locuteur.
- Etre faciles à mesurer.
- Etre stables dans le temps.
- Apparaître naturellement et fréquemment dans la parole.
- Etre peu modifiables par un changement de l'environnement.
- Ne pas être imitables.

Ce sont ces considérations qui vont guider le choix des paramètres lors de l'élaboration du système.

1.6 Domaines d'applications

Les applications des systèmes de reconnaissance automatique du locuteur sont nombreuses et diversifiées. Néanmoins, elles peuvent être regroupées en trois catégories : applications sur sites géographiques, applications juridiques et applications téléphoniques.

1.6.1 Applications sur sites géographiques

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. On peut citer :

- Le verrouillage automatique pour la protection de domiciles, garages, bâtiments, etc.
- La sécurisation accrue des cartes d'accès et le contrôle d'accès à des zones protégées.
- Les validations des transactions sur site (au niveau des distributeurs bancaires).

1.6.2 Applications téléphoniques

C'est la catégorie la plus importante car elle permet de vérifier ou d'identifier un locuteur à longue distance. Parmi ces applications on cite :

- Validation des transactions bancaires par téléphone.
- Accès à des bases de données pour plus de sécurité et plus de protection.
- Accès à des services téléphoniques.
- Le commerce électronique.

1.6.3 Applications juridiques

Dans cette catégorie, la reconnaissance automatique du locuteur est utilisée, par exemple, pour :

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

1.7 Conclusion

Les moyens biométriques permettent une authentification sûre car ils sont basés sur l'individu lui-même. Il est alors indispensable de caractériser l'individu par une empreinte afin de le différencier des autres sans aucune ambiguïté. Cette empreinte est une clé codant l'identité de la personne sans redondance ni variabilité. La plupart des indices biométriques, comme les empreintes digitales, répondent à ces critères.

Il en est différemment pour la voix dont la disposition à varier est inscrite dans sa nature. Si nous voulons vraiment parler d'empreinte vocale, il faut tenir compte du fait que la variabilité interlocuteur est plus importante que la variabilité intra locuteur. La voix devient donc un indice biométrique intéressant à exploiter et surtout via le réseau téléphonique, contrairement aux autres indices.

Chapitre 2

Paramétrisation et extraction des vecteurs acoustiques

L'objectif d'un système de paramétrisation est d'extraire les informations caractéristiques du signal de la parole en éliminant la redondance. Un tel système prend un signal en entrée et retourne un vecteur de paramètres (appelé vecteur acoustique ou vecteur d'observations). Les vecteurs de paramètres doivent être pertinents (précis, de taille restreinte et sans redondance), robustes (aux bruits) et réalisant au mieux la fonction de discrimination entre les locuteurs.

Il existe un certain nombre d'approches pour la paramétrisation. Nous présentons ici celles utilisées le plus couramment.

2.1 Analyse et paramétrisation du signal vocal

L'analyse acoustique du signal de parole consiste à extraire l'information pertinente et à réduire au maximum la redondance. Généralement, on calcule les coefficients acoustiques à des intervalles de temps réguliers, sur des blocs de signal de longueur fixe. Les techniques de paramétrisation acoustique sont nombreuses. Néanmoins, on peut les regrouper en trois grandes familles :

- Analyse par bancs de filtres.
- Analyse par transformée de Fourier.
- Analyse par prédiction linéaire.

2.1.1 Prétraitement acoustique

Le calcul des paramètres acoustiques passe par une phase de prétraitement contenant deux étapes, la préaccentuation et le fenêtrage (figure 2.1).

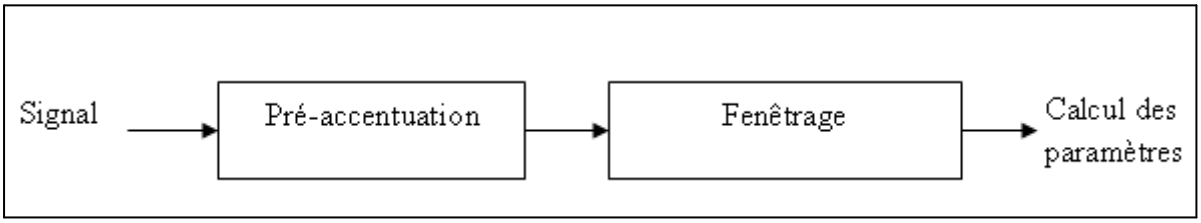

Fig. 2.1 Prétraitement acoustique.

2.1.1.1 Préaccentuation

L'onde acoustique sortante des lèvres subit, à cause de la désadaptation entre les deux milieux intérieur et extérieur, une distorsion assimilable à une désaccentuation de 6 dB par octave sur tout le spectre [2]. Pour pouvoir compenser cette distorsion, et accentuer les hautes fréquences, on applique un filtre de préaccentuation passe haut de transmitance :

$$H(z) = 1 - \alpha z^{-1} \quad (2.1)$$

avec :

$0.9 \leq \alpha \leq 1$

2.1.1.2 Fenêtrage

L'étape de fenêtrage consiste à appliquer au signal vocal une fenêtre glissante de durée limitée, et ce afin de limiter le nombre d'échantillons et de réduire les effets de bords (phénomène de Gibbs). Parmi les différentes fenêtres de pondération, les plus utilisées sont : la fenêtre rectangulaire, la fenêtre de Hamming, la fenêtre de Hanning et la fenêtre de Blackmann. En traitement de la parole, la fenêtre de Hamming est la plus utilisée. La fenêtre de Hamming est donnée par l'expression :

$$w(\mathrm{n}) = 0.54 - 0.46 \cos\left(\frac{2\pi \mathrm{n}}{\mathrm{N}-1}\right) ; \ 0 \leq n \leq N\text{-}1 \quad (2.2)$$

N : Le nombre d'échantillons.

2.1.2 Paramètres acoustiques

2.1.2.1 Paramétrisation par la méthode de prédiction linéaire

Le signal de parole n'étant pas complètement aléatoire, les échantillons successifs sont corrélés. Peut-on utiliser cette corrélation pour réduire la quantité de données ?

2.1.2.1.1 Modélisation autorégressive du signal vocal

La modélisation du signal vocal *s(n)* consiste en l'estimation des paramètres d'un filtre linéaire *H(z)* qui, soumis à une excitation particulière *u(n)*, reproduit ce signal le plus fidèlement possible [1].

L'objectif de cette modélisation étant la réduction du nombre de paramètres décrivant le signal *s(n)*, simplifiant ainsi son enregistrement, transmission ou sa reproduction.

Le modèle AR est une modélisation mathématique basée sur la mise en équation simplifiée du modèle physique de production de la parole et aboutissant à une transmittance *H(z)*, dite tous-pôles, du système. L'excitation du conduit vocal, idéalisée, est soit un bruit blanc (sons non voisés), soit un train périodique d'impulsion (sons voisés). Le conduit vocal est modélisé par une succession de tubes acoustiques, c'est à dire une cascade de résonateurs. Au final, le modèle AR consiste à dire que le son *S* est le résultat du filtrage par un filtre tous-pôles *H* d'une source *U* qui est soit un bruit blanc gaussien centré, soit un train périodique d'impulsion de fréquence : le pitch. On obtient en termes de transmittance :

$$S(z) = U(z)H(z) \tag{2.3}$$

avec :

$$H(z) = \frac{\sigma}{A(z)} \tag{2.4}$$

où :

σ : le gain du modèle.

$$A(z) = \sum_{i=0}^{p} a_p(i)\, z^{-i}\ ;\ a_p(0) = 1 \tag{2.5}$$

$a_p(i)$: coefficients de prédiction linéaire.

p : ordre du modèle.

Ce modèle de production d'un signal est appelé modèle autorégressif. En effet, l'équation (2.3) correspond dans le domaine temporel à la récurrence linéaire suivante :

$$s(n) + \sum_{i=0}^{p} a_p(i)\, s(n-i) = \sigma\, u(n) \qquad (2.6)$$

Cette équation montre qu'un échantillon quelconque du signal de la parole *s(n)* peut s'exprimer sous forme d'une combinaison linéaire des *p* échantillons qui le précèdent plus le terme d'excitation.

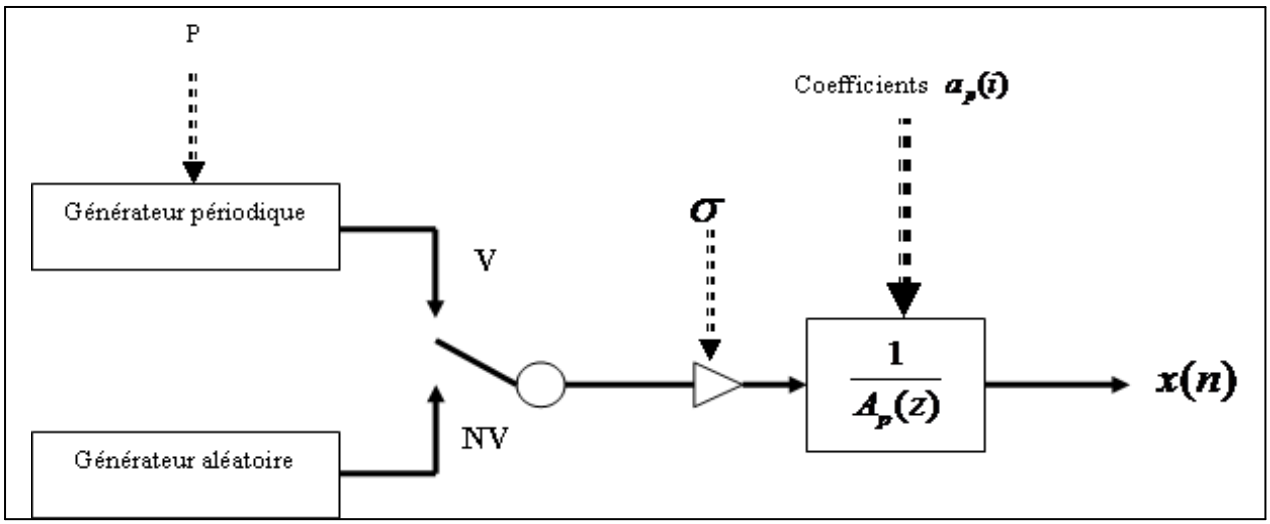

Fig. 2.2 Modèle autorégressif de la production de la parole [1].

Comme l'illustre la figure 2.2, la définition du modèle AR décrit plus haut revient à chercher les paramètres suivants : le pitch, la décision V/NV, le gain et les coefficients de prédiction.

La modélisation autorégressive du signal vocal n'est valable que dans la mesure où la condition de stationnarité est vérifiée. En raison que le signal vocal ne peut être considéré comme quasi stationnaire que sur des intervalles de temps de durée limitée, on est amené à considérer des tranches successives et à estimer un modèle AR pour chacune d'elles ; une procédure usuelle consiste à effectuer l'analyse sur des tranches de 20ms avec décalage de 10ms d'une tranche à la suivante.

2.1.2.1.2 Coefficients de prédiction linéaire LPC (Linear Prediction Coefficients)

Les coefficients LPC découlent directement du modèle AR de production de la parole. Chaque échantillon de la parole *s(n)* est constitué par une combinaison linéaire finie des *p* échantillons précédents [1]. Un seul jeu de coefficients du prédicteur est déterminé en minimisant les différences entre les échantillons réels et ceux prédits.

La modélisation étant faite, il convient à présent d'estimer les coefficients de prédiction $a_p(i)$ ainsi que le gain σ du système. L'estimation est fondée soit sur la méthode dite de la covariance, soit sur la méthode dite de l'autocorrélation qui est développée dans l'annexe A.

2.1.2.2 Paramètres LSP (Line Spectral Pair ou Line Spectral Frequencies LSF)

Les paramètres LSP (Line Spectral Pair) ont été présentés la première fois par Itakura comme représentation alternative d'information spectrale du LPC. Ils contiennent exactement la même information que les coefficients LPC [10].

En analyse par prédiction linéaire, un segment de parole est supposé être généré comme sortie d'un filtre tous pôles $H(z) = \sigma/A(z)$. Où $A(z)$ est un polynôme appelé le filtre inverse dont l'expression est donnée par:

$$A(z) = 1 + a_1 z^{-1} + \cdots + a_p z^{-p} \tag{2.7}$$

Par définition, $H(z)$ est un filtre stable si tous ses pôles sont à l'intérieur du cercle unité. Par conséquence, son filtre inverse est à minimum de phase, parce qu'il ne possède aucun zéro à l'extérieur du cercle unité. Le polynôme $A_p(z)$ associé à l'ordre p de l''analyse LPC vérifie la relation suivante:

$$A(z) = \frac{1}{2}\,[P(z) + Q(z)\,] \tag{2.8}$$

avec :

$$P(z) = A(z) + z^{-(p+1)}\,A(z^{-1}) \tag{2.9}$$

$$Q(z) = A(z) - z^{-(p+1)}\,A(z^{-1}) \tag{2.10}$$

Les LSP sont les fréquences des racines des polynômes $P(z)$ et $Q(z)$. Les polynômes $P(z)$ et $Q(z)$ possèdent des racines sous la forme e^{jw_i} . Les paramètres $\{w_i\}$, pour $i=0,1,\ldots, p+1$, définissent alors les " Line Spectral Frequencies" (LSF ou LSP). Il est important de noter que $w_0 = 0$ et $w_{p+1} = \pi$ sont des racines fixes des polynômes $Q(z)$ et $P(z)$ respectivement et seront exclus de l'ensemble des paramètres LSF.

Les polynômes *P(z)* et *Q(z)* possèdent des propriétés très intéressantes et importantes :

- les racines des polynômes *P (z)* et *Q (z)* sont sur le cercle unité.
- les racines des polynômes *P (z)* et *Q (z)* sont entrelacées dans un ordre ascendant et se trouvent dans le premier et le second quadrant du plan complexe, ce qui se traduit par la relation suivante:

$$w_0\,(Q) < w_1\,(P) < w_2\,(Q) < \cdots < w_P\,(Q) < w_{p+1}\,(P) \quad (2.11)$$

2.1.2.3 Coefficients cepstraux de prédiction linéaire LPCC

Les coefficients cepstraux de prédiction linéaire peuvent être dérivés directement des coefficients LPC. Les coefficients LPCC (Linear Prediction Cepstral Coefficients) sont obtenus par [11] :

$$c_k = -a_k - \sum_{i=1}^{k-1}\left(1 - \frac{i}{k}\right) a_i\, c_{k-i} \;;\; 1 \le k \le N \quad (2.12)$$

N: nombre de coefficients LPCC considérés.

Si $N > p$ (ordre de la prédiction linéaire), alors : $c_k=0$ pour $p < k \le N$.

2.1.2.4 Coefficients MFCC (Mel Frequency Cepstral Coefficients)

Les coefficients cepstraux issus d'une analyse par transformée de Fourier (leur fondement est basé sur l'analyse homomorphique [12] [6]) caractérisent bien la forme du spectre et permettent de séparer l'influence de la source glottique de celle du conduit vocal.

Le cepstre du signal de la parole est défini comme étant la transformée de Fourier inverse du logarithme de la densité spectrale de puissance. Pour ce signal, la source d'excitation glottique est convoluée avec la réponse impulsionnelle du conduit vocal.

$$s(t) = e(t) * h(t) \quad (2.13)$$

où *s(t)* est le signal de la parole, *e(t)* est la source d'excitation glottique et *h(t)* est la réponse impulsionnelle du conduit vocal.

L'application à l'équation (2.13) du logarithme du module de la transformée de Fourier donne :

$$\mathrm{Log|S(f)|} = \mathrm{Log|E(f)|} + \mathrm{Log|H(f)|} \quad (2.14)$$

Par une transformée de Fourier inverse, on obtient :

$$\mathrm{s'(cef)} = \mathrm{e'(cef)} + \mathrm{h'(cef)} \quad (2.15)$$

La dimension du nouveau domaine est homogène à un temps et s'appelle la quéfrence (cef), le nouveau domaine s'appelle donc : le domaine quéfrentiel. Un filtrage dans ce domaine s'appelle liftrage.

Ce domaine est intéressant pour faire la séparation des contributions du conduit vocal et de la source d'excitation dans le signal de parole. En effet, si les contributions relevant du conduit vocal et les contributions de la source d'excitation évoluent avec des vitesses différentes dans le temps, alors il est possible de les séparer par application d'une simple fenêtre dans le domaine quéfrentiel (liftrage passe-bas pour le conduit vocal).

Les coefficients cepstraux les plus répandus sont les MFCC (Mel Frequency Cepstral Coefficients) en plus des LPCC. Ils présentent l'avantage d'être faiblement corrélés entre eux, et qu'on peut donc approximer leur matrice de covariance par une matrice diagonale. La procédure de calcul des coefficients MFCC est illustrée sur la figure 2.3. Pour simuler le fonctionnement du système auditif humain, les fréquences centrales du banc de filtres sont réparties sur une échelle perceptive. Plus la fréquence centrale d'un filtre est élevée, plus sa bande passante est large. Cela permet d'augmenter la résolution dans les basses fréquences, zone qui contient le plus d'information utile dans le signal de la parole. Les échelles perceptives les plus utilisées sont l'échelle Mel et l'échelle Bark.

- Echelle Mel

$$\text{Mel(f)} = 2595 \log\left(1 + \frac{\text{f}}{700}\right) \tag{2.16}$$

- Echelle Bark

$$\text{Bark(f)} = 6 \operatorname{arcsh}\left(\frac{\text{f}}{1000}\right) \tag{2.17}$$

f : la fréquence en Hertz.

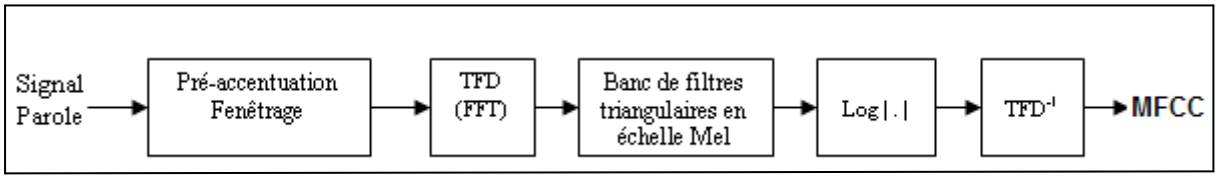

Fig. 2.3 Calcul des coefficients MFCC [2].

La figure 2.4 présente un banc de filtres triangulaires en échelle Mel sur l'échelle linéaire.

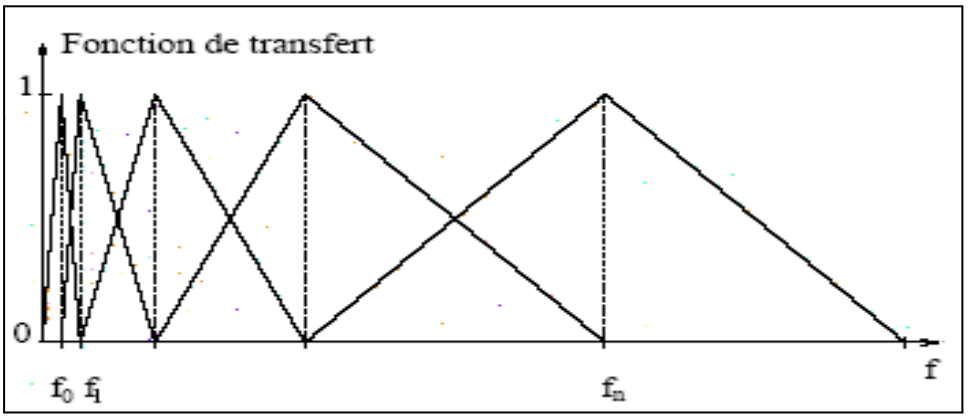

Fig. 2.4 Banc de filtres sur l'échelle linéaire [2].

Les coefficients cepstraux de fréquence en échelle Mel (MFCC) peuvent être obtenus par une transformée de Fourier inverse à partir des énergies d'un banc de filtres. Les *d* premiers coefficients cepstraux peuvent être calculés directement à partir du logarithme des énergies E_i issues d'un banc de *M* filtres par la transformée en cosinus discrète DCT définie comme :

$$c_k = \sum_{i=1}^{M} logE_i \cos\left[\frac{\pi k}{M}\left(i - \frac{1}{2}\right)\right]; \quad 1 \leq k \leq d \tag{2.18}$$

Ceci permet d'obtenir des coefficients peu corrélés. Le coefficient c_0 qui est la somme des logarithmes des énergies n'est pas utilisé ; il est éventuellement remplacé par le logarithme de l'énergie totale *E* calculée dans le domaine temporel et normalisée.

2.1.2.5 Les coefficients LFCC (Linear Frequency Cepstral Coefficients)

Aux coefficients MFCC s'ajoute un autre type de paramètres, les LFCC (Linear Frequency Cepstral Coefficients) qui sont calculés de la même manière que les MFCC, mais avec la seule différence que les fréquences des filtres sont uniformément réparties sur l'échelle linéaire des fréquences, et non pas sur une échelle perceptive de type Mel.

2.2 Conclusion

Les méthodes d'analyse de la parole dépendent essentiellement de l'application envisagée. La prédiction linéaire nous fournit des coefficients qui peuvent caractériser le conduit vocal. L'analyse cepstrale issue du modèle de perception de l'oreille nous fournit des paramètres discriminants étendus sur tout le spectre de la bande de perception de l'oreille humaine.

Dans le cadre de notre travail, nous allons évaluer l'utilisation des coefficients LSP, LPCC et MFCC pour un système d'identification du locuteur.

Chapitre 3

Modélisation et classification des locuteurs

Comme dans le cas de la reconnaissance de la parole, le problème de la reconnaissance du locuteur peut se formuler selon un problème de classification. Différentes approches ont été développées, néanmoins on peut les classer en quatre grandes familles [16] :

- L'approche vectorielle : le locuteur (son signal de parole) est modélisé par un ensemble de vecteurs de paramètres dans l'espace acoustique. Ses principales techniques sont la reconnaissance à base de la quantification vectorielle et de l'alignement temporel dynamique DTW (Dynamic Time Warping).
- L'approche statistique : consiste à représenter chaque locuteur par une densité de probabilité dans l'espace des paramètres acoustiques. Elle couvre les techniques de modélisation par modèles de Markov cachés HMM (Hidden Markov Models) et par modèles de mélanges des gaussiennes GMM (Gaussian Mixture Models).
- L'approche relative : il s'agit de modéliser un locuteur non pas de façon absolue mais relativement par rapport à d'autres locuteurs de référence, dont les modèles sont bien déterminés.
- L'approche connexionniste : consiste principalement à modéliser les locuteurs par des réseaux de neurones.

Dans ce chapitre, nous nous intéressons essentiellement à la modélisation des locuteurs en utilisant les réseaux de neurones qui fournissent de bonnes performances.

3.1 Approche vectorielle

3.1.1 Alignement Temporel Dynamique DTW

La reconnaissance par DTW repose sur le principe que chaque mot est représenté par une prononciation de référence. Utilisée en mode dépendant du texte, cette technique effectue la comparaison entre la forme d'entrée à reconnaître et une ou plusieurs formes 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. Cet alignement est réalisé par un algorithme ou technique de programmation dynamique. Malgré les bonnes performances obtenues par cette technique, elle reste très sensible à la qualité de l'alignement [1] [16] et notamment le choix du point de départ des deux formes à comparer.

3.1.2 Quantification vectorielle (QV)

Cette technique permet une compression considérable des données. Il s'agit de représenter l'espace acoustique par un nombre fini de vecteurs acoustiques formant un dictionnaire, et ce en faisant un partitionnement de cet espace en régions ou classes, qui seront représentées par leurs vecteurs centroïdes. Ainsi, chaque locuteur va être représenté par son dictionnaire de quantification [16]. 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 lent.

3.1.2.1 Définition

La quantification vectorielle consiste à représenter tout vecteur x de dimension k par un autre vecteur y_i de même dimension appartenant à un ensemble fini D de L vecteurs. Les y_i sont appelés les codes vecteurs. D est appelé le dictionnaire ou catalogue des formes. La quantification vectorielle permet d'avoir une constellation qui minimise la distorsion moyenne pour un dictionnaire de taille k donnée. La quantification vectorielle peut fournir un décodage rapide en utilisant une table simple d'identification. La figure 3.1 illustre ce principe.

Un quantificateur vectoriel de dimension k et de taille L peut être défini mathématiquement comme une application Q de R^k vers D :

$$Q: R^k \rightarrow D$$

$$x \qquad Q(x) = y_i$$

$$D = \{y_i \in R^k / i=1, 2, \ldots, L\}$$

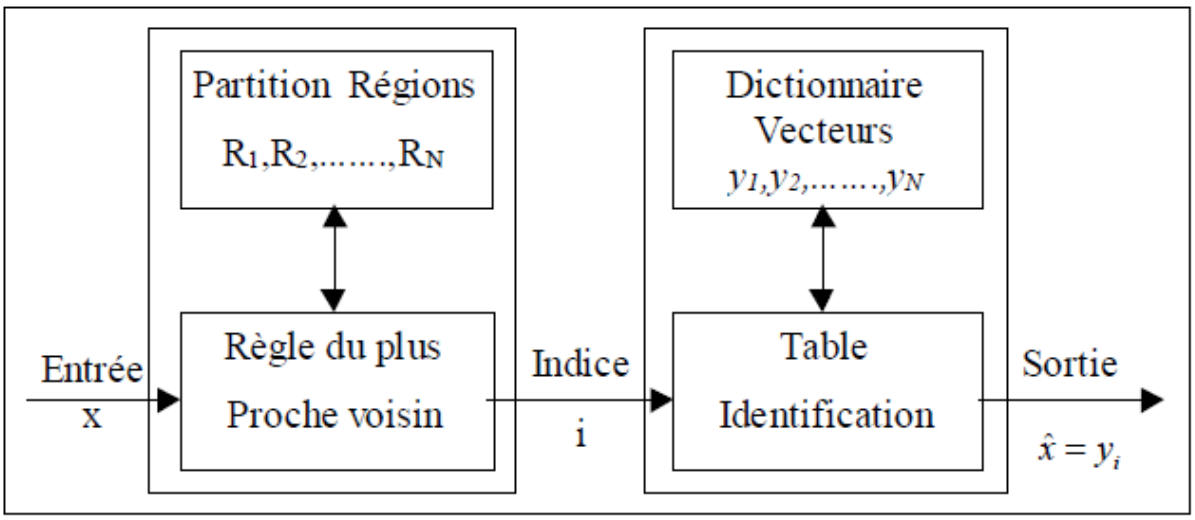

Fig. 3.1 Modèle du quantificateur vectoriel.

Cette application Q détermine implicitement une partition de l'espace R^k en L régions C_i. Ces régions, encore appelées classes ou régions de Voronoï**,** sont déterminées par :

$C_i = \{x \in R^k / Q(x) = y_i\}$

En supposant que la grandeur d'entrée est un vecteur aléatoire distribué selon une loi $p(x)$, les performances du quantificateur peuvent être mesurées par la distorsion moyenne D_Q introduite, c'est à dire par l'espérance mathématique de la distance d :

$$D_Q = E\left[d\left(x, Q(x)\right)\right] = \int d\left(x, Q(x)\right) p(x)\, dx \qquad (3.1)$$

Dans la pratique, la distribution des points d'entrée étant généralement inconnue, on approximera D_Q par une distorsion moyenne calculée sur un large nombre d'échantillons $\{x_1, x_2, \ldots, x_N\}$ de vecteurs d'entrée :

$$D_Q = \frac{1}{N} \sum_{j=1}^{N} d(x_j, Q(x_j)) \qquad (3.2)$$

On appelle centroïde de la classe C_i, le vecteur c_i tel que sa distance moyenne à tous les éléments de la classe soit minimale (en géométrie euclidienne, le centroïde correspond au centre de gravité).

Etant donné une distance et une taille du dictionnaire, il existe un quantificateur qui minimise la distorsion moyenne : c'est le quantificateur optimal.

3.1.2.2 Quantificateur Vectoriel Optimal

Un quantificateur se décompose en deux applications : un codeur et un décodeur. Le quantificateur optimal est alors celui réunissant le codeur optimal et le décodeur optimal.

- Le codeur optimal : étant donné un dictionnaire $\{y_1,y_2,...,y_L\}$, la meilleure partition est celle qui vérifie : $R_i = \{ x \in R^k / d(y_i,x) \leq d(y_j,x) \ \forall j \in \{1,...,L\} \}$ pour $i=1,2,...,L$. C'est la règle dite du plus proche voisin.
- Le décodeur optimal : étant donné une partition $\{R_1,R_2,...,R_L\}$, les meilleurs représentants sont obtenus par la condition du centroïde (centre de gravité c_i de la partie de la densité de probabilité placée dans la région R_i).
- Une troisième condition est nécessaire : il faut que la probabilité qu'un vecteur à coder se trouve à la même distance de deux représentants soit nulle, sinon ce vecteur source est affecté à l'un des deux représentants, et dans ce cas, la partition de l'espace n'est plus optimale. Si les vecteurs source sont à amplitude continue, cette troisième condition est toujours vérifiée.

Ces trois conditions conduisent à la conception d'un algorithme qui réalise, à partir d'une séquence d'apprentissage représentative de la statistique de la source à coder, la construction d'un dictionnaire optimal.

3.1.2.3 Algorithme LBG

L'algorithme LBG, du nom de ses inventeurs Linde, Buzo et Gray, est l'extension au cas vectoriel de l'algorithme de Lloyd-Max du cas scalaire. Il permet de déterminer les L meilleurs représentants (les codes vecteurs du dictionnaire) d'une distribution inconnue à partir d'un ensemble de N réalisations dit ensemble d'apprentissage $\{x_1,x_2,...,x_N\}$. En pratique $N >> L$.

Les étapes de l'algorithme sont :

1. Initialiser le dictionnaire.
2. Construction des L régions de Voronoï associées aux L représentants du dictionnaire.
3. Calcul des L nouveaux centroïdes (représentants) correspondants à partir de la relation : $$c_i = \frac{1}{N_i} \sum_{\{k;x_k \in R_i\}} x_k \quad (3.3)$$
4. Répéter les étapes 2 et 3 tant que la croissance de la distorsion moyenne reste importante.

L'algorithme présente le problème lié à l'initialisation du dictionnaire : il n'est pas sûr que l'algorithme converge vers le minimum global [8]. Pour éviter ce problème, l'algorithme LBG possède la procédure d'initialisation efficace dite par dichotomie vectorielle (Split Binary Method) que voici (figure 3.2) :

1. Calculer le centroïde de l'ensemble de toute la séquence d'apprentissage.

2. A partir de chaque centroïde c_i, on fabrique, par un petit déplacement dans deux directions opposées, deux nouveaux vecteurs :
   $c_i^+ = c_i\,(1+\varepsilon)$ et $c_i^- = c_i\,(1-\varepsilon)$ avec : $0.01 \leq \varepsilon \leq 0.05$
   Cela a pour conséquence de doubler le nombre de vecteurs représentants.
3. Déterminer les régions de Voronoï associées à l'ensemble des vecteurs obtenus précédemment, ainsi que leurs centroïdes respectifs.
4. Répéter les étapes 2 et 3 jusqu'à l'obtention d'un ensemble de L centroïdes.
   La taille du dictionnaire L est une puissance de 2.

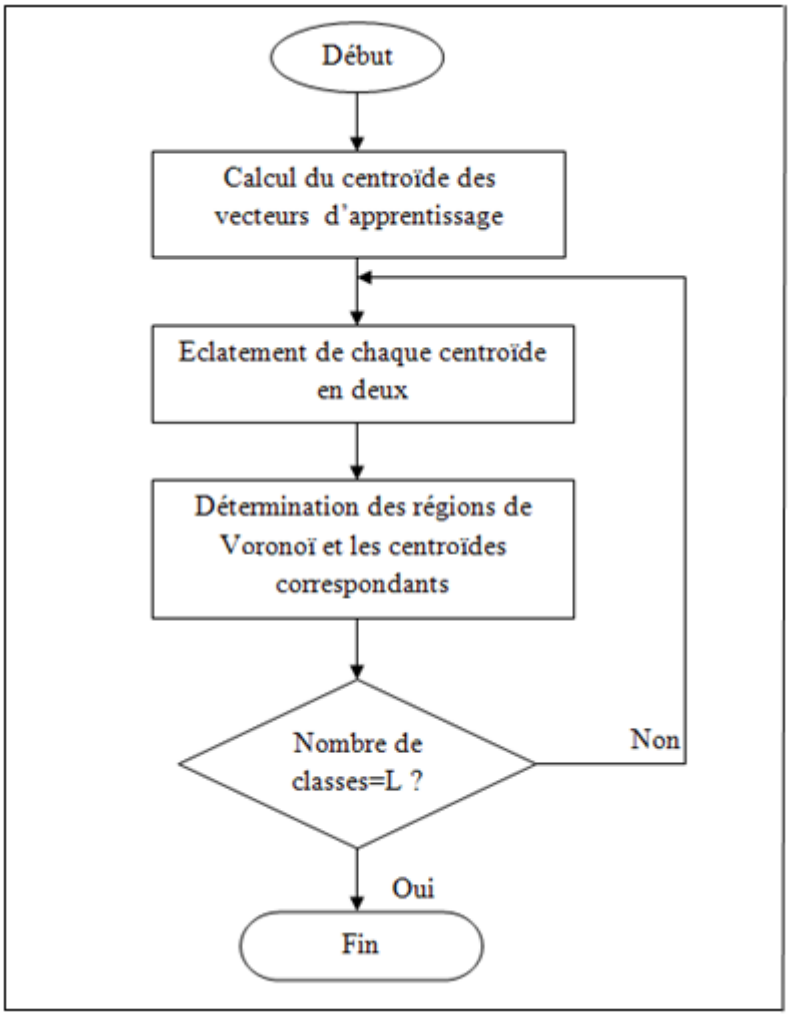

Fig. 3.2 Initialisation de l'algorithme LBG.

3.2 Approche statistique

3.2.1 Modèles de Markov Cachés HMM

3.2.1.1 Définition

Le modèle HMM a été d'abord introduit dans un cadre purement statistique, il s'est ensuite imposé en reconnaissance automatique de la parole avant d'être appliqué en reconnaissance automatique du locuteur. Le modèle HMM présente différents avantages : clarté, rigueur, efficacité et généralité. Un modèle HMM se caractérise par un système à états comportant deux processus.

- Les réalisations du premier processus sont des chaînes cachées $Q= (q_1q_2...q_T)$ des états du système avec un état initial q_1 et un état final q_T.
- Les réalisations du second processus sont des chaînes externes ou observations $O= (o_1o_2...o_T)$ où chaque o_t est un élément d'un espace d'observation Ω.

Dans une modélisation par HMM, on suppose que la suite des vecteurs acoustiques d'observation est stationnaire par blocs. Ainsi, les vecteurs acoustiques d'un bloc suivent la même loi de probabilité. La modélisation d'un bloc de vecteurs acoustiques représente un état du modèle HMM. Dans cette approche, chaque entité est modélisée par une machine d'états appelée machine Markovienne et qui est composée d'un ensemble d'états et de transitions qui permettent de passer d'un état à un autre. Un modèle HMM est un modèle statistique séquentiel qui suppose que les caractéristiques observées forment une succession d'états distincts.

Soit λ un modèle Markovien de N états et $Q= (q_1q_2...q_T)$ une séquence d'états correspondant à l'observation $O= (o_1o_2...o_T)$, où q_t est l'état atteint par le processus à l'instant t. L'état du modèle de Markov λ qui correspond à o_t n'étant pas directement observable, on dit qu'il est caché (d'où le nom de modèles de Markov cachés). La figure 3.3 représente un exemple de modèle de Markov. Un tel modèle est définit par :

- Un ensemble d'états cachés $\{s_1,s_2,...,s_N\}$.
- Un ensemble d'observations $\{v_1,v_2,...,v_M\}$.
- Probabilités de transition $a_{ij} =P(q_{t+1} =s_j \mid q_t =s_i)$.
- Probabilités d'observation $b_j(k) =P(o_t =v_k \mid q_t =s_j)$ qui sont en général des mélanges de gaussiennes.
- Un ensemble de probabilités initiales $\pi = \{\pi_i / \pi_i =P (q_1=s_i) ; i=1,...,N \}$.

Un modèle de Markov caché est donc spécifié par le triplet $\lambda = \{A,B, \pi\}$ où A est la matrice des probabilités de transition, B la matrice des probabilités d'observation et π les probabilités initiales.

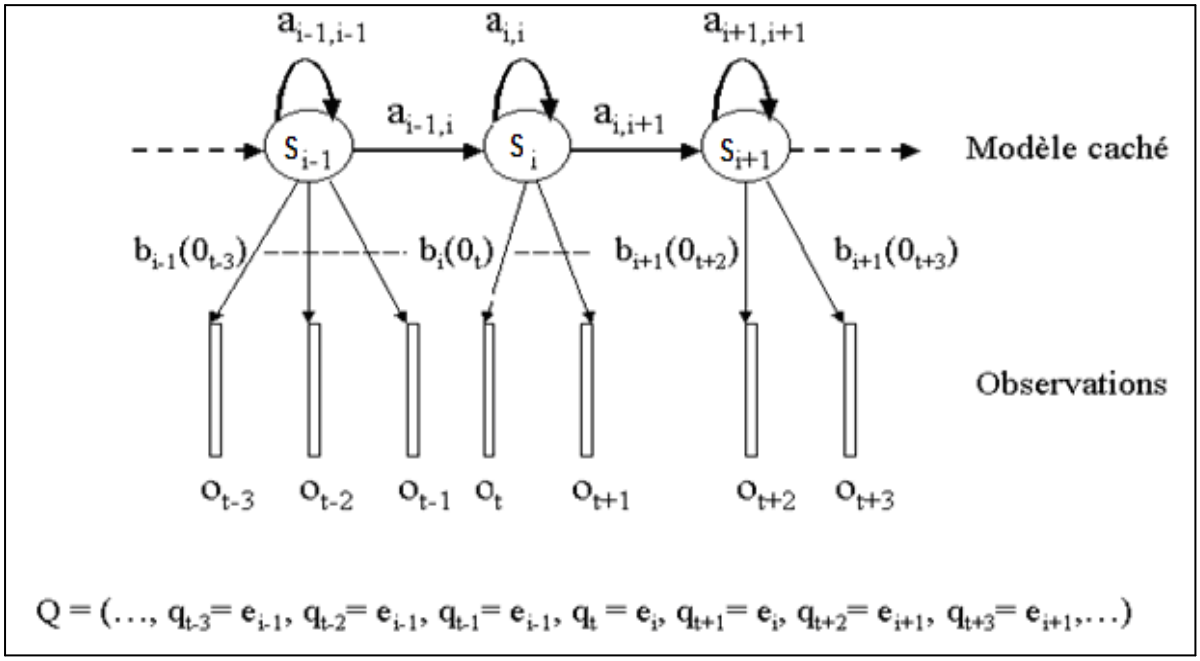

Fig. 3.3 Exemple d'une machine Markovienne [15].

3.2.1.2 Problèmes des modèles HMM

- **Evaluation**

Etant donné une séquence d'observations $O = (o_1 o_2 \ldots o_T)$ et un modèle $\lambda = \{A,B, \pi\}$, comment déterminer la probabilité que l'observation ait été engendrée par le modèle $P(O|\lambda)$? Il existe deux méthodes pour résoudre ce problème. La méthode dite directe et qui consiste à calculer cette probabilité en énumérant toutes les séquences d'états possibles de même longueur que la séquence d'observation. Cette technique demande beaucoup de temps de calcul. Un moyen plus rapide pour calculer cette probabilité est l'utilisation des algorithmes de programmation dynamique [16].

- **Séquence d'états optimale**

Etant donné une séquence d'observations $O = (o_1 o_2 \ldots o_T)$ et un modèle $\lambda = \{A,B,\pi\}$, quelle est la séquence d'états $Q = (q_1 q_2 \ldots q_T)$ optimale qui maximise la probabilité $P(O,Q|\lambda)$? Pour cela, l'algorithme de Viterbi est le plus utilisé [16]. Il permet de chercher la séquence d'états

cachés la plus probable en ne gardant que les états s_i qui maximisent la probabilité à chaque instant t (jusqu'à T).

- **Apprentissage**

C'est le problème principal d'un modèle HMM. En effet, la qualité d'un système utilisant une modélisation HMM dépend principalement de la qualité de ses modèles. C'est pourquoi l'étape d'apprentissage qui consiste à estimer les paramètres des modèles HMM ($\lambda = \{A,B,\pi\}$) est très importante et ainsi la plus difficile. Il existe plusieurs méthodes pour résoudre ce problème, les plus utilisées sont :

- L'algorithme de Viterbi associé à des estimateurs empiriques : l'algorithme de Viterbi sert à déterminer la séquence d'états cachés la plus vraisemblable, correspondant aux données d'apprentissage. Les paramètres des densités de probabilité de chaque état peuvent être alors ré-estimés en utilisant des estimateurs empiriques et les observations associées à chaque état le long du chemin de Viterbi [1].
- L'algorithme EM (Expectation-Maximisation) : Cet algorithme permet de résoudre le problème d'apprentissage en estimant de manière itérative les paramètres d'un modèle au sens du maximum de vraisemblance [15] [16].

3.2.1.3 Phase de reconnaissance

La phase de reconnaissance consiste, étant donné une observation, à évaluer la probabilité qu'elle soit engendrée par chacun des modèles et sélectionner celui qui est le plus probable. Le principal avantage de l'approche HMM est sa grande capacité d'apprendre les propriétés statistiques. En reconnaissance du locuteur, le choix le plus fréquent consiste à utiliser un modèle dont la distribution conditionnelle dans chaque état (probabilité d'observation) est un mélange de gaussiennes. L'utilisation de ces modèles est plus importante dans le mode dépendant du texte parce qu'en mode indépendant du texte l'information supplémentaire apportée par les transitions entre états n'améliore pas les performances de la reconnaissance du locuteur.

3.2.2 Modèles de mélanges des gaussiennes GMM

3.2.2.1 Définition

Un mélange de gaussiennes est une somme pondérée de M densités gaussiennes (figure 3.4). Soit un locuteur s et un vecteur acoustique x de dimension D, le mélange de gaussiennes est défini comme suit :

$$p(x|\lambda_s) = \sum_{m=1}^{M} \pi_m^s \; b_m^s(x) \tag{3.4}$$

où les $b_m^s(x)$ représentent des densités gaussiennes paramétrées par un vecteur de moyenne μ_m^s et une matrice de covariance Σ_m^s :

$$b_m^s(x) = \frac{1}{(2\pi)^{D/2}\,|\Sigma_m^s|^{1/2}} \exp\left[-\frac{1}{2}\,(x-\mu_m^s)^T(\Sigma_m^s)^{-1}(x-\mu_m^s)\right] \tag{3.5}$$

et les π_m^s représentent les poids du mélange, avec :

$$\sum_{m=1}^{M} \pi_m^s = 1$$

Un locuteur est donc modélisé par un ensemble de paramètres noté$\lambda_s = \{\pi_m^s, \mu_m^s, \Sigma_m^s\}$; pour *m=1,2,…,M*.

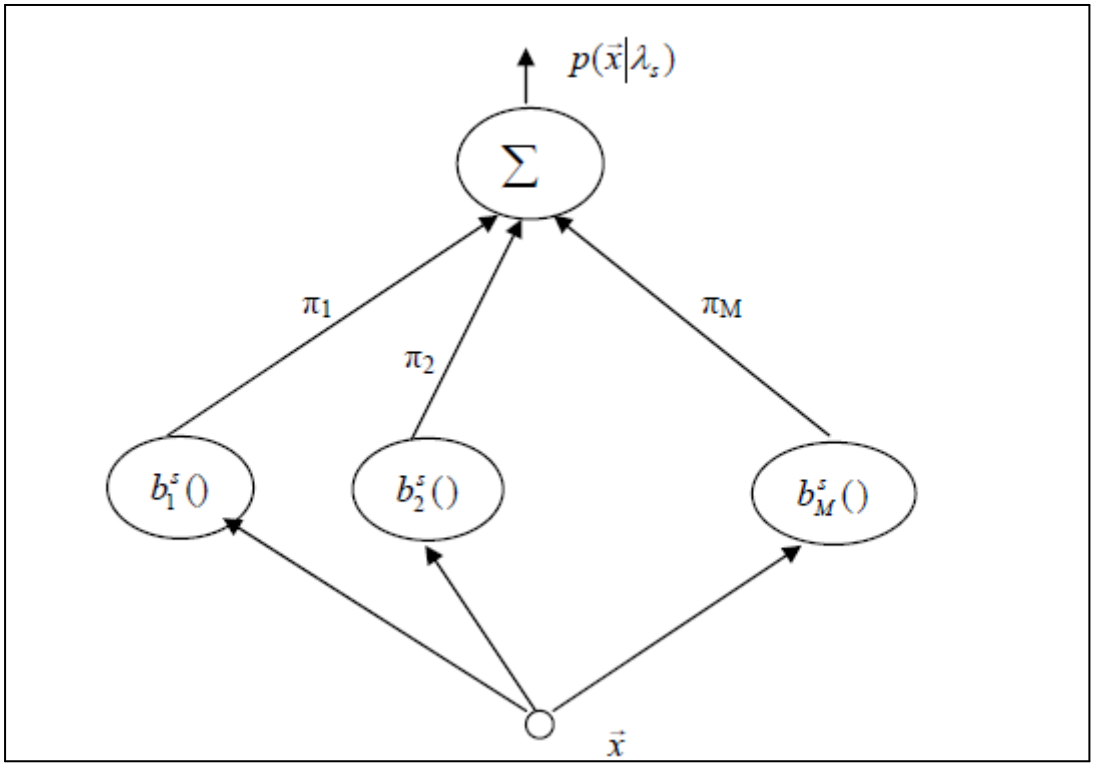

Fig. 3.4 Modèle de GMM [15].

Le modèle GMM peut prendre plusieurs formes, notamment en ce qui concerne les matrices de covariance. On peut utiliser une matrice de covariance pour chaque gaussienne, ou bien une matrice de covariance globale commune à toutes les gaussiennes.

3.2.2.2 Apprentissage du modèle

La phase d'apprentissage consiste à estimer l'ensemble λ des paramètres d'un modèle GMM pour chaque locuteur, et ce à partir des vecteurs acoustiques issus de la phase de paramétrisation. La méthode conventionnelle est celle du Maximum de vraisemblance (MV) dont le but est de déterminer les paramètres du modèle qui maximisent la vraisemblance des données d'apprentissage en utilisant l'algorithme efficace EM [15] [16]. Pour une séquence de *N* vecteurs d'apprentissage $X= (x_1x_2...x_N)$ suffisamment indépendants, la vraisemblance du modèle GMM est donnée par :

$$p(X|\lambda) = \prod_{n=1}^{N} p(x_n|\lambda) = \prod_{n=1}^{N} \sum_{m=1}^{M} p(x_n|\pi_m, \mu_m, \Sigma_m) \quad (3.6)$$

3.2.2.3 Décision

Soit un groupe de *S* locuteurs, représentés par des modèles GMM ou HMM : $\lambda_1,\lambda_2,\ldots,\lambda_S$. L'objectif de la phase d'identification est de trouver, à partir d'une séquence observée $X= (x_1x_2...x_N)$, le modèle qui a la probabilité a posteriori maximale, c'est-à-dire :

$$\hat{s} = argmax_{1\leq s\leq S} p(\lambda_s|X) \quad (3.7)$$

Ce qui donne, d'après la loi de Bayes :

$$\hat{s} = argmax_{1\leq s\leq S} [\ \frac{p(X|\lambda_s)}{p(X)}\ p(\lambda_s)\ ] \quad (3.8)$$

En supposant l'équiprobabilité d'apparition des locuteurs $p(\lambda_s) = \frac{1}{S}$, la loi de classification devient:

$$\hat{s} = argmax_{1\leq s\leq S} [\ p(X|\lambda_s)\ ] \quad (3.9)$$

En utilisant le logarithme et l'indépendance entre les observations x_n, le système d'identification calcule le score suivant :

$$\hat{s} = argmax_{1\leq s\leq S} [\ \sum_{n=1}^{N} \log p(x_n|\lambda_s)\ ] \quad (3.10)$$

3.3 Approche relative

C'est une nouvelle technique qui consiste à modéliser un locuteur non plus de façon absolue, mais relativement à un ensemble de locuteurs bien appris. En effet, chaque locuteur est représenté par sa localisation dans un espace de référence. Cette technique trouve son application lorsqu'on dispose de peu de données d'apprentissage. Il faut donc estimer avec très peu de données un modèle robuste du locuteur qui permettra sa reconnaissance. Cette approche a donné naissance à la notion d'espace de locuteurs, où un locuteur est

représenté par une combinaison linéaire des modèles de référence, ce qui réduit considérablement le nombre de paramètres [16].

3.4 Approche connexionniste

Les réseaux de neurones ont été récemment et largement utilisés en reconnaissance du locuteur. Ils offrent en effet une bonne alternative au problème de la discrimination entre les locuteurs. Cette approche comprend une grande famille de méthodes différentes. Chaque méthode est représentée par un réseau qui implémente une fonction de transfert globale spécifiée par l'architecture et les fonctions élémentaires du réseau. Dans cette approche, un locuteur est représenté par un ou plusieurs réseaux de neurones appris directement des trames obtenues en phase de paramétrisation et permettant de le discriminer par rapport à un ensemble de locuteurs.

Les réseaux de neurones sont capables d'implanter des techniques discriminantes très efficaces et offrent des outils de classification qui permettent la séparation des classes de façon non linéaire.

3.4.1 Historique

Les recherches menées dans le domaine du connexionnisme [19] ont démarré avec la présentation en 1943 par W. McCulloch et W. Pitts d'un modèle simplifié de neurone biologique communément appelé neurone formel. Ils montrèrent également théoriquement que des réseaux de neurones formels simples peuvent réaliser des fonctions logiques, arithmétiques et symboliques complexes.

En 1949, D. Hebb initie, dans son ouvrage "The Organization of Behavior", la notion d'apprentissage. Deux neurones entrant en activité simultanément vont être associés (c'est-à dire que leurs contacts synaptiques vont être renforcés). On parle de loi de Hebb et d'associationnisme.

En 1958, F. Rosenblatt développe le modèle du Perceptron. C'est un réseau de neurones inspiré du système visuel. Il possède deux couches de neurones : une couche de perception (sert à recueillir les entrées) et une couche de décision. C'est le premier modèle pour lequel un processus d'apprentissage a pu être défini.

S'inspirant du perceptron, Widrow et Hoff, développent, dans la même période, le modèle de l'Adaline (Adaptive Linear Element). Ce dernier sera, par la suite, le modèle de base des réseaux de neurones multicouches.

En 1969, Les recherches sur les réseaux de neurones ont été pratiquement abandonnées lorsque M. Minsky et S. Papert ont publié leur livre « Perceptrons » (1969) et démontré les limites théoriques du perceptron, en particulier, l'impossibilité de traiter les problèmes non linéaires.

En 1982, Hopfield développe un modèle qui utilise des réseaux totalement connectés basés sur la règle de Hebb pour définir les notions d'attracteurs et de mémoire associative. En 1984, c'est la découverte des cartes de Kohonen avec un algorithme non supervisé basé sur l'auto-organisation et suivi une année plus tard par la machine de Boltzman (1985).

Une révolution survient alors dans le domaine des réseaux de neurones artificiels : une nouvelle génération de réseaux de neurones, capables de traiter avec succès des phénomènes non-linéaires : le perceptron multicouche ne possède pas les défauts mis en évidence par Minsky. Proposé pour la première fois par Werbos, le Perceptron Multicouche apparaît en 1986 introduit par Rumelhart, et, simultanément, sous une appellation voisine, chez Le Cun (1985). Ces systèmes reposent sur la rétropropagation du gradient de l'erreur dans des systèmes à plusieurs couches.

De nos jours, l'utilisation des réseaux de neurones dans divers domaines ne cesse de croître. Les applications en sont multiples et variées. Ils sont principalement utilisés pour résoudre des problèmes de contrôle et commande de processus, de classification et reconnaissance de formes, de prévision, de mémorisation comme une alternative à l'intelligence artificielle et en relation plus ou moins étroite avec la modélisation de processus cognitifs réels (capables de connaître ou faire connaître).

3.4.2 Neurone biologique

Le neurone biologique est composé de quatre parties distinctes (figure 3.5) :

- le corps cellulaire, qui contient le noyau de la cellule nerveuse; c'est en cet endroit que prend naissance l'influx nerveux, qui représente l'état d'activité du neurone;
- les dendrites, ramifications tubulaires courtes formant une espèce d'arborescence autour du corps cellulaire; ce sont les entrées principales du neurone, qui captent l'information venant d'autres neurones;
- l'axone, longue fibre nerveuse qui se ramifie à son extrémité; c'est la sortie du neurone et le support de l'information vers les autres neurones;
- la synapse, qui communique l'information, en la pondérant par un poids synaptique, à un autre neurone; elle est essentielle dans le fonctionnement du système nerveux.

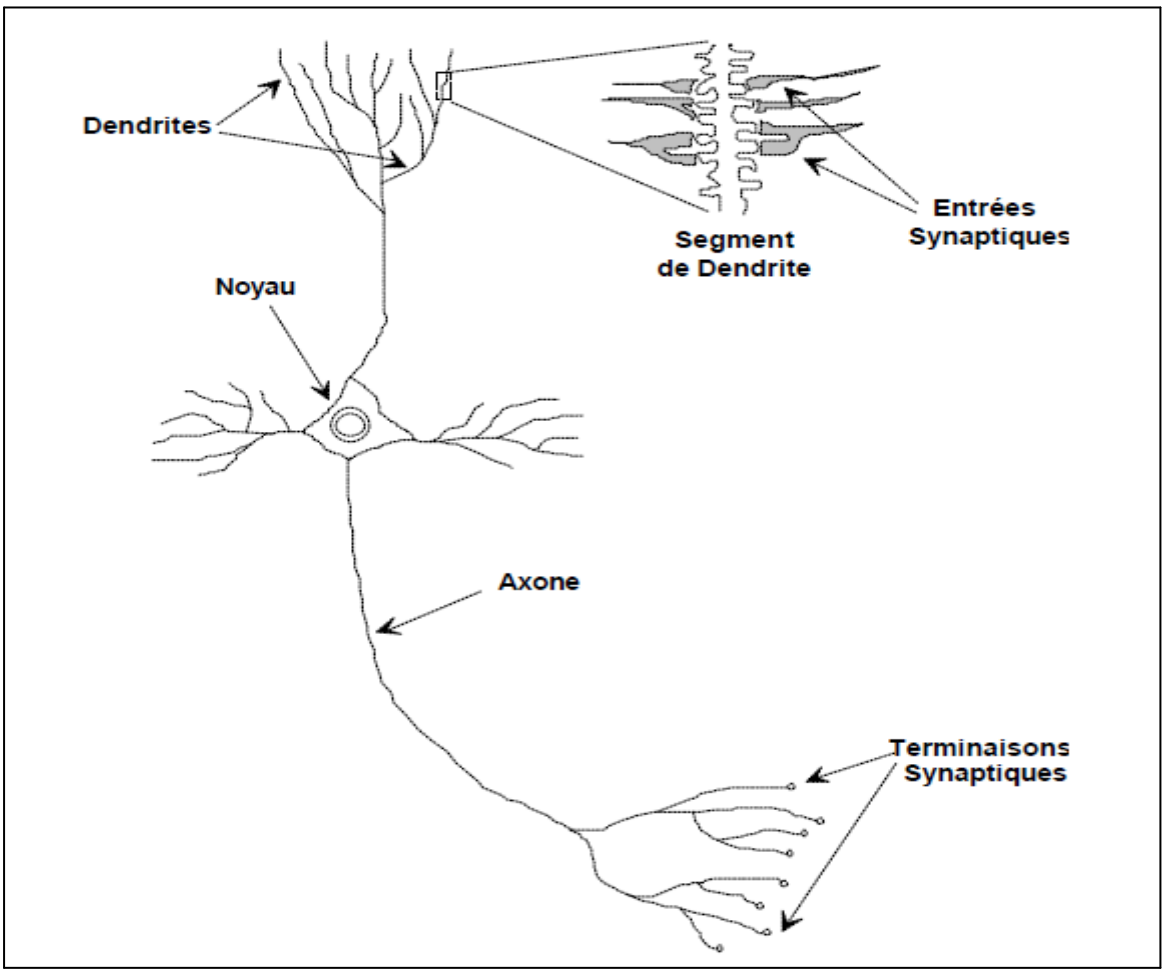

Fig. 3.5 Schéma d'un neurone biologique [17].

Chaque neurone réalise une opération très simple, qui est en fait une somme pondérée de ses entrées. Le résultat est comparé à un seuil et le neurone devient excité si ce seuil est dépassé. L'information contenue dans le cerveau est représentée par les poids donnés aux entrées de chaque neurone. Du fait du grand nombre de neurones et de leurs interconnections, ce système possède une propriété de tolérance aux fautes. Ainsi, la défectuosité d'un élément mémoire (neurone) n'entraînera aucune perte réelle d'information, mais seulement une faible dégradation en qualité de toute l'information contenue dans le système.

3.4.3 Neurone formel (artificiel)

Chaque neurone artificiel est un processeur élémentaire. Il reçoit un nombre variable d'entrées en provenance de neurones en amont ou des capteurs composant la machine dont il fait partie. A chacune de ses entrées est associé un poids représentatif de la force de la connexion. Chaque processeur élémentaire est doté d'une sortie unique, qui se ramifie ensuite pour alimenter un nombre variable de neurones en aval. A chaque connexion est associé un poids. Il est commode de représenter graphiquement un neurone comme indiqué sur la figure 3.6.

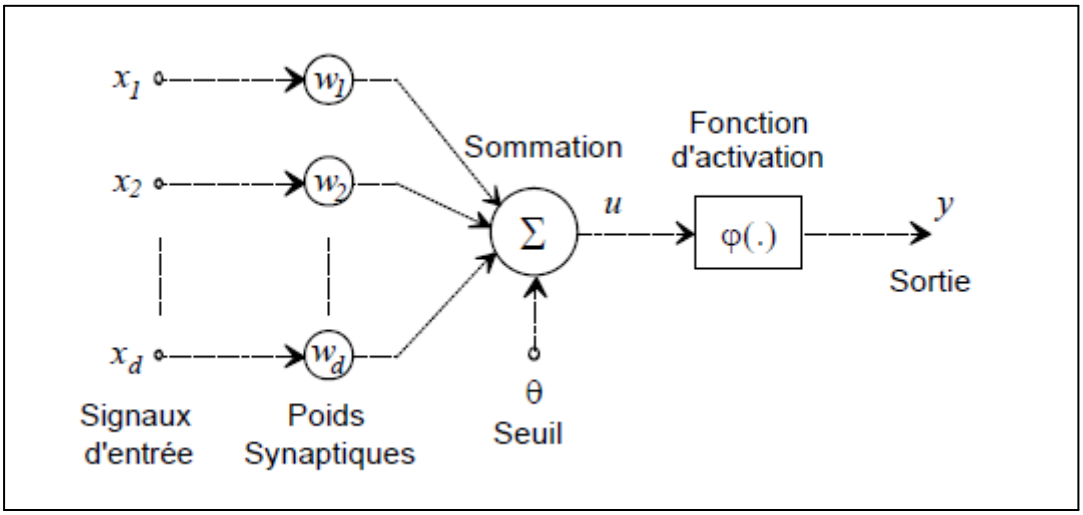

Fig. 3.6 Modèle du neurone formel [17].

Le neurone réalise alors trois opérations sur ses entrées :

- Pondération : multiplication de chaque entrée par un paramètre appelé poids de connexion.
- Sommation : une sommation des entrées pondérées est effectuée.
- Activation : passage de cette somme dans une fonction, appelée fonction d'activation.

3.4.4 Modélisation d'un neurone formel

Les réseaux de neurones formels sont à l'origine d'une tentative de modélisation mathématique du cerveau humain. Les premiers travaux datent de 1943 et sont l'œuvre de McCulloch et Pitts. Ils présentent un modèle assez simple pour les neurones et explorent les possibilités de ce modèle. La modélisation consiste à mettre en œuvre un système de réseaux neuronaux sous un aspect non pas biologique mais artificiel, cela suppose que d'après le principe biologique on aura une correspondance pour chaque élément composant le neurone biologique, donc une modélisation pour chacun d'entre eux.

On pourra résumer cette modélisation par le tableau 3.1 qui nous permettra de voir clairement la transition entre le neurone biologique et le neurone formel [22].

Neurone biologique	**Neurone artificiel**
Synapse	Poids de la connexion
Axone	Signal de sortie
Dendrite	Signal d'entrée
Noyau	Fonction d'activation

Tab. 3.1 Analogie entre le neurone biologique et le neurone artificiel [22].

Ainsi, la sortie du neurone artificiel de la figure 3.6 s'écrit :

$$y = f(\sum_{i=1}^{R} w_i x_i + b) \quad (3.11)$$

On note:

x_i ($i=1,...,R$) : les entrées du neurone.

w_i ($i=1,...,R$) : les poids des connexions.

b : le seuil ou biais.

f : la fonction d'activation du neurone.

y : la sortie du neurone.

- **Fonction d'activation**

Cette fonction permet de définir l'état interne du neurone en fonction de son entrée totale. Elle peut être une fonction à seuil, une fonction linéaire ou une fonction sigmoïde (figure 3.7).

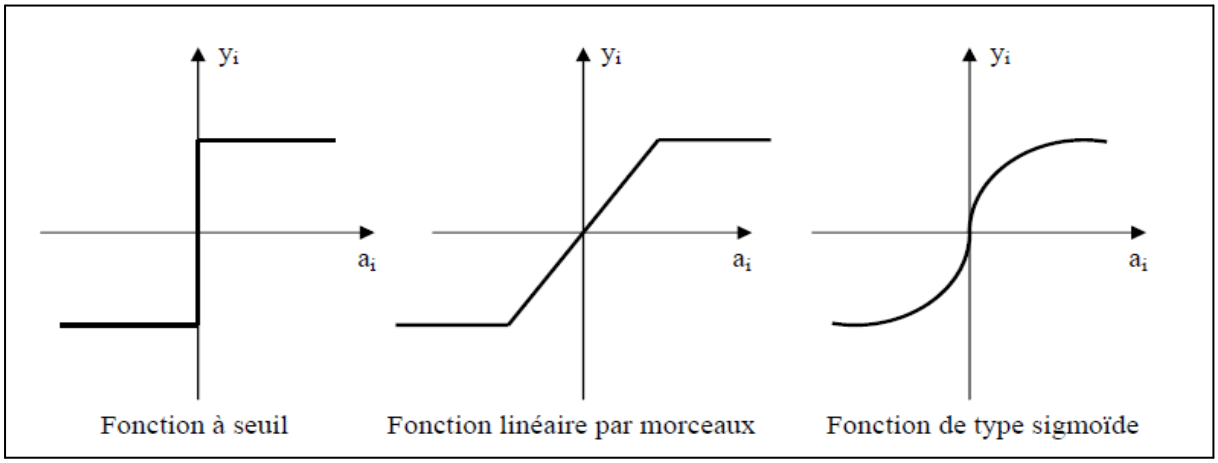

Fig. 3.7 Différents types de fonctions d'activation pour le neurone artificiel [19].

La fonction sigmoïde présente l'avantage d'être monotone, continûment dérivable et bornée. On distingue deux types :

$$f(x) = \tanh(\beta x) \; ; \quad \beta > 0 \quad (3.12)$$

ou :

$$f(x) = \frac{1}{1+e^{-\beta x}} \; ; \quad \beta > 0 \quad (3.13)$$

3.4.5 Définition du réseau de neurones artificiel RNA

Haykin en propose la définition suivante [20]:

« Un réseau de neurones est un processus distribué de manière massivement parallèle, qui a une propension naturelle à mémoriser des connaissances de façon expérimentale et de les rendre disponibles pour l'utilisation. Il ressemble au cerveau en deux points:

- *la connaissance est acquise à travers d'un processus d'apprentissage;*
- *les poids des connections entre les neurones sont utilisés pour mémoriser la connaissance.»*

Ainsi, un réseau de neurones artificiel peut être considéré comme un modèle mathématique de traitement réparti, composé de plusieurs éléments de calcul non linéaire (neurones), opérant en parallèle et connectés entre eux par des poids. Chaque neurone (processeur élémentaire) calcule une sortie unique sur la base des informations qu'il reçoit.

3.4.6 Apprentissage des réseaux de neurones

Le point crucial du développement d'un réseau de neurones est son apprentissage. Il s'agit d'une procédure adaptative par laquelle les connexions des neurones sont ajustées face à une source d'information. Dans le cas des réseaux de neurones artificiels, on ajoute souvent à la description du modèle (topologie) l'algorithme d'apprentissage. Le modèle sans apprentissage présente en effet peu d'intérêt.

On appelle apprentissage des réseaux de neurones la procédure qui consiste à estimer les paramètres des neurones du réseau afin que celui-ci remplisse au mieux la tâche qui lui est affectée [22].

Dans la majorité des algorithmes d'apprentissage actuels, les variables modifiées pendant l'apprentissage sont les poids des connexions. L'apprentissage est la modification des poids du réseau dans l'optique d'accorder la réponse du réseau aux exemples et à l'expérience. Les poids sont initialisés avec des valeurs aléatoires. Puis des exemples expérimentaux, représentatifs du fonctionnement du procédé dans un domaine donné, sont présentés au réseau de neurones. Ces exemples sont constitués de couples expérimentaux de vecteurs d'entrée et de sortie. Une méthode d'optimisation modifie les poids au fur et à mesure des itérations pendant lesquelles on présente la totalité des exemples, afin de minimiser l'écart entre les sorties calculées et les sorties expérimentales désirées. Afin d'éviter les problèmes de surapprentissage, la base d'exemples est divisée en deux parties : la base d'apprentissage et la base de test. L'optimisation des poids se fait sur la base d'apprentissage, mais les poids retenus sont ceux pour lesquels l'erreur obtenue sur la base de test est la plus faible. En effet,

si les poids sont optimisés sur tous les exemples de l'apprentissage, on obtient une précision très satisfaisante sur ces exemples mais on risque de ne pas pouvoir généraliser le modèle à des données nouvelles. A partir d'un certain nombre d'itérations, le réseau ne cherche plus l'allure générale de la relation entre les entrées et les sorties du système, mais s'approche trop prés des points et apprend le bruit [22].

Sur la figure 3.8 ci-dessous, on peut observer qu'au début de l'apprentissage, pour les premières itérations, l'erreur sur la base d'apprentissage est grande et peut légèrement augmenter étant donné que les poids initiaux sont choisis aléatoirement. Ensuite, cette erreur diminue avec le nombre d'itérations. L'erreur sur la base de test diminue puis augmente à partir d'un certain nombre d'itérations. Les poids et le nombre d'itérations retenus sont ceux qui minimisent l'erreur sur la base de test.

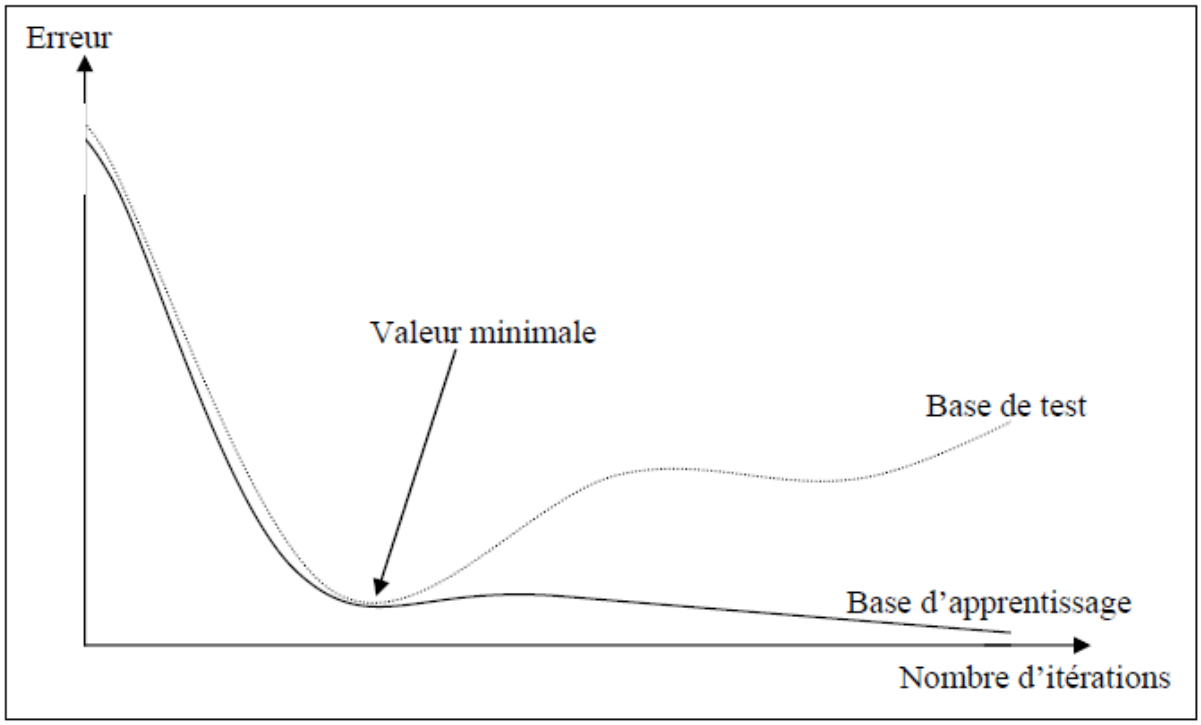

Fig. 3.8 Erreur sur la base d'apprentissage et la base du test en fonction du nombre d'itérations [22].

- **Surapprentissage**

Il arrive qu'à faire apprendre un réseau de neurones toujours le même échantillon, celui-ci devient inapte à reconnaître autre chose que les éléments présents dans l'échantillon. Le réseau ne cherche plus l'allure générale de la relation entre les entrées et les sorties du système, mais cherche à reproduire les allures de l'échantillon. On parle alors de surapprentissage : le réseau est devenu trop spécialisé et ne généralise plus correctement.

Ce phénomène apparaît aussi lorsqu'on utilise trop d'unités cachées (de connexions), la phase d'apprentissage devient alors trop longue (trop de paramètres réglables dans le système) et les performances du réseau en généralisation deviennent médiocres [22].

3.4.6.1 Types d'apprentissage

Il existe de nombreux types de règles d'apprentissage qui peuvent être regroupées en deux grandes catégories: les règles d'apprentissage supervisé et non supervisé.

- **Apprentissage supervisé**

Un apprentissage est dit supervisé lorsqu'on force le réseau à converger ou tendre vers un état final précis et désiré, en même temps qu'on lui présente un motif. Ce genre d'apprentissage est réalisé à l'aide d'une base d'apprentissage constituée de plusieurs exemples entrées-sorties (les entrées du réseau et les résultats désirées).

La procédure usuelle dans le cadre de la prévision (régression) est l'apprentissage supervisé (figure 3.9) qui consiste à associer une réponse spécifique désirée à chaque entrée. La modification des poids s'effectue progressivement jusqu'à ce que l'erreur entre les sorties calculés et les résultats désirés soit minimale.

Cet apprentissage n'est possible que si un large jeu de données est disponible et si les solutions sont connues pour les exemples de la base d'apprentissage.

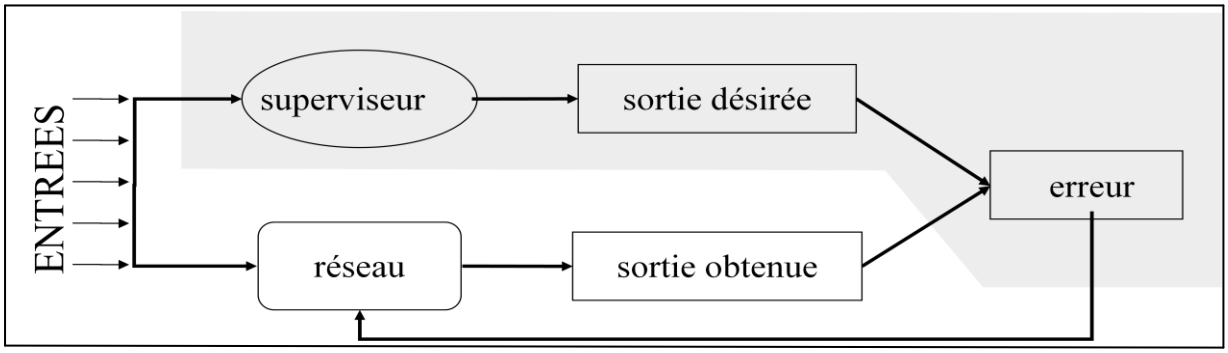

Fig. 3.9 Apprentissage supervisé.

- **Apprentissage non supervisé**

L'apprentissage non supervisé consiste à ajuster les poids à partir d'un ensemble d'apprentissage formé uniquement de données (figure 3.10). Aucun résultat désiré n'est fourni au réseau. L'apprentissage consiste à détecter les similarités et les différences dans l'ensemble d'apprentissage. Les poids et les sorties du réseau convergent, en théorie, vers les représentations qui capturent les régularités statistiques des données [22]. Ce type d'apprentissage est également dit compétitif. L'avantage de ce type d'apprentissage réside dans sa grande capacité d'adaptation reconnue comme une auto-organisation.

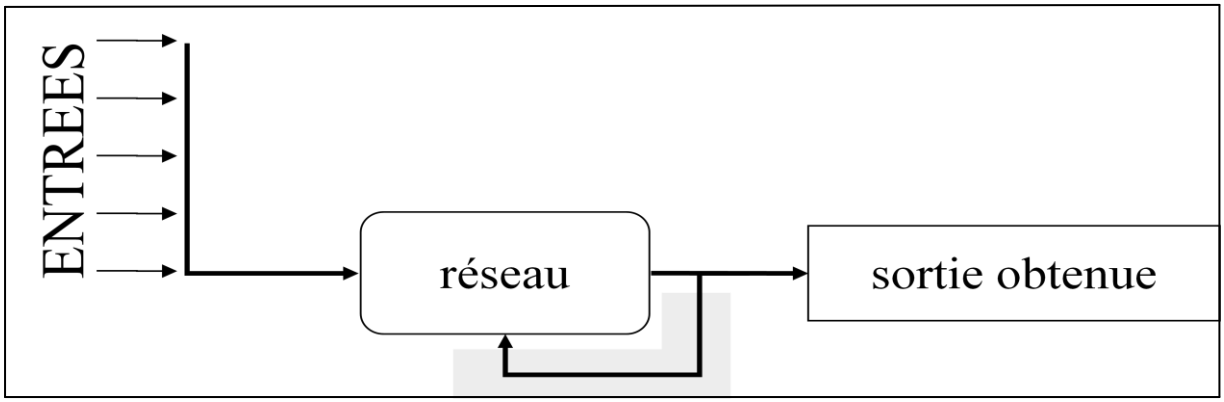

Fig. 3.10 Apprentissage non supervisé.

3.4.7 Architecture des réseaux de neurones

Les connexions entre les neurones qui composent le réseau décrivent la topologie du modèle. Elle peut être quelconque, mais le plus souvent il est possible de distinguer une certaine régularité. On peut classer les RNA en deux grandes catégories: les réseaux «feed-back» et les réseaux «feed-forward». Dans le cadre de notre travail, nous nous intéressons essentiellement aux réseaux «feed-forward», plus particulièrement aux perceptrons multicouches MLP (Multi-Layer Perceptron) et aux réseaux de neurones probabilistes PNN (Probabilistic Neural Networks), qui s'adaptent bien à l'identification du locuteur et donnent de très bonnes performances.

3.4.7.1 Réseaux «feed-back»

Appelés aussi "réseaux récurrents", ce sont des réseaux dans lesquels il y a un retour en arrière de l'information.

- **Cartes Auto-Organisatrices**

Ce modèle a été présenté par Kohonen en 1984 en se basant sur des constatations biologiques [21]. Les cartes de Kohonen sont réalisées à partir d'un réseau à deux couches, une en entrée et une en sortie. Notons que les neurones de la couche d'entrée sont entièrement connectés à la couche de sortie. Les neurones de la couche de sortie sont placés dans un espace à une ou deux dimensions en général. Chaque neurone possède donc des voisins dans cet espace et possède des connexions latérales récurrentes dans sa couche (le neurone inhibe les neurones éloignés et laisse agir les neurones voisins).Ce sont des réseaux à apprentissage non supervisé.

- **Réseaux de Hopfield**

Les réseaux de Hopfield sont des réseaux récurrents et entièrement connectés. Dans ce type de réseaux, chaque neurone est connecté à chaque autre neurone et il n'y a aucune

différenciation entre les neurones d'entrée et de sortie. Ils fonctionnent comme une mémoire associative non linéaire et sont capables de trouver un objet stocké en fonction de représentations partielles ou bruitées. L'application principale des réseaux de Hopfield est l'entrepôt des connaissances mais aussi la résolution des problèmes d'optimisation [22]. Le mode d'apprentissage utilisé ici est le mode non supervisé.

3.4.7.2 Réseaux «feed-forward»

Appelés aussi "réseaux de type Perceptron", ce sont des réseaux dans lesquels l'information se propage de couche en couche sans retour en arrière possible.

- **Perceptron monocouche**

C'est historiquement le premier RNA, c'est le Perceptron de Rosenblatt. C'est un réseau simple puisqu'il ne se compose que d'une couche d'entrée et d'une couche de sortie (figure 3.11). Il est calqué, à la base, sur le système visuel et de ce fait a été conçu dans un but premier de reconnaissance des formes. Cependant, il peut aussi être utilisé pour faire de la classification et pour résoudre des opérations logiques simples. Sa principale limite est qu'il ne peut résoudre que des problèmes linéairement séparables. Il suit généralement un apprentissage supervisé selon la règle des moindres carrés de l'erreur [22].

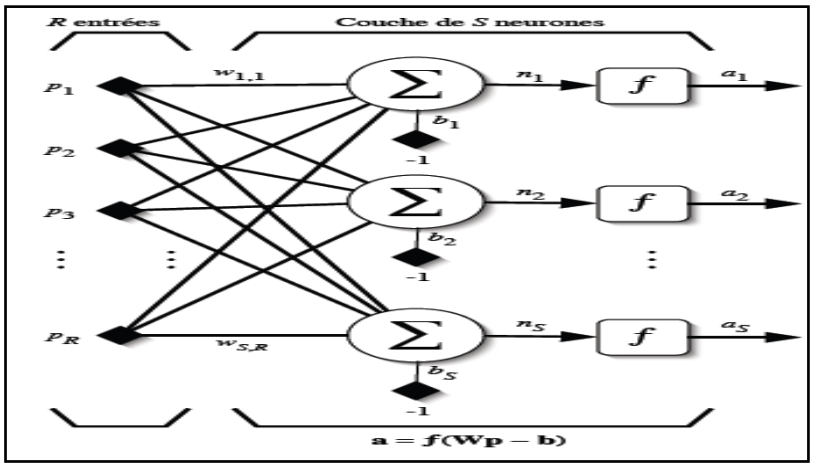

Fig. 3.11 Perceptron monocouche [18].

- **Perceptron multicouche MLP**

C'est une extension du modèle précédent, avec une ou plusieurs couches «cachées » entre l'entrée et la sortie. Chaque neurone dans une couche est connecté à tous les neurones de la couche précédente et de la couche suivante (excepté pour les couches d'entrée et de sortie) et il n'y a pas de connexion entre les cellules d'une même couche (figure 3.12). Les fonctions d'activations utilisées dans ce type de réseaux sont principalement les fonctions linéaires et sigmoïdes. Il peut résoudre des problèmes non linéairement séparables (le cas de la reconnaissance du locuteur) et des problèmes logiques plus compliqués [18]. Il suit aussi un

apprentissage supervisé selon plusieurs règles, la plus utilisée est la rétropropagation du gradient de l'erreur [18] développée dans l'annexe B (qui est une généralisation de la règle des moindres carrés).

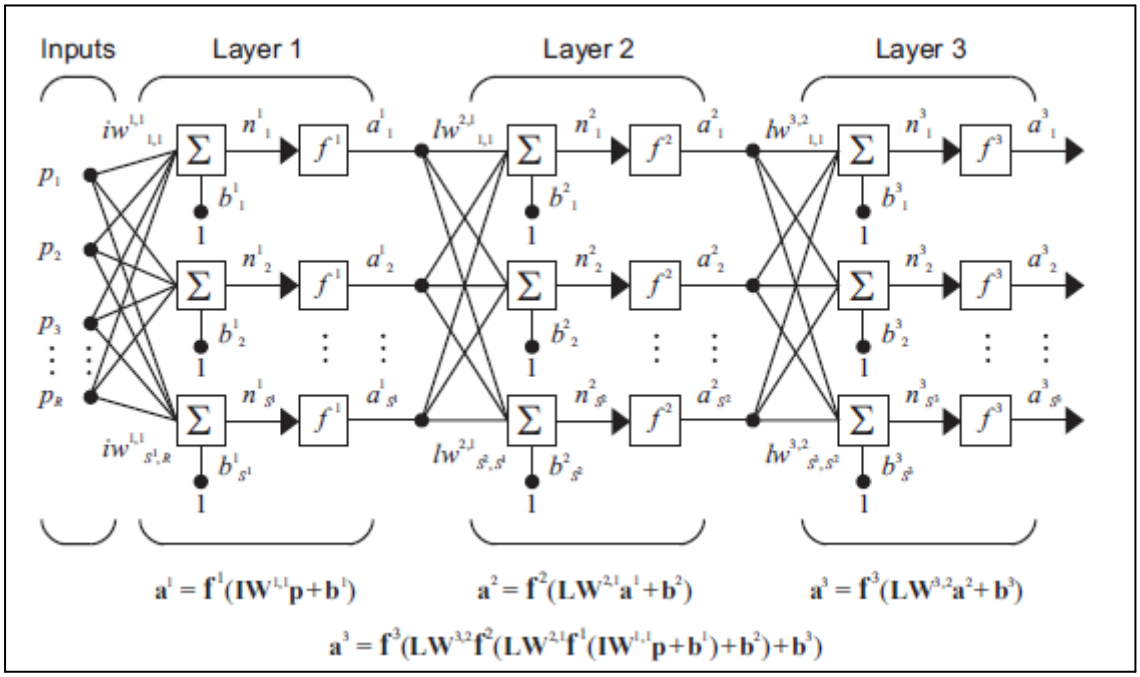

Fig. 3.12 Perceptron multicouche à deux couches cachées [18].

- **Réseaux de neurones probabilistes PNN**

Ces réseaux sont principalement utilisés en classification et reconnaissance des formes. Ce type de réseaux contient trois couches : une couche d'entrée, une couche cachée à fonction radiale de base RBF (Radial Basis Function) et une couche compétitive de sortie dont le nombre de sorties est égal au nombre de classes *K* (figure 3.13).

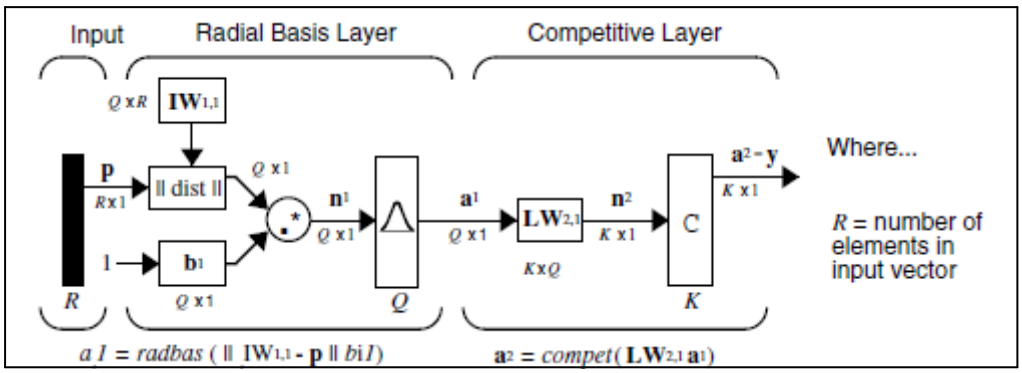

Fig. 3.13 Architecture des PNN [18].

Dans la couche cachée RBF, la distance euclidienne entre le vecteur d'entrée et les poids associés à chaque neurone constituant cette couche (l'apprentissage est supervisé et le nombre de neurones *Q* dans la couche RBF est égal au nombre des paires entrées-sorties désirées de la phase d'apprentissage) est calculée. Le vecteur résultant, de *Q* composantes, est multiplié

(multiplication scalaire) par le vecteur des biais correspondant à chaque neurone. La fonction d'activation de cette couche est exponentielle (figure 3.14). Elle est continue, dérivable, bornée et paire. Les *Q* composantes du vecteur de sortie a^1, de la couche RBF, vont être regroupées en *K* classes (*Q*>>*K*).

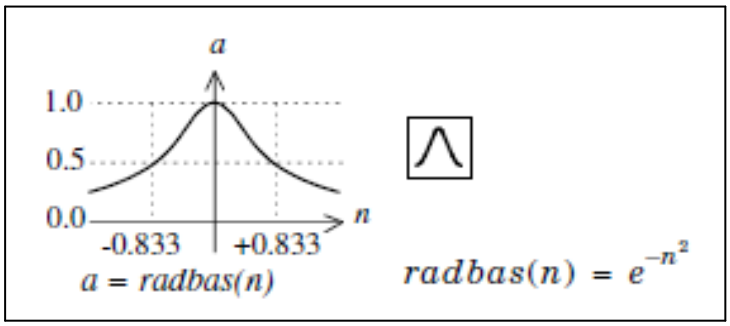

Fig. 3.14 Fonction radiale de base [18].

La couche compétitive de sortie attribue à la sortie correspondant à la classe à laquelle appartient le vecteur d'entrée la valeur 1 et la valeur 0 pour toutes les autres sorties [18]. En effet, la fonction d'activation de cette couche attribue, en sortie, la valeur 1 pour la plus grande valeur du vecteur n^2, ayant *K* composantes, et la valeur 0 pour toutes les autres valeurs inférieures.

3.5 Conclusion

Nous avons décrit les différents modèles de classification. Nous nous sommes basés essentiellement sur l'approche vectorielle (QV) et l'approche connexionniste (RNA). Cette description a pour objet l'application de ces différentes approches sur une base de données, et évaluer leur performance en fonction des paramètres choisis.

Chapitre 4

Evaluation expérimentale

Cette partie présente l'évaluation expérimentale des deux approches : QV et RNA (MLP et PNN) dans l'identification du locuteur en mode indépendant du texte (ensemble fermé). En premier lieu, nous décrivons la base de données utilisée. Ensuite, nous rappelons l'analyse acoustique appliquée. Enfin, nous détaillons les résultats obtenus en donnant les commentaires nécessaires.

4.1 Base de données utilisée

Nous avons pris un échantillon de 38 locuteurs (33 hommes et 5 femmes) extrait de la base de données de l'Ecole Nationale Polytechnique. C'est une base acoustique dédiée seulement à la reconnaissance du locuteur. Elle est constituée de 45 locuteurs (37 hommes et 8 femmes) algériens. Les données sont échantillonnées avec 16KHz, sur 16 bits. L'enregistrement a été fait dans le laboratoire du signal et communication de l'Ecole Nationale Polytechnique sous les conditions :

- La chambre d'enregistrement est un peu bruyante ;
- Le type de parole : phrases continues en Arabe ;
- Le microphone utilisé : bidirectionnelle.

Pour chaque locuteur, nous disposons de 10 phrases chacune de 3 secondes en moyenne. Nous avons concaténé 7 phrases pour l'apprentissage et les 3 autres phrases sont utilisées pour le test.

4.2 Analyse acoustique

L'analyse de la parole consiste à extraire l'information pertinente et à réduire au maximum la redondance. Nous nous intéressons essentiellement à extraire des coefficients représentant l'information relative à l'identité du locuteur. Les coefficients cepstraux (MFCC ou LPCC) permettent une parfaite déconvolution de la contribution du conduit vocal et celle de la source d'excitation. Les paramètres LSP contiennent des informations sur le conduit vocal. Nous allons évaluer l'utilisation de ces trois types de coefficients (qui sont d'ailleurs les plus utilisés dans les systèmes d'identification du locuteur).

Dans nos expériences, une analyse est appliquée, toutes les 10 ms, sur des fenêtres de 20 ms (par glissement et recouvrement des fenêtres d'analyse). A chaque trame (fenêtre d'analyse), nous associons un vecteur de coefficients acoustiques (MFCC, LPCC ou LSP). Avant de faire l'extraction des coefficients, deux étapes s'imposent et qui sont : le prétraitement et l'élimination du silence.

- **Prétraitement acoustique**

La phase du prétraitement contient deux étapes :

- L'étape de préaccentuation acoustique qui consiste à filtrer le signal vocal par un filtre passe haut de transmittance : $H(z) = 1 - 0.95\, z^{-1}$.
- L'étape du fenêtrage qui consiste à multiplier le signal vocal par une fenêtre de pondération. Nous avons utilisé la fenêtre de Hamming (figure 4.1) de durée de 20 ms avec déplacement de 10 ms.

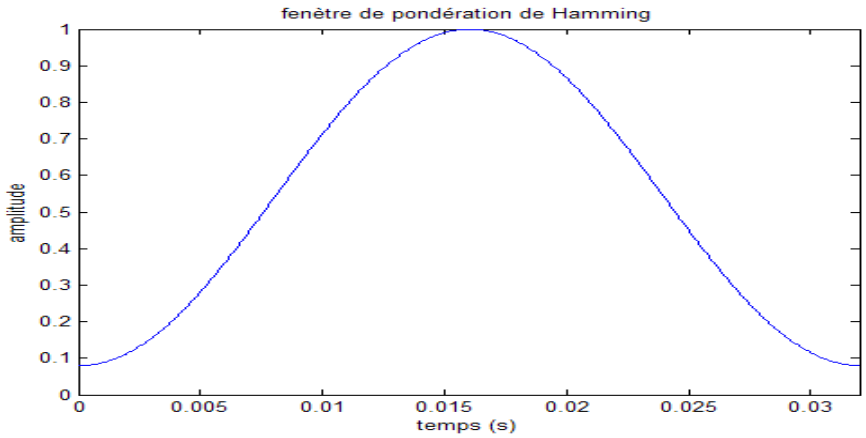

Fig. 4.1 Fenêtre de pondération de Hamming.

- **Détection et élimination du silence**

Les périodes du silence ne portent aucune information et peuvent diminuer les performances d'un système de reconnaissance. Pour cela, nous avons effectué une étude

statistique sur la base de données utilisée, à partir de laquelle nous avons déterminé un seuil d'énergie. Toute trame du signal de niveau énergétique inférieur au seuil prédéterminé sera considérée comme étant un silence et sera éliminée.

4.3 Protocole d'évaluation

La performance du système d'identification est évaluée en calculant le taux d'identification correcte défini par :

$$I_c\,(\%) = \frac{Nombre\ de\ segments\ de\ test\ correctement\ identifiés}{Nombre\ total\ de\ segments\ de\ test} \times 100 \qquad (4.1)$$

Le test est effectué sur l'ensemble de tous les locuteurs (38 locuteurs), chaque locuteur a trois segments (phrases) de test, soit un total de 114 tests.

4.4 Langage utilisé

Nous avons utilisé **MATLAB 7.5.0 (R2007b)** qui possède des boites à outils spécialisées. L'ensemble des fonctions de ces boites à outils facilitent beaucoup la simulation. Dans ce travail, nous avons utilisé principalement deux boites à outils, la première «Signal Processing Toolbox» orientée traitement du signal et la seconde « Neural Networks Toolbox » orientée réseaux de neurones.

4.5 Evaluation expérimentale

4.5.1 Paramètres d'évaluation

Nous allons étudier l'influence des paramètres suivants sur la performance du système (taux d'identification correcte) :

- **Ordre du modèle**

Nous faisons varier l'ordre du modèle (taille du dictionnaire de la quantification vectorielle) de 1 à 128.

- **Dimension des vecteurs acoustiques**

Pour voir l'apport de la dimension du vecteur acoustique sur le taux d'identification, nous faisons varier le nombre de coefficients (LPCC, LSP ou MFCC) de 4 à 40.

- **Type des vecteurs acoustiques**

Nous allons évaluer l'utilisation des vecteurs LPCC, LSP et MFCC dans un système d'identification du locuteur.

- **Qualité des données d'apprentissage et de test**

Nous commençons à travailler avec une fréquence d'échantillonnage de 16 KHz. Ensuite, nous introduisons une dégradation sur les données à utiliser afin d'approcher la qualité du réseau téléphonique commuté RTC (filtrage dans la bande téléphonique [300Hz-3400KHz], sous-échantillonnage à 8KHz et l'ajout d'un bruit blanc gaussien avec un rapport signal sur bruit SNR variant de 10dB à 100dB).

Remarque

Dans le cas des RNA, l'architecture du réseau va aussi influencer sur les performances du système (nombre de couches cachées, nombre de neurones dans chaque couche, etc).

4.5.2 Quantification vectorielle QV

Chaque locuteur est représenté ou modélisé par son dictionnaire de quantification (phase d'apprentissage). Lors de la phase du test, les vecteurs acoustiques extraits du segment de test (sur 3 secondes) sont appliqués à tous les dictionnaires afin d'être respectivement quantifiés. Le dictionnaire pour lequel la distance cumulée est minimale désigne le locuteur identifié.

4.5.2.1 Influence du l'ordre du modèle

La fréquence d'échantillonnage est 16KHz. Nous faisons varier la taille L des dictionnaires en maintenant le nombre de coefficients à 12. Les résultats obtenus sont dans le tableau ci-dessous :

L	1	2	4	8	16	32	64	128
I_c(LPCC)	69.3	72.8	86.8	90.4	93.0	94.0	94.0	94.0
I_c(LSP)	62.3	72.8	86.8	92.1	94.0	95.6	95.6	95.6
I_c(MFCC)	74.5	74.5	79.8	87.7	89.5	92.1	92.1	92.1

Tab. 4.1 Influence du l'ordre du modèle QV.

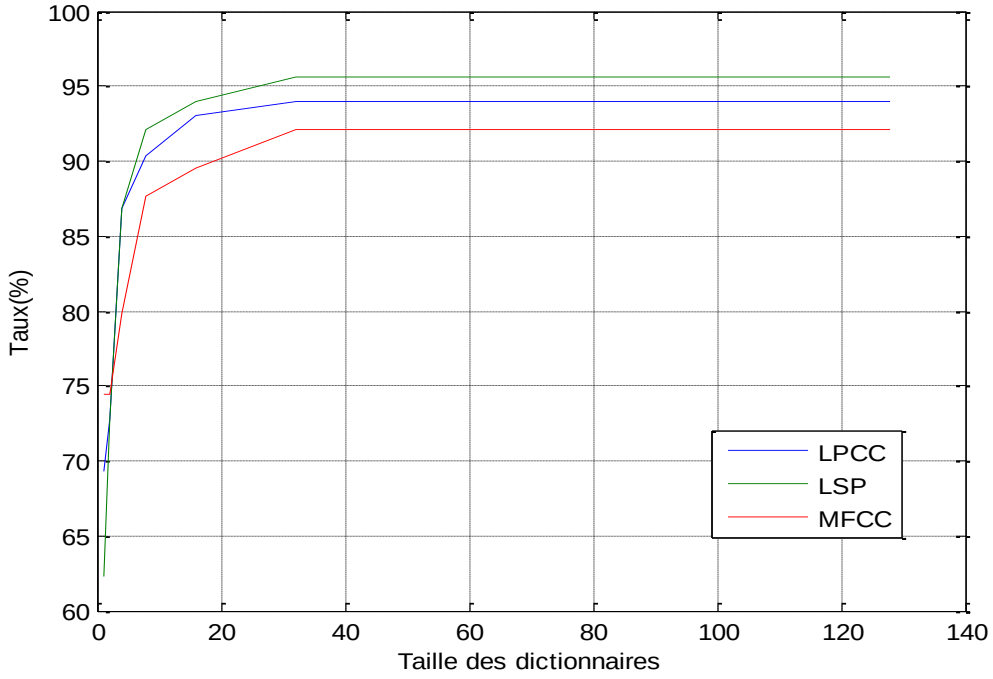

Fig.4.2 Influence du l'ordre du modèle QV.

D'après ces résultats, nous constatons que la quantification vectorielle est très performante, notant que les conditions d'enregistrement de la base de données ne sont pas assez bonnes (bruit ambiant, microphone utilisé, adaptation homme-machine, etc). Le taux d'identification augmente avec l'ordre du modèle pour les trois courbes (figure 4.2). Il atteint sa valeur maximale à partir de L=32 et devient constant. Nous constatons que les coefficients LSP donnent la meilleure performance (très proche de celle fournie par les coefficients LPCC).

4.5.2.2 Influence de la dimension des vecteurs acoustiques

La fréquence d'échantillonnage est 16KHz. Nous faisons varier la dimension P des vecteurs acoustiques en maintenant la taille des dictionnaires à 32. Les résultats obtenus sont dans le tableau ci-dessous :

P	4	8	12	16	20	24	28	32	40
I_c(LPCC)	82.4	92.1	94.0	94.0	94.8	94.8	94.8	94.8	94.8
I_c(LSP)	86.0	93.0	95.6	95.6	96.5	96.5	96.5	96.5	96.5
I_c(MFCC)	84.2	89.5	92.1	92.1	92.1	92.1	92.1	92.1	93.0

Tab. 4.2 Influence de la dimension des vecteurs acoustiques QV.

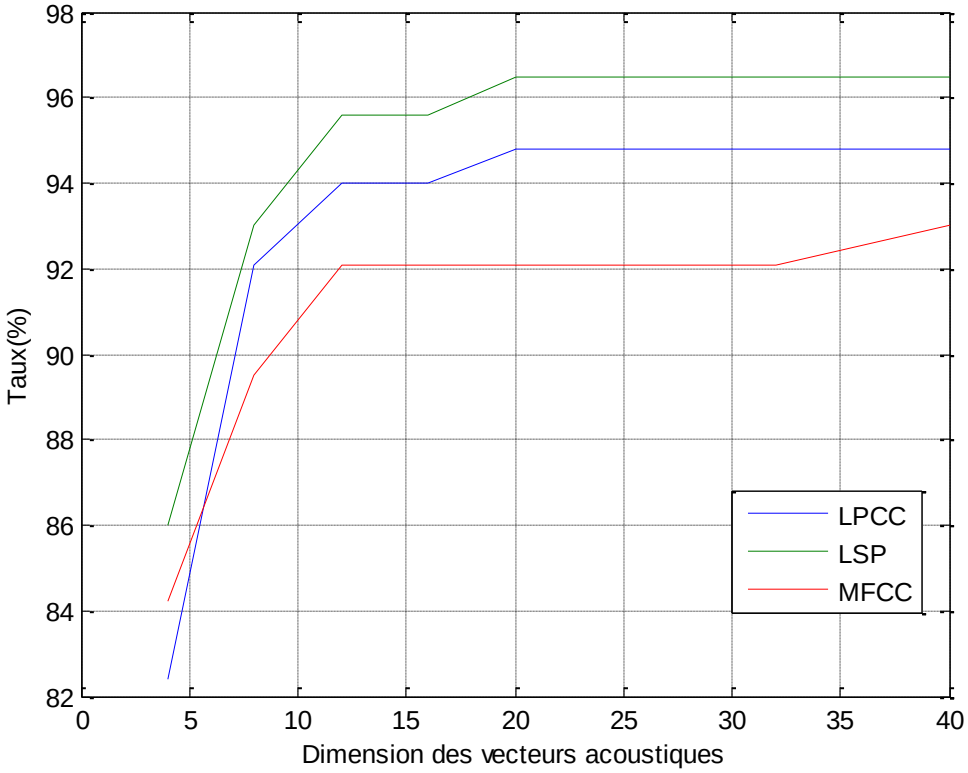

Fig. 4.3 Influence de la dimension des vecteurs acoustiques QV.

Le taux d'identification augmente avec le nombre de coefficients des vecteurs acoustiques pour les trois courbes (figure 4.3). Il devient pratiquement constant à partir de P=12. Les vecteurs acoustiques de dimension supérieure (P>12) n'apportent pas un plus d'informations remarquable sur l'identité des locuteurs. Les coefficients LSP fournissent les meilleurs résultats et ils sont suivis par les coefficients LPCC.

Conclusion

Dans un milieu non bruité (avec : Fe=16KHz), la quantification vectorielle est très performante. Cette performance s'améliore avec l'augmentation de L et P, mais les durées d'apprentissage et surtout du test risquent de devenir longues (augmentation du temps de calcul et de l'espace mémoire nécessaire pour le stockage des références). Pour remédier à ce problème, il faut ajuster les paramètres L et P de façon à garder une bonne performance, avoir une durée de test courte et réduire l'espace mémoire nécessaire. Pour L=32, P=12 et des vecteurs acoustiques LSP, ces contraintes sont satisfaites. Nous utiliserons ces paramètres pour la suite des expériences.

4.5.2.3 Qualité des données d'apprentissage et du test

Nous introduisons une dégradation sur les données à utiliser afin d'approcher la qualité du réseau téléphonique commuté RTC. La fréquence d'échantillonnage devient égale à 8KHz. Nous faisons varier le rapport signal sur bruit SNR de 10dB à 100dB (nous travaillons avec les coefficients LSP, L=32, P=12). Les résultats obtenus sont dans le tableau ci-dessous :

SNR	10	20	30	40	50	60	70	80	90	100
Ic	14.0	37.7	61.4	81.6	89.5	92.1	92.1	92.1	92.1	92.1

Tab. 4.3 Influence de la qualité des données QV.

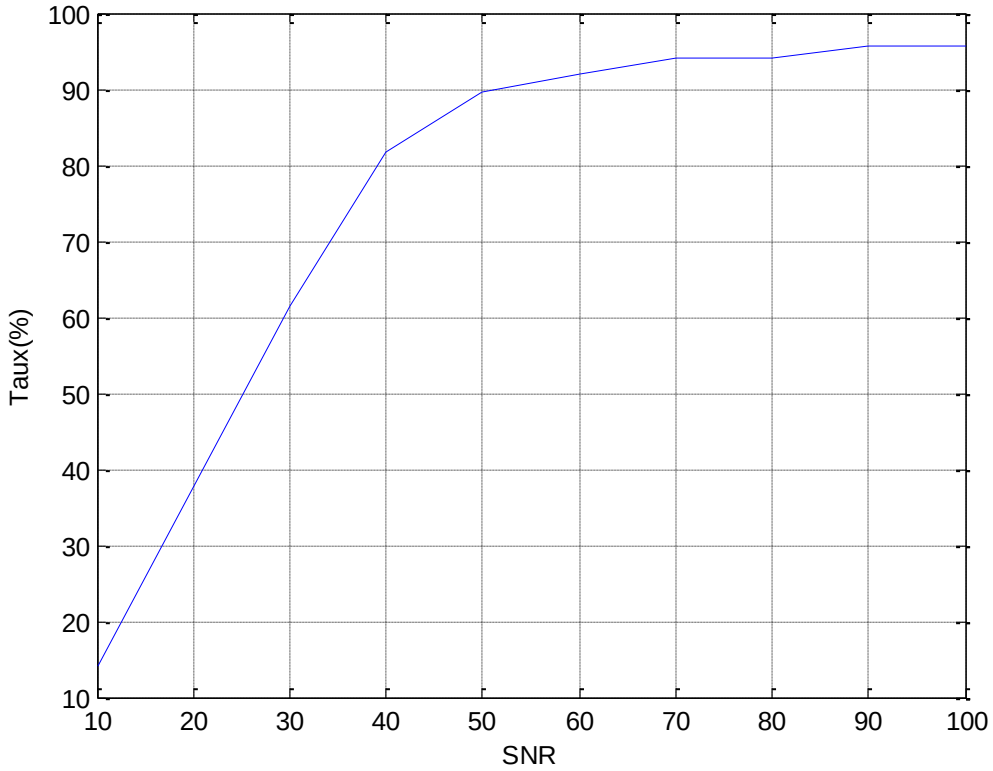

Fig. 4.4 Influence de la qualité des données QV.

D'après ces résultats (figure 4.4), la quantification vectorielle est sensible aux milieux fortement bruités (pour le RTC, la valeur typique du SNR est 40dB). La QV n'est pas robuste au bruit, cependant la performance du système s'améliore en améliorant la qualité des données. La QV est adaptée aux milieux faiblement bruités.

4.5.3 Réseaux de neurones RNA

Nous allons évaluer la performance des MLP et PNN dans un système d'identification du locuteur. L'architecture du réseau et la façon de modéliser les locuteurs, ainsi que la quantité des données d'apprentissage vont jouer un rôle très important.

4.5.3.1 Perceptrons multicouches MLP

Chaque locuteur est modélisé par un MLP ayant une seule sortie. Lors de la phase d'apprentissage, les vecteurs acoustiques extraits du locuteur concerné auront la sortie égale à 1 et les vecteurs de tous les autres locuteurs (37 restants) auront la sortie égale à 0. Le nombre d'entrées R est fixé par la dimension des vecteurs acoustiques. Les fonctions d'activation utilisées sont des fonctions sigmoïdales ($f(x) = \frac{1}{1+e^{-x}}$). L'algorithme d'apprentissage est la rétropropagation du gradient de l'erreur. Lors de la phase du test, les vecteurs acoustiques extraits du segment de test vont être appliqués à tous les MLP des différents locuteurs. Le réseau MLP pour lequel la valeur de la sortie cumulée est la plus grande désigne le locuteur identifié (probabilité à posteriori maximale). L'architecture optimale des MLP qui donne la meilleure performance (nombre de couches cachées, nombre de neurones dans chaque couche, etc) est déterminée de façon expérimentale. L'inconvénient majeur du MLP est sa longue durée d'apprentissage qui augmente avec la quantité d'apprentissage, mais la durée du test est extrêmement courte.

1. Nous avons diminué la durée des segments d'apprentissage jusqu'à 5 secondes pour tous les locuteurs (MLP à une seule couche cachée). Avec des vecteurs acoustiques LSP et R=P=12, le taux d'identification correcte est égal à 72% (performance réduite). Ceci est dû au fait que lors de la phase d'apprentissage, le nombre de vecteurs qui inhibent le MLP (mise à 0) est beaucoup plus supérieur au nombre de vecteurs qui l'excitent (mise à 1). Le réseau va en effet apprendre que des zéros.
2. Pour remédier à ce problème, nous introduisons non pas les vecteurs acoustiques mais les codes vecteurs des dictionnaires de la QV (L=32) de tous les locuteurs qui inhibent le MLP du locuteur concerné. Pour le locuteur concerné, nous utilisons toute la quantité d'apprentissage disponible (7 phrases). Nous avons utilisé des MLP à une seule couche cachée. Nous faisons varier le nombre N de neurones dans la couche cachée. Les résultats obtenus sont dans le tableau ci-dessous :

N	16	32	64	128	256
I_c(%)	83.3	83.3	83.3	84.2	79.8

Tab. 4.4 Influence du nombre de neurones dans la couche cachée MLP.

Le taux d'identification est pratiquement constant de N=16 à N=128. Il diminue pour N=256. Puisque la complexité de l'architecture du réseau fait augmenter le temps d'apprentissage, N=16 assure une bonne performance et une durée plus courte.

3. Nous faisons varier le nombre des entrées R (dimension des vecteurs acoustiques) en maintenant N=16 et L=32. Pour R=24, le taux d'identification correcte devient égal à 98.0% et il reste constant pour R=32, notant que le temps d'apprentissage n'est pas très affecté. C'est un excellent résultat qui démontre la grande utilité et performance des MLP dans l'identification du locuteur. En effet, ce sont d'excellents classificateurs qui garantissent une bonne discrimination et réalisent une séparation non linéaire entre les différentes classes (locuteurs).
4. L'utilisation de deux couches cachées n'apporte rien en plus. Tout au contraire, la performance du système se dégrade (pour N1=N2=16, I_c=70.1%).
5. Nous pouvons penser à modéliser tous les locuteurs par un seul MLP ayant 38 sorties, chaque sortie est associée à un locuteur (les vecteurs acoustiques d'un locuteur donné auront la sortie associée à ce locuteur égale à 1 et toutes les autres sorties égales à 0). La performance du système se dégrade ne dépassant pas 65%. Ceci n'est pas vrai en ce qui concerne les PNN.

4.5.3.2 Réseaux de neurones probabilistes PNN

Les PNN sont des réseaux dédiés à la classification. Leur apprentissage est immédiat (problème des MLP), mais par contre la durée du test est grande vu la complexité de leur architecture. En effet, le nombre de neurones dans la couche cachée à fonction radiale de base est égal au nombre des entrées-sorties désirées de la phase d'apprentissage. Il faut donc diminuer la quantité d'apprentissage tout en obtenant une bonne performance. Nous utilisons un seul réseau PNN ayant R=12 entrées et 38 sorties(les vecteurs acoustiques LSP d'un locuteur donné auront la sortie associée à ce locuteur égale à 1 et toutes les autres sorties égales à 0). Lors de la phase du test, les vecteurs acoustiques extraits du segment de test vont être appliqués à l'entrée du réseau. La sortie dont la valeur cumulée est la plus grande désigne le locuteur identifié.

1. Nous faisons varier la durée des segments d'apprentissage T de 5 à 15 secondes. Les résultats obtenus sont dans le tableau suivant :

T(s)	5	10	15
I_c(%)	83.3	87.7	90.3

Tab. 4.5 Influence de la durée d'apprentissage PNN.

Le taux d'identification augmente (bonne performance) avec T mais de la durée du test devient plus grande.

2. Au lieu d'utiliser les vecteurs acoustiques extraits, nous utilisons les codes vecteurs des dictionnaires de la QV (L=32, P=R=12). Le taux d'identification correcte devient égal à 98.0% (valeur maximale). En plus, la durée du test est courte. Ceci démontre la grande adaptation des PNN à la classification. Avec un seul réseau, nous avons obtenu une excellente performance. Les PNN sont très efficaces dans la classification.
3. Nous pouvons, en plus, diminuer le nombre de vecteurs d'apprentissage (complexité du réseau) en diminuant la taille L des dictionnaires, tout en gardant un très bon taux d'identification. Néanmoins, il faut augmenter le nombre d'entrées R. Pour L=16 et R=24, I_c=94.0%. Pour L=8 et R=32, I_c=93%.

4.5.3.3 Qualité des données d'apprentissage et du test

Nous introduisons une dégradation sur les données à utiliser afin d'approcher la qualité du réseau téléphonique commuté RTC. La fréquence d'échantillonnage devient égale à 8KHz. Nous faisons varier le rapport signal sur bruit SNR de 10dB à 100dB (nous travaillons avec les coefficients LSP, L=32, R=12). Les résultats sont dans le tableau ci-dessous :

SNR	10	20	30	40	50	80	100
I_c(MLP)	13.2	40.5	81.6	89.5	92.1	92.1	94.0
I_c(PNN)	12.3	22.0	40.5	66.6	86.0	89.5	92.1

Tab. 4.6 Influence de la qualité des données MLP et PNN.

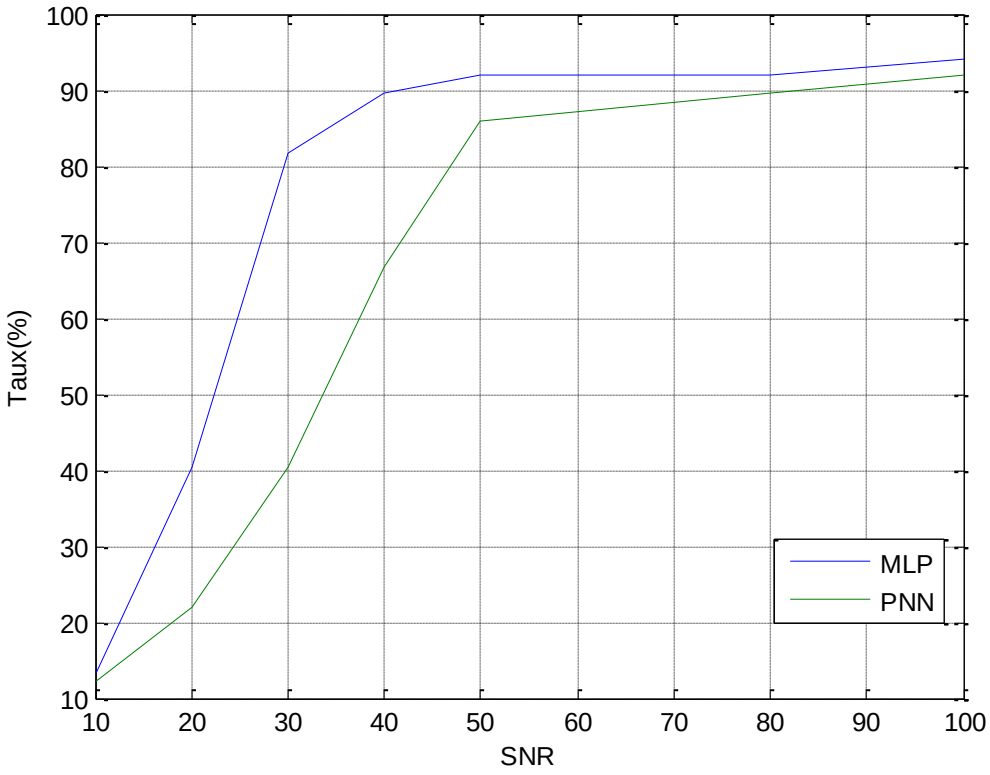

Fig. 4.5 Influence de la qualité des données MLP et PNN.

D'après ces résultats (figure 4.5), les MLP sont bien adaptés à la qualité du RTC et sont plus robustes que la QV et les PNN. Les PNN ont de faibles performances vis-à-vis des données bruitées.

Conclusion

Les réseaux de neurones MLP et PNN améliorent la performance du système (bonne discrimination, séparation non linéaire entre les différentes classes). Pour les MLP, il faut utiliser 38 réseaux. Pour les PNN, il suffit d'utiliser un seul. Les MLP sont bien adaptés aux milieux bruités et offrent une meilleure capacité de généralisation.

4.6 Conclusion

- Les coefficients LSP et LPCC donnent les meilleurs résultats.
- La QV est coûteuse en termes de temps de calcul et l'espace mémoire nécessaire pour le stockage des références des locuteurs.
- Les MLP sont sensibles à la durée d'apprentissage. Leurs paramètres sont difficiles à ajuster, mais leur capacité de généralisation est très bonne.

- Les PNN sont sensibles à la durée du test. Leurs paramètres sont faciles à régler, mais leur capacité de généralisation est moins bonne (pour les données bruitées).
- Si nous voulons augmenter le nombre de locuteurs de la base de données, il faut refaire la phase d'apprentissage des réseaux de neurones (construire de nouveaux réseaux). Dans le cas de la QV, il faut faire l'apprentissage des nouveaux locuteurs seulement (ajouter les dictionnaires de quantification des nouveaux locuteurs).

Conclusion générale

Au cours de ce travail, nous avons traité le problème de l'identification du locuteur en mode indépendant du texte. Il s'agit d'extraire des vecteurs acoustiques, à partir des signaux de paroles prononcés par les locuteurs de la base de données, qui servent à l'entraînement (apprentissage) des modèles représentants chaque locuteur. Nous avons évalué l'utilisation des paramètres MFCC, LSP et LPCC avec deux approches de modélisation (QV et RNA).

Nous avons montré que les vecteurs LSP et LPCC donnent les meilleurs résultats et garantissent une bonne discrimination entre les locuteurs. L'utilisation des RNA fournit une excellente performance du système. Les MLP occupent un large espace mémoire et leur durée d'apprentissage est importante. Les PNN sont sensibles aux données bruitées et leur durée de test peut être longue. Ces contraintes citées apparaissent, principalement, au cours de l'implémentation suivant le type d'application.

Plusieurs points peuvent faire l'objet d'améliorations notables. Nous pouvons utiliser des coefficients acoustiques qui caractérisent le système auditif (modèle de l'oreille gamma-chirp) plutôt de ceux qui caractérisent le système phonatoire. Nous pouvons aussi utiliser, pour la modélisation, l'approche hybride statistique/connexionniste (GMM/RNA) qui est actuellement la plus utilisée. Elle réduit l'espace mémoire nécessaire et elle est très robuste aux milieux bruités.

Annexe A

Estimation des coefficients de prédiction linéaire

Le modèle AR de production de la parole est décrit par :

$$s(n) = \sum_{k=1}^{p} a_k \, s(n-k) + G \, u(n)$$

Ainsi, chaque échantillon de la parole *s(n)* est constitué par une combinaison linéaire des *p* échantillons précédents. Le prédicteur est défini comme un système dont la sortie est:

$$\tilde{s}(n) = \sum_{k=1}^{p} a_k \, s(n-k)$$

L'erreur de prédiction est donnée par :

$$e(n) = s(n) - \tilde{s}(n)$$

On cherche à trouver un ensemble de coefficients a_k de façon à minimiser l'énergie résiduelle de prédiction dans un certain intervalle [1].

L'énergie résiduelle de prédiction est donnée par :

$$E = \sum_n e(n)^2$$

En faisant :

$$\frac{\partial E}{\partial a_i} = 0 \; ; \quad \text{pour } i=1,\ldots, p$$

avec :

$$\frac{\partial \mathrm{E}}{\partial \mathrm{a_i}} = -2\sum_{\mathrm{n}}\{[\mathrm{s(n)} - \sum_{k=1}^{p} a_k\ s(n-k)\ ]\ \mathrm{s(n-i)}\} = 0$$

Cette dernière équation nous conduit à écrire :

$$\sum_{\mathrm{n}} \mathrm{s(n)s(n-i)} = \sum_{n} \sum_{k=1}^{p} a_k\, s(n-k)\, s(n-i)$$

On définit:

$$\varphi(i,k) = \sum_{n} s(n-i)\, s(n-k)$$

Il résulte :

$$\sum_{k=1}^{p} a_k\, \varphi(i,k) = \varphi(i,0)\ ;\ \text{pour } i=1,\ldots,p \qquad (1)$$

Cet ensemble de p équations à p inconnus peut être résolu d'une manière efficace et ainsi trouver les coefficients de prédiction inconnus $\{a_i\}$.

On suppose que le segment de parole est nul en dehors de l'intervalle $0 < n < La-1$, où La est la longueur de la fenêtre d'analyse (méthode dite de l'autocorrélation). Ceci est équivalent à multiplier le signal parole d'entrée par une fenêtre de longueur finie.

$e(n)$ est non nulle uniquement sur l'intervalle $0 < n < La + p - 1$.

Ainsi:

$$\varphi(i,k) = \sum_{n=0}^{La+p-1} s(n-i)\ s(n-k)\ ;\ i=1,\ldots,p;\ \ k=0,\ldots,p$$

On pose: $m = (n - i)$

$$\varphi(i,k) = \sum_{m=0}^{La-(i-k)-1} s(m)\ s(m+i-k)$$

Donc, $\varphi(i,k)$ est l'autocorrélation de $s(m)$ évaluée sur $(i - k)$

$$\varphi(i,k) = R(i-k)$$

Finalement, la relation (1) devient :

$$\sum_{k=1}^{p} a_k\, R(|i-k|) = R(i);\ i=1,\ldots,p$$

On obtient :

$$\begin{pmatrix} R(0) & R(1) & R(2) & \cdots & R(p-1) \\ R(1) & R(0) & R(1) & \cdots & R(p-2) \\ R(2) & R(1) & R(0) & \cdots & R(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & R(p-3) & \cdots & R(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(p) \end{pmatrix}$$

La matrice, de dimension $p \times p$, est une matrice de Toeplitz symétrique, tous les éléments d'une diagonale donnée sont égaux. Cette propriété peut être exploitée pour obtenir un algorithme efficace de résolution du système d'équations.
La solution la plus efficace est une méthode itérative connue sous le nom de l'algorithme de Levinson Durbin [1], [7].

$$\begin{cases} E_0 = R_0 \\ k_i = -\dfrac{R_i + \sum_{j=1}^{i-1} a_j^{(i-1)} R_{i-j}}{E_{i-1}} \\ a_i^{(i)} = k_i \\ a_j^{(i)} = a_j^{(i-1)} + k_i a_{i-j}^{(i-1)} \; ; 1 \le j \le i-1 \\ E_i = (1 - k_i^2) E_{i-1} \end{cases} \quad ; \quad 1 \le i \le p$$

Ainsi :

$$a_i = a_i^{(p)} \; ; \quad 1 \le i \le p$$

et :

$$\sigma = E_p$$

H(z) se met sous la forme :

$$H(z) = \frac{\sigma}{A(z)}$$

avec :

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

Annexe B

Apprentissage des perceptrons multicouches MLP par rétropropagation du gradient

L'algorithme de rétropropagation ou de propagation arrière (Backpropagation en anglais) est l'exemple d'apprentissage supervisé le plus utilisé, à cause de l'écho médiatique de certaines applications spectaculaires, telles que la démonstration de Sejnowski et Rosenberg (1987) dans laquelle l'algorithme est utilisé dans un système qui apprend à lire un texte.

L'algorithme de rétropropagation exige une architecture ayant au moins une couche cachée, de plus la fonction de transfert qui transforme l'activation en réponse au niveau d'une couche cachée doit être non linéaire.

Le perceptron multicouche MLP procède par un entraînement avec des exemples connus. L'ensemble des vecteurs d'apprentissage est défini par $\{(x_p; y_p)$; $p=1,2,...,P\}$. A chaque vecteur d'entrée x_p, le réseau calculera le vecteur de sortie o_p après propagation en avant du stimulus à travers les couches cachées du réseau. Le vecteur o_p est comparé à celui que l'on désire obtenir y_p. On déduit le vecteur d'erreur $(y_p\text{-}o_p)$. On se base sur le critère de minimisation d'une fonction du coût qui est la somme des carrés des erreurs. Elle est exprimée de la manière suivante :

$$E = \frac{1}{2} \sum_{p=1}^{P} \sum_{k=1}^{N} (y_{pk} - o_{pk})^2$$

où :

P : le nombre total des vecteurs d'apprentissage.

N : le nombre de neurones dans la couche de sortie.

L'erreur calculée sera propagée en arrière à travers le réseau, et les poids w_{ij} seront modifiés. L'algorithme de rétropropagation utilise la descente du gradient pour minimiser la distance entre la sortie désirée et la sortie obtenue par le réseau :

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}}$$

η : le pas d'apprentissage.

La formule de mise à jour (procédure itérative) des poids synaptiques sera :

$$w_{ij}^{(t+1)} = w_{ij}^{(t)} + \Delta w_{ij}$$

t : l'indice de l'itération.

L'algorithme de rétropropagation peut être résumé par les étapes suivantes :

1. Initialiser les poids du réseau.
2. Introduire les vecteurs d'apprentissage dans le réseau.
3. Propagation en avant des d'entrée à travers le réseau.
4. Calculer le signal d'erreur entre la sortie réelle et la sortie désirée.
5. Envoyer le signal d'erreur en arrière à travers le réseau.
6. Mise à jour des poids pour minimiser la fonction du coût.
7. Répéter les étapes 2 à 6 jusqu'à ce que l'erreur soit suffisamment petite ou jusqu'a ce que le nombre d'itérations maximal, fixé au préalable, soit atteint.

- **Considérations pratiques**

- Les poids du réseau doivent être initialisés à de petites valeurs aléatoires.
- La valeur du taux d'apprentissage η a un effet significatif sur les performances du réseau, si ce taux est petit l'algorithme converge lentement, par contre s'il est grand l'algorithme risque de générer des oscillations.
- Généralement, η doit être compris entre 0 et 1 pour assurer la convergence de l'algorithme vers une solution optimale.
- Il n'existe pas de règles permettant de déterminer le nombre de couches cachées dans un réseau donné ni le nombre de neurones dans chacune d'elles.

Références bibliographiques

[1] R.BOITE, H.BOURLARD, T.DUTOIT, J.HANCQ, H.LEICH, "Traitement de la parole", Presses polytechniques et universitaires romandes, Lausanne, 2000.
[2] CALLIOPE, "La parole et son traitement automatique ", Masson, Paris, 1989.

[3] Y.GRENIER, Thèse de doctorat "Identification du locuteur et adaptation au locuteur d'un système de reconnaissance phonémique", ENST -E- 77005, Octobre 1977.
[4] J.J. WOLF, "Efficient acoustic parameters for speaker recognition ", JASA – vol 51 – part 2 – June 1972.
[5] J.P.HATON, J.M.PIERREL, G.PERENNOU, J.CAELEN, J.L.GAUVAIN, "Reconnaissance automatique de la parole", Dunod, 1991.

[6] M.KUNT, "Traitement numérique des signaux", Presses polytechniques et universitaires romandes, Lausanne, 1980.

[7] J.P.CAMPBELL, "Speaker recognition: a tutorial", IEEE transactions on speech and audio processing, vol.85, no.9, September 1997.

[8] G.BLANCHET, M.CHARBIT, "Signaux et images sous Matlab", HERMES science publications, Paris, 2001.

[9] M.KUNT, M.BELLANGER, F. COULON, C.GUEGUEN, M.HASLER, N.MOREAU, M.VETTERLI, "Techniques modernes de traitement numérique des signaux", Presses polytechniques et universitaires romandes, Lausanne, 1991.

[10] F.ITAKURA, "Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals ", J.Acoust.Soc.Am, 57, 535(a), s35 (A), 1975.

[11] K.R.FARRELL, R.J.MAMMONE, K.T.ASSALEH, "Speaker recognition using neural networks and conventional classifiers", IEEE transactions on speech and audio processing, vol.2, no.1, part 2, January 1994.
[12] A.V.OPPENHEIM, R.W.SHAFFER, "Digital signal processing", Prentice Hall, New Jersey, 1975.
[13] H.TAKHEDMIT, N.AIT SAADI, Projet de fin d'études "Identification du locuteur en mode indépendant du texte", Département d'Electronique, ENP, Juin 2005.

[14] M.BOUCHAMEKH, H.HADJ-ALI, Projet de fin d'études "Identification du locuteur en mode indépendant du texte", Département d'Electronique, ENP, Juin 2004.

[15] L.RABINER, "A tutorial on hidden Markov models and the speech signal", Proceedings of the IEEE, vol.77, no.2, 1989.

[16] L.RABINER, B-H.JUANG, "Fundamentals of speech recognition", Prentice Hall, New Jersey, 1993.

[17] B.GOSSELIN, Thèse de doctorat "Application des réseaux de neurones artificiels à la reconnaissance de caractères manuscrits", Sciences appliquées, Faculté Polytechnique de Mons, 1996.

[18] H.DEMUTH, M.BEALE, M.HAGAN, "Neural networks toolbox: user's guide", The Math Works, 2009.

[19] M.Y.AMMAR, Thèse de doctorat "Mise en œuvre de réseaux de neurones pour la modélisation de cinétiques réactionnelles en vue de transposition batch/continu", Génie des procédés et de l'environnement, Institut National Polytechnique de Toulouse, Juillet 2007.

[20] S.HAYKIN, "Neural networks - A comprehensive foundation", Macmillan College Publishing Company, New York, 1994.
[21] R. LIPPMANN, "An Introduction to computing with neural nets", IEEE ASSP magazine, volume 4, n° 2, pp 4-22, Avril 1987.
[21] T. KOHONEN, "The Self-Organising Map", Proc. of the IEEE, vol 78, n°9, pp 1464-1480, September 1990.
[22] G.DREYFUS, J.M.MARTINEZ, J.M.M.SAMUELIDES, M.B.GORDON, F.BADRAN, S.THIRIA, L.HERAULT, "Réseaux de neurones : Méthodologie et applications", Eyrolles, 2001.