UNIVERSITE SAÂD DEHLEB DE BLIDA

Faculté Des Sciences de l'Ingénieur

Département D'électronique

MÉMOIRE DE MAGISTER

Spécialité : Image et Parole

RECONNAISSANCE DES FORMES PHONÉMIQUES EN ARABE

Par

Mr FERROUGA Mohamed

Soutenu le 20 Mai 2010

Devant le jury composé de :

Mr A. GUESSOUM	Professeur	USD-Blida	Président
M^{me} M. GUERTI	Professeur	ENP Alger	Rapporteur
Mr N. KHORISSI	Chargé de cours	USD-Blida	Co-rapporteur
M^{me} N. BENBLIDIA	Maître de conférences	USD-Blida	Examinatrice
Mr Z. BENSLAMA	Maître de conférences	USD-Blida	Examinateur

ملـخص

إن الإدراك الصوتي كان موضوع لعدة بحوث خاصة في اللغات الأجنبية، و بالأخص اللغة الانجليزية. لكن اللغة العربية بقيت بعيدة نوعا ما عن هذه البحوث، و لإدراك منا لضرورة هذه البحوث، ارتأينا أن نقوم بمحـاولة إنشـاء برنـامج كمبيوتر يسـمح بإدراك الصوت و الصورة معـا و ذلك فقـط لـ12 لفـظ من ألفاظ اللغة العربية من مجموعة أصوات مكونة من 2160 عينة مسجلة على أساس 11025 Hz و مقسمة يدويا.

قسمنا هذا العمل إلى ثلاثة أقسام أساسية :

- أولا، إدراك الصوت فقـط، حيث تم استخـراج العوامـل الصوتية MFCC و تم الإدراك الصوتي بواسطة الشبكات العصبونية من نوع MLP-TDNN.
- ثانيا، الإدراك البصـري للألفـاظ، حيث تم تحديد كل الصور الموافقة للألفاظ أو ما يسمى باللغة الأجنبية بـ viseme، وذلك باستعـمال الشبكـات العصبونية من نـوع MLP وAutoassociator من مجـموعة صور مكونة من 4853 عينة.
- ثالثا، ألإدراك السمعي البصري للألفاظ، حيث تم الإدماج بين العوامل الصوتية و البصرية فيما يعرف بالإدماج المباشر و هذا باستعمال الشبكات العصبونية من نوع MLP-TDNN.

نسبة الإدراك الصوتي بلغت 76,69% من مجموع 2076 عينة. فيما بلغت نسبة الإدراك البصري 84,98% و 84,69% بالنسبة لـ MLP و Autoassociator تباعا. أما الإدراك السمعي البصري، فقد بلغت نسبته 85,83%.

كلمات مفاتيح: إدراك الأشكال، المعالجة الآلية للغة، اللغة العربية، الإدراك السمعي البصري للكلام، صورة اللفظة (فيزام)، الشبكات العصبونية.

RÉSUMÉ

La reconnaissance audiovisuelle de la parole fait l'objet de plusieurs travaux de recherche, cependant peu de recherches ont été faites pour la langue Arabe. De ce fait, nous avons opté pour réaliser un système de reconnaissance audiovisuelle de 12 phonèmes en Arabe Standard (AS) constituant un corpus audio de 2160 échantillons à 11025 Hz. La segmentation a été faite manuellement.

Le travail est divisé en trois grandes parties, la reconnaissance :

- acoustique où les paramètres acoustiques MFCC ont été extraits. Les Réseaux de Neurones Artificiels à délai et multicouches ou Time Delay Neural Network et Multi Layer Perceptron (TDNN-MLP) ont été utilisés pour la reconnaissance;
- visuelle où nous avons d'abord défini les visèmes de l'AS. Cette partie est aussi faite par les RNA de type MLP et Autoassociateur, et ceci sur un corpus de 4853 échantillons de visèmes arabes;
- audiovisuelle où nous avons fait appel à une technique d'intégration précoce dite intégration directe au niveau des paramètres acoustiques et visuels. La reconnaissance audiovisuelle est basée sur les RNA de type TDNN-MLP.

Le taux de reconnaissance pour l'acoustique seul est de 76,69% d'un corpus de 2076 échantillons. Pour la reconnaissance visuelle, le taux de reconnaissance est de 84,98% et 84,69% pour MLP et Autoassociateur respectivement. Quant au taux pour la reconnaissance audiovisuelle est de 85,83%.

Mots clés : Reconnaissance De Formes, Traitement Automatique de la Parole, Arabe Standard, Reconnaissance Automatique de la Parole Audio-Visuelle, Visème, Réseaux de Neurones Artificiels.

ABSTRACT

Audiovisual speech recognition is widely studied in many works for many languages. Alas, Arabic is deprived from such works. Therefore, and as Arabic is our mother tongue, we have seen that it is our duty to realize such system. In fact, it consists to recognize 12 standard Arabic phonemes recorded on 11025 Hz constituting a corpus of 2160 samples manually segmented.

This work is divided in three important parts:

- acoustic recognition, where MFCCs (acoustics features) are firstly extracted , then recognized by means of artificial neural network TDNN and MLP;
- visual recognition, where firstly we described Arabic visemes. We used MLP and autoassociator neural network to recognize 4853 samples of Arabic visèmes;
- audiovisual recognition, in which acoustic parameters (MFCCs) are combined to visual ones (image of visemes) in precocious integration known as direct integration and then recognized by neural network TDNN-MLP.

The acoustic recognition rate is 76.69% for corpus constituted of 2076 samples. Visual recognition rate is 84.98% and 84.69% using MLP and Autoassociator respectively. Regarding audiovisual recognition, the rate is 85.83%.

Keywords: Pattern recognition, automatic speech processing, Arabic language, acoustic speech recognition, audiovisual speech recognition, viseme, neural network

DEDICACES

Je dédie ce modeste travail à mes chers parents

A tous mes frères et sœurs

A toute ma famille.

A tous mes amis.

REMERCIEMENTS

Je remercie tout d'abord Le Bon Cher Dieu Allah qui m'a donné la force pour achever ce travail.

Mes sincères remerciements à ma directrice du mémoire de Magister, Mme Guerti Mhania, Professeur à l'École Nationale Polytechnique d'Alger, qui était très patiente et m'a donné un peu de son précieux temps pour m'expliquer les objectifs de ce travail et qui progressivement l'a augmenté jusqu'à son heureux achèvement, où un moment donné j'étais sur le point de tout abandonner.

Je tiens à remercier aussi M Nassr-Eddine Khorissi, Chargé de cours au Département d'Électronique à l'USD de Blida, mon codirecteur, pour son patience, ces nombreux conseils et orientations qui ont été fructueux pour ce travail.

Je remercie Mr. A. Guessoum, Professeur et Directeur du Laboratoire de Recherche en Traitement de Signal et Imagerie (LATSI) au Département d'Électronique à l'USD de Blida, qui m'a fait l'honneur de présider ce jury.

Je remercie également Mlle N. Benblidia et Mr. Z.A. Benslama, Maîtres de conférences au Département d'Électronique à l'USD de Blida pour avoir accepté d'évaluer ce travail et pour leurs remarques constructives.

TABLE DES MATIERES

CHAPITRE 4 :
RECONNAISSANCE DES PHONÈMES ET RESULTATS

LISTE DES ILLUSTRATIONS, FIGURES ET TABLEAUX

LISTE DES ABREVIATIONS

API	Alphabet Phonétique International
AS	Arabe Standard
ASR	Automatic Speech Recognition
DAP	Décodage Acoustique Phonétique
DFT	Discret Fourier Transform
DTW	Dynamic Time Wrapping
EM	Estimation Maximisation
HM	Homme Machine
HMM	Hidden Markov Model
ID	Intégration Directe
IS	Intégration Séparée
K-PPV	K Plus Proche Voisins
LPC	Langage Parlée Complétée
MFCC	Mel Frequency Cepstral Coefficient
MLP	Multi Layer Perceptron
OCR	Optical Character Recognition
RAP	Reconnaissance Automatique de la Parole
RAPAV	Reconnaissance Automatique de la Parole Audio Visuelle
RDF	Reconnaissance Des Formes
RN	Réseau de Neurones
RNA	Réseau de Neurones Artificiels
S/B	Signal/Bruit
TAL	Traitement Automatique du Language
TAP	Traitement Automatique de la Parole
TDNN	Time Delay Neural Network
TF	Transformée de Fourier
TFD	Transformée de Fourier Discrète
TFI	Transformée de Fourier Inverse
TOP	Transcription Orthographique Phonétique
[VCV]	Voyelle Consonne Voyelle
DARPA	Defense Advanced Research Agency

INTRODUCTION

La parole est le moyen de communication le plus important entre les êtres humains. Les personnes parlent entre elles d'une manière assez simple et naturelle. Cette spontanéité de communication est derrière la motivation des chercheurs à concevoir des systèmes de Reconnaissance Automatique de la Parole (RAP) aussi naturels que possible de telle sorte qu'il soit possible de communiquer avec son ordinateur comme c'est le cas avec une autre personne. La conception de tels systèmes nécessite un traitement automatique spécifique et complexe de la parole brute pour en extraire toutes les informations utiles et pertinentes. Le traitement automatique de la parole est une des technologies déterminantes pour le développement d'interfaces Homme-Machine avancées. Toutefois, malgré les avancées très importantes pendant ces dernières années, les systèmes sont encore très loin d'égaler les capacités langagières des êtres humains.

Le Traitement Automatique de la Parole (TAP) est un domaine en progression rapide, dont la partie la plus visible a été l'apparition des logiciels commerciaux de dictée vocale. Les nouvelles applications émergentes sont les systèmes de dialogue évolués et l'indexation multimédia, en combinaison avec l'extension aux documents audio des techniques de recherche d'information [1]. Le dialogue Homme-Machine (HM) essaie de mettre en place un traitement automatique de la langue qui peut servir d'interface entre la machine, ou une application, et l'Homme [2]. Un système de dialogue oral HM a pour objectif de donner à l'utilisateur l'information qu'il recherche en s'aidant de diverses sources de connaissances, et le système de question-réponse de rechercher une réponse précise à une question. Le terme de système de dialogue indique généralement un système permettant une interaction entre un être humain et un système dans un cadre restreint. Toutefois, notamment dans le cas des travaux sur les systèmes de question-réponse, le cadre tend à s'élargir. Un dialogue est une suite d'échanges entre interlocuteurs dans un contexte donné. Un système de dialogue HM interprète les requêtes de l'utilisateur en fonction de la tâche à accomplir, de l'histoire du dialogue et du comportement de l'utilisateur. Son objectif est de donner à l'utilisateur les informations recherchées tout en assurant une interaction efficace et naturelle.

Actuellement, Les systèmes de dialogue orales concernent des domaines restreints tels que l'information horaire de moyens de transports (trains, avions, cinéma) ou les informations

touristiques, par exemple le projet européen Le 3-Arise (horaires de train), le projet américain ATIS du DARPA Communicator (voyages en avions), et le projet français Techno langue MEDIA (informations touristiques). Ces systèmes permettent une interaction orale relativement naturelle : l'utilisateur peut à tout moment changer d'avis et revenir sur des choix exprimés, interrompre la réponse du système en prenant la parole, le système peut lui aussi changer de stratégie d'interaction en fonction des réactions de l'utilisateur. Un système de dialogue utilise des sources de connaissances diverses et complexes : connaissances acoustiques, phonétiques, lexicales, morphologiques, syntaxiques et sémantiques, pragmatiques, ainsi que des connaissances sur le dialogue, la tâche à réaliser et sur l'interlocuteur [3].

Diverses solutions peuvent être employées pour améliorer l'ergonomie et faciliter l'utilisation des applications où l'homme interagit avec la machine. Dans de nombreuses situations, l'utilisation de la modalité acoustique peut être avantageuse par rapport à l'utilisation d'un clavier ou d'une souris [4]. Cependant, il est possible d'améliorer cette interaction en ajoutant à la modalité acoustique la modalité visuelle de la parole. Autrement dit, le système de dialogue est non seulement appelé à faire la reconnaissance vocale, mais de faire en plus la reconnaissance visuelle de la parole. De ce fait, s'il est appelé à rechercher par exemple une information audiovisuelle, il prend l'information auditive et l'information visuelle qui est en l'occurrence les images vidéos accompagnant la parole.

Les premiers travaux de RAP pour la langue arabe ont été focalisés sur la reconnaissance de l'Arabe standard moderne. Certains de ces travaux ont été faits dans le cadre de développement des applications de dictée vocale comme le système ViaVoice d'IBM. Récemment, BBN Arabic Broadcast, ont développé un système pour la reconnaissance des mots lors de lancement des informations diffusées atteignant un taux d'erreur de 15%.

BBN ont également développé un système de reconnaissance d'une conversation téléphonique ou le dialecte égyptien a été utilisé avec un taux de reconnaissance compris entre 61,6% et 71,1%. Un autre travail plus récent, en Octobre 2004, a été fait par l'équipe de Perceptual Science Laboratory à l'Université de California aux USA dont S. Ouni fait partie de l'équipe. Ils ont fait une extension d'un système existant d'une tête parlante connu sous le nom de Baldi. Il s'agit d'une application de synthèse de la parole à partir du texte et l'animation faciale en plusieurs langues. Une langue peut être ajoutée en définissant ses

phonèmes et visèmes correspondants. Le plus important c'est qu'ils ont fait intégrer la langue Arabe parmi les autres langues ; donc une application importante pour la synthèse de la parole visuelle de la langue Arabe. C'est l'unique travail de recherche que nous avons trouvé dans la littérature où la synthèse de la parole visuelle en Arabe a été utilisé en définissant les visèmes en Arabe à partir des autres langues comme par exemple l'Anglais où la synthèse de la parole visuelle à été largement utilisée et bien avancée [5].

Le travail que nous présentons dans ce document a donc pour but, de concevoir un système de dialogue HM en AS où les deux modalités de la parole sont intégrées pour améliorer le niveau d'intelligibilité et donc faciliter la compréhension de ce qu'une personne dit par la machine. Il s'agit d'un nouvel axe de recherche pour l'AS englobant à la fois le Traitement Automatique de la Parole (TAP) et de l'image présentée dans les séquences vidéo. Dans ce mémoire n'est présenté que la partie s'occupant de la Reconnaissance de la Parole. Pour cela, nous avons choisi de diviser le travail en 4 chapitres :

Le premier est consacré à une description brève de la Reconnaissance Des Formes (RDF) et les différentes techniques utilisées : les Réseaux de Neurones (RN), la décision bayésienne, les chaînes de Markov cachées ou Hidden Markov Model HMM, …, etc.

Le second présente des généralités sur la nature de la parole, sa production, son traitement automatique, sa présentation dans la machine ainsi que les différents outils techniques pour l'analyser et le représenter de manière à garder l'information utile et filtrer l'information redondante. Une courte introduction est faite sur l'Arabe Standard (AS).

Le troisième expose un état de l'art sur la Reconnaissance de la Parole, les différentes techniques utilisées, la Reconnaissance Automatique de la parole Audio-Visuelle (RAPAV), les différentes intégrations des modalités acoustique et visuelle.

Le dernier où nous avons décrit en détail la technique que nous avons choisie pour la RAPAV qui est basée sur les RN de type TDNN/MLP et bien entendu une présentation des résultats obtenus.

Enfin, nous terminons notre travail par des conclusions générales et des perspectives.

CHAPITRE 1 :
GÉNÉRALITÉS SUR LA RECONNAISSANCE DE FORMES ET RESEAUX DE NEURONES

1.1. Introduction

L'idée de construire des machines capables de simuler des êtres humains afin de les aider dans certaines tâches, voir de les remplacer, était antérieure aux ordinateurs. Leur apparition a permis d'étendre le spectre des tâches à simuler en ajoutant celles dont l'exécution relève de facultés mentales comme la perception et le raisonnement.

De nos jours, il y a beaucoup de motivations pour concevoir des systèmes de traitement automatique de données (documents, images, sons, etc.). Des étapes géantes ont été faites pendant la dernière décennie, en termes d'appuis technologiques et dans des produits de logiciels. L'emploi des ordinateurs permet de se délivrer des contraintes de type mathématique (on peut calculer approximativement l'intégrale d'une fonction sans avoir la primitive), et de ce fait, facilite l'automatisation de traitements [54].

Ce premier chapitre présente des généralités sur la Reconnaissance Des Formes et quelques techniques d'analyse et de reconnaissance.

1.2. Reconnaissance Des Formes (RDF)

La RDF, est la science qui réunit l'ensemble des techniques informatiques de représentation et de décision permettant aux machines de simuler quelques comportements similaires à ceux des humains en leur offrant le privilège d'être qualifiés comme étant des machines intelligentes [54].

La RDF a donc pour objet général, d'une part, de capter et de décrire en mémoire des formes, c'est-à-dire, les manifestations de l'univers extérieur auxquelles la machine a été rendue sensible et, d'autre part, de prendre sur la représentation mémoire ainsi obtenue une décision d'identification par référence à un ensemble d'apprentissage décrit dans une représentation analogue. La partie essentielle est évidemment le choix de cette représentation mémoire qui doit être à la fois assez informative pour permettre une bonne précision à l'identification, et assez condensée pour éviter une redondance inutile qui se traduirait par un temps de calcul exagéré au moment de la décision [6].

La RDF fait appel à de nombreuses disciplines : des mathématiques (probabilités, statistiques, topologie, traitement du signal, etc.), et les problèmes à résoudre sont très variés : reconnaissance de caractères et plus généralement de l'écriture, manuscrite ou imprimée, en direct (avec une tablette graphique, par exemple), reconnaissance optique de documents

numérisés Optical Character Recognition (OCR), vision en robotique, reconnaissance vocale (dictée vocale, reconnaissance des mots, identification des personnes), etc. [7].

Parallèlement aux travaux sur les méthodes de reconnaissance, se développaient le traitement d'images, la vision par ordinateur et le traitement de la parole. Ces domaines ont focalisé le problème de reconnaissance sur des données spécifiques, mais par ailleurs, ils ont permis de situer la reconnaissance dans un processus plus vaste d'interprétation d'image ou de compréhension de la parole, impliquant des niveaux de perception et des connaissances propres au domaine [8].

Une forme est une représentation simplifiée du monde extérieur acceptable par l'ordinateur (fig.1.1).

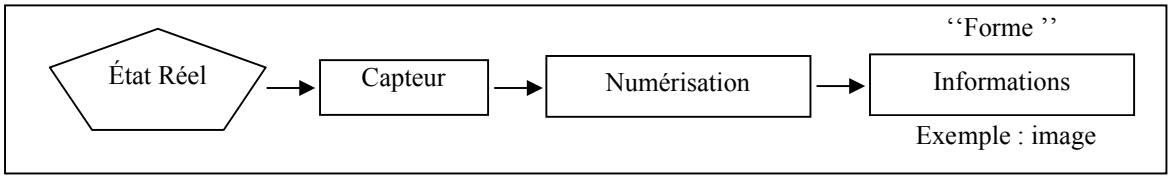


Figure 1.1 : Extraction de la forme

1.2.1. Système de RDF

Un système de RDF peut comporter une phase d'apprentissage dans laquelle le système est appelé à apprendre à reconnaître des formes sur la base d'échantillons. Lorsque cette phase est achevée, le système sera alors prêt à fonctionner pour reconnaître des formes inconnues qui lui seront soumises, c'est la phase de reconnaissance.

Mais un système de RDF peut être aussi un système qui trie (fait des paquets homogènes suivant certains critères) un ensemle de formes inconnues. Il n'y a alors pas d'apprentissage à proprement parler.

Nous pouvons schématiser en général un système de RDF avec une phase d'apprentissage (fig.1.2).

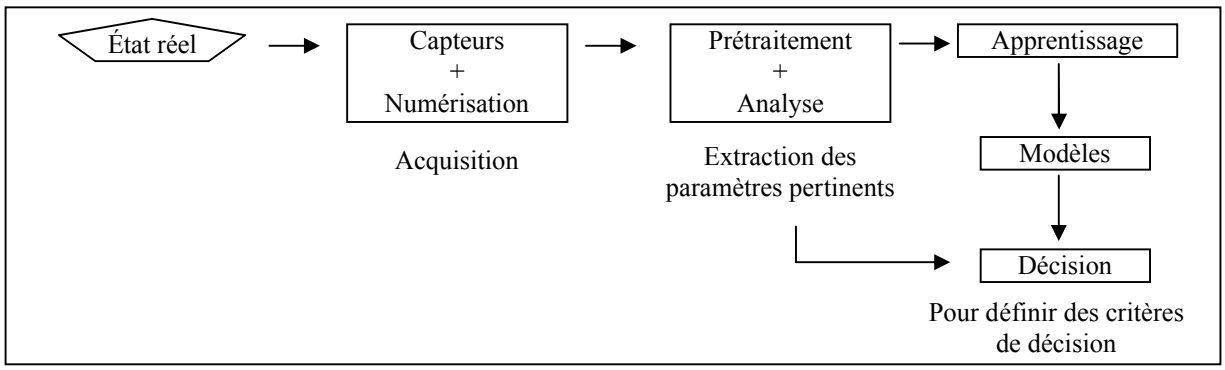


Figure 1.2 : Schéma général d'un système de RDF [9]

Dans la pratique, un système de RDF s'éloigne souvent de ce schéma. Des traitements en amont sont souvent nécéssaires pour isoler la forme à reconnaître de son contexte, ce qui est, en soi, un problème de reconnaissance (segmentation objet/fond, délimation d'une forme dans un ensemble). Des traitements ultérieurs sont utiles pour valider les décisions et éventuelllement les remettre en cause.

1.2.1.1. Prétraitement et analyse

Les données brutes issues des capteurs sont les représentations initiales des données à partir desquelles des traitements permettent de construire celles qui seront utilisées pour la reconnaissance. Elles sont souvent bruitées, contennant des informations parasites, elles n'explicitent pas les informations utiles pour la reconnaissance. Pour les prétraitements, le concepteur s'aide des connaissances qu'il possède sur les capteurs, les types de données, le problème posé et les méthodes d'apprentissage et de reconnaissance qu'il utilisera. Les prétraitements sont utiles pour éliminer des bruits qui peuvent être dûs aux capteurs ou à des interférences avec d'autres sources de signaux (la parole en milieu sonore, l'encre du verso qui traverse le papier et dont la trace est visible sur la feuille du manuscrit, les fonds imagés des chèques, etc.). L'analyse sert à représenter la forme suivant la méthode de la reconnaisssance et sous une forme simplifiée, c'est l'extraction des paramètres, paramètres contour pour une forme image , par exemple [8].

1.2.1.2. Apprentissage

Dans le cas d'apprentissage, il s'agit en fait, de fournir au système un ensemble de formes connues ou inconnues (les classes sont connues) pour que le système trouve pour chaque classe des exemples d'entrainement un modèle ou un représentant. Ce modèle doit capturer les attributs caractéristiques de la classe [10]. Si on prend le cas de la figure 1.3, il s'agit de trouver un modèle de l'état réel en prenant l'erreur entre la sortie actuelle et l'entrée comme paramètre à optimiser.

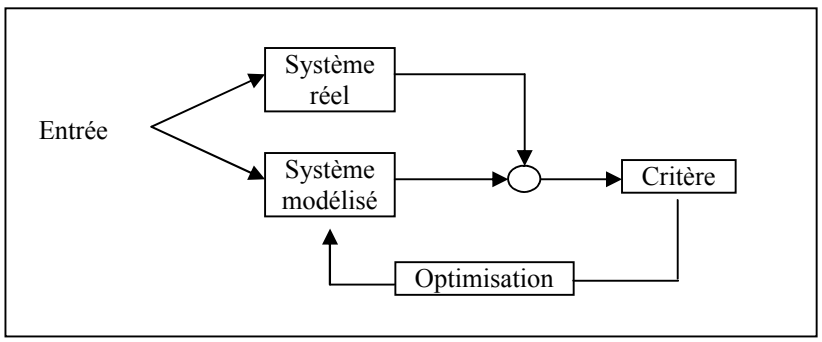


Figure 1.3 : Exemple d'un système d'apprentissage

On peut différencier deux types d'apprentissage :

- supervisé ou avec professeur, dans lequel le nombre de classes est connu et on connaît la classe de chaque forme de l'ensemble d'échantillons ;
- non supervisé ou sans professeur, dans lequel le nombre de classes est connu ou non, et on ne connaît pas la classe des échantillons (ce qu'on sait c'est seulement que l'ensemble des échantillons représente ce qu'on doit connaître) [11].

1.2.1.3. Décision

Dans le cas de la simulation d'une reconnaissance humaine, le maître existe et la décision peut être prise après avoir fait une caractérisation du motif inconnu et par la suite la discrimination de ce même motif par rapport à d'autres motifs des autres classes. La caractérisation consiste à savoir si la forme possède les caractéristiques d'une classe ou non (exemple : un électrocardiogramme appartient à la classe "normale ", rechercher un mot clé dans une séquence parlée). Pour la discrimination, il s'agit d'assigner une forme inconnue à une des classes possibles (exemple de la reconnaissance des caractères, des phonèmes).

La décision se fait par comparaison des résultats de la reconnaissance automatique aux étiquettes données par le maître. On en tire donc les taux de reconnaissance et les taux d'erreurs. On peut aussi avoir des taux de rejet qui correspondent à la décision de ne pas classer la forme. Le système, en évaluant un critère de décision, peut assigner une forme à une classe, mais il peut aussi déterminer avec quelle confiance il effectue cette décision. Si le critère de décision prend des valeurs très proches pour plusieurs classes, la confiance dans la décision est faible [8].

1.2.2. Représentation de la forme

Le système de RDF doit possèder des représentations des formes à classer et des classes. Les types de représentation sont déterminés par les méthodes utilisées. Les contenus de représentations sont déterminés par les buts de reconaissance. Certains types de représentation sont mieux appropriés à exprimer certains contenus. Par exemple, une variabilité interclasse qui se présente comme une répartition aléatoire des caractéristiques de la forme autour de valeurs fortement probables, sera bien représentée par des modèles statistiques. Par contre, les structures (les relations entre composantes d'une forme) peuvent plus facilement s'exprimer par des représentations structurelles.

On appelle caractéristique ou descripteur une information qui peut être mesurée sur la donnée à reconnaître. Par exemple : l'amplitude moyenne d'un signal sur une fenêtre

temporelle, l'energie dans une bande de fréquence, le rapport hauteur sur largeur d'un caractère manuscrit, le niveau de gris moyen d'une zone d'image, etc.
On appelle primitive une composante élémentaire d'une forme. Les primitives ne sont pas décomposables. Par exemple : un segment de droite, une boucle, etc [8].

1.2.3. Les méthodes de la RDF

Les méthodes de RDF sont souvent regroupées en grandes classes identifiées par : statistiques, structurelles et hybrides. À ces classes correspondent différentes manières de représenter les exemplaires et les classes et différentes méthodes pour l'apprentissage et la reconnaissance. Mais elles correspondent aussi aux différentes façons d'aborder le problème de la RDF. On s'intéresse ici à la simulation d'un être humain dans une tâche de reconnaissance.

1.2.4.1. Méthodes statistiques

On considère ici la situation où il existe un ensemble de réalisations étiquettées permettant un apprentissage supervisé (par opposition aux méthodes qui partent d'une connaissance nulle sur les étiquettes des réalisations et cherchent à les grouper sur des critères de ressemblance). Les exemplaires des classes correspondent aux observations d'une variable aléatoire X. Chaque réalisation x est représentée par un vecteur de R^n. Chaque composante du vecteur correspond à un discripteur.

L'objectif est toujours d'assigner une réalisation inconnue à sa classe d'appartenance en minimisant l'erreur de décision. Ce problème peut être résolu de nombreuse manières, le choix de la méthode dépend en partie des connaissances a priori que l'on a sur les distributions de probabilité des exemplaires des classes.

Le cas paramétrique correspond à une connaissance a priori sur les lois de probabilité dont il faut estimer les paramètres. La théorie de la décision bayésienne est la théorie fondamentale des méthodes stochastiques. Le point central de cette théorie est la règle de Bayes qui permet en fait de choisir l'hypothèse ayant la probabilité la plus élevée [8].
La décision qui minimise la probabilité d'erreur globale est celle qui associe à chaque point x de R^n la classe dont la densité de probabilité en x est la plus forte.
Règle de Bayès :

$$P(\omega_i / x) = \frac{p(x/\omega_i)\ P(\omega_i)}{p(x)} \qquad (1.1)$$

Avec $$p(x) = \sum_{i=1}^{s} p(x/\omega_i)\ P(\omega_i) \qquad (1.2)$$

$\{\omega_i\}, i = 1, \ldots, s$ représente les classes.

x$\in R^n$:space de représentation.

Problèmes à estimer $P(\omega_i)$ *et* $p(x/\omega_i)$.

Les méthode statistiques peuvent être divisées en :

- bayésiennes paramétriques pour lesquelles on estime les paramètres des lois de probabilité connues à priori (lois normale en général) ;
  estimateur de la moyenne :
  $$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad (1.3)$$
  et estimateur de la matrice de covariance :
  $$\hat{S} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{\mu})(x_i - \hat{\mu})^t \qquad (1.4)$$
- bayésiennes non paramétriques dans lesquelles les échantillons permettent d'éstimer les paramètres de la loi de densité de probabilité de chaque classe, mais toutes les lois de distribution classiques sont unimodales, elles ne peuvent donc prétendre représenter toutes les possibilités pour les classes le cas des fenêtres de Perzen [12].

  D'autre part, faire des hypothèses erronées sur une classe peut amener à des résultats catastrophiques, comme dans la figure 1.4, où l'on dispose des échantillons répartis dans un espace des observations supposé l'ensemble des nombre réels *R*, et où l'on fait l'hypothèse que la loi suivie par la classe est normale, alors qu'en réalité les échantillons sont répartis autour de deux échantillons différents.

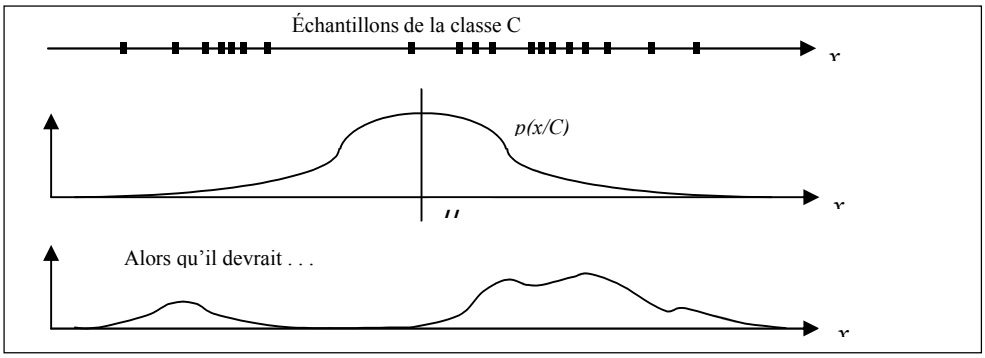


Figure 1.4 : L'effet des hypothèses erronées sur une classe

Les méthodes non paramétriques prennent en compte les échantillons et sutout leur répartition dans l'éspace des paramètres pour obtenir des lois de densité de probabilités plus proche de la réalité de la classe [8].

- paramétriques non bayésiennes ; il s'agit de construire des surfaces séparatrices des classes sans estimer les probabilités. La surface séparatrice a une forme connue, il faut estimer les paramètres qui optimisent la décision par la méthode de gradient par exemple fig.1.5.

  l'apprentissage consiste à determiner un hyperplan defini par :

  $\forall x \in \omega_1, \sum_i a_i x_i + a_0 > 0$

  $\forall x \in \omega_2, \sum_i a_i x_i + a_0 < 0$

  $\omega_1 et\ \omega_2$ deux classes différentes.

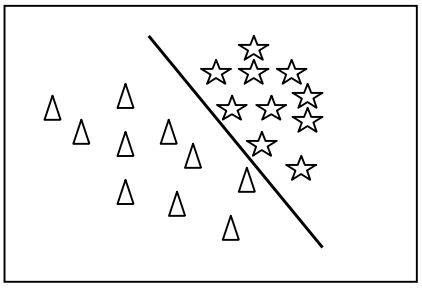


Figure 1.5 : Surface séparatrice

  décision : positionnement du point inconnu par rapport aux surfaces [13].

- non paramétriques et non bayésiennes ; il s'agit de rechercher une partition de l'espace de représentation sans rechercher les lois de probabilités ni les surfaces de séparation. La méthode des k-plus-proches-voisins (k-ppv) en est un exemple :

  dans l'espace de représentation R^n, les k-ppvs de la réalisation inconnue définissent le sous-espace de décision. La décision se fait par un vote, la réalisation inconnue hérite de l'étiquette de la classe la plus représentée dans ce sous-espace [11].

On peut récapituler les diffréntes méthodes statistiques dans ce tableau :

Tableau 1.1 : Les méthodes statistiques [13]

	Méthodes bayésiennes	Méthodes non bayésiennes
Méthodes paramétriques	Estimation de gaussienne	Séparation lineaire
Méthodes non paramétriques	Fenêtre de Parzen	K plus proches voisins

1.2.4.2. Méthodes structurelles

On peut dire, pour dégager l'essentiel de la diférrence entre méthodes statistiques et méthodes structurelles, que d'une part, une donnée est représentée par des vecteurs de caractéristiques, et que d'autre part, elle est représentée par une description sous forme de composantes et de relations entre ces composantes [8].

La notion de structure, quoique sujette à de nombreuses définitions, fait toutefois apparaître l'existence d'une décomposition du toute la structure en parties et de relations entre

ces parties. Les méthodes structurelles de RDF sont plus appropriées pour traiter des structures visuelles où les composantes sont liées par des relations spaciales. La décomposition d'une forme peut faire apparaître une hiérarchie de composantes. On appelle primitives les composantes élémentaires qui ne peut plus être décomposée ; le niveau d'information de ces primitives est fixé par le concepteur du système de reconnaissance.

Une première phase de la reconnaissance sera donc de ramener la donnée brute à un ensemble de composantes, c'est-à-dire, de la segmenter en fragments, suivant un ou plusieurs niveaux de décomposition, fig.1.6. Les traits descriptifs d'une structure sont de type fragment mais aussi des traits qui expriment des configuration comme la symétrie ou la fermeture, ou des événements qui se produisent sur la forme comme un changement de direction au niveau d'un point anguleux ou encore des caractéristiques comme le rapport hauteur sur largeur, les moments, le centre de gravité, etc.

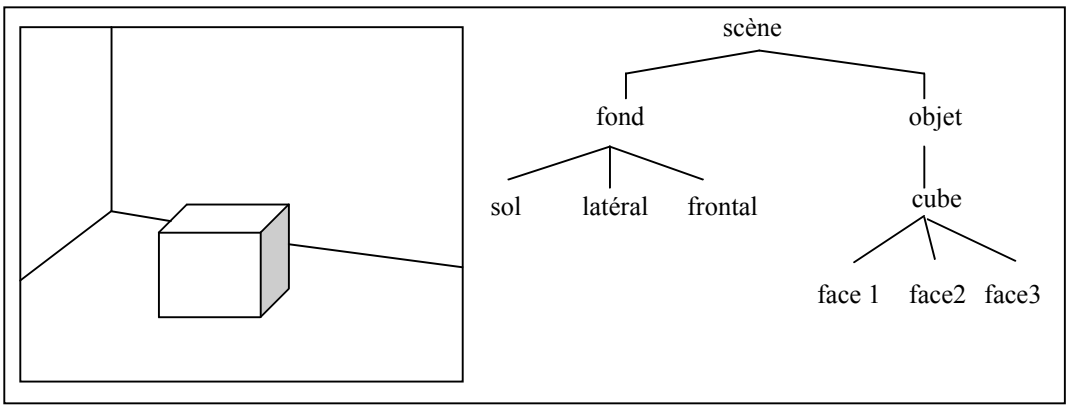


Figure 1.6 : Décomposition d'une forme en éléments primitifs [14]

Les relations entre les composantes peuvent être explicitées par différents formalismes de représentation. Le plus simple est la chaîne de composantes, particulièrement adaptée au cas des signaux monodimensionnels et des contours. Pour des relations spaciales dans l'espace 2D ou 3D, des représentations sous forme d'arbres ou plus généralement de graphes sont définis[8].

1.3. Application de la RDF

Les applications de la RDF sont nombreuses, elles se résument généralement en deux grandes catégories :

1.3.1. Reconnaissance de signaux

Le traitement du signal fournit des paramètres très utilisables pour pousser plus loin l'analyse, et décider quel est le signal émis, en fonction d'un répertoire de signaux possibles ; reconnaissance de la parole, par exemple.

Un autre domaine est extrêmement utile est celui des signaux biomédicaux ; la reconnaissance des formes permettrait d'automatiser ou de simplifier des tâches à la fois très complexes et très répétitives, comme par exemple assurer une surveillance automatique sur des mesures prises en temps réel.

Si l'on quitte les signaux physiologiques, la RDF s'interesse aussi aux mesures de signaux d'origine artificielle : surveillance de machine, interprétation d'écho, etc. Un bon exemple est la détection d'objets sur signal radar [7].

1.3.2. Reconnaissance des images

L'autre domaine prépondérant est celui de l'analyse et l'interprétation des images. Depuis les dessins les plus simples, comme les chiffres dactylographiés, jusqu'aux images multispectrales complexes issues de satellites, le champ des applications est immense. Pour le premier exemple, sont déjà opérationnels des lecteurs pratiques aussi bien pour les caractères dactylographiés ou imprimés que pour les caractères manuscrits. Pour l'imagerie médicale, on trouve les problèmes de comptage de cellules ou de chromosomes, de sélection de radiographies, d'interprétation des résultats de tous les systèmes d'imagerie.

Un grand nombre d'images proviennent du domaine de la robotique, en particulier industrielles : reconnaissance de pièces pour saisie, par exemple. L'analyse des paysages est également très utile (photos aériennes, guidage en temps réel d'engins, etc.). La reconnaissance d'objets, ou la detection d'eventuelles ressources sur les photograpies prises par des satellites font également l'objet d'un très grand nombre d'études [7].

1.4. Généralités sur les Réseaux de Neurones Artificiels (RNA)

Les RNA sont des réseaux fortement connectés de processeurs élémentaires fonctionnant en parallèle. Chaque processeur élémentaire calcule une sortie unique sur la base des informations qu'il reçoit. Un processeur élémentaire constitue un modèle neuronal artificiel représenté dans la figure 1.7.

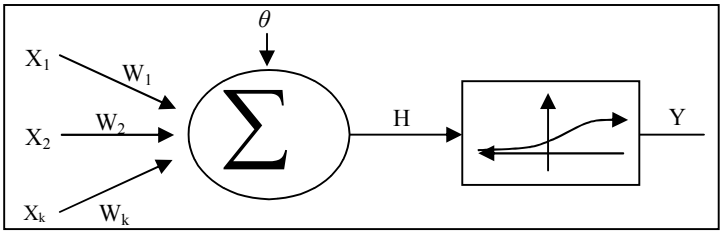


Figure 1.7 : Neurone formel [54]

Pour un nombre compris entre $j = 1$ et un nombre quelconque n, le neurone formel calcule la somme de ses entrées $(X_1, ..., X_n)$, pondérées par les poids synaptiques $(W_1, ..., W_n)$, et la comparer à son seuil θ. Si le résultat est supérieur au seuil, alors la valeur renvoyée est 1, sinon la valeur renvoyée est 0, c'est le cas pour une fonction d'activation seuil [54].

$$y = f(\sum_{j=1}^{n} \omega_j \ x_j - \theta) \qquad (4.7)$$

Il existe de nombreuses formes possibles pour la fonction de transfert. Les plus courantes sont présentées sur la figure 1.8.

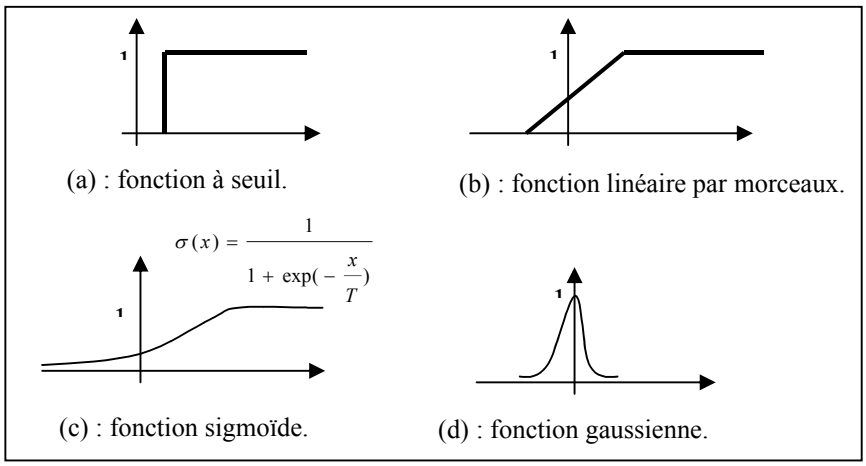


Figure 1.8 : Quelques fonctions d'activation [54]

1.4.1. Principe de RN multicouches ou Multi Layer Perceptron (MLP)

Le MLP (Multi Layer Perceptron) est un RN multicouches, chaque couche contient plusieurs neurones non connectés entre eux, néanmoins, la connexion est établie entre couches. Un MLP peut contenir une ou plusieurs couches cachées (fig.1.9).

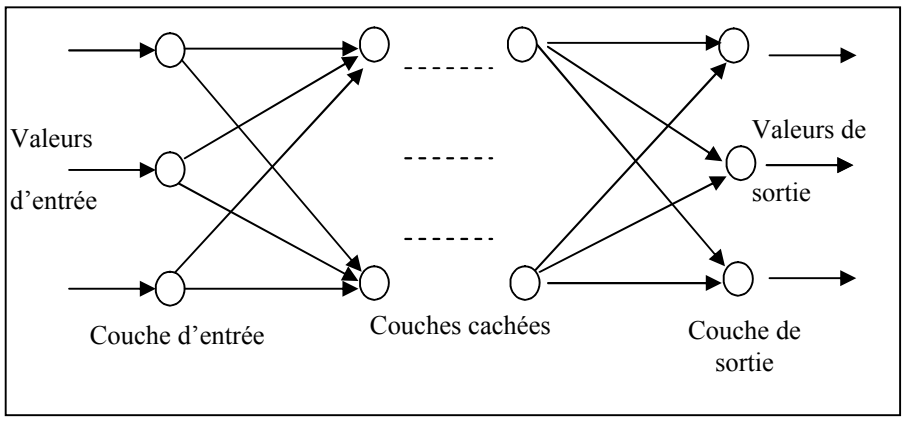


Figure 1.9 : Réseaux multicouches [54]

La figure 1.10 montre la possibilité de configurer le réseau MLP pour faire la reconnaissance. En effet, les vecteurs acoustiques issus d'une analyse spectrale de la parole

sont présents à la couche d'entrée du RN. Quant à la couche de sortie, elle est configurée de telle sorte que ses neurones codent les classes (phonèmes, mots,…etc.).

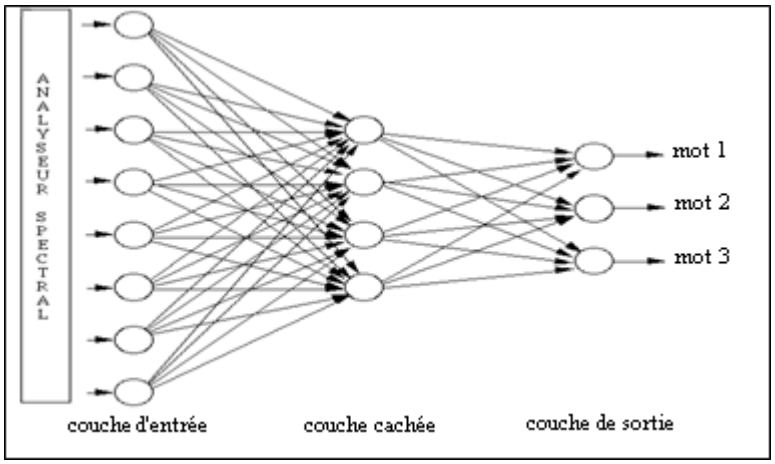


Figure 1.10 : Structure d'un MLP à trois couches

L'apprentissage dans ce type de réseau nécessite l'introduction des paramètres de toutes les classes à la fois ; ce qui est coûteux de point de vue calcul. Le problème devient de plus en plus complexe s'il s'agit des données importantes [15].

Le MLP peut aussi prendre une architecture dite Autoassociateur. Dans la phase d'apprentissage, le réseau est forcé pour que les sorties suivent les entrées. Dans la plupart des cas, le nombre de neurones de la couche cachée est inférieur au nombre de neurones de la couche d'entrée (de sortie) (fig.1.11) [16].

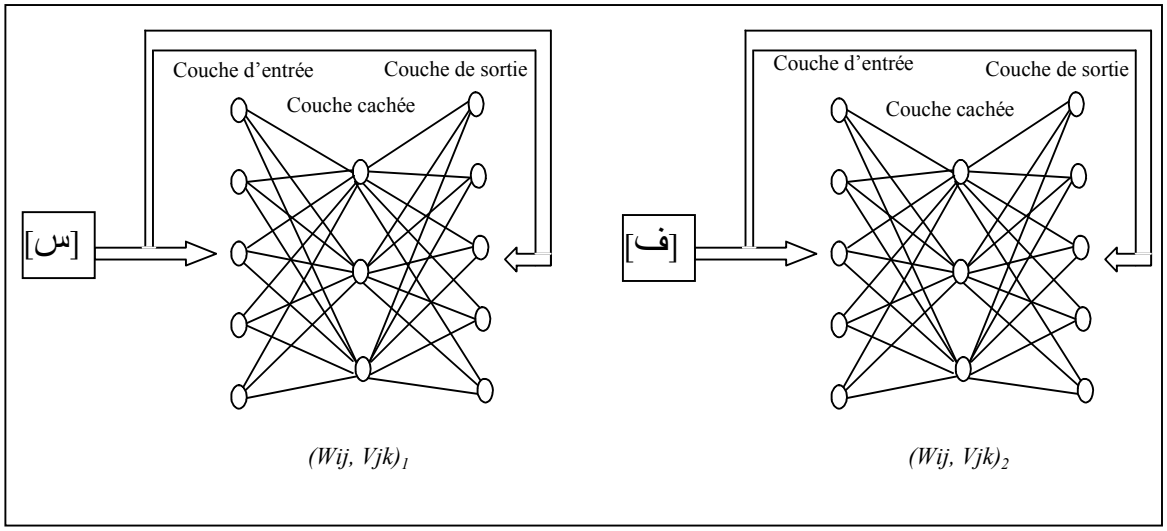


Figure 1.11 : Architecture de l'autoassociateur [54]

1.4.2. Principe de RN à délai ou Time Delay Neural Network (TDNN)

Le TDNN est un type de réseau de neurone semblable au MLP. Il est connu par sa puissance à exploiter l'aspect spatiotemporel du vecteur attribut, ce qui est convenable à la

parole. L'architecture retenue comporte deux parties principales. La première, correspondant aux couches basses, elles sont structurées comme un registre à décalage ; la liaison synaptique partagées entres elle, permet de transformer progressivement les caractéristiques en grandeurs de plus en plus signifiantes.

La seconde correspond à un MLP classique, elle reçoit en entrée l'ensemble des sorties de la partie TDNN, c'est donc l'intégration de l'ensemble des caractéristiques issues des couches précédentes. Ces deux blocs sont complètement paramétrables suivant le nombre des vecteurs attributs et bien sûr le nombre des classes (fig.1.12) [17].

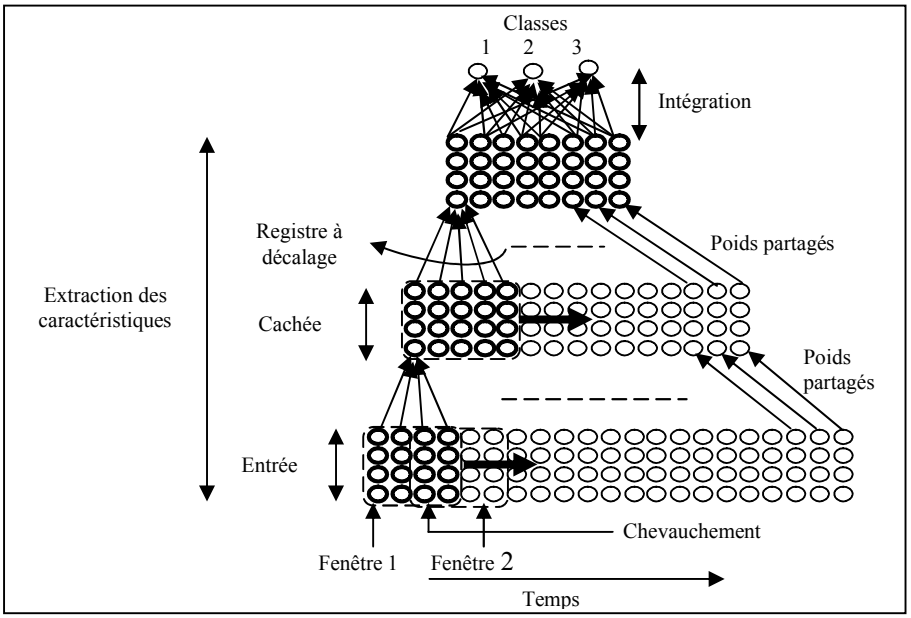


Figure 1.12 : Réseau de neurones à délai TDNN

Néanmoins, le RN utilisé dans notre application est un mélange entre les deux types précédemment cités. Il consiste à faire l'apprentissage en premier lieu avec uniquement le TDNN. Un deuxième apprentissage a été fait pour intégrer les résultats du premier. C'est donc un RN à double apprentissage et cela pour faciliter le calcul des poids synaptiques. Donc nous avons des poids synaptiques TDNN et des poids synaptiques MLP.

1.5. Conclusion

Dans ce chapitre nous avons présenté brièvement quelques notions sur la RDF ainsi que les principales méthodes et techniques disponibles pour mettre en œuvre un système à base de reconnaissance de formes. Nous avons présenté également une description sur les réseaux de neurones puisque nous les avons utilisés pour notre application. Enfin nous avons cité quelques applications réalisées ou en voie de réalisation où on fait appel à la RDF tels que la reconnaissance des signaux ou des images.

CHAPITRE 2 :
GÉNÉRALITÉS SUR LE TRAITEMENT AUTOMATIQUE DE LA PAROLE ET DE L'ARABE STANDARD

2.1. Introduction

La parole est sans doute le moyen de communication le plus simple et le plus efficace chez les humains. Depuis le début de la recherche dans le domaine du traitement de signal, les chercheurs ont toujours eu une attention particulière pour le signal de la parole. Le TAP recouvre les activités liées à l'analyse de la parole, à son codage, à sa synthèse, à sa reconnaissance et à sa compréhension, en vue de l'appliquer dans le cadre d'un dialogue HM, par exemple. Il comprend également la reconnaissance du locuteur et la reconnaissance de la langue parlée.

Dans ce chapitre nous présentons une brève description sur le traitement automatique de la parole et sur les particularités de l'Arabe standard.

2.2. Appareil phonatoire et production de la parole

La parole est bien considérée comme une activité propre à l'homme. Elle met en jeu des organes de phonation et est une véritable gymnastique des muscles de larynx, du pharynx, de la langue et des parois de la cavité buccale d'une façon générale. Elle nécessite donc une coordination musculaire qui passe par un apprentissage, même si, dans le cas général, cet apprentissage se fait naturellement dans le milieu familiale depuis les tous premiers temps après la naissance.

L'organe essentiel de la phonation est le larynx. Les cordes vocales vibrent au passage de l'air provenant des poumons et produisent des sons audibles. Les crico-aryténoïdiens, muscles intrinsèques, déterminent la hauteur des sons en augmentant la tension longitudinale des cordes vocales. Les muscles extrinsèques, reliant le larynx aux structures anatomiques voisines, modifient aussi la forme du larynx et modulent la voix. On parle aussi d'appareil vocal subglottique pour désigner les poumons et les muscles de la cage thoracique, de l'abdomen, du dos et de la poitrine. Le pharynx, parti de la gorge située entre la bouche et l'œsophage, ainsi que les cavités buccales et nasales agissent comme des résonateurs qui atténuent certaines fréquences (fig.2.1) [18].

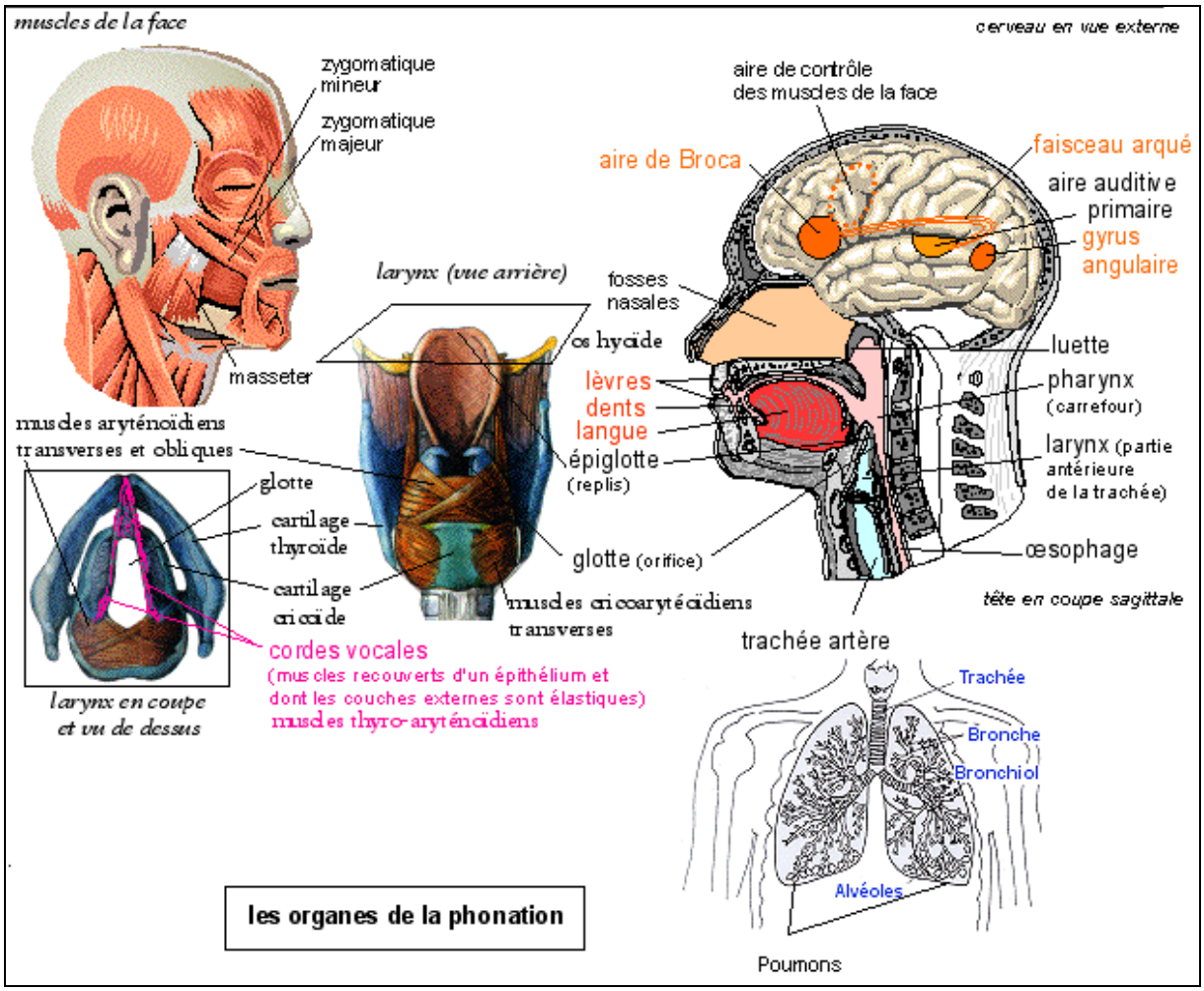


Figure 2.1 : Les organes de la phonation [18]

2.3. Système auditif et perception de la parole

Dans le cadre du TAP, une bonne connaissance des mécanismes de l'audition et des propriétés perceptuelles de l'oreille est aussi importante qu'une maîtrise des mécanismes de production. En effet, tout ce qui peut être mesuré acoustiquement ou observé par la phonétique articulatoire n'est pas nécessairement perçu.

Les processus complexes par lesquels un auditeur comprend un message oral émis par un locuteur peuvent être fonctionnellement décomposés en deux grandes phases :

Dans une première phase, l'oreille transforme l'information contenue dans le signal acoustique et la transmet ensuite au cerveau par l'intermédiaire du nerf auditif. La deuxième phase correspond à la reconnaissance du message linguistique sur la base de cette représentation du signal de parole. Cette deuxième phase peut elle-même être décomposée en deux niveaux, l'un correspondant à l'interprétation d'indices fournis à l'issue de prétraitement auditif sans référence à la signification, le deuxième réalise l'accès au sens [19].

L'appareil auditif comprend l'oreille externe, l'oreille moyenne et l'oreille interne (fig.2.2).

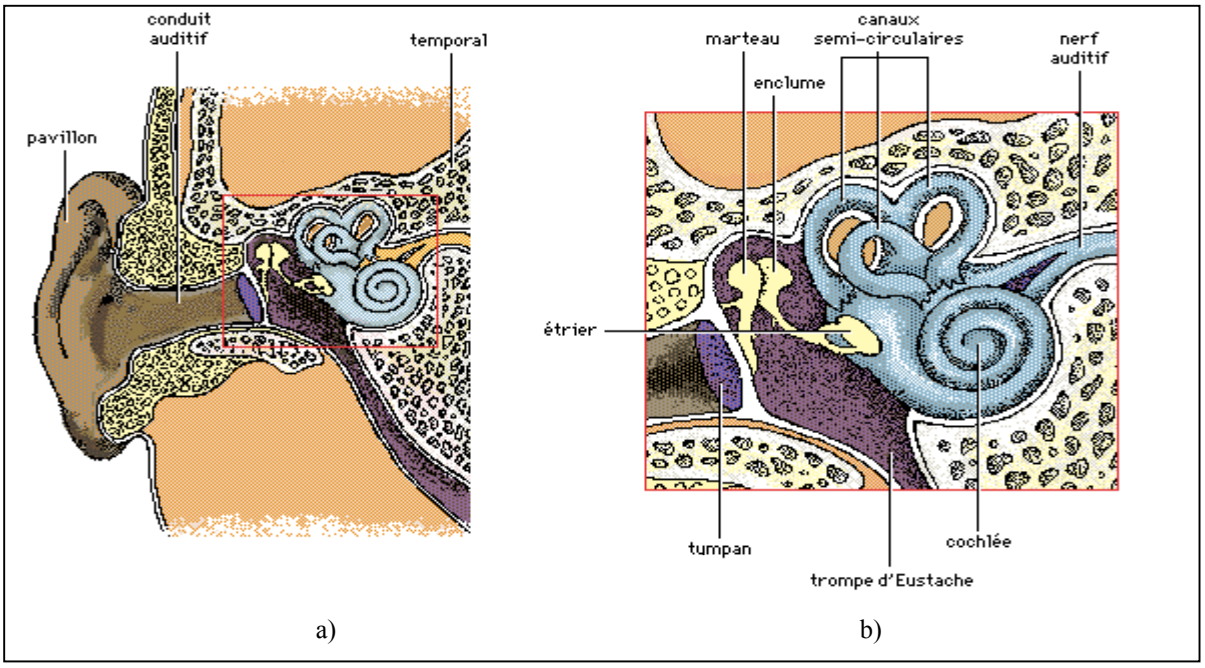


Figure 2.2 : a) anatomie de l'oreille ; b) oreille interne [7]

2.3.1. Système de transmission

L'oreille externe (pavillon et conduit auditif) permet de recueillir les sons et de les orienter vers l'oreille moyenne. L'oreille moyenne (tympan et osselets) assure la fonction de transmission proprement dite, qui inclut une transformation d'ondes sonores aériennes en ondes liquidiennes, mais sans d'importantes pertes d'intensité que l'on observait si l'on passait directement de l'air au liquide. C'est au mécanisme composé de marteau, étrier et enclume qu'appartient l'opération adaptation d'impédance entre l'air et le milieu liquide de l'oreille interne. Les vibrations de l'étrier sont transmises au liquide de la cochlée. L'oreille moyenne joue aussi un rôle d'accommodation auditive. La position des osselets les uns par rapport aux autres assure l'amplification des sons. La perméabilité du tube auditif est, en outre, indispensable à une bonne réception des sons.

L'oreille interne, dans sa partie antérieure (la cochlée) et le nerf auditif assurent la fonction de perception. Comme pour d'autres organes sensoriels, le processus commence par l'élaboration d'un message nerveux à partir d'un phénomène non nerveux, en l'occurrence la vibration d'un liquide, grâce à l'intervention de cellules spécialisées [7].

2.3.2. Aire d'audition

L'homme est en effet très limité dans ses capacités de perception auditive vis-à-vis d'autres membres du règne animal. Il lui est ainsi impossible de distinguer des sons de plus de 20 kHz, les ultrasons, alors que certains animaux qui lui sont familiers peuvent percevoir des sons allant jusqu'à 50 kHz. De même il lui est impossible de distinguer des sons d'une fréquence inférieure à 20-25 Hz, les infrasons. À l'intérieur de cet espace fréquentiel existe un sous-espace délimité par les niveaux d'énergie des sons. Il existe une limite d'énergie au-delà de laquelle l'homme ne percevra pas un son d'une fréquence appartenant pourtant au spectre de l'audition. Cette limite d'énergie est appelée seuil d'audition et il est variable en fonction de la fréquence. Inversement, il existe une limite d'énergie maximale. Cette limite ne doit pas être franchie car la cochlée, et plus particulièrement les cellules ciliées, peuvent être irrémédiablement endommagées. Cette limite s'appelle le seuil de douleur et il est aussi variable en fonction de la fréquence.

L'espace de fréquences et d'énergies ainsi défini (fige 2.3) constitue la zone d'audition à l'intérieur de laquelle l'homme peut recevoir des informations de son environnement. C'est bien sûr à l'intérieur de cet espace que se trouve le champ de la musique qui circonscrit lui-même le champ de la parole [2].

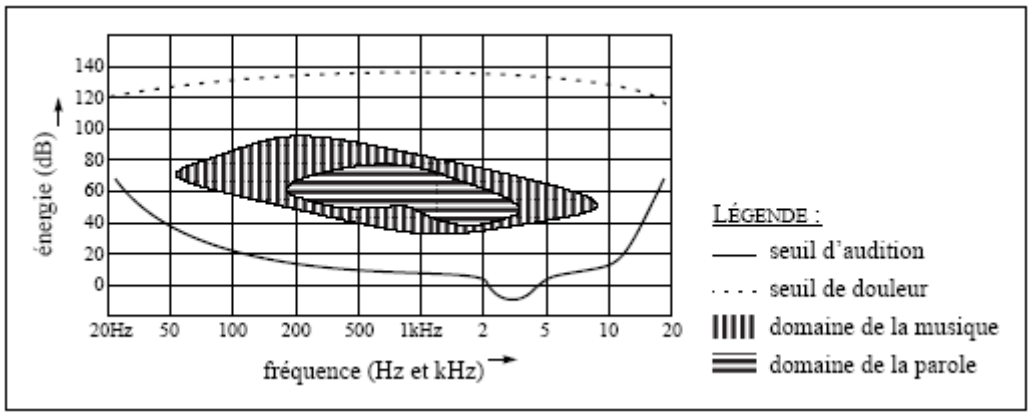


Figure 2.3 : L'aire d'audition [2]

2.4. Contenu du signal de la parole

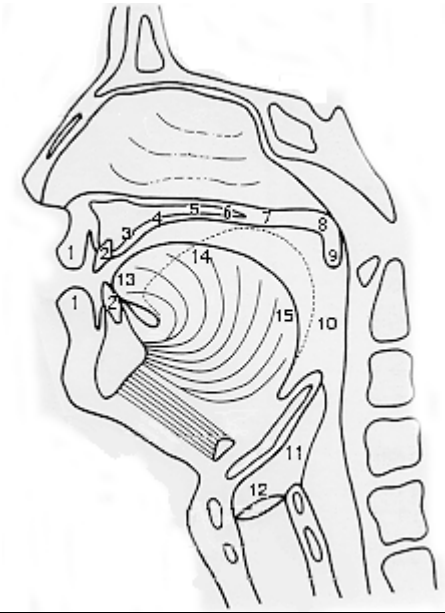
Le signal de la parole est un signal très complexe. Il contient une quantité importante d'informations imbriquées entre elles. La parole transmet l'information phonétique, une information sur le locuteur (homme, femme ou enfant), sur son état psychologique (joyeux, en colère, etc.), sur son état physique (en bonne santé, fatigué, malade, par exemple) [20].

L'information portée par le signal de parole peut être analysée sur plusieurs niveaux de description : phonétique, phonologique, prosodique, morphologique, syntaxique, sémantique et pragmatique [21], [22].

2.4.1. Sur le plan phonétique

La phonétique est une science qui concerne l'étude des caractéristiques physiques des sons de la parole en liaison avec la langue. On peut analyser les sons sur trois plans complémentaires :

- perceptif ou auditif : relève de la physiologie et de la neurologie et concerne le processus de réception des sons à travers le nerf auditif jusqu'au cerveau ;
- articulatoire : étudie l'appareil phonatoire, son fonctionnement et ses possibilités générales ; s'occupe de la production des sons et a pour but d'inventorier et de décrire les sons des langues du monde (fig.2.4);
- acoustique : elle mesure les phénomènes physiques (ondes sonores) qui sont les traits de la parole : le timbre, la durée, la fréquence et l'intensité de son [24].



	Organe anatomique		Nomenclature phonétique correspondante	
1	lèvres		labiales	
2	dents		dentales	
3	alvéoles		alvéolaires	
4	palais dur		pré-palatales	
5			médio-palatales	
6			post-palatales	
7	voile du palais		pré-vélaires	
8			post-vélaires	
9	luette (*uvula*)		uvulaires	
10	pharynx		pharyngales	
11	larynx		laryngales	
12	glotte		glottales	
13	apex	de la langue	apicales (pré-dorsales)	dorsales
14	dos		médio-dorsales	
15	racine		radicales (post-dorsales)	

Figure 2.4 : Principaux lieux d'articulation [23]

2.4.1.1. Transcription Orthographique Phonétique (TOP)

La TOP ou phonétisation est une étape essentielle pour la synthèse de la parole à partir du texte. Elle permet la prononciation correspondante au texte en entrée sous la forme d'une liste de phonèmes. Les phonéticiens symbolisent les sons du langage au moyen de signes divers auxquels on attribue un code conventionnel [22].

Diverses transcriptions phonétiques sont utilisées (Tab.2.1 et Tab2.2).

Tableau 2.1 : TOP du Français [2]

symbole phonétique	exemple en langue française	classe	phonétique
a	plat		voyelles
ɑ	mât		
i	pile		
y	rue		
ɔ	bol		
o	pôt		
ə	le		
ɛ	lait		
e	blé		
ø	peu		
œ	heure		
u	roue		
ɑ̃	blanc		voyelles nasales
ɔ̃	bon		
ɛ̃	lin		
œ̃	brun		
j	hier		semi-consonnes
ɥ	huit		
w	oui		
l	lent		liquides
R	rue		
m	masse		nasales
n	nous		
ɲ	signal		
f	fer	sourdes	fricatives
s	assis		
ʃ	chou		
v	verre	sonores	
z	Asie		
ʒ	joue		
p	passe	sourdes	occlusives
t	toux		
k	cou		
b	basse	sonores	
d	doux		
g	goût		

Tableau 2.2 : TOP de l'Anglais [2]

phonème API	phonème ARPABET	exemple en langue anglaise	classe phonétique
i	IY	beat	voyelles
ɪ	IH	bit	
ɛ	EH	bet	
æ	AE	bat	
ɑ	AA	bob	
ɔ	AO	bought	
ʊ	UH	book	
u	UW	boot	
ʌ	AH	but	
ɚ	ER	bird	
ə	UR	neighbour	
əʊ	AX	about	
χ	IX	roses	
α^y	AY	my	diphtongues
$ɔ^y$	OY	boy	
e^y	EY	bait	
o^w	OW	boat	
α^w	AW	down	
j	Y	you	semi-voyelles
w	W	wit	
l	L	let	liquides
r	R	rent	
m	M	met	nasales
n	N	net	
ŋ	NX	bang	
h	HH	hat	fricatives
f	F	fat	
θ	TH	thin	
s	S	sat	
š	SH	shut	
v	V	vat	
ð	DH	that	
z	Z	zoo	
ž	ZH	azure	
č	CH	church	affriquées
ǰ	JH	judge	
p	P	pet	occlusives
t	T	ten	
k	K	kit	
b	B	bet	
d	D	den	
g	G	get	

2.4.1.2. Classes phonétiques

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

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

Les différentes classes phonétiques présentes en français et en anglais sont :

- les voyelles, cette classe correspond, à quelques nuances supplémentaires près, aux voyelles de l'écrit. Elles se caractérisent principalement par le voisement qui crée des formants. Ces formants, qui sont des zones fréquentielles de forte énergie, correspondent à une résonance dans le conduit vocal de la fréquence fondamentale produite par les cordes vocales. Ces formants peuvent s'élever jusqu'à des fréquences de 5 kHz mais ce sont principalement les formants en basses fréquences qui caractérisent les voyelles. Cette caractéristique permet d'ailleurs de distinguer grossièrement les voyelles en fonction de leur F_1 et F_2 [2] ;
- les consonnes, cette classe est en fait constituée, pour simplification, du regroupement des trois sous-classes que sont les semi-consonnes, les liquides et les nasales.
- les semi-consonnes (ou semi-voyelles), elles ont la structure acoustique des voyelles mais ne peuvent en jouer le rôle car elles ne sont que des transitions vers d'autres voyelles qui sont les véritables noyaux syllabiques. D'un point de vue syntaxique, une règle stricte de la langue française veut que deux voyelles ne puissent jamais se suivre. Cette règle est très largement respectée dans la construction des mots mais présente, comme toute règle, quelques exceptions. La classe des semi-consonnes a été créée pour pallier ces exceptions de manière gracieuse. Les semi-consonnes sont sonores.
- les liquides, les liquides sont très similaires aux voyelles et aux semi-consonnes mais leur durée et leur énergie sont généralement plus faibles, elles sont sonores.
- les nasales, les phonèmes sont formés par passage de l'air dans le conduit vocal depuis les cordes vocales. Ce passage exclut normalement toute connexion du conduit normal, le conduit buccal, avec le conduit nasal. Ce dernier peut cependant être employé, dans un nombre limité de cas puisque sa physionomie ne permet pas de créer des sons autrement qu'en modifiant le volume de la caisse de résonance qu'il constitue par l'intermédiaire de la langue, faisant occlusion dans le conduit buccal. Les nasales sont donc produites de la même manière que les occlusives nasales mais l'air n'est pas, cette fois, comprimé dans le conduit vocal. Le vélum est en effet abaissé pour permettre à l'air d'être expiré. Les nasales sont voisées. Il est à noter que certaines voyelles possèdent également un caractère de nasalité.
- les occlusives, les phonèmes de cette classe se caractérisent oralement par la fermeture du conduit vocal, fermeture précédant un brusque relâchement. Les occlusives sont

donc constituées de deux parties successives : une première partie de silence, correspondant à l'occlusion effective, et une deuxième partie d'explosion, au moment du relâchement. Les occlusives peuvent être voisées, à la manière des voyelles, ou sourdes, c'est à dire non voisées. Les occlusives voisées peuvent également être appelées occlusives sonores.

- les fricatives, dans cette classe sont regroupés les sons produits par la friction de l'air dans le conduit vocal lorsque celui-ci est rétréci au niveau des lèvres, des dents ou de la langue. Cette friction produit un bruit de hautes fréquences et peut être voisée ou sourde.
- les diphtongues, cette classe phonétique est propre à l'anglo-américain. Les phonèmes qui composent cette classe se caractérisent par deux états stables formantiques et par la transition entre ces deux états.
- les affriquées, cette classe est, elle aussi, propre à l'anglo-américain mais les affriquées peuvent également être observées dans le français québécois. Les affriquées sont composées d'un occlusive immédiatement suivie par une fricative de durée cependant plus faible que celle des véritables fricatives [2].

2.4.2. Sur le plan phonologique

La phonologie a comme objectif d'étudier les variantes phonétiques contextuelles. En reconnaissance de la parole, la phonologie regroupe l'ensemble des modules de traitement des altérations possibles d'un phonème ou d'un mot dans un contexte donné.

La phonologie regroupe trois types d'altérations :

- phonologiques dans le mot, (variantes de prononciation) ;
- phonologiques dues aux flexions en fin de mot (conjugaisons des verbes, pluriels des noms et des adjectifs) ;
- qui apparaissent à la jonction de deux mots (liaison).

A ces trois types d'altérations on peut ajouter les altérations qui se produisent artificiellement à cause des erreurs du Décodage Acoustique Phonétique (DAP):

- les délétions de phonèmes dans les cas de sous-segmentation du signal de la parole ;
- les insertions de phonèmes dans les cas de sur-segmentation du signal de la parole ;
- les substitutions de phonèmes dans les cas de mauvaises identifications.

Un module phonologique doit tenter dans la mesure du possible de compenser ces erreurs [21].

2.4.3. Sur le plan prosodique

La prosodie peut être considérée comme une sorte de “ponctuation acoustique” de la parole. Elle recouvre les aspects liés à la hauteur de la voix (liée essentiellement à la fréquence fondamentale F_0), à l'intensité (concerne l’amplitude et l’énergie du son) et à la durée des segments syllabiques (correspond à la durée acoustique ou temps d’émission) [22].

2.4.4. Sur le plan morphologique

La morphologie est la branche de la linguistique qui étudie comment les formes lexicales sont obtenues à partir d’un ensemble réduit d’unités porteuses de sens, appelées morphèmes. On distingue les morphèmes lexicaux des morphèmes grammaticaux, qui apportent aux premiers des nuances de genre, nombre, mode, temps, personne, etc. Tout comme le phonème, le morphème est une unité abstraite. Elle peut être réalisée en pratique sous diverses formes appelées allomorphes, fonction de leur contexte morphémique. Ainsi le morphème grammatical du pluriel se manifeste sous la forme d’un ‘*s’* dans ‘*pommes’*, d’un ‘*x’* dans ‘*jeux’* et d’un ‘*nt’* dans ‘*jouent*’ [21].

2.4.5. Sur le plan syntaxique

Du point de vue de la langue, la syntaxe est l’ensemble des règles écrites ou orales contraignant l’ordre des mots dans la phrase qui sont souvent appelés règles grammaticales. Les grammaires ne servent qu’à dresser la frontière entre les phrases régulièrement constituées ou pas. Elles permettent également de décrire l’organisation hiérarchique des phrases, leur structure syntaxique.

Dans un système de compréhension, le but de la syntaxe est de réduire le nombre de phrases autorisées à partir du vocabulaire choisi. Par exemple, on peut construire 250^8 phrases de 8 mots à partir d’un vocabulaire de 250 mots, mais seules 250^4, par exemple, d’entre elles ont un sens. On a donc divisé l’espace de recherche par la moitie [22].

2.4.6. Sur le plan sémantique

La sémantique est définie d'un point de vue linguistique, comme la relation entre la forme des signes linguistiques, ou "signifiants", et ce qui est signifié, ou "signifiés". On peut distinguer plusieurs types de sémantique : descriptive, générative, interprétative, différentielle, etc.

2.4.7. Sur le plan pragmatique

Au contraire du sens sémantique, que l’on qualifie souvent d’indépendant du contexte, le sens pragmatique est défini comme dépendant du contexte. Tout ce qui se réfère au contexte, souvent implicite, dans lequel une phrase s’inscrit et à la relation entre le locuteur et de son auditoire, a quelque chose à voir avec la pragmatique. Son étendue couvre l’étude de

sujets tels que les présuppositions, les implications de dialogue, les actes de parole indirects, etc. Elle est malheureusement bien moins développée encore que la sémantique [21].

2.5. Propriétés du signal vocal

Un signal vocal est essentiellement caractérisé par ses propriétés de continuité et ses propriétés de variabilité.

Contrairement au langage écrit où les mots sont séparés par des blancs dans les textes manuscrits ou par des espaces dans les textes dactylographiés, les séparateurs, symbolisés par les silences entre les mots, sont parfois très difficiles à repérer, ce qui fait la continuité du signal de la parole.

Le terme de variabilité, qui est assez générique, peut englober plusieurs problèmes qui sont cependant totalement indépendants du point de vue des techniques actuellement utilisées pour les résoudre. Il est ainsi possible d'isoler une variabilité du signal de parole relativement aux classes phonétiques définies. Il est aussi possible d'isoler la variabilité de l'environnement sonore d'un système de reconnaissance. À un niveau beaucoup plus abstrait, celui de la sémantique, il est également possible de parler de variabilité, certaines phrases ne pouvant pas être comprises lorsqu'elles sont considérées hors contexte, imposant ainsi de définir des mécanismes de gestion de l'historique du dialogue.

On peut faire une distinction entre variabilité :

- intralocuteur qui identifie les différences dans le signal produit par une même personne. Cette variation peut résulter de l'état physique ou moral du locuteur. Une maladie des voies respiratoires peut ainsi dégrader la qualité du signal de parole de manière à ce que celui-ci devienne totalement incompréhensible, même pour un être humain. L'humeur ou l'émotion du locuteur peut également influencer son rythme d'élocution, son intonation ou sa phraséologie.
- Interlocuteurs où la cause principale est de nature physiologique. La parole est principalement produite grâce aux cordes vocales qui génèrent un son à une fréquence fondamentale. Cette fréquence sera différente d'un individu à l'autre et plus généralement d'un sexe à l'autre, une voix d'homme étant plus grave qu'une voix de femme, la fréquence du fondamental étant plus faible. La variabilité interlocuteur trouve également son origine dans les différences de prononciation qui existent au sein d'une même langue et qui constituent les accents régionaux. Ces différences s'observeront d'autant plus facilement qu'une communauté de langue occupera un espace géographique très vaste, sans même tenir compte de l'éventuel rayonnement international de cette communauté et donc de la probabilité qu'a la langue d'être

utilisée comme seconde ou, pire, troisième langue par un individu de langue maternelle étrangère.

- liée à l'environnement, elle peut, parfois, être considérée comme une variabilité intralocuteur mais les distorsions provoquées dans le signal de parole sont communes à toute personne soumise à des conditions particulières. La variabilité due à l'environnement peut également provoquer une dégradation du signal de parole sans que le locuteur ait modifié son mode d'élocution. Le bruit ambiant peut provoquer une déformation du signal de parole en obligeant le locuteur à accentuer son effort vocal. Enfin, le stress et l'angoisse que certaines personnes finissent par éprouver lors de longs voyages peuvent également être mis au rang des contraintes environnementales susceptibles de modifier le mode d'élocution [2].

2.6. Outils pour le traitement du signal vocal

En acoustique, un son se définit classiquement au moyen de son amplitude, de sa durée, et de son timbre. La parole qui est un son particulièrement complexe, n'échappe pas de cette définition. Le traitement du signal a pour but précisément de quantifier ces trois grandeurs pour faire correspondre à l'onde sonore (temporelle) une description multidimensionnelle. Les notions d'amplitude et de timbre n'ont véritablement de sens que si le son considéré est stationnaire ou pour le moins stable. La parole, au contraire, est généralement non stationnaire, continu, d'énergie finie. Sa structure est complexe et variable dans le temps : tantôt périodique (pseudopériodique) pour les sons voisés, tantôt apériodique pour les sons non voisés.

Joseph Fourier a montré que toute onde physique peut être représentée par une somme de fonctions trigonométriques appelée série de Fourier. Elle comporte un terme constant et des fonctions sinusoïdales d'amplitudes diverses. Ainsi un son sinusoïdal ne comporte qu'une seule raie spectrale correspondant à la fréquence de sa fonction sinus. Un son complexe est composé d'une multitude de ces raies spectrales qui représentent sa composition fréquentielle [25].

Dans le cas d'une séquence d'échantillons, il est alors possible de calculer une Transformée de Fourier Discrète (TFD, Discret Fourier Transform - DFT - en Anglais).

La formule *(2.1)* donne le calcul de la TFD pour une séquence x*(n)* comportant *N* échantillons :

$$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{-jn2\pi\frac{k}{N}} \qquad (2.1)$$

Le spectrogramme est un outil de visualisation utilisant la technique de la transformée de Fourier et donc du calcul de spectres. Il a commencé à être largement utilisé en 1947, à l'apparition du sonagraphe, et est devenu l'outil incontournable des études en phonétique pendant de nombreuses années. L'apparition de l'informatique puis d'écrans graphiques de bonne qualité a permis d'abandonner tout matériel comme le sonagraphe mais la technique du spectrogramme est encore aujourd'hui largement utilisée du fait de sa simplicité de mise en œuvre et du grand nombre d'études qui ont déjà été réalisées (fig.2.5).

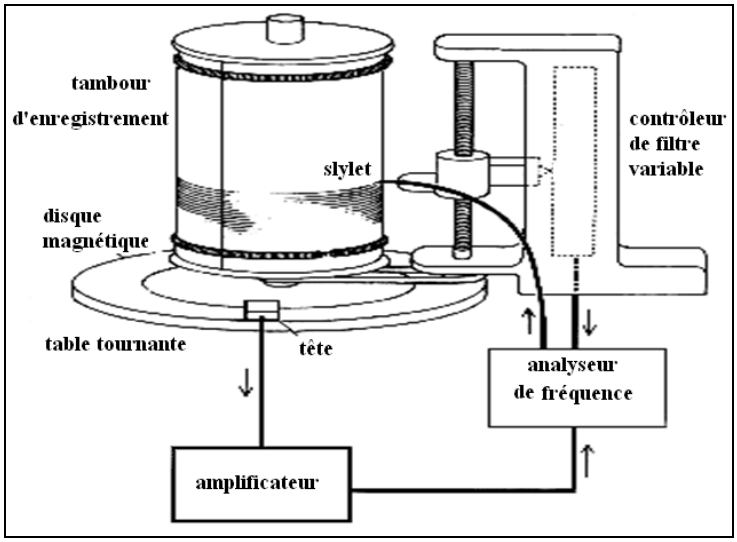


Figure 2.5 : Spectrographe analogique

Le spectrogramme permet de mettre en évidence les différentes composantes fréquentielles du signal à un instant donné, une TFD étant régulièrement calculée à des intervalles de temps rapprochés. Avant le calcul des transformées successif, le signal doit d'abord être préaccentué par un filtre du premier ordre pour égaliser les hautes fréquences dont l'énergie est toujours plus faible que celle des basses fréquences.

La formule de préaccentuation d'un signal est de la forme :

$$H(z)=1-a.z^{-1} \qquad (2.2)$$

Où a le facteur de préaccentuation, pris communément à 0.97.

Ensuite le signal va subir une opération de fenêtrage, où il est considéré comme indéfiniment stable et constitué d'une somme invariable de fonctions sinusoïdales de fréquences différentes. Pour contourner cette contrainte théorique d'invariabilité du signal, il faut convoluer le signal avec une fenêtre temporelle qualifiée de glissante puisque chaque calcul de spectre nécessite de convoluer le signal avec la fenêtre temporelle à un instant particulier. Différentes fenêtres temporelles existent mais chacune introduit une erreur résiduelle plus ou moins importante dans le spectre obtenu du fait de la forme choisie qui peut être, dans le pire des cas, triangulaire ou carrée (fig.2.6).

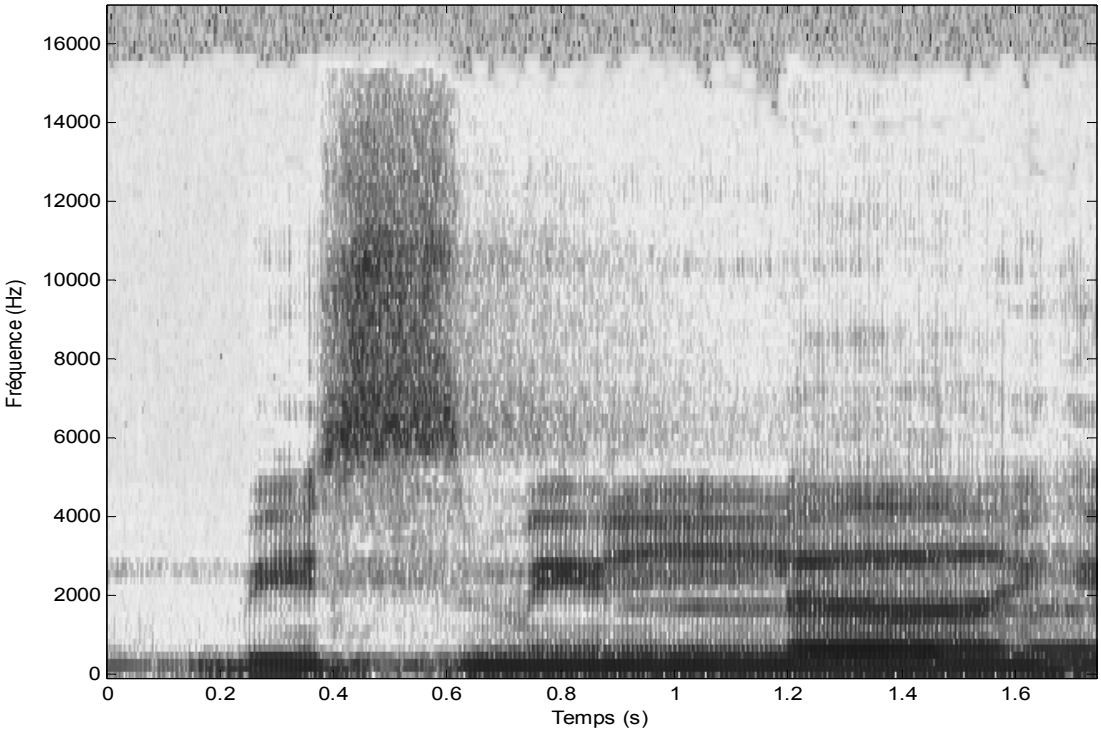


Figure 2.6 : Spectrogramme ou sonagramme de la phrase ‘‘بسم الله’’

Le choix de la taille de la fenêtre, en nombre de points de convolution, est également important vis-à-vis de la qualité de l'analyse fréquentielle obtenue. Ainsi, une fenêtre de petite taille (avec un nombre de 128 points, par exemple) permettra d'obtenir une bonne analyse dans le domaine temporel, du fait de son étroitesse, mais ne permettra pas d'obtenir une bonne information fréquentielle, la taille de la fenêtre étant alors trop petite pour ne pas tronquer les phénomènes de basses fréquences. À l'inverse, une fenêtre de grande taille (plus de 512 points) permettra d'obtenir une bonne information fréquentielle mais ne permettra pas d'obtenir une bonne information temporelle car tout événement, même de courte durée, est jugé présent sur l'ensemble du pas de temps analysé puisque la théorie de la transformée de Fourier considère les signaux indéfiniment stables.

Une fois la convolution effectuée, la transformée de Fourier est calculée sur la totalité de la fenêtre, le reste du "signal" étant alors égal à zéro. Ce processus permet d'obtenir un spectre qui correspond à une trame, un ensemble de trames calculées à intervalles réguliers permettant d'obtenir le spectrogramme désiré [2].

Le plus souvent en pratique, on utilise des fenêtres de durée de 25 à 30 ms, tandis que le pas de glissement et de l'ordre de 10 ms. Ces valeurs ont été choisies empiriquement ; elles sont liées au caractère quasi-stationnaire du signal de la parole (fig.2.7).

Le fenêtrage se fait généralement par la fenêtre de Hamming :

$$w(n) = 0{,}54 + 0{,}46.\cos(2\pi\frac{n}{N}) \qquad (2.3)$$

Pour *n=0, 1,2,…, N-1.*

N : le nombre de points dans chaque fenêtre.

L : le nombre de points de glissement.

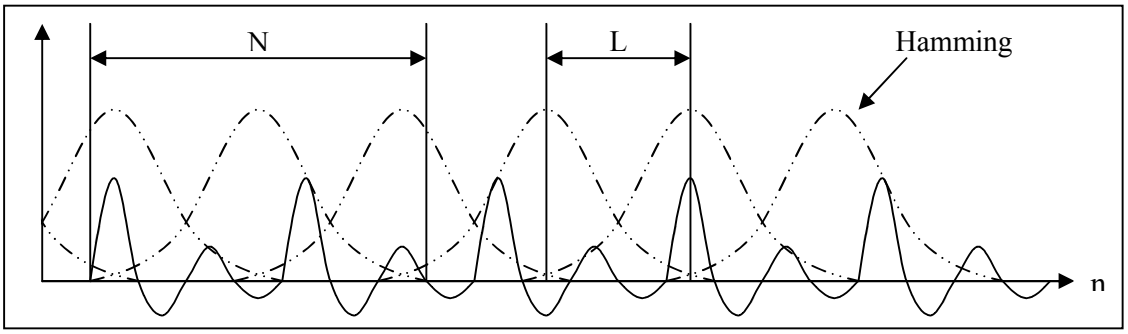


Figure 2.7 : Découpage d'une séquence en tranches avec recouvrement

Soit *x(n)* le signal de parole discrétisé, alors sa Transformée de Fourier (TF) sur une fenêtre de *N* échantillons est donnée par :

$$X(k) = \sum_{m=0}^{N-1}\left(\sum_{n=0}^{N-1} x(n)\, w(m-n)\right) e^{-j2\pi\frac{mk}{N}} \qquad (2.4)$$

Pour séparer les deux informations présentes dans le signal de parole qui sont la F_0 et la transformation, supposée linéaire, effectuée par le conduit vocal, il est nécessaire d'effectuer une déconvolution a posteriori du signal. Ceci va nous permet de connaître la contribution des cordes vocales et du conduit vocal lors de la génération du signal qui a, par la suite, été observé en entrée du système. Cette déconvolution peut être effectuée grâce au cepstre.

Le cepstre est une méthode qui se fonde sur la TF mais qui, grâce à une méthode efficace, permet d'isoler la F_0 de la transformation qui a été opérée par le conduit vocal. Comme pour le calcul du spectrogramme, le signal est préaccentué puis convolué avec une fenêtre glissante. Une première TF est alors calculée pour obtenir un spectre du signal. Ces coefficients sont ensuite transformés par le logarithme module. La convolution étant un opérateur multiplicatif, ce passage par les logarithmes permet de passer les coefficients dans un espace additif. Une Transformée de Fourier Inverse (TFI) permet alors d'obtenir un cepstre dont un coefficient représente le fondamental, les autres coefficients permettant d'obtenir le spectre de la convolution effectuée sur le fondamental. Cette méthode de calcul des cepstres est élémentaire (fig.2.8) [2].

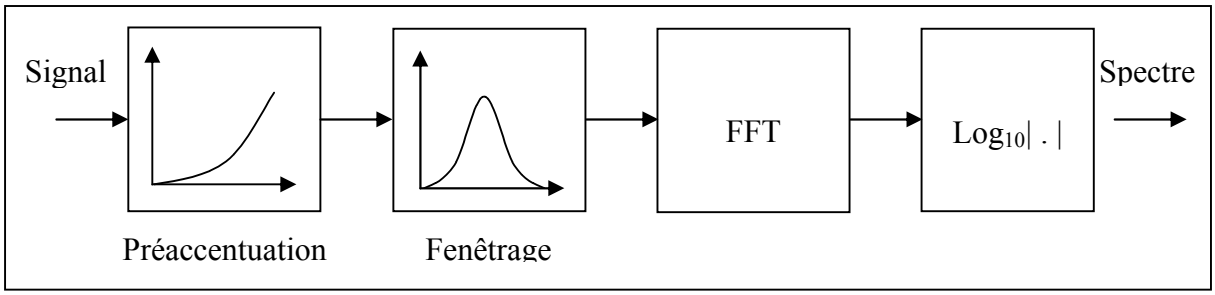


Figure 2.8 : Processus de calcul de spectre d'un signal de la parole

Les travaux de Stevens ont permis la mise en évidence de la *loi de puissance* ou *loi de Stevens* selon laquelle l'intensité de la perception d'un stimulus n'augmente pas linéairement en fonction de sa puissance mais de façon exponentielle en tenant aussi compte des modalités de l'expérimentation. Les coefficients MFCCs pour (*Mel-scaled Frequency Cepstral Coefficients)* dans la littérature, sont donc basés sur une échelle de perception appelée Mel, non linéaire. Celle-ci peut être définie par la relation suivante entre la fréquence en Hertz et sa correspondance en mels :

$$M_{mels} = x.\log(1 + \frac{f_{Hz}}{y}) \qquad (2.5)$$

Plusieurs valeurs sont utilisées pour *x* et *y*. On trouve dans *x=1000/log (2)* et *y=1000* [14].
Cependant les valeurs les plus couramment utilisées sont *x=2595* et *y=700* [25].
Le but final de l'extraction des paramètres est de modéliser la parole, un phénomène très variable. Par exemple, même si elle a de l'importance, la simple valeur de l'énergie n'est pas suffisante pour donner toute l'information portée par ce paramètre. Il est souvent nécessaire de recourir à des informations sur l'évolution dans le temps de ces paramètres. Pour cela, les dérivées première et seconde sont calculées pour représenter la variation ainsi que l'accélération de chacun des paramètres. Même si la robustesse de la représentation obtenue est accrue, cela implique aussi de multiplier par 3 l'espace de représentation (fig.2.9) [26].

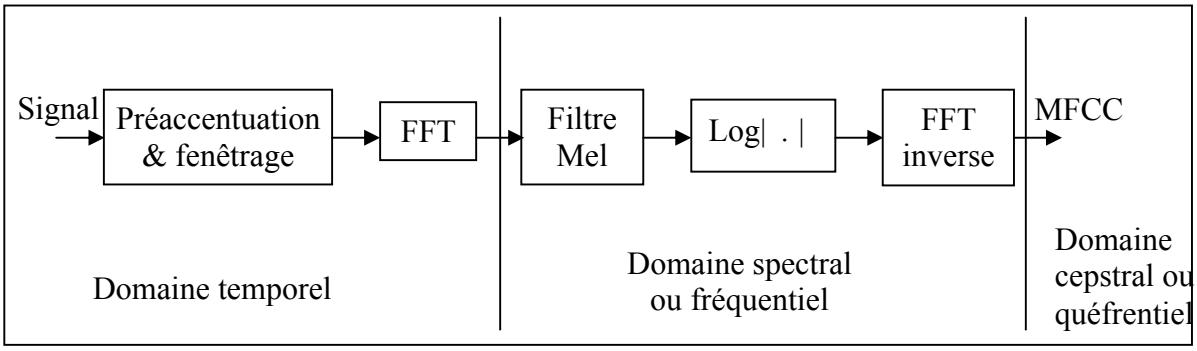


Figure 2.9 : Algorithme de calcul des MFCCs

2.8. Notions fondamentales sur les sons de l'Arabe Standard (AS)

L'Arabe appartient à la famille des langues sémitiques comme l'Akkadien, l'Hébreu, l'Araméen et le sud arabique. Au sein de cet ensemble, il appartient à un sous groupe particulier : le sémitique méridional. L'Arabe va dans une expansion extrêmement rapide reliée à un immense empire recouvrant le Proche-Orient, l'ensemble de la bordure méditerranéenne de l'Afrique, l'Espagne, la Sicile, plus de 22 pays, environ 250 millions d'arabophones dans le monde entier dont 195 millions l'Arabe est une première langue et 55 millions comme deuxième langue [27], [28].

L'expansion et le développement de cette langue sont intimement liés à la naissance et à la diffusion de l'Islam. L'Arabe s'est imposé dans toute l'aire arabo-musulmane comme langue religieuse mais plus encore comme langue d'administration d'empires successifs, langue de la culture, de la pensée, des sciences et des techniques, coexistant avec les langues locales. Ce développement s'est accompagné d'une rapide et profonde évolution (en particulier dans la syntaxe et l'enrichissement lexical). Une abondante poésie antéislamique est en outre parvenue jusqu'à nous, recueillie par les philologues arabes du Moyen-âge qui nous ont fourni ainsi des renseignements sur la situation de la langue à l'apparition de l'islam [27].

Par ses propriétés morphologiques et syntaxiques l'Arabe est considérée comme une langue difficile à maîtriser dans le domaine du Traitement Automatique du Langage (TAL). L'Arabe doit sa formidable expansion à partir du VII^{ème} siècle grâce à la propagation de l'Islam et la diffusion du Coran. Les recherches pour le traitement automatique de l'arabe ont débuté vers les années 1960. Les premiers travaux concernaient notamment les lexiques et la morphologie [29].

2.8.1. Variétés d'Arabe

L'Arabe est un terme générique de nombreuses variétés :

L'Arabe est un terme générique de nombreuses variétés :

- classique : la langue du Coran, parlée à l'époque de l'expansion arabo-musulmane ;
- arabe littéral : une forme modernisée mais peu différenciée de l'Arabe classique, qui est la langue écrite commune de tous les pays arabophones ;
- les dialectes arabes : langues orales parlées dans les pays arabes, issues de l'Arabe classique, avec des substrats, superstrats et emprunts différents selon les régions. Les dialectes peuvent être assez différents les uns des autres : un Irakien par exemple pourrait avoir du mal à comprendre le dialecte marocain à premier abord. Cependant, et même si

ces différences existent, des locuteurs arabes de différents pays peuvent se comprendre sans trop de difficultés.

Les dialectes de l'Arabe peuvent être divisés en deux groupes : L'Arabe occidental, qui inclut les dialectes parlés au Grand Maghreb, et l'Arabe oriental, qui peut être subdivisé en Égyptien, Levantine, et arabe du Golfe. Ces divers dialectes diffèrent considérablement de l'un à l'autre et de l'AS (fig.2.10) [28], [30].

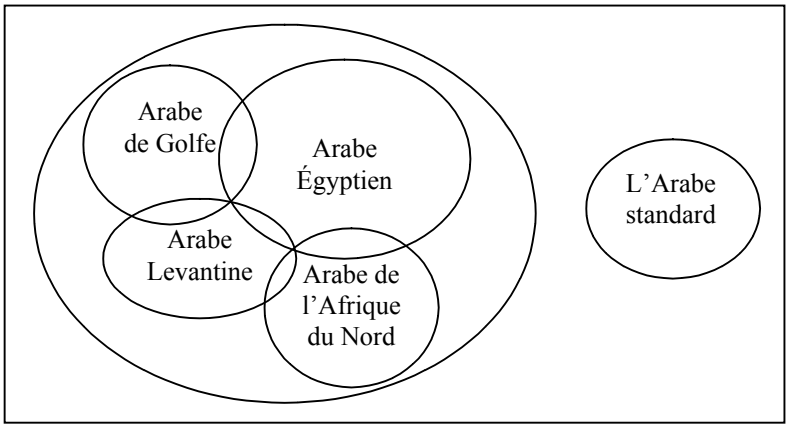


Figure 2.10 : Variétés linguistiques arabes

2.8.2. Alphabet arabe

L'alphabet arabe est utilisé principalement pour écrire la langue arabe. Bien que très souvent désigné comme un alphabet, à la manière de l'écriture d'autre langues sémitiques, c'est en fait un ''abjad'', terme décrivant un système d'écriture ne notant que les consonnes de la langue [31].

L'écriture arabe va de la droite à la gauche et lie les lettres de son alphabet selon des règles de ligatures bien définies, et ceci dans les deux modes manuscrit ou imprimé. Son alphabet comporte 28 lettres et 3 voyelles longues ou courtes.

La graphie des lettres est différente selon leur position dans le mot. Ainsi, la lettre ع[ɛ] est transcrit عَادَ [ɛaada] en début de mot, لَعِبَ [laɛiba] en milieu de mot, مَعَ [maɛa] en fin de mot, وَدَّعَ [waddaɛa] isolé en fin de mot. Il résulte 78 formes graphiques à partir des 28 lettres. Par ailleurs, la distinction minuscule-majuscule n'existe pas [32].

2.8.2.1. Les signes diacritiques

Les voyelles brèves sont figurées par des symboles appelés signes diacritiques. Ces symboles sont absents à l'écrit dans la majorité des textes arabes (fig.2.11). Ces symboles sont transcrits de la manière suivante :

- la fatha [a] est symbolisée par un petit trait sur la consonne (بَ [ba]).
- la damma [u] est symbolisée par un crochet au-dessus de la consonne (بُ [bu]).

- la kasra [i] symbolisée par un petit trait au-dessous de la consonne (بِ [bi]).

Un petit rond (°) symbolisant la *sukuun* (سكون) est apposé sur une consonne lorsque celle-ci n'est liée à aucune voyelle (بَعْدَ [baɛda]).

2.8.2.2. Le tanwin

Le signe de tanwin est ajouté à la fin des mots indéterminés. Il est en relation d'exclusion avec l'article de détermination (ال) placé en début de mot. Les symboles de tanwin sont au nombre de trois et sont constitués par dédoublement des signes, ce qui se traduit par l'ajout du phonème (n) au niveau phonétique (fig.2.11).

[an] : signe ً (بً [ban]).

[un] : signe ٌ (بٌ [bun]).

[in] : signe ٍ (بٍ [bin]).

2.8.2.3. La chadda

Le signe de la chadda peut être placé au-dessus de toutes les consonnes en position non initiale. La consonne qui la reçoit est alors analysée en une séquence de deux consonnes identiques (fig.2.11) [32]:

Signe ّ عَدَّلَ [ɛaddala]

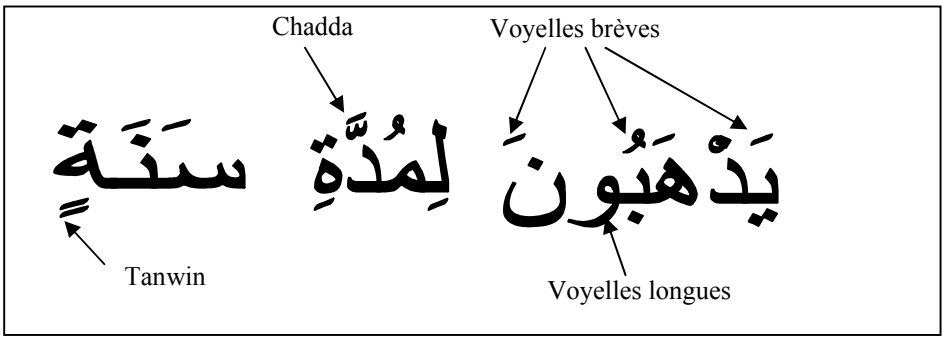


Figure 2.11 : Exemple d'une phrase voyellée [yavhabuuna limuddati sanatin]

Le tableau 2.3 représente l'ensemble des caractères de l'arabe avec pour chaque graphème, sa transcription phonétique.

Tableau 2.3: Correspondance graphème phonème de la langue arabe suivant API [30]

Isolé	Debut	Milieu	Fin	Nom	Phonème
ا	ا	ا	ا	'alif	[a:]
ب	بـ	ـبـ	ـب	baa'	[b]
ت	تـ	ـتـ	ـت	taa'	[t]
ث	ثـ	ـثـ	ـث	thaa'	[θ]
ج	جـ	ـجـ	ـج	gym	[ʤ]
ح	حـ	ـحـ	ـح	Haa'	[ʜ]
خ	خـ	ـخـ	ـخ	khaa'	[x]
د	د	ـد	ـد	daal	[d]
ذ	ذ	ـذ	ـذ	dhaal	[ð]
ز	ز	ـز	ـز	zayn	[z]
ر	ر	ـر	ـر	raa	[r]
س	سـ	ـسـ	ـس	syn	[s]
ش	شـ	ـشـ	ـش	shyn	[ʃ]
ص	صـ	ـصـ	ـص	Saad	[ʂ]

ض	ضـ	ـضـ	ـض	Daad	[d]
ط	طـ	ـطـ	ـط	Taa'	[t]
ظ	ظـ	ـظـ	ـظ	Zaa'	[z]
ع	عـ	ـعـ	ـع	'ayn	[ʕ]
غ	غـ	ـغـ	ـغ	ghayn	[ʁ]
ك	كـ	ـكـ	ـك	kaaf	[k]
ق	قـ	ـقـ	ـق	qaaf	[q]
ف	فـ	ـفـ	ـف	faa'	[f]
ل	لـ	ـلـ	ـل	laam	[l]
ن	نـ	ـنـ	ـن	nuwn	[n]
م	مـ	ـمـ	ـم	mym	[m]
ه	هـ	ـهـ	ـه	haa'	[h]
و	و	ـو	ـو	waaw	[u:]
ي	يـ	ـيـ	ـي	yaa'	[i:]
ء	أ	ؤ	ئ	hamza	[ʔ]

2.8.3. Phonétique et phonologie de la langue arabe

À côté des trois voyelles brèves [a] [u] [i], il existe trois voyelles longues [aa] [uu] [I] qui s'opposent aux précédentes par une durée plus importante sur le plan temporel. L'ensemble des voyelles brèves et longues est dit oral, car elles sont émises sans l'intervention de la cavité nasale. L'AS contient 28 consonnes qui correspondent chacune à un phonème. La hamza ء [ʔ] a un statut particulier en ce sens que certains grammairiens la considèrent comme le 29ème (tab.2.4)

Tableau 2.4 : Classement des consonnes arabes selon leur mode d'articulation [33]

	Occlusives	Emphatiques	Fricatives	Nasales	Liquides	Glides (semi-voyelles)
Labiales	ب b		ف f	م m		و w
Interdentales		ظ V	ذ v ث f			
Dentales	د d ت t	ض D ط T		ن n	ر r ل l	
Sifflantes		ص S	ز z س s			
Palatales	ج J		ش X			ي y
Vélaires	ك k		غ g خ h			
Uvulaire	ق q					
Pharyngales			ع e ح H			
Glottales	ء '		ه h			

À l'instar des autres langues, les consonnes de l'Arabe sont classées selon leur mode d'articulation (occlusif, fricatif, nasal, glissant ou liquide), leur lieu d'articulation (labial, dental ou vélo-palatal) et leur voisement (sonore ou sourd), les phonèmes (fig.2.12):

- spécifiques à l'Arabe qui n'ont pas d'équivalent dans les langues européennes :ص [S], ض [D], ط [T], ظ [Z], ع [ε], ق [q], ح [H], ء[?].
- qui ont des équivalents dans langue française, par exemple : ت [t], ز [z],د [d], س [s], ش [O], ك [k], ف [f], ب [b], ل [l], م [m], ن [n], و [w], ي [y].
- qui ont des équivalents dans plusieurs langues telles que l'espagnol, l'allemand ou l'anglais : ر [r], ذ [v], ه [h], خ [x],ث [c] [30].

Dans ce mémoire nous avons opté pour la notation décrite dans le tableau 2.4 pour faciliter la rédaction.

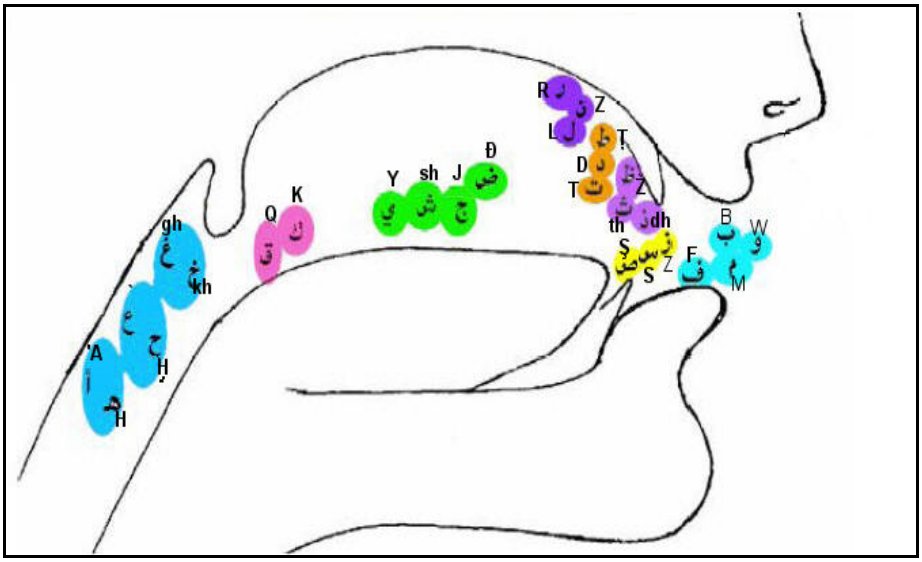


Figure 2.12 : Lieux d'articulation des phonèmes arabes [34]

2.8.4. Problèmes du Traitement Automatique de l'Arabe

Un des aspects complexes de la langue arabe est l'absence des voyelles dans le texte, qui risque de générer une certaine ambiguïté à deux niveaux :

- Sens du mot (problème de voyellation des mots)

  L'ambiguïté vient du mot **العلم** *la science* ou *drapeau* alors que voyellé on aura **العِلْمُ** pour la science et **العَلَمُ** pour le drapeau. Cette ambiguïté pourrait, dans certains cas, être levée soit par une analyse plus profonde de la phrase ou des statistiques (par exemple, il est plus probable d'avoir **العلم الوطني** le drapeau national que la science nationale (fig2.13).

- agglutination des mots ou difficulté à identifier sa fonction dans la phrase (différencier entre le sujet et le complément, etc.).

  Ceci peut influencer les fréquences des mots étant donné qu'elles sont calculées après la détection de la racine des mots qui est basée sur la suppression de préfixes et suffixes. Lors de calcul des scores à partir des titres, il peut arriver que des mots soient considérés comme dérivants d'un même concept alors qu'ils ne le sont pas.

  De plus la capitalisation n'est pas employée dans l'Arabe, ce qui rend l'identification des noms propres, des acronymes, et des abréviations encore plus difficiles [29].

العنوان: أثر العلم **1- العلماء.....** **2- علميا.....** **3- في المحاضرة ليس العلم الوطني ولكن العلم لكل الدول**	Titre : impact de la science 1- les scientifiques…. 2- Scientifiquement… 3- À la conférence non seulement le drapeau national, mais aussi le drapeau de chaque pays

Figure 2.13 : Effet du mot non voyellé **العلم** dans les phrases

2.9. Conclusion

Étant donnée que notre objectif est de réaliser un système de reconnaissance de formes phonémiques en AS, nous n'avons décrit dans ce chapitre, et d'une manière assez brève, que quelques techniques d'analyse acoustique qui sont les plus utilisées pour extraire les caractéristiques pertinentes d'un signal de la parole, et ceci après avoir expliqué la manière de produire la parole chez les être humains, ainsi que les différents niveaux de son analyse. Enfin, nous avons conclu ce chapitre par une description sommaire de l'Arabe standard et son traitement automatique.

CHAPITRE 3 :
RECONNAISSANCE AUTOMATIQUE DE LA PAROLE

3.1. Introduction

La Reconnaissance Automatique de la Parole (RAP) par une machine a fait l'objet d'un effort important de recherche depuis près de cinquante ans. Malgré des progrès significatifs effectués au cours de ces dernières années, les performances des systèmes réalisés sont encore loin d'égaler celles de l'être humain, notamment dans les conditions réalistes de fonctionnement : parole spontanée, présence de bruit ambiant, etc. En bref, reconnaître et comprendre la parole demeure un des grands défis de la reconnaissance des formes et de l'intelligence artificielle, pour lesquels de nombreuses méthodes ont été développées.

Dans ce présent chapitre, nous allons décrire le principe de fonctionnement d'un système de RAP et les différentes méthodes et techniques couramment utilisées dans ce domaine. Par la suite, nous allons voir la reconnaissance de la parole audiovisuelle qui est l'objectif principal de ce mémoire, tout en décrivant la multimodalité de la parole et les différentes approches de fusionner les informations acoustiques et visuelles accompagnant la parole.

3.2. Reconnaissance Automatique de la Parole (RAP)

La RAP est une discipline récente. Vers 1950 apparut le premier système de reconnaissance de chiffres, appareil entièrement câblé et très imparfait. Vers 1960, l'introduction des méthodes numériques et l'utilisation des ordinateurs changent la dimension des recherches. Néanmoins, les résultats demeurent modestes car la difficulté du problème avait été largement sous-estimée, en particulier en ce qui concerne la parole continue. Vers 1970, la nécessité de faire appel à des contraintes linguistiques dans le décodage automatique de la parole avait été jusque-là considérée comme un problème d'ingénierie. La fin de la décennie 70 voit se terminer la première génération des systèmes commercialisés de reconnaissance des mots. Les générations suivantes, mettant à profit les possibilités sans cesse croissantes de la micro-informatique, posséderont des performances supérieures (systèmes multilocuteurs, parole continue).

On peut résumer en quelques dates les grandes étapes de la RAP :

- 1952 : reconnaissance des 10 chiffres, pour un monolocuteur, par un dispositif électronique câblé ;
- 1960 : utilisation des méthodes numériques ;

- 1965 : reconnaissance des phonèmes en parole continue ;
- 1968 : reconnaissance des mots isolés par des systèmes implantés sur gros ordinateurs (jusqu'à 500 mots) ;
- 1969 : utilisation d'informations linguistiques ;
- 1971 : lancement du projet ARPA aux USA (15 millions de dollars) pour tester la faisabilité de la compréhension automatique de la parole continue avec des contraintes raisonnables ;
- 1972 : premier appareil commercialisé de reconnaissance des mots;
- 1976 : fin du projet ARPA ; les systèmes opérationnels sont HARPY, HEARSAY I et II et HWIM ;
- 1978 : commercialisation d'un système de reconnaissance à microprocesseurs sur une carte de circuits imprimés ;
- 1981 : utilisation de circuits intégrés VLSI (Very Large Scale Integration) spécifiques du traitement de la parole ;
- 1981 : système de reconnaissance des mots sur un circuit VLSI ;
- 1983 : première mondiale de commande vocale à bord d'un avion de chasse en France ;
- 1985 : commercialisation des premiers systèmes de reconnaissance de plusieurs milliers de mots ;
- 1986 : lancement du projet japonais ATR de téléphone avec traduction automatique en temps réel ;
- 1988 : apparition des premières machines à dicter par mots isolés ;
- 1989 : recrudescence des modèles connexionnistes neuromimétiques;
- 1990 : premières véritables applications de dialogue oral homme-machine ;
- 1994 : IBM lance son premier système de reconnaissance vocale sur PC ;
- 1997 : lancement de la dictée vocale en continu par IBM.
- 1999 : développement d'un traducteur à usage militaire médical par DARPA.
- 2003 : DARPA a mis un phraselator au profit de l'armée américaine en Iraq et Afghanistan pour traduire des phrases de l'anglais vers 30 langues. Ainsi soldat peut parler en anglais et l'appareil traduit vers la langue locale des citoyens ou bien plus simplement il choisit une phrase et l'appareil la prononce en langue locale [35].

L'appellation RAP (ASR pour Automatic Speech Recognition en anglais) se réfère à plusieurs types de systèmes dont la mission est de décoder l'information portée par le signal vocal. Selon l'information à extraire, on distingue deux types :

- la reconnaissance du locuteur, dont le but est de reconnaître la personne qui parle parmi une population de locuteurs (identificateur) ou de vérifier son identité (vérificateur) ;
- la reconnaissance de la parole, dont le but est de transcrire l'information symbolique exprimée par le locuteur. On distingue les cas de reconnaisseur monolocuteur, multilocuteurs ou indépendant du locuteur. Une distinction est également faite entre reconnaisseur des mots isolés, des mots connectés et de la parole continue [36].

3.2.1. Système de RAP

La RAP est fondée dans la plupart des systèmes actuels sur une approche probabiliste. Les systèmes sont généralement constitués de deux unités principales, le module de Décodage Acoustique Phonétique (DAP) et le module de modélisation du langage.

Le premier permet, à partir d'une analyse paramétrique du signal à reconnaître, de définir quel est l'élément acoustique qui est le plus probablement produit. Cet élément peut être de différents types : phonèmes, diphones, syllabes, etc. Cette étape franchie, il est nécessaire de mettre en correspondance une suite d'éléments acoustiques avec une forme lexicale. C'est ici qu'intervient le second module. Il permet d'obtenir une information a priori sur le positionnement d'un mot dans le signal à reconnaître par différentes techniques de modélisation soit à base de grammaire, soit purement statistique, soit à base d'approches mixtes telles que les grammaires probabilistes (fig.3.1) [37].

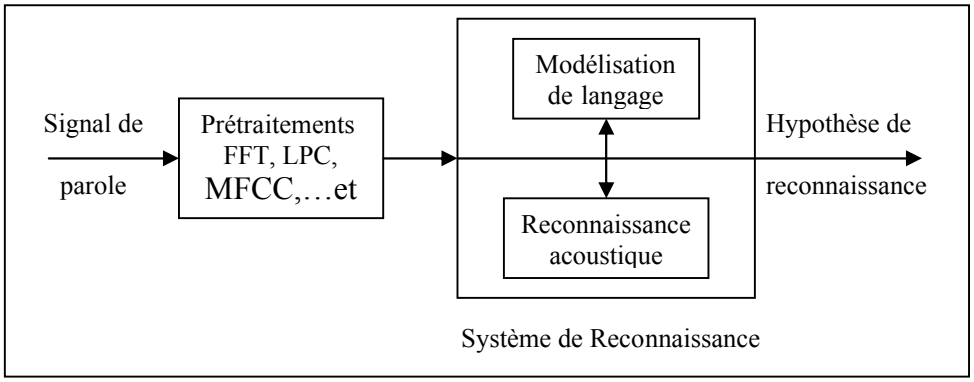


Figure 3.1 : Description symbolique d'un système de reconnaissance de la parole

3.2.2. Reconnaissance et compréhension automatique de la parole

On peut distinguer deux grands types de systèmes :

- de reconnaissance de la parole qui ont pour objectif de décoder les phrases prononcées mot par mot voire phonème par phonème, sans comprendre le sens de la phrase (comprendre s'entend ici par "fournir une représentation sémantique de l'énoncé") ;
- de compréhension de la parole qui ont pour objectif de comprendre la phrase prononcée même si la phrase d'entrée n'est pas reconnue précisément (il y a des erreurs au niveau du DAP).

Il y a deux principales méthodes de RAP, globale et analytique.

3.2.2.1. La méthode globale

Cette méthode considère le plus souvent le mot ou le phonème comme unité de reconnaissance minimale, c'est-à-dire indécomposable. Dans ce type de méthode on compare globalement le message d'entrée (mot, phrase) aux différentes références stockées dans un dictionnaire en utilisant des algorithmes de programmation dynamique ou des modèles de Markov. Cette méthode a pour avantage d'éviter l'explicitation des connaissances relatives aux transitions qui apparaissent entre les phonèmes.

La généralisation de la méthode à des unités enchaînés présente un certain intérêt car les unités phonétiques sont représentées par des modèles et les connaissances phonétiques, lexicales et syntaxiques sont compilées dans un seul réseau, ce qui rend le système de reconnaissance très homogène, des niveaux acoustiques jusqu'aux niveaux linguistiques. La reconnaissance consiste alors à trouver le meilleur chemin dans le réseau global pour reconnaître une phrase prononcée.

Ce type de méthode est utilisé dans les systèmes de reconnaissance :

- des mots isolés ;
- d'unités enchaînées ;
- de la parole dictée avec pauses entre les mots [19].

3.2.2.2. La méthode analytique

Cette méthode fait intervenir un modèle phonétique du langage. Il y a plusieurs unités minimales pour la reconnaissance qui peuvent être choisies (syllabes, demi-syllabes, diphones, phonèmes, phones homogènes, etc.). Le choix parmi ces unités dépend des performances des méthodes de segmentation utilisées. La reconnaissance dans cette méthode, passe par la segmentation du signal de la parole en unités de décision puis par l'identification

de ces unités en utilisant des méthodes de reconnaissance des formes (classification statistique, réseau de neurones, etc.) ou des méthodes d'intelligence artificielle (systèmes experts, par exemple). Cette méthode est beaucoup mieux adaptée pour les systèmes à grand vocabulaire et pour la parole continue. Les problèmes qui peuvent apparaître dans ce type de système sont dus en particulier aux erreurs de segmentation (délétions, insertions, substitutions, recouvrements) et d'étiquetage phonétique. C'est pourquoi le DAP est fondamental dans une telle approche.

3.2.3. Facteurs de complexités

La RAP pour la voix d'un seul locuteur est déjà un problème élémentaire, en raison de la variabilité intralocuteur inhérente au processus humain de production de la parole. La reconnaissance multilocuteurs est un problème encore plus difficile dans la mesure où à la variabilité intra s'ajoute la variabilité interlocuteur.

Si le locuteur marque une pause après chaque mot de l'énoncé, la complexité du problème est réduite, puisque les frontières des mots sont alors disponibles, contrairement au cas de la parole continue. De plus, les mots parlés peuvent éventuellement être considérés comme des entités globales et non comme une suite d'entités élémentaires.

Il sera plus difficile de travailler sur un vocabulaire très étendu (quelques milliers ou dizaines de milliers de mots) que sur un vocabulaire très restreint (quelques dizaines de mots). La taille du vocabulaire est cependant un paramètre insuffisant, un ensemble de mots très différents les uns des autres étant plus facile à traiter que des mots proches phonétiquement.

La prise en compte du langage produit par l'utilisateur (et éventuellement de sa sémantique, et sa pragmatique dans le cadre d'une application) sera plus facile pour un langage rigide, très contraint, que si toute la souplesse de la langue naturelle parlée peut être rencontrée.

L'environnement acoustique et les conditions de prise de son constituent un facteur important : la présence de bruit, même stationnaire, dégrade en général fortement les performances des systèmes de reconnaissance. De plus, si ce bruit est intense, il induit une augmentation de la variabilité [19].

3.3. Reconnaissance Automatique la Parole AudioVisuelle (RAPAV)

Les énormes progrès réalisés ces dernières années dans les domaines du traitement du signal, de la reconnaissance des formes et de l'intelligence artificielle ont abouti à des systèmes de RAP à élocutions continues tirées d'un grand vocabulaire multilingue qui fonctionnent indépendamment du locuteur. Malgré ces progrès, la plupart des systèmes de RAP manquent de robustesse. Ainsi, leurs performances décroissent considérablement quand

les conditions de test s'éloignent beaucoup des conditions d'apprentissage. En effet, il existe un décalage important entre les expérimentations des systèmes en laboratoire et les applications réelles.

Le manque de robustesse est dû au fait que dans une application réelle, l'environnement sonore n'étant pas protégé; le signal de la parole est susceptible d'interférer avec le bruit. L'effet 'cocktail party', par exemple, survenant dans le cas d'une communication orale donnée au milieu d'une conversation de groupe, perturbe énormément le processus de reconnaissance parce que le bruit de foule peut avoir une signification linguistique interprétable par le système de RAP. Le bruit environnant peut, de plus, prendre des formes variées ; si le système doit travailler dans un milieu bruité pour lequel il n'a pas été entraîné, ses performances chutent en dessous de la limite acceptable.

Les méthodes de RAP bruitées prennent effet, la plupart du temps, au niveau du processus acoustique et consistent à extraire des indices acoustiques résistant au bruit ou à rehausser le signal acoustique. Ces méthodes permettent l'obtention de bons résultats de reconnaissance quand le bruit est stationnaire et non corrélé avec le signal de la parole. Une méthode fondée sur le découpage du spectre de parole en bandes dans lesquelles l'identification des unités de la parole est indépendante, ce qui permet d'améliorer la décision finale de reconnaissance.

Pour augmenter les performances et la robustesse d'un système de RAP, d'autres méthodes reposent sur une meilleure prise en compte des phénomènes de perception de la parole. Dans ce sens, il a été démontré à travers des études et des expérimentations, qu'une meilleure liaison entre la perception de la parole chez l'homme et dans la machine permettrait, d'une part, une meilleure compréhension du processus de communication orale chez l'homme, et d'autre part, l'amélioration de la performance et de la robustesse du processus de perception chez la machine. Cette dernière peut se traduire par l'utilisation conjointe, dans les systèmes de RAP, du signal de parole et d'autres sources de connaissances qui sont par exemple :

- le sujet de conversation, puisque sa connaissance facilite l'accès au lexique puis la compréhension du message oral ;
- les modules de langage, étant donnée que la compréhension de la parole repose sur l'utilisation des règles de langage sous forme de grammaire et de sémantique ;
- les indices prosodiques, car l'accentuation et la mélodie affectent le processus de perception.

Les informations visuelles devraient aussi constituer une source de connaissances pertinentes puisque, d'une part, les lèvres participent au processus de production de la parole et que, d'autre part, la lecture labiale est une partie intégrante du processus de perception de la parole. En effet, les interlocuteurs utilisent non seulement des informations auditives, mais aussi des informations visuelles portant principalement sur la forme et le mouvement des lèvres du locuteur. L'importance des informations visuelles pour l'identification d'un message oral dépend de la dégradation du signal acoustique et de la complexité linguistique du message [30].

3.3.1. Multimodalité de la parole

Il est clair que la communication parlée Homme-Homme (HH) est bimodale en nature, contenant les parties acoustique et visuelle. Alors que la plupart d'entre nous croient que la perception de la parole est accomplie en exploitant uniquement les propriétés acoustiques de signal de la parole, plusieurs résultats de recherches ont montré que cette perception est complétée en exploitant de plus les propriétés visuelles [39].

Cet aspect bimodal de la perception de la parole se retrouve également dans la production. La parole est produite par les vibrations des cordes vocales et par certains organes articulatoires tels que la trachée artère, les cavités buccale et nasale, les dents, le palais et les lèvres. Comme certains des ces organes sont visibles, il doit exister une relation implicite entre la parole produite et la parole vue [40].

Une série d'expériences a été réalisée afin d'évaluer la contribution de différentes parties du visage dans la perception visuelle de la parole. Ces expériences ont montré que les lèvres portent à elles seules plus de la moitié de l'information visuelle que procure un visage complet. Elles ont aussi montré que la vue des dents augmente l'intelligibilité du message, en désambiguïsant des sons qui diffèrent dans la position de la mâchoire. Cependant, l'information de parole visuelle se trouve partout sur le visage du locuteur, et plus on présente d'indices visuels, plus l'information apportée est importante (fig.3.2) [41].

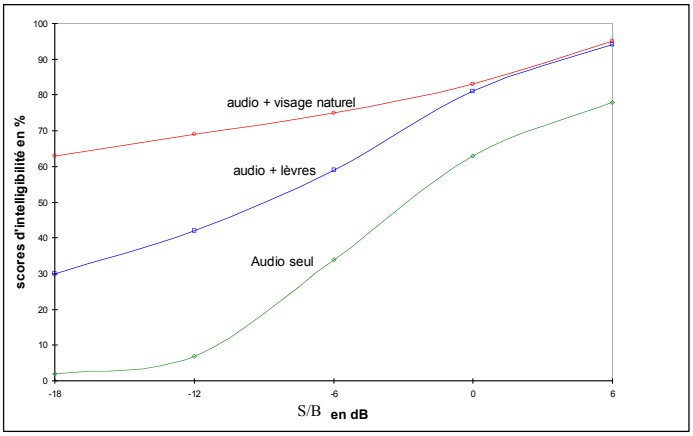


Figure 3.2 : Tests d'intelligibilité réalisés avec différents rapports S/B

Pour les sourds, ils n'ont pas le choix de la modalité ; ils doivent se servir de la vue pour comprendre leur interlocuteur, c'est la lecture labiale. La lecture labiale seule permet d'appréhender environ 40 à 60% des phonèmes d'une langue donnée, et 10 à 20 % des mots ; les meilleurs lecteurs labiaux atteignent des scores supérieurs à 60 %. Il y a des cas rare où des sujets sourds-aveugles pourront percevoir la parole par le toucher du conduit vocal de l'interlocuteur cette méthode est appelée la méthode Tadoma.

Dans le cas d'une communication perturbée, la vue de l'interlocuteur vient en renfort de l'audition. Il a été démonté que lorsqu'on voit son interlocuteur l'intelligibilité de la parole augmente à un débruitage équivalent à 11dB.

Même en l'absence de dégradation acoustique, la lecture labiale favorise la compréhension du message. Ceci est évident lorsque nous essayons de répéter un message un peu complexe, à ce moment, nous aurons tendance à suivre les mouvements des lèvres de l'interlocuteur. Ce même type de performance apparaît pour la compréhension d'une langue étrangère [42].

3.3.2. Lecture labiale

La lecture labiale permet de comprendre le message parlé en mettant en relation les mouvements des lèvres et du visage. On peut la définir comme étant la capacité à comprendre la pensée d'un locuteur en regardant les mouvements de son visage et de son corps tout en utilisant le contexte. Outre le travail de "déchiffrage" des mots, un travail de suppléance mentale est nécessaire pour que les mots aient un sens.

Trois éléments jouent un rôle important dans une bonne lecture labiale : le locuteur, le lecteur labial et l'environnement :

- Le locuteur, c'est à dire celui ou celle qui s'exprime, doit :
  - se mettre obligatoirement face à la personne malentendante ;
  - parler normalement (à trop vouloir articuler ou parler trop fort) ;
  - éviter de parler trop vite ou, à l'inverse, trop lentement ;
  - ne pas parler avec un objet dans sa bouche ;
  - avoir une expression faciale et des gestes appropriés.
- Le lecteur labial, celui ou celle qui lit sur les lèvres de son interlocuteur, doit :
  - avoir suivi une formation adaptée ;
  - pouvoir se concentrer suffisamment longtemps face au locuteur ;
  - être capable de s'adapter à différents locuteurs ;
  - avoir une bonne vision.

- L'environnement, pour sa part, doit être calme et bien éclairé. Les discussions de groupe sont à éviter car un lecteur labial ne peut regarder qu'un locuteur à la fois [43].

3.3.3. Langage Parlé Complété (LPC)

Le Langage Parlé Complété (LPC) est un codage : la main près du visage complète syllabe par syllabe, tout ce que vous dites. Chaque syllabe se code en mettant la main à la position correspondant à la voyelle, les doigts réalisant la clé de la consonne (fig. 3.3 et 3.4).

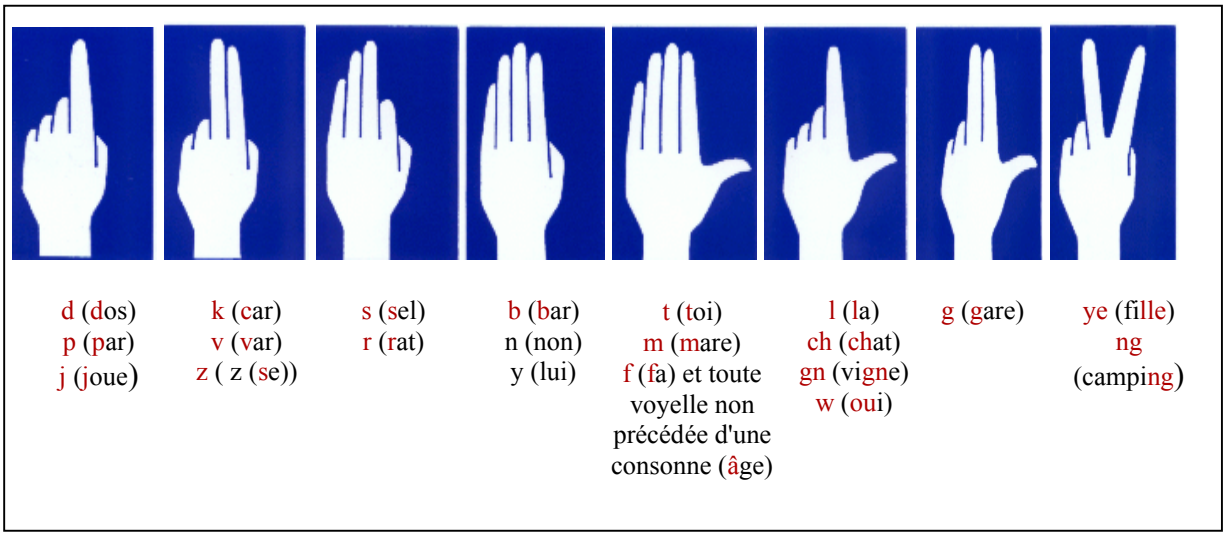


Figure 3.3 : Les différentes configurations des doigts codent les consonnes [44]

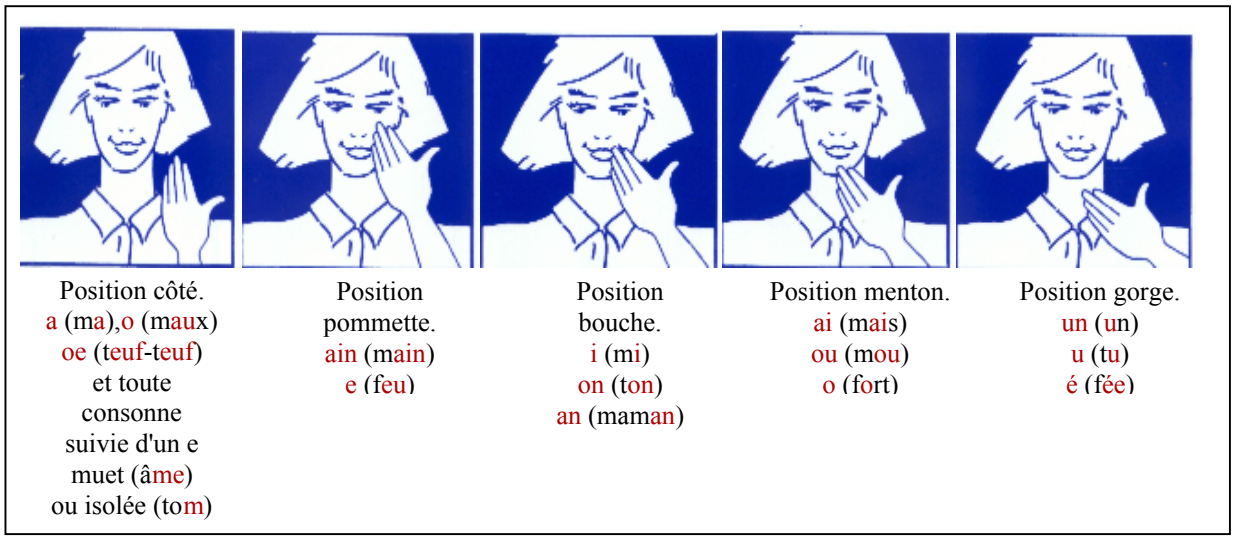


Figure 3.4 : Les cinq positions de la main pour coder les voyelles [44]

Un enfant sourd ou malentendant n'entend pas les sons, ou les perçoit plus ou moins affaiblis et déformés : il s'aide de la lecture labiale, mais de nombreuses confusions sont possibles. Essayer de dire devant une glace " mama", puis " papa " ; vous verrez que le mouvement des lèvres est le même. Le codage de la main attire l'attention de l'enfant, lui

permet de différencier les mouvements des lèvres et d'y associer les sons correspondants (fig.3.5).

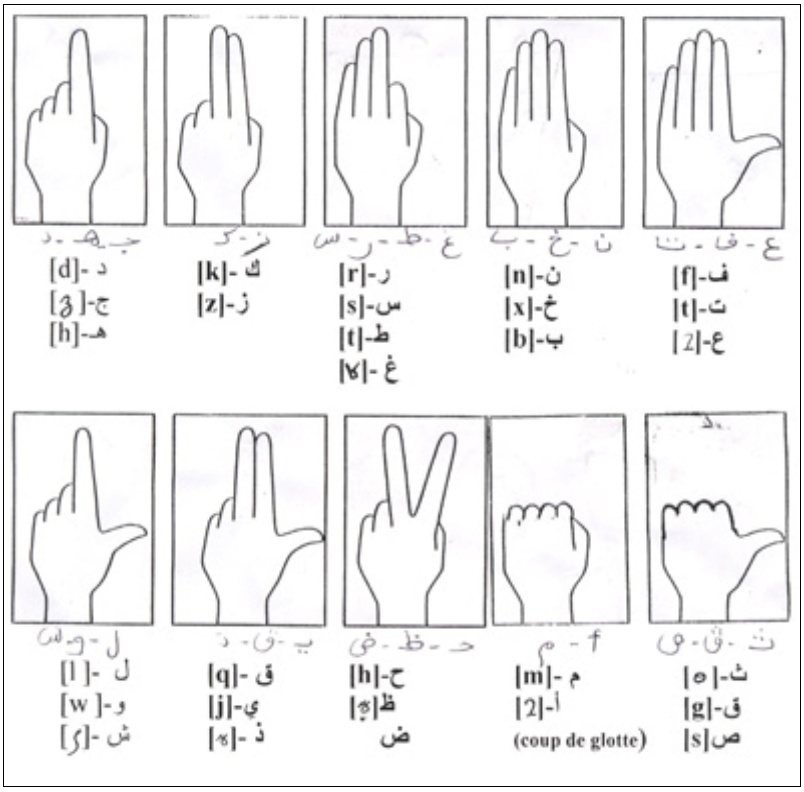


Figure 3.5 : Les dix configurations de doigt codant les consonnes de la langue arabe [45]

Le LPC facilite la compréhension du langage et l'apprentissage du vocabulaire et de la syntaxe. Il aide l'enfant à contrôler son articulation. Il permet une acquisition naturelle de l'expression orale [45].

3.3.4. Effet de McGurk

Dans beaucoup de situations dans lesquelles la communication parlée entre les humains se produit, l'auditeur non seulement entend son locuteur, mais aussi le voit. Bien que la parole soit souvent considérée comme un processus purement auditif, elle est influencée aussi par la vision. Cette réalisation est bien montrée par l'effet de McGurk où dans une expérience il a enregistré les séquences : [baba], [gaga], [papa] et [kaka]. Après une segmentation de ces signaux, il les a arrangés pour obtenir d'autres séquences telles que [ba] audio + [ga] vidéo, [ga] audio + [ba] vidéo, [pa] audio + [ka] vidéo et [ka] audio + [pa] vidéo. Ensuite, il a exposé ces séquences à des groupes de sujets de différents âges en leur demandant de répéter ce qu'ils viennent d'entendre. Il a constaté deux types de réponses :

- illusion de type fusion : [ba] audio + [ga] vidéo et [pa] audio + [ka] vidéo produisent des séquences telles que [dada] et [tata] respectivement. Autrement dit, les sujets ont perçu des séquences qui n'ont pas été présentées dans les deux modalités audio et vidéo ;
- quant à [ga] audio + [ba] vidéo et [ka] audio + [pa] vidéo produisent des séquences telles que [bagba], [gabga] et [kapka], [pakpa]…etc. il s'agit d'une illusion de type combinaison (combinaison des deux séquences).

Ces deux phénomènes forment ce que l'on appelle depuis l'effet de McGurk où il a été prouvé qu'avec audition normale sous bonnes conditions d'écoute, la lecture labiale a été utilisée [38].

3.3.5. Sources des informations visuelles

L'une des principales questions posées dans la perception de la parole audiovisuelle est, quelle partie du visage est considérée comme source des informations visuelles ? La partie visible de l'articulation de la parole peut être trouvée un peu partout dans le visage, néanmoins, la moitie inférieure du visage porte la grande majorité des ces informations. Les facteurs principaux sont les articulateurs visibles composés aussi bien de la bouche incluant les lèvres, la langue et les dents que la partie inférieure de la mâchoire [45].

La contribution des informations visuelles à la perception de la parole dépend de la situation de communication orale, étant moins importante dans une situation normale que dans une situation difficile. Ainsi, dans une situation de communication orale difficile, la contribution des informations visuelles dépend :

- du couple langue étrangère/langue maternelle, dans le cas où les interlocuteurs utilisent une langue étrangère ;
- de la qualité relative des informations acoustiques de la parole et aussi de l'évaluation que le locuteur fait sur cette qualité. Ainsi, il a été prouvé que la contribution du visuel pour la perception des logatomes de types [VCV] (Voyelle Consonne Voyelle) augmente avec la dégradation des informations acoustiques et décroît avec la dégradation des informations visuelles. Outre la qualité des informations, s'ajoute l'estimation que fait l'interlocuteur sur celle-ci.

En revanche, dans une situation de communication orale normale, la contribution des informations visuelles peut varier :

D'une langue à l'autre, en raison des différences d'ordre phonétique et phonologique ; il y a moins d'illusions de McGurk en Français qu'en Américain et en Japonais qu'en Américain.

- d'un interlocuteur à un autre, en fonction des habitudes culturelles, de la familiarité avec le locuteur, de l'acuité acoustique et visuelle et par conséquent en fonction de la motivation et de l'expérience de la lecture labiale ;
- inter et intralocuteur, en fonction des modes d'articulation du locuteur ;
- d'une unité à l'autre, puisque les réalisations visuelles des phonèmes tant au niveau des mouvements que des configurations articulateurs visibles peuvent être plus marqués pour certaines classes, [p], [b], [m] que pour d'autres [t], [d], [k], [g] par exemple [38].

3.3.6. Conversion phonèmes-visèmes

Un visème, acronyme de « visual » et de « phoneme » en Anglais, est une unité visuelle de la parole. La forme des lèvres est associée au visème pour représenter un son particulier. Un visème peut être utilisé pour décrire visuellement un ensemble de phonèmes. Il peut être utile pour améliorer la performance de reconnaissance de la parole aussi bien pour l'homme (lecture labiale) que pour la machine (reconnaissance visuelle de la parole) [46].

Des phonèmes différents peuvent avoir des réalisations visuelles similaires : [p], [b] et [m]. Ces phonèmes forment alors un visème. Par conséquent, les visèmes représentent les sous ensemble de phonèmes, chaque sous-ensemble regroupe les phonèmes qui ne peuvent être distingués visuellement. Les différences intervisèmes sont donc significatives, informatives et catégoriques pour la perception de la parole, contrairement aux différences intravisèmes. Le degré de l'intelligibilité de la parole visible est alors directement proportionnel au nombre de visèmes [38].

Tableau 3.1 : Visèmes définis pour les voyelles (a), les consonnes (b) de l'Anglais [38]

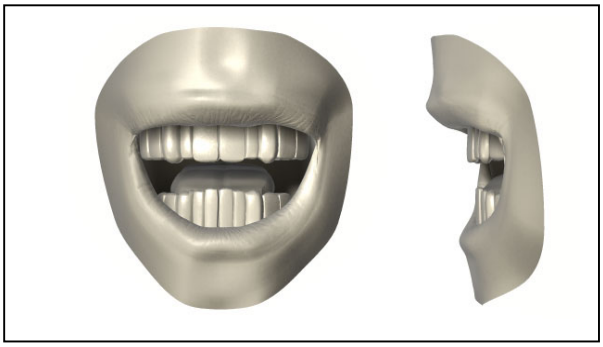
Visème	Description	Composition
{iy}	high front rounded	[iy,ih]
{ɛ}	non high front	[æ,ey,ay]
{a}	lower back	[a, ao, ah]
{uh}	central	[uh,ɛ,ə]
{uw}	high back rounded	[uw]
{oy}	diphtongues	[oy]
{aw}	diphtongues	[aw]
{ow}	diphtongues	[ow]

(a)

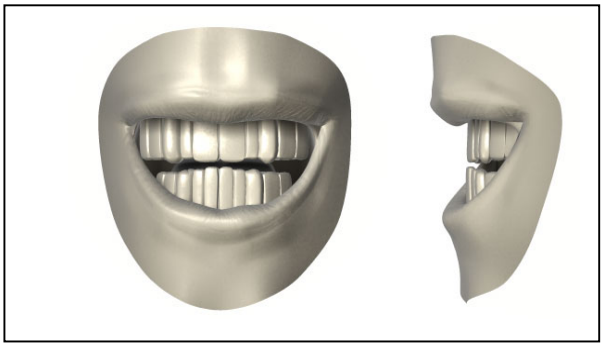
Visème	Description	Nitchie 1950	Burchett 1965	Hazard 1971	Walden 1977	Summerfield 1987	Massaro 1993
{p}	labial stop	[p,b,m]	[p,b,m]	[p,b,m]	[p,b,m]	[p,b,m]	[p,b,m]
{f}	labiodental fricative	[f,v]	[f,v]	[f,v]	[f,v]	[f,v]	[f,v]
{th}	interdental fricative	[th]	[th]		[th,dh]	[th]	[th,dh]
{d}	low visibility	[d,n,t]	[d,n,t]	[d,n,t]	[d,n,t,k, g,j]	[d,t]	[d,n,t,k, g,j,h,ng]
{k}	low visibility	[k,g]	[k,g]	[k,g]		[k,g,n]	
{s}	alveolar fricative	[s,z]	[s,z]	[s,z]	[s,z]	[s,z]	[s,z]
{l}	lateral	[l]	[l]		[l]	[l]	[l]
{r}	retroreflex		[r]		[r]	[r]	[r]
{ch}	palato alveolar	[ch,ch,jh,zh]	[ch,j,sh,zh]	[ch,j,sh]	[sh,zh]	[sh,zh]	[ch,zh,jh]
{w}	labial approximant		[w]	[w]	[w]	[w]	[w]
{y}		[y]				[y]	
{h}		[h]	[h]				

(b)

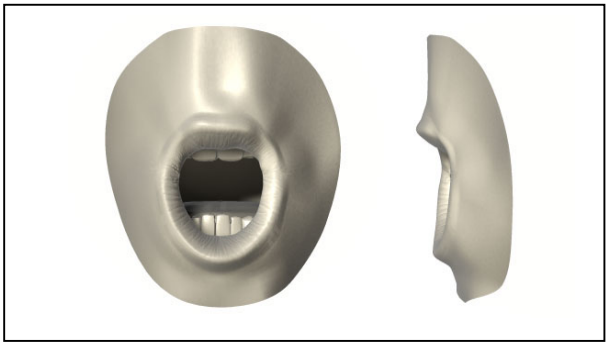
Dans les figures suivantes sont présentés en 3D des visèmes anglais, à gauche vue de face, à droite vue de profile.



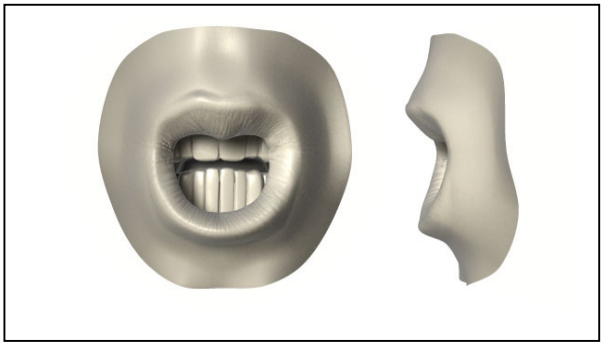
A et I



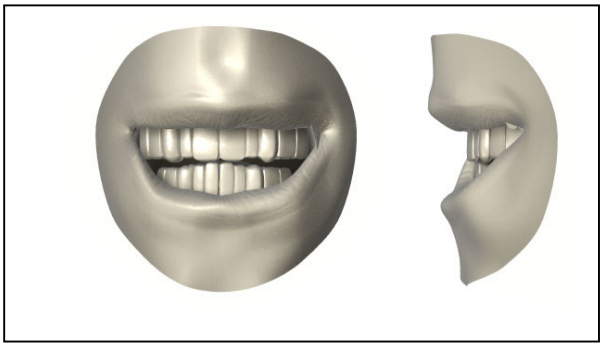
E



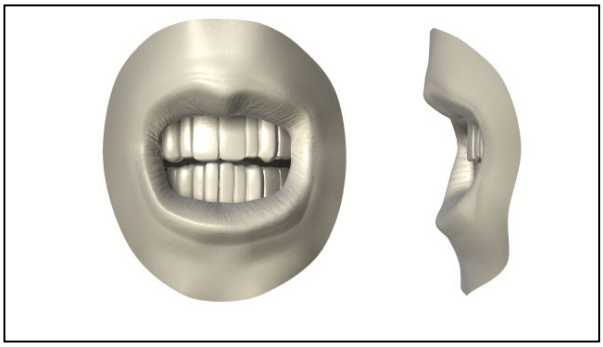
O



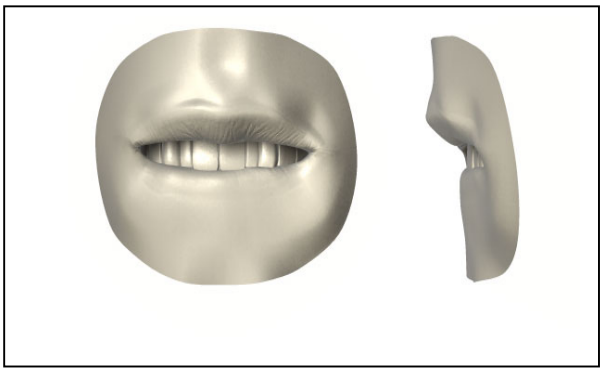
U



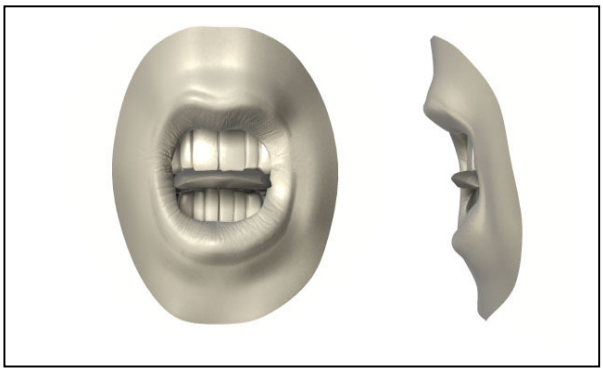
C D G K N R S Y Z



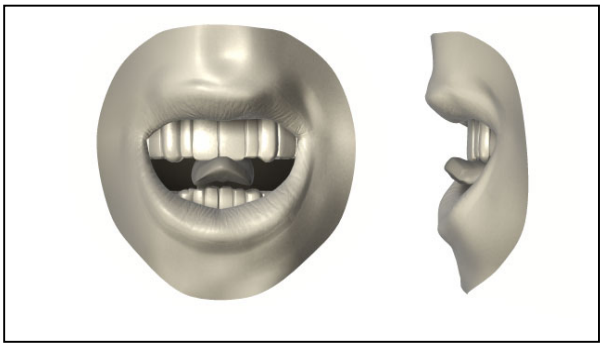
C D G J K N R S Y Z



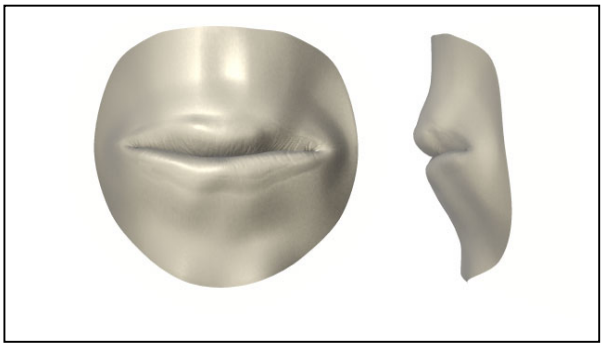
F V



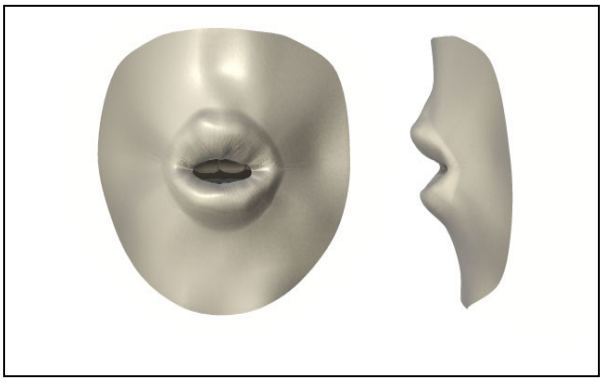
TH



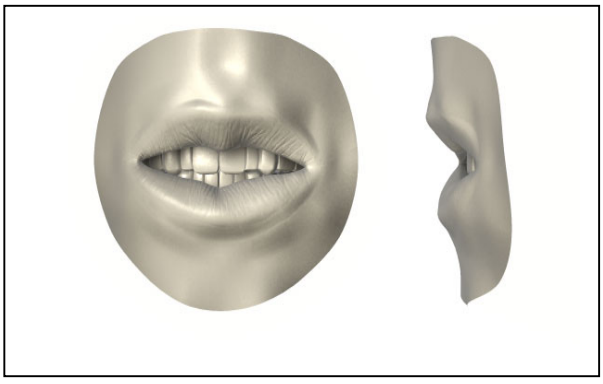
L



B M P



W Q



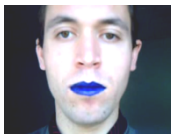
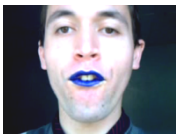
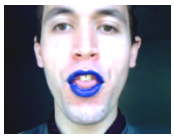
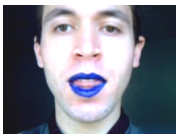
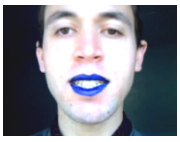
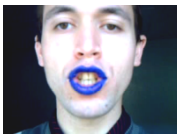
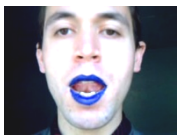
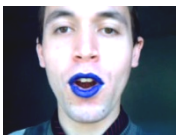
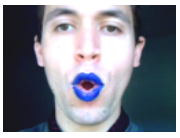
Reste

Figure 3.6 : Représentation 3D de face et de profil des visèmes anglais [47]

Malheureusement la définition des visèmes pour réaliser un système de reconnaissance audiovisuelle de la parole pour la langue arabe n'est pas disponible à grand public comme c'est le cas pour l'Anglais ou le Français. Cependant, dans [5], la définition des visèmes arabes a été faite à l'aide des visèmes anglais, et ce dans le but de réaliser un système de synthèse de la parole à partir de texte prononcé à l'aide d'une tête parlante représentée en 3D. Ce même système pourrait être utilisé pour aider des malentendants, des enfants ou même des gens qui veulent apprendre la langue arabe comme langue étrangère, à comprendre le texte prononcé et même à s'entraîner à bien articuler les différents phonèmes.

En se basant principalement sur cette étude et les différents lieux d'articulation des phonèmes arabes, et bien d'autres définitions des visèmes anglais trouvés dans la littérature [42], [48], [49], [50], [51], [52], [53], nous proposons dans le tableau suivant une représentation des différents visèmes de la langue arabe.

Tableau 3.2 : Représentation des visèmes de l'AS

1	2	3	4	5	6
سكوت ب م	ف	ظ ض ذ ث	ط د ت	س ص ز	ش ج
					
7	8	9	10	11	
ل ن	ر	خ غ ح ه ع أ ك ق	أ و	إ ي	
					

3.3.7. Système de RAPAV

Les premières tentatives d'intégration des indices visuels dans les systèmes de reconnaissance de la parole furent lancées pour résoudre les situations où l'audition n'était pas suffisante pour assurer la compréhension de la parole. Bien souvent, il s'agissait de rajouter quelques informations visuelles à l'intérieur d'un système de RAP permettant la détection de phonèmes séparées ou d'enchaînement [VCV]. La parole visuelle resta un certain temps une force d'appoint au système de perception auditive. L'échec des méthodes purement auditives pour la RAP en environnement bruité, ainsi que l'accroissement des puissances de calcul, ont favorisé l'apparition de systèmes audiovisuels de reconnaissance automatique de la

parole. Les recherches ont d'abord tenté de mettre en évidence l'apport de l'image pour la reconnaissance en environnement bruité. La variabilité des méthodes employées est à la mesure de la variabilité des architectures possibles de reconnaissance [40].

Un système de reconnaissance automatique de la parole audiovisuelle pourrait se présenter comme l'indique la figure 3.7.

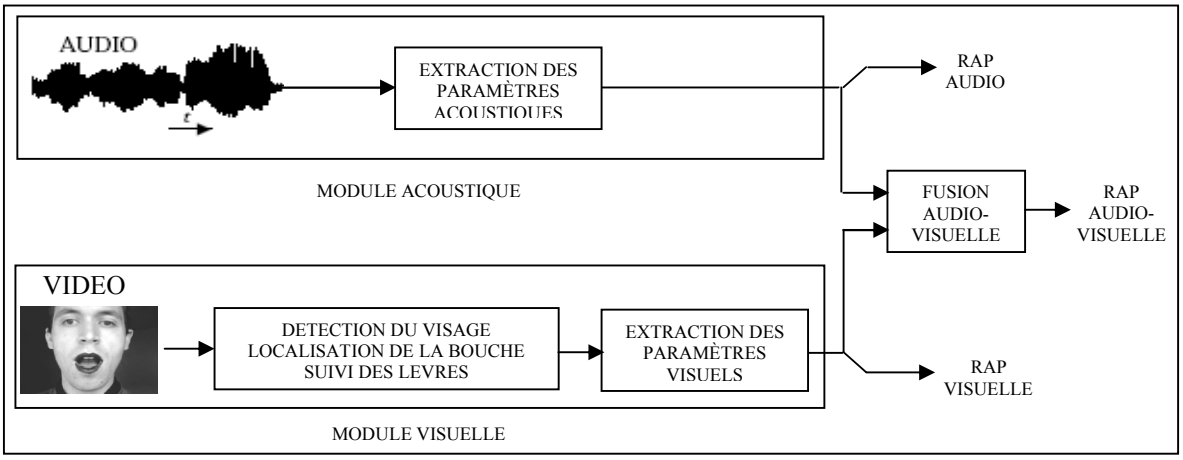


Figure 3.7 : Schéma général d'un système audiovisuel de RAP

Dans cette figure, le sous-système supérieur (module acoustique est réalisé généralement de la même manière que dans un système de RAP purement acoustique. Le sous-système inférieur (module visuel) et le module de fusion des deux parties sont les deux différences des systèmes audiovisuels de RAP par rapport aux systèmes purement acoustiques. C'est dans le module de fusion que se fait l'interaction des informations acoustiques avec les informations visuelles pour produire enfin une décision audiovisuelle.

Fusionner des données issues des capteurs pour prendre une décision avec une sécurité optimale est un enjeu d'importance croissante dans le domaine du traitement de l'information, notamment pour la robotique. Le processus de décision multicapteurs peut être simplifié en trois principales configurations (fig.3.8).

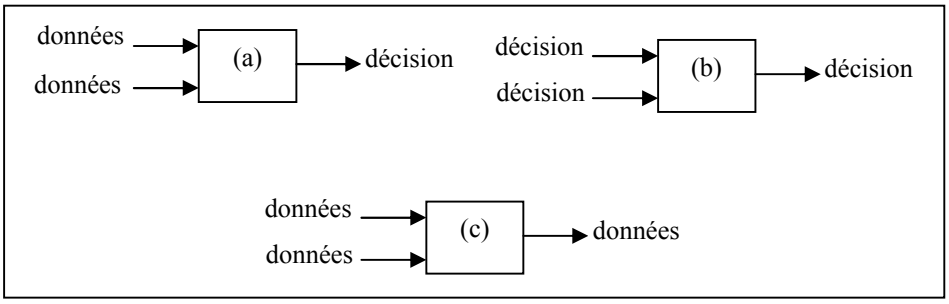


Figure 3.8 : Trois architectures pour la fusion de capteurs en traitement de l'information [42].

Il n'y a pas dans ce cas de processus de fusion proprement dit ; c'est la prise de décision globale qui effectue le regroupement de données.

On peut également effectuer une première prise de décision, partielle ou intermédiaire, sur chaque flux de données, puis prendre une décision globale à partir de ces éléments de décision monocapteur (fig. 3.8 (b)).

Enfin, on peut commencer par réaliser une fusion de données compactant les flux individuels en un flux global, sur lequel s'appliquera ensuite le processus de décision (fig. 3.8 (c)) [43].

Lorsqu'on cherche à fusionner les données auditives et visuelles pour un système de RAPAV, la première question qui se pose est où cette fusion de données se réalise ? Différentes architectures ont été proposées dans la littérature, elles sont classées en quatre modèles de base.

Dans le premier modèle (fig.3.9), l'intégration se fait au niveau des paramètres, ceci dit que les paramètres audio et vidéo sont directement combinés dans un seul vecteur de paramètres. Ce modèle est connu aussi sous le nom de l'Intégration Directe (ID).

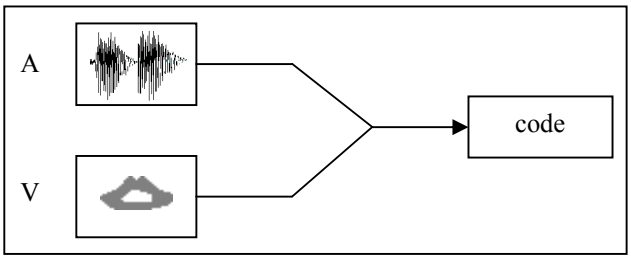


Figure 3.9 : Architecture Intégration Directe (ID)

Contrairement au modèle précédent, la fusion peut également avoir lieu après une identification indépendante de chaque flux de données (fig. 3.10). Par conséquent, la fusion est plutôt une fusion des résultats d'identification.

C'est un processus de fusion tardive, car elle suit l'accès au code dans chaque modalité. Ce type de fusion s'appelle l'Intégration Séparée (IS).

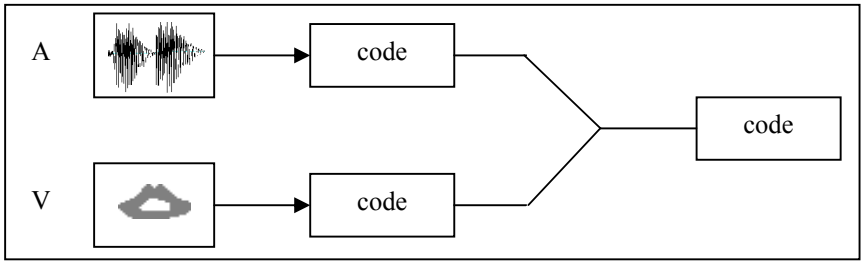


Figure 3.10: Architecture Intégration Séparée (IS)

Le modèle ID se différencie des autres au niveau de la représentation commune aux deux modalités, puisqu'il est le seul à ne pas contenir une telle étape, dans laquelle son et image peuvent être d'une certaine manière comparés, détection de désynchronisation dans un film doublé par exemple. Des études ont prouvé même que des bébés de 4 mois savent orienter leur regard vers une autre bouche articulant une voyelle correspondant au son qu'ils entendent, plutôt que vers une autre bouche produisant une voyelle non cohérente.

Le modèle IS n'est pas optimal de point de vue traitement de l'information, puisqu'il prend des décisions trop tôt sans exploiter les cohérences audiovisuelles préphonétiques. Si l'on considère, par exemple, le cas de deux classes phonétiques qui ne s'opposeraient ni de leur allure auditive moyenne, ni par leur allure visuelle moyenne, mais simplement par leur covariance audiovisuelle, le modèle IS serait incapable de les discriminer. C'est en réalité ce qui advient dans le cas d'une expérience dans laquelle il a été présenté à des sujets le mouvement d'ouverture des lèvres correspondant à une séquence [ba] et [pa], et un son basse fréquence synchronisé avec la mise en action du voisement ; dans ce cas, rien ne permet d'opposer [b] de [p] par l'audition seule ou la lecture labial. Par contre, la coordination temporelle entre son et image permet aux sujets de discriminer [b] de [p] [38], [40], [46], [55], [56].

3.3.8. Extraction des paramètres visuels

Il est clair que les lèvres présentent la partie la plus importante du visage qui contient l'information essentielle de la parole visuelle. À ce stade, la question qui se pose est comment extraire les caractéristiques pertinentes susceptibles à être utilisées dans un système de RAPAV.

L'une des approches les plus utilisées s'appelle l'approche géométrique où les informations visuelles sont représentées par l'étirement intérolabiale ''L'', la séparation intérolabiale ''H'' et l'aire intérolabiale ''S'' comme l'indique la figure 3.11.

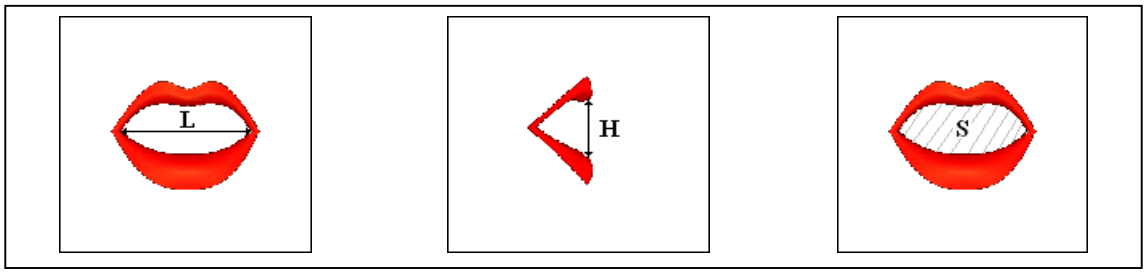


Figure 3.11 : Les trois caractéristiques pertinentes utilisées dans l'approche géométrique

Outre ces trois paramètres, on peut trouver dans la littérature d'autres paramètres tels que ceux utilisés par Lallouache, fig. 3.12.

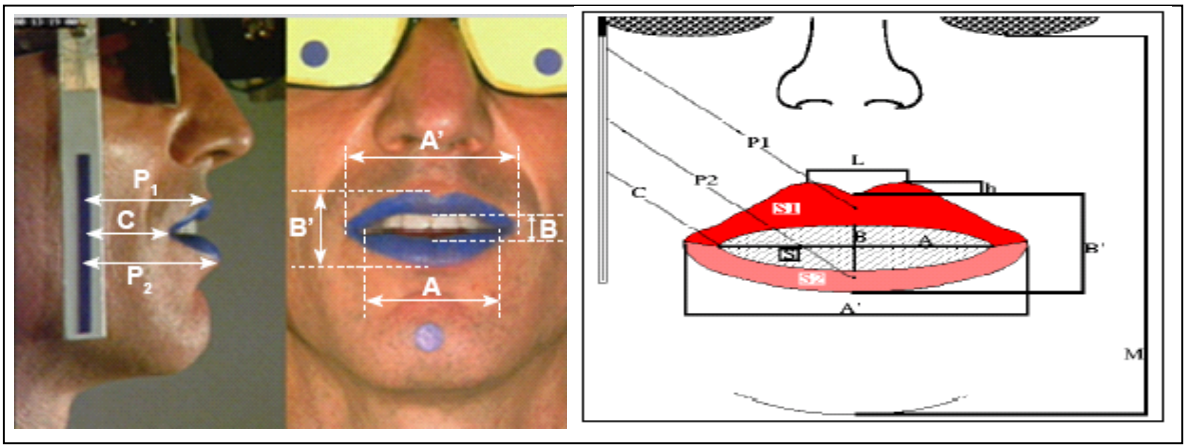


Figure 3.12 : Paramètres visuels utilisés par T. Lallouache (1991) [3]

Les lèvres sont maquillées en bleu pour faciliter l'extraction.

En 1991, Lallouache a mis au point un système automatique de mesure, le poste visage-parole, capable de mesurer de nombreux paramètres de face et de profil. Dans la figure 3.12, les paramètres extraits sont :

- les trois paramètres de face A, B et S ;
- la largeur aux commissures A' : la distance horizontale entre les points extrêmes du contour externe des lèvres ;
- la hauteur du contour externe au centre de la bouche sous l'arc du Cupidon B' : la distance verticale entre le point le plus bas de l'arc du Cupidon et le point le plus bas du contour externe du vermillon des lèvres ;
- les surfaces de la lèvre supérieure S1 et inférieure S2 ;
- la hauteur h et la largeur L de l'arc du Cupidon ;
- la hauteur M entre un point fixe de référence et le menton ;
- les trois paramètres de profile de Abry et Boï C, P1 et P2 (fig. 3.13) ;

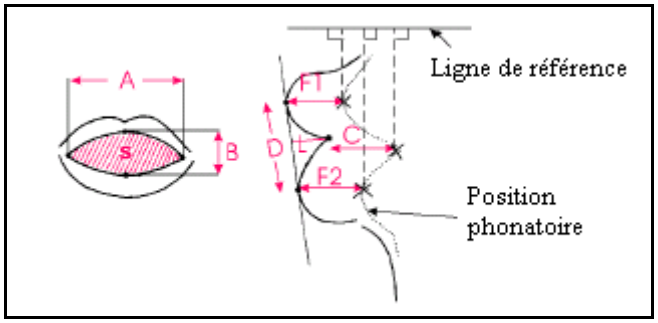


Figure 3.13 : Paramètres utilisés par C. Abry et L.J. Boë en 1986 [4]

Où :

- F1 est la protrusion de la lèvre supérieure ;
- F2 celle de la lèvre inférieure ;
- C est protrusion du contact labial ;
- D est l'aperture du pavillon labial ;
- L sa profondeur.

La deuxième approche utilisée pour extraire les indices visuels est l'approche image (fig.3.14), où les niveaux de gris des pixels ou leur transformation ou même l'image en couleur, sont utilisés comme paramètres pertinents. Dans la plupart des cas on considère les pixels appartenant aux lèvres inférieure et supérieur. Cette approche est connue aussi sous le nom de l'approche d'apparence (apparence des lèvres).

(a) (b)

Figure 3.14 : L'approche image. (a) image en niveaux de gris,
(b) lèvres détectées

Dans la troisième approche, nous faisons appel à l'extraction des contours interne et externe des lèvres comme le montre la figure 3.15. Cette approche est connue sous le nom de l'approche modèle ou modèle de forme (forme des contours des lèvres).

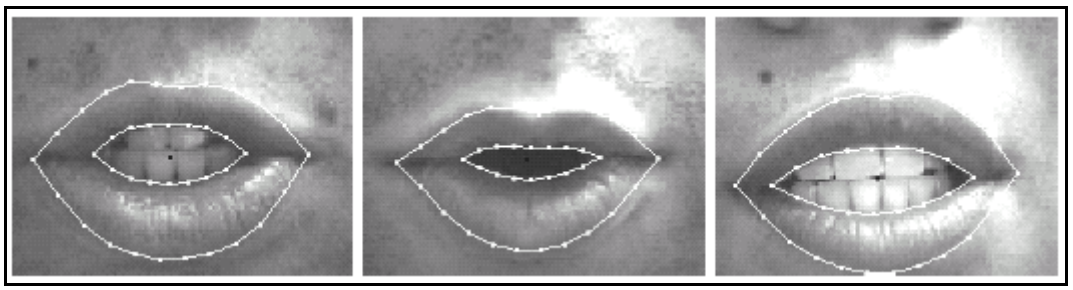


Figure 3.15 : Exemple d'extraction de contour interne et externe.

Une autre approche exploit l'aspect dynamique du mouvement des lèvres, c'est les dérivées premières et/ou secondes des paramètres précédemment cités, c'est l'approche mouvement. Cette approche peut être intégrée avec les approches précédentes [57], [49]

3.4. Conclusion

Dans ce chapitre nous avons vu d'une manière assez brève le principe de fonctionnement d'un système de reconnaissance de la parole RAP après avoir cité les principales stations historiques de ce magnifique domaine, tout en évoquant les différentes techniques utilisées. Ensuite nous avons attaqué un autre aspect de la parole où nous avons, à l'aide des expériences trouvées dans la littérature, expliqué le fait que la parole est non seulement acoustique mais aussi visuelle ; et le fameux effet McGurk l'a bien montré. En effet, la multimodalité de la parole nous impose à intégrer d'autres paramètres visuels dans un système classique de RAP, tout en relatant les différentes manières de fusion de données acoustiques et visuels pour en déduire un système de RAP audiovisuel, et cela dans le but d'améliorer la fiabilité des systèmes de reconnaissance de la parole.

CHAPITRE 4 :
RECONNAISSANCE DES PHONÈMES ET RESULTATS

4.1. Introduction

Dans ce chapitre, nous décrivons nos expérimentations pour construire un système de reconnaissance audiovisuelle, et cela en faisant combiner la modalité acoustique avec la modalité visuelle de la parole. L'intégration des deux modalités se fera en adoptant l'intégration directe ou ID. Pour la réalisation de tel système nous avons choisi d'utiliser les réseaux de neurones.

Pour notre travail, nous allons utiliser uniquement 12 phonèmes comme source de la modalité acoustique avec les 12 visèmes correspondants comme source de la modalité visuelle.

4.2. Reconnaissance acoustique de la parole

L'enregistrement a été fait avec une fréquence d'échantillonnage de 11025Hz en prononçant des mots divers. Les phonèmes qui ont été segmentés manuellement et pris comme base de donnée sont ceux qui ne sont pas voyellés, c.-à-d., les phonèmes présentant un [sukuun].

4.2.1. Traitements acoustiques

Le traitement acoustique consiste à extraire les coefficients MFCC. Ces coefficients constituent l'entrée d'un type de réseau de neurone dit TDNN.

La motivation d'extraction de coefficients MFCC est leur correspondance avec la réponse en fréquence d'une oreille humaine. Pour transformer une fréquence linéaire en une fréquence Mel (fig. 4.1) [58].

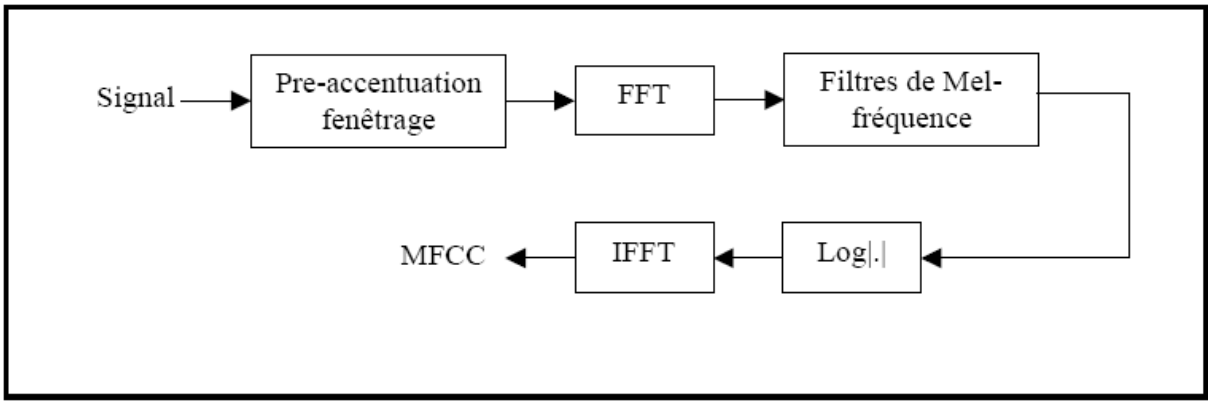


Figure 4.1 : Calcul des MFCCs

Soit un signal numérique de la parole *s(n,i)* sur une fenêtre i.

Avec : $0 \le n \le N-1,\ 1 \le i \le I$.

N est le nombre d'échantillons contenus dans cette fenêtre. Dans notre travail, *N=256* points.

I est le nombre des fenêtres.

s(n) est transformé dans le domaine fréquentiel par la TFD :

$$S(k,i) = \sum_{n=0}^{N-1} s(n,i)\ \exp\left(\frac{-2\pi j\, n\, k}{N}\right) \qquad (4.1)$$

Puis le spectre d'amplitude du signal est filtré par une suite des filtres triangulaires dont les bandes passantes sont de même taille dans l'échelle Mel. Chaque filtre va donner un coefficient cepstral (fig.4.2). Pour notre cas nous avons choisi à utiliser 12 coefficients cepstraux.

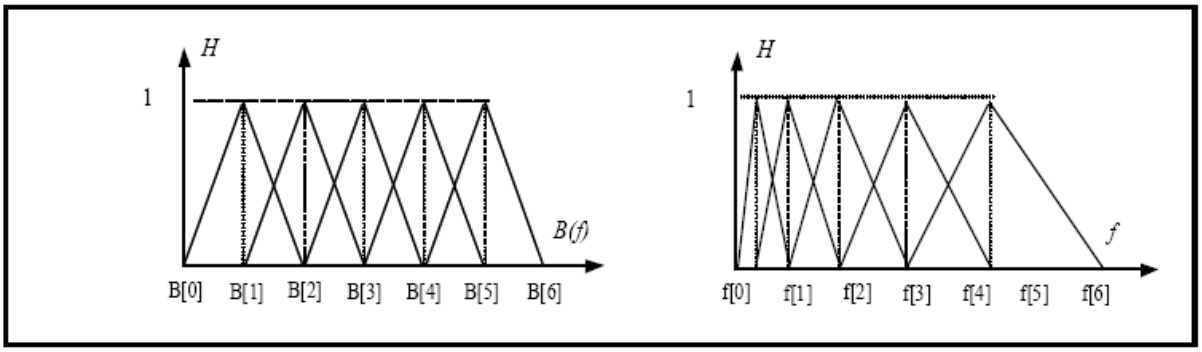


Figure 4.2 : Les filtres triangulaires passe-bande en Mel-Fréq *B(f)* et en fréquence *(f)*

Les points frontières B[m] des filtres en melfréquence peuvent être calculés ainsi :

$$B[m] = B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1} \qquad 0 \le m \le M+1 \qquad (4.2)$$

M est le nombre de filtres.

Les points *f[m]* correspondants dans le domaine de fréquence réelle se calculent par :

$$f[m] = \left(\frac{N}{F_s}\right) B^{-1}\left(B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1}\right) \qquad (4.3)$$

- f_h est la fréquence la plus haute du signal ;
- f_l est la fréquence la plus basse du signal ;
- *Fs* est la fréquence d'échantillonnage.

Les coefficients de chaque filtre se déterminent par :

$$H_m[k] = \begin{cases} 0 & si \quad k \le f[m-1] \\ \dfrac{k - f[m-1]}{f[m] - f[m-1]} & si \quad f[m-1] \le k \le f[m] \\ \dfrac{f[m+1] - k}{f[m+1] - f[m]} & si \quad f[m] \le k \le f[m+1] \\ 0 & si \quad k \ge f[m+1] \end{cases} \qquad (4.4)$$

Ensuite toutes les énergies *S[k]* sont multipliées par les coefficients $H_m(k)$ et nous calculons leur logarithme :

$$E[m] = \log\left[\sum_{k=0}^{N-1} |S[k]|^2 H_m[k]\right] \quad 0 \le m < M \qquad (4.5)$$

Enfin, les coefficients cepstraux de melfréquence (MFCCs) peuvent être obtenus par une transformée en cosinus inverse à partir des coefficients de sortie des filtres. La formule de calcul est la suivante :

$$c[n] = \sum_{m=0}^{M-1} E[m] \cos\left(\frac{\pi n (m+\frac{1}{2})}{M}\right) \text{ avec } 0 \le n < M \qquad (4.6)$$

4.2.2. Configuration de RN

Après avoir calculé les coefficients MFCC et obtenir des vecteurs acoustiques, pour notre cas de 12 coefficients pour chaque vecteur, ces vecteurs seront configurés pour qu'ils constituent des registres à décalage dans le temps chevauchant entre eux. Chaque registre constitue l'entrée du RN de type TDNN (Time Delay Neural Network) capable d'exploiter l'aspect temporel de la parole.

Pour le TDNN, la couche d'entrée est constituée de 48 neurones représentant un registre à décalage de quatre vecteurs acoustiques concaténés l'un après l'autre. Chaque vecteur acoustique représente les 12 coefficients MFCC. Un décalage de deux vecteurs acoustiques permet d'avoir un deuxième registre chevauchant avec le premier de deux vecteurs acoustiques constituant ainsi une deuxième entrée pour le même phonème.

De la même manière un ou plusieurs registres peuvent être créés pour le même phonème. Un phonème peut contenir de 1 jusqu'à 14 registres. Le nombre 14 correspond au phonème le plus long. Deux enregistrements du même phonème peuvent ne pas avoir la même longueur (fig.4.3).

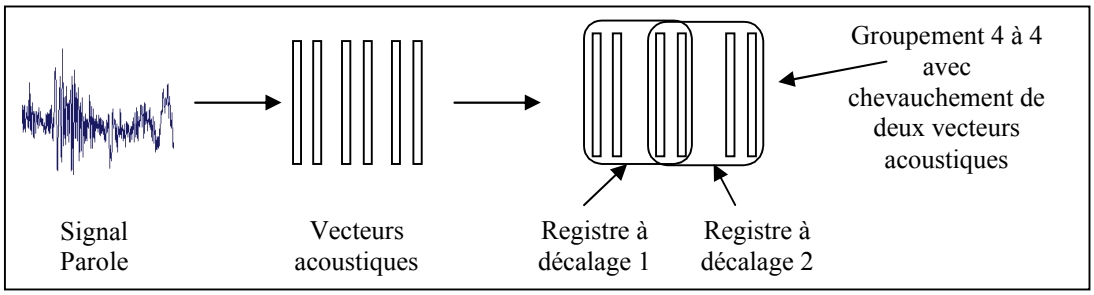


Figure 4.3 : Chaîne de traitements acoustiques

Le nombre de neurones de la couche cachée est choisi à 20 neurones. Quant à la couche de sortie, le nombre est défini à 12 suivant le nombre de phonèmes utilisés.

Les registres de même phonème correspondent à la sortie désirée de ce phonème, ainsi l'apprentissage a été fait avec 7 exemplaires pour chaque phonème. A la fin d'apprentissage, il y aura autant de sorties que de registres.

Pour constituer l'ensemble des sorties désirées, chaque phonème est codé suivant le tableau 4.1.

Tableau 4.1 : Codage des phonèmes pour l'apprentissage du TDNN-MLP

Neurone de sortie / Phonèmes	1	2	3	4	5	6	7	8	9	10	11	12
ب	1	0	0	0	0	0	0	0	0	0	0	0
ف	0	1	0	0	0	0	0	0	0	0	0	0
س	0	0	1	0	0	0	0	0	0	0	0	0
ح	0	0	0	1	0	0	0	0	0	0	0	0
ث	0	0	0	0	1	0	0	0	0	0	0	0
ش	0	0	0	0	0	1	0	0	0	0	0	0
ط	0	0	0	0	0	0	1	0	0	0	0	0
و	0	0	0	0	0	0	0	1	0	0	0	0
ر	0	0	0	0	0	0	0	0	1	0	0	0
ع	0	0	0	0	0	0	0	0	0	1	0	0
ل	0	0	0	0	0	0	0	0	0	0	1	0
ي	0	0	0	0	0	0	0	0	0	0	0	1

Quant à la deuxième partie, elle est constituée de 14 systèmes de MLP puisqu'il y a au maximum 14 registres correspondant au phonème le plus long. Le système MLP le plus court correspond à un seul registre. Un registre peut représenter un phonème. Un registre comme c'est décrit précédemment contient 4 vecteurs acoustiques. Un phonème peut être représenté par un seul vecteur acoustique, de ce fait, ce même vecteur est dupliqué 3 fois pour avoir un

registre de 4 vecteurs acoustiques. Ainsi si un phonème est représenté par deux vecteurs acoustiques, ceux-ci seront dupliqués ; s'il est composé de 3 vecteurs acoustiques, seulement le dernier vecteur sera dupliqué.

L'apprentissage de ces 14 systèmes MLP se fait indépendamment, ainsi chaque système aura ces propres poids synaptiques. Le tableau 4.2 résume le nombre de neurones choisi pour chaque système.

Tableau 4.2 : Configuration des 14 MLPs

Sous système MLP / *N° neurones dans les 3 couches*	1	2	3	4	5	6	7	8	9	10	11	12	13	**14**
Entrée	12	24	36	48	60	72	84	96	108	120	132	144	156	168
Cachée	7	10	20	20	20	20	20	20	20	20	20	20	20	20
Sortie	12	12	12	12	12	12	12	12	12	12	12	12	12	12

4.2.3. Phase d'apprentissage de RN

Par exemple pour le système MLP2, les prototypes peuvent être soit des sorties directes issues de système TDNN; ou bien des vecteurs pris de telle sorte qu'ils représentent tous les phonèmes.

v1= (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) ; v2= (0.99, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0), ces deux vecteurs concaténés peuvent constituer un prototype d'apprentissage pour le phonème 1.

v1= (0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) ; v2= (0.01, 0.98, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0), ces deux vecteurs concaténés peuvent constituer un prototype d'apprentissage pour le phonème 2.

v1= (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1) ; v2= (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.97), ces deux vecteurs concaténés peuvent constituer un prototype d'apprentissage pour le phonème 12.

De la même manière se fait l'apprentissage de tous les autres systèmes. Nous signalons que l'apprentissage a été fait avec sept exemplaires pour chaque classe.

Pour les deux systèmes (TDNN et MLP), l'apprentissage a été fait en suivant l'algorithme de rétro-propagation de l'erreur. La fonction d'activation utilisée est la fonction sigmoïde. Cet algorithme avec la technique d'apprentissage dite partielle sont récapitulés dans ce qui suit :

1. Initialisation des poids synaptiques à des valeurs aléatoires de faible grandeur.

2. Sélection du premier exemplaire de chaque phonème de la base d'apprentissage, et attribution des sorties désirées suivant l'entrée pour chaque phonème en constituant ainsi une matrice *X* de 12 colonnes dont chacune contient les paramètres de chaque exemplaire.

3. Appliquer l'entrée *X* sur le réseau et calculer les sorties actuelles de chaque couche.

$$Y_j(k) = \frac{1}{1 + \exp(\sum_{i=1}^{N} X_i(k).\omega ij)} \qquad (4.8)$$

W_{ij} est la liaison synaptique entre la couche d'entrée et la couche cachée.

Y_i *(k)* est la sortie actuelle de la couche cachée. *K=1,…,12.*

N : nombre de neurones de la couche d'entrée.

$$S_i(k) = \frac{1}{1 + \exp(\sum_{j=1}^{M} Yj(k).vji)} \qquad (4.9)$$

M : nombre de neurones de la couche de cachée.

V_{ij} est la liaison synaptique entre la couche cachée et la couche de sortie.

Si(k) est la sortie actuelle de la couche de sortie.

4. Calcul de la somme des erreurs quadratiques au niveau des neurones de la couche de sortie par :

$$E(k) = \frac{1}{2}\sum_{j=1}^{N} (Dj(k) - Sj(k))^2 \qquad (4.10)$$

K=1,…,12.

D_j est la sortie désirée de j^{ème} neurone et *Si* réponse actuelle du j^{ème} neurone.

5. Calculer l'erreur totale des premiers exemplaires pour tous les phonèmes :

$$ET = \sum_{i=1}^{12} E(k) \qquad (4.11)$$

Si *ET* est inférieure à une valeur seuil fixée au départ, alors aller à 9.

Sinon, aller à l'étape 6.

6**.** Pour chaque sortie du réseau calculer le terme d'erreur :

$$\delta_i^1 = (Dj - Sj).(1 - Sj).Sj \qquad (4.12)$$

Pour le neurone j de la couche cachée, calculer l'erreur :

$$\delta_i^2 = (1 - Y_i).Y_i.\sum \delta_j^1.vij \qquad (4.13)$$

7. Mettre à jour chaque poids synaptique du réseau :

$$wij(k+1) = wij(k) + \alpha.\delta_j^2.Xi + \xi.(wij(k) - wij(k-1)) \qquad (4.14)$$

$$vij(k+1) = vij(k) + \alpha.\delta_j^1.Xi + \xi.(vij(k) - vij(k-1)) \qquad (4.15)$$

Avec :

- α : le coefficient d'apprentissage compris entre 0 et 1 ;
- ξ : coefficient de viscosité ;
- w_{ij} : matrice des poids synaptiques entre la couche d'entrée et la couche de cachée ;
- v_{ij} : matrice des poids synaptiques entre la couche cachée et la couche de sortie.

8. Retourner en 1 avec les nouveaux poids synaptiques [54].

9. Sauvegarder les poids synaptiques de cette première étape.

10. Faire entrer les deuxièmes exemplaires de chaque phonème avec les premiers en une seule matrice. Il est possible de réduire encore l'erreur seuil pour bien apprendre les exemplaires.
11. Recharger à nouveau les nouveaux poids synaptiques et aller à l'étape 3.
12. Faire toutes les étapes précédentes jusqu'à épuiser la base d'apprentissage.

L'intérêt d'appliquer une telle technique d'apprentissage dit partiel ou par partie, c'est de faciliter le calcul et réduire considérablement le temps du calcul par la sauvegarde des poids synaptique dans chaque étape. Ainsi l'erreur totale peut atteindre rapidement l'erreur seuil. Celle-ci peut être réduite à chaque stage d'apprentissage, ainsi lors de l'apprentissage le système atteint l'erreur seuil progressivement. De plus cette technique nous permet au plus tard l'ajout de nouveaux exemplaires à apprendre par le système, ainsi le système peut être facilement mis à jour.

4.2.4. Phase de reconnaissance par RN

La reconnaissance se fait d'abord par le premier système qui est le TDNN, la sortie de celui-ci passe par un des 14 systèmes MLP suivant le nombre des registres déterminé dans la première phase (fig.4.4).

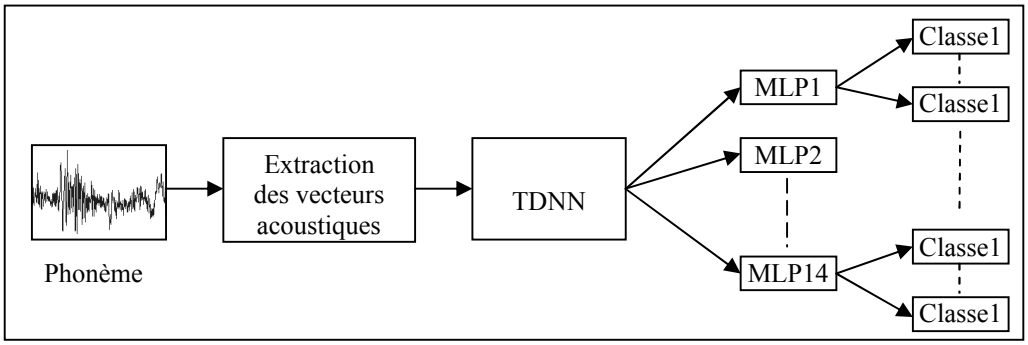


Figure 4.4 : Reconnaissance acoustique des phonèmes

4.2.5. Principe de Rejet/Erreur

Il consiste à exprimer la puissance du système en terme de nombre d'erreurs d'une part, et en terme de nombre de prototypes rejetés en d'autre part. Donc le système fait soit une bonne classification ou il peut commettre des erreurs, mauvaise classification ou bien ni l'une ni l'autre, il fait un rejet du prototype, pas de classification. Ce principe permet dans certains points de fonctionnement de diminuer le nombre d'erreurs au profit du rejet, et donc au profit de la fiabilité du système.

Pour reconnaître un prototype inconnu X, le système calcule sa sortie correspondante Y_j, cette dernière est comparée avec la sortie désirée correspondante pour calculer les erreurs de chaque neurone de sortie.

En effet si S_i est la sortie désirée correspondante à la sortie actuelle Y_i, alors pour chaque neurone de la sortie nous calculons l'erreur *Ei* comme suit :

$$Ei=|Y_i-S_i| \qquad (4.16)$$

Avec $i=1,...,12$.

Alors le prototype *X* est reconnu si :

$$\frac{\xi_h - \xi_k}{\xi_h} \rangle R_a \qquad (4.17)$$

Sinon *X* est rejeté.

Avec R_a un certain coefficient seuil choisi par l'utilisateur caractérisant le système à réseau de neurones. C'est le point de fonctionnement du système.

$$\xi_k = \min_{j=1,...,12}[E_j(X)] \; et \; \xi_h = \min_{\substack{j=1,...,12 \\ j\neq k}}[E_j(X)] \qquad (4.18)$$

ξ_k et ξ_h sont respectivement les plus petites erreurs parmi les 12 [59].

Nous signalons que le taux de reconnaissance se calcule de la manière suivante

$$Taux = 100 \times \frac{Nbr\ de\ phonèmes - (Nbr\ rejet + Nbr\ erreur)}{Nbre\ de\ phonèmes} \qquad (4.19)$$

Nous faisons la remarque que si un phonème est reconnu, il peut être mal classé, c'est là que le système fait une erreur de classification, et il est à réduire au maximum cette erreur [54].

Les résultats de la reconnaissance pour le rejet et l'erreur sont présentés dans le tableau 4.3.

Tableau 4.3 : Résultats erreur/rejet pour la reconnaissance acoustique

Ra	0,001	0,01	0,3	0,6	0,7	0,8	0,9	0,95	0,985	0,99	0,995	0,999
Er	340	335	317	303	294	285	277	265	254	248	238	208
Rj	7	12	36	59	73	90	108	131	160	173	205	276
Taux	83,28	83,28	82,99	82,56	82,32	81,93	81,45	80,92	80,05	79,72	78,66	76,69

La courbe figure montre le comportement du système aux différents points de fonctionnement choisis empiriquement. Il est clair que lorsque l'erreur augmente, le rejet diminue et vice versa (fig. 4.5).

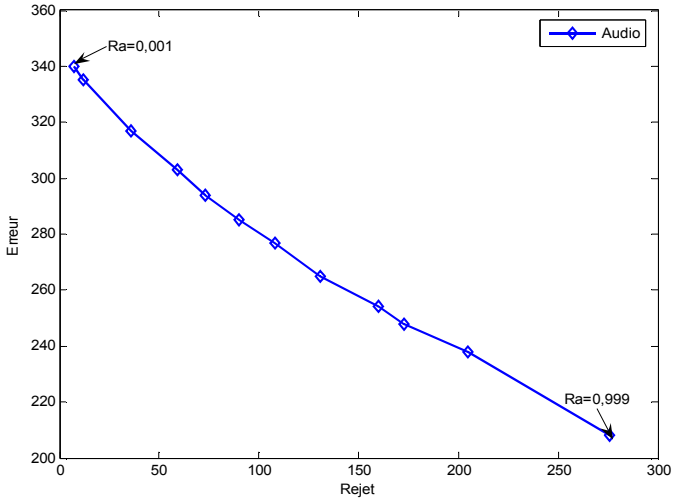


Figure 4.5 : Erreur/rejet pour la reconnaissance acoustique.

Tableau 4.4 : Matrice de confusion pour la reconnaissance acoustique, Ra=0,999

Classes	ب	ف	س	ح	ث	ش	ط	و	ر	ع	ل	ي	Erreur	Rejet	Taux (%)
ب	**153**	0	0	0	0	0	0	1	0	0	0	0	1	19	88.44
ف	0	**161**	1	0	2	0	0	0	0	0	1	0	4	8	93.06
س	0	2	**159**	0	0	0	0	0	0	0	0	0	2	12	91.91
ح	0	2	0	**130**	0	15	0	0	0	0	0	0	15	28	75.14
ث	0	80	0	1	**58**	1	0	0	0	0	0	0	82	33	33.53
ش	0	0	0	14	0	**150**	0	0	0	0	0	0	14	9	86.71
ط	0	1	0	0	0	0	**166**	0	0	0	0	0	1	6	95.95
و	0	0	0	0	0	0	3	**133**	7	0	0	0	10	30	76.88
ر	0	0	0	0	0	0	14	1	**105**	1	13	0	29	39	60.69
ع	0	0	0	3	0	0	4	1	2	**126**	0	0	10	37	72.83
ل	1	0	0	0	0	1	1	14	3	0	**96**	15	35	42	55.49
ي	0	0	0	0	0	0	0	1	0	0	4	**155**	5	13	89.60
													208	**276**	**76 ,96**

La reconnaissance globale montre un taux de reconnaissance de 76,69% de la base de test constituée de 2076 exemplaires. Néanmoins pour le phonème [θ] le taux de reconnaissance est faible de 33,53%, confondu par le phonème [f] de 80 échantillons. En effet ces phonèmes sont acoustiquement semblables (tab.4.4).

4.3. Reconnaissance visuelle de la parole

Dans cette approche, le pixel joue un rôle important pour définir la discrimination entre classes, cette approche est appelée également approche pixel. Dans notre cas, nous avons utilisé deux approches pour extraire les caractéristiques pertinentes des différentes classes. La première consiste à exploiter la valeur couleur des pixels, c'est donc l'approche couleur. Dans cette méthode l'image va subir quelques traitements sur les trois niveaux de couleur RGB (Red, Green, Blue). Le but est de séparer la bouche en quatre zones indépendantes qui sont les lèvres, la langue, l'intérieur de la bouche et enfin les dents et la peau constituent une seule zone.

La deuxième est basée sur le niveau de gris de l'image, c'est l'approche niveau de gris. En premier lieu, les lèvres seules ont été prises comme paramètres, elles ont été séparées du reste de la bouche, donc les lèvres constituent l'objet et le reste est considéré comme le fond de l'image en blanc. En second lieu, toute l'image de la bouche (zone d'intérêt) est prise en niveau de gris comme paramètres pertinentes.

4.3.1. Acquisition et traitements des images

Tout d'abord j'ai teinté mes lèvres en bleu pour bien faire la séparation entre les lèvres et le reste de la bouche et pour faciliter la localisation de la zone d'intérêt (de traitement). De plus nous allons examiner la contribution des lèvres seules à la reconnaissance. Quant à l'acquisition, elle a été faite en prononçant des séquences de types [CV] (Consonnes Voyelle) au moyen d'une camera numérique en raison de 25 frames par seconde. Toutes les consonnes ont été prononcées avec la voyelle courte ([a] ou fetha). Entre deux séquences il y a une pause de quelques secondes pour faciliter la segmentation manuelle. Les trois voyelles courtes ont été prononcées séparément [a], [u] et [i]. Ensuite vient l'étape de segmentation de ces séquences vidéo au moyen d'un logiciel de traitement des vidéos.

Le choix des images prototypes a été fait de la manière suivante (fig.4.6):

Par exemple pour prononcer la séquence **[فَ]** :

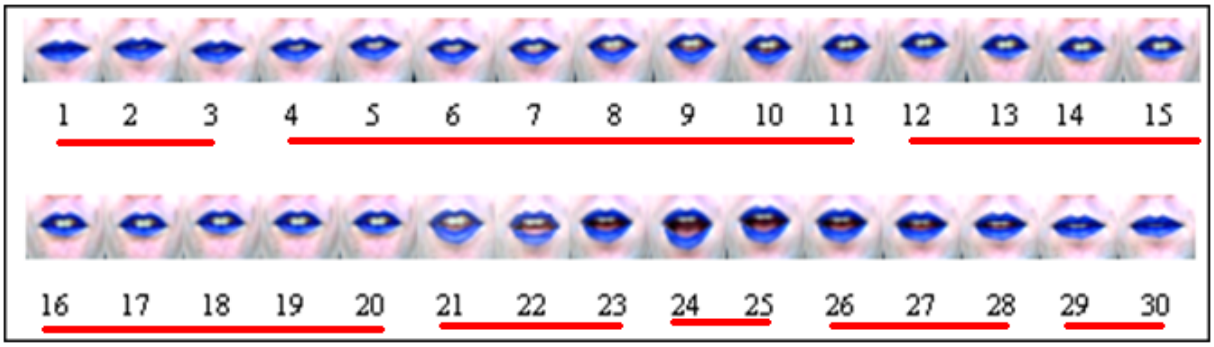


Figure 4.6 : Une séquence de [fa].

- 1,2 et 3 appartiennent au visème « *Silence* ».
- De 4 à 11 ignorées puisqu'elles sont considérées comme des images intermédiaires entre le visème « Silence » et le visème « F ». Alors elles n'appartiennent ni au «Silence » ni au visème « F » ;
- De 12 à 20 appartiennent au visème « F » ;
- 21 et 23 intermédiaires, alors ignorées ;
- 24 et 25 appartiennent au visème « A » ;
- De 26 à 28 intermédiaires, alors ignorées ;
- 29 et 30 appartiennent au visème « Silence ».
- De cette manière une base de données des images a été créée pour tous les autres visèmes.

Les différentes étapes de traitement sont présentées dans l'organigramme suivant (fig.4.7):

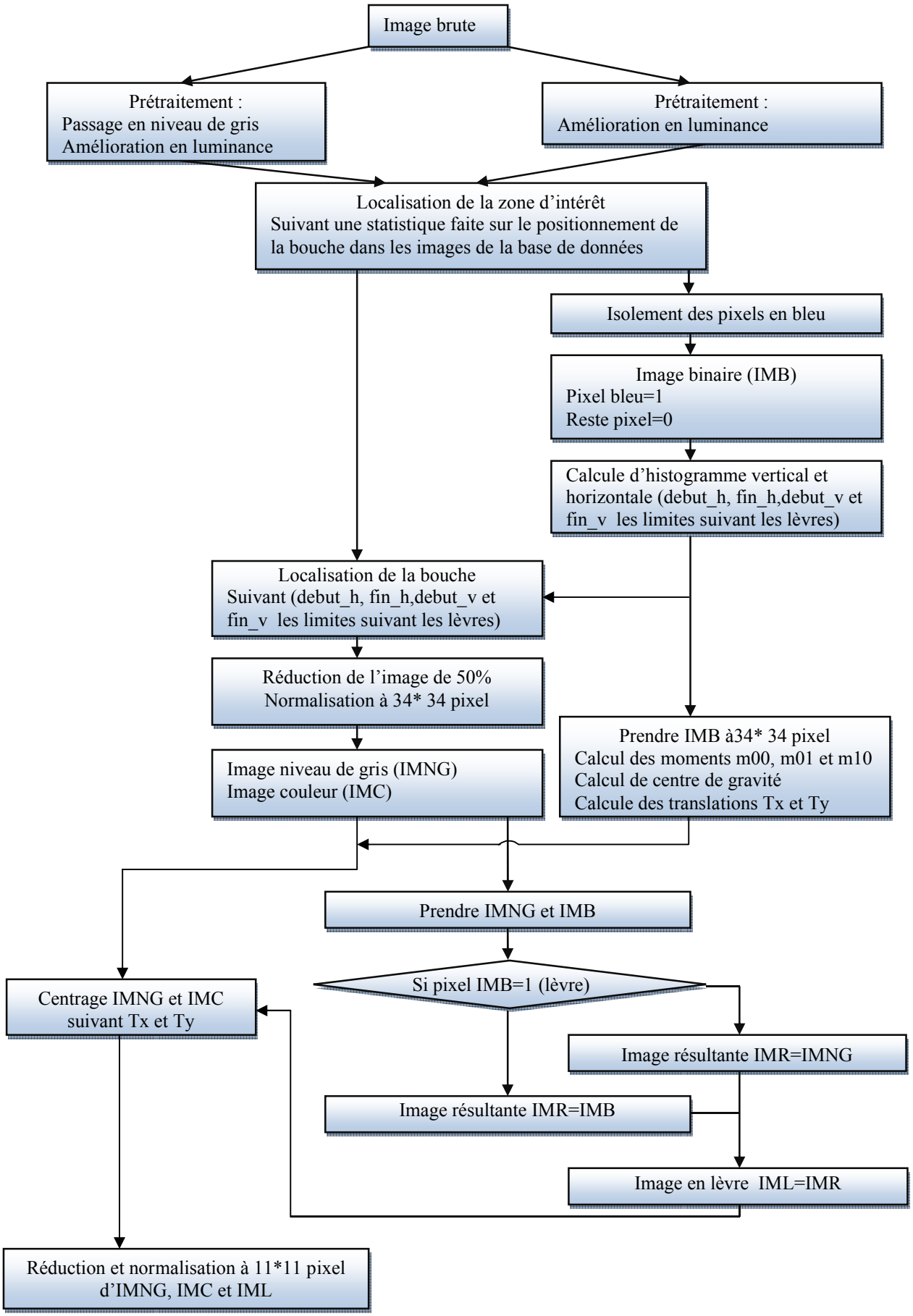


Figure 4.7 : Organigramme de processus de traitement des images

Le processus de traitement pour l'approche couleur est résumé dans la figure 4.8.

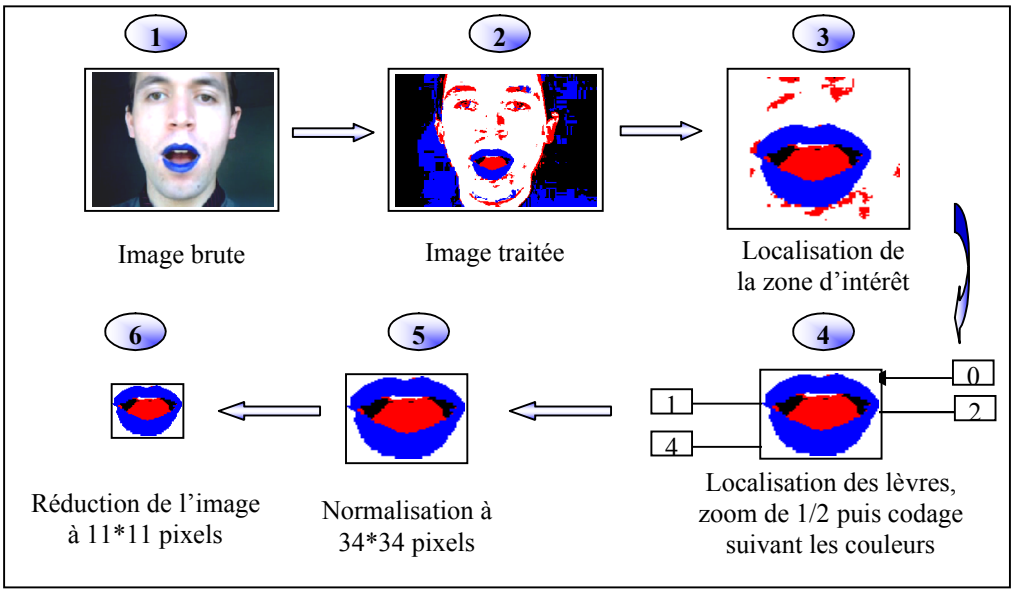


Figure 4.8 : Chaîne de traitements des images

Avec une matrice de 34*34 nous pouvons construire un vecteur de 1156 éléments, ce qui est énorme pour l'apprentissage. Ce qui nous a conduit à réduire les dimensions de l'image à 11*11 soit 121 éléments ce qui allège le calcul (fig.4.8).

4.3.2. Apprentissage par RN de type autoassociateur

Le réseau que nous utilisons est composé de 121 neurones à l'entrée et 121 à la sortie. Quant à la couche cachée, le nombre de neurone a été choisi égal à 50.

Pour chaque classe l'autoassociateur est exercé pour réaliser une fonction propre à la classe. Autrement dit, dans la phase d'apprentissage, le réseau est appelé à calculer les poids synaptiques pour chacune des 11 classes (11 visèmes), donc chaque classe aura sa propre matrice des poids synaptiques.

Soit $X \in R^m$ une entrée quelconque.

La sortie désirée qui correspond à X est X elle-même, pour chaque autoassociateur Ai, le processus d'apprentissage est appelé à minimiser l'erreur quadratique:

$$E_i(x) = \frac{1}{2}\sum_{j=1}^{m}(x_j - o_j^i)^2 \qquad (4.20)$$

Où : O^i_j sont les sorties actuelles du réseau Ai [54].

L'apprentissage a été fait avec 7 exemplaires pour chaque visème en suivant l'algorithme précédemment cité dans la partie acoustique en tenant compte cette fois sur le fait que chaque visème est caractérisé par ces propre poids synaptiques.

4.3.3. Reconnaissance visuelle

La classification d'un visème inconnu est obtenue en comparant les erreurs quadratiques produites de chacun des 11 réseaux A_1, A_2,..., A_{11} qui sont associés aux 11 classes où Ai est principalement caractérisé par sa propre matrice des poids synaptiques (fig.4.9).

Soit l'entrée X, pour chaque autoassociateur A_i on calcule l'erreur quadratique E_i (X), la décision est prise suivant le critère du rejet et de l'erreur. Le motif X est accepté par l'autoassociateur si :

$$\frac{\xi_h - \xi_k}{\xi_h} \rangle R_a \qquad (4.21)$$

Sinon X est rejeté.

$$\xi_k = \min_{j=1,\ldots,12}[E_j(X)] \; et \; \xi_h = \min_{\substack{j=1,\ldots,12 \\ j\neq k}}[E_j(X)] \qquad (4.22)$$

R_a un coefficient seuil choisi [59].

ξ_k et ξ_h sont respectivement les plus petites erreurs parmi les 11 [59].

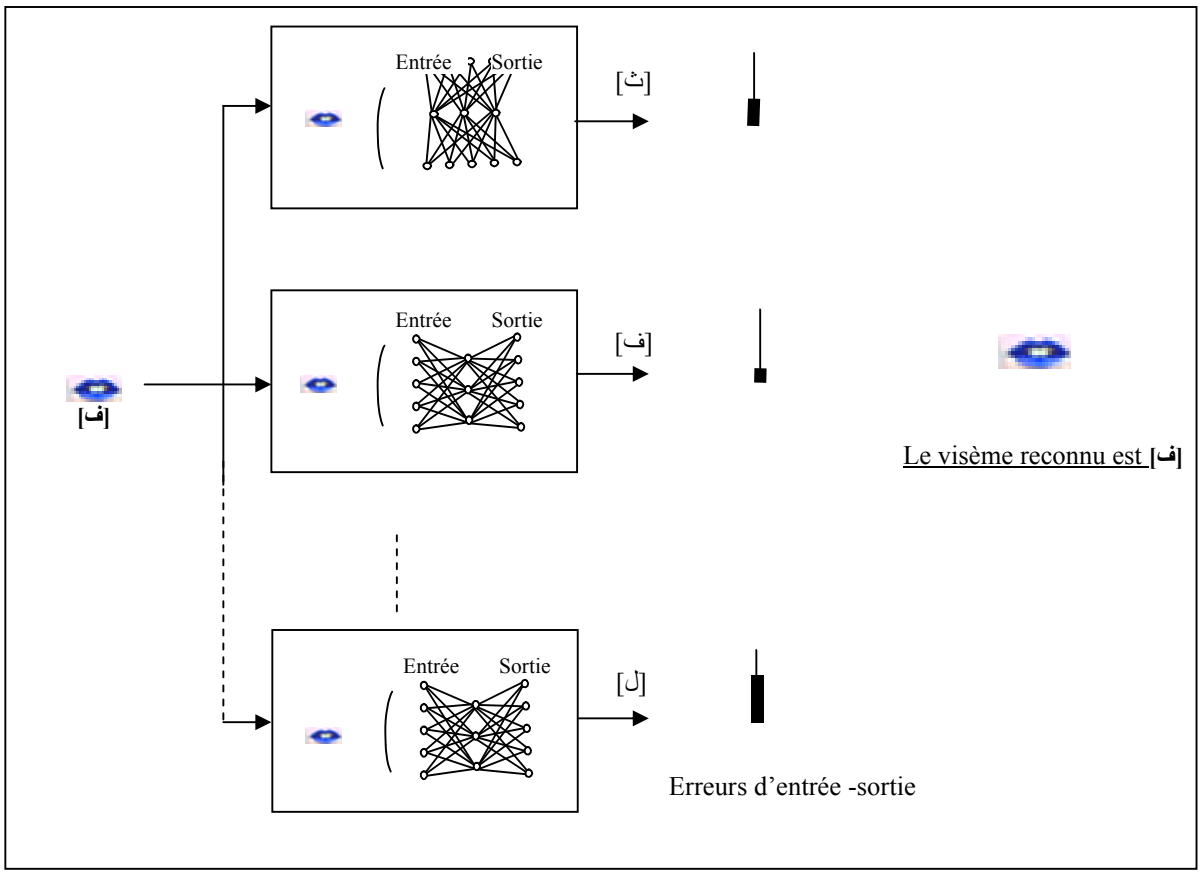


Figure 4.9 : Principe de reconnaissance par l'autoassociateur [59]

Tableau 4.5 : Matrice de confusion, approche couleur, pour Ra=0,001.

Classes	أ	ب	و	ي	د	ث	ش	ز	ل	ف	ر	Erreur	Rejet	Taux (%)
أ	**399**	1	3	0	77	1	0	0	137	0	84	301	1	**56,92**
ب	0	**988**	2	0	11	0	0	0	0	30	0	43	0	**95,83**
و	4	0	**348**	0	0	0	0	0	2	0	30	36	0	**90,63**
ي	31	3	0	**289**	102	0	0	0	5	4	1	146	0	**66,44**
د	16	1	0	26	**386**	0	0	0	2	1	2	48	0	**88,94**
ث	2	0	1	0	11	**280**	19	0	18	0	163	214	0	**56,68**
ش	2	0	0	0	6	27	**225**	0	12	2	44	93	0	**70,75**
ز	1	0	1	169	50	0	1	**219**	0	0	0	222	0	**49,66**
ل	115	0	0	3	31	5	1	0	**125**	0	1	154	1	**44,64**
ف	0	1	0	0	5	4	0	0	0	**179**	1	11	1	**93,72**
ر	6	0	46	0	1	0	0	0	2	0	**93**	55	0	**62,83**
												1323	**3**	**72,68**

Pour une erreur seuil de 0,001 nous avons obtenu à un taux de reconnaissance de 72,68%, c'est pas mal mais la matrice de confusion montre que le taux de reconnaissance pour le visème **ل**, **ز** ou encore **ث** est trop faible par rapport au reste des visèmes. En effet, le visème **ز** ressemble le plus souvent au visème **ي**, il en est en moins pour le visème **د.** De même le visème **ي** s'est confondu au visème **د** par le système. Idem pour le visème **ث** confondu au **ر** (tab.4.5). Le résultat de prétraitement de l'image a créé ce genre de similarité entre ces visèmes, notamment pour celles qui ont été mal classées au départ et plus précisément les images dites intermédiaires.

Ce faible taux de reconnaissance pour certaines visèmes nous a poussés à examiner une autre approche, c'est l'approche niveaux de gris.

4.3.4. Image prise en niveaux de gris

- Approche lèvres uniquement en niveau de gris :

  Cette fois-ci nous avons pris uniquement les lèvres en niveaux de gris, c'est l'approche lèvres (fig.4.10).

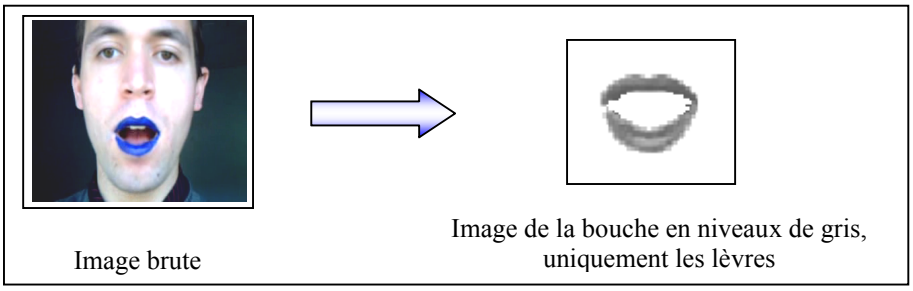


Figure 4.10 : Isolement des lèvres en niveau de gris

Tableau 4.6 : Matrice de confusion, approche lèvres seules, pour Ra=0,001

Classes	أ	ب	و	ي	د	ث	ش	ز	ل	ف	ر	Erreur	Rejet	Taux (%)
أ	**483**	7	56	15	32	8	15	0	0	9	75	216	1	**69,00**
ب	0	**992**	1	3	0	0	0	0	0	35	0	39	0	**96,22**
و	3	12	**357**	0	0	0	0	0	0	0	11	26	0	**93,21**
ي	2	7	3	**354**	37	12	1	7	2	10	0	81	0	**81,38**
د	0	1	0	108	**266**	3	0	47	2	5	0	166	2	**61,29**
ث	2	0	5	33	6	**413**	37	0	0	0	0	81	0	**83,60**
ش	1	0	14	24	3	29	**241**	0	0	3	0	77	0	**75,79**
ز	0	0	1	61	30	0	0	**338**	0	11	0	103	0	**76,64**
ل	0	0	0	13	39	93	41	0	**92**	0	0	186	0	**33,09**
ف	0	3	0	0	0	0	0	0	0	**187**	1	4	0	**97,91**
ر	2	0	65	0	0	0	0	0	0	0	**81**	67	0	**54,73**
												1046	**3**	**78,38**

Le taux globale est nettement amélioré de 72,68% à 78,38% soit de 5,7% ce qui est appréciable de point de vue reconnaissance. Cependant, ce même taux est trop faible pour le visème ل qui s'est confondu fortement avec le visème ث, de moins avec le visème د et ش. Quant au visème ز, une nette amélioration a été réalisée notamment avec la confusion avec le visème ي qui a diminué de 169 dans le premier test à 61 dans le deuxième test. Le visème ر a enregistré une baisse en taux de reconnaissance de 10% et la confusion reste toujours forte avec le visème ي (tab.4.6).

- Approche toute la bouche en niveau de gris :

  Ici au lieu de prendre uniquement les lèvres, on prend de plus les dents, la langue et l'intérieur de la bouche tous en niveaux de gris.

  Dans ce cas les mêmes étapes que l'approche couleur sauf qu'au lieu de prendre l'image en couleur traitée, on prend l'image initiale puis on la transforme en niveaux de gris (fig.4.11).

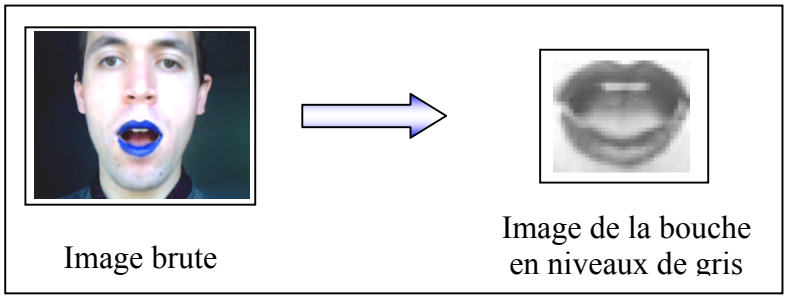


Figure 4.11 : Passage de couleur en niveaux de gris

Tableau 4.7 : Matrice de confusion, approche toute la bouche, pour Ra=0,001

Classes	أ	ب	و	ي	د	ث	ش	ز	ل	ف	ر	Erreur	Rejet	Taux (%)
أ	**585**	10	6	8	22	0	3	0	10	6	51	116	1	**83,29**
ب	0	**1011**	2	0	2	0	0	0	0	20	0	24	0	**97,67**
و	11	0	**360**	0	0	0	0	0	0	0	11	22	1	**93,99**
ي	3	5	0	**275**	92	2	0	40	10	7	0	159	1	**63,22**
د	0	1	0	17	**407**	3	0	3	2	0	0	26	1	**93,78**
ث	4	0	4	0	37	**417**	25	0	7	0	0	77	0	**84,41**
ش	2	0	0	0	37	8	**268**	0	0	2	0	49	1	**84,28**
ز	0	0	1	17	105	0	0	**318**	0	0	0	123	0	**72,11**
ل	0	0	0	0	57	29	0	0	**192**	0	0	86	0	**69,06**
ف	1	4	0	0	0	0	0	0	0	**184**	0	7	0	**96,34**
ر	5	0	44	0	0	32	0	0	**0**	0	**104**	49	0	**66,89**
												738	**5**	**84,69**

Ce tableau montre bien que le taux de la reconnaissance globale a été nettement amélioré par rapport aux deux tests précédents. Néanmoins pour les visèmes (ي, ز et ف) ce taux est légèrement diminué. La grande amélioration a été réalisée pour les visèmes (ل qui a passé de 33,09% à 69,06%, د de 61,29% à 93,78% et ر de 54,73% à 66,89%) (tab.4.7).

La conclusion que nous pouvons tirer de ces résultats c'est que toute la bouche (lèvres, dents, langue et intérieur de la bouche) contribue à l'amélioration de la reconnaissance et donc la discrimination entre visèmes. D'autre part, la prise des images en niveaux de gris permet de réduire la possibilité de détruire les caractéristiques pertinentes de chaque visèmes, comme c'était le cas pour le premier test.

Ce résultat positif nous a encouragés à faire d'autres tests pour d'autres erreurs seuils *Ra.* Le tableau 4.8 résume les résultats obtenus. La courbe erreur (rejet) est présentée dans a figure 4.12.

Tableau 4.8 : Résultats erreur/rejet pour la reconnaissance visuelle par autoassociateur

Ra	0,001	0,05	0,1	0,15	0,2	0,3
Erreur	738	639	547	449	357	230
Rejet	5	161	346	540	738	1108
Taux	84,69	83,52	81,60	79,62	77,44	72,43

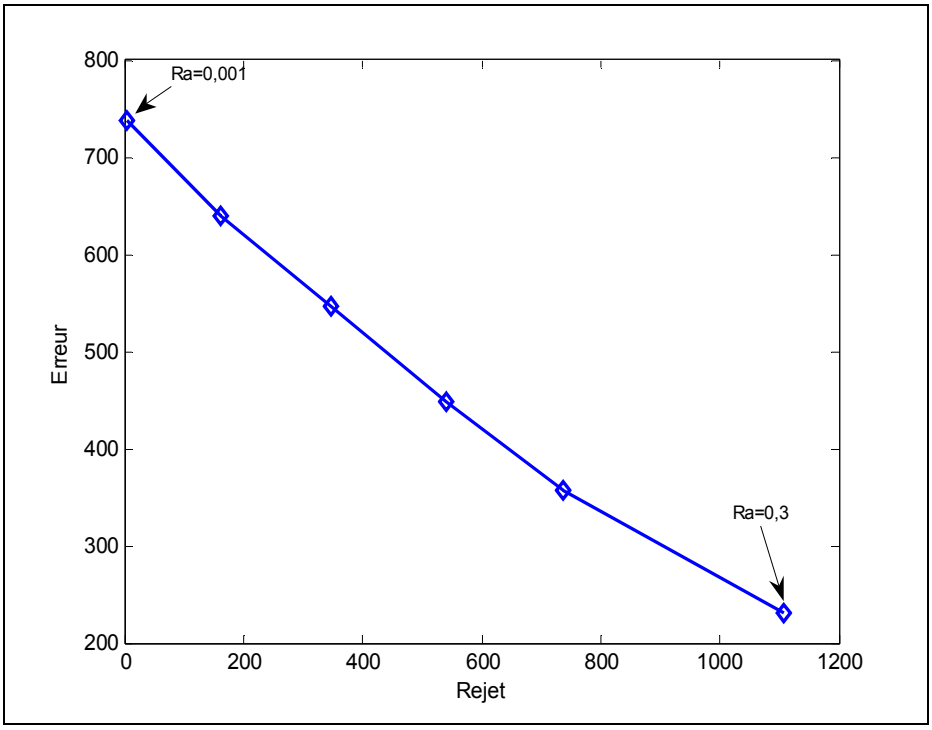


Figure 4.12 : Erreur/rejet pour la reconnaissance visuelle par l'Autoassociateur

Un autre test pour la même base de données (image en niveaux de gris) a été fait par un autre type de RN qui est le MLP, avec 121 neurones pour la couche d'entrée, 50 neurones pour la couche cachée et 11 neurones pour la couche de sortie représentant les 11 visèmes. Les visèmes sont codés suivant le tableau 4.9.

Tableau 4.9 : Codage des visèmes pour l'apprentissage du MLP

Neurone de sortie / Classes	1	2	3	4	5	6	7	8	9	10	11
أ	**1**	0	0	0	0	0	0	0	0	0	0
ب	0	**1**	0	0	0	0	0	0	0	0	0
و	0	0	**1**	0	0	0	0	0	0	0	0
ي	0	0	0	**1**	0	0	0	0	0	0	0
د	0	0	0	0	**1**	0	0	0	0	0	0
ث	0	0	0	0	0	**1**	0	0	0	0	0
ش	0	0	0	0	0	0	**1**	0	0	0	0
ز	0	0	0	0	0	0	0	**1**	0	0	0
ل	0	0	0	0	0	0	0	0	**1**	0	0
ف	0	0	0	0	0	0	0	0	0	**1**	0
ر	0	0	0	0	0	0	0	0	0	0	**1**

Tableau 4.10 : Matrice de confusion pour la reconnaissance visuelle par MLP, Ra=0,001

Classes	أ	ب	و	ي	د	ث	ش	ز	ل	ف	ر	Erreur	Rejet	Taux (%)
أ	**572**	26	19	23	5	0	20	0	2	2	25	122	6	**81,71**
ب	0	**1014**	3	0	0	0	0	0	0	13	0	16	1	**98,35**
و	2	3	**371**	0	0	0	2	0	0	0	5	12	0	**96,87**
ي	7	7	1	**292**	25	0	7	35	1	33	0	116	27	**67,13**
د	1	1	0	1	**346**	9	7	27	4	35	0	85	3	**79,72**
ث	1	0	2	0	2	**412**	22	3	12	23	0	65	17	**83,40**
ش	0	0	0	0	0	16	**288**	1	0	9	0	26	4	**90,57**
ز	0	0	1	4	55	0	1	**374**	2	2	0	65	2	**84,81**
ل	0	0	0	0	24	30	23	0	**185**	0	0	77	16	**66,55**
ف	0	2	0	0	0	0	0	0	0	**181**	8	10	0	**94,76**
ر	4	6	48	0	0	0	0	0	0	0	**89**	58	1	**60,14**
												652	**77**	**84,98**

Ces résultats montrent que pour certaines visèmes le taux de reconnaissance a été amélioré, et pour d'autre, il est légèrement diminué. Cependant, dans l'ensemble le taux global de la reconnaissance est légèrement amélioré passant de 84,69% à 84,98% pour *Ra*=0,001 (tab.4.10).

D'autres tests pour d'autres erreurs seuils *Ra* dans le tableau 4.11. Les valeurs de l'Erreur (Rejet) sont la somme des erreurs (rejets) de tous les visèmes, fig.4.13.

Tableau 4.11: Résultats erreur/rejet pour la reconnaissance visuelle par MLP

Ra	0,001	0,01	0,1	0,4	0,7	0,9
Erreur	652	576	440	303	207	123
Rejet	77	195	446	736	1010	1293
Taux	84,98	84,11	81,74	78,59	74,92	70,82

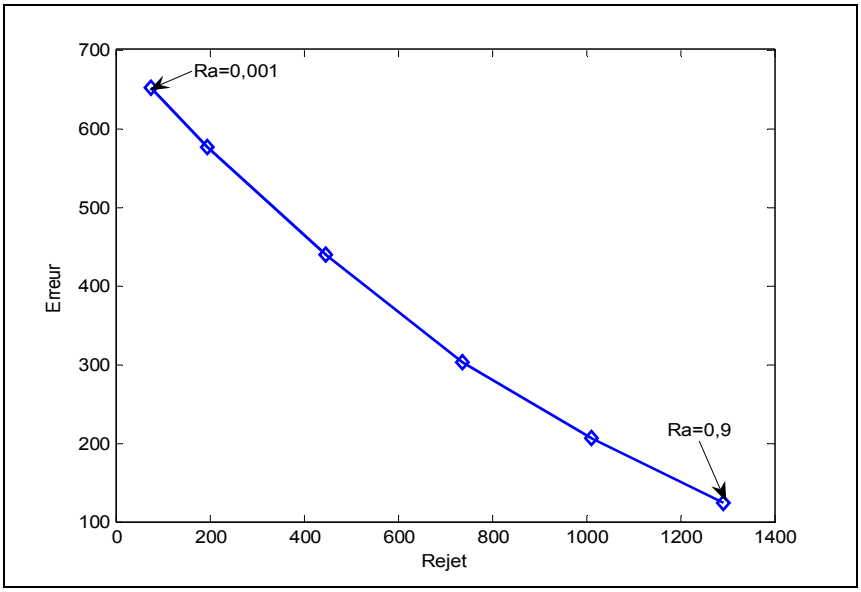


Figure 4.13 : Erreur/rejet pour la reconnaissance visuelle par MLP

La figure 4.14 présente une comparaison des résultats obtenus pour l'autoassociateur et le MLP.

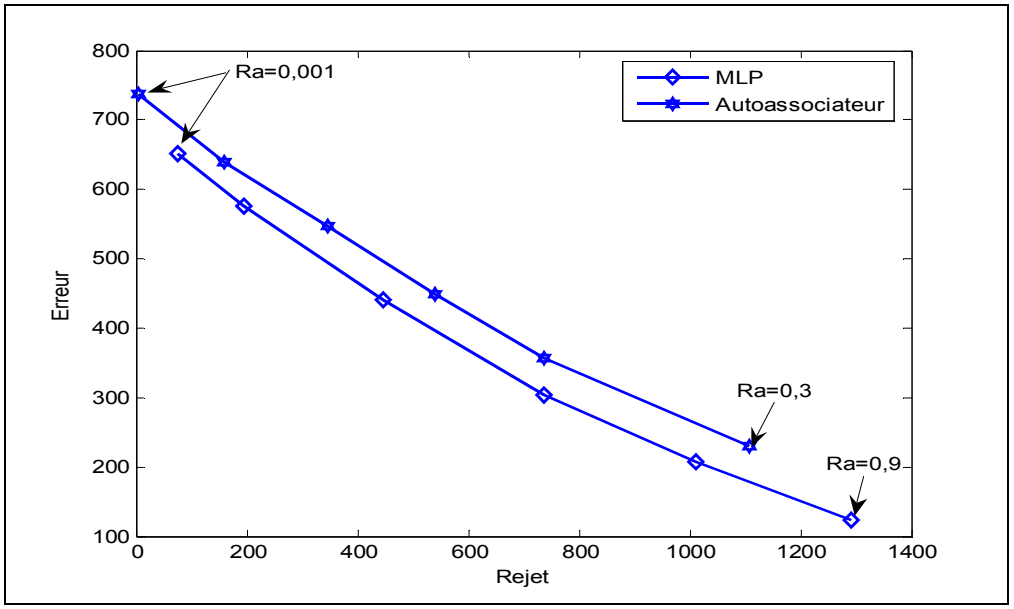


Figure 4.14 : Erreur/rejet pour la reconnaissance visuelle par le MLP et l'Autoassociateur

Plus la courbure tend vers le point (0,0), meilleur est le système. C'est ce que nous remarquons sur cette figure où la courbe Erreur/Rejet pour le système MLP est en dessous de celle de l'autoassociateur, ce qui nous amène à conclure que le MLP est meilleur que l'autoassociateur. De plus, le nombre de neurones de la couche cachée pour le MLP est 20, et celui de la couche de sortie est 11 permet de réaliser un temps de calcul fortement réduit par rapport à l'autoassociateur qui compte 50 neurones dans la couche cachée et 121 neurones dans la couche de sortie.

Le gain dans le temps de calcul est réalisé donc pour l'apprentissage et la reconnaissance. De plus, le fait que le taux global de la reconnaissance est légèrement amélioré, il nous a permis de conclure que les lèvres contient l'information la plus importante et la plus pertinente ; le reste de la bouche (dents, langue, intérieur de la bouche) constitue une information complémentaire à celle portée par les lèvres, ce qui est claire puisqu'une bonne articulation des visèmes dépend fortement des lèvres.

4.4. Reconnaissance audiovisuelle de la parole

Le problème qui se pose dans cette partie importante dans notre travail c'est la manière d'intégrer les paramètres acoustiques avec les paramètres visuels et la décision acoustique avec la décision visuelle. Notre choix s'est porté sur l'ID où nous avons développé une méthode créative et efficace.

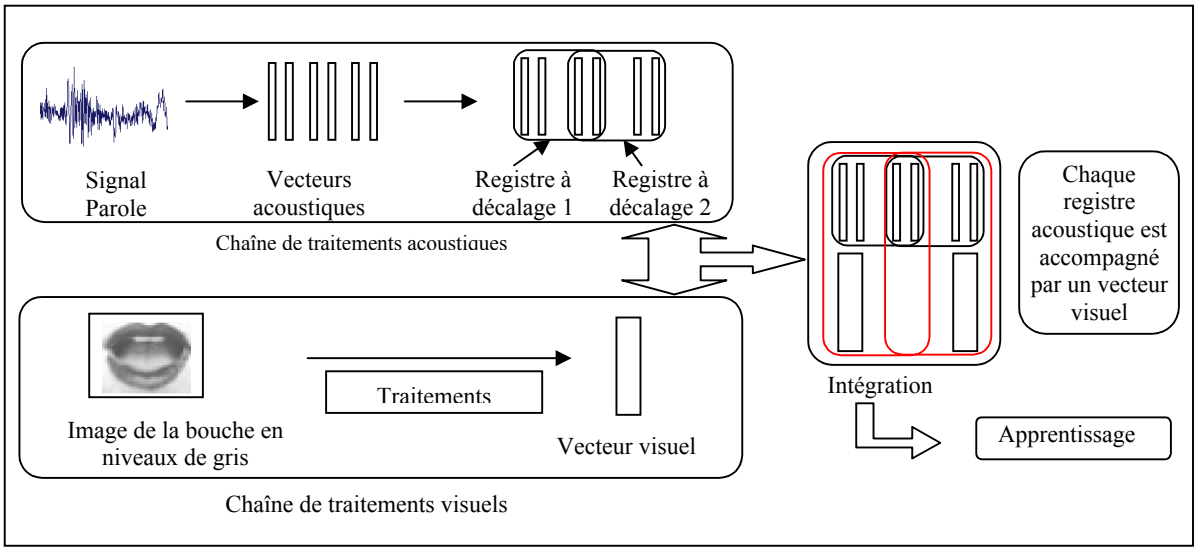


Figure 4.15 : Chaîne de traitements pour l'Intégration Directe ID

Dans cette méthode les paramètres acoustiques et visuels ont été intégrés dans un seul vecteur paramètre. Les paramètres visuels constituent un seul vecteur de 121 éléments, alors que la partie acoustique est une suite de registres à décalage de 48 éléments. La technique consiste à faire concaténer chaque registre à un vecteur visuel en constituant ainsi un vecteur résultant de 169 éléments. Ainsi le nombre des vecteurs résultants est le même que le nombre des registres à décalage déterminés dans la partie acoustique. Par cette configuration, à chaque instant de signal acoustique est accompagné un signal visuel du visème correspondant.

De la même manière que la partie acoustique, l'apprentissage se fait en partie, la partie TDNN ensuite la partie MLP. Pour la partie TDNN, le nombre de neurones de la couche d'entrée est donc de 169 neurones, celui de la couche cachée est choisi égale à 50 et pour la couche de sortie est bien évidement est égale à 12. L'apprentissage a été fait sur base de sept exemples pour chaque phonème\visème.

Concernant la partie MLP nous avons opté à utiliser les poids synaptiques calculés dans la phase acoustique puisqu'il s'agit uniquement d'intégrer les résultats du TDNN composés de plusieurs vecteurs de 12 éléments chacun (fig. 4.15).

Les résultats de la reconnaissance audiovisuelle pour l'intégration ID sont présentés dans le tableau suivant :

Tableau 4.12 : Résultats erreur/rejet pour la reconnaissance audiovisuelle

Ra	**0,001**	**0,01**	**0,3**	**0,6**	**0,8**	**0,985**	**0,999**
Erreur	203	197	180	163	150	137	107
Rejet	12	22	53	69	87	125	187
Taux	89,64	89,45	88,77	88,14	87,88	87,37	85,83

La figure 4.16 montre l'évolution du taux de reconnaissance pour différents points de fonctionnement du système (*Ra*).

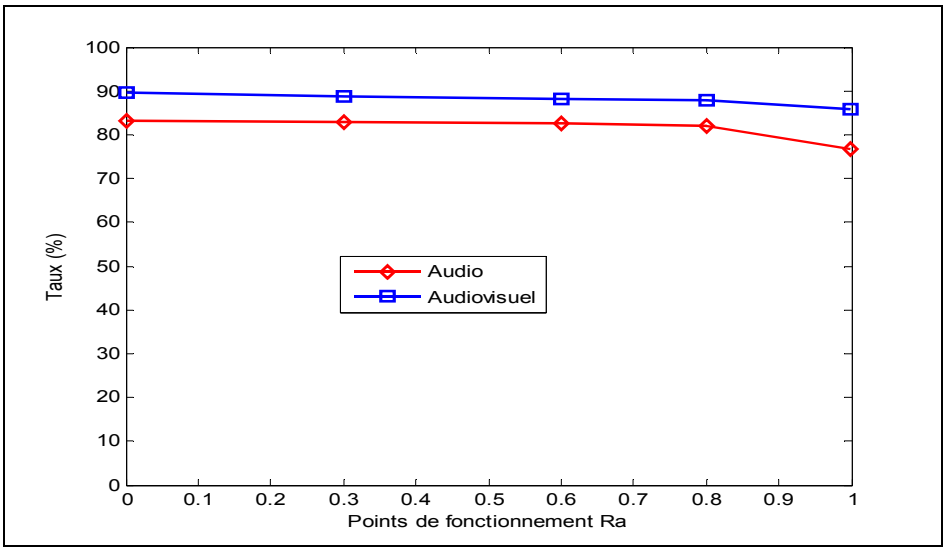


Figure 4.16 : Taux de reconnaissance en fonction du seuil *Ra*

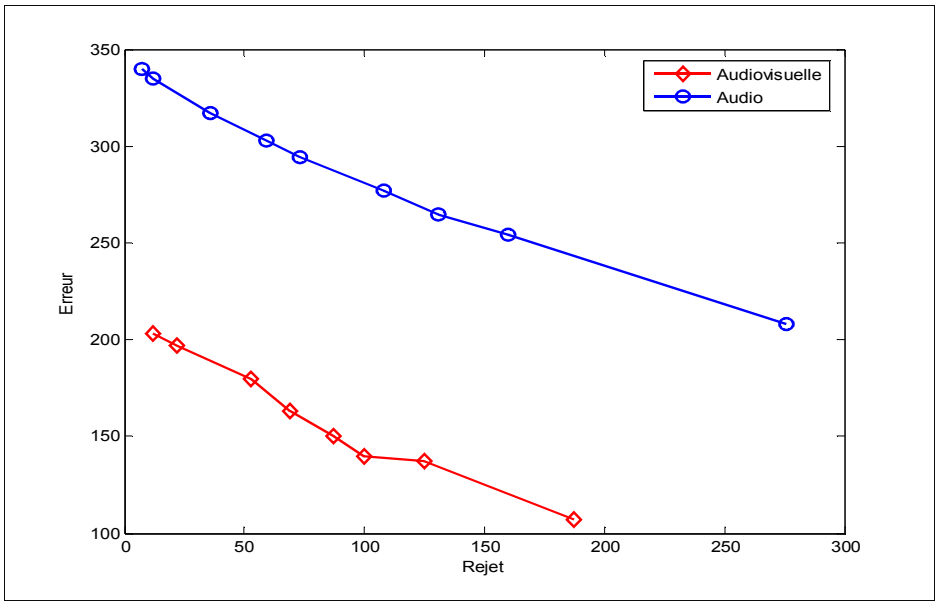


Figure 4.17 : Erreur/rejet pour la reconnaissance audiovisuelle

Il est clair dans cette figure que la courbe en rouge qui represente la reconnaissance audiovisuelle présente une courbure plus proche au point (0,0) qui est le point idéal du fonctionnement (fig.4.17). Ce qui preuve que l'aspect visuel de la parole non seulement contribue à bien faire la distinction entre les phonèmes, mais il présente une source d'information importante et fait ainsi une partie intégrante de la parole, c'est bien donc la bimodalité de la parole.

Tableau 4.13 : Matrice de confusion pour la reconnaissance audiovisuelle, Ra=0,999

Classes	ب	ف	س	ح	ث	ش	ط	و	ر	ع	ل	ي	Erreur	Rejet	Taux (%)
ب	**172**	0	0	0	0	0	0	0	0	0	0	0	0	1	99,42
ف	0	**159**	0	0	2	0	0	0	0	0	0	0	2	12	91,90
س	0	0	**160**	0	0	0	0	0	0	0	0	0	0	13	92,48
ح	0	3	0	**144**	0	7	0	0	0	0	0	0	10	19	83,23
ث	0	0	1	11	**75**	32	0	0	0	0	0	0	44	54	43,35
ش	0	0	0	0	0	**173**	0	0	0	0	0	0	0	0	100,00
ط	0	0	0	0	0	0	**169**	0	0	0	0	0	0	4	97,68
و	0	0	0	0	0	0	2	**157**	1	0	0	0	3	13	90,75
ر	0	0	0	0	0	0	1	0	**130**	0	21	0	22	21	75,14
ع	0	0	0	2	0	0	0	0	0	**142**	0	0	2	29	82,08
ل	1	0	0	0	0	0	0	18	2	0	**129**	2	23	18	76,30
ي	0	0	0	0	0	0	0	0	0	0	1	**169**	1	3	97,68
													107	187	85,83

Outre le taux de reconnaissance que nous avons réalisé (passant de 76,69% en acoustique seul à 85,83% en audiovisuel), nous pouvons dire que les RN de type TDNN n'ont pas seulement un rôle de bon classifieur, donc un bon moyen d'extraire les caractéristiques pertinentes, mais aussi ils peuvent être considérés comme bon réducteur de quantité de données. En effet, la souplesse de la gestion des données dynamiques dans le temps permet d'exploiter l'ensemble des échantillons de la forme dans différents points en même temps. Ce qui est bien clair lorsque nous avons passé d'une représentation d'un phonème donné de 6 vecteurs acoustiques de 12 coefficients chacun à une représentation de uniquement 2 vecteurs de 12 coefficients chacun, puisque d'après l'algorithme utilisé, 6 vecteurs acoustiques peuvent construire 2 registres à décalages (de 72 coefficients (6x12) à 24 coefficients (2*12) soit 3 fois moins). Ainsi si une information est commune entre deux formes différentes, elle sera bien classée lorsqu'elle est traitée avec d'autres informations de la même forme que si elle est traitée seule.

4.5. Conclusion

Dans ce chapitre nous avons présenté notre expérience de concevoir un système de reconnaissance audiovisuelle de la parole en commençant par la reconnaissance uniquement acoustique de la parole tout en décrivant la base de donnée utilisée ainsi que les paramètres acoustiques pertinents, sans oublier le système à base de RN de type TDNN. Arrivant à ce stade là, nous avons traité aussi la partie visuelle de système de reconnaissance où nous avons présenté tous les visèmes de la langue arabe. Pour cette partie nous avons essayé deux types de RN qui sont le MLP et l'autoassociateur. Nous avons montré expérimentalement que les RNs de type MLP donne de bon résultat par rapport à l'autoassociateur

Une fois les étapes ci-dessus franchi, nous avons passé à l'intégration directe des deux modalités de la parole dans un seul système de reconnaissance en montrant l'amélioration du taux de reconnaissance par rapport à l'acoustique seul ou le visuel seul. La technique utilisée pour l'intégration directe a été bien décrite dans cette partie de chapitre.

CONCLUSIONS GENERALES ET PERSPECTIVES

1. Conclusions

Les travaux menés depuis quelques années ont montré que les modèles neuronaux présentent un intérêt certain pour le TAP, non seulement en reconnaissance de la parole, mais aussi en synthèse à partir du texte, en vérification du locuteur ou en acquisition du langage, essentiellement pour leur capacité d'apprentissage discriminant.

Ces modèles ont d'abord été utilisés dans des reconnaisseurs statiques et la façon dont ils prennent en compte la structure temporelle de la parole n'est pas encore totalement satisfaisante, malgré les diverses solutions proposées. De nombreuses recherches sont en cours sur ce sujet, y compris en ce qui concerne la conception de nouveaux modèles inspirés de la réalité neurobiologique et utilisant les données de la perception auditive. D'autres voies de recherche sont également prometteuses, notamment dans la mise en œuvre de modèles hybrides stochastiques connexionnistes efficaces.

Il est à signaler que le développement d'un système à base de RN est une tâche délicate et qui nécessite beaucoup d'expérience. De nombreux problèmes se posent en effet concernant le choix et le dimensionnement du réseau, les paramètres à ajuster, le contrôle du système, etc., même avec les outils logiciels de développement maintenant disponibles.

Dans notre cas l'utilisation du réseau de type TDNN a permis de résoudre le problème de fait que la parole est un signal dynamique et dont il faut l'exploiter tel qu'il est pour donner plus de signification à l'ensemble des informations extraites. D'autre part, l'utilisation de plusieurs systèmes de type MLP nous a permis de résoudre le problème de la taille non identique pour tous les phonèmes, puisque la variabilité inter et intra locuteur permet de générer ce genre de problème. L'avantage du couplage des deux types de réseau de neurones est que leur apprentissage se fait d'une manière indépendante, d'où un gain important de point de vue encombrement d'informations lors de la mise à jour des poids synaptiques. Il s'agit dans ce cas d'une modélisation modulaire, module TDNN et module MLP complémentaires en eux. Le principal rôle du module TDNN est bien d'assurer le groupement d'informations pertinentes pour un même phonème en associant l'ancienne information à *(t-1)*, par exemple, avec la nouvelle à *(t),* et donc plus de discrimination entre phonèmes. Quant au module MLP, il assure, outre l'intégration des informations issues du premier module, l'aiguillage entre les 14 sous-systèmes MLP suivant la taille du phonème.

L'utilisation des coefficients MFCC pour la modélisation des phonèmes n'est pas arbitraire. L'extraction de ces coefficients se base sur la perception du signal acoustique l'oreille humain. D'autre part, dans plusieurs travaux de recherche le choix s'est porté sur ces coefficients et par suite ont montré leur capacité à la discrimination et à la bonne représentation de l'information parole et par suit ils permettent d'obtenir un taux de reconnaissance élevé.

En ce qui concerne la partie visuelle, nous avons choisi d'utiliser encore une fois les réseaux de neurones. Les résultats obtenus ont permis de réaliser que les réseaux de neurones de type MLP sont meilleurs que les réseaux de neurones de type autoassociateur pour l'approche image que nous avons adoptée. Le choix de l'approche image vient du fait que l'image peut contenir plus d'informations que l'approche contour ou géométrique. Exploiter plus d'information c'est obtenir plus de discrimination. De plus, l'information contour et les informations géométriques seront inclues dans toute l'image, d'où les visèmes seront bien représentés par l'approche image que par d'autre approches.

La réalisation d'un système de reconnaissance audiovisuelle de la parole en utilisant une intégration directe nous a permis de confirmer le fait que l'aspect de la parole est bimodal et que la modalité visuelle joue le rôle du complément de la modalité acoustique, et cela même en l'absence du bruit et dans un environnement calme.

2. Perspectives

Parler et dialoguer avec une machine sont des défis majeurs qui ne sont pas encore complètement résolus. Cependant, l'utilisation des interfaces en parole naturelles est extrêmement séduisante et devrait changer profondément la nature des interactions dans la communication Homme-Machine. Ces interfaces sont d'ores et déjà sollicitées pour l'enseignement assisté par ordinateur et pour le multimédia. Les technologies qui conditionnent la réalisation de telles applications sont regroupées autour du codage audio, de la synthèse et de la RAP, du dialogue, ainsi que de la compréhension du langage parlé.

Afin de rendre les interfaces en parole naturelle plus fiables, une solution est d'augmenter les modalités pouvant être perçues par la machine qui sont la modalité acoustique et visuelle de la parole. L'objectif principal de notre travail, comme nous l'avons déjà signalé, c'est de réaliser de tel système de dialogue HM en AS.

L'approche que nous proposons pour de futurs travaux c'est de réaliser une tête parlante assurant l'interfaçage entre l'homme et la machine représentée par un ordinateur. L'utilisation de ce système sera principalement focalisé vers le e-Learning pour

l’apprentissage correcte et efficace de la langue arabe à travers la bonne articulation des différents phonèmes par l’écoute et la vue des différents articulation des visèmes ainsi que les différents gestes et mimiques que peut produire une personne lorsqu’elle parle, et éventuellement une application de la lecture labiale ou une personne malentendante peut comprendre un texte introduit par clavier ou un extrait audio enregistré de telle sorte que le système à travers une tête parlante traduit le texte ou le fichier audio en langage labiale, donc il s’agit de la synthèse visuelle de la parole à partir du texte et des fichiers audio. La réalisation de tel système de dialogue fait appel à la modélisation en 3D du visage en le décomposant en différentes parties, et cela pour le pouvoir animer facilement et indépendamment, c’est un domaine qui fait appel aux techniques de l’animation faciale [16].

REFERENCES BIBLIOGRAPHIQUES

1. http://www.lris.fr/dep-inf

2. L. Buniet, ‘‘Traitement automatique de la parole en milieu bruité : étude de modèles connexionnistes statiques et dynamiques’’, Thèse de Doctorat de l’Université Henri Poincaré - Nancy 1, France, Février 1997.

3. O. Galibert, G. Illouz and S. Rosset ‘‘Dialogue Homme-Machine à domaine ouvert’’ LIMSI – CNRS, TALN Dourdan, France, 6-10 juin 2005.

4. P. Daibias, ‘‘Modèle a posteriori de la forme et de l’apparence des lèvres pour la reconnaissance automatique de la parole audiovisuelle’’, Thèse de Doctorat, Spécialité Informatique, Université du Maine, France, Décembre 2002.

5. S. Ouni, M.M. Cohen and D.W. Massaro, ‘‘Training Baldi to be multilingual: A case study for an Arabic Badr’’, Journal of Speech communication, vol.45, issue 2, pp.115-137. 2005.

6. http://users.info.unicaen.fr/~giguet/java/textes/formes.html

7. Collection Microsoft ® Encarta ® 2005

8. http://www.tsi.enst.fr\~cfaure\intro\intro2.html

9. http://ltswww.epfl.ch/~coursrf/intro.pdf

10. A. Deneche, S. Meshoul et M. Batouche, “Une approche hybride pour la reconnaissance des formes en utilisant un système immunitaire artificiel”. Proc. graphic computer science, Biskra, Algeria, 2005.

11. http://www.univ-ag.fr/grimaag/ bios/jdesachy/Publis/RF-encours.pdf

12. http://www.hds.utc.fr/sy19/documents/cours/non_param.pdf

13. http://www.univ-rouen.fr/psi/heutte/rdf/ClassifStatistique.pdf

14. M. Kunt, G.Coray, G.H. Granlund and J-P. Haton,"Reconnaissance des formes et analyse de scènes",Publié par PPUR presses polytechniques, 2000.

15. G. Dreyfus, J.-M. Martinez, M. Samuelides, M.B. Gordon, F.Badran, S. Thiria and L. Hérault, ‘‘Réseaux de neurones, Méthodes et applications’’, Edition EYROLLES, Deuxième tirage 2002.

16. E. Cosatto, ‘‘Sample-Based Talking-Head Synthesis’’, PhD Thesis, Swiss Federal Institute of Technology, Switzerland, October 2002.

17. E. Poisson and C. Viard-Gaudin, ‘‘Réseaux de neurones à convolution, reconnaissance de l’écriture manuscrite non contrainte’’, *Valgo* (ISSN 1625-9661), n° 01-02, Oct. 2001.

18. http://pst.chez-alice.fr\svtiufm\jeparlejechante.html

19. Calliope, ‘‘La parole et son traitement automatique’’, Ed. Masson, 1989.

20. D. Raggai, ‘’Reconnaissance automatique de la parole par les modèles de Markov cachés’’, Mémoire de Magister, Département d’électronique, Université Saâd Dahleb, Blida, Algérie, Septembre 2000.

21. http://tcts.fpms.ac.be/cours/1005-07-08/speech/parole.pdf

22. F. Ykhlef, ‘‘Modification de la fréquence fondamentale en vue de la synthèse de la parole à partir du texte de l’Arabe standard’’, Mémoire de Magister, Département d’électronique, Université Saâd Dahleb, Blida, Algérie, Septembre 2005.

23. http://www.alis.isoc.org/glossaire/ phonetique.fr.html

24. http://lesla.univ-lyon2.fr/IMG/pdf/doc-587.pdf

25. Bellanger, ‘‘Traitement Numérique du Signal, Théorie et pratique’’, Ed. Masson 1980.

26. D. Vaufreydaz, ‘‘Modélisation statistique du langage à partir d’Internet pour la Reconnaissance Automatique de la Parole’’, Thèse de Doctorat, Université Joseph Fourier, Grenoble I, France, Janvier 2002.

27. http://www.eduscol.education.fr/D0033/algerie_actedumas.pdf

28. M. Barkat-Defradas, R. Hamdi and F. Pellegrino, ‘‘De la caractéristique linguistique à l’identification automatique des dialectes arabes’’, Acte de Workshop MIDL, Carré des sciences, Paris 29-30 Novembre 2004.

29. D. Fouad, ‘‘Résumé automatique de texte arabe’’, Mémoire de Master, Département d'informatique et de recherche opérationnelle, Université de Montréal, Canada, Septembre 2004.

30. K. Kirchhoff, ‘‘Novel Speech Recognition Models for Arabic’’, Workshop final report, University of Washington, USA, 2002.

31. http://fr.wikipedia.org/wiki/Alphabet_arabeWikipédia

32. S. Baloul, ‘‘Développement d’un système automatique de synthèse de la parole à partir de texte arabe standard voyellé’’, Thèse de Doctorat de l’Université du Maine, Académie de Nantes, France, Mai 2003.

33. http://www.primlangues.education.fr/php/download.php/PhonoAra.pdf

34. http://transliteration.org

35. http://www.vieartificielle.com

36. J. Rachedi, ‘‘Reconnaissance et classification de phonèmes’’, Mémoire pour le Master Sciences et Technologie de l’UPMC, Paris, France, Août 2005.

37. http://www-prima.imag.fr/Vaufreydaz/These/Reconnaissance.html

38. A. Rogozane, ‘‘Étude de fusion de données hétérogènes pour la Reconnaissance Automatique de la Parole Audio Visuelle’’, Thèse de Doctorat de l’Université d’Orsay Paris XI, France, Juillet 1999.

39. S. Fu, ‘‘Audio/Visual Mapping Based on Hidden Markov Models’’, Master of Science in Computer Engineering, B.S. Beijing University of Posts and Telecommunications, China, 1994.

40. N. Eveno, ‘‘Segmentation des lèvres par un modèle déformable analytique’’, Thèse de Doctorat de l’Institut National Polytechnique de Grenoble, France, Novembre 2003.

41. http:// www.icp.inpg.fr/ICP/publis/synthese/_jv/pub98.doc
Un synthétiseur audiovisuel 3D du Langage Parlé Complété (LPC) pour le Français.

42. J. Mariani, ‘’Reconnaissance automatique de la Parole’’, Ed. Hermès, 2002.

43. http://alize3.finances.gouv.fr/criph/lacces/lalecturelabiale.html

44. http://perso.wanadoo.fr/ecouter/lcp.html

45. http://www.asdadjerba.org.tn/Orthophonie.html

46. R. Goecke, ‘‘A Stereo Vision Lip Tracking Algorithm and Subsequent Statistical Analyses of the Audio-Video Correlation in Australian English’’, PhD Thesis, the Australian National University, Australia, Janvier 2004.

47. http://www.garycmartin.com/phoneme_examples.html

48. S. Lee and D. Yook, ‘’Viseme Recognition Experiment Using Context Dependent Hidden Markov Models’’, Proceedings of the Third International Conference on Intelligent Data Engineering and Automated Learning, Vol. 2412. pp.557 – 561, London, UK, 2002.

49. G. Ernten, Ph.D, ‘’Hierarchical Hidden Markov Model (HMM) Topologies for Robust Object Recognition’’, IC Tech, Inc, USA, 2000.

50. J. Dongmei, X. Lei, Z. Rongchun, W. Verhelux, I. Ravyse and H.Sahli, ‘‘Acoustic Viseme Modelling For Speech Driven Animation: A Case Study’’, Proc. 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002), Leuven, Belgium, pp.49-52, November 15, 2002.

51. M. Visser, M. Poel and A. Nijholt, ‘‘Classifying Visemes for Automatic Lipreading’’, Inernational Workshop Text, Speech and Dialogue (TSD99), Plzen, Czech republic, pp.349-352, 13-17 September 1999.

52. M. Ouhyoung, I-C. Lin and D.S.D. Lee, ‘‘Web-enabled Speech Driven Facial Animation’’, Int Conf Artif Real Telexistence, vol.9, pp.23-28, Japan 1999.

53. P. Lucey, T. Martin and S. Sridharan, ‘’Confusability of Phonemes Grouped According to their Viseme Classes in Noisy Environments’’, 9th Int Conf on Speech science & Technology, Maquarie University, Sydney, 9-10 December 2004.

54. M. Ferrouga and S. Kati, ‘‘Réalisation d'un OCR basé sur la combinaison série pour la reconnaissance de caractères imprimés dégradés ’’, Projet de Fin d’Étude d’Ingéniorat, Département électronique, Université Saâd Dahled, Blida, Algérie, Juin 2004.

55. M. Heckmann, F. Berthommier and K. Kroschel, ‘’Noise Adaptive Stream Weighting in Audio-Visual Speech Recognition’’, EURASIP Journal on Applied Signal Process. vol. 1, pp. 1260–1273, 2002.

56. V.A. Dubesset, “La langue française parlée complétée (LPC) : Production et perception“, Thèse de Doctorat de l’Institut National Polytechnique de Grenoble, Spécialité – Sciences cognitives, France, Novembre 2005.

57. G. Bailly, E. Vatikiotis-Bateson and P. Perrier, “Issues in Visual and Audio-Visual Speech Processing”, Eds., MIT Press, 2004.

58. M.A. Bencherif, ‘‘ Pathologie du langage arabe, Etude du stigmatisme occlusif ’’, Mémoire de Magister, Département électronique, Université Saâd Dahled, Blida, Algérie, Septembre 2005.

59. E. Francesconi, M. Gori and S. Marini ‘‘A serial combination of connectionist based classifiers for OCR’’, IJDAR 3:160-168, 2001.