N° d'ordre : 16/2008, M/EL

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université des Sciences et de la Technologie
HOUARI BOUMEDIENE

Faculté d'Électronique et d'Informatique

Mémoire
Présentée pour l'obtention du diplôme de Magister
En ÉLECTRONIQUE
Spécialité : Communication Parlée

Par : ***M^elle^ Nadia BAKIR***

Sujet

Reconnaissance automatique de la parole par fusion audiovisuelle dans un milieu reél

Soutenu le 06/07/2008, devant le jury composé de :

Mr. M. ATTARI	Professeur	USTHB	President
Mr. M. DEBYECHE	Maître de Conférences	USTHB	Directeur de Thèse
Mr. A. AMROUCHE	Maître de Conférences.	USTHB	Examinateur
Mr. H. TEFFAHI	Maître de Conférences	USTHB	Examinateur
Mr. Y. CHIBANI	Professeur	USTHB	Invité

Remerciements

Mes premiers remerciements vont à mes promoteurs Dr. Mohamed DEBYECHE et Pr. Youcef CHIBANI, pour leur soutien de toujours et l'intérêt constant avec lequel ils ont suivi mon travail.

Je tiens également remercier ma sœur et collègue Hassiba Nemmour, Maître assistante à la faculté d'Electronique et d'Informatique, pour son soutien durant la période de mon travail.

Je tiens à remercier Monsieur Mokhtar ATTARI, Professeur à la faculté d'Electronique et d'Informatique, pour avoir accepté la présidence de ce jury.

Je remercie également, Mr. Houcine TEFFAHI, Maître de conférence à la faculté d'Electronique et d'Informatique. Mr. A. AMROUCHE, Maître de conférence à la faculté d'Electronique et d'Informatique, pour avoir accepter de faire partie de ce jury.

Je remercie également Mr. AMAR DJERADI, Directeur du laboratoire LCPTS et Professeur à la faculté d'Electronique et d'Informatique et Mr. Amrane HOUACINE, Professeur à la faculté d'Electronique et d'Informatique.

A tous les membre du laboratoire LCPTS

Qu'il me soit permis enfin d'associer dans une même pensée amicale tous ceux qui ont contribué à la réalisation de ce travail. Qu'ils trouvent tous ici l'expression de ma très vive sympathie.

A mes très chers parents.

Résumé

L'utilisation d'informations supplémentaires conjointement à celles extraites du signal acoustique est une nouvelle méthode utilisée afin d'améliorer les performances et la robustesse des systèmes de reconnaissance automatique de la parole. De nombreux travaux sur la perception de la parole ont montré l'importance des informations visuelles dans le processus de reconnaissance chez l'homme. L'utilisation de données sur la forme et le mouvement des lèvres du locuteur semble donc être une voie prometteuse pour la reconnaissance de la parole.

Notre travail, dans le cadre de ce magister, concerne la mise en œuvre d'un système de reconnaissance automatique de la parole (RAP) audiovisuelle pour les chiffres arabes en milieu réel. Il s'agit d'une intégration des informations visuelles aux informations acoustiques.

Tout d'abord, se pose le problème du niveau dans lequel se fera la fusion : est-ce au niveau des données ou bien au niveau des résultats. Ensuite intervient le problème d'adaptation des contributions des deux modalités acoustique et visuelle.

Le système audiovisuel que nous avons mis en œuvre utilise les modèles de Markov cachés continus (CHMM) comme moteur de reconnaissance aussi bien pour la modalité acoustique que pour la modalité visuelle. Le résultat final de reconnaissance est obtenu après fusion des scores issus de chaque reconnaisseur. Les CHMM constituent l'approche la plus performante actuellement pour la RAP. L'estimation des paramètres des modèles par maximum de vraisemblance nécessite des procédures itératives, chaque itération mettant elle-même en jeu des parcours récursifs, visant à calculer la loi conditionnelle des variables cachées. La difficulté consiste à obtenir des algorithmes efficaces, interprétables de manière probabiliste.

La fusion des scores est basée sur l'utilisation de réseaux de neurones de type Perceptron MultiCouches (PMC)

Nous avons testé les performances de notre système sur un corpus audiovisuel des chiffres arabes en mode mono locuteur. Ce corpus a été enregistré par nos soins au niveau de notre laboratoire. Les paramètres auditifs utilisés sont les coefficients cepstraux dans l'échelle Mel (Mel Frequency Cepstral Coefficients) et les paramètres visuels sont eux basés sur la DCT (Discrete Cosine Transform). Les tests réalisés, dans un milieu réel, ont montré que de bonnes performances sont obtenues pour le système acoustique (TMBR = 69.33%) par rapport au système visuel (TMBR = 59.33%), elle meilleures pour la reconnaissance audiovisuelle (TMBR = 87%).

Les expériences réalisées nous ont permis de constater que l'information visuelle intégrée à l'information acoustique peut constituer une alternative pour augmenter la performance des systèmes de reconnaissance en milieu réel (nécessairement bruité).

Les performances de notre système restent à être évaluées dans un contexte plus étendu, par exemple un corpus de plus grande taille prononcé par plusieurs locuteurs. Nous comptons tester notre système dans le cas des signaux bruités à différents bruits. Nous pensons également utiliser d'autres types de fusion telles que la fusion au niveau des paramètres ou la fusion hybride.

Sommaire

Chapitre 2 : *Modèles de Markov Cachés Continus*

Chapitre 3 *: Fusion audiovisuelle pour la reconnaissance automatique de la parole*

Chapitre 4 : *Expériences et résultats*

Liste des figures

Liste des tableaux

Liste des abréviations

RAP	Reconnaissance Automatique de la Parole
HMM	Hidden Markov Model
CHMM	Continuous Hidden Markov Model
LPC	Linear Prediction Coding
LPCC	Linear Prediction Cepstral Coefficients
MFCC	Mel Frequency Cepstral Coefficients
SRAP	Système de Reconnaissance Automatique de la Parole
SRAPAV	Système de Reconnaissance Automatique de la Parole AudioVisuelle
FD	Fusion Directe
FS	Fusion Séparée
FH	Fusion Hybride
FI	Fusion Intermédiaire
ID	Identification Directe
IS	Identification Séparée
DCT	Discrete Cosine Transform
PMC	Perceptron MultiCouches
EM	Expectation-Maximization
MLE	Maximum Likelihood Estimate
LCPTS	Laboratoire de Communication Parlée et Traitement de Signal
TMBRG	Taux Moyen de Bonne Reconnaissance Global

INTRODUCTION GENERALE

La Reconnaissance Automatique de la Parole Audio-Visuelle est un domaine de recherche qui a connu un intérêt grandissant durant ces dernières années. Elle se situe à l'intersection de deux grands axes de recherche : *la modélisation de la perception de la parole audiovisuelle et la reconnaissance automatique des formes*.

La parole est un moyen de communication audiovisuelle. Le message parlé est plus intelligible quand il est accompagné de la vision des lèvres du locuteur, surtout quand le milieu de transmission est dégradé.

Depuis plusieurs siècles, l'observation des mouvements labiaux est utilisée pour la compréhension de la parole par l'humain en l'absence de paramètres acoustiques. Il s'agit alors d'une reconnaissance purement visuelle de la parole également appelée lecture labiale**.** D'ailleurs, la lecture labiale permet aux malentendants de restituer une grande partie de l'intelligibilité de la parole acoustique. Chez les personnes ayant une présence auditive, les lèvres permettent de mieux comprendre un message complexe.

La reconnaissance automatique de la parole audiovisuelle se propose d'utiliser, en plus de l'information acoustique, une information visuelle, reposant principalement sur les mouvements et la forme des lèvres du locuteur, comme source complémentaire d'information. Deux questions introductives importantes se posent dans ce domaine de recherche :

- *Quelle information visuelle utiliser et comment l'extraire des images ?*
- *Comment combiner les paramètres acoustiques et visuels ?*

La seconde question, se pose pratiquement dans les mêmes termes qu'en reconnaissance de la parole multi-bandes [Mirghafori et Morgan, 1999], avec notamment les difficultés liées à l'existence d'une désynchronisation entre les modalités acoustique et visuelle.

Travailler sur l'intégration des scores acoustiques et visuels reste d'actualité, mais, en l'absence de corpus de la parole audiovisuelle d'entraînement et de validation de grande taille, la généralisabilité des résultats obtenus reste limitée.

C'est pourquoi que, dans ce mémoire, nous sommes intéressé à ces deux questions : Extraction d'informations visuelles et méthodes d'intégrer les scores acoustiques et visuels.

Dans les systèmes automatiques, pour assurer l'extraction des paramètres visuels fiables, le maquillage [Auckenthaler et al., 1999a] et /ou un dispositif d'acquisition spécifique avec des conditions d'éclairage contrôlées [Liévin et Luthon, 1999] ont longtemps été et sont encore utilisés. Ces artefacts permettent d'étudier les paramètres labiaux et ont permis par le passé de prouver l'utilité de la modalité visuelle, mais restent difficilement envisageables dans les applications réelles. Pour pouvoir utiliser la parole audiovisuelle dans les applications réelles, il semble nécessaire d'étudier l'extraction des paramètres labiaux sur des images, que nous qualifierons par la suite de naturelle, acquises sans préparation du locuteur (sans maquillage ou dispositif d'acquisition solidaire de la tête du locuteur), dans un environnement réaliste soumis à des variations d'éclairement.

Dans notre travail se pose alors le problème d'intégrer des informations de nature différente : acoustique et visuelle. C'est précisément cette intégration d'informations : acoustiques et visuelles, en vue de leur exploitation pour la Reconnaissance Automatique de la Parole Audio-Visuelle (RAPAV) qui fait l'objet de ce mémoire.

Ainsi, ce manuscrit est organisé en quatre chapitres :

Le premier chapitre, présente quelques notions générales sur la reconnaissance automatique de la parole. Nous décrirons le fonctionnement d'un système de RAP. Après,

nous abordons quelques problèmes de la reconnaissance de la parole. Puis, Nous détaillerons la façon de représenter le signal de parole par le biais de sa paramètrisation.

La méthode utilisée pour la reconnaissance audio et la reconnaissance vidéo est abordée dans le *chapitre2*.

Dans le *troisième chapitre*, nous décrivons le schéma de fonctionnement général d'un système audiovisuel, puis nous présentons les différents modèles d'intégration audiovisuelle proposés dans la littérature, et les différentes techniques de fusion utilisées. Nous terminons par la paramétrisation des signaux acoustique et visuel de parole.

Le *dernier chapitre* se termine par une évaluation des performances de notre système. Nous décrivons la base de données audiovisuelle sur laquelle notre système a été expérimenté. Ensuite, nous définissons un ensemble de visèmes adapté à la tache de reconnaissance et aux données visuelles du locuteur. Puis nous testons la pertinence de cette définition pour la RAP avec la méthode de reconnaissance fondée sur des modèles de Markov cachés, et des réseaux de neurones

Chapitre 1

Notions générales sur la reconnaissance automatique de la parole

Résumé :

Ce chapitre a pour objet de présenter les notions générales sur la parole et de son traitement automatique. Pour commencer, nous évoquerons le principe général de la reconnaissance automatique de la parole (RAP) et en particulier le système le plus utilisé de nos jours à savoir le système bayesien utilisant les modèles de Markov cachés (HMM : Hidden Markov Model). Nous aborderons ensuite, la paramétrisation du signal de parole en vu de la reconnaissance. Avant de terminer, nous présenterons les différents problèmes de la RAP Finalement, nous donnons la définition générale sur la reconnaissance automatique de la parole audiovisuelle.

1.1. Introduction

La parole est l'un des moyens les plus naturels par lequel des personnes communiquent. C'est un son émis par le *locuteur*, c'est à dire une variation de pression acoustique plus ou moins rapide et plus ou moins forte qui est captée par un microphone placé à proximité. La possibilité de reconnaître la phrase sera donc très dépendante des *conditions d'enregistrement*: qualité du microphone, distance au locuteur, niveau du bruit environnemental ...etc.

La Reconnaissance Automatique de la Parole (RAP) a pour objet la transformation automatique du signal acoustique en une séquence de mots (ou de phrases) qui, idéalement, correspond aux mots (phrases) prononcés par le locuteur. La reconnaissance constitue donc un processus de transformation d'un signal de parole en mots ou séquence de mots de la langue naturelle.

La présence de différents bruits peut affecter significativement la qualité de la RAP. La reconnaissance audio-visuelle propose d'améliorer les performances des systèmes de RAP, principalement dans le cas ou le canal audio est corrompu, en ajoutant de l'information provenant de la modalité visuelle, sous la forme d'images vidéo du locuteur.

1.2. La lecture labiale

La lecture labiale permet de comprendre le message parlé en mettant en relation les mouvements des lèvres et du visage. Outre le travail de *déchiffrage* des mots, un travail est nécessaire pour que les mots aient un sens.

Pour avoir une bonne lecture labiale, trois éléments sont nécessaires :

- Le locuteur, c'est-à-dire celui qui s'exprime, doit :
  - se mettre obligatoirement face à la personne malentendante ;
  - parler normalement (à trop vouloir articuler ou parler trop fort, on déforme les mots) ;
  - éviter de parler trop vite ou, 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.

Quelques réflexions sur la lecture labiale

Chacun lit des mots sur les lèvres de son interlocuteur. Une telle affirmation ne va pas sans poser certaines questions :
- Quand est ce que l'on se sert de la lecture labiale pour percevoir et comprendre la parole?
- Pour quelles raisons la lecture labiale est-elle alors efficace?
- comment utilise-t-on les informations visuelles sur les lèvres du locuteur?

1.3. Signal de parole acoustique

1.3.1. Définition

Le signal de parole acoustique (capté par un microphone) véhicule des informations de nature différentes, ceci impose aux systèmes de reconnaissance de n'extraire que l'information nécessaire à son application. L'information sur celui qui a émis le message pour la reconnaissance du locuteur, dans quelle langue le massage a été émis pour la reconnaissance de la langue et enfin l'information linguistique du message émis pour la reconnaissance de la parole. Ces différentes informations sont portées par des paramètres tels que : le fondamental F_0, les formants, le codage par prédiction lineaire *LPC* (LPC : Linear Prediction Cepstral), les coefficients cepstraux *MFCC* (MFCC : Mel Frequency Cepstral Coefficients) ... etc.

1.3.2. Caractéristiques du signal de parole

Le signal de parole n'est pas un signal ordinaire, il s'inscrit dans le cadre de la communication parlée. Il présente plusieurs caractéristiques spécifiques qui rendent son traitement extrêmement complexe. Parmi ces caractéristiques : particularité de la langue, redondance de l'information, variabilité du signal, les interférences ... etc. [Debyeche, 20007]

1.3.3. Du signal de parole à l'observation acoustique

Cette partie présente le principe de traitement dans un Système de Reconnaissance Automatique de la Parole (SRAP).

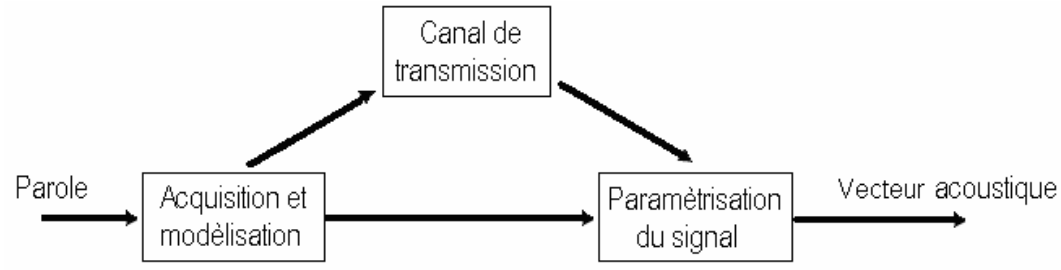

Figure 1.1 : *Chaîne de traitement acoustique d'un système de RAP*

Comme le montre la *figure 1.1*, le signal de parole est d'abord numérisé puis modélisé sous une forme généralement fréquentielle. Avant d'obtenir ces mesures, le signal a subit des modifications dues à l'environnement dans lequel se trouve le locuteur, à l'influence du système d'acquisition. Ces modifications sont souvent regroupées sous le terme *canal de transmission*. Le module suivant, dans la chaîne de traitement acoustique, est celui qui extrait les paramètres pertinents pour la reconnaissance de la parole. Ces paramètres sont ensuite envoyés au module de reconnaissance acoustique qui identifie les sons présents dans le signal.

Détaillons chacun de ces modules pour comprendre l'enchaînement allant du signal de parole à l'observation acoustique

1.3.3.1. Acquisition et modélisation du signal

Pour que le signal soit utilisable par un ordinateur, il doit être numérisé. Cette opération tend à transformer un phénomène temporel analogique, en une suite d'éléments discrets, les échantillons. La numérisation repose sur l'échantillonnage et la quantification du signal.

1.3.3.2. Canal de transmission

Comme dans la tache de reconnaissance de la parole, l'une des deux entités de la chaîne de communication est un ordinateur. De ce fait, il est nécessaire de prendre en compte le canal de transmission entre l'être humain et la machine, car celui-ci introduit des distorsions qui sont de nature à perturber suffisamment le signal de parole pour le rendre difficilement reconnaissable pour la machine. Ce canal de transmission est en général assimilé à un filtre. Il est possible d'inclure dans ce canal des informations comme la réponse impulsionnelle où l'enregistrement est effectué, ou encore le bruit de fond.

1.3.3.3. Extraction des paramètres

A partir des échantillons de chaque trame (observation), un traitement acoustique extrait des coefficients caractéristiques qui sont rassemblés en un vecteur descriptif de cette tranche, désigné sous le nom de *vecteur acoustique*. La définition et le nombre de coefficients différent selon les systèmes de reconnaissance, on rencontre principalement : les coefficients cepstraux dans l'échelle Mel (MFCC : Mel Frequency Cepstral Coefficient), les coefficients LPC (Linear Predictive Coding), l'énergie, le fondamental ... etc. [Dupont, 1996]

1.4. Reconnaissance automatique de la parole

Cette partie est dédiée à la présentation du principe de la RAP. Le principe général a beaucoup évolué car, en passant d'une reconnaissance fondée sur l'exemple à une reconnaissance fondée sur le modèle, la suite de traitements s'est allongée. De façon générale, les systèmes de RAP actuels, à base de modélisation statistique, suivent le schéma représenté par la *figure 1.2 :*

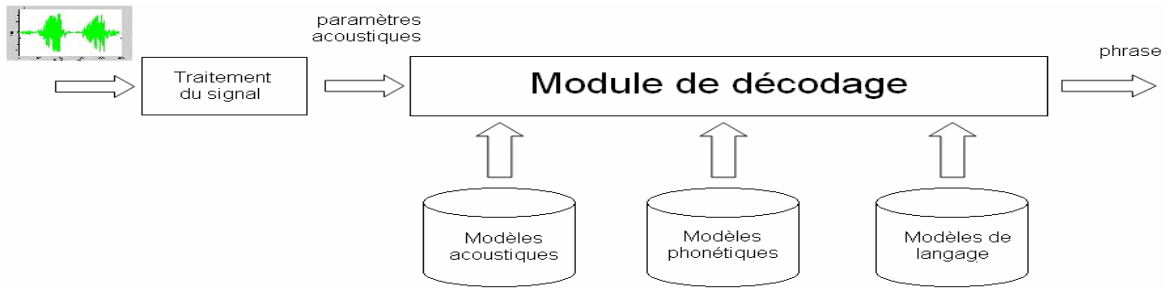

Figure 1.2 : *Système général de RAP acoustique par modélisation statistique.*

À partir d'un signal de parole, le premier traitement consiste à extraire les paramètres acoustiques. Ces paramètres sont mis en entrée d'un module de décodage. Ce décodage peut produire une ou plusieurs hypothèses phonétiques associées en général à une probabilité pour chaque segment (une fenêtre ou une trame) de signaux de parole. Ce générateur d'hypothèses utilise des modèles statistiques d'unités élémentaires de parole, par exemple le phonème [Dutoit et al., 2002].

Le générateur d'hypothèses interagit pour forcer le décodage acoustico-phonétique à ne reconnaître que des mots. Les modèles phonétiques sont représentés par un dictionnaire de prononciation (dictionnaire phonétique) ou par des automates probabilistes qui sont capables d'associer une probabilité à chaque prononciation possible d'un mot.

Pour la reconnaissance automatique de la parole continue grand vocabulaire, le générateur d'hypothèses de mot interagit pour forcer le reconnaisseur à intégrer des contraintes qui sont souvent formalisées par des modèles de langage [Le Viet Bac, 2006].

Approche Bayesienne de la reconnaissance automatique de la parole

La formule générale, dans le cadre d'un système entièrement probabiliste, s'exprime sous la forme d'une équation *Bayesienne (1.1)*. Le but du système est de trouver l'hypothèse W^* qui maximise pour toutes les séquences de mot W possibles et pour une observation acoustique A.

$$W^* = \arg\max_W P(W|A) = \arg\max_W \frac{P(W).P(A/W)}{P(A)} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1.1)$$

Dans cette équation, nous pouvons identifier plusieurs facteurs :

- $P(A)$: est la probabilité de l'observation acoustique A. Celle-ci est constante pour toutes les séquences de mot W, d'où l'approximation finale de l'équation précédente est donnée par la formule *(1.2)*. Pour générer cette observation, l'unité de décodage acoustique doit, dans un premier temps, analyser le signal de parole, ensuite définir quelle est la suite d'éléments acoustiques la plus probable.

$$W^* \approx \arg\max_W \left(P(A/W).P(W)\right) \quad \text{..... (1.2)}$$

- $P(AIW)$: est la probabilité de l'observation acoustique A connaissant une séquence de mots W.
- $P(W)$: est la probabilité *a priori* de la séquence de mots W, sans aucune notion d'acoustique, dans le langage considéré. C'est la probabilité générée par le modèle de langage.

Généralement, pour classer les systèmes de reconnaissance automatique de la parole, on a recours aux critères suivants :

- Le mode d'élocution.
- La taille du vocabulaire.
- La dépendance plus ou moins grande vis-à-vis du locuteur.
- L'environnement protégé. [Dupont, 1996]

1.5. Principales méthodes d'analyse du signal de parole

La Reconnaissance Automatique de la Parole a pour objet la transformation automatique d'un signal acoustique en une séquence de phonèmes ou de mots qui, idéalement, correspond au mot ou à la phrase prononcée par un locuteur. La reconnaissance constitue donc un processus de traduction d'un signal acoustique en une séquence de phrases de la langue orale telle qu'elle est prononcée par un locuteur, reconnu par un système automatique.

Le traitement acoustique regroupe l'acquisition du signal vocal, son filtrage, son échantillonnage et l'extraction des coefficients caractéristiques de ce signal. Au niveau acoustique, la parole est donc représentée par une suite de vecteurs constitués des coefficients qui caractérisent le signal [Calliope, 1989].

Pour obtenir une représentation numérique, on échantillonne le signal acoustique de parole. La fréquence d'échantillonnage est choisie conformément au théorème d'échantillonnage (au-delà de 8KHz pour un signal dont la bande passante aurait été limité par la qualité téléphonique).

Les deux familles de coefficients acoustiques les plus utilisées en RAP sont issus de deux analyses différentes pour obtenir le spectre :

- ✓ Lorsque le spectre d'amplitude résulte d'une FFT sur le signal de parole prétraité, lissé par une suite de filtres triangulaires repartis selon l'échelle MEL, les coefficients sont appelés Mel Frequency Cepstral Coefficients (MFCC)(voir *figure :1.3*).
- ✓ Lorsque le spectre correspond à une analyse LPC (LPC : Linear Prediction Cepstral), les coefficients se déduisent des coefficients LPC par développement de Taylor, d'où leur nom de Linear Prediction Cepstral Coefficients (LPCC) (voir *figure : 1.4*)

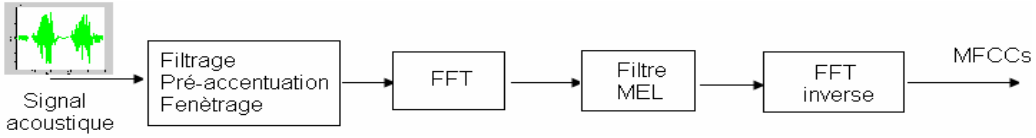

Figure 1.3 : *Calcul des coefficients acoustiques de type MFCCs*

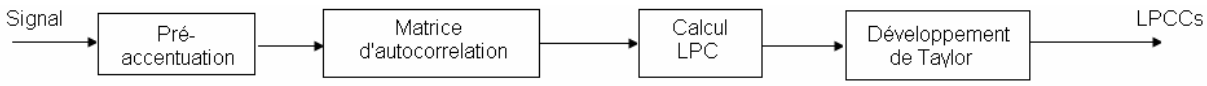

Figure 1.4 : *Calcul des coefficients acoustiques de type LPCC*

L'information acoustique permanente du signal de parole se situe principalement dans la bande passante [50 Hz – 8 KHz], la fréquence d'échantillonnage devrait donc au moins être égale à 16 kHz, selon le théorème de Shannon [Kunt, 1991], mais elle peut varier en fonction du domaine d'application ou des besoins ou des contraintes matérielles.

La trame acoustique est un ensemble de coefficients ou paramètres calculés sur un bloc d'échantillonnage. Dans la plus part des applications, ce bloc d'analyse est de taille fixe, il correspond à un temps de parole de 10 à 30 ms. La suite de vecteurs d'analyse est obtenue en déplaçant ce bloc de 5 à 15 ms ; il y a recouvrement de blocs, ce qui apparente cette analyse à une analyse de type fenêtre glissante.

En reconnaisse de la parole, les paramètres extraits doivent être :

- **Pertinents :** extraits de mesures suffisamment fines, ils doivent être précis mais leur nombre doit rester raisonnable afin de ne pas avoir de coût de calcul trop important dans les modules de décodage.
- **Discriminants :** ils doivent donner une représentation caractéristique des sons de base et les rendre facilement séparables.
- **Robustes :** ils ne doivent pas être trop sensibles à des variations de niveau sonore ou un bruit de fond.

De nombreuse paramétrisation ont été utilisées en parole ; les méthodes issues du traitement de signal sont classiquement répertoriées en trois catégories :

A) Les méthodes fondées sur un modèle de perception.
B) Les transformées non paramétriques usuelles telles que la transformé de Fourier.
C) Les méthodes paramétriques qui s'appuient sur un modèle simplifié de production de la parole et qui exploitent le couplage '' source/conduit'' (codage prédictif linéaire et cepstre).
D) Les paramètres ''coefficients cepstraux : *MFCC*''.

A) Méthode de calcul du modèle de perception

Des modèles de perception ont pu être obtenus à partir d'études de perception et d'études psycho-acoustiques. Ils consistent à définir des bandes critiques de perception, correspondant à la distribution fréquentielle de l'oreille humaine. Les coefficients sont les sorties de bancs de filtres calibrés à partir de ces résultats : cette technique est celle utilisée dans les vocodeurs à canaux [Zurcher, 80].

B) Méthodes non paramétriques

Ce type de paramétrisation fait appel aux techniques classiques utilisées en traitement de signal : les transformées temps fréquence et temps échelle [Flandrin, 1993]. Malgré quelques tentatives récentes d'exploitation des transformées de type ondelettes [Malbos, 1995], la transformée la plus utilisée en parole reste la transformée de Fourier discrète. La Transformée de Fourier Rapide (FFT) permet d'obtenir des spectres en temps. La description acoustique des sons qui s'appuie sur cette représentation se réalise de la façon suivante :

- Un filtre de pré-accentuation est appliqué afin d'égaliser les aigus toujours plus faibles que les graves.

- Un fenêtrage de type Hamming est effectué sur chaque bloc d'analyse de façon à diminuer les effets de bords dus au décodage en fenêtre.
- Une FFT est calculée; seul son module est retenu, la phase de transformé de Fourier numérique du signal de la parole ne contient pas d'information suffisamment pertinente pour la reconnaissance de la parole.

Cependant, d'une part, la période fondamentale fait apparaître de nombreuses harmoniques sur le spectre d'amplitude ainsi obtenu, et d'autre part, l'information reste redondante. Il est donc courant d'effectuer des lissages dans le domaine spectral. Pour tenir compte de la perception humaine, le spectre est ramené à une échelle non linéaire Mel, donnée par la formule suivante :

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad \text{..... (1.3)}$$

Afin de réduire l'information, une suite de filtres triangulaire est appliquée dans le domaine spectral selon l'une des échelles précédemment décrites. La *figure 1.5* donne un exemple de répartition d'une suite de filtres selon l'échelle Mel, couramment utilisée.

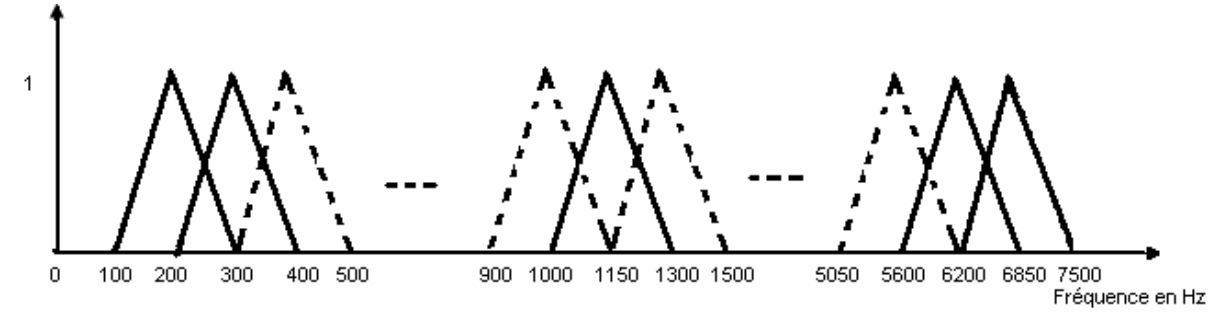

Figure 1.5 : *Répartition fréquentielle de filtre triangulaire (utilisée au CNET)*

C) Méthodes paramétriques

Ces méthodes s'appuient sur un modèle linéaire simplifié de production de la parole [Jacob, 1995]. Le signal vocal est considéré comme la sortie d'un filtre excité par une source. Le filtre modélise le conduit nasal, le conduit vocal et le rayonnement aux lèvres, tandis que la source correspond à un signal périodique ou un bruit aléatoire. L'analyse LPC (Linear Prediction Coding) simplifie ce modèle de production en supposant que le filtre ne comporte que des pôles [Markel et al., 76]. Les paramètres sont alors les coefficients du filtre, ils décrivent la fonction de transfert du conduit vocal.

D) Les coefficients cepstraux

Dans cette partie nous verrons plus précisément la représentation que nous avons choisie pour caractériser le signal audio : *les coefficients cepstraux*.

L'analyse cepstrale résulte du modèle de production ; son but est d'effectuer la déconvolution ‘'source/conduit'' par une transformation homomorphique : les coefficients sont obtenus en appliquant une transformé de Fourier numérique inverse ou logarithme du spectre d'amplitude. Le signal ainsi obtenu est représenté dans un domaine appelé cepstre ou quéfrenciel ; les échantillons se situant en basse quéfrences correspondent à la contribution du conduit vocal et donnent les paramètres utilisés en reconnaissance automatique de la parole, tandis que la contribution de la source n'apparaît qu'en hautes quéfrences.

Lorsque le spectre d'amplitude résulte d'une FFT sur le signal de parole prétraité, lissé par une suite de filtres triangulaires repartis selon l'échelle Mel, les coefficients sont appelés : Mel Frequency Cepstral Coefficients (MFCC).

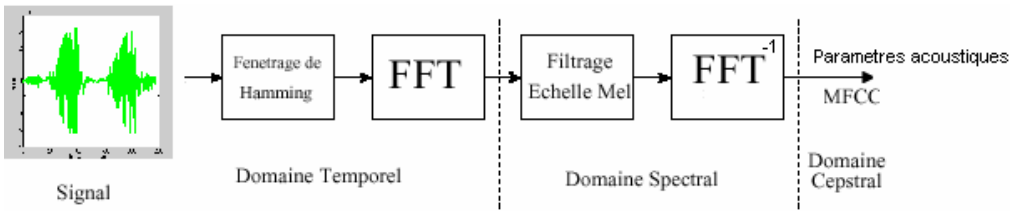

Figure 1.6 : *Schéma de calcul des coeficients cepstraux MFCCs.*

La paramétrisation présentée ci-dessus est la paramétrisation de base. De nombreuses variantes ont été utilisées dans le but principal d'accroître la robustesse et la résistance du bruit [Mokbel, 1992]. Ces coefficients caractérisent la forme statique du spectre. Afin d'introduire une information de nature dynamique sur la parole, signal composé de zones stables et transitoires, il est courant de calculer, dans le cas des coefficients cepstraux, les coefficients de régression associés qui sont les dérivées temporelles premières et secondes (ΔMFCC et ΔΔMFCC).

1.6. Problèmes de la reconnaissance automatique de la parole :

Pour bien appréhender le problème de la reconnaissance de la parole, il est bon de comprendre les différents niveaux de complexité et les différents facteurs qui en font un problème difficile.

Le système est-il dépendant du locuteur ou indépendant du locuteur (Pouvant reconnaître n'importe quel utilisateur) ?

Les systèmes dépendants du locuteur sont plus faciles à développer et sont caractérisés par de meilleurs taux de reconnaissance que les systèmes indépendants du locuteur étant donné que la variabilité du signal de parole est plus limitée. Cette dépendance au locuteur est cependant acquise aux prix d'un entraînement spécifique à chaque utilisateur. Ceci n'est cependant pas toujours possible. Par exemple, dans le cas d'applications téléphoniques, il est évident que les systèmes doivent pouvoir être utilisés par n'importe qui et doivent donc être indépendants du locuteur. Bien que la méthodologie de base reste la même, cette indépendance au locuteur est cependant obtenue par l'acquisition de nombreux locuteurs qui sont utilisés simultanément pour l'entraînement de modèles susceptibles d'en extraire toutes les caractéristiques majeures. Une solution intermédiaire parfois utilisée est de développer des systèmes capables de s'adapter (de façon supervisée ou non supervisée) rapidement au nouveau locuteur.

Le système reconnaît-il des mots isolés ou de la parole continue ?

Evidemment, il est plus simple de reconnaître des mots isolés bien séparés par des périodes de silence que de reconnaître la séquence de mots constituant une phrase. En effet, dans ce dernier cas, non seulement la frontière entre mots n'est plus connue mais, de plus, les mots deviennent fortement articulés (c'est-à-dire que la prononciation de chaque mot est affectée par le mot qui précède ainsi que par celui qui suit : - Un exemple très simple et bien connu étant les liaisons du français).

Dans le cas de la parole continue, le niveau de complexité varie selon qu'il s'agisse de texte lu, de texte parlé ou, beaucoup plus difficile, de langage naturel avec ses hésitations, phrases grammaticalement incorrecte, faux départ ...etc. Un autre problème, qui commence à être bien maîtrisé, concerne *la reconnaissance de mot clés* en parole. Dans ce cas, le vocabulaire à reconnaître est relativement petit et bien défini mais le locuteur n'est pas contraint de parler en mots isolés. Par exemple, si un utilisateur est invité à répondre par « *OUI* » ou « *NON* », il peux répondre « oui, s'il vous plait ». Dans ce contexte, un problème qui reste particulièrement difficile est de rejet de phrases ne contenant aucun mots clés.

La taille du vocabulaire et son degré de confusion sont également des facteurs importants. Les petits vocabulaires sont évidemment plus faciles à reconnaître que les grands vocabulaires, étant donné que dans ce dernier cas, les possibilités de confusion augmentent. Certains petits vocabulaires peuvent cependant s'avèrent particulièrement difficiles à traiter ; ceci est le cas, par exemple, pour l'ensemble des lettres de l'alphabet.
Le système est il robuste, c'est-à-dire, capable de fonctionner proprement dans des *conditions difficiles* ? En effet, de nombreuses variables pouvant affecter significativement les systèmes de reconnaissance ont été identifiées :

- Les bruits d'environnement tels que bruit additifs stationnaires ou non stationnaires (par exemple, dans une voiture ou dans une usine).
- Acoustique déformée et bruits (additifs) corrélés avec le signal de parole utile (par exemple distorsions non linéaires et réverbérations).
- Utilisation de différents microphones et différentes caractéristiques (fonction de transfert) du système d'acquisition du signal (filtres), conduisant généralement à du bruit de convolution.
- Bande passante fréquentielle limitée (par exemple dans le cas des lignes téléphoniques pour les quelles les fréquences transmises sont naturelles limitées entre environ 300Hz et 3400Hz [Boite et al, 1999]).
- Elocution inhabituelle ou altérée, comprenant entre autre : l'effet Lombard, (qui désigne toutes les modifications, souvent inaudibles, du signal acoustique lors de l'élocution en milieu bruité), le stresse physique ou émotionnel, une vitesse d'élocution inhabituelle, ainsi que les bruits de lèvres ou de respiration.

Certains systèmes peuvent être plus robustes que d'autres à l'une ou l'autre de ces perturbations, mais en règle générale, les reconnaisseurs de parole actuels restent encore très sensibles à ces paramètres. [Boite et al., 1999]

Particularités de la langue : La langue arabe est mise en regard de cinq autres langues selon six axes dans le *tableau* qui suit. Son originalité réside dans la combinaison de quatre caractéristiques : inflexion, liaison, dérivation et agglutination.

- La *dérivation* : est l'opération morphologique la plus complète qui permet de rattacher un mot sa racine après dépouillement de ses éventuelles préfixes ou suffixes. L'ensemble de dérivés rassemble ce qu'il est convenu d'appeler communément les mots de la même famille comme par exemple ''*KARAA*'', ''*KIRAA*'' ou ''*KARIA*''.
- Les *inflexions* d'un mot correspondent à toutes les formes obtenues par conjugaison de ce mot. Par exemple, ''YADHHABOU'',''DHAHABATA'' sont deux flexions du lemme ''DHAHABA'', mais ''KATABA'' et ''KITAB'' ne proviennent pas d'un même lemme bien qu'étant de la même famille.

	arabe	français	allemand	italien	espagnol	anglais
inflexion	+	+	+	+	–	+
composition	+–	+	–	–	–	–
apostrophe	+	–	+	–	–	–
liaison	+	–	–	–	–	+
dérivation	+	+	+	+	+	+

Tableau 1.1 : *Comparaison de l'arabe avec d'autres langues.*

1.7. Reconnaissance Automatique de la Parole Audio-Visuelle (RAPAV)

17.1. Fusion audiovisuelle

Malgré de nombreux progrès dans la conception des systèmes de reconnaissance, leur fiabilité reste toujours insuffisante quand les conditions de test sont différentes de celles utilisées pour l'apprentissage. Les résultats montrent une chute des taux de reconnaissance quand le locuteur se trouve dans un milieu bruité [Adjoudani, 1993]. Pour pallier ce problème, des chercheurs ont commencé à utiliser les informations visuelles comme source secondaire à l'appui des informations acoustiques pour réduire les effets de bruit et améliorer les scores de reconnaissance [Stork et al., 1992 ; Silsbee, 1993 ; Bregler et al., 1993].

L'objectif de ce type de reconnaissance est de développer un prototype qui extrait des caractéristiques visuelles (ex.: forme de la bouche pour les chiffres arabes), et les fusionner avec les caractéristiques acoustiques (ex. : les coefficients cepstraux MFCC) fournies par le système de reconnaissance de la parole audiovisuelle. Plusieurs facteurs affectant les performances seront alors à considérer : la détection et le suivi de l'image et de la bouche, la pose de la tête, la présence de plusieurs images, les conditions d'illumination, l'encodage des caractéristiques visuelles, le prétraitement des données, la technique de fusion et le temps de traitement, en sont des exemples. Il existe trois approches de fusion d'après la *figure 1.7* :

1. fusion des caractéristiques audio et visuelles de bas niveau.
2. fusion au niveau des décisions phonétiques.
3. combinaison des deux approches précédentes.

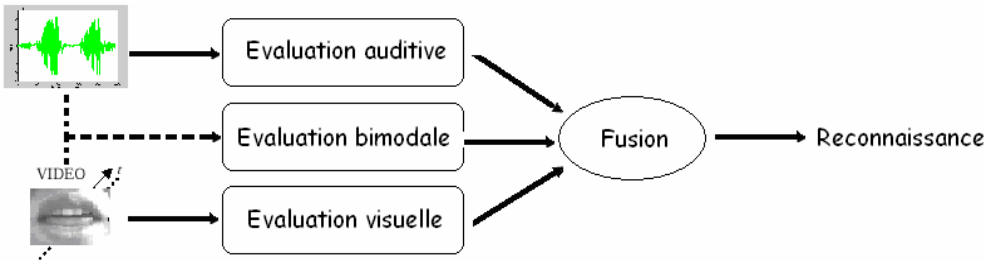

Figure 1.7 *: Description de système de Reconnaissance Automatique de la Parole Audio-Visuelle.*

1.7.2. Chaîne audio

Un système de paramétrisation du signal, appelé aussi prétraitement acoustique, se décompose en trois étapes, un filtrage analogique, une conversion analogique/numérique et un calcul de coefficients. Son rôle est de fournir et d'extraire des informations caractéristiques et pertinentes du signal pour une représentation moins redondante de la parole. Le signal analogique est fourni en entrée et une suite discrète de vecteurs, appelés trame acoustique, est obtenue en sortie.
Le processus de calcul des coefficients acoustiques est donné par le schéma de la *figure* suivante :

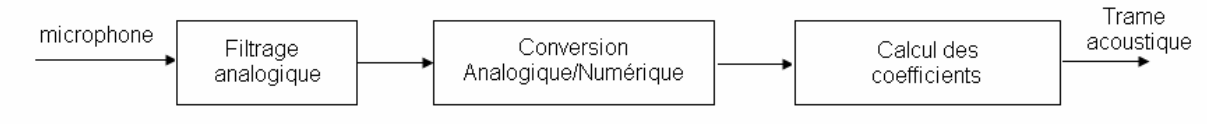

Figure 1.8 : *Schéma général d'un traitement acoustique.*

1.7.3. Chaîne vidéo

La reconnaissance visuelle de la parole est un processus complexe en raison de la diversité des traitements qu'elle requière, depuis l'extraction des informations visuelles jusqu'à leur interprétation phonétique, mais aussi en raison de la variabilité interlangues, inter et intralocuteur et selon le contexte phonétique, du signal visuel de parole. Pour faciliter ce processus de reconnaissance, le système doit utiliser une parametrisation visuelle adaptée aux phénomènes physiques qu'il doit observer. Ces phénomènes sont, dans ce cas, la forme et les mouvements des lèvres. Nous avons choisi d'utiliser seulement les paramètres mesurés sur la forme des lèvres. Ces paramètres bien que peu nombreux constituent une représentation visuelle pertinente comme l'attestent différentes études tant en perception qu'en reconnaissance automatique.

Sur cette parametrisation visuelle, nous devons construire une méthode de reconnaissance adaptée aux événements de parole. Cette méthode provient, comme pour le système acoustique, d'une modélisation par chaîne de Markov cachée des effets visuels des mots. L'utilisation des mêmes méthodes pour l'acoustique et le visuel diminue les problèmes de mise en correspondance des points de vue acoustiques et visuels lors de leur fusion pour la reconnaissance bimodale de la parole.

La *figure 1.9* présente le schéma général d'un traitement vidéo :

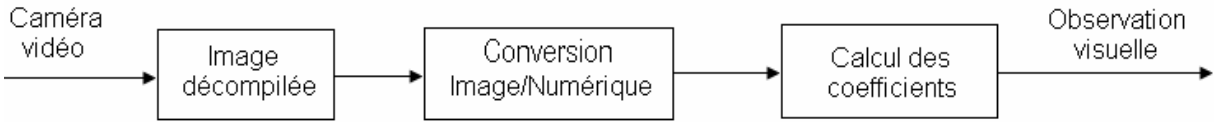

Figure 1.9 : *Schéma général d'un traitement vidéo.*

1.8. Conclusion

Dans ce chapitre, nous nous somme intéressés aux bases de la Reconnaissance Automatique de la Parole (RAP), ensuite nous avons présenté les problèmes de la *RAP*. Nous avons également mis en évidence les particularités du signal de parole et l'approche de

reconnaissance la plus utilisée de nos jours : l'approche Bayesienne. Enfin, nous avons évoqué le principe général d'un Système de Reconnaissance Automatique de la Parole AudioVisuelle (S*RAPAV*).

Chapitre 2

Modèles de Markov cachés continus

Résumé :

Ce chapitre est consacré à l'approche globale basée sur les modèles de Markov cachés (HMM : Hidden Markov Model). Il nous permettra d'exposer les différents outils théoriques utilisés au cours de la mise en œuvre de cette approche stochastique de reconnaissance de la parole, en particulier la théorie de la modélisation Markovienne. Donc, nous allons présenter, dans un premier temps, la définition des HMMs. Ensuite, nous poserons les différents problèmes liés aux HMMs. Et nous terminons par la présentation des algorithmes fondamentaux à ces modèles.

2.1. Introduction

Les modèles de Markov cachés (par définition en anglais : ''HMM'' Hidden Model Markov) ont été introduits par Baum et ses collaborateurs dans les années 1960-70. D'ailleurs, ils sont utilisés en reconnaissance de la parole à partir des années 80 [Rabiner, 1989]. Le modèle de Markov caché est fortement apparenté aux automates probabilistes.

Cet automate est un graphe probabilisé dans lequel chaque nœud est censé produire un ou plusieurs segments stables ou transitoires du signal vocal. A chaque état ou nœud est associée une distribution de probabilité d'émettre un vecteur spectral. Un automate probabiliste est une structure composée d'états et de transitions, et un ensemble de distributions de probabilités de transitions.

Les HMMs sont devenus la méthode la plus utilisée pour la modélisation des signaux de la parole dans les applications suivantes : reconnaissance automatique de la parole, suivi de la fréquence fondamentale et des formants, synthèse vocale, traduction automatique, étiquetage syntaxique, compréhension du langage oral.

2.2. Définition

Les modèles HMMs tentent de modéliser les unités représentatives de la parole par des modèles statistiques. Ils supposent que la suite de vecteurs acoustiques $X = x_1, x_2, ...x_T$ représentatifs du signal de parole est stationnaire par morceaux, ce qui signifie que, par morceau, les vecteurs acoustiques suivent la même loi de probabilité. On associe donc au processus X un processus caché Y où Y_t est une indicatrice de la loi correspondant à X_t. Pour modéliser l'évolution temporelle de la parole, la loi du processus Y est donnée par une chaîne de Markov homogène, généralement d'ordre 1, ce qui signifie que, le saut d'état est toujours de l'état i à l'état $i+1$. On représente habituellement le processus Y sous forme d'un automate stochastique comme illustré par la *figure 2.1*, une densité de probabilité étant associée à chacun des états de l'automate. L'automate étant capable, après apprentissage, d'estimer la probabilité qu'une séquence d'observations ait été générée par ce modèle

Formellement, un modèle de Markov caché est défini par le nombre d'états N de l'automate et de l'ensemble des paramètres λ suivants :

$$\lambda = \left\lfloor A=(a_{ij}\ ; i,j=1,.....,N),\ B=b_j(.),\ \Pi=(\pi_i\ , i=1,.....,N) \right\rfloor \quad ..... \quad (2.1)$$

Où π_i est la distribution des probabilités initiales des états, a_{ij} la probabilité de transiter de l'état i à l'état j, soit $a_{ij}=P(q_t=j\,|\,q_{t-1}=i)$ et $b_j(x_i)$ la fonction densité de probabilité associée à l'état j. Les fonctions densités de probabilités associées aux états déterminent le type de modèle c'est ainsi qu'on parle de HMM discrets, continus, semi continus, etc. Nous allons définir les caractéristiques du modèle de Markov caché de type continu (CHMM : Continuous Hidden Makov Model) utilisé tout au long de ce travail.

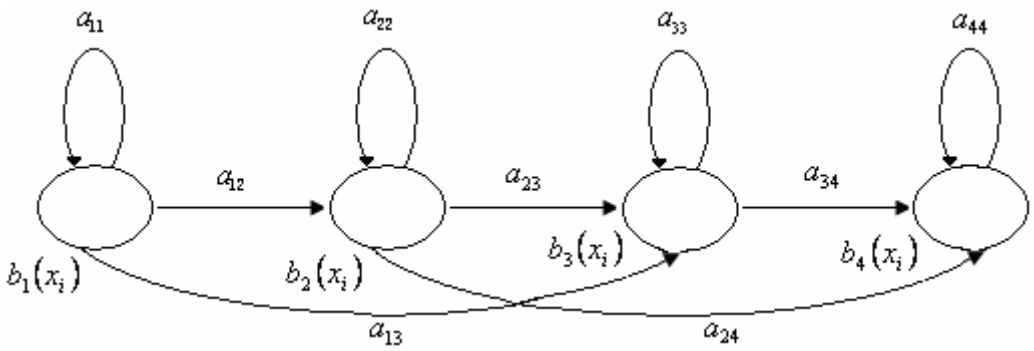

Figure 2.1 : *Un exemple de modèle de Markov caché à 4 états de type gauche droite.*

2.3. Topologie du Modèle de Markov Caché

Différentes topologies des modèles HMMs ont été testées et sont actuellement utilisées en pratique. Dans le cas de modèle de mot, ceux-ci sont souvent représentés par des modèles HMM à plusieurs états (typiquement entre 5 et 10 états), dont le nombre est parfois proportionnel au nombre de phonèmes dans le mot ou à la longueur de celui-ci.

Dans le cas de modèles de phonèmes, les models HMMs sont souvent à 3 états comme montre la *figure 2.2*. Le but initial étant d'avoir l'état central modélisant la partie stable du phonème alors que les deux états extrêmes modélisent la partie transitoire. Il a alors été observé qu'il est souvent préférable de ne pas permettre le saut d'état, de façon à également introduire une certaine contrainte sur la durée minimale de chaque unité phonétique.

Un exemple de chaîne de Markov, appelée modèle de Bakis, est présenté à la *figure 2.2*. Elle est formée d'une suite finie d'états et les séquences d'états admises sont définies par l'ensemble des transitions complétant le modèle. Dans l'exemple considéré, deux types de transitions sont possibles : le bouclage sur un état (donnant lieu à la répétition de cet état) et le passage à l'état suivant. Une séquence de transitions de l'état initial à l'état final est appelée un chemin. Une chaîne de Markov est généralement utilisée comme modèle de phonème. Un modèle de mot est alors obtenu par la concaténation des modèles des phonèmes constituant ce mot.

La topologie de type « strictement gauche-droite » d'ordre 1 dite de Bakis est utilisé tout au long de ce travail. Cette topologie est la plus adaptée à la modélisation du signal de parole [Rabiner, 1989, Boite et al.,1999].

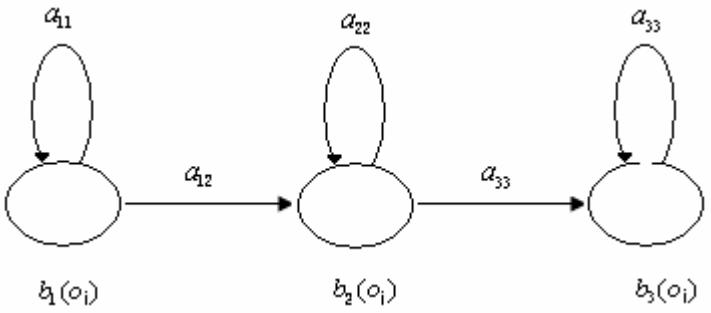

Figure 2.2 : *Automate de Bakis « gauche-droite » d'ordre 1 à trois états.*

Avec :

$b_j(o_i)$: est la probabilité d'émission des observations dans chaque état. Elle est de type gaussien dans le cas continu, définie par les vecteurs moyens, les matrices de covariance et des poids associés à chaque gaussienne.

Chaque état est caractérisé par une distribution de probabilité (par exemple la probabilité de l'observation O sachant l'état q_j : $P(O/q_j)$). Les transitions d'un état à un autre sont caractérisées par une probabilité de transition notée $P(q_j/q_i)$. L'architecture du modèle pourrait être plus générale et contenir, par exemple, les sauts d'états. Le modèle présenté ci-dessus est tout le long de notre travail.

2.4. Caractéristiques du Modèle de Markov Caché Continu

Les processus réels produisent généralement des sorties en forme de signaux. Ces derniers peuvent être de nature :

- *Discrète* : la probabilité d'émission b_i est une distribution de probabilité discrète.
- *Continue* : la probabilité d'émission b_i est une fonction de densité de probabilité définie sur R.

Pour des applications de traitement de la parole, les deux modèles ont procuré de bons résultats. Dans notre travail, le signal de parole est modélisé par une modélisation stochastique de type Modèles de Markov Cachés Continus CHMM (CHMM : Continuous Hidden Markov Model).

Nous fournirons une définition formelle du HMM. Nous emploierons la densité en continu parce qu'elle est la plus générale, et d'autres types peuvent être traité en tant que cas spéciaux. La présentation sera en termes de mélange des distributions gaussiennes en tant que probabilités des états. Le choix des mélanges gaussiens a plusieurs motivations. D'abord elles peuvent rapprocher des fonctions plus générales de densités et elles sont plus populaires pour des applications de reconnaissance de la parole.

Le modèle CHMM est constitué d'un ensemble Q de N états, tel que : $Q = \{q_i / 1 \le i \le N\}$ et q_i l'état i. Les états sont caractérisés par une première distribution (distribution initiale) $\pi = [\pi_i]$, avec $1 \le i \le N$, de probabilité et ils sont commandés par une première chaîne de Markov d'ordre 1 décrite par la matrice de transition A : matrice de la

distribution gaussienne définie sur chaque q_i d'état. Le mélange à M composants est défini par une distribution de poids $C=[c_{ik}], 1\le i\le N$, $1\le k\le M$ et un ensemble de densités de composantes gaussiennes $N(o_t,\mu_{ik},\Sigma_{ik})$ où μ_{ik} et Σ_{ik} sont le vecteur moyen et la matrice de covariance de la K^{eme} composante de la i^{eme} état. Par définition des CHMMs, $\lambda(A,C,\mu,\Sigma,\pi)$ est caractérisé par les paramètres suivants :

A : Matrice contenant les probabilités de transition d'un état i à un état j : $A=[a_{ij}]$.

C : Matrice contenant les probabilités de distribution de poids $C=[c_{ik}]$.

μ : Vecteur moyen $\mu=[\mu_{ik}]$.

Σ : Matrice covariance $\Sigma=[\Sigma_{ik}]$.

μ et Σ caractérisent la distribution de la gaussienne k définie dans tous les états q_i.

π : Vecteur contenant les probabilités d'initialisation des états $\pi=[\pi_i]$.

Sous les contraintes :

$$\begin{aligned}&\Sigma a_{ij}=1;\Sigma c_{ik}=1;\Sigma\pi_i=1.\\&a_{ij}\ge 0;c_{ik}\ge 0;\pi_i\ge 0.\\&(i,j)\in[1,N]^2;k\in[1,M].\end{aligned}\qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.2)$$

Et nous nous référons au CHMM par l'ensemble global de paramètres $\lambda(A,C,\mu,\Sigma,\pi)$.

Maintenant que nous nous référons à un état au temps t par q_t et une observation au temps t par o_t. Pour le mélange gaussien, la probabilité d'observer o_t pour $q_t=q_i$ peut être écrite comme :

$$b(o_t/q_t=q_i)=\sum_{k=1}^{M}c_{ik}N(o_t,\mu_{ki},\Sigma_{ki}) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.3)$$

Où :

$$N(o_t,\mu_{k,i},\Sigma_{k,i})=\frac{1}{(2\pi)^{\sigma/2}\left|\Sigma_{k,i}\right|^{\frac{1}{2}}}\exp\left(-(o_t-\mu_{k,i})^T\Sigma_{k,i}^{-1}(o_t-\mu_{k,i})\right) \qquad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2.4)$$

Pour une séquence des observations de longueur T $O=(o_1,o_2,...,o_T)$, et une séquence d'état $Q=(q_1,q_2,...,q_T)$, nous avons :

$$P(O/Q,\lambda)=\prod_{t=1}^{T}a_{q_{t-1}q_t}b(o_t/q_t) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.5)$$

Où : $a_{q_0q_1}=\pi_{q_1}$

Additionnant au-dessus de toutes les séquences d'état possibles :

$$P(O/\lambda)=\sum_{Q}\prod_{t=1}^{T}a_{q_{t-1}q_t}b(o_t/q_t) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2.6)$$

2.5. Problèmes liés aux modèles de Markov cachés

L'utilisation des HMMs dans un système de reconnaissance suppose de pouvoir résoudre les trois problèmes fondamentaux [Rabiner et Juang, 1993] :

- **Le problème d'*évaluation* :** Etant donné une séquence d'observation $O = \{o_1, o_2, ..., o_T\}$ et le modèle $\lambda(A, C, \mu, \Sigma, \pi)$. Comment calculer $P(O/\lambda)$, probabilité que le modèle λ génère la séquence d'observation O? Ce problème est résolu par l'application de l'algorithme de *FORWARD*.

- **Le problème de *décodage* :** Etant donné une séquence d'observation $O = \{o_1, o_2, ..., o_T\}$ et le modèle $\lambda(A, C, \mu, \Sigma, \pi)$. Qu'elle est la séquence d'états $Q = \{q_1, q_2, ..., q_T\}$ la plus probable pour un modèle et une séquence d'observation donnée ? On utilise l'algorithme de *VITERBI* pour effectuer cette tache.

- **Le problème d'*apprentissage*** : Comment peut-on ajuster les paramètres du modèle $\lambda(A, C, \mu, \Sigma, \pi)$ pour maximiser $P(O|\lambda)$ la vraisemblance (probabilité jointe) de génération d'une séquence d'observation ? Les algorithmes de *BAUM-WELCH* et de *VITERBI* permettent d'effectuer l'apprentissage.

L'ensemble de ces algorithmes permettant de résoudre les problèmes liés aux HMMs sera décrit dans cette partie qui suit.

2.6. Modèles de Markov Cachés dans la RAP

Les Modèles de Markov Cachés sont utilisés de manière intensive en reconnaissance automatique de la parole. Dans ce domaine, les signaux sont codés comme des variations temporelles de spectres de courte durée. Un HMM est un double processus stochastique dont un processus sous-jacent est non observable mais peut être estimé à partir d'un ensemble de processus qui produisent une séquence d'observations. Les HMMs peuvent être utilisés pour le traitement de problèmes dans lesquels l'information est incertaine et incomplète. Leur utilisation nécessite deux étapes :
Une étape d'apprentissage au cours de laquelle le processus stochastique est estimé à partir d'observations extensives, et une étape de reconnaissance où le modèle peut être utilisé pour obtenir les séquences de probabilités maximales. Les HMMs tirent leur succès de l'existence de nombreux algorithmes efficaces [Cerf-Danon et al., 1991]. Pour l'étape d'apprentissage nous citerons les algorithmes de Baum-Welch et les algorithmes de Viterbi qui sont devenus très populaires du fait de leur sûreté [Rabiner, 1989]. Basés sur le principe du maximum de vraisemblance et du maximum du critère a posteriori, ces algorithmes ont été étendus pour prendre en compte des fonctions de distributions de probabilités tant continues que discrètes.

De nombreux algorithmes basés sur d'autres critères sont également disponibles. Toutes ces méthodes commencent par une évaluation initiale et affinent les paramètres du modèle de façon itérative jusqu'à l'obtention d'un certain point critique (en général une solution optimale).

2.6.1. Phase d'apprentissage

Le but principal de l'apprentissage est d'optimiser le vecteur des paramètres du modèle $\lambda(A, C, \mu, \Sigma, \pi)$. Pour calculer l'estimateur de ce modèle, deux méthodes peuvent être

utilisées, la procédure de Baum-Welch, ou celle de Viterbi. En générale, on utilise la méthode de Viterbi car elle est la plus simple, moins coûteuse en mémoire et en temps de calcul et donne des résultats comparables à ceux de Baum-Welch. [Youcefi et Meziane, 2002]

2.6.1.1. Initialisation des paramètres

Comme dans tout algorithme de maximisation, le problème du choix de la valeur initiale est essentiel. Ici une façon simple d'opérer consiste à découper le signal de parole correspondant au mot à parler, en N portions égales. Ce qui consiste à supposer que les N états sont d'égales durées. Nous estimons alors, dans chaque portion, les moyennes et les variances respectives de chaque composante. Cela donne les N vecteurs moyens initiaux μ et les N matrices diagonales initiales Σ. En ce qui concerne la matrice de transition, on choisit, comme valeur initiale, une matrice de dimension $N \times N$, avec $a_{ij} = 0$ pour $i > j$ (HMM gauche-droite) et $a_{i,i-1} = a_{ii} = a_{i,i+1} = 1/3$ (équirépartition des deux états à droite). Rappelons que le vecteur de probabilité de l'état initial est supposé connu et égal à $\pi = [1,0,...,0]$.

2.6.1.2. Estimation des paramètres du modèle

Un des problèmes majeurs, mais aussi une des propriétés les plus remarquables des modèles HMM, est la possibilité de déterminer automatiquement l'ensemble de leurs paramètres sur simple présentation d'un ensemble suffisamment grand de séquences O et de leurs modèles HMM associés.

La distribution de séquences O est supposée être caractérisée par l'ensemble λ des paramètres HMM, à savoir : la matrice de probabilités de transition $A = [a_{ij}]$ de dimension $(N \times N)$, et B les paramètres de probabilités d'émission dans le cas discret. Et dans le cas continu la distribution des séquences O est caractérisée par : vecteur de probabilités d'initialisation des états $\pi = [\pi_i]$ de dimension N, vecteur de probabilités moyen $\mu = [\mu_{ik}]$ de dimension M, matrice de probabilités de covariance $\Sigma = [\Sigma_{ik}]$ de dimension $(M \times M)$, matrice de probabilités de distribution de poids $C = [c_{ik}]$ de dimension $(N \times M)$.

L'estimation selon le critère du Maximum de Vraisemblance (ML) de l'ensemble de paramètres $\lambda(A, C, \mu, \Sigma, \pi)$ ($\lambda(A, B)$ dans le cas discret) du modèle se fait alors selon :

$$\lambda^* = \arg\max_{\lambda} P(O/\lambda) \qquad \text{..... (2.7)}$$

Où O représente l'ensemble de vecteur d'entraînement (généralement repartis en plusieurs séquences d'entraînement).

Il n'y a malheureusement pas de solution analytique à ce problème, mais il existe cependant une solution itérative [Boite et all, 1999], connue sous le nom d'*algorithme EM* (Expectation-Maximization) qui garantit une convergence vers un optimum local en itérant :

- L'estimation de la fonction de vraisemblance pour l'ensemble des paramètres fixes, calculée comme l'espérance mathématique des vraisemblances (pondérées par la distribution a posteriori des variables latentes qui restent cachée mais peuvent manifester à tout instant) ;

- La maximisation de cette fonction dans l'ensemble de paramètres (maximisation).

Cet algorithme est un cas particulier d'une méthode de gradient dans laquelle le pas de gradient à chaque itération est optimal. [Boite et all, 1999].

Le principe général de l'entraînement des modèles CHMMs est basé sur l'algorithme itératif EM et peut se résumer comme suit. A partir de valeurs de paramètres $\lambda^t(A^t, C^t, \mu^t, \Sigma^t, \pi^t)$ à l'itération t, il sera nécessaire d'estimer les probabilités de transition, de vecteur moyen, de covariance et de distribution de poids, connaissant toute la séquence d'observation O, c'est-à-dire :

- ✓ La probabilité d'être sur l'état q_j à l'instant $(n-1)$ et sur l'état q_i à l'instant n :

$$P^{(t)}\left(q_i^n, q_j^{n-1} \mathrm{I} O\right) = P\left(q_i^n \mathrm{I} q_j^{n-1}, O, \lambda^{(t)}\right) P\left(q_j^{n-1} \mathrm{I} O, \lambda^{(t)}\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots \textit{(2.8)}$$

- ✓ La probabilité d'être sur l'état q_i à l'instant n :

$$P^{(t)}\left(q_i^n, q_j^{n-1} \mathrm{I} O, \lambda^{(t)}\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \textit{(2.9)}$$

Dans l'étape de maximisation, on peut montrer que les nouveaux estimateurs de paramètres $\lambda^{(t+1)}$ maximisant $P(O \mathrm{I} \lambda)$ peuvent être calculés à partir des distributions à posteriori *(2.24)* et *(2.25)*.

Pour les probabilités des transitions $A^{(t+1)}$, nous aurons alors :

$$P^{(t+1)}\left(q_i \mathrm{I} q_j\right) = \frac{\sum_{n=1}^{N} P^{(t)}\left(q_i^n, q_j^{n-1} \mathrm{I} O\right)}{\sum_{n=1}^{N} P^{(t)}\left(q_j^{n-1} \mathrm{I} O\right)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \textit{(2.10)}$$

Où la somme sur n est utilisée pour calculer la probabilité de transition de l'état q_i à l'état q_j à n'importe quel instant n. L'équation *(2.10)* représente la probabilité jointe moyenne (sur le temps) de deux états successifs divisée par la probabilité moyenne d'être passé sur l'état q_j.

Pour les probabilités d'émission, et dans le cas discret, nous aurons :

$$P^{t+1}\left(v_i / q_j\right) = \frac{\sum_{n=1}^{N} P^{(t)}\left(q_j^n, v_i / O\right)}{\sum_{n=1}^{N} P^{(t)}\left(q_j^n / O\right)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \textit{(2.11)}$$

Où les termes $P^{(t)}\left(q_j^n, v_i / O\right)$ du numérateur reprennent les probabilités $P^{(t)}\left(q_i^n / O\right)$ uniquement pour le cas où les observations O_n ont été associées à l'étiquette $v_i \in A$ lors de la quantification vectorielle.

Dans le cas d'observations continues, on suppose une forme particulière de la fonction de densité $P\left(o_n / q_j\right)$ et les observations sont directement utilisées pour estimer les paramètres

des probabilités d'émission. Par exemple, dans le cas gaussien, le nouveau estimateur de la moyenne associée à q_j et maximisant $P(O/\lambda)$ sera donné par :

$$\mu_j^{(t+1)} = \frac{\sum_{n=1}^{N} o_n P^{(t)}\left(q_j^n / O\right)}{\sum_{n=1}^{N} P^{(t)}\left(q_j^n / O\right)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(2.12)}$$

A chaque itération de l'algorithme EM, les quantités *(2.8)* et *(2.9)* sont donc calculées en utilisant les paramètres courants $\lambda^{(t)}$ (B dans le cas discret). Dans le cas particulier des modèles HMM. Le calcul ces probabilités fait intervenir des récurrences « *avant* » et « *arrière* » rapides (*Forward-Backward algorithm*). Sur la base des probabilités, les valeurs de paramètres $\lambda^{(t+1)}$ des probabilités de transition *(2.10)* et d'émission *(2.11 ou 2.12)* sont mises à jour de façon à garantir que $P\left(O/\lambda^{(t+1)}\right) > P\left(O/\lambda^{(t)}\right)$. La convergence de ce processus itératif est arrêté lorsqu'une nouvelle itération ne conduit plus à une modification significative des paramètres [Boite et all, 1999].

➢ *Algorithme de Baum-Welch*

Il s'agit d'une adaptation de l'algorithme EM (Expectation-Maximization) au sens du critère de maximum de vraisemblance (MLE, Maximum Likelihood Estimate)

On obtient les paramètres par une procédure itérative qui maximise $P(O/\lambda)$. C'est une instance de l'algorithme de EM.

Soit $A_t(i,j) = P(q_t = q_i, q_{t+1} = q_j / O, \lambda)$ la probabilité de transition de i vers j sachant l'observation O et le modèle λ.

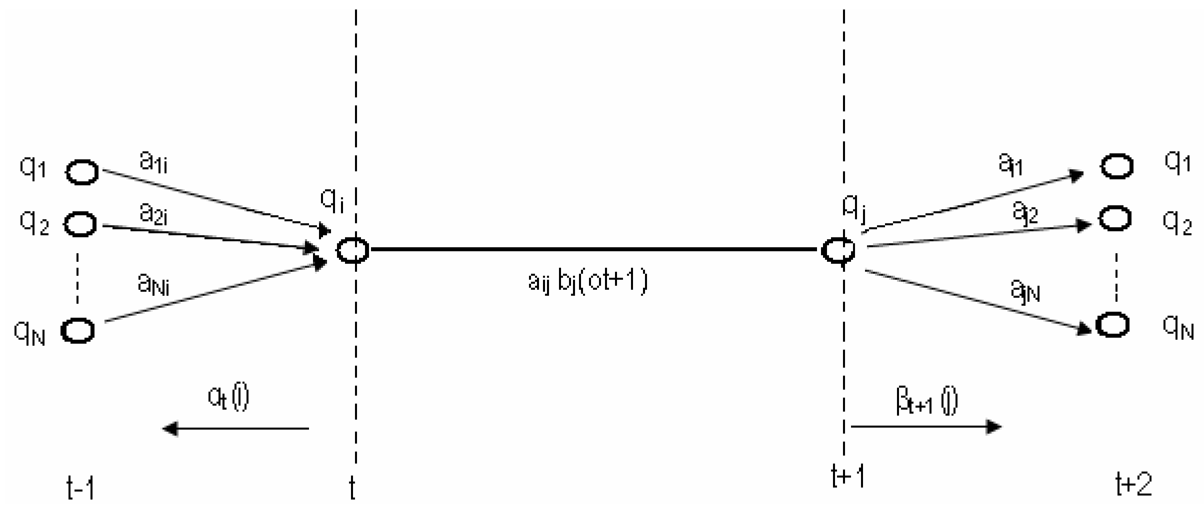

Figure : 2.3 : L'utilisation des récurrences « *avant* » et « *arrière* » pour le calcul de la distribution à posteriori des transitions.

$$A_t(i,j) = \frac{P(q_t = q_i, q_{t+1} = q_j, O/\lambda)}{P(O/\lambda)} \qquad \text{…… (2.13)}$$

$$= \frac{\alpha_t(i) a_{ij} b_j(o_t+1) \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o_t+1) \beta_{t+1}(j)} \qquad \text{……...…. (2.14)}$$

Note : $\gamma_t(i) = P(q_t = q_i / O, \lambda)$ d'où : $\gamma_t(i) = \sum_{j=1}^{N} A_t(i,j)$

$\sum_{t=1}^{T-1} \gamma_t(i)$ = Le nombre espéré de transition depuis q_i, sachant O et le modèle λ.

$\sum_{t=1}^{T-1} A_t(i,j)$ = Le nombre espéré de transitions depuis q_i Vers q_j, sachant O et le modèle λ.

Avec un peu de calcul, on peut trouver les équations de ré-estimation pour chaque paramètre :

$\overline{\pi_i}$ = Nombre espéré de fois ou au temps 1 on est en $q_i = \gamma_1(i)$.

$$\overline{a_{ij}} = \frac{\text{nb. espéré de transitions de } q_i \text{ vers } q_j}{\text{nb. espéré de transitions depuis } q_i} = \frac{\sum_{t=1}^{T-1} A_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \qquad \text{……..... (2.15)}$$

$$\overline{b_j(k)} = \frac{\text{nb. espéré de fois où on est en } q_i \text{ et en observe } v_k}{\text{nb. espéré de fois ou est dans } q_j} = \frac{\sum_{t=1;o_t=v_k}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \qquad \text{… (2.16)}$$

En 1972, Baum a démontré la convergence de cet algorithme.

La recette EM passe par le calcul d'une espérance des données jointes (la véritable observation et la variable cachée). Cette espérance est calculée sur la variable cachée, en utilisant nos estimées des paramètres à un instant donné. On recherche ensuite les paramètres qui maximisent cette espérance.

Dans le cas des modèles de Markov, la variable cachée est la séquence d'états q et la fonction auxiliaire A est :

$$A(\lambda, \lambda') = \sum_{q \in Q} P(O, q/\lambda) \log P(O, q/\lambda) \text{A} \qquad \text{……...…….. (2.17)}$$

Reste à faire les calculs de maximisation (sur λ), pour extraire nos nouvelles estimées… Pour cela posons : $Q = q_1, ..., q_T$ et $O = o_1, ..., o_T$,

$$P(O, q/\lambda) = \pi_{q_1} b_{q_1}(o_1) \prod_{t=2}^{T} a_{q_{t-1} q_t} b_{q_t}(o_t) \qquad \text{……..……. (2.18)}$$

On peut décomposer notre fonction auxiliaire en 3 termes indépendants (au regard de la maximisation).

$$A(\lambda, \lambda') = \sum_{q \in Q} (\log \pi_{q_1} P(O, q/\lambda') + \log b_1(o_1) P(O, q/\lambda'))$$

$$= \sum_{q \in Q} \left(\sum_{t=2}^{T} \log a_{q_{t-1} q_t} \right) P(O, q/\lambda') + \sum_{q \in Q} \left(\sum_{t=2}^{T} \log b_{q_t}(o_t) \right) P(O, q/\lambda') \qquad \text{…….. (2.19)}$$

Pour connaître l'estimée des π_i (les probabilités initiales), alors il suffit de dériver le premier terme (par rapport à chaque π_i) et à résoudre à 0. De même pour les autres paramètres.

Maximiser (sur π_i) le premier est équivalent à maximiser seulement $\sum_{q\in Q}\log\pi_{q_1}P(O,q/\lambda')$ qui revient à maximiser seulement $\sum_{i=1}^{N}\log\pi_i O(O,q_1=i\mathrm{I}\lambda')$.

Ne pas oublier la contrainte $\sum_{j=1}^{N}\pi_j=1$ que l'on peut intégrer dans le terme à maximiser en introduisant un multiplicateur de Lagrange (μ : multiplicateur de Lagrange) :

$$\frac{\delta}{\delta\pi_i}\left(\sum_{i=1}^{N}\log\pi_i P(O,q_1=i/\lambda')-\mu(\sum_{j=1}^{N}\pi_j-1)\right)=0 \qquad \text{…… (2.20)}$$

$$\frac{P(O,q_1=i/\lambda')}{\pi_i}-\mu=0;\forall i\in[1,N] \qquad \text{…….... (2.21)}$$

Soit :

$$\pi_i=\frac{P(O,q_1=i/\lambda')}{\mu};\forall i\in[1,N] \qquad \text{…… (2.22)}$$

Or :

$$\sum_{i=1}^{N}\pi_i=1=\sum_{i=1}^{N}\frac{P(O,q_1=i/\lambda')}{\mu}\Rightarrow\mu=\sum_{i=1}^{N}P(O,q_1=i/\lambda') \qquad \text{…….. (2.23)}$$

D'où :

$$\pi_i=\frac{P(O,q_1=i/\lambda')}{\sum_i P(O,q_1=i/\lambda')}=\gamma_1(i) \qquad \text{…… (2.24)}$$

En retombe sur notre estimée inspirée.

2.6.2. Phase de reconnaissance

La phase de reconnaissance consiste à calculer pour une suite d'observations O, soit sa vraisemblance par rapport à un modèle λ, soit la probabilité du chemin optimal l'ayant générée. Deux algorithmes ont été développés pour résoudre ses deux problèmes, l'algorithme de ForWard ou BackWard pour le calcul de vraisemblance et l'algorithme de Viterbi pour le calcul du chemin optimal.

2.6.2.1. Evaluation directe

Soit $Q=q_1q_2...q_T$ une séquence d'états pouvant expliquer O.

$$P(O/\lambda)=\sum_Q P(O,Q/\lambda)=\sum_Q P(O/Q,\lambda)P(Q/\lambda) \qquad \text{…… (2.25)}$$

Avec :

$$P(O/Q,\lambda)=\prod_{t=1}^{T}P(o_t/q_t,\lambda)=b_{q_1}(o_1)\times b_{q_2}(o_2)...b_{q_t}(o_T) \qquad \text{…….. (2.26)}$$

$$P(Q/\lambda)=\pi_{q_1}\times a_{q_1q_2}\times...\times a_{q_{T-1}q_T} \qquad \text{…….... (2.27)}$$

D'où :

$$P(O/\lambda) = \sum_{q_1 \dots q_T} \pi_{q1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \dots a_{q_{T-1} q_T} b_{q_T}(o_T) \quad \dots\dots\dots\dots (2.28)$$

Complexité : $(2 \times T - 1) \times N^T$ multiplications et $N^T - 1$ additions.
Exemple : $N = 5$ (états), $T = 100$ (observations), alors on doit faire de l'ordre de $2 \times 100 \times 5^{100} \approx 10^{72}$ options !

2.6.2.2. Algorithme ForWard (*en avant ''α''*)

Soit $\alpha_t(i) = P(o_1 \dots o_t, q_t = q_i / \lambda)$ la probabilité jointe de générer $(o_1 \dots o_t)$ et de se trouver dans l'état q_i à l'instant t.

- Init:

$$\alpha_1(i) = \pi_i b_i(o_1), \ \forall i \in [1, N] \quad \dots\dots\dots\dots (2.29)$$

- Récurrence :

$$\alpha_{t+1}(j) = \left[\sum_{i+1}^{N} \alpha_{t+1}(i) a_{ij} \right] b_j(o_{t+1}), \quad \dots\dots\dots\dots (2.30)$$

pour tout : $t \in [1, T-1]$, $j \in [1, N]$

- Terminaison :

$$P(O/\lambda) = \sum_{i=1}^{N} \alpha_T(i) \quad \dots\dots\dots\dots (2.31)$$

Soit *la figure* suivant :

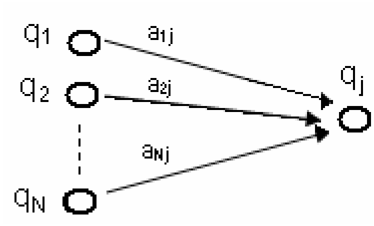

Figure 2.4 : *L'utilisation de la récurrence « avant » pour le calcul de la distribution à posteriori des transitions.*

Complexité : de l'ordre de $N^2 \times T$ opération au lieu $2 \times T \times N^T$.
Exemple : $N = 5$, $T = 100$ $\Rightarrow$ environ 3000 opérations.

2.6.2.3. Algorithme BackWard (*en arrière ''β''*)

Soit $\beta_t(i) = P(o_{t+1} \dots o_T / q_t = q_i, \lambda)$ la probabilité de générer la séquence d'observation $o_{t+1} \dots o_T$ sachant qu'on se trouvait dans l'état q_i au temps t.

- Init:

$$\beta_T(i) = 1, \forall i \in [1, N]. \quad \dots\dots\dots\dots (2.32)$$

- Récurrence:

$$\beta_T(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \quad \dots\dots\dots\dots (2.33)$$

pour tout $t \in [1, T-1]$ et pour tout $i \in [1, N]$

Soit *la figure* suivante :

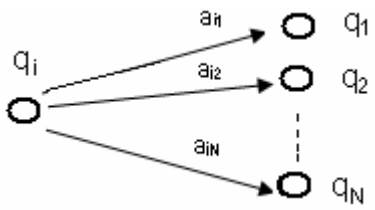

Figure 2.5 : *L'utilisation de la récurrence « arrière » pour le calcul de la distribution à posteriori des transitions.*

2.6.2.4. Algorithme de Viterbi :

On cherche à maximiser $(\text{sur}\, Q)$: $P(Q/O,\lambda)$ ce qui revient au même que de maximiser $P(O,Q/\lambda) = P(Q/O,\lambda) \times P(O/\lambda)$.

Pour cela, définissons la probabilité maximale d'une séquence au temps t qui se termine dans l'état q_i.

$$\delta_t(i) = \max_{q_1 \ldots q_{t-1}} P(q_1 q_2 \ldots q_t = q_i, o_1 \ldots o_t / \lambda) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \textit{(2.34)}$$

Par induction on a :

$$\delta_{t+1}(j) = \left[\max_i \delta_t(i) a_{ij}\right] \times b_j(o_{t+1}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \textit{(2.35)}$$

On conservons ce max pour chaque t et chaque i : $\Phi_t(j)$, on obtient l'algorithme de *Viterbi*.

- Init :

$$\delta_1(i) = \pi_i b_i(o_1) \text{ et } \Phi_1(i) = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \textit{(2.36)}$$

- Récurrence :

$$\delta_t(j) = \max_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij}\right] b_j(t) \text{ ; pour : } 2 \le t \le T \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \textit{(2.37)}$$

$$\Phi_t(j) = \arg\max_{1 \le i \le N} \left[\delta_{t-1}(i) a_{ij}\right] \text{ ; pour } \quad 1 \le t \le T \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \textit{(2.38)}$$

- Terminaison :

$$\hat{p} = \max_{1 \le i \le N} \delta_T(i) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \textit{(2.39)}$$

$$\hat{q}_T = \arg\max_{1 \le i \le N} \delta_T(i) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \textit{(2.40)}$$

- Meilleure séquence : $\hat{q}_t = \Phi_{t+1}(\hat{q}_{t+1}), t = T-1, T-2, \ldots, 1$

2.7 Conclusion

Dans ce chapitre nous avons décrit les modèles de Markov cachés ainsi que tous les algorithmes qui leur sont associés. Une attention particulière a été réservée aux modèles de Markov cachés continus. Etant donné que se sont les modèles qui ont été utilisés dans notre travail comme moteur de reconnaissance aussi bien pour la modalité vocale que pour la modalité visuelle. Dans le chapitre suivant, nous allons voir comment cette approche CHMM est appliquée dans notre système.

Chapitre 3

Fusion audiovisuelle pour la Reconnaissance Automatique de la Parole

Résumé :

Nous introduisons dans ce chapitre la reconnaissance audiovisuelle, nous décrivons ensuite les modèles de fusion audiovisuelle dans les systèmes de RAP proposés dans la littérature. Après avoir donné le schéma de fonctionnement général d'un tel système de reconnaissance audiovisuelle, nous classons ces systèmes selon le modèle cognitif développé et la méthode de fusion employée. Enfin, nous présentons le système audiovisuel mis en œuvre en explicitant chacune de ces étapes.

3.1. Introduction

Plusieurs méthodes ont été développées dans le but d'améliorer la reconnaissance automatique de la parole en milieu réel. On distingue trois principales approches :

- Le débruitage du signal.
- L'adaptation du système à l'environnement.
- L'Intégration de l'information visuelle à l'information acoustique en considérant la forme et le mouvement des lèvres.

Cette dernière approche que nous allons utiliser dans notre travail repose sur l'idée que la parole est un moyen audiovisuel de communication. En effet, le message parlé est plus intelligible quand il est accompagné de la vision de la bouche du locuteur.

3.2. Reconnaissance audiovisuelle de la parole

Dans le schéma représenté dans la *figure 3.1*, les parties d'analyse acoustique et de décodage des informations de parole se font généralement à l'aide des mêmes méthodes que dans les systèmes purement acoustiques de la RAP. Les différences des systèmes audiovisuels par rapport aux systèmes acoustiques (audio) se situent, premièrement dans la paramétrisation visuelle et deuxièmement dans l'interaction des informations acoustiques avec les informations visuelles tant au niveau de la représentation que du décodage.

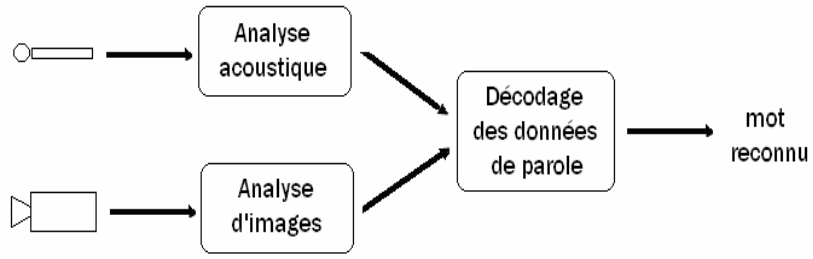

Figure 3.1 *: Schéma de fonctionnement général d'un système audiovisuel de RAP.*

La plupart des modèles de perception audiovisuelle de la parole se sont focalisés sur une interaction sensorielle de type fusion ou intégration. A ce niveau, reste posée la question du où et comment cette fusion des modalités acoustique et visuelle se passe-t-elle chez l'homme. Pour répondre à cette question, il existe plusieurs modèles cognitifs qui diffèrent de par leur lieu d'intégration des informations acoustiques et visuelles et par la manière de représentation de ces informations en vue de leur intégration. [Rogozan, 1999 ; Adjoudani, 1993]

3.3. Modèles de fusion audiovisuelle dans les systèmes de RAP

Considérant la classification par les HMMs, l'intégration des informations auditives et visuelles peut se faire de différentes manières [Potamianos et al., 2004 ; Rogozan, 1999 ; Adjoudani, 1993].
On peut classer les modèles d'intégration audiovisuelle en trois principales catégories [Robert-Ribes, 1995]. Les sections qui suivent décrivent ces quatre architectures de fusion d'informations :

- Modèle de fusion directe.
- Modèle de fusion séparée.
- Modèle de fusion immédiate.
- Modèle de fusion hybride.

3.3.1. Modèle de fusion directe

Certains systèmes audiovisuels proposent le modèle à Identification Directe (*ID*) par simple concaténation des entrées auditives et visuelles. Il s'agit d'ajouter des paramètres visuels à ceux issus de l'analyse de signal acoustique de parole au moyen d'une fusion sensorielle. C'est la manière la plus simple pour intégrer dans un système de reconnaissance les informations visuelles au moyen d'une paramétrisation étendue.

Ce modèle est directement inspiré du modèle de [Klatt, 1979]. Les deux sources d'information sont directement transmises vers un classifieur bimodal. Il n'y a donc aucun traitement avant la fusion des deux sources. La *Figure 3.2* montre le principe de ce modèle :

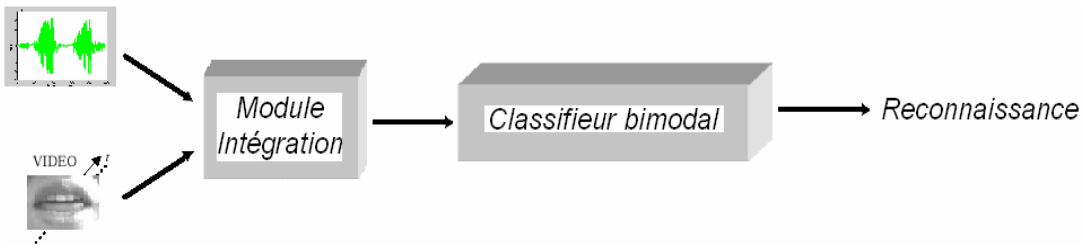

Figure 3.2 : *Schéma de principe du modèle de fusion directe (FD).*

Pour cette fusion, la part de réglages des paramètres est déterminante pour l'obtention de bonnes performances.

Par ailleurs, cette stratégie de fusion, fondée sur le modèle *ID*, nécessite une synchronisation parfaite des trames d'image sur le signal audio, lors de l'acquisition des données audiovisuelles [Rogozan, 1999].

❖ *Avantages :*

- C'est un modèle qui est facile à implémenter. Il suffit de concaténer les indices visuels à ceux du canal acoustique pour former une seule observation audiovisuelle.
- Si la taille du corpus d'apprentissage est importante, il est possible d'obtenir automatiquement le poids relatif attribué à chaque canal dans le processus de fusion [Silsbee et Su, 1996].
- La coordination temporelle entre les modalités est conservée au cours de la fusion.

❖ *Inconvénients :*

- Les systèmes fondés sur ce modèle nécessitent un corpus de plus grande taille que les systèmes utilisant le modèle à identification séparée [Jacob et Sénac, 1996].
- Pour adapter le système à différents niveaux de bruit du canal acoustique, l'apprentissage doit être effectué pour chaque niveau de dégradation [Silsbee et Su, 1996].
- La topologie des deux sources doit être identique.
- Le problème de déphasage n'est pas géré.

3.3.2. Modèle de fusion séparée

Suivant ce modèle, les informations visuelles et auditives sont traitées séparément et sont transmises chacune vers un classifieur. Les résultats en sortie de chacun de ces processus de reconnaissance sont fusionnés dans un module d'intégration qui donne le résultat final.

Dans cette configuration, chacun des systèmes de reconnaissance auditive et visuelle fournit en sortie une probabilité de ressemblance à un candidat donné. Ces informations sont ensuite utilisées pour prendre la décision finale qui peut suivre une loi bayesienne. *La Figure : 3.3* montre l'architecture du modèle d'Identification séparée (*IS*).

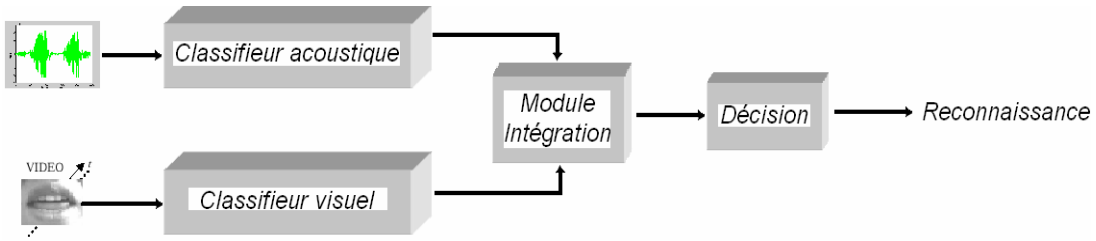

***Figure 3.3** : Architecture de la Reconnaissance Audio-Visuelle de la Parole fondée sur le modèle de fusion séparée (FS).*

❖ *Avantages :*

- Comme chaque modalité a un réseau indépendant, l'apprentissage nécessite un corpus moins important que le modèle à identification directe.
- Il est possible de gérer l'asynchronie dans le cadre d'un mot (entre son état initial et final).
- Les deux modalités n'ont pas nécessairement la même architecture de reconnaissance. La pondération des deux modalités peut être obtenue en fonction de la décision dans chaque modalité, sans apprentissage préalable sur des données bruitées [Adjoudani et Benoît, 1996].
- Cette pondération peut même être étendue aux cas où les paramètres visuels sont bruités.

❖ *Inconvénients :*

- Le bloc d'intégration peut être complexe et dépendant du corpus [Petajan et al, 1987].

3.3.3. Modèle de fusion intermédiaire

Quelques systèmes audiovisuels proposent un modèle dans lequel les informations auditives interagissent avec les informations visuelles avant catégorisation comme pour une ID, mais leurs traitements s'effectuent jusqu'à un certain niveau séparément comme pour une IS.

Ce modèle a été utilisé dans un cadre neuronal en effectuant la fusion non pas en sortie des réseaux de neurones acoustique et visuel mais au niveau des couches. Dans ces systèmes, la dépendance temporelle reste forte puisqu'il y a une synchronisation trame par trame.

3.3.4. Modèle de fusion Hybride

Le modèle hybride est une combinaison du modèle à identification directe et du modèle à identification séparée. L'intégration est faite après la classification dans chaque modalité (modèle à identification séparée) tout en prenant en compte la décision du bloc mixte dans lequel les données audiovisuelles sont traitées simultanément (modèle à identification directe). Le modèle hybride se présente donc selon la *Figure 3.4*

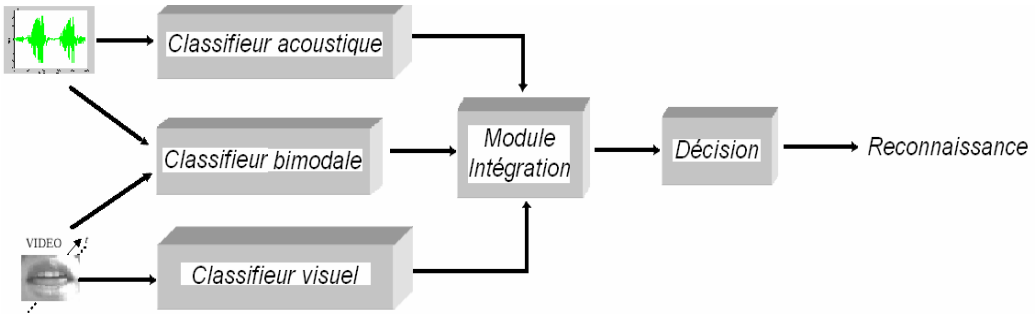

***Figure 3.4** : Architecture de la Reconnaissance Audio-Visuelle de la Parole fondée sur le modèle de fusion hybride (FH).*

L'intégration s'effectue sur les indices de sortie des blocs d'évaluation auditive, visuelle mais aussi audiovisuelle qui représentent le degré du support à un candidat. Comme dans la version antérieure de ce modèle, le bloc "décision" calcule la probabilité d'avoir la classe à reconnaître en sortie.

3.4. Techniques de fusion

L'objectif d'un système de reconnaissance audiovisuelle est de combiner aux mieux les performances de deux systèmes audio et vidéo afin d'améliorer les performances de la reconnaissance automatique de la parole, en particulier en présence de bruit. Classiquement, on distingue deux types de fusions : fusion des *paramètres* et fusion des *scores*.

3.4.1. Fusion des paramètres

Cette fusion est réalisée au moment de la paramètrisation des signaux audio et vidéo. Une fois les paramètres de chaque modalité sont extraits, les vecteurs audio $O_{a,t}$ (par exemple MFCC) et vidéo $O_{v,t}$ (par exemple la DCT) de dimension d_a et d_v respectivement, sont concaténés à chaque instant t pour ne former qu'un seul vecteur de paramètres audiovisuels $O_{av,t} = [O_{v,t}, O_{a,t}]$ de dimension $d_a + d_v$.

3.4.2. Fusion des scores

La fusion de scores ou de décision est possible lorsque l'on dispose de systèmes séparés (audio et vidéo) et que leur fusion est réalisée au moment de la décision, par combinaison de leurs scores respectifs. C'est la technique utilisée dans notre travail.

3.5. Fusion audiovisuelle pour la RAP

3.5..1. Système audiovisuel

La parole audiovisuelle propose d'étudier la forme et le mouvement des lèvres du locuteur. Le canal visuel présente l'intérêt de ne pas être soumis au bruit acoustique et il apporte une information complémentaire à celle véhiculée par l'acoustique. Donc, le but d'améliorer la RAP en milieu réel est l'intégration de l'information visuelle à l'information acoustique en considérant la forme et le mouvement des lèvres.

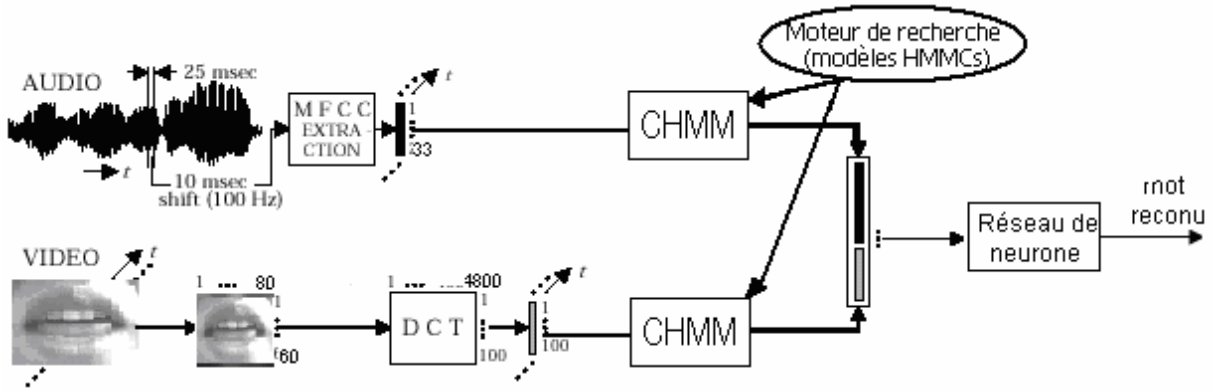

***Figure 3.5** : Description d'u n système de Reconnaissance Automatique de la Parole Audio-Visuelle.*

La plupart des études comparatives entre les approches donnent des meilleurs résultats avec le modèle d'identification séparée [Robert-Ribes et al., 1996 ;Su et Silsbee, 1996], comme le montre la figure ci-dessus.

3.5.2 Analyse des données :

3.5.2.1. Analyse des données acoustiques

Le vecteur audio est constitué des coefficients cepstraux sur l'échelle Mel (MFCCs), calculés sur des fenêtres de Hamming de 20 ms, avec un espacement de 10 ms. La première et la deuxième dérivée sont calculées pour chaque trame et sont ensuite concaténées pour constituer le vecteur acoustique. Pour le calcul des dérivées temporelles, les mêmes algorithmes que pour les fichiers de paramètres ont été appliqués.

Pour l'analyse des bancs de filtres Mel, nous avons appliqué la formule de l'équation (1.3). Des filtres triangulaires uniformément répartis sur l'échelle de Mel ont été appliqués sur la fenêtre d'analyse (voir figure : 1.5).

Nous avons utilisé la formule ci-dessous pour le calcul les coefficients cepstraux :

$$C_i = \sum_{j=1}^{P} m_j \cos\left(\frac{\pi_i}{P}(j-0.5)\right), 1 \le i \le L \qquad \text{.....(3.1)}$$

Dans laquelle m_j est la sortie du filtre j, P est l'ordre d'analyse et L est le nombre de coefficients calculés.

En tenant compte de la largeur de bande du signal acoustique et des caractéristiques spectrales de la parole, nous fixons l'ordre d'analyse P. Dans notre cas, 11 coefficients pour chaque vecteur acoustique (plus leurs dérivées première et deuxième) sont suffisants pour représenter ces informations spectrales. Pour minimiser les effets de bords, une fenêtre de Hamming de 20 ms est appliquée sur les échantillons audio avant l'analyse cepstrale. *La figure 3.6* montre l'allure de cette fenêtre.

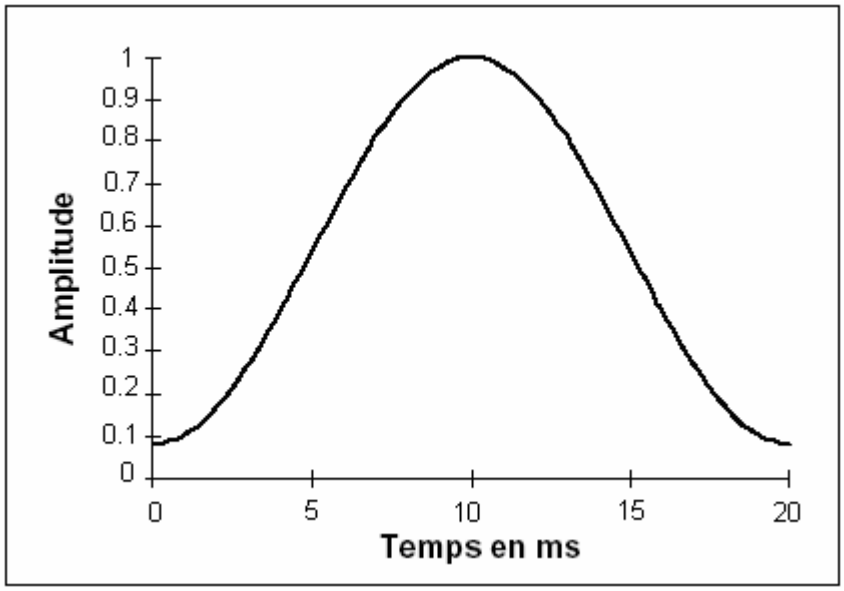

Figure 3.6 : *La fenêtre de Hamming sur 20 ms.*

Afin d'avoir le même nombre de vecteurs de coefficients cepstraux et de paramètres visuels pour l'implémentation du modèle d'identification séparée *(IS)*, nous avons fixé le décalage de la fenêtre de Hamming à 50 %, ce qui donne une analyse toutes les 10 ms. Ainsi, pour chaque seconde du signal acoustique, 50 vecteurs de coefficients cepstraux sont obtenus.

3.5.2.2 Analyse des données visuelles

Différentes techniques d'extraction de l'information visuelle peuvent être utilisées pour l'extraction des paramètres labiaux qui se divisent en deux grandes catégories [Revéret et Benoît, 1997] : les algorithmes orientés "*image*" et ceux orientés "*modèle*".

Dans les méthodes orientées *image*, les traitements s'effectuent directement sur l'ensemble des pixels de l'image. Ainsi, dans une fenêtre d'analyse qui contient les lèvres, on détecte la forme des lèvres, ce qui permet ensuite d'extraire les paramètres (exemple : la DCT).

L'approche orientée *modèle* est fondée sur des connaissances à priori sur les lèvres et leur mouvement. Un modèle de lèvres, qui prend en compte ces contraintes physiologiques est appliqué sur l'image. Grâce à une fonction d'optimisation, le modèle est ajusté au mieux sur l'image réelle des lèvres et les informations nécessaires sont extraites.

Dans notre travail nous avons consacré à l'étude de la première approches qu'on va détailler dans le *chapitre 4.*

3.5.3.3. La fusion

3.5.3.3.1 Les réseaux de neurones

Depuis une vingtaine d'année, les réseaux de neurones (type Perceptron MultiCouches ''PMC'') constituent une technique utilisée dans les systèmes de Reconnaissance Automatique de la Parole.

Leur utilisation est largement répondue dans les domaines pour résoudre des problèmes de classifications, de reconnaissance des formes (traitement d'image), etc. Les réseaux de

neurones possèdent des propriétés, très appréciées en RAP, d'associations de classification, de représentation, d'apprentissage et de généralisation.

Ils sont basés sur une modélisation grossière du neurone biologique (neurone formel). Tout comme le neurone biologique, le neurone formel calcule son activation en fonction des signaux qu'il reçoit d'autres neurones et d'une fonction d'activation plus ou moins complexe. De tels réseaux sont composés d'états appelés *neurones* et de transitions appelées *synapses*.

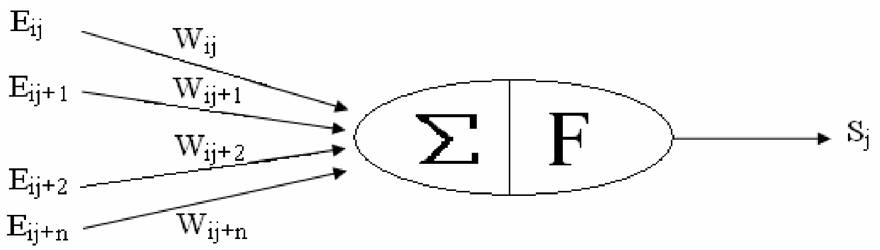

Figure 3.7 : *Schéma d'un neurone formel.*

a- Fonctionnement d'un neurone

Chaque neurone a pour tache la propagation d'un paramètre S_j. Le paramètre est la somme penderée des paramètres envoyés par les neurones auxquels il est connecté.

$$S_j = F\left(\sum_i \omega_{ij} \times E_j\right) \times T_j \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(3.2)}$$

Où

F : est une fonction généralement sigmoïde.

Et

T_j : un seuil.

Chaque neurone a un seuil propre T_i, de réponse à un paramètre et à chaque synapse est associé un poids ω_{ij}

b- Fonctionnement d'un réseau de neurones à plusieurs niveaux

Les réseaux multicouches sont composés :

- ✓ d'une couche d'entrée.
- ✓ d'une ou plusieurs couches dites ‘’ cachées ‘’.
- ✓ d'une couche de sortie.
- ✓

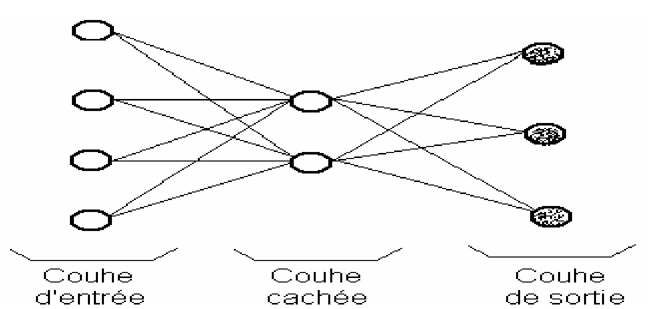

Figure 3.8: *Schéma général d'un réseau multicouche.*

3.5.3.3.2 Fusion par réseaux de neurones

a- Phase d'apprentissage

La phase d'apprentissage est généralement réalisée à partir de l'algorithme de retropropagation du gradient qui consiste à comparer les résultats de la couche de sortie avec les résultats hypothéqués, en calculant la valeur de l'erreur et en la minimisant. Cette erreur est retropropagée dans le réseau de façon à ajuster les poids et les seuils.

Evaluation et ajustement sont répétés jusqu'à l'obtention des résultats escomptés ou stabilité des résultats obtenus. L'apprentissage est fait par présentation unique ou répétée de chaque observation, de façon supervisée ou non [Austin, 92].

b- Phase de reconnaissance

La reconnaissance consiste simplement à injecter les observations dans le réseau. Elles sont propagées jusqu'à la couche de sortie donnant le résultat. Généralement le résultat est :

- soit *direct* : le neurone en sortie donnant le meilleur score correspond à la forme reconnue.
- soit *indirect* : l'ensemble des sorties rend compte d'une forme test qui va être comparée à toutes les formes ou références apprises.

Dans la reconnaissance de phonèmes, ils donnent de bons résultats. Leur très grand pouvoir de discrimination permet de distinguer des phonèmes ayant des comportements acoustiques très proches [Bourlard, 92].

Leur principal inconvénient est qu'ils nécessitent un apprentissage *supervisé* avec un volume de données très important et traitent assez mal la dimension temporelle de la parole.

3.6. Conclusion

Nous avons décrit dans ce chapitre les différents modèles développés des systèmes de Reconnaissance Automatique de la Parole Audio-Visuelle proposés dans la littérature. Après avoir donné le schéma de fonctionnement général d'un tel système, nous avons présenté les techniques utilisées pour paramétrer les deux signaux de parole audio et vidéo. Enfin nous avons présenté la méthode de reconnaissance employée que nous avons choisi pour la fusion des deux signaux audio et vidéo.

Chapitre 4

Expériences et résultats

Résumé :

Dans ce chapitre il s'agit de présenter l'évaluation du système de reconnaissance audiovisuel mis en œuvre. Dans un premier temps, nous parlerons de la base de données audiovisuelle élaborée, les traitements acoustiques et visuels utilisés dans le système développé sont également décrits. Il est clair que pour concevoir un système de reconnaissance audiovisuel capable de fusionner au mieux les informations acoustique et visuelles, une base de données est construite. Ensuite, nous expérimentons notre modèle d'intégration selon que celle-ci intervient dans le système de RAP par identification séparée.

4.1. Introduction

La reconnaissance de la parole audiovisuelle nécessite l'élaboration d'une base de données spécifique (audio + vidéo), d'un moteur de reconnaissance et d'une méthode de fusion. C'est ainsi, que nous avons choisi comme approche de reconnaissance l'approche stochastique basée sur les modèles de Markov cachés continus. Des expériences ont été menées afin de trouver le nombre de mixture optimal pour notre moteur de reconnaissance Après le choix des paramètres acoustique et visuels, des tests de reconnaissance dans le mode mono locuteur en utilisant une fusion sur les scores par une approche neuronale de type PMC ont été réalisées.

4.2. Objectifs

Notre travail consiste à identifier l'architecture d'intégration audiovisuelle parmi les modèles présentés dans la littérature. Dans ce sens, le contexte de notre travail a été limité à la reconnaissance audiovisuelle des chiffres arabes de zéro à neuf prononcés par une seule locutrice.

Pour l'ensemble de nos expériences, nous avons utilisé, l'outil de programmation, *''MATLAB''*.

En tenant compte de la tâche de reconnaissance et de la durée moyenne de chaque mot, nous avons utilisé un modèle de Markov caché continu à trois états émetteurs pour toutes nos

expériences. Chaque mot est ainsi représenté par un modèle de Markov à mélange de gaussiennes représentant la distribution de probabilité des observations

4.3. Protocole expérimental

4.3.1. Caractéristiques du système

Le schéma général du système implémenté est donné par la *figure 4.1 :*

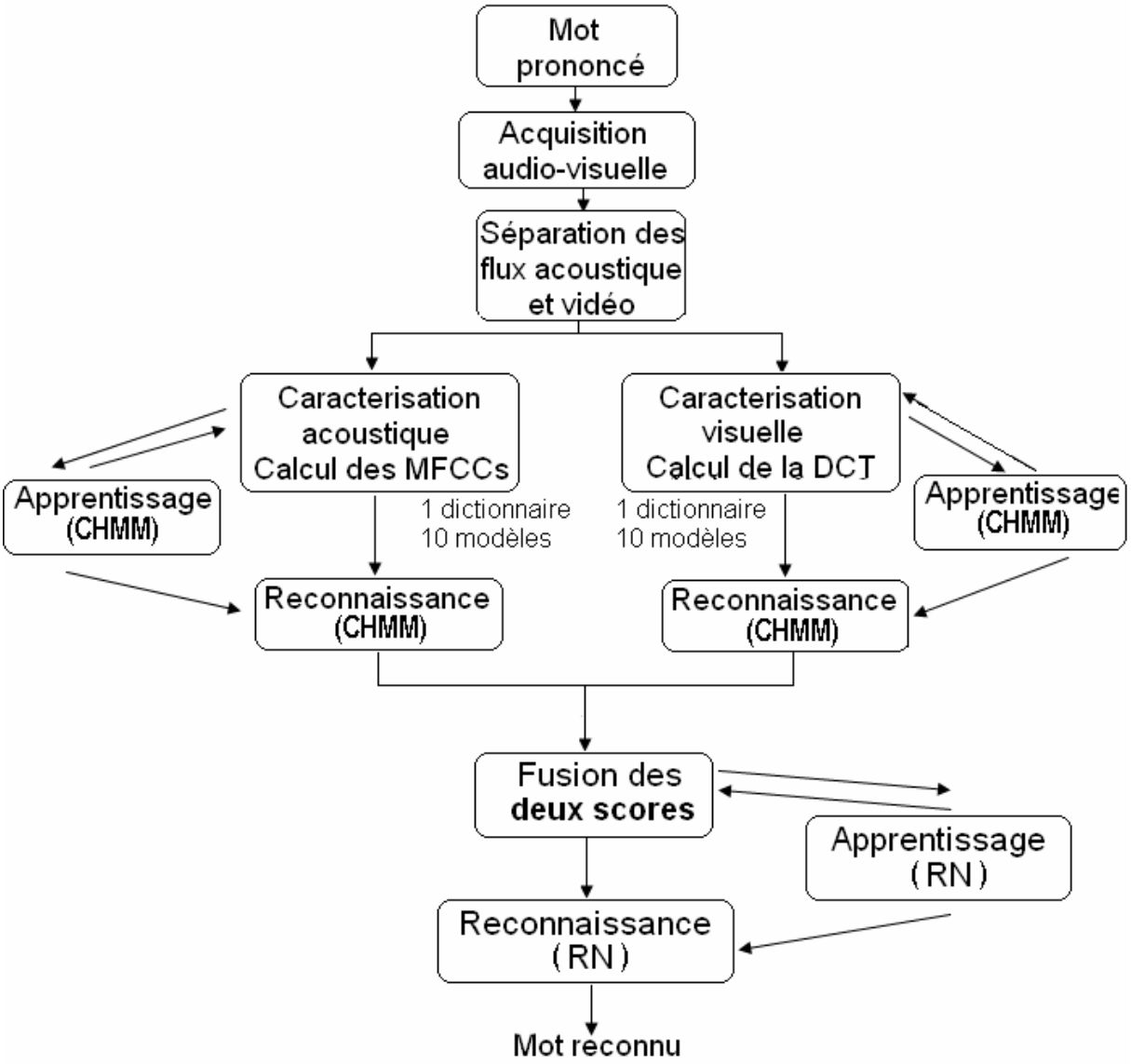

Figure 4.1 : *Schéma synoptique du système implémenté.*

4.3.2. Nature de la base de données

Dans le cadre de notre travail, il est clair que pour concevoir un système de reconnaissance audiovisuelle capable d'intégrer au mieux les informations auditives et visuelles, une base de données réelle est réalisable.

La base de données utilisée est une base audiovisuelle comportant des chiffres arabes isolées prélevés à une fréquence d'échantillonnage de 16 KHz. Elle est constituée de 25 répétitions des 10 mots (*siffer, wahed, ithnani, thalatha, arbaa, khamssa, sitta, sabaa, thamania, tissaa*) prononcés par une seule locutrice arabisante. Ainsi, la base construite est une base monolocuteur. Elle a été apprise sur le mot à reconnaître pour le même style de corpus de parole. A savoir des séries de dix mots des chiffres arabes tirées aléatoirement, sans répétition, appelées en élocution continue. L'utilisation d'un modèle monolocuteur avait pour but de minimiser des différentes déformations liées au changement de locuteur.
Pour nos expériences, nous avons construit cette base de données audiovisuelle au sein de notre laboratoire (LCPTS). Les vidéos sont enregistrées à 25 images/s et l'audio à 16 KHz. Pour extraire la région d'intérêt, la région de la bouche est détectée de sorte que toutes les bouches ont approximativement la même taille et orientation.

4.3.3. Prétraitement des données audiovisuelles

4.3.3.1. Acquisition du signal de parole

Notre objectif principal pour le corpus étant qu'il correspond à des conditions naturelles, nous n'avons donc pas utilisé d'éclairage artificiel complémentaire pour compenser une éventuelle sous-exposition du locuteur. Nous avons effectué les enregistrements lors de journées ou le soleil était visible, pour éviter d'avoir une luminosité trop faible. La camera était située à coté d'une fenêtre, à une distance d'environ un demi mètre du locuteur, à leur hauteur pour qu'ils n'aient ni besoin de relever la tête vers l'arrière, ni de la pencher vers l'avant. On peut d'ailleurs remarquer que le plus grand corpus de parole audiovisuelle actuel [Neti et all., 2000] a également été enregistré dans ce type de condition (camera vidéo distante de locuteur et prise de vue de face, mais éclairage artificiel).

Nous avons conçu ce corpus dans le but de l'utiliser pour la RAP Audio-Visuelle, en ne prévoyant pas son utilisation pour la reconnaissance du locuteur, cependant comme le montre les expériences de Jourlin et Luetin [Jourlin et all., 1997a], qui n'utilise que la zone des lèvres, il serait également envisageable d'utiliser notre corpus dans ce but. Nous avons focalisé notre attention sur la zone des lèvres, et non sur le visage dans son ensemble.

La *figure* suivante présente le schéma synoptique du banc d'acquisition :

Figure 4.2 : *Schéma synoptique du banc d'acquisition.*

Le système d'acquisition de la parole audiovisuelle se compose d'un ensemble comprenant :

a - Une camera Webcam :

C'est une caméra numérique, se branche sur un port USB 1 d'un ordinateur. Cette caméra numérique constitue la façon la plus simple d'enregistrer des films ou des images statiques de haute qualité de manière très simple.

b - Un ordinateur :

Pour l'enregistrement de la base de données, la camera Webcam est reliée à une carte son intégré insérée dans un micro-ordinateur compatible ''Intel P4 CPU 1.6 GHz de RAM de 256Mo DDL''.

c – Les logiciels:

- ***Windows movie maker :***
  Ce type de logiciel permet l'acquisition et la séparation des séquences vidéo.
- ***Gold Wave :***
  Il permet la séparation audio de la vidéo.
- ***BPS Vidéo Converter & Decompiler :***
  Ce type de logiciel permet la conversion des fichiers vidéo AVI, ASF, WMV, RM et MPEG ; et aussi la séparation des séquences vidéo.

4.3.3.2. Le système de séparation audiovisuelle :

Une fois l'enregistrement des séquences vidéo du locuteur est réalisé à l'aide du logiciel *Windows Movie Maker*, avec l'extension *''.wmv''*, la première opération consiste à convertir les séquences vidéo de l'extension *''.wmv''* vers l'extension *''.avi''* avec le logiciel *BPS (Vidéo Converter & Decompiler)*. Puis, on passe à la séparation des deux flux audio et vidéo. On va extraire le flux audio sous forme d'un signal, a l'aide du logiciel *Gold Wave*, avec l'extension *''.wave''* d'une part ; et d'extraire à partir du flux vidéo des images fixes de la séquence, à l'aide du logiciel *BPS Vidéo Converter & Decompiler*, d'autre part.

La deuxième opération consiste à construire la base de données audio et vidéo, d'où l'organigramme suivant, qui présente les étapes nécessaires pour construire les deux fichiers audio/vidéo :

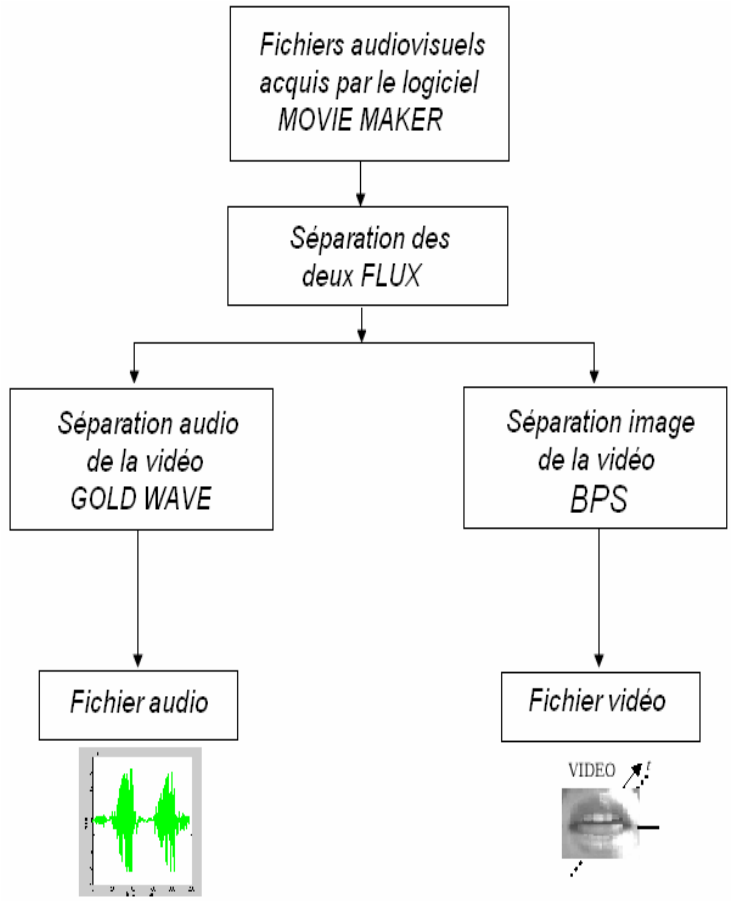

Figure 4.3 : *L'organigramme de la séparation audio / vidéo.*

4.3.3.3. Extraction d'images fixes à partir de la vidéo

Une fois le flux audio/vidéo est réalisé, le flux vidéo est compressé avec une résolution de 80×60. L'extraction des images fixes est réalisée en décomposant le signal vidéo en trames successives, ces trames constituent les images de résolution 80×60 qui vont représenter notre base de données images.

Dans notre corpus, la taille du fichier vidéo est moyenne de l'ordre de 25 images /secondes. Les images obtenues sont sauvegardées au format *''.jpeg''*.

4.3.3.4. Prétraitement acoustique

Pour chaque fichier de paramètres visuels, le fichier audio correspondant a été numérisé sur 16 bits et sur un seul canal, à une fréquence d'échantillonnage Fe = 16 KHz. Cette configuration a permis d'avoir une bonne qualité acoustique et une bande passante assez large.

Une fois numérise, le signal subit une opération de préaccentuation, nous utilisons simplement un filtre de réponse impulsionnelle finie $(1, a)$ avec $a = 0.95$. Si $s(n)$ désigne le signal de parole et $s_p(n)$ le signal pré-accentué, on a :

$$s_p(n) = s(n) - 0.95 \times s(n-1) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \textit{(4.1)}$$

D'autre part, nous avons à traiter les données sur des trames de taille fixes consécutives. Cette façon de procéder revient à appliquer une fenêtre de Hamming glissante de durée finie sur l'ensemble du signal. Cette opération donne la trame fenêtrée :

$$S_w(n) = s_p(n)W(n) \qquad \text{…….....……..…….. } (4.2)$$
$$W(n) = 0.54 - 0.46\cos(2n/(N-1))$$
$$\text{avec } 0 \le n \le N-1$$

La longueur N d'une trame est choisie d'une façon à avoir des trames dont la durée est de l'ordre de 20ms. Enfin l'opération de découpage en trame de longueur N comporte un recouvrement de 50% entre deux trames successifs. En général, les fenêtres d'analyse successives se superposent de façon à avoir le maximum d'information du signal, résultant généralement en un vecteur acoustique toutes les 10 ms. En conséquence, pour un mot de durée de 1s (comme par exemple la liste des chiffres arabes dans notre cas : *siffer, wahed,…, tissaa*), cela donne entre 40 à 50 trames à traiter par mot.

Ce vecteur acoustique est souvent obtenu en étendant les coefficients (X) à leurs dérivées (temporelles) premières et secondes *(4.3)*. Ces paramètres sont souvent appelés *coefficients delta* et *coefficients delta delta* et peuvent être estimés selon :

$$\Delta x_n = \frac{1}{\sum_{k=-c}^{+c} k^2} \sum_{k=-c}^{c} k x_{n+k} \qquad \text{……..….. } (4.3)$$

Où X_n : représente le vecteur acoustique à l'instant n .

Les paramètres les plus couramment adoptés pour l'analyse acoustique en vue de la reconnaissance sont les cœfficient cepstraux dans l'échelle Mel MFCCs (Mel Frequency Cepstral Coefficient). Cette analyse cepstrale est définie comme la transformée de Fourier du spectre logarithmique, calculée à partir d'un spectre non uniforme espacé selon l'échelle « Mel » correspondant aux bandes critiques du système auditif. La motivation de cette représentation « Mel » est de tenir compte de certaines propriétés de l'oreille humaine qui traite les sons selon une échelle de fréquence **non** uniforme [Boite et all, 1999]**.**
Nous utilisons comme paramètres acoustiques onze coefficients MFCCs ainsi que leurs vitesses et accélérations.

4.3.3.5. Prétraitement visuels (La DCT)

Pour notre cas, nous nous intéressons à l'extraction des paramètres visuels sur des images que nous qualifions de ''naturelles'', c'est-à-dire acquises dans des conditions que nous pensons réalistes, sans utilisation d'artefacts (maquillage ou pastilles réfléchissantes) ; ni prise de vue ou utilisation de dispositif d'acquisition spécifique et sans contrôle particulier sur les conditions d'éclairage (utilisation de la lumière solaire ambiante). Nous cherchons tout d'abord à valider les modèles que nous avons construit pour extraire les paramètres, nous nous intéressons également aux paramètres obtenus à l'aide de ces modèles pour déterminer quel peut être leur apport dans le cadre de le RAP Audio-Visuelle.

Pour caractériser les signaux vidéo nous utilisons la DCT (Discrete Cosine Transform), ou transformé en cosinus, qui permet de représenter une image de dimension $m \times n$ en une transformée de dimension égale, dont les coefficients sont classés dans l'ordre croissant.

Définition de la DCT :

La DCT, permet de représenter une image de dimension $m \times n$ en une transformée de dimension égale. Cette technique est très utilisée dans les codeurs d'image ou vidéo de type JPEG et MPEG [BOVIK 2000]. La DCT est semblable à la transformée de Fourier, car elle transpose le domaine temporel dans le domaine fréquentiel. Par contre, contrairement à la transformée de Fourier, elle comporte seulement les coefficients réels.

La transformée en cosinus en deux dimensions est définie par la relation :

$$H(u,v) = \frac{2}{\sqrt{MN}} C(u)C(v) \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} h(x,y) \cos\left[\frac{(2x+1)u\pi}{2M}\right] \cos\left[\frac{(2y+1)v\pi}{2N}\right] \quad \text{…….. } (4.4)$$

La transformée inverse (IDCT) est exprimée par la relation *4.5* et elle est souvent utile pour vérifier si l'implémentation logicielle de la DCT est correcte. Dans ce cas, la transformée inverse (IDCT) des coefficients obtenus par la DCT doivent redonner l'image originale (*Figure 4.5*).

$$h(x,y) = \frac{2}{\sqrt{MN}} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} C(u)C(v)H(u,v) \cos\left[\frac{(2x+1)u\pi}{2N}\right] \cos\left[\frac{(2y+1)v\pi}{2N}\right] \quad \text{……. } (4.5)$$

$$C(\lambda) = \begin{cases} \frac{1}{\sqrt{2}} pour \lambda = 0 \\ 1 \, pour \lambda > 0 \end{cases} \quad \text{……...…… } (4.6)$$

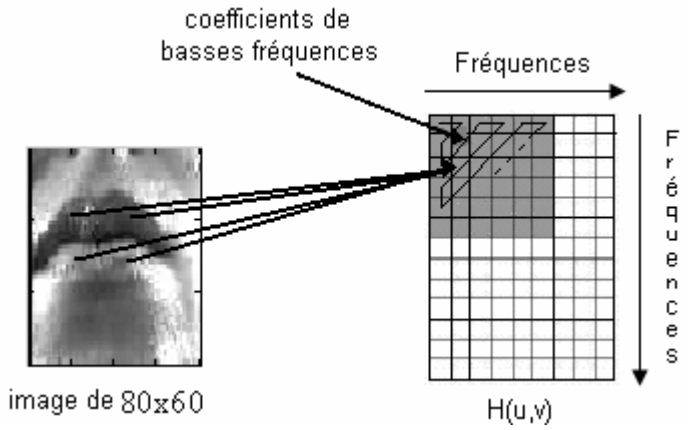

Figure 4.4 : *Représentation graphique de la transformée en cosinus (DCT)*

Les vecteurs d'entrées sont formés des coefficients basses fréquences qui se trouvent dans le coin supérieur gauche de la matrice résultante, comme montré la *figure : 4.4*. Dans notre cas, l'image qui doit être transformée dépend du maximum des amplitudes supérieures par valeurs absolues, comme montre les résultats de nos expériences (voir annexe). En effet, la DCT est effectuée sur la région de couleur noire, de dimension variable. *La figure : 4.4* montre un symbole de dimension 80x60, où sont conservés seulement 100 coefficients de basses fréquences ou de hautes amplitudes dans notre cas.

Le nombre de coefficients de basses fréquences (hautes amplitudes) conservés après la transformation par la DCT est choisi de manière à conserver un maximum d'énergie totale dans les coefficients de basses fréquences (hautes amplitudes) qui sera suffisant pour reconstituer les caractéristiques principales de l'image. L'énergie totale E de l'image est calculée (théorème de Parseval) à partir des coefficients de la DCT par la relation suivante :

$$E = \sum_{u=0}^{M-1}\sum_{v=0}^{N-1}\left|H(u,v)\right|^2 \qquad \text{..... (4.7)}$$

Figure : 4.5 : *Reconstitution d'une image à partir de 100 coefficients de hautes amplitudes de dimension* 80 x 60

La *figure : 4.5* montre la reconstitution de l'image à partir des coefficients de hautes amplitudes de dimension 80 x 60. Dans ce cas, nous pouvons voir que conserver la totalité de l'énergie avec les 100 coefficients nous donne une image très semblable à l'image originale. L'idée principale de l'algorithme pour encoder l'image par la DCT est de ne pas utiliser 100% des coefficients, pour limiter la taille mémoire et les calculs nécessaires pour l'entraînement et la reconnaissance par les modèles de Markov cachés continus CHMM. Par ailleurs, il serait possibles d'utiliser moins de coefficients de la DCT (donc mois d'énergie) pour représenter l'image et obtenir de bon résultats. Et pour cela, nous avons pris 100 coefficients.

4.3.3.6. Algorithmes de prétraitement

4.3.3.6.1. Algorithme de l'analyse du signal acoustique

```
Initialisation :
        Trame = 0 ;      % Nombre de trames
        Ouverture du fichier signal ''.wave''
Début
    Tant que non fin fichier (.wave)
    Faire
        Fenêtrage [Trame]. % prendre une fenêtre du signal
        Calcul des coefficients cepstraux MFCCs : MFCC[trame](i).
        Calcul de la première dérivée des MFCCs : ΔMFCC[trame](i).
        Calcul de la deuxième dérivée des MFCCs : ΔΔMFCC[trame](i).
        Trame++.
    Fait++
FIN
```

4.3.3.6.2. Algorithme de l'analyse du signal visuel

```
Initialisation
        Image = 0 ; % Nombre d'image
        Ouverture du fichier vidéo ''.avi''      % prendre un clip vidéo.
Début
    Tant que non fin fichier (.avi)
    Faire
        Ouverture du fichier image ''.jpeg''.    % prendre une image de la vidéo.
        Calcul de la DCT.
        Image ++.
    Fait++
FIN
```

4.3.4. Implémentation du Modèle de Markov Cachée Continu

4.3.4.1. Topologie du Modèle de Markov Caché Continu utilisé

Dans le cas de modèles de mots, les modèles HMM sont souvent à 3 états, comme la montre la *figure : 2.2 (voir section : 2.3)*. Le but étant d'avoir l'état central modélisant la partie stable du mot alors que les deux états extrêmes modélisent la partie transitoire. Il a alors été observé qu'il est souvent préférable de ne pas permettre le saut d'état.

La topologie employée dans notre cas est celle d'un modèle gauche-droite d'ordre 1 à trois états émetteurs dit de Bakis, c'est le modèle le plus adapté à la modélisation du signal de la parole [Rabiner, 1989; Boite et all, 1999]

Chaque état est caractérisé par une distribution de probabilité (par exemple $P(O/q_j)$). Les transitions d'un état à un autre sont caractérisées par une probabilité de transition notée $P(q_j/q_i)$. L'architecture du modèle pourrait être plus générale et contenir, par exemple, les sauts d'états. Le modèle présenté ci-dessus est cependant celui souvent utilisé pour la modélisation d'unités phonétiques (phonème, mot, etc.).

4.3.4.2. Apprentissage par CHMM

L'apprentissage des modèles acoustiques et visuels se fait par estimation de leurs paramètres sur un corpus dit *''apprentissage''* qui doit être disjoint du corpus de *''test''*. Nous avons utilisé 40% de notre base de donnée pour l'apprentissage et 60% restant pour le test.

Plus précisément, l'apprentissage a été effectué selon les étapes suivantes :

- Lors de la première étape, les valeurs initiales des moyennes et des variances des distributions de chacun des modèles, sont estimées. Ces valeurs, obtenues à l'issue d'un processus itératif, maximisent la probabilité d'émission des exemples extraits du corpus d'apprentissage. Pour chacun de ces exemples, la suite d'observations est alignée sur les états du modèle correspondant au moyen de l'algorithme de Viterbi [Viterbi, 1967].

- Dans une deuxième étape, ces valeurs sont optimisées en tenant compte de tous les alignements possibles de chacune des suites d'observations acoustiques (et les observations visuelles) sur les états du modèle correspondant au moyen de l'algorithme de BaumWelch

[Baum-Welch, 1972]. Les valeurs de probabilités de transition d'un modèle au moyen de mesures statistiques sur l'utilisation des transitions sont aussi estimées.

- Lors de la troisième étape, les paramètres des modèles sont optimisés de manière itérative afin de maximiser la probabilité d'émission de l'ensemble d'observations du corpus d'apprentissage. Les paramètres des modèles utilisés pour chacun des mots d'apprentissage sont réestimés simultanément au moyen de l'algorithme de BaumWelch sans tenir compte des frontières des unités de parole qui la composent.

- Enfin, l'apprentissage des modèles est complété par des informations statistiques de chaque mot et d'une loi de probabilité sur cette durée. Cette dernière est exprimée au moyen d'une fonction gaussienne inverse par la formule *(2.3)*.

Dans laquelle o_t représente la séquence d'observation, les paramètres μ et Σ représentent successivement la moyenne et la covariance. Pour chaque modèle, les estimations statistiques sont effectuées, suivant la taille du mot, 40 à 50 trames par mot. L'estimation des paramètres, fondée sur la maximisation de la vraisemblance, est effectuée au moyen des formules suivantes :

$$\mu = \bar{x} \qquad \text{…….. (4.8)}$$

$$\Sigma = \frac{n}{\sum_{i=1}^{n} x_i^{-1} - \left(\bar{x}\right)^{-1}} \qquad \text{……. (4.9)}$$

Cette loi de probabilité a l'avantage d'être positive, vérifiant ainsi le fait d'exprimer une durée. De plus, l'estimation des paramètres μ et Σ est relativement aisée.

Puisque les informations statistiques sur les contraintes temporelles n'ont pas été incluses dans les modèles au niveau des états, les algorithmes de décodage classiques doivent être modifiés pour en tenir compte.

4.3.4.2.1. Construction du dictionnaire de données audiovisuelles

Après analyse des deux signaux audio et vidéo sur l'ensemble des mots, on a obtenu trois fichiers : de vecteurs acoustiques *MFCC*, de vecteurs visuels *DCT* et de vecteurs audiovisuels *DONNEES*.

Ce dictionnaire est donc composé des trois fichiers *MFCC, DCT* et *DONNEES :*

- Le fichier *MFCC* est composé de 33 coefficients cepstraux (11 coefficients cepstraux et leurs dérivées premières et secondes)
- Le fichier *DCT* est composé de 100 coefficients de la DCT

- Le fichier *DONNEE* lui-même est composé de 20 éléments, chaque élément est composé des paramètres d'un modèle $CHMM_{ij}$, ou i est la donnée audio ou vidéo, et j est le chiffre (0,…,9).

4.3.4.2.2. Modèle initial

Les matrices initiales A et Σ, et le vecteur initial π sont stochastiques définis comme suite :

$$\Sigma = \begin{bmatrix} 1 & 0 & . & . & 0 \\ 0 & 1 & 0 & . & . \\ . & 0 & 1 & 0 & . \\ . & . & 0 & 1 & 0 \\ 0 & . & . & 0 & 1 \end{bmatrix} ; \; A = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \end{bmatrix} ; \; \pi = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

4.3.4.2.3. Algorithme d'estimation des modèles

Entrée :

- Un dictionnaire de référence.
- Modèle initial $\lambda_0(A_0, C_0, \mu_0, \Sigma_0, \pi_0)$.

Sortie :

- 10 modèles $\lambda_k(A_k, C_k, \mu_k, \Sigma_k, \pi_k)$ reéstimés.

Initialisation :

- Maxmodel = 10 *% nombre totale de modèles.*
- k = 1. *% nombre de modèles.*

Début
Tant que (k < Maxmodel)
Faire

- Initialisation du modèle $\lambda_k(A_k, C_k, \mu_k, \Sigma_k, \pi_k)$.
- Ouverture de fichier de dictionnaire. *% un mot*
- Lire une suite de vecteurs acoustiques ou visuels. *% MFCC ou DCT*
- Lancer l'algorithme de Baum-Welch pour le calcul de $\lambda_k(A_k, C_k, \mu_k, \Sigma_k, \pi_k)$.
- Sauvegarde du modèle $\lambda_k(A_k, C_k, \mu_k, \Sigma_k, \pi_k)$ dans le fichier *MODEL*.

Fait.
FIN

4.3.4.3. Reconnaissance par CHMM

La reconnaissance se fait par les étapes suivantes :

- Le locuteur prononce un mot qui contient plusieurs phonèmes, l'élocution est enregistrée dans un fichier ''*.avi*''.
- Le mot d'élocution est séparé en deux signaux différents : signal *audio* ''*.wave*'' et signal *vidéo* ''*.jpeg*''.
- Analyse des deux signaux acoustique par trame et visuel par image.
- Application de l'algorithme de reconnaissance.

Algorithme de reconnaissance

Entrée :

Mots d'élocution du signal audio ou vidéo inconnus.

10 fichiers *MODEL* contiennent les modèles calculés. *% modèles des chiffres arabes (0 à 9)*

Dictionnaire de référence.

Sortie :

Mots reconnus. *% chiffre à reconnaître*.

Début

Ouverture du dictionnaire de référence.

Tant que nbmot *% nombre de mot à reconnaître.*

Faire

- Prendre le fichier audio (ou vidéo) du mot.
- Appel à l'algorithme de l'analyse du signal audio (ou vidéo).
- Sauvegarde des vecteurs acoustiques (ou visuels) issus de l'analyse du signal dans un fichier *COEFF* (ou *DCT*).
- Tant que non fin de fichier (*MODEL*)
  Faire
  - Lire le modèle $\lambda_k(A_k, C_k, \mu_k, \Sigma_k, \pi_k)$
  - Application de la procédure Forward *% Pour le calcul de la probabilité de production de la séquence d'observations O par le modèle* λ_k, $P(O/\lambda_k)$.
  - Application de l'algorithme de Viterbi. *% Pour trouver le chemin optimal parcouru par la séquence d'observations.*

  Fait
- Afficher le mot reconnu. *% le mot reconnu est le mot qui correspond au modèle* λ_k *tel que* $P(O/\lambda_k)$ *soit maximale.*

Fait

Fin

4.4. Expériences et résultats

Les expériences seront présentées sous forme de figures ou matrices de confusion indiquant chaque fois le ''Taux de Bonne Reconnaissance''

4.4.1. Evaluation des résultats

4.4.1.1. Critère d'évaluation : ''*matrice de confusion''*

4.4.1.1.1. Définition

Un excellent critère d'évaluation de la reconnaissance est la matrice de *confusion*. Cette matrice permet d'estimer non seulement les taux de bonne reconnaissance, mais également les erreurs d'inclusion ou d'exclusion. Il s'agit d'un tableau indiquant le nombre de mots affectés à chaque modèle. Nous choisissons arbitrairement de disposer en ligne la référence et en colonne le résultat de reconnaissance.

Les tableaux représentés ci-dessous présentent les résultats de la matrice de confusion pour les taux de reconnaissance acoustique et visuel globaux.

4.4.1.1.2. Normalisation de la matrice de confusion :

Il existe deux types de normalisations :

- *Normalisation en lignes* :

La démarche traditionnelle propose une normalisation en lignes, c'est relativement aux nombres totaux de mots de chaque catégorie de référence. La matrice ainsi obtenue indique la probabilité d'un mot de référence soit classé dans telle ou telle modèle. Ces valeurs sont souvent appelé taux producteur, car elles indiquent au producteur de la reconnaissance de quelle façon chaque catégorie est reconnue.

- *Normalisation en colonne :*

A l'inverse de la normalisation en ligne, une normalisation en colonnes, c'est-à-dire relativement aux nombres totaux de mots classés dans chaque catégorie, fournit des Taux utilisateurs. Ces taux donnent à l'utilisateur de la reconnaissance la probabilité qu'un mot classé dans une catégorie y appartienne réellement.

4.4.1.1.3. Calcul des taux de bonne reconnaissance et de confusion

a- Taux de bonne reconnaissance

Ce taux est égal au rapport du nombre de mots correctement reconnus pour un modèle donné, au nombre total de mots test de ce modèle.
Pour un modèle i, on calcul :

$$Taux\,\mathrm{Re}connaissance = \frac{\mathrm{Nombre\,de\,mots\,correctement\,reconnus\,pour\,le\,modèle\,i}}{\mathrm{Nombre\,total\,de\,mots\,test\,du\,modèle\ i}} \times 100 \quad \ldots \textit{(4.10)}$$

b- Taux de confusion

La confusion est l'effet d'affecter un mot test du modèle donné à un autre modèle.
Pour chaque modèle i, on calcule :

$$TauxConfusion = \frac{\mathrm{nombre\ de\ mots\ reconnus\ dans\ les\ autres\ modèles}}{\mathrm{nombre\ total\ de\ mots\ test\ du\ modèle}} \times 100 \quad \ldots\ldots\ldots\ldots \textit{(4.11)}$$

4.4.1.2 . Evaluation des résultats d'apprentissage et de test

Plusieurs paramètres sont pris en considération pendant de l'apprentissage des modèles de Markov cachés et pendant de l'apprentissage des réseaux de neurones. Dans notre application, les paramètres sont choisis pour offrir le meilleur résultat possible pour chaque algorithme.

Les résultats d'apprentissage et de test pour évaluer la reconnaissance audio, la reconnaissance vidéo et la reconnaissance audiovisuelle sont présentés dans les tableaux suivants sous forme des matrices de confusion :
Les paramètres choisis pour les CHMMs sont :
$M = 5$ et $N = 3$ [voir ANNEXE]

Modèles	Chiffres										
	siffer	***wahed***	***ithnani***	***thalatha***	***arbaa***	***khamssa***	***sitta***	***sabaa***	***thamania***	***tissaa***	
HMM0 (%)	**100**	0	0	0	0	0	0	0	0	0	
HMM1 (%)	0	**100**	0	0	0	0	0	0	0	0	
HMM2 (%)	0	0	**100**	0	0	0	0	0	10	0	
HMM3 (%)	0	0	0	**90**	0	0	0	0	10	0	
HMM4 (%)	0	0	0	0	**100**	0	0	10	0	0	
HMM5 (%)	0	0	0	0	0	**90**	0	0	0	0	
HMM6 (%)	0	0	0	0	0	10	**90**	0	0	0	
HMM7 (%)	0	0	0	0	0	0	0	**90**	0	0	
HMM8 (%)	0	0	0	10	0	0	0	0	**80**	0	
HMM9 (%)	0	0	0	0	0	0	10	0	0	**100**	
TMBRG (%)											***94***

Tableau 4.1 : *Matrice de confusion pour la base d'apprentissage (acoustique)*

Modèles	Chiffres										
	siffer	***wahed***	***ithnani***	***thalatha***	***arbaa***	***khamssa***	***sitta***	***sabaa***	***thamania***	***tissaa***	
HMM0 (%)	**86,66**	6,66	6,66	0	0	0	0	0	0	0	
HMM1 (%)	0	**66,66**	0	0	0	0	0	0	0	0	
HMM2 (%)	0	0	**53,33**	6,66	0	0	0	0	0	0	
HMM3 (%)	0	0	20	**73,33**	0	0	0	0	20	0	
HMM4 (%)	0	13,33	0	0	**53,33**	0	0	20	0	6,66	
HMM5 (%)	0	0	0	0	0	**60,00**	0	0	0	0	
HMM6 (%)	13,33	0	0	0	0	20	**66,66**	0	0	6,66	
HMM7 (%)	0	0	0	0	26,66	13,33	13,33	**80,00**	0	13,33	
HMM8 (%)	0	13,33	20	20	0	0	0	0	**80,00**	0	
HMM9 (%)	0	0	0	0	20	6,66	20	0	0	**73,33**	
TMBRG (%)											***69,33***

Tableau 4.2 : *Matrice de confusion pour la base de test (acoustique)*

Modèles	Chiffres									
	siffer	**wahed**	**ithnani**	**thalatha**	**arbaa**	**khamssa**	**sitta**	**sabaa**	**thamania**	**tissaa**
HMM0 (%)	**100**	0	0	0	0	0	0	0	0	0
HMM1 (%)	0	**80**	0	0	0	0	0	0	0	0
HMM2 (%)	0	10	**90**	0	0	0	0	0	0	0
HMM3 (%)	0	10	10	**90**	0	0	0	0	10	0
HMM4 (%)	0	0	0	0	**80**	0	0	20	0	0
HMM5 (%)	0	0	0	0	0	**90**	0	0	0	0
HMM6 (%)	0	0	0	0	0	0	**90**	0	0	10
HMM7 (%)	0	0	0	0	20	0	0	**70**	0	0
HMM8 (%)	0	0	0	10	0	0	0	0	**90**	0
HMM9 (%)	0	0	0	0	0	10	10	10	0	**90**
TMBRG (%)										***87***

Tableau 4.3 *: Matrice de confusion pour la base d'apprentissage (vidéo)*

Modèles	Chiffres										
	siffer	***wahed***	***ithnani***	***thalatha***	***arbaa***	***khamssa***	***sitta***	***sabaa***	***thamania***	***tissaa***	
HMM0 (%)	**73,33**	0	6,66	0	0	0	0	0	0	0	
HMM1 (%)	0	**60**	0	0	0	0	0	0	0	0	
HMM2 (%)	0	0	**46,66**	13,33	0	0	0	0	13,33	0	
HMM3 (%)	0	0	20	**60,00**	0	0	0	0	20	0	
HMM4 (%)	0	0	0	0	**46,66**	0	0	26,66	0	6,66	
HMM5 (%)	0	0	0	0	6,66	**53,33**	6,66	0	0	0	
HMM6 (%)	20	13,33	0	0	0	20	**60**	0	0	6,66	
HMM7 (%)	6,66	13,33	0	0	26,66	13,33	13,33	**53,33**	0	13,33	
HMM8 (%)	0	0	26,66	26,66	0	0	0	0	**66,66**	0	
HMM9 (%)	0	13,33	0	0	20	13,33	20	20	0	**73,33**	
TMBRG (%)											***59,33***

Tableau 4.4 : *Matrice de confusion pour la base de test (vidéo)*

4.4.2. Reconnaissances audio et vidéo : *''Influence du nombre de mixtures sur le taux de reconnaissance''*

Ces tests sont effectués pour voir l'influence du nombre de mixtures de gaussiennes M sur les taux de bonne reconnaissance audio et vidéo TMBRG (Taux Moyen de Bonne Reconnaissance Glogal) où tous les autres paramètres du modèle CHMM sont fixes et seul le nombre de mixtures de gaussiennes M qui varie de 1 jusqu'au 9.
Nous avons choisi un nombre d'états fixe pour tous les tests : $N = 3$ [Atoui et al., 2005]

4.4.2.1. Reconnaissance par mot (chiffre)

Nous comparons dans cette section les taux de reconnaissance entre les chiffres arabes. Les *figures* présentées ci-dessous montrent l'influence de nombre de mixtures M sur les taux de reconnaissance, test et apprentissage, acoustique et visuel pour chaque chiffre.

Nous montrons les valeurs du taux de reconnaissance acoustiques et visuels obtenus, pour chaque chiffre, sur l'ensemble des mots d'apprentissages et de tests dans un but de comparaison.

Les résultats ainsi obtenus montrent que les taux de reconnaissance visuels dans le cas de notre système *IS* sont plus petits que les taux de reconnaissance acoustiques. Cependant, la parole audio donne une meilleure reconnaissance. Nous remarquons, également qu'au niveau de chaque chiffre le visuel est moins important que l'acoustique.

Chaque chiffre a été représenté par plusieurs modèles selon le nombre de gaussiennes (1 à 9) pour un nombre d'états fixe $N = 3$.

D'après la majorité des figures ci-dessous on voit bien que les meilleurs taux de reconnaissance, acoustique et visuel, correspondent à un nombre de gaussienne $M = 5$ (d'après les chiffres : *siffer, wahed, arbaa, khamssa et tissaa* pour l'apprentissage et les chiffres : *siffer, wahed, ithnani,thamania et tissaa* pour le test).

Les figures des chiffres *thalatha* et *tissaa* présentent le chevauchement des deux courbes audio et vidéo sur toute la variation de nombre de mixtures. Cela, peut s'expliquer par le fait

que soit le son n'est pas très bien défini pendant l'enregistrement, soit la forme des lèvres n'est pas très claire pendant la vidéo.

Les résultats des figures des chiffres *wahed* et arb*aa,* montrent que les taux de reconnaissance, acoustique et visuel, sont presque indépendant sur toute la variation du nombre de mixtures.

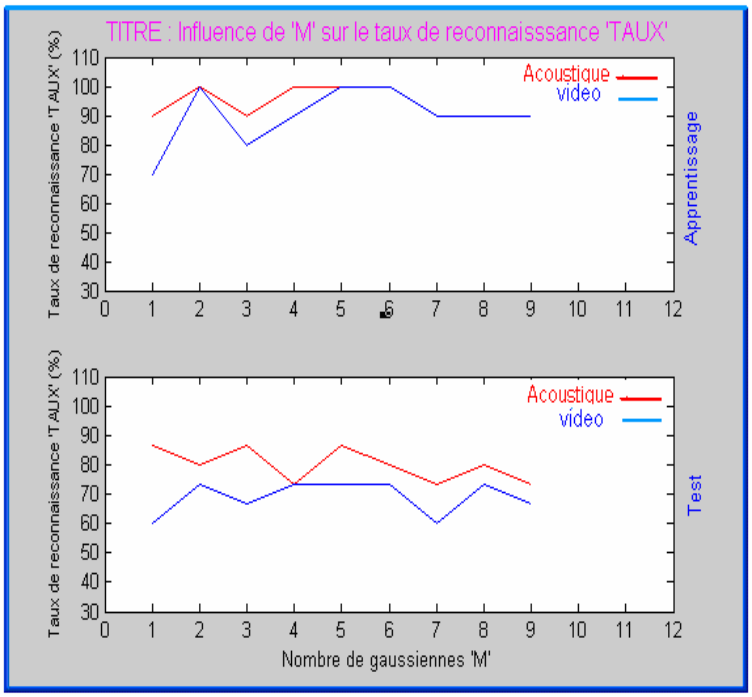

Figure a *: Chiffre siffer.*

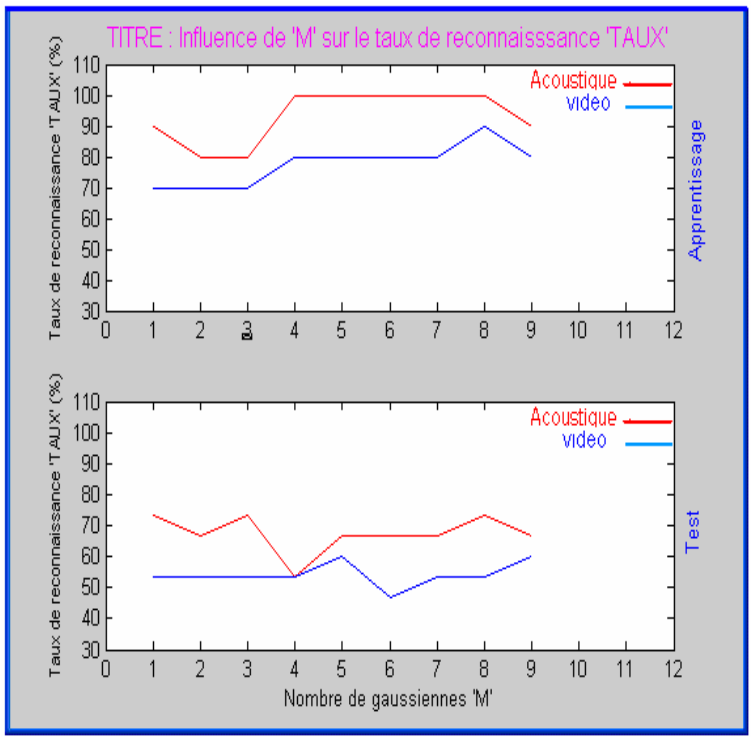

Figure b *: Chiffre wahed*

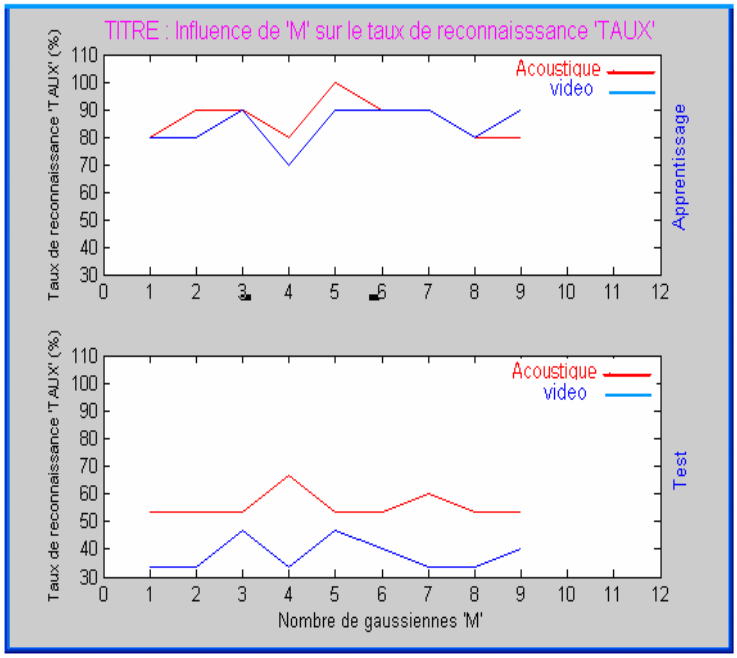

Figure c : *Chiffre ithnani*

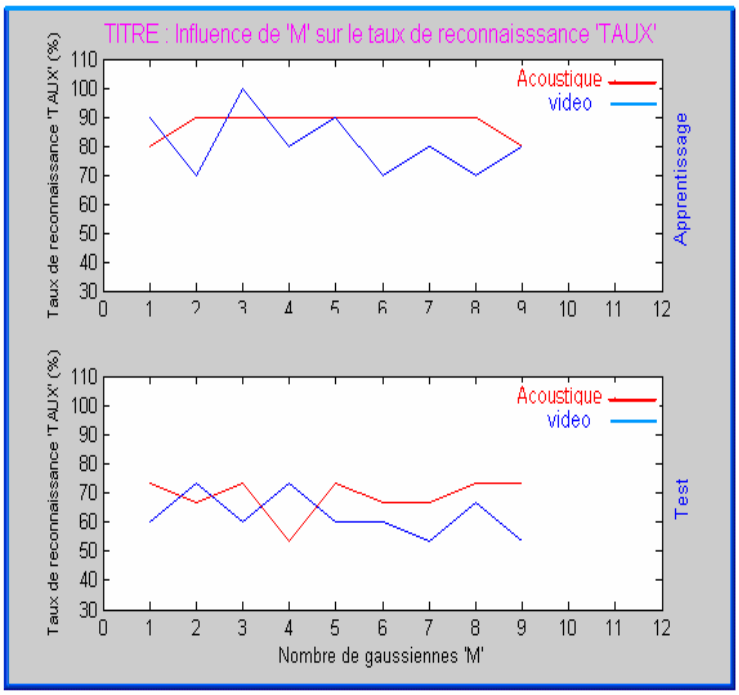

Figure d *: Chiffre thalatha*

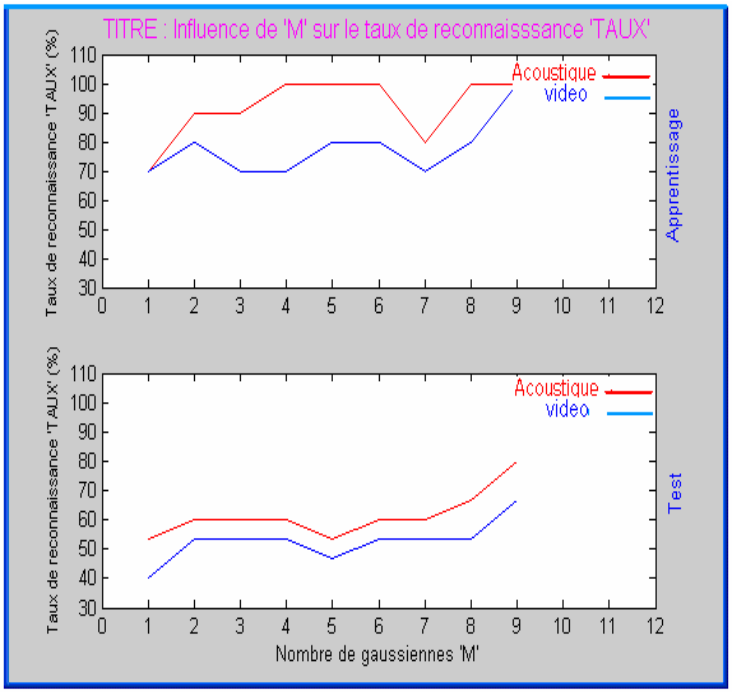

Figure e : *Chiffre arbaa*

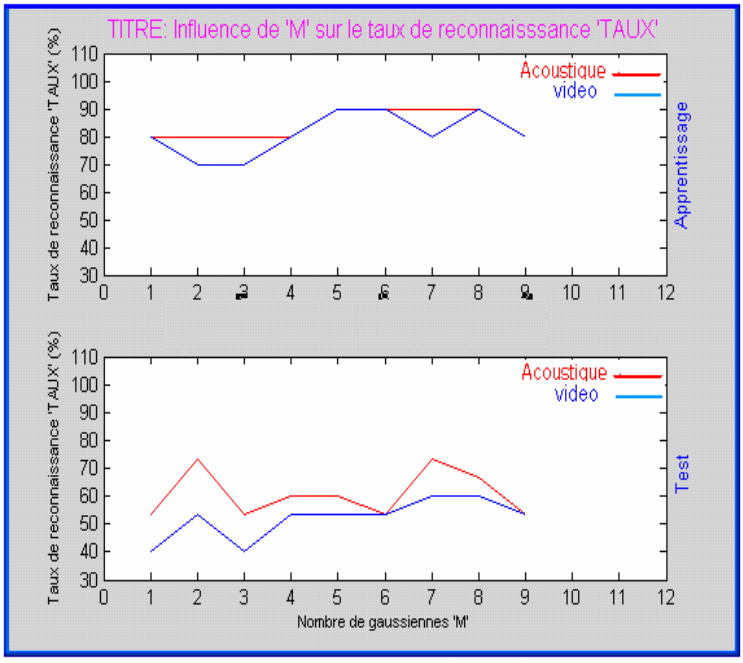

Figure f *: Chiffre khamssa*

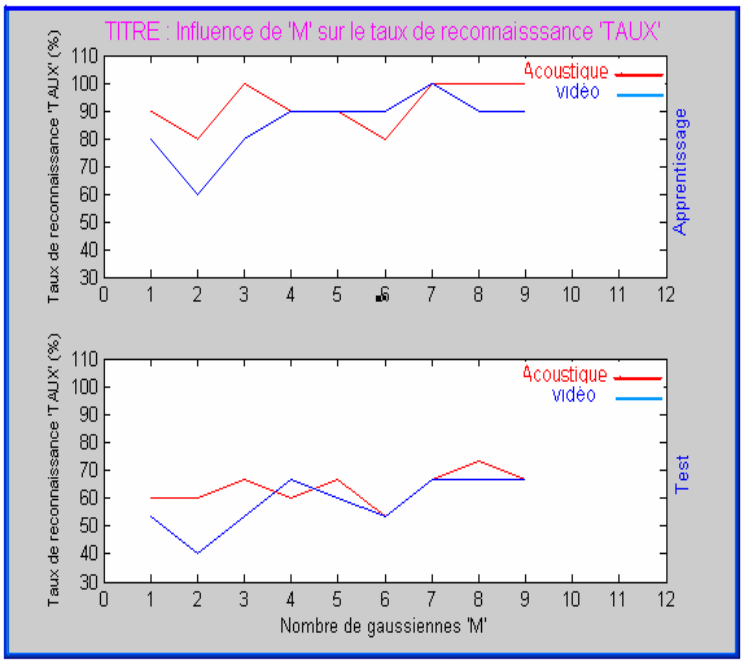

Figure g *: Chiffre sitta*

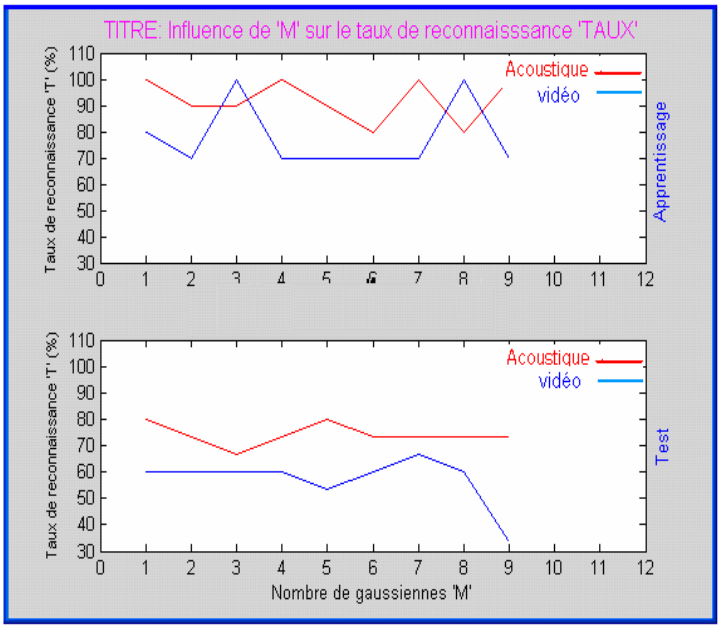

Figure h : *Chiffre sabaa*

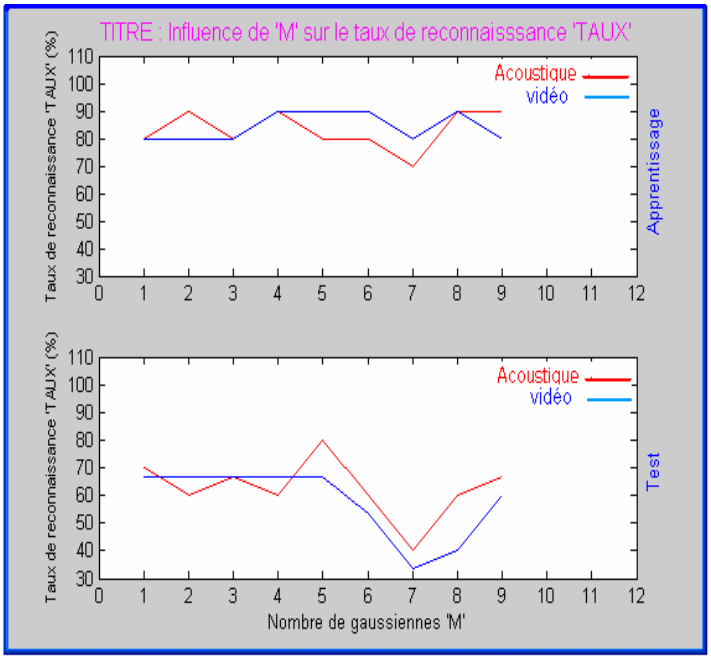

Figure i : *Chiffre thamania*

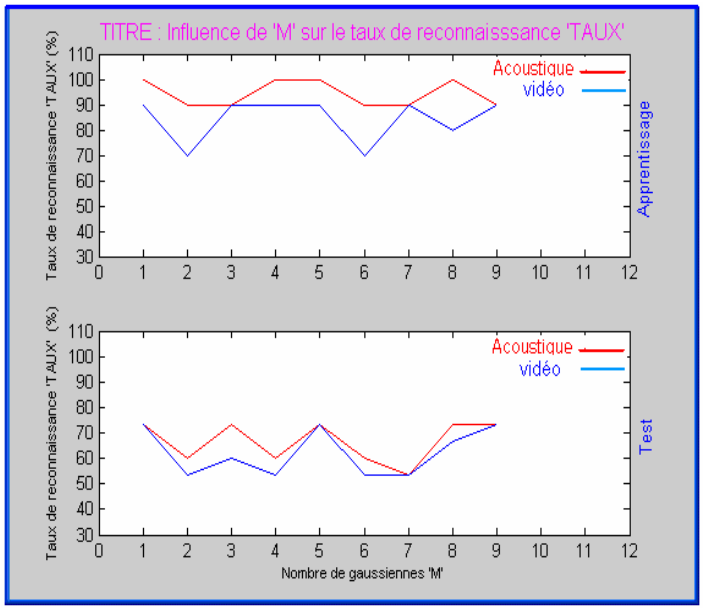

Figure j : *Chiffre tissaa*

Figure : 4.6 : *Influence de nombre de mixtures sur les taux de bonne reconnaissance pour chaque chiffre*

4.4.2.2. Reconnaissance globale

Le *tableau 4.5* ci-dessous montre l'évaluation globale de notre système en fonction du nombre de mixture M. On remarque que :

- Le taux de reconnaissance *acoustique* global varie entre *86%* et *94%* pour l'*apprentissage*, et entre *61.998%* et *69.330%* pour le *test*.
- Le taux de reconnaissance visuelle global varie entre *75%* et *87%* pour l'*apprentissage*, et *53.33%* et *59.33%* pour le *test*.

		M									
		1	2	3	4	5	6	7	8	9	
Résultats acoustiques (%)	***apprentissages***	86	88	88	93	94	90	90	92	90	***90,11***
	tests	67,66	65,33	67,32	62	69,33	62,66	63,33	69,33	67,99	***66,11***
Résultats Visuels (%)	***apprentissages***	79	75	83	81	87	83	83	86	85	***82,44***
	tests	53,99	55,99	54,66	58,66	59,33	54,66	53,33	57,33	57,33	***56,14***

Tableau 4.5 : *Représentation du taux de bonne reconnaissance global*

La *figure 4.7* suivante présente également l'influence de nombre de mixture M sur les taux de reconnaissance globaux, test et apprentissage, acoustique et visuel. Nous remarquons que les résultats du taux de reconnaissance acoustique sont meilleurs que ceux du visuel quelque soit le nombre de mixture, sauf que les meilleurs taux de reconnaissances sont obtenus pour $M = 5$, dont les taux sont :

- ✓ Taux acoustique d'apprentissage = *90.111%.*
- ✓ Taux acoustique de test = *66.107%.*
- ✓ Taux visuel d'apprentissage = *82.444%.*
- ✓ Taux visuel de test = *56.145%.*

Ceci prouve que la reconnaissance de la parole audio est meilleur que la reconnaissance de la parole vidéo dans un milieu réel.

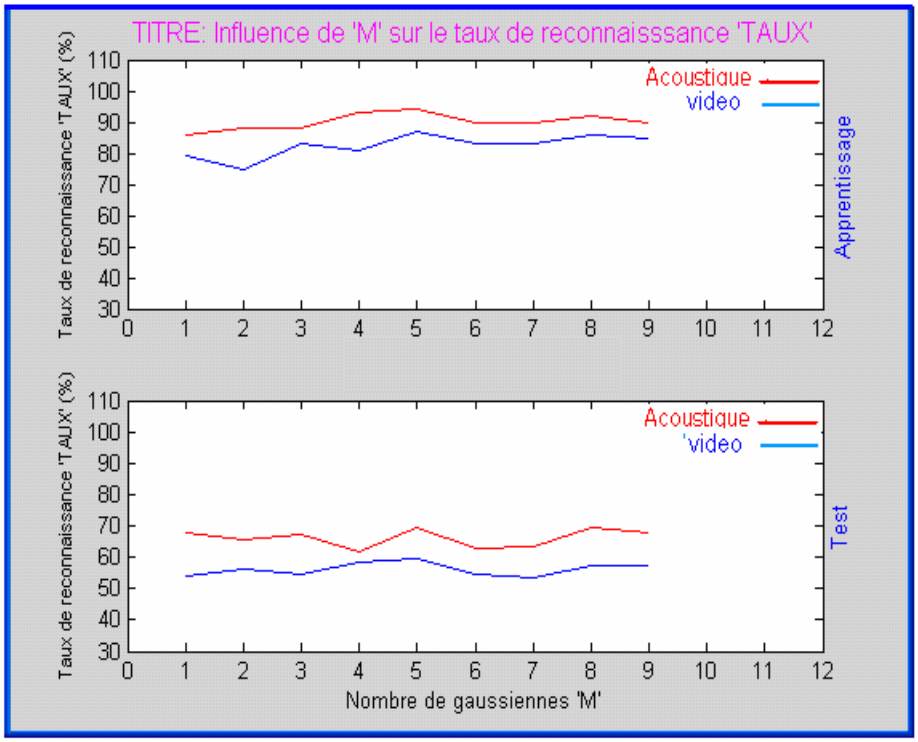

Figure 4.7 : *Influence de nombre de mixtures sur le taux de bonne reconnaissance global.*

4.4.3. Reconnaissance audiovisuelle

Le réseau utilisé pour la fusion des scores est de type PMC ayant les caractéristiques suivantes :
Une couche d'entrée à 20 entrées.
Une couche cachée avec 100 nœuds.
Une couche de sortie.

Les résultats des expériences sont présentées sous forme de tableaux indiquant chaque fois les ''Taux de bonne reconnaissance audiovisuel'' pour la base d'apprentissage et pour la base de test.

4.4.3.1. Tests d'apprentissage et de reconnaissance :

Les résultats des tests d'apprentissage et de reconnaissance sont présentés dans les tableaux ci-dessous.

4.4.3.1.1. Influence du nombre de nœuds

Pour voire l'importance du *nombre de nœuds* cachés et son influence sur le taux de bonne reconnaissance automatique de la parole audiovisuelle, on a effectué les tests suivants ou chaque fois nous fixons tous les autres paramètres du réseau , et seul le nombre de noeuds varie.

- *Nombre d'itération = 100000*
  Pas d'apprentissage = 0.1
  Momentum = 0.9

Nombre de nœuds	*TMBRG d'apprentissage (%)*	*TMBRG de Test (%)*
100	97	82
50	99	83
75	***99***	***86***

Tableaux 4.6 : *Influence de nombre de nœuds sur le taux de bonne reconnaissance audiovisuelle.*

4.4.3.1.2. Influence du nombre d'itérations

Ces tests sont effectués pour voir l'influence du *nombre d'itérations* sur le taux de bonne reconnaissance automatique de la parole audiovisuelle. Pour un même réseau nous varions seulement le nombre d'itérations, et les autres paramètres sont fixes :

- *Pas d'apprentissage = 0.1*
  Nombre de nœuds = 100
  Momentum = 0.9

Nombre d'itérations	*TMBRG d'apprentissage (%)*	*TMBRG de Test (%)*
2000	89	75
5000	90	76
10000	90	79
25000	91	77
50000	95	80
75000	96	81
100000	97	82
200000	***99***	***84***

Tableaux 4.7 : *Influence de nombre d'itérations sur le taux de bonne reconnaissance audiovisuelle.*

4.4.3.1.3. Influence du pas d'apprentissage

Les tests suivants présentent l'influence du paramètre *pas d'apprentissage* sur le taux de bonne reconnaissance automatique de la parole audiovisuelle, nous fixons tous les paramètres du réseau, et seul le pas d'apprentissage varie.

- *Nombre de nœuds = 100*
  Nombre d'itération = 100000
  Momentum = 0.9

Pas d'apprentissage	*TMBRG d'apprentissage (%)*	*TMBRG de Test (%)*
0,1	***97***	***82***
0,01	95	81
0,001	89	79

Tableaux 4.8 : *Influence du pas d'apprentissage sur le taux de bonne reconnaissance audiovisuelle*

4.4.3.1.4. Influence du momentum

Dans ce type de test, nous fixons tous les paramètres du réseau et nous varions le paramètre *momentum* pour voir son influence sur le taux de bonne reconnaissance automatique de la parole audiovisuelle, d'où les résultats suivants :

- *Nombre de nœuds = 100*
  Nombre d'itération = 100000
  Pas d'apprentissage = 0.1

Momentum	*TMBRG d'apprentissage (%)*	*TMBRG de Test (%)*
0,9	***97***	***82***
0,99	95	81
0,999	93	79

Tableaux 4.9 : *Influence du momentum sur le taux de bonne reconnaissance audiovisuelle.*

4.4.3.2. Interprétation des résultats

Pour tous les tests d'apprentissage et de reconnaissance effectués précédemment les paramètres de l'algorithme du réseau de neurones ont été choisis de façon à obtenir les meilleurs résultats. Chaque fois on prend les mêmes paramètres du réseau pour les deux tests *apprentissage* et *reconnaissance* pour voir l'influence du visuel sur l'acoustique.

Les tests effectués sur la variation du nombre de *nœuds* dans le réseau d'après *le tableau 4.6* montrent que le nombre de nœuds cachés possède une grande influence sur le comportement du réseau dont le TMBR augmente à chaque fois que l'on augmente ce nombre.

Nous remarquons aussi, à partir des tests, représentés dans *le tableau 4.7* (influence du nombre d'*itération)*, que ce réseau nécessite un nombre suffisamment grand d'itérations pour fournir une bonne augmentation du TMBR.

Les tableaux 4.8 et 4.9 représentent les tests effectués sur l'influence des paramètres *pas d'apprentissage* et *momentum* sur le taux de bonne reconnaissance, le réseau garde pratiquement les même taux de reconnaissance; d'où ces paramètres n'ont pas une grande influence sur les taux de reconnaissance.

4.4.3. Résultats Comparatifs

Les *tableaux 4.10* et *4.11* résument les résultats comparatifs globaux obtenus.

Chiffres	Résultats apprentissages				Résultats tests		
	audio	***vidéo***	**audiovisuelle**		***audio***	***vidéo***	**audiovisuelle**
siffer (%)	100	100	100		86,66	73,33	89.3
wahed (%)	100	80	100		66,66	60	84.6
ithnani (%)	100	90	99		53,33	46,66	84.0
thalatha (%)	90	90	98		73,33	60	88.0
arbaa (%)	100	80	99		53,33	46,66	82.6
khamsa (%)	90	90	98		60	53,33	88.0
sitta (%)	90	90	99		66,66	60	88.0
sabaa (%)	90	70	98		80	53,33	86.6
thamania (%)	80	90	99		80	66,66	88
tissaa (%)	100	90	100		73,33	73,33	87.3
TMBRG (%)	***94***	***87***	***99***		***69,33***	***59,33***	***86.6***

Tableau 4.10 : *Résultats d'expérience de reconnaissance de la parole des chiffres arabes.*

	TMBRG acoustique	***TMBRG vidéo***	***TMBRG audiovisuelle***
apprentissage (%)	94	87	99
test (%)	69,33	59,33	86.6

Tableau 4.11 : *Résultats globaux de reconnaissance de la parole.*

- ***Discussions***

Les tableaux 4.10 et 4.11 regroupent les performances, exprimées pour les systèmes acoustique, visuel et audiovisuels en terme du Taux Moyen de Bonne Reconnaissance Global (TMBRG).
Les résultats de reconnaissance exprimés dans le tableau 4.10, représentent ces TMBRG par mot (chiffre) de notre corpus pour des deux bases d'apprentissage et de test.
Les résultats de reconnaissance auditive nous servent de référence pour évaluer les performances du système audiovisuel mis en oeuvre.
Le tableau 4.11 représente les résultats globaux, ces résultats montrent qu'en utilisant un modèle à identification séparée, le score global est nettement supérieur aux scores aussi bien visuel qu'acoustique seuls (59.33% pour la reconnaissance visuelle seule, 69.33% pour la reconnaissance acoustique seule et 87% pour la reconnaissance audiovisuelle).
Ceci montre qu'une intégration audiovisuelle fondée sur le modèle d'identification séparée permet d'améliorer les performances de reconnaissance de la parole audiovisuelle par rapport à la reconnaissance acoustique seule ou à la reconnaissance visuelle seule.

4.5. Conclusion

Dans ce chapitre, nous avons présenté les expériences d'évaluation du système de reconnaissance de la parole audiovisuelle mis en oeuvre.

Plusieurs expériences ont été réalisées pour aboutir à la configuration optimale de notre système aussi bien au niveau du modèle de reconnaissance (CHMM) que du modèle de fusion réseau de neurones (ANN). Une base de données audiovisuelle à été également construite et paramétrée, les coefficients MFCC pour les paramètres acoustiques et les coefficients DCT pour les paramètres vidéo.

Notre système audiovisuel par identification séparée fondée sur une modélisation markovienne des réalisations acoustique et visuelle des mots suivie d'une fusion sur les scores peut constituer une alternative aux limites des systèmes de reconnaissance acoustique seul dans les milieux réels.

CONCLUSION GENERALE

Notre travail présenté dans ce document a porté sur la fusion des informations acoustiques et visuelle dans un système de reconnaissance de la parole utilisable en milieu réel. Nous avons ainsi abordé les principaux problèmes de la reconnaissance de la parole audiovisuelle, à savoir la paramétrisation des informations de parole, la nature du système de reconnaissance dans chacune des deux modalités, ainsi que le type de processus de fusion des informations auditives et visuelles.

Nous avons choisi de résoudre ces problèmes en nous appuyant sur des études réalisées dans le domaine de la perception audiovisuelle de la parole.

Nous nous sommes intéressés, en premier, à l'extraction des paramètres acoustiques et des paramètres visuels. Pour les paramètres acoustiques nous avons utilisé les coefficients cepstraux dans l'échelle Mel à savoir les MFCC, paramètres qui représentent bien l'enveloppe spectrale du signal acoustique. Les paramètres visuels sont eux calculés sur des images fixes basées sur la forme des lèvres, et paramétrées par les cœfficients DCT par ordre décroissent (Discret Cosine Transform).

Nous avons ensuite mis en œuvre le module de reconnaissance basé sur l'approche stochastique utilisant les modèles de Markov cachées continus. Ce module de reconnaissance est utilisé aussi bien pour la modalité acoustique que pour la modalité visuelle. C'est ainsi que des expériences de reconnaissance ont été réalisées pour chaque modalité prise séparément.
Deux constations sont à déduire des résultats de ces expériences :

- la modalité visuelle peut être utilisée pour la reconnaissance de la parole (par exemple : pour les malentendants).
- La modalité acoustique reste la plus performante (elle donne les meilleurs taux de reconnaissance).

Nous avons ensuite mis en oeuvre un système de reconnaissance combinant les deux modalités acoustique et visuelle en optant pour une fusion basée sur le modèle d'intégration séparée (*IS*). Ce système a été testé sur notre corpus audiovisuel, constitué de séquences des chiffres arabes prononcés par une seule locutrice arabisante. Les résultats des expériences réalisées dans un milieu réel nous amènent à conclure sur l'apport de la modalité visuelle dans un système de reconnaissance. En effet, un TMBRG (Taux Moyen de Bonne Reconnaissance Globaux), égale à 69.33%, a été obtenu pour la modalité acoustique seule, un TMBRG égale à 59.33% pour la modalité visuelle seule et un TMBRG égale à 87% pour les deux modalités combinées.

Ces expériences nous permettent de constater que l'information visuelle intégrée à l'information acoustique peut constituer une alternative afin d'augmenter la performance des systèmes de reconnaissance en milieu réel (nécessairement bruité).
Les objectifs de notre travail nous semblent atteint dans la mesure ou nous avons mis au point un système de reconnaissance audiovisuel testé dans un milieu réel. Pour sa mise en œuvre nous avons eu à étudier la reconnaissance par l'approche stochastique basée sur les modèles de Markov cachés continus et la fusion séparée utilisant les réseaux de neurones artificiels.
Le système mis en œuvre reste bien étendu perfectible. Les performances de ce système doivent être réévaluer dans un contexte plus étendu, par exemple sur un corpus de plus grande taille, prononcé par plusieurs locuteurs, ou un corpus fortement bruité à différents RSB (Rapport Signal Bruit). Comme perspectives également de ce travail d'autres types de fusion peuvent être aussi testées comme par exemple la fusion directe ou bien la fusion hybride.

REFERENCES BIBLIOGRAPHIQUES

[Adjoudani, 1993] : Ali. Adjoudani. ''*Élaboration d'un modèle de lèvres 3D pour animation en temps réel''*. Mémoire de D.E.A. Signal Image Parole, Institut National Polytechnique de Grenoble,France ; 1993.

[Adjoudani et Benoit, 1996] : A. Adjoudani et C. Benoit. *''On the integration of auditory and visual parameters in an HMM-based ASR''*. In Speechreading by humans and machines (D. G. Stork and M. E. Hennecke, eds.), pp. 461-471, Springer, 1996.

[Auckenthaler et al., 1999a] : Roland Auckenthale, Jason Brand, John S. D. Masson, Fazzin Deravi, and Claude Chilufya Chibeluchi. ''*Lip signatures of automatic person recognition''*. In Proc. 2nd AVBPA, page 142-147, Washington, DC, USA, March 22-23, 1999.

[Auckenthaler et al., 1999b] : Roland Auckenthale, Jason Brand, John S. D. Masson, Fazzin Deravi, and Claude Chilufya Chibeluchi. ''*Lip signatures of automatic person recognition''*. In Submited to IEE 99, 1999

[Austin, 92] : S. Austin, G. Zavaliagkos, J. Makhoul, R. Schwartz: ''*Speech recognition using segmental neural nets''* , IEEE 1992.

[Barreaud, 2004] : Vincent Barreaud. *''Reconnaissance automatique de la parole continue : Compensation de bruits par transformation de la parole''*. Thèse de doctorat. Ecole doctorale IAEM Lorraine de l'université Henri Poincaré - Nancy 1, 9 Novembre 2004.

[Baum, 1972] : L. Baum. *"An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes"*. In Inequalities. Pages 3:1-8, 1972.

[Boite et al, 1999] : R. Boite, H. Bourlard, T. Dutoit, J. Hancq et H. Leich. *''Traitement de la parole''*. Collection Electricité, presses polytechniques et universitaires remandes,EPLF-Centre Midi, CH-1015 Lausanne.

[Bourlard, 92] : H. Bourlard : ''*Reconnaissance automatique de la parole : modèles stockastiques et/ou modèles connexionistes ?''*, XIX JEP, Bruxelles, pp 499-506, Mai 1992

[Bovik, 2000]: A. Bovik:*''Handbouk of Image and Video Processing''*,Academic Press,p891.

[Bregler et al., 1993] : C. Bregler, H. Hild, S. Manke, and A. Waibel**. ''***Improving connected letter recognition by lipreading''*. In proc. of the International Conference on acoustics, Speech, and Signal Processing, volume 1, page 557-560, Minneapolis, USA

[Calliope,1989]:CALLIOPE:'*La parole et son traitement automatique,*édition Masson, 1989.

[Debyeche, 20007] : Mohamed DEBYECHE. *'' Reconnaissance Automatique de la Parole Appliquée à la Langue Arabe.* Thèse de doctorat de l'Université des Sciences et de la Technologies Houari Boumediene USTHB. ALGERIE, 2007

[Dupont, 1996] : Pierre DUPONT. *''Utilisation et apprentissage de modele de langage pour la reconnaissance de la parole continue''*. Thèse de doctorat de l'ENST, Paris, 1996

[Dutoit et al., 2002] : T. Dutoit, L. Couvreur, F. Malfrère, V. Pagel, C. Ris, ''*Synthèse Vocale et Reconnaissance de la Parole : Droites Gauches et Mondes Parallèles''*, CFA'02, Lille, France, Avril 2002.

[Flandrin, 1993] : P. Flandrin : ''*Temps-Fréquences'', Editions* Hermès, 1993

[Gerasimos et al, 2003]: Gerasimos Potamianos, Chalapathy Neti, and Sabine Deligne ''*Joint Audio-Visual Speech Processing for Recognition and Enhancement''*; IEEE, proceeding of the Auditory-Visual Speech Processing Tutorial and Research Workshop (AVSP), pp. 95-104, St. Jorio,n France, Septembre 2003

[Green et Kuhl, 1989] : K.P. Green, P.K. Kuhl. **''***The role of visual information in the processing of place and manner features in speech perception''*. Perception and Psychophysics, 45; 34-42.

[Haton et all., 1991] : J-P Haton , J-M Pierrel, G Pérennou , J Caelen , J-L Gauvain , ‘’*Reconnaissance Automatique de la Parole’’,* AFCET, édition DUNOD Informatique, 1991

[Jacob, 1995] : Bruno JACOB. *‘’Un outil informatique de gestion Modèles de Markov Cachés : expérimentations en Reconnaissance Automatique de la Parole’’*. Thèse de doctorat de l'Université Paul Sabatier de Toulouse III. Septembre 1995

[Jacob et Sénac, 1996] : B. Jacob and C. Sénac**. ‘’***Un modèle maître-esclave pour la fusion des données acoustiques et articulatoires en reconnaissance automatique de la parole’’*. Actes des 21èmes Journées d'Etude de la Parole, pages 363-366, Avignon, France

[Klatt, 1979]: D. H. Klatt**.** *‘’Speech perception: ‘’A model of acoustic-phonetic analysis and lexical’’*. Access. Journal of Phonetics’’, 7:279:312.

[Kunt, 1991] : M. Kunt,: *‘’Techniques modernes de traitement numérique des signaux’’,* Presses polytechniques et universitaires Romandes, 1991.

[Lallouache, 1991] : T. Lallouache. ‘’*Un poste visage-parole : acquisition et traitement automatique des contours labiaux’’*. Thèse de doctorat, Institut National Polytechnique Grenoble, France, 1991

[Liberman et Mattingly, 1985] : A.M. Liberman et I.G. Mattingly. **‘’***The motor theory of speech perception revised’’*. Cognition, 21:1-36.

[Liévin et Luthon, 1999] : Marc Liévin et Franck Luthon. ‘’*Unsupervised lip segmentation under natural conditions’’*. In; volume 6, page 3065-3068, Phonix, AZ, USA, March 1999.

[Lisker et Rossi, 1992] : L. Lisker and M. Rossi. **‘’***Auditory and visual cueing of the [± rounded] feature of vowels’’*. Language and Speech, 35(4), pp 391-417

[Le Viet Bac, 2006] : Le Viet Bac. *‘’Reconnaissance automatique de la parole pour des langues peu dotées’’*. Thèse PHD. Université Joseph Fourier - GRENOBLE 1, 2006

[Levinson, 1985] : S. Levinson : *‘’Structural methods in automatic speech recognition’’* , proc. IEEE, Vol. 73, ndeg.11, 1985.

[Malbos, 1995] : F. Malbos : *‘’Détection et identification des occlusives à l'aide de la transformée en ondelettes’’,* Thèse de doctorat 3deg.cycle, Paris XI Orsay, Janvier 1995.

[Markel et al., 1976]: J.D. Markel, A.H. Gray : ‘’*Linear prediction in speech’’*, Editions Springer-Verley, Berlin, Heidelberg, New-York, 1976.

[Massaro, 1996] : D. W. Massaro. ‘’*Bimodal speech perception’’*: A progress report. In D. G. Stork and M. E. Hennecke editors, Speech reading by Humans and Machines, Models, Systems, and Applications, volume 150, pages 79-102. Springer-Verlag, Berlin.

[Massaro, 1987a] : D. W. Massaro. ‘’*Integrating multiple sources of information in listening and reading’’*. In Language Perception and Production, Academic press, New York

[Mokbel, 1992] : Mokbel : *‘’Reconnaissance automatique de la parole dans le bruit’’,* Thèse de doctorat-ingénieur, Telecom Paris, 1992

[Montacié et al., 1996] : Claude Montacié, Regine André-Obrecht, Louis Jean-Boé, Marie-José Caraty, Paul Deléglise, Isabelle Herlin, and Henry Meloni. ‘’*Application multimodales pour interfaces et bornes evolués (AMIBE’’)*, rapport final. Technical report, GDR-PRC Communication Homme-Machine, 1996).

[Neti et all., 2000] : Chalapathy Nety, Gerassimos Potamianos, Juergen Luettin, Lain Matthews, Hervé Glotin, Dimitra Vergyri, June Sison, Azad Mashari, et Jie Zhou. *‘’Audio-visual speech recognition’’*. Technical report Worshop 2000, Internationnal Computer Science Institute, center for language and speech processing (CLSP), The Johns Hopkins University, Baltimore, MD, USA, October 12 2000

[Nocera, 1992] : P. Nocera : *‘’Utilisation conjointe de réseaux neuronaux et de connaissances explicites pour le décodage acoustico-phonétique’’,* Thèse de l'Université d'Avignon et des Pays du Vaucluse, 1992.

[Petajan et al., 1987] : E. D. Petajan, B. J. Bischoff, D. A. Bodoff, and N. M. Brooke**. ''***An improved automatic lipreading system to enhance speech recognition''*. Technical Report TM 11251-871012-11, Bell Labs.

[Potamianos et al., 1998] : G. Potamianos, H. P. Graf, and E. Cosatto : ''*An image transform approach for HMM based automatic lip reading*", in Proceedings of the International Conference on Image Processing, vol. 3, pp. 173–177, 1998

[Potamianos et all., 2004] : G. Potamianos , C. Neti , J. Luttin, et I. Matthews. *''Audio-visual automatic speech recognition : an overview'',* In issues in audio-visual speech processing (G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, eds.), MIT Press, 2004

[Rabiner, 1989] : L. R. Rabiner. *''A tutorial on Hidden Markov Models and selected applications in speech recognition"*. Proceeding of the IEEE, vol. 77(2), 1989

[Rabiner et Juang, 1993] : L. Rabiner et B.H. Juang. ''*Fundamentals of speech recognition''*. Prentice Hall, 1993.

[Reichl et Ruske, 1995] : W. Reichl et G. Ruske. *''Descriminative training for continuous speech recognition''*. In eurospeech, 1995.

[Reilly et Scanlon, 2001] : R. Reilly and P. Scanlon : ''*Feature analysis for automatic speech reading''*, Proc. Workshop on Multimedia Signal Processing, pp. 625–630, 2001

[Revéret et Benoît, 1997]:L. Revéret and C. Benoît.**''***A viseme-based approach to labiometrics''*.In J. Bigün, G.Chollet and G.Borgefors, editors,1st International Conference on Audio and Video based Biometrics for Person Authentication,Crans-Montana, pages 335-343

[Robert-Ribes, 1995] : J. Robert-Ribes. **''***Modèles d'intégration audiovisuelle de signaux liguistiques : de la perception humaine à la reconnaissance automatique des voyelles''*. Thèse de doctorat, Signal Image Parole, Institut National Polytechnique de Grenoble, France.

[Robert-Ribes et al., 1996] : J. Robert-Ribes, M. Piquemal, J. L. ''Schwartz. *Exploiting sensor fusion architectures and stimuli complementary in AV speech recognition''*. In D. G. Stork and M. E. Hennecke, editors, Speech reading by Humans and Machines, Models, Systems, and Applications, volume 150, pages 193-210. Springer-Verlag, Berlin

[Rogozan, 1999] : Alexandrina ROGOZAN. *''Etude de la fusion des données hétérogènes pour la reconnaissance automatique de la parole audiovisuelle''* .Thèse PHD. Ecole doctorale en électronique de l'université d'Orsay, Paris, 1999

[Sakoë et all., 1978] : H. Sakoë, S. Chiba : *''Dynamic programming optimisation for spoken word recognition'',* Proc. IEE Trans. on ASSP, Vol. ASSP-26, p. 43, Février 1978

[Silsbee, 1993] : P. L. Silsbee. ''*Computer lipreading for improved accuracy in automatic speech recognition''*. PhD thesis, University of Texas.

[Silsbee, 1994] : P. L. Silsbee. **''***Sensory in tegration in audiovisual automatic speech recognition''*. In 28th Annual Asilomar Conference on Signals, Systems, and Computers, volume 1, pages 561-565

[Silsbee et Su, 1996] : P. L. Silsbee and Q. Su**. ''***Audiovisual sensory integration using hidden Markov models''*. In D. G. Stork and M. E. Hennecke, editors, Speech reading by Humans and Machines, Models, Systems, and Applications, volume 150, pages 489-496. Springer-Verlag, Berlin.

[Simon, 1985] : J.C. Simon : ''*Invariance en reconnaissance des formes'',* Colloque COGNITIVA 1985, pp 119-130.

[Stork et al., 1992] : D. G. Stork, G. J. Wolf, and E. P. Levine**. ''***Neural Network lipreading system for improved speech recognition''*. In proc. of the International Joint Conference on Neural Networks, volume 2, pages 289-295, IEEE, Baltimore.

[Su et Silsbee, 1996] : Q. Su and P. L. Silsbee. ''*Robust audiovisual integration using semicontinuous Hidden Markov Models''*. In proc. of the 4th International Conference on Spoken Language Processing, Philadelphia, USA

[Summerfield et McGrath, 1984] : Q. Summerfield and M. McGrath. **''***Detection and resolution of audio-visual incompatibility in the perception of vowels''*. Quarterly Journal of Experimental Psychology, 36A(1):51-74.

[Viterbi, 1967] : A. Viterbi. *''Error bounds for convolutional codes and an asymptotically optimum decoding algorithm''*. In IEEE Transactions on Information Theory, pages IT-13(2): 260-269, 1967.

[Youcefi et Meziane, 2002]: Abdellah Youcefi et Abdelwafi Meziane. *'' Introduction de l'enrgie dans un modele de reconnaissance automatique de la parole''*. XXIV^{èmes} journées d'étude sur la parole, Nancy 24-27 juin 2002

[Zurcher, 1980] : Zurcher : ''*Le vocodeur à canaux : une nouvelle jeunesse''*, Recherche/Acoustique CNET, 1980, Vol 5, pp21-40.

ANNEXE 1

Les tableaux suivants représentent les résultats des *taux maximums* de reconnaissance automatique de la parole *acoustiques et* visuelles, pour un nombre d'états fixe *''N = 3''* et un nombre de mixtures *''M''* différents (1 à 9).

Meilleurs résultats des taux de reconnaissance

	Résultats acoustiques				Résultats vidéo			
Chiffres	***App***	***Test***	***N***	***M***	***App***	***Test***	***N***	***M***
siffer	100%	86,66%	3	5	100%	73,33%	3	5
wahed	100%	66,66%	3	3	80%	60,00%	3	3
ithnani	90%	60,00%	3	7	90%	33,33%	3	7
thalatha	100%	73,33%	3	3	100%	60,00%	3	3
arbaa	100%	60,00%	3	4	80%	60,00%	3	4
khamsa	90%	73,33%	3	7	80%	60,00%	3	7
sitta	100%	66,66%	3	7	100%	66,66%	3	7
sabaa	100%	80,00%	3	1	80%	60,00%	3	1
thamania	100%	80,00%	3	4	90%	66,66%	3	4
tissaa	100%	73,33%	3	1	90%	73,33%	3	1
taux	***98%***	***71,997%***			***89%***	***61,331%***		

- ***M = 1***

	Résultats acoustiques				Résultats vidéo			
Chiffres	***App***	***Test***	***N***	***M***	***App***	***Test***	***N***	***M***
siffer	90%	86,66%	3	1	70%	60,00%	3	1
wahed	90%	73,33%	3	1	70%	53,33%	3	1
ithnani	80%	53,33%	3	1	80%	33,33%	3	1
thalatha	80%	73,33%	3	1	90%	60,00%	3	1
arbaa	70%	53,33%	3	1	70%	40,00%	3	1
khamsa	80%	53,33%	3	1	80%	40,00%	3	1
sitta	90%	60,00%	3	1	80%	53,33%	3	1
sabaa	100%	80,00%	3	1	80%	60,00%	3	1
thamania	80%	70,00%	3	1	80%	66,66%	3	1
tissaa	100%	73,33%	3	1	90%	73,33%	3	1
taux	***86%***	***67,664%***			***79%***	***53,998%***		

➢ ***M = 2***

	Résultats acoustiques				Résultats vidéo			
Chiffres	***App***	***Test***	***N***	***M***	***App***	***Test***	***N***	***M***
siffer	100%	80,00%	3	2	100%	73,33%	3	2
wahed	80%	66,66%	3	2	70%	53,33%	3	2
ithnani	90%	53,33%	3	2	80%	33,33%	3	2
thalatha	90%	66,66%	3	2	70%	73,33%	3	2
arbaa	90%	60,00%	3	2	80%	53,33%	3	2
khamsa	80%	73,33%	3	2	70%	53,33%	3	2
sitta	80%	60,00%	3	2	60%	40,00%	3	2
sabaa	90%	73,33%	3	2	70%	60,00%	3	2
thamania	90%	60,00%	3	2	80%	66,66%	3	2
tissaa	90%	60,00%	3	2	70%	53,33%	3	2
taux	***88%***	***65,331%***			***75%***	***55,997%***		

➢ ***M = 3***

	Résultats acoustiques				Résultats vidéo			
Chiffres	***App***	***Test***	***N***	***M***	***App***	***Test***	***N***	***M***
siffer	90%	86,66%	3	3	80%	66,66%	3	3
wahed	80%	73,33%	3	3	70%	53,33%	3	3
ithnani	90%	53,33%	3	3	90%	33,33%	3	3
thalatha	90%	73,33%	3	3	100%	60,00%	3	3
arbaa	90%	60,00%	3	3	70%	53,33%	3	3
khamsa	80%	53,33%	3	3	70%	40,00%	3	3
sitta	100%	66,66%	3	3	80%	53,33%	3	3
sabaa	90%	66,66%	3	3	100%	60,00%	3	3
thamania	80%	66,66%	3	3	80%	66,66%	3	3
tissaa	90%	73,33%	3	3	90%	60,00%	3	3
taux	***88%***	***67,329%***			***83%***	***54,664%***		

➢ ***M = 4***

	Résultats acoustiques				Résultats vidéo			
Chiffres	***App***	***Test***	***N***	***M***	***App***	***Test***	***N***	***M***
siffer	100%	73,33%	3	4	90%	73,33%	3	4
wahed	100%	53,33%	3	4	80%	53,33%	3	4
ithnani	80%	66,66%	3	4	70%	33,33%	3	4
thalatha	90%	53,33%	3	4	80%	73,33%	3	4
arbaa	100%	60,00%	3	4	70%	53,33%	3	4
khamsa	80%	60,00%	3	4	80%	53,33%	3	4
sitta	90%	60,00%	3	4	90%	66,66%	3	4
sabaa	100%	73,33%	3	4	70%	60,00%	3	4
thamania	90%	60,00%	3	4	90%	66,66%	3	4
tissaa	100%	60,00%	3	4	90%	53,33%	3	4
taux	***93%***	***61,998%***			***81%***	***58,663%***		

- ***M = 5***

	Résultats acoustiques				Résultats vidéo			
Chiffres	***App***	***Test***	***N***	***M***	***App***	***Test***	***N***	***M***
siffer	100%	86,66%	3	5	100%	73,33%	3	5
wahed	100%	66,66%	3	5	80%	60,00%	3	5
ithnani	100%	53,33%	3	5	90%	46,66%	3	5
thalatha	90%	73,33%	3	5	90%	60,00%	3	5
arbaa	100%	53,33%	3	5	80%	46,66%	3	5
khamsa	90%	60,00%	3	5	90%	53,33%	3	5
sitta	90%	66,66%	3	5	90%	60,00%	3	5
sabaa	90%	80,00%	3	5	70%	53,33%	3	5
thamania	80%	80,00%	3	5	90%	66,66%	3	5
tissaa	100%	73,33%	3	5	90%	73,33%	3	5
taux	***94%***	***69,330%***			***87%***	***59,330%***		

- ***M = 6***

	Résultats acoustiques				Résultats vidéo			
Chiffres	***App***	***Test***	***N***	***M***	***App***	***Test***	***N***	***M***
siffer	100%	80,00%	3	6	100%	73,33%	3	6
wahed	100%	66,66%	3	6	80%	46,66%	3	6
ithnani	90%	53,33%	3	6	90%	40,00%	3	6
thalatha	90%	66,66%	3	6	70%	60,00%	3	6
arbaa	100%	60,00%	3	6	80%	53,33%	3	6
khamsa	90%	53,33%	3	6	90%	53,33%	3	6
sitta	80%	53,33%	3	6	90%	53,33%	3	6
sabaa	80%	73,33%	3	6	70%	60,00%	3	6
thamania	80%	60,00%	3	6	90%	53,33%	3	6
tissaa	90%	60,00%	3	6	70%	53,33%	3	6
taux	***90%***	***62,664%***			***83%***	***54,664%***		

- ***M = 7***

	Résultats acoustiques				Résultats vidéo			
Chiffres	***App***	***Test***	***N***	***M***	***App***	***Test***	***N***	***M***
siffer	90%	73,33%	3	7	90%	60,00%	3	7
wahed	100%	66,66%	3	7	80%	53,33%	3	7
ithnani	90%	60,00%	3	7	90%	33,33%	3	7
thalatha	90%	66,66%	3	7	80%	53,33%	3	7
arbaa	80%	60,00%	3	7	70%	53,33%	3	7
khamsa	90%	73,33%	3	7	80%	60,00%	3	7
sitta	100%	66,66%	3	7	100%	66,66%	3	7
sabaa	100%	73,33%	3	7	70%	66,66%	3	7
thamania	70%	40,00%	3	7	80%	33,33%	3	7
tissaa	90%	53,33%	3	7	90%	53,33%	3	7
taux	***90%***	***63,330%***			***83%***	***53,330%***		

- ***M = 8***

	Résultats acoustiques				Résultats vidéo			
Chiffres	***App***	***Test***	***N***	***M***	***App***	***Test***	***N***	***M***
siffer	90%	80,00%	3	8	90%	73,33%	3	8
wahed	100%	73,33%	3	8	90%	53,33%	3	8
ithnani	80%	53,33%	3	8	80%	33,33%	3	8
thalatha	90%	73,33%	3	8	70%	66,66%	3	8
arbaa	100%	66,66%	3	8	80%	53,33%	3	8
khamsa	90%	66,66%	3	8	90%	60,00%	3	8
sitta	100%	73,33%	3	8	90%	66,66%	3	8
sabaa	80%	73,33%	3	8	100%	60,00%	3	8
thamania	90%	60,00%	3	8	90%	40,00%	3	8
tissaa	100%	73,33%	3	8	80%	66,66%	3	8
taux	***92%***	***69,330%***			***86%***	***57,330%***		

- ***M = 9***

	Résultats acoustiques				Résultats vidéo			
Chiffres	***App***	***Test***	***N***	***M***	***App***	***Test***	***N***	***M***
siffer	90%	73,33%	3	9	90%	66,66%	3	9
wahed	90%	66,66%	3	9	80%	60,00%	3	9
ithnani	80%	53,33%	3	9	90%	40,00%	3	9
thalatha	80%	73,33%	3	9	80%	53,33%	3	9
arbaa	100%	80,00%	3	9	100%	66,66%	3	9
khamsa	80%	53,33%	3	9	80%	53,33%	3	9
sitta	100%	66,66%	3	9	90%	66,66%	3	9
sabaa	100%	73,33%	3	9	70%	33,33%	3	9
thamania	90%	66,66%	3	9	80%	60,00%	3	9
tissaa	90%	73,33%	3	9	90%	73,33%	3	9
taux	***90%***	***67,996%***			***85%***	***57,330%***		

ANNEXE 2

********** *RESULTATS ACOUSTIQUES* **********

Les tableaux suivants représentent les résultats des taux de Reconnaissance Automatique de la Parole *acoustiques*, sous forme *matrices de confusion*, pour un nombre d'états fixe *''N = 3''* et un nombre de mixtures *''M''* différents (1 à 9).

- ***M = 1***

CHIFFRES	Modèles HMMi										
	HMM0	***HMM1***	***HMM2***	***HMM3***	***HMM4***	***HMM5***	***HMM6***	***HMM7***	***HMM8***	***HMM9***	
siffer	**90,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	
wahed	0,00%	**90,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	
ithnani	10,00%	0,00%	**80,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	¨10%	0,00%	
thalatha	0,00%	0,00%	0,00%	**80,00%**	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**70,00%**	0,00%	0,00%	20,00%	0,00%	10,00%	
khamssa	0,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	10,00%	0,00%	0,00%	10,00%	
sitta	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	10,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**100,00%**	0,00%	0,00%	
thamania	0,00%	0,00%	10,00%	10,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**100,00%**	
Taux	90,00%	90,00%	80,00%	80,00%	70,00%	80,00%	90,00%	100,00%	80,00%	100,00%	***86,000%***

Matrice de confusion APPRENTISSAGE

CHIFFRES	Modèles HMMi										
	HMM0	***HMM1***	***HMM2***	***HMM3***	***HMM4***	***HMM5***	***HMM6***	***HMM7***	***HMM8***	***HMM9***	
siffer	**86,66%**	0,00%	0,00%	0,00%	0,00%	0,00%	13,33%	0,00%	0,00%	0,00%	
wahed	0,00%	**73,33%**	13,33%	0,00%	0,00%	6,66%	0,00%	0,00%	6,66%	0,00%	
ithnani	0,00%	0,00%	**53,33%**	26,66%	0,00%	0,00%	0,00%	6,66%	13,33%	0,00%	
thalatha	0,00%	0,00%	13,33%	**73,33%**	0,00%	0,00%	0,00%	0,00%	13,33%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**53,33%**	0,00%	0,00%	33,33%	0,00%	13,33%	
khamssa	6,66%	0,00%	0,00%	0,00%	0,00%	**53,33%**	20,00%	6,66%	0,00%	13,33%	
sitta	6,66%	0,00%	0,00%	0,00%	0,00%	6,66%	**60,00%**	13,33%	0,00%	13,33%	
sabaa	0,00%	0,00%	0,00%	0,00%	13,33%	0,00%	0,00%	**80,00%**	0,00%	6,66%	
thamania	0,00%	0,00%	6,66%	13,33%	6,66%	0,00%	0,00%	0,00%	**73,33%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	6,66%	13,33%	6,66%	0,00%	**73,33%**	
Taux	86,66%	73,33%	53,33%	73,33%	53,33%	53,33%	60,00%	80,00%	73,33%	73,33%	***67,997%***

Matrice de confusion TEST

- ***M = 2***

CHIFFRES	Modèles HMMi										
	HMM0	***HMM1***	***HMM2***	***HMM3***	***HMM4***	***HMM5***	***HMM6***	***HMM7***	***HMM8***	***HMM9***	
siffer	**100,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
wahed	0,00%	**80,00%**	20,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
ithnani	0,00%	0,00%	**90,00%**	10,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
thalatha	0,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	0,00%	0,00%	10,00%	
khamssa	10,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	10,00%	0,00%	0,00%	0,00%	
sitta	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	0,00%	0,00%	20,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	
thamania	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	**90,00%**	
Taux	100,00%	80,00%	90,00%	90,00%	90,00%	80,00%	80,00%	90,00%	90,00%	90,00%	***88,000%***

Matrice de confusion APPRENTISSAGE

CHIFFRES	Modèles HMMi										
	HMM0	***HMM1***	***HMM2***	***HMM3***	***HMM4***	***HMM5***	***HMM6***	***HMM7***	***HMM8***	***HMM9***	
siffer	**80,00%**	0,00%	0,00%	0,00%	0,00%	6,66%	13,33%	0,00%	0,00%	0,00%	
wahed	6,66%	**66,66%**	20,00%	6,66%	0,00%	0,00%	0,00%	0,00%	6,66%	0,00%	
ithnani	0,00%	6,66%	**53,33%**	13,33%	0,00%	0,00%	0,00%	0,00%	26,66%	0,00%	
thalatha	0,00%	0,00%	13,33%	**66,66%**	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**60,00%**	0,00%	0,00%	26,66%	0,00%	13,33%	
khamssa	0,00%	0,00%	0,00%	0,00%	0,00%	**73,33%**	13,33%	0,00%	0,00%	13,33%	
sitta	6,66%	0,00%	0,00%	0,00%	0,00%	6,66%	**60,00%**	13,33%	0,00%	13,33%	
sabaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	6,66%	**73,33%**	0,00%	20,00%	
thamania	0,00%	6,66%	13,33%	20,00%	0,00%	0,00%	0,00%	0,00%	**60,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	6,66%	13,33%	20,00%	0,00%	**60,00%**	
Taux	80,00%	66,66%	53,33%	66,66%	60,00%	73,33%	60,00%	73,33%	60,00%	60,00%	***65,331%***

Matrice de confusion TEST

- ***M = 3***

CHIFFRES	Modèles HMMi										
	HMM0	***HMM1***	***HMM2***	***HMM3***	***HMM4***	***HMM5***	***HMM6***	***HMM7***	***HMM8***	***HMM9***	
siffer	**90,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	
wahed	20,00%	**80,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
ithnani	0,00%	0,00%	**90,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
thalatha	0,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	10,00%	0,00%	0,00%	
khamssa	0,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	20,00%	0,00%	0,00%	0,00%	
sitta	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**100,00%**	0,00%	0,00%	0,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	
thamania	0,00%	0,00%	0,00%	20,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	**90,00%**	
Taux	90,00%	80,00%	90,00%	90,00%	90,00%	80,00%	100,00%	90,00%	80,00%	90,00%	***88,000%***

Matrice de confusion APPRENTISSAGE

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**86,66%**	0,00%	0,00%	0,00%	0,00%	0,00%	13,33%	0,00%	0,00%	0,00%	
wahed	13,33%	**73,33%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	6,66%	6,66%	
ithnani	6,66%	0,00%	**53,33%**	13,33%	0,00%	0,00%	0,00%	0,00%	26,66%	0,00%	
thalatha	0,00%	0,00%	6,66%	**73,33%**	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	
arbaa	6,66%	0,00%	0,00%	0,00%	**60,00%**	0,00%	0,00%	20,00%	0,00%	13,33%	
khamssa	0,00%	0,00%	0,00%	0,00%	0,00%	**53,33%**	6,66%	13,33%	13,33%	13,33%	
sitta	6,66%	0,00%	0,00%	0,00%	0,00%	6,66%	**66,66%**	0,00%	0,00%	20,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	0,00%	**66,66%**	0,00%	13,33%	
thamania	0,00%	0,00%	13,33%	20,00%	0,00%	0,00%	0,00%	0,00%	**66,66%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	13,33%	0,00%	0,00%	13,33%	0,00%	**73,33%**	
Taux	86,66%	73,33%	53,33%	73,33%	60,00%	53,33%	66,66%	66,66%	66,66%	73,33%	***67,329%***

Matrice de confusion TEST

➢ ***M = 4***

CHiFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**100,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
wahed	0,00%	**100,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
ithnani	0,00%	0,00%	**80,00%**	10,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
thalatha	0,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**100,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	
khamssa	0,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	20,00%	0,00%	0,00%	0,00%	
sitta	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	10,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**100,00%**	0,00%	0,00%	
thamania	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**100,00%**	
Taux	100,00%	100,00%	80,00%	90,00%	100,00%	80,00%	90,00%	100,00%	90,00%	100,00%	***93,000%***

Matrice de confusion APPRENTISSAGE

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**73,33%**	0,00%	0,00%	0,00%	0,00%	0,00%	13,33%	6,66%	0,00%	6,66%	
wahed	13,33%	**53,33%**	13,33%	0,00%	0,00%	13,33%	0,00%	6,66%	0,00%	0,00%	
ithnani	0,00%	0,00%	**66,66%**	20,00%	0,00%	0,00%	0,00%	0,00%	13,33%	0,00%	
thalatha	0,00%	0,00%	13,33%	**53,33%**	0,00%	0,00%	0,00%	0,00%	33,33%	0,00%	
arbaa	0,00%	0,00%	6,66%	0,00%	**60,00%**	0,00%	0,00%	26,66%	0,00%	6,66%	
khamssa	6,66%	0,00%	0,00%	0,00%	0,00%	**60,00%**	20,00%	0,00%	0,00%	13,33%	
sitta	6,66%	0,00%	0,00%	0,00%	0,00%	0,00%	**60,00%**	13,33%	0,00%	20,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	0,00%	**73,33%**	0,00%	13,33%	
thamania	0,00%	0,00%	13,33%	26,66%	0,00%	0,00%	0,00%	0,00%	**60,00%**	0,00%	
tissaa	6,66%	0,00%	0,00%	0,00%	0,00%	6,66%	13,33%	13,33%	0,00%	**60,00%**	
Taux	73,33%	53,33%	66,66%	53,33%	60,00%	60,00%	60,00%	73,33%	60,00%	60,00%	***61,998%***

Matrice de confusion TEST

- ***M = 5***

CHIFFRES	Modèles HMMi										
	HMM0	***HMM1***	***HMM2***	***HMM3***	***HMM4***	***HMM5***	***HMM6***	***HMM7***	***HMM8***	***HMM9***	
siffer	**100,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
wahed	0,00%	**100,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
ithnani	0,00%	0,00%	**100,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
thalatha	0,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**100,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	
khamssa	0,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	10,00%	0,00%	0,00%	0,00%	
sitta	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	10,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	
thamania	0,00%	0,00%	10,00%	10,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**100,00%**	
Taux	100,00%	100,00%	100,00%	90,00%	100,00%	90,00%	90,00%	90,00%	80,00%	100,00%	***94,000%***

Matrice de confusion APPRENTISSAGE

CHFFRES	Models HMMi										
	HMM0	***HMM1***	***HMM2***	***HMM3***	***HMM4***	***HMM5***	***HMM6***	***HMM7***	***HMM8***	***HMM9***	
siffer	**86,66%**	0,00%	0,00%	0,00%	0,00%	0,00%	13,33%	0,00%	0,00%	0,00%	
wahed	6,66%	**66,66%**	0,00%	0,00%	13,33%	0,00%	0,00%	0,00%	13,33%	0,00%	
ithnani	6,66%	0,00%	**53,33%**	20,00%	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	
thalatha	0,00%	0,00%	6,66%	**73,33%**	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**53,33%**	0,00%	0,00%	26,66%	0,00%	20,00%	
khamssa	0,00%	0,00%	0,00%	0,00%	0,00%	**60,00%**	20,00%	13,33%	0,00%	6,66%	
sitta	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**66,66%**	13,33%	0,00%	20,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	0,00%	**80,00%**	0,00%	0,00%	
thamania	0,00%	0,00%	0,00%	20,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	6,66%	0,00%	6,66%	13,33%	0,00%	**73,33%**	
Taux	86,66%	66,66%	53,33%	73,33%	53,33%	60,00%	66,66%	80,00%	80,00%	73,33%	***69,330%***

Matrice de confusion TEST

- ***M = 6***

CHIFFRES	Modèles HMMi										
	HMM0	***HMM1***	***HMM2***	***HMM3***	***HMM4***	***HMM5***	***HMM6***	***HMM7***	***HMM8***	***HMM9***	
siffer	**100,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
wahed	0,00%	**100,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
ithnani	0,00%	0,00%	**90,00%**	10,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
thalatha	0,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**100,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	
khamssa	0,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	10,00%	0,00%	0,00%	0,00%	
sitta	10,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	0,00%	0,00%	10,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	**80,00%**	0,00%	10,00%	
thamania	0,00%	0,00%	20,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	
Taux	100,00%	100,00%	90,00%	90,00%	100,00%	90,00%	80,00%	80,00%	80,00%	90,00%	***90,000%***

Matrice de confusion APPRENTISSAGE

CHFFRES	Models HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**80,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	13,33%	0,00%	0,00%	6,66%	
wahed	13,33%	**66,66%**	0,00%	0,00%	6,66%	6,66%	0,00%	0,00%	0,00%	6,66%	
ithnani	0,00%	6,66%	**53,33%**	20,00%	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	
thalatha	0,00%	0,00%	13,33%	**66,66%**	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**60,00%**	0,00%	0,00%	26,66%	0,00%	13,33%	
khamssa	6,66%	0,00%	0,00%	0,00%	0,00%	**53,33%**	20,00%	6,66%	0,00%	13,33%	
sitta	0,00%	0,00%	0,00%	0,00%	0,00%	6,66%	**53,33%**	13,33%	0,00%	26,66%	
sabaa	0,00%	0,00%	0,00%	0,00%	13,33%	0,00%	0,00%	**73,33%**	0,00%	13,33%	
thamania	0,00%	0,00%	13,33%	26,66%	0,00%	0,00%	0,00%	0,00%	**60,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	6,66%	13,33%	20,00%	0,00%	**60,00%**	
Taux	80,00%	66,66%	53,33%	66,66%	60,00%	53,33%	53,33%	73,33%	60,00%	60,00%	***62,664%***

Matrice de confusion TEST

➢ ***M = 7***

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**90,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	
wahed	0,00%	**100,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
ithnani	0,00%	0,00%	**90,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	
thalatha	0,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**80,00%**	0,00%	0,00%	20,00%	0,00%	0,00%	
khamssa	0,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	10,00%	0,00%	0,00%	0,00%	
sitta	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**100,00%**	0,00%	0,00%	0,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**100,00%**	0,00%	0,00%	
thamania	0,00%	0,00%	0,00%	20,00%	0,00%	0,00%	0,00%	10,00%	**70,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	**90,00%**	
Taux	90,00%	100,00%	90,00%	90,00%	80,00%	90,00%	100,00%	100,00%	70,00%	90,00%	***90,000%***

Matrice de confusion APPRENTISSAGE

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**73,33%**	6,66%	0,00%	0,00%	0,00%	0,00%	13,33%	0,00%	0,00%	6,66%	
wahed	13,33%	**66,66%**	0,00%	0,00%	0,00%	0,00%	0,00%	6,66%	0,00%	13,33%	
ithnani	0,00%	6,66%	**60,00%**	13,33%	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	
thalatha	0,00%	0,00%	13,33%	**66,66%**	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**60,00%**	0,00%	0,00%	20,00%	0,00%	20,00%	
khamssa	6,66%	0,00%	0,00%	0,00%	0,00%	**73,33%**	13,33%	0,00%	0,00%	6,66%	
sitta	6,66%	0,00%	0,00%	0,00%	0,00%	0,00%	**66,66%**	13,33%	0,00%	13,33%	
sabaa	6,66%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**73,33%**	0,00%	20,00%	
thamania	0,00%	6,66%	13,33%	40,00%	0,00%	0,00%	0,00%	0,00%	**40,00%**	0,00%	
tissaa	6,66%	6,66%	0,00%	0,00%	0,00%	0,00%	13,33%	20,00%	0,00%	**53,33%**	
Taux	73,33%	66,66%	60,00%	66,66%	60,00%	73,33%	66,66%	73,33%	40,00%	53,33%	***63,330%***

Matrice de confusion TEST

- ***M = 8***

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**90,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	
wahed	0,00%	**100,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
ithnani	0,00%	0,00%	**80,00%**	20,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
thalatha	0,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**100,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	
khamssa	0,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	10,00%	0,00%	0,00%	0,00%	
sitta	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**100,00%**	0,00%	0,00%	0,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	**80,00%**	0,00%	10,00%	
thamania	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**100,00%**	
Taux	90,00%	100,00%	80,00%	90,00%	100,00%	90,00%	100,00%	80,00%	90,00%	100,00%	***92,000%***

Matrice de confusion APPRENTISSAGE

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**80,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	13,33%	6,66%	0,00%	0,00%	
wahed	13,33%	**73,33%**	6,66%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	6,66%	
ithnani	6,66%	0,00%	**53,33%**	20,00%	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	
thalatha	0,00%	0,00%	6,66%	**73,33%**	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**66,66%**	0,00%	0,00%	26,66%	0,00%	6,66%	
khamssa	6,66%	0,00%	0,00%	0,00%	0,00%	**66,66%**	13,33%	0,00%	0,00%	13,33%	
sitta	0,00%	0,00%	0,00%	0,00%	0,00%	6,66%	**73,33%**	6,66%	0,00%	13,33%	
sabaa	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	0,00%	**73,33%**	0,00%	6,66%	
thamania	0,00%	0,00%	13,33%	26,66%	0,00%	0,00%	0,00%	0,00%	**60,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	6,66%	6,66%	13,33%	0,00%	**73,33%**	
Taux	80,00%	73,33%	53,33%	73,33%	66,66%	66,66%	73,33%	73,33%	60,00%	73,33%	***69,330%***

Matrice de confusion TEST

- ***M = 9***

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**90,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	
wahed	0,00%	**90,00%**	10,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
ithnani	0,00%	10,00%	**80,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
thalatha	0,00%	0,00%	0,00%	**80,00%**	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**100,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	
khamssa	0,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	10,00%	0,00%	0,00%	10,00%	
sitta	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**100,00%**	0,00%	0,00%	0,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**100,00%**	0,00%	0,00%	
thamania	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	**90,00%**	
Taux	90,00%	90,00%	80,00%	80,00%	100,00%	80,00%	100,00%	100,00%	90,00%	90,00%	***90,000%***

Matrice de confusion APPRENTISSAGE

CHIFFRES	Modèles HMMi										
	HMM0	***HMM1***	***HMM2***	***HMM3***	***HMM4***	***HMM5***	***HMM6***	***HMM7***	***HMM8***	***HMM9***	
siffer	**73,33%**	6,66%	0,00%	0,00%	0,00%	0,00%	13,33%	0,00%	0,00%	6,66%	
wahed	13,33%	**66,66%**	0,00%	0,00%	0,00%	13,33%	0,00%	0,00%	0,00%	6,66%	
ithnani	0,00%	0,00%	**53,33%**	20,00%	0,00%	0,00%	0,00%	0,00%	26,66%	0,00%	
thalatha	0,00%	0,00%	6,66%	**73,33%**	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**80,00%**	0,00%	0,00%	20,00%	0,00%	0,00%	
khamssa	6,66%	0,00%	0,00%	0,00%	0,00%	**53,33%**	20,00%	6,66%	0,00%	13,33%	
sitta	6,66%	0,00%	0,00%	0,00%	0,00%	0,00%	**66,66%**	6,66%	0,00%	20,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	0,00%	**73,33%**	0,00%	6,66%	
thamania	0,00%	0,00%	13,33%	20,00%	0,00%	0,00%	0,00%	0,00%	**66,66%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	6,66%	20,00%	0,00%	**73,33%**	
Taux	73,33%	66,66%	53,33%	73,33%	80,00%	53,33%	66,66%	73,33%	66,66%	73,33%	***67,996%***

Matrice de confusion TEST

********** *RESULTATS VIDEOS* **********

Les tableaux suivants représentent les résultats des taux de reconnaissance automatique de la parole *visuelles*, sous forme *matrices de confusion*, pour un nombre d'états fixe *''N = 3''* et un nombre de mixtures *''M''* différents (1 à 9).

- ***M = 1***

CHIFFRES	Modèles HMMi										
	HMM0	***HMM1***	***HMM2***	***HMM3***	***HMM4***	***HMM5***	***HMM6***	***HMM7***	***HMM8***	***HMM9***	
Siffer	**70,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	20,00%	10,00%	0,00%	0,00%	
wahed	20,00%	**70,00%**	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	0,00%	
ithnani	10,00%	0,00%	**80,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
thalatha	0,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
Arbaa	0,00%	0,00%	0,00%	0,00%	**70,00%**	0,00%	0,00%	20,00%	0,00%	10,00%	
khamssa	0,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	10,00%	0,00%	0,00%	10,00%	
Sitta	10,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	0,00%	0,00%	10,00%	
Sabaa	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	**80,00%**	0,00%	10,00%	
thamania	0,00%	0,00%	10,00%	10,00%	0,00%	0,00%	0,00%	10,00%	**80,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	**90,00%**	
Taux	70,00%	70,00%	80,00%	90,00%	70,00%	80,00%	80,00%	80,00%	80,00%	90,00%	***79,000%***

Matrice de confusion APPRENTISSAGE

CHIFFRES	Modèles HMMi										
	HMM0	***HMM1***	***HMM2***	***HMM3***	***HMM4***	***HMM5***	***HMM6***	***HMM7***	***HMM8***	***HMM9***	
siffer	**60,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	13,33%	20,00%	0,00%	6,66%	
wahed	0,00%	**53,33%**	20,00%	0,00%	0,00%	13,33%	6,66%	0,00%	6,66%	0,00%	
ithnani	0,00%	0,00%	**33,33%**	26,66%	13,33%	0,00%	0,00%	6,66%	20,00%	0,00%	
thalatha	6,66%	0,00%	20,00%	**60,00%**	0,00%	0,00%	0,00%	0,00%	13,33%	0,00%	
arbaa	13,33%	6,66%	0,00%	0,00%	**40,00%**	0,00%	0,00%	26,66%	0,00%	13,33%	
khamssa	6,66%	0,00%	0,00%	0,00%	0,00%	**40,00%**	20,00%	13,33%	0,00%	20,00%	
sitta	6,66%	0,00%	0,00%	0,00%	0,00%	13,33%	**53,33%**	13,33%	0,00%	13,33%	
sabaa	0,00%	6,66%	0,00%	0,00%	20,00%	0,00%	0,00%	**60,00%**	0,00%	13,33%	
thamania	0,00%	0,00%	13,33%	13,33%	6,66%	0,00%	0,00%	0,00%	**66,66%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	6,66%	13,33%	6,66%	0,00%	**73,33%**	
Taux	60,00%	53,33%	33,00%	60,00%	40,00%	40,00%	53,33%	60,00%	66,66%	73,33%	***53,998%***

Matrice de confusion TEST

- ***M = 2***

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**100,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
wahed	10,00%	**70,00%**	20,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
ithnani	0,00%	0,00%	**80,00%**	10,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
thalatha	0,00%	0,00%	20,00%	**70,00%**	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**80,00%**	0,00%	0,00%	10,00%	0,00%	10,00%	
khamssa	10,00%	0,00%	0,00%	0,00%	0,00%	**70,00%**	20,00%	0,00%	0,00%	0,00%	
sitta	20,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**60,00%**	0,00%	0,00%	20,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	0,00%	**70,00%**	0,00%	10,00%	
thamania	0,00%	0,00%	10,00%	10,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	20,00%	0,00%	**70,00%**	
Taux	100,00%	70,00%	80,00%	70,00%	80,00%	70,00%	60,00%	70,00%	80,00%	70,00%	***75,000%***

Matrice de confusion APPRENTISSAGE

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**73,33%**	0,00%	0,00%	0,00%	0,00%	6,66%	20,00%	0,00%	0,00%	0,00%	
wahed	13,33%	**53,33%**	13,33%	6,66%	0,00%	6,66%	0,00%	0,00%	6,66%	0,00%	
ithnani	13,33%	6,66%	**33,33%**	26,66%	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	
thalatha	6,66%	0,00%	6,66%	**73,33%**	0,00%	0,00%	0,00%	0,00%	13,33%	0,00%	
arbaa	0,00%	6,66%	0,00%	0,00%	**53,33%**	0,00%	0,00%	20,00%	0,00%	20,00%	
khamssa	6,66%	0,00%	0,00%	0,00%	0,00%	**53,33%**	20,00%	6,66%	0,00%	13,33%	
sitta	13,33%	0,00%	0,00%	0,00%	0,00%	6,66%	**40,00%**	26,66%	0,00%	13,33%	
sabaa	0,00%	0,00%	0,00%	0,00%	13,33%	0,00%	6,66%	**60,00%**	0,00%	20,00%	
thamania	0,00%	6,66%	13,33%	13,33%	0,00%	0,00%	0,00%	0,00%	**66,66%**	0,00%	
tissaa	6,66%	0,00%	0,00%	0,00%	0,00%	6,66%	13,33%	20,00%	0,00%	**53,33%**	
Taux	73,33%	53,33%	33,33%	73,33%	53,33%	53,33%	40,00%	60,00%	66,66%	53,33%	***55,997%***

Matrice de confusion TEST

- ***M = 3***

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**80,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	10,00%	0,00%	0,00%	
wahed	20,00%	**70,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
ithnani	0,00%	0,00%	**90,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
thalatha	0,00%	0,00%	0,00%	**100,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**70,00%**	0,00%	0,00%	20,00%	0,00%	10,00%	
khamssa	20,00%	0,00%	0,00%	0,00%	0,00%	**70,00%**	10,00%	0,00%	0,00%	0,00%	
sitta	10,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	10,00%	0,00%	0,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**100,00%**	0,00%	0,00%	
thamania	0,00%	0,00%	10,00%	10,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	**90,00%**	
Taux	80,00%	70,00%	90,00%	100,00%	70,00%	70,00%	80,00%	100,00%	80,00%	90,00%	***83,000%***

Matrice de confusion APPRENTISSAGE

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**66,66%**	0,00%	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	0,00%	13,33%	
wahed	20,00%	**53,33%**	6,66%	0,00%	0,00%	0,00%	0,00%	0,00%	13,33%	6,66%	
ithnani	13,33%	6,66%	**33,33%**	26,66%	0,00%	6,66%	0,00%	0,00%	13,33%	0,00%	
thalatha	0,00%	6,66%	13,33%	**60,00%**	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	
arbaa	6,66%	6,66%	0,00%	0,00%	**53,33%**	0,00%	0,00%	20,00%	0,00%	13,33%	
khamssa	6,66%	0,00%	6,66%	0,00%	0,00%	**40,00%**	6,66%	13,33%	6,66%	20,00%	
sitta	6,66%	0,00%	0,00%	0,00%	0,00%	6,66%	**53,33%**	13,33%	0,00%	20,00%	
sabaa	0,00%	6,66%	0,00%	0,00%	20,00%	0,00%	0,00%	**60,00%**	0,00%	13,33%	
thamania	0,00%	0,00%	13,33%	20,00%	0,00%	0,00%	0,00%	0,00%	**66,66%**	0,00%	
tissaa	0,00%	0,00%	0,00%	6,66%	6,66%	0,00%	0,00%	26,66%	0,00%	**60,00%**	
Taux	66,66%	53,33%	33,33%	60,00%	53,33%	40,00%	53,33%	60,00%	66,66%	60,00%	***54,664%***

Matrice de confusion TEST

➢ ***M = 4***

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**70,00%**	10,00%	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	0,00%	0,00%	
wahed	10,00%	**80,00%**	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	0,00%	
ithnani	0,00%	0,00%	**70,00%**	10,00%	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	
thalatha	0,00%	0,00%	10,00%	**80,00%**	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**70,00%**	0,00%	0,00%	20,00%	0,00%	10,00%	
khamssa	0,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	20,00%	0,00%	0,00%	0,00%	
sitta	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	10,00%	
sabaa	10,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	**70,00%**	0,00%	10,00%	
thamania	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	0,00%	
tissaa	10,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	
Taux	70,00%	80,00%	70,00%	80,00%	70,00%	80,00%	90,00%	70,00%	90,00%	90,00%	***79,000%***

Matrice de confusion APPRENTISSAGE

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**73,33%**	0,00%	0,00%	0,00%	0,00%	0,00%	13,33%	6,66%	0,00%	6,66%	
wahed	20,00%	**53,33%**	6,66%	0,00%	0,00%	13,33%	0,00%	6,66%	0,00%	0,00%	
ithnani	6,66%	6,66%	**33,33%**	20,00%	0,00%	0,00%	0,00%	0,00%	26,66%	6,66%	
thalatha	0,00%	0,00%	13,33%	**73,33%**	0,00%	0,00%	0,00%	0,00%	33,33%	0,00%	
arbaa	0,00%	0,00%	0,00%	6,66%	**53,33%**	0,00%	0,00%	20,00%	0,00%	13,33%	
khamssa	6,66%	6,66%	0,00%	0,00%	0,00%	**53,33%**	13,33%	6,66%	0,00%	13,33%	
sitta	6,66%	0,00%	0,00%	0,00%	0,00%	0,00%	**66,66%**	6,66%	0,00%	20,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	26,66%	0,00%	0,00%	**60,00%**	0,00%	13,33%	
thamania	0,00%	0,00%	13,33%	20,00%	0,00%	0,00%	0,00%	0,00%	**66,66%**	0,00%	
tissaa	6,66%	0,00%	0,00%	0,00%	13,33%	6,66%	6,66%	13,33%	0,00%	**53,33%**	
Taux	73,33%	53,33%	33,33%	73,33%	53,33%	53,33%	66,66%	60,00%	66,66%	53,33%	***58,663%***

Matrice de confusion TEST

- ***M = 5***

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**100,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
wahed	0,00%	**80,00%**	10,00%	10,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
ithnani	0,00%	0,00%	**90,00%**	10,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
thalatha	0,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**80,00%**	0,00%	0,00%	20,00%	0,00%	0,00%	
khamssa	0,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	0,00%	10,00%	
sitta	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	10,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	0,00%	**70,00%**	0,00%	10,00%	
thamania	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	**90,00%**	
Taux	100,00%	80,00%	90,00%	90,00%	80,00%	90,00%	90,00%	70,00%	90,00%	90,00%	***87,000%***

Matrice de confusion APPRENTISSAGE

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**73,33%**	0,00%	0,00%	0,00%	0,00%	0,00%	20,00%	6,66%	0,00%	0,00%	
wahed	0,00%	**60,00%**	0,00%	0,00%	0,00%	0,00%	13,33%	13,33%	0,00%	13,33%	
ithnani	6,66%	0,00%	**46,66%**	20,00%	0,00%	0,00%	0,00%	0,00%	26,66%	0,00%	
thalatha	0,00%	0,00%	13,33%	**60,00%**	0,00%	0,00%	0,00%	0,00%	26,66%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**46,66%**	6,66%	0,00%	26,66%	0,00%	20,00%	
khamssa	0,00%	0,00%	0,00%	0,00%	0,00%	**53,33%**	20,00%	13,33%	0,00%	13,33%	
sitta	0,00%	0,00%	0,00%	0,00%	0,00%	6,66%	**60,00%**	13,33%	0,00%	20,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	26,66%	0,00%	0,00%	**53,33%**	0,00%	20,00%	
thamania	0,00%	0,00%	13,33%	20,00%	0,00%	0,00%	0,00%	0,00%	**66,66%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	6,66%	0,00%	6,66%	13,33%	0,00%	**73,33%**	
Taux	73,33%	60,00%	46,66%	60,00%	46,66%	53,33%	60,00%	53,33%	66,66%	73,33%	***59,330%***

Matrice de confusion TEST

- ***M = 6***

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**100,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
wahed	10,00%	**80,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	
ithnani	0,00%	0,00%	**90,00%**	10,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
thalatha	0,00%	0,00%	20,00%	**70,00%**	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**80,00%**	0,00%	0,00%	20,00%	0,00%	0,00%	
khamssa	0,00%	0,00%	0,00%	0,00%	0,00%	**70,00%**	20,00%	10,00%	0,00%	0,00%	
sitta	10,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	0,00%	0,00%	10,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	
thamania	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	20,00%	0,00%	**70,00%**	
Taux	100,00%	80,00%	90,00%	70,00%	80,00%	70,00%	80,00%	90,00%	90,00%	70,00%	***82,000%***

Matrice de confusion APPRENTISSAGE

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**73,33%**	6,66%	0,00%	0,00%	0,00%	0,00%	13,33%	0,00%	0,00%	6,66%	
wahed	13,33%	**46,66%**	0,00%	0,00%	6,66%	20,00%	0,00%	0,00%	6,66%	6,66%	
ithnani	0,00%	6,66%	**40,00%**	26,66%	0,00%	6,66%	0,00%	0,00%	20,00%	0,00%	
thalatha	0,00%	0,00%	13,33%	**60,00%**	6,66%	0,00%	0,00%	0,00%	20,00%	0,00%	
arbaa	6,66%	0,00%	0,00%	0,00%	**53,33%**	0,00%	0,00%	26,66%	0,00%	13,33%	
khamssa	13,33%	6,66%	0,00%	0,00%	0,00%	**53,33%**	13,33%	6,66%	0,00%	13,33%	
sitta	0,00%	0,00%	0,00%	0,00%	0,00%	6,66%	**53,33%**	13,33%	0,00%	26,66%	
sabaa	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	6,66%	**60,00%**	0,00%	13,33%	
thamania	6,66%	0,00%	13,33%	26,66%	0,00%	0,00%	0,00%	0,00%	**53,33%**	0,00%	
tissaa	6,66%	0,00%	0,00%	0,00%	0,00%	6,66%	13,33%	20,00%	0,00%	**53,33%**	
Taux	73,33%	46,66%	40,00%	60,00%	53,33%	53,33%	53,33%	60,00%	53,33%	53,33%	***54,664%***

Matrice de confusion TEST

➢ ***M = 7***

CHFFRES	Models HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**90,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	
wahed	10,00%	**80,00%**	0,00%	10,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
ithnani	0,00%	0,00%	**90,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	
thalatha	0,00%	0,00%	10,00%	**80,00%**	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**70,00%**	0,00%	0,00%	20,00%	0,00%	10,00%	
khamssa	0,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	10,00%	10,00%	0,00%	0,00%	
sitta	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**100,00%**	0,00%	0,00%	0,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	30,00%	0,00%	0,00%	**70,00%**	0,00%	0,00%	
thamania	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	10,00%	**80,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	**90,00%**	
Taux	90,00%	80,00%	90,00%	80,00%	70,00%	80,00%	100,00%	70,00%	80,00%	90,00%	***83,000%***

Matrice de confusion APPRENTISSAGE

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**60,00%**	13,00%	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	0,00%	6,66%	
wahed	13,33%	**53,33%**	0,00%	0,00%	6,66%	0,00%	0,00%	13,33%	0,00%	13,33%	
ithnani	0,00%	6,66%	**33,33%**	26,66%	0,00%	6,66%	6,66%	0,00%	20,00%	0,00%	
thalatha	0,00%	6,66%	13,33%	**53,33%**	0,00%	6,66%	0,00%	0,00%	20,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	6,66%	**53,33%**	0,00%	0,00%	20,00%	0,00%	20,00%	
khamssa	6,66%	0,00%	0,00%	0,00%	0,00%	**60,00%**	13,33%	13,33%	0,00%	6,66%	
sitta	6,66%	0,00%	0,00%	0,00%	0,00%	0,00%	**66,66%**	13,33%	0,00%	13,33%	
sabaa	6,66%	0,00%	0,00%	0,00%	6,66%	0,00%	0,00%	**66,66%**	0,00%	20,00%	
thamania	0,00%	6,66%	13,33%	40,00%	0,00%	6,66%	0,00%	0,00%	**33,33%**	0,00%	
tissaa	6,66%	6,66%	0,00%	0,00%	0,00%	0,00%	13,33%	20,00%	0,00%	**53,33%**	
Taux	60,00%	53,33%	33,33%	53,33%	53,33%	60,00%	66,66%	66,66%	33,33%	53,33%	***53,330%***

Matrice de confusion TEST

- ***M = 8***

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**90,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	
wahed	10,00%	**90,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
ithnani	0,00%	0,00%	**80,00%**	20,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
thalatha	0,00%	0,00%	20,00%	**70,00%**	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**80,00%**	0,00%	0,00%	20,00%	0,00%	0,00%	
khamssa	0,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	10,00%	0,00%	0,00%	0,00%	
sitta	10,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	0,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**100,00%**	0,00%	0,00%	
thamania	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	10,00%	0,00%	**80,00%**	
Taux	90,00%	90,00%	80,00%	70,00%	80,00%	90,00%	90,00%	100,00%	90,00%	80,00%	***86,000%***

Matrice de confusion APPRENTISSAGE

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**73,33%**	6,66%	0,00%	0,00%	0,00%	0,00%	13,33%	6,66%	0,00%	0,00%	
wahed	13,33%	**53,33%**	6,66%	0,00%	6,66%	13,33%	0,00%	0,00%	0,00%	6,66%	
ithnani	6,66%	0,00%	**33,33%**	33,33%	0,00%	0,00%	0,00%	0,00%	26,66%	0,00%	
thalatha	0,00%	0,00%	6,66%	**66,66%**	0,00%	6,66%	0,00%	0,00%	20,00%	0,00%	
arbaa	0,00%	6,66%	0,00%	0,00%	**53,33%**	0,00%	0,00%	26,66%	0,00%	13,33%	
khamssa	6,66%	0,00%	0,00%	0,00%	0,00%	**60,00%**	13,33%	6,66%	0,00%	13,33%	
sitta	0,00%	0,00%	0,00%	0,00%	0,00%	13,33%	**66,66%**	6,66%	0,00%	13,33%	
sabaa	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	6,66%	**60,00%**	0,00%	13,33%	
thamania	6,66%	13,33%	13,33%	26,66%	0,00%	0,00%	0,00%	0,00%	**40,00%**	0,00%	
tissaa	6,66%	0,00%	0,00%	0,00%	0,00%	6,66%	6,66%	13,33%	0,00%	**66,66%**	
Taux	80,00%	73,33%	53,33%	73,33%	66,66%	66,66%	73,33%	73,33%	60,00%	73,33%	***69,330%***

Matrice de confusion TEST

- ***M = 9***

CHIFFRES	Modèles HMMi										
	HMM0	*HMM1*	*HMM2*	*HMM3*	*HMM4*	*HMM5*	*HMM6*	*HMM7*	*HMM8*	*HMM9*	
siffer	**90,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	0,00%	
wahed	10,00%	**80,00%**	10,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	
ithnani	0,00%	0,00%	**90,00%**	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	
thalatha	0,00%	0,00%	0,00%	**80,00%**	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**80,00%**	0,00%	0,00%	10,00%	0,00%	10,00%	
khamssa	0,00%	0,00%	0,00%	0,00%	0,00%	**100,00%**	0,00%	0,00%	0,00%	0,00%	
sitta	10,00%	0,00%	0,00%	0,00%	0,00%	0,00%	**90,00%**	0,00%	0,00%	0,00%	
sabaa	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	0,00%	**70,00%**	0,00%	10,00%	
thamania	0,00%	0,00%	10,00%	10,00%	0,00%	0,00%	0,00%	0,00%	**80,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	10,00%	0,00%	0,00%	**90,00%**	
Taux	90,00%	80,00%	90,00%	80,00%	100,00%	80,00%	90,00%	70,00%	80,00%	90,00%	***85,000%***

Matrice de confusion APPRENTISSAGE

CHIFFRES	Modèles HMMi										
	HMM0	***HMM1***	***HMM2***	***HMM3***	***HMM4***	***HMM5***	***HMM6***	***HMM7***	***HMM8***	***HMM9***	
siffer	**66,66%**	13,33%	0,00%	0,00%	0,00%	0,00%	13,33%	0,00%	0,00%	6,66%	
wahed	13,33%	**60,00%**	0,00%	0,00%	6,66%	13,33%	0,00%	0,00%	0,00%	6,66%	
ithnani	0,00%	6,66%	**40,00%**	20,00%	0,00%	6,66%	0,00%	0,00%	26,66%	0,00%	
thalatha	6,66%	0,00%	20,00%	**53,33%**	0,00%	0,00%	0,00%	0,00%	20,00%	0,00%	
arbaa	0,00%	0,00%	0,00%	0,00%	**66,66%**	0,00%	0,00%	20,00%	0,00%	13,33%	
khamssa	6,66%	0,00%	0,00%	0,00%	0,00%	**53,33%**	20,00%	6,66%	0,00%	13,33%	
sitta	6,66%	0,00%	0,00%	0,00%	0,00%	0,00%	**66,66%**	6,66%	0,00%	20,00%	
sabaa	13,33%	6,66%	0,00%	0,00%	26,66%	6,66%	0,00%	**33,33%**	0,00%	20,00%	
thamania	0,00%	0,00%	13,33%	20,00%	6,66%	0,00%	0,00%	0,00%	**60,00%**	0,00%	
tissaa	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	6,66%	20,00%	0,00%	**73,33%**	
Taux	66,66%	60,00%	40,00%	53,33%	66,66%	53,33%	66,66%	33,33%	60,00%	73,33%	***57,330%***

Matrice de confusion TEST

ANNEXE 3

Les tableaux suivants présentent les résultats des taux de reconnaissance automatique de la parole *acoustiques et visuelles*, pour un nombre d'états fixe *''N = 3''* et un nombre de mixtures *''M''* différents (1 à 9) ; utilisant des paramètres acoustiques de type MFCC et des paramètres visuelles de type DCT (basant sur l'amplitude).

- ***M = 1***

	Résultats acoustiques				Résultats vidéo			
chiffres	***App***	***Test***	***N***	***M***	***App***	***Test***	***N***	***M***
siffer	70%	66,66%	3	1	60%	53,33%	3	1
wahed	70%	46,66%	3	1	60%	46,66%	3	1
ithnani	60%	40,00%	3	1	70%	26,66%	3	1
thalatha	70%	60,00%	3	1	70%	70,00%	3	1
arbaa	50%	46,66%	3	1	50%	33,33%	3	1
khamsa	60%	40,00%	3	1	60%	46,66%	3	1
sitta	90%	66,66%	3	1	70%	53,33%	3	1
sabaa	80%	60,00%	3	1	80%	53,33%	3	1
thamania	60%	53,33%	3	1	70%	60,00%	3	1
tissaa	80%	46,66%	3	1	70%	53,33%	3	1
taux	**69%**	**52,663%**			**66%**	**49,663%**		

- ***M =2***

	Résultats acoustiques				Résultats vidéo			
chiffres	***App***	***Test***	***N***	***M***	***App***	***Test***	***N***	***M***
siffer	80%	70,00%	3	2	90%	73,33%	3	2
wahed	90%	70,00%	3	2	80%	60,00%	3	2
ithnani	80%	46,66%	3	2	70%	40,00%	3	2
thalatha	70%	60,00%	3	2	60%	66,66%	3	2
arbaa	80%	53,33%	3	2	80%	46,66%	3	2
khamsa	90%	66,66%	3	2	80%	53,33%	3	2
sitta	60%	53,33%	3	2	70%	46,66%	3	2
sabaa	80%	73,33%	3	2	60%	53,33%	3	2
thamania	70%	53,33%	3	2	80%	60,00%	3	2
tissaa	90%	60,00%	3	2	70%	53,33%	3	2
taux	**79%**	**60,664%**			**74%**	**55,330%**		

- ***M =3***

	Résultats acoustiques				Résultats vidéo			
chiffres	*App*	*Test*	*N*	*M*	*App*	*Test*	*N*	*M*
siffer	90%	73,33%	3	3	60%	66,66%	3	3
wahed	70%	53,33%	3	3	60%	46,66%	3	3
ithnani	80%	46,66%	3	3	90%	26,66%	3	3
thalatha	50%	60,00%	3	3	70%	53,33%	3	3
arbaa	80%	53,33%	3	3	70%	53,33%	3	3
khamsa	70%	60,00%	3	3	60%	46,66%	3	3
sitta	80%	46,66%	3	3	50%	53,33%	3	3
sabaa	80%	40,00%	3	3	70%	46,66%	3	3
thamania	70%	66,66%	3	3	80%	53,33%	3	3
tissaa	60%	60,00%	3	3	70%	53,30%	3	3
taux	**73%**	**55,997%**			**68%**	**49,992%**		

- ***M =4***

	Résultats acoustiques				Résultats vidéo			
Chiffres	*App*	*Test*	*N*	*M*	*App*	*Test*	*N*	*M*
siffer	90%	66,66%	3	4	80%	53,33%	3	4
wahed	70%	60,00%	3	4	70%	46,66%	3	4
ithnani	90%	53,33%	3	4	80%	53,33%	3	4
thalatha	70%	73,33%	3	4	70%	46,66%	3	4
arbaa	80%	53,33%	3	4	60%	40,00%	3	4
khamsa	70%	66,66%	3	4	70%	33,33%	3	4
sitta	80%	60,00%	3	4	80%	53,33%	3	4
sabaa	80%	60,00%	3	4	80%	46,66%	3	4
thamania	90%	73,33%	3	4	60%	60,00%	3	4
tissaa	90%	46,66%	3	4	90%	66,66%	3	4
taux	**81%**	**61,330%**			**74%**	**49,996%**		

- ***M =5***

	Résultats acoustiques				Résultats vidéo			
Chiffres	*App*	*Test*	*N*	*M*	*App*	*Test*	*N*	*M*
siffer	100%	73,33%	3	5	80%	66,66%	3	5
wahed	80%	73,33%	3	5	70%	53,33%	3	5
ithnani	90%	53,33%	3	5	80%	46,66%	3	5
thalatha	80%	60,00%	3	5	60%	73,33%	3	5
arbaa	90%	46,66%	3	5	70%	60,00%	3	5
khamsa	70%	60,00%	3	5	90%	40,00%	3	5
sitta	90%	46,66%	3	5	80%	53,33%	3	5
sabaa	80%	73,33%	3	5	80%	73,33%	3	5
thamania	60%	60,00%	3	5	70%	53,33%	3	5
tissaa	80%	60,00%	3	5	80%	53,33%	3	5
taux	**82%**	**60,664%**			**76%**	**57,330%**		

➢ ***M =6***

	Résultats acoustiques				Résultats vidéo			
Chiffres	**App**	**Test**	**N**	**M**	**App**	**Test**	**N**	**M**
siffer	90%	53,33%	3	6	80%	53,33%	3	6
wahed	80%	53,33%	3	6	70%	33,33%	3	6
ithnani	70%	60,00%	3	6	60%	40,00%	3	6
thalatha	80%	53,33%	3	6	60%	53,33%	3	6
arbaa	90%	46,66%	3	6	70%	46,66%	3	6
khamsa	90%	53,33%	3	6	80%	40,00%	3	6
sitta	70%	40,00%	3	6	80%	53,33%	3	6
sabaa	80%	66,66%	3	6	50%	53,33%	3	6
thamania	60%	53,33%	3	6	70%	33,33%	3	6
tissaa	80%	46,66%	3	6	50%	53,33%	3	6
taux	**79%**	**52,663%**			**67%**	**45,997%**		

➢ ***M =7***

	Résultats acoustiques				Résultats vidéo			
Chiffres	***App***	***Test***	***N***	***M***	***App***	***Test***	***N***	***M***
siffer	60%	60,00%	3	7	70%	46,66%	3	7
wahed	80%	53,33%	3	7	60%	53,33%	3	7
ithnani	90%	60,00%	3	7	90%	33,33%	3	7
thalatha	80%	53,33%	3	7	80%	46,66%	3	7
arbaa	90%	73,33%	3	7	70%	40,00%	3	7
khamsa	70%	66,66%	3	7	80%	53,33%	3	7
sitta	80%	53,33%	3	7	60%	46,66%	3	7
sabaa	60%	53,33%	3	7	50%	66,66%	3	7
thamania	70%	33,33%	3	7	80%	40,00%	3	7
tissaa	90%	46,66%	3	7	90%	40,00%	3	7
taux	**77%**	**55,330%**			**73%**	**46,663%**		

➢ ***M =8***

	Résultats acoustiques				Résultats vidéo			
Chiffres	***App***	***Test***	***N***	***M***	***App***	***Test***	***N***	***M***
siffer	70%	73,33%	3	8	70%	66,66%	3	8
wahed	90%	60,00%	3	8	90%	46,66%	3	8
ithnani	80%	46,66%	3	8	80%	33,33%	3	8
thalatha	60%	73,33%	3	8	60%	53,33%	3	8
arbaa	90%	53,33%	3	8	70%	46,66%	3	8
khamsa	70%	46,66%	3	8	80%	60,00%	3	8
sitta	80%	66,66%	3	8	50%	60,00%	3	8
sabaa	90%	60,00%	3	8	80%	53,33%	3	8
thamania	80%	53,33%	3	8	90%	26,66%	3	8
tissaa	90%	53,33%	3	8	70%	53,33%	3	8
taux	**80%**	**58,663%**			**74%**	**49,996%**		

- ***M =9***

	Résultats acoustiques				Résultats vidéo			
Chiffres	**App**	**Test**	**N**	**M**	**App**	**Test**	**N**	**M**
siffer	80%	66,66%	3	9	80%	40,00%	3	9
wahed	60%	53,33%	3	9	70%	60,00%	3	9
ithnani	70%	40,00%	3	9	80%	33,33%	3	9
thalatha	80%	53,33%	3	9	70%	46,66%	3	9
arbaa	90%	73,33%	3	9	80%	53,33%	3	9
khamsa	60%	53,33%	3	9	90%	60,00%	3	9
sitta	90%	60,00%	3	9	70%	53,33%	3	9
sabaa	80%	66,66%	3	9	60%	26,66%	3	9
thamania	90%	53,33%	3	9	80%	46,66%	3	9
tissaa	80%	60,00%	3	9	60%	53,33%	3	9
taux	**78%**	**57,997%**			**74%**	**47,330%**		