Réalisation 6

Comme tout modèle qui doit être expérimenté, le présent chapitre constitue un cadre d'expérimentation et d'argumentation du chapitre précédent.

Nous allons présenter dans ce chapitre les expérimentations que nous avons menées pour aller vers une collaboration des processus de reconnaissance automatique de la parole et de reconnaissance visuelle de la parole.

Nous présentons à présent les différents tests que nous avons effectués afin d'analyser les mérites des méthodes retenues dans le chapitre précédent. Les plus performantes seront validées par comparaison avec des algorithmes d'apprentissage classiquement utilisés dans la littérature.

6.1 Architecture général du système de reconnaissance

Dans nos expérimentations nous évaluons la performance des modèles audio-visuels HMM appris en utilisant les GA par rapport à l'apprentissage standard des HMM en utilisant une estimation du maximum de vraisemblance (EM).

Comme l'a fait remarquer (Alpaydin 2004), nous devons toujours garder à l'esprit que les conclusions que nous tirons de l'analyse est conditionnée par l'ensemble de données. Ainsi, nous ne comparons pas les modèles et les algorithmes d'apprentissage d'une manière indépendante de domaine. Tout résultat nous présentons n'est valable que pour l'application particulière de AVASR et pour l'ensemble de données utilisé. Comme indiqué dans le Non déjeuner théorème de gratuit (Wolpert and Macready 1997) il n'y a pas une telle chose comme le "meilleur" algorithme d'apprentissage en général. Pour n'importe quel algorithme d'apprentissage, il y aura un ensemble de données où il est très précis et une autre ou il est très faible. Ainsi, nos résultats ne sont valables que pour l'application particulière d'AVASR et en particulier pour les corpus de données que nous avons choisis. Ces corpus de données sont discutés par la suite.

Dans notre travail nous allons appliquer l'algorithme de clustering K-means sur les BDD audiovisuelles CUAVE et notre propre BDD arabe (AVARB), les résultats de cette opération seront en suite introduits au HMM pour faire l'apprentissage. Afin d'augmenter la performance du système de reconnaissance proposé, nous avons utilisé une nouvelle méthode basée sur l'hybridation des deux paradigmes HMM et GA.

Pour réaliser ce système de reconnaissance, il fallait :

- Détection de visage et Localisation des lèvres dans les scènes vidéo en utilisant la méthode Viola-Jones.
- Extraction de paramètres acoustiques avec la méthode RASTA-PLP.
- Extraction de paramètres visuels avec la méthode DCT.
- Réaliser une quantification vectorielle et dégager des classes, en utilisant l'approche suivante : K-means.
- Phase d'apprentissage en utilisant les modèles HMM, et GA/HMM.
- Comparaison des taux de reconnaissance obtenus pour tirer la méthode la plus performante de reconnaissance.

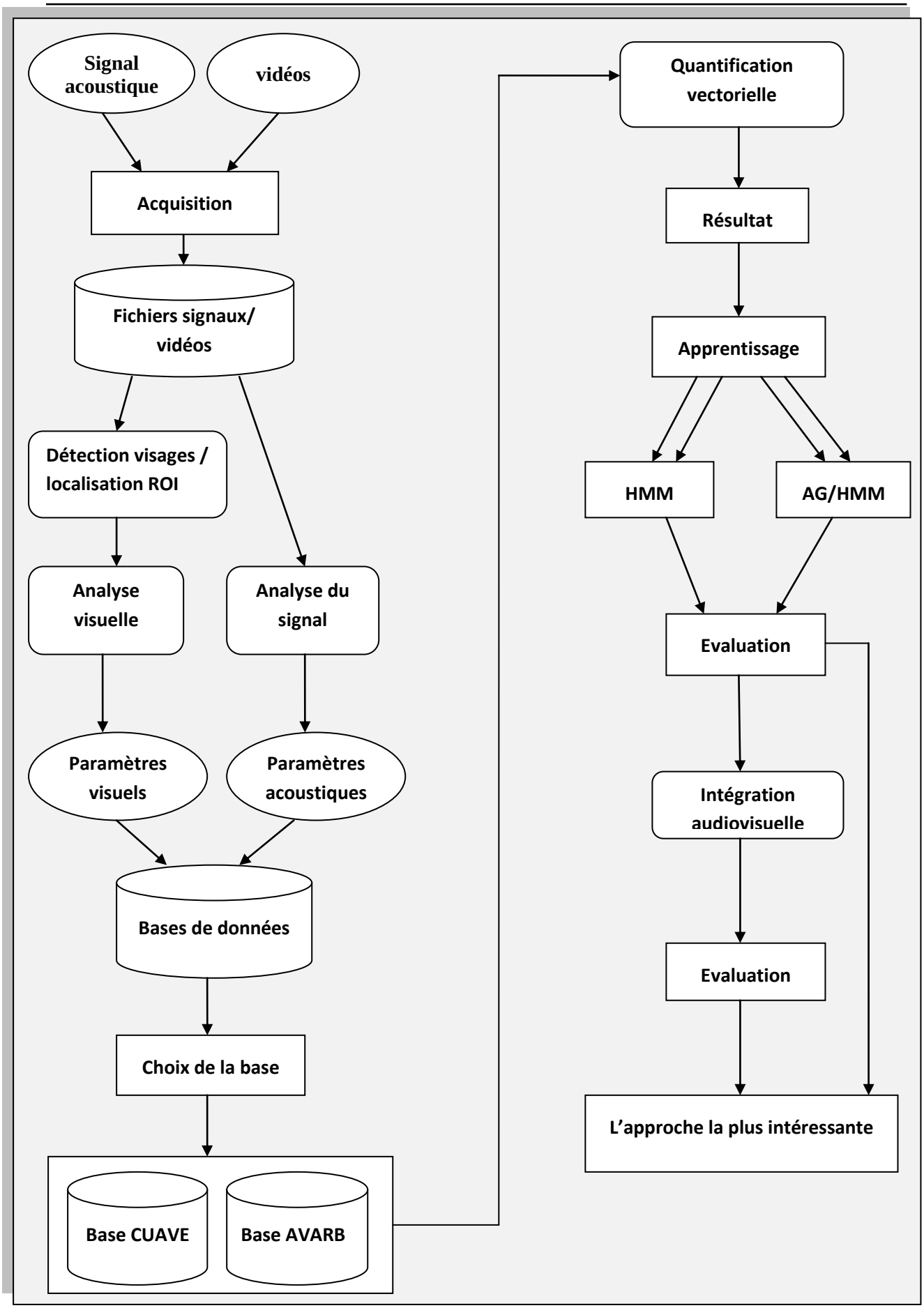


Figure 6.1 – Architecture générale du système proposé.

6.2 Base de données utilisée

6.2.1 Les bases de données audiovisuelle arabe

Dans notre travail nous avons utilisé notre propre base de données audiovisuelle de parole arabe : cette base de données multi-locuteurs a été enregistrée dans un milieu réel (une salle de cours très bruyante), Nous visons de plus la diversité des données pour un apprentissage bien amélioré, les vidéos sont capturées La à une distance moyenne égale à 16.5 cm avec une résolution de 690×450 pixel et à 30 trames/sec et avec des variations de pose (vue de profil, de face) pour un ensemble de 18 locuteurs (16 garçon et 2 filles) sauvegardées avec l'extension « .avi », alors que les fichiers audio sont sauvegardé avec l'extension « .wav », l'échantillonnage standard après des testes réalisés au sein de notre laboratoire est 16 KHz MONO (à un canal unique) car il est optimal de calculer les coefficients issus d'un signal acoustique à paramètres unique.

Notre base AVARB contient 2 corpus, le premier corpus contient des prononciations des chiffres arabes isolés (de zéro (0) à neuf (9)), alors que le deuxième corpus contient un ensemble commandes en arabe (25 mots), comme il est illustré dans le tableau 6.1 :

Corpus chiffre				***Corpus commandes***			
code	*Prononciation*	*Ecriture arabe*	*glossaire français*	*code*	*Prononciation*	*Ecriture arabe*	*Glossaire français*
1	Siffer	صفر	Zéro	1	Marhaban	مرحبا	Bienvenue
2	Wahed	واحد	Un	2	Ebdaa	ابدأ	Démarrer
3	Ithnani	اثنان	Deux	3	Iqaf	إيقاف	Arrêter
4	Thalatha	ثلاثة	Trois	4	Eftah	افتح	Ouvrir
5	Arbaa	أربعة	Quatre	5	Arliq	أغلق	Fermer
6	Khamssa	خمسة	Cinq	6	Takbir	تكبير	Agrandir
7	Sitta	ستة	Six	7	Tasrir	تصغير	Réduire
8	Sabaa	سبعة	Sept	8	Tashril	تشغيل	Fonctionnement
9	Thamania	ثمانية	Huit	9	Elraa	إلغاء	Annuler
10	Tissaa	تسعة	Neuf	10	Bahth	بحث	Recherche
				11	Ekhtiyar	اختيار	Sélection
				12	Aaouda	عودة	Retour
				13	Edhar	إظهار	Affichage
				14	Qaima	قائمة	Liste
				15	Mouafiq	موافق	Accepter

				16	Doukhoul	دخول	Se connecter
				17	Khourouj	خروج	Quitter
				18	Nasskh	نسخ	Copier
				19	Qass	قص	Couper
				20	Lasq	لصق	Coller
				21	Tarjama	ترجمة	Traduire
				22	Khasaiss	خصائص	Propriétés
				23	Tatbiq	تطبيق	Application
				24	Tenfid	تنفيذ	Exécution
				25	Tahmil	تحميل	Chargement

Table 6.1 – Notre deux corpus proposés de chiffres et commandes arabes.

Les locuteurs sont de différentes régions dialectes algériennes, et chaque locuteur prononce chaque mot 9 fois avec différentes modes de prononciation (normal, lente, et rapide). Dans notre corpus basic qui contient que des mots isolés, la taille de chaque enregistrement est 2 secondes qui est un temps suffisant pour prononcer un mot lentement en arabe. La figure suivante montre quelques trames de notre base AVARB :

Figure 6.2 – quelques exemples de trames de notre base audiovisuelle AVARB.

6.2.2 La base de données CUAVE

Elle se compose de 36 locuteurs, 19 hommes et 17 femmes, poussant chiffres isolés et continue. Les vidéos des orateurs sont enregistrées en profil frontal, et pendant le mouvement. La base de données CUAVE contient environ 3 heures de parole enregistrées par une caméra Mini DV. La Vidéo a ensuite été compressée en MPEG-2 fichiers (audio stéréo à un taux d'échantillonnage 44 kHz, 16-bit). Il comprend également des fichiers audio vérifiés pour la synchronisation (taux de mono de 16 kHz, 16-bit) et des fichiers d'annotation (Patterson et al. 2002).

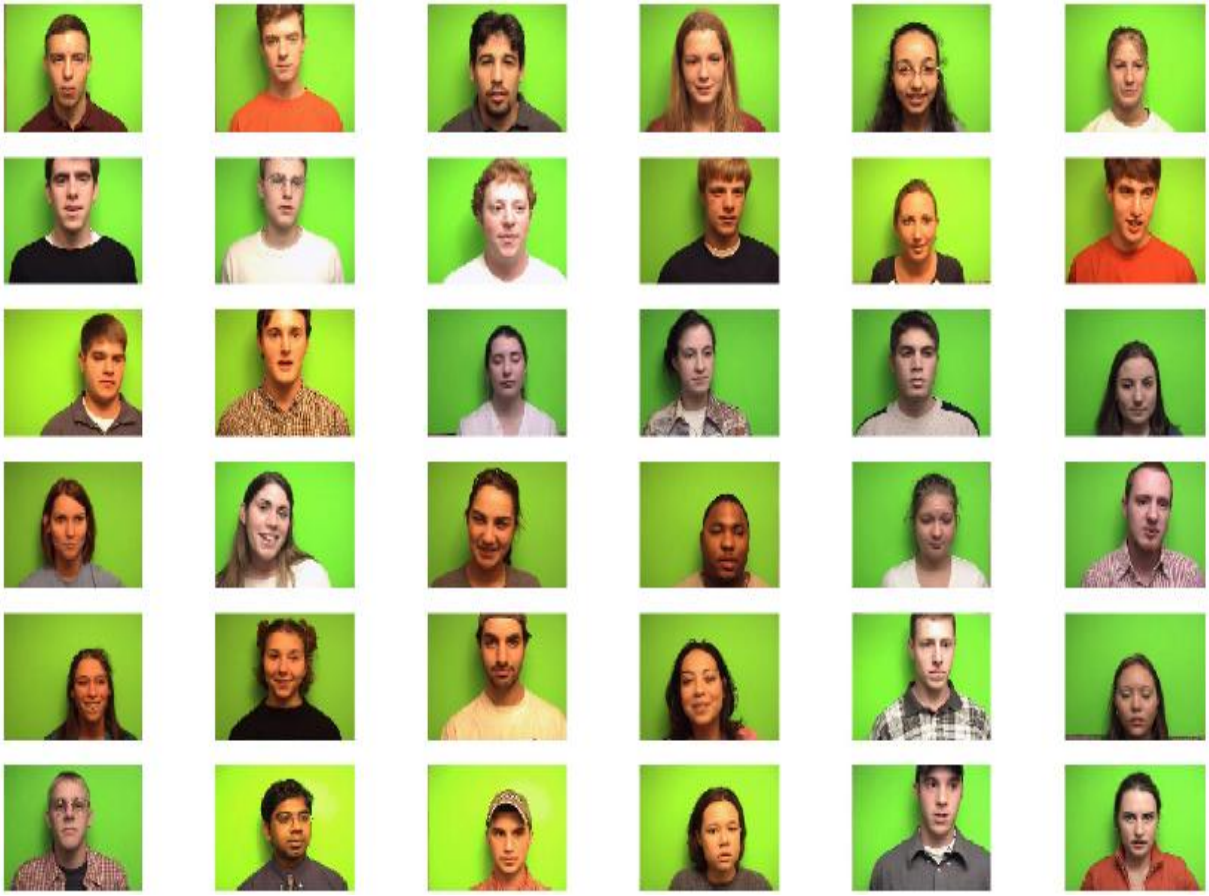


Figure 6.3 – Exemples de trames de la base CUAVE.

6.3 Validation du système

Une étape importante et très consommatrice en temps de développement d'un système de transcription est l'expérimentation. Il s'agit de tester les différents modules du système pour ajuster leurs paramètres. De bonnes valeurs de paramètres peuvent apporter beaucoup au niveau du taux de reconnaissance. Chaque module a ses propres paramètres et il est nécessaire de les ajuster de façon plus ou moins optimale. Ajuster les paramètres de tous les modules en même

temps est une tache irréalisable puisque le nombre de combinaisons de paramètres à tester serait très grand et donc le temps d'expérimentation serait énorme. En général, l'expérimentation est effectuée module par module pour économiser du temps. Puis le système complet est testé également.

6.4 Traitement des données audiovisuelles

6.4.1 Séparation audiovisuelle

Une fois l'enregistrement des séquences vidéo du locuteur est réalisé à l'aide d'un appareil photo numérique Sony Cyber-Shot DSC-W530 14.1 Méga Pixel avec un zoom optique 4x grand-angle Zoom optique et 2.7 pouces moniteur LCD. La première opération consiste à la séparation des deux flux audio et vidéo. Le flux audio est extrait sous forme d'un signal à l'aide du logiciel Gold Wave de l'extension ''.wav'', et à partir du flux vidéo on extrait, à l'aide du logiciel BPS, des images fixes de la séquence. On passe ensuite à la construction des bases de données audio et vidéo.

6.4.2 Données visuels

Après la détection de visage avec l'utilisation de l'algorithme de Viola-Jones (voir l'exemple dans la figure 6.4), nous avons localisé la région de la bouche de chaque locuteur comme il est illustré dans les exemples dans la figure 6.5.

(a) (b) (c)

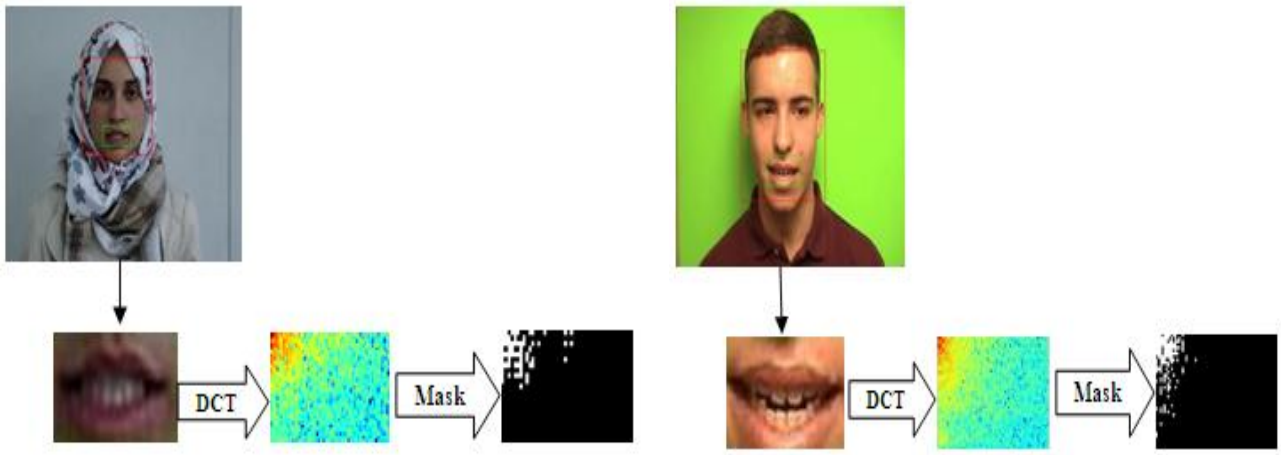
Figure 6.4 – Un exemple de détection de visage : (a) image originale (b) détection de peau avec suppression de bruit (c) résultat de détection de visage.

(a)

(b)

Figure 6.5 – Exemples de la région de la bouche détectée à partir de : (a) la base AVARB (b) la base CUAVE.

L'extraction des caractéristiques vidéo est effectuée avec la 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 par la figure 6.6. Dans cette figure, nous avons conservé uniquement les 100 premiers coefficients de hautes amplitudes d'une image, donc le vecteur visuel dans ce cas est composé des 100 éléments. Le nombre de coefficients hautes amplitudes conservés après la transformation par la DCT est choisi de manière à conserver un maximum d'énergie totale dans les coefficients hautes amplitudes qui sera suffisant pour reconstituer les caractéristiques principales de l'image (Makhlouf et al. 2013a ; 2013b). L'énergie totale E de l'image est calculée (théorème de Parseval, à partir des coefficients de la DCT).



(a) (b)

Figure 6.6 – Le processus de sélection des coefficients DCT avec un échantillon à partir: (a) la base AVARB (b) la base CUAVE.

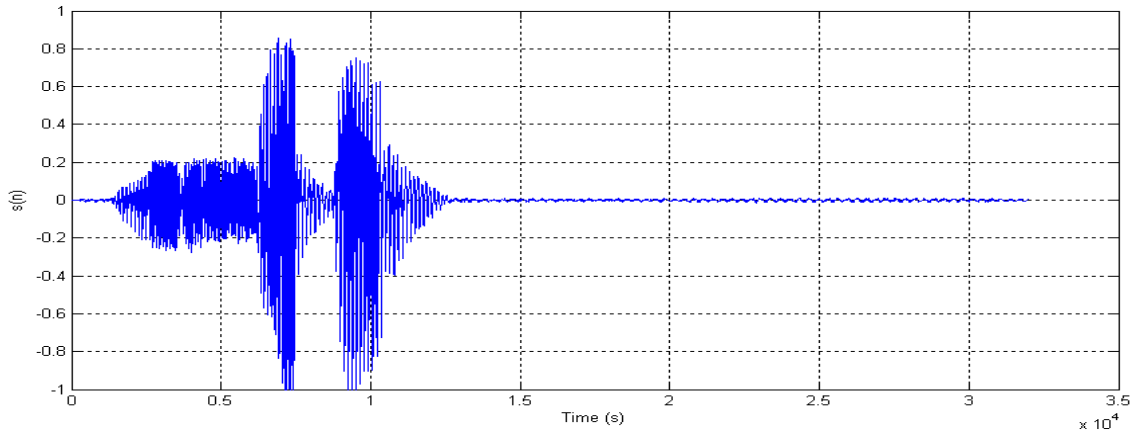
L'idée principale de l'algorithme pour encoder l'image par la DCT est de ne pas utiliser la totalité des coefficients (310500 coefficients), afin de limiter la taille mémoire et les calculs nécessaires pour l'entraînement et la reconnaissance par les modèles proposés dans notre système. Dans notre travail nous avons gardé les cent (100) premiers coefficients pour représenter l'image.

6.4.3 Données acoustiques

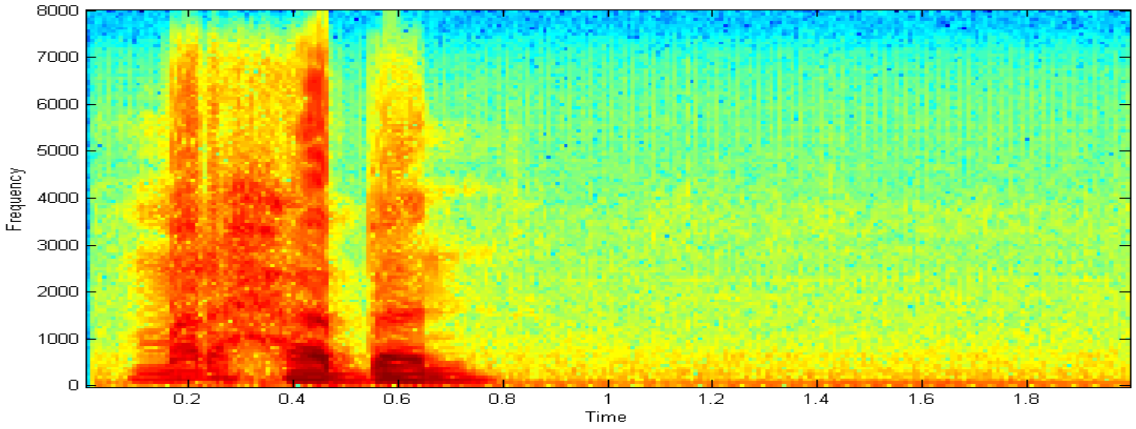
L'objectif d'un système de paramétrisation est d'extraire les informations caractéristiques du signal de parole en éliminant au maximum les parties redondantes. Pour la réalisation de cette phase d'extraction des paramètres, nous avons utilisé la technique RASTA-PLP (comme il est mentionné dans le chapitre 5).

Pour chaque signal vocal et avec la méthode RASTA-PLP, on extrait 9 paramètres du signal acoustique de 98 trames d'échantillonnage à 16kHz, et d'une taille de fenêtre 0.025 secondes et d'un pas de 0.010 secondes. En intégrant la première et la deuxième dérivé des paramètres, on obtient des matrice de 27 paramètres organisé comme suit : Pour chaque corpus multilocuteur, si on prend le corpus commandes par exemple, on a pour les tests 25 occurrences de commandes vocal répétés 3 fois chacune de 18 locuteurs, donc 25*3*18*27= 36450 et 98 trames est la taille de la matrice, même pour l'apprentissage, sauf que l'ordre de l'occurrence entre les locuteurs sont organisés les uns après les autres.

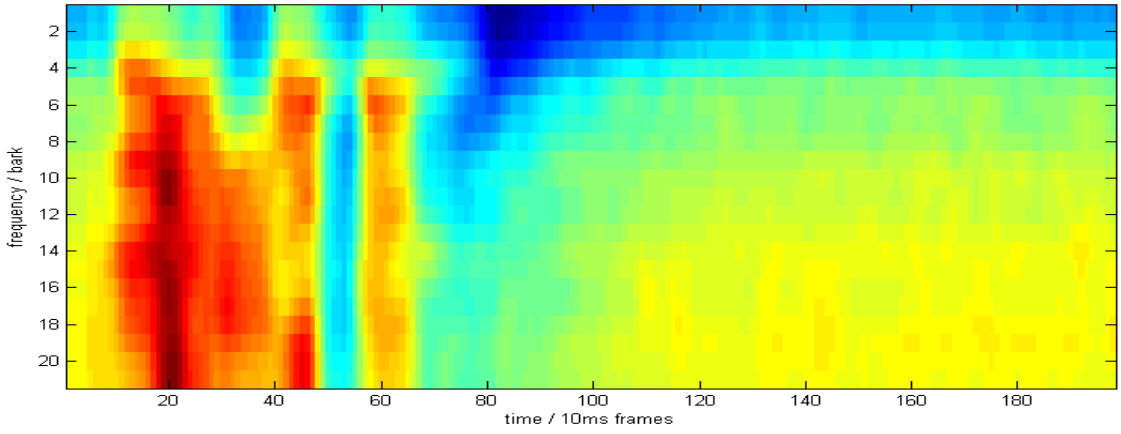
Un exemple de calcul de quelques paramètres du signal de parole utilisant cette méthode d'extraction est illustré par la figure 6.7.



(a)



(b)



(c)

Figure 6.7 – Exemple d'un signal de parole du mot arabe "/ marhaban /" (a) son spectrogramme (b) et l'ensemble des caractéristiques spectrales RASTA-PLP (c).

6.5 Modélisation par GA/HMM

Après avoir défini formellement notre approche, il est nécessaire de la tester afin de la valider.

6.5.1 Résultats obtenus et discussion

Cet algorithme optimise à la fois les paramètres (probabilités) de HMM. Il repose sur une recherche génétique d'un bon modèle parmi une population hétérogène de HMM et une optimisation par un algorithme de gradient (Baum-Welch).

Pour l'apprentissage, nous avons utilisé un nombre m des HMM de type gauche-droite avec un nombre m d'états dont m est le nombre des mots dans chaque corpus, afin de représenter les m classes.

6.5.1.1. Expérimentations avec des bruits sonore et visuel additifs

Dans cette section, nous présentons les résultats des expériences menées en utilisant des signaux audio et vidéo bruyants.

Nous avons utilisé deux types de bruit vidéo pour examiner la robustesse de notre système AVASR contrairement à audio seule ASR. Les types de bruit que nous avons implémenté sont la diminution des trames**,** et le bruit aléatoire gaussien. Ces types de bruit imitent des scénarios typiques où il existe une distorsion soit depuis un appareil photo défectueux ou d'un signal de transmission vidéo. De plus, La diminution de la fréquence de trames (FPS) et le bruit de bloc peut simuler la perte d'information à la suite des mouvements abrupts de la bouche et la parésie d'une partie de la bouche ou des lèvres qui peut être causée par un problème de santé. Par conséquent, ce type de bruit présente un intérêt dans des environnements d'assistance envahissants.

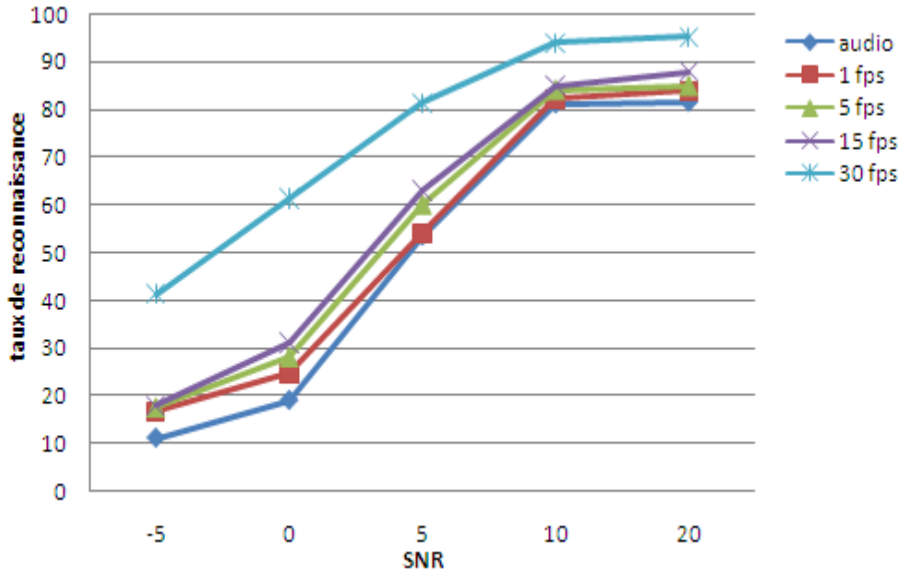
Le taux de reconnaissance est affecté par la qualité du signal (i.e. diminution du rapport signal sur bruit (Signal-to-Noise Ratio (SNR))). Nous examinons d'abord le cas de d'image perdue (**Frame-Dropped)**. La fréquence des trames initiale était 30 fps, donc nous avons réduit à 15, 5 et 1fps puis l'interpolée de nouveau à 100fps afin de correspondre au taux de caractéristique audio. Nos mesures sont présentées dans la figure 6.8(a).

Nous présentons aussi nos résultats expérimentaux sur notre système AVASR au cours d'une gamme de niveaux de bruit. Nous avons utilisé le bruit aléatoire gaussien pour dégrader la qualité de l'image. La valeur moyenne du bruit est 0 et l'écart type était 15, 30, 50 et

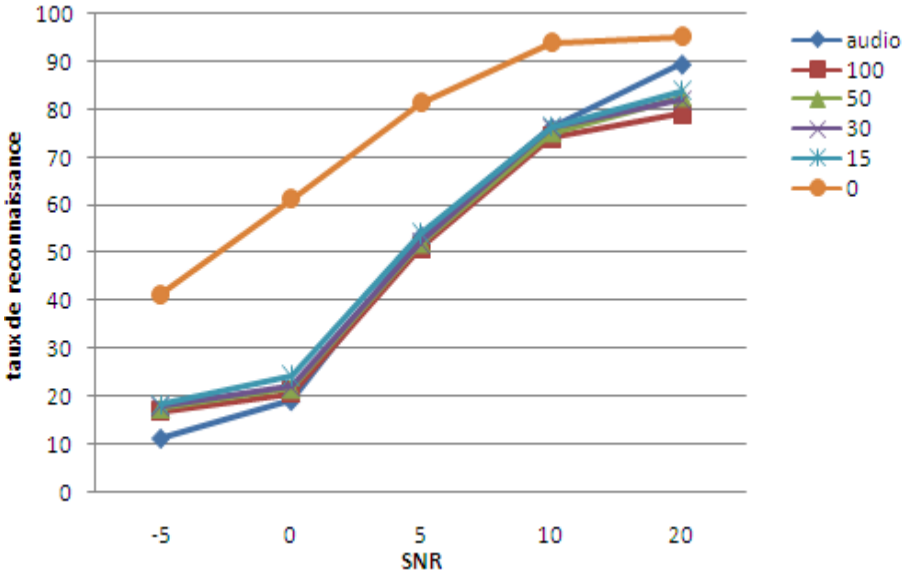
100 respectivement. L'effet du bruit sur la ROI peut être vu dans la figure 6.8 et les résultats dans la figure 6.8(b).

(A) (B) (C) (D) (E)

Figure 6.8 – ROI avec bruit gaussien, l'écart type =(A) 0 (B) 15 (C) 30 (D) 50 et (E) 100.



(a)



(b)

Figure 6.9 – La performance du système AVASR : (a) sous une fréquence des trames vidéo réduite (b) pour un bruit aléatoire gaussien.

Comme nous pouvons voir, les caractéristiques visuelles augmentent le taux de reconnaissance, même à 1fps. Plus précisément, le la performance est supérieur pour 5 (de 56.1% à 1FPS) et 0 db (24.8% à 1FPS) à celle du reconnaisseur audio-seul (53.5% et 19.1% respectivement). Comme le montre le graphique dans 6.8(b), le taux de reconnaissance pour le système AVASR est réduit pour 10db mais pour des valeurs plus basses du SNR, le système AVASR surpasse le système de reconnaissance audio-seul. Même à un écart type de 100, le système fonctionne mieux pour 0 et 5db atteindre un taux 19.1% et 57,3% respectivement.

6.5.1.2. Expérimentations avec un bruit réel

Nous avons présenté différentes sortes d'instance avec des paramètres de contrôle différents de GA qui ont été résolus par notre algorithme pour évaluer la performance du système proposé. Nous avons exécuté chaque instance 15 fois avec un nombre différent de clusters, des valeurs de probabilité de croisement entre 0.5-0.9, et une probabilité de mutation avec la valeur 0,01. De plus, nous prenons un nombre maximum d'itérations pour l'algorithme de Baum-Welch égale à 40, les valeurs moyennes de $P(o|\lambda)$ obtenue valeurs après 150 générations (le nombre d'itérations idéale pour des meilleurs performance) sont listés dans les Tables 6.2 et 6.3 comme suit:

Nombre de clusters	**P_c**	**P_m**	**Average $P(o|\lambda)$**
3	0.5	0.01	-2.3630
5	0.6	0.01	-1.5838
7	0.7	0.01	-1.1396
9	0.8	0.01	-3.3185
12	0.9	0.01	-4.0122

(a)

Nombre de clusters	**P_c**	**P_m**	**Average $P(o|\lambda)$**
3	0.5	0.01	-3.7416
5	0.6	0.01	-3.2604
7	0.7	0.01	-3.4235
9	0.8	0.01	-3.9134
12	0.9	0.01	-4.3637

(b)

Table 6.2 – paramètres GA pour l'entraînement du HMM pour l'audio seul: (a) base AVARB (b) base CUAVE.

Nombre de clusters	**P_c**	**P_m**	**Average $P(o|\lambda)$**
3	0.5	0.01	-7.7629
4	0.5	0.01	-7.0046
7	0.8	0.01	-7.1555
9	0.8	0.01	-7.6595
12	0.9	0.01	-7.8234

(a)

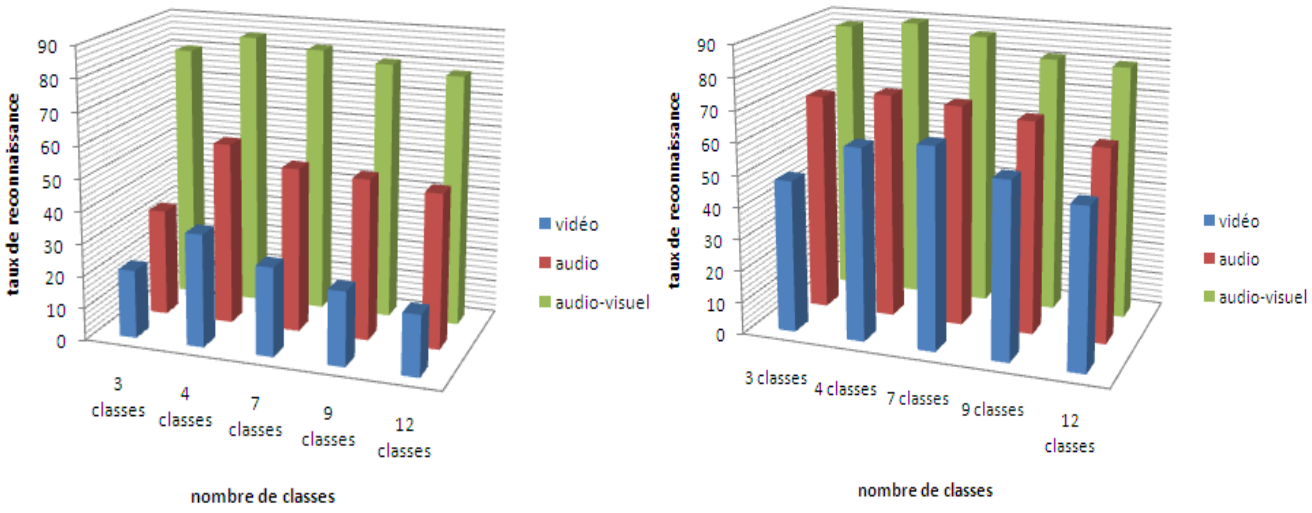
Nombre de clusters	**P_c**	**P_m**	**Average $P(o|\lambda)$**
3	0.5	0.01	-5.1860
5	0.6	0.01	-5.2987
7	0.7	0.01	-5.4743
9	0.8	0.01	-5.8747
12	0.9	0.01	-6.0890

(b)

Table 6.3 – paramètres GA pour l'entraînement du HMM pour le vidéo seul: (a) base AVARB (b) base CUAVE.

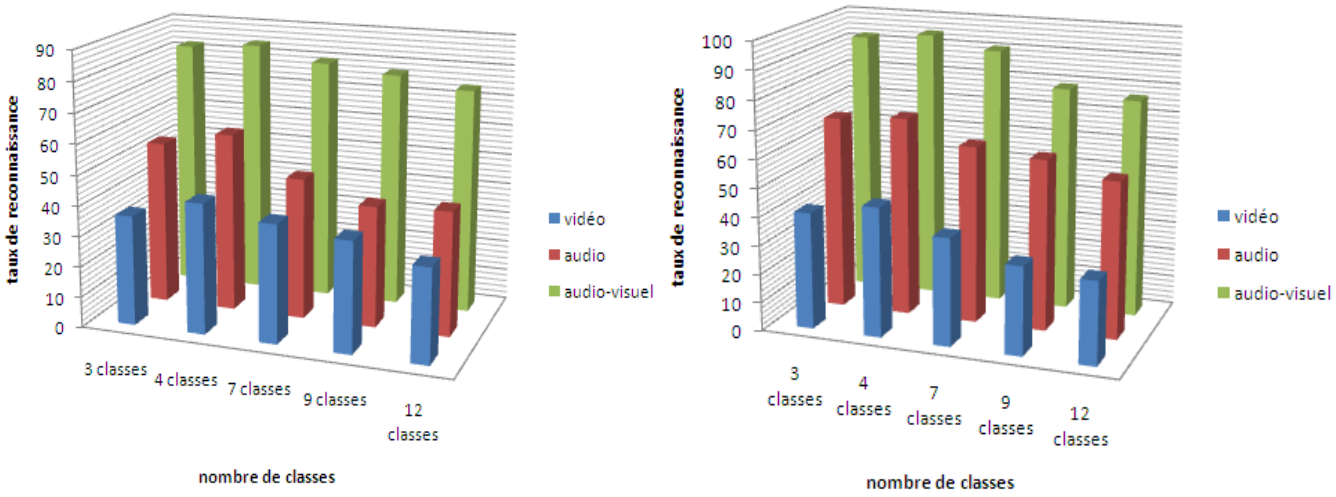
Nous observons que les résultats varient en fonction des paramètres d'entrainement de l'AG, également aux le nombre de clusters obtenu par la phase de quantification vectorielle, par exemple, avec 7 clusters, Pc = 0.7 et Pm = 0.01, pour la base AVARB audio et 5 clusters, Pc = 0.6 et Pm = 0.01 pour la base AVARB visuelle sont supérieures à toutes les autres approches dans tous les cas. Par conséquent, nous les utilisons dans notre GA/HMM. Les mêmes observations pour la base de données audio CUAVE avec 4 clusters, P_c = 0.6 et P_m = 0.01, et pour la base de données visuelle CUAVE la meilleure performance est obtenue avec 3 clusters, P_c = 0.5 et P_m = 0.01.

Les figures 10 et 11 donnent le taux de reconnaissance moyennes par rapport au nombre de clusters utilisés dans l'expérience.



(a) (b)

Figure 6.10 – Comparaison entre les taux de reconnaissances audio, vidéo, et audiovisuel, on utilisant : (a) HMM standard (b) GA/HMM pour la BDD AVARB.



(a) (b)

Figure 6.11 – Comparaison entre les taux de reconnaissances audio, vidéo, et audiovisuel, on utilisant : (a) HMM standard (b) GA/HMM pour la BDD CUAVE.

En se basant sur les figures 6.10 et 6.11, nous constatons que les taux de reconnaissance obtenus avec notre GA/HMM sont meilleurs dans la plupart des cas par rapport à ceux obtenus avec le HMM standard (Les figures ci-dessus indiquent également que le système AVASR avec une fusion des scores dépassent significativement en atteignant des taux de reconnaissance les plus élevés. Dans la figure 6.10, nous avons noté presque les mêmes observations précédentes avec notre base de données de AVARB, c'est à dire que nous avons trouvé le meilleur taux moyen de reconnaissance égale à 93,7% et 97,6% en utilisant le HMM standard (Young et al. 2006) et le modèle hybride GA/HMM respectivement, et avec 7 classes à la fois.

Pour la base de données CUAVE les résultats montrent que le taux moyen de reconnaissance atteint un meilleur taux avec 86,8% en utilisant le modèle HMM standard avec 5 classes pour la phase de classification, et 98,1% en utilisant le modèle GA/HMM avec 3 classes.

Plus généralement, nous avons trouvé une augmentation du pourcentage variant de presque 5% à 28% des résultats de nos tests, mais cette augmentation dans les taux de reconnaissance donnés n'est pas fixe, ainsi que avec l'augmentation de la taille de la population. Il se peut donner des taux pire ou les mêmes de celle du HMM standard avant les optimisations. Cela est dû à la caractéristique de la méthode GA qui est aléatoire et aussi que ce système utilise le processus général de remplacement standard.

6.6 Conclusion

Dans ce chapitre, nous avons présenté les caractéristiques techniques et les performances du système AVASR proposé. Les différents blocs matériels ainsi leur fonctionnement ont été détaillés.

Les résultats de l'évaluation (calcul d'erreur et les tests de reconnaissance) sont très satisfaisants et témoignent d'une grande fiabilité de mesures obtenues par ce système.

Les scores de reconnaissance obtenus ont montré que l'intégration des deux modalités acoustiques est visuelles sont supérieurs à ceux obtenus avec chaque modalité prise séparément, dans toutes les conditions expérimentales (niveau de bruit).