



REVUE ALGERIENNE DES SCIENCES

Section -A-
Sciences de la Nature et de la Vie
Sciences techniques

<http://univ-eltarf.dz/fr/>



Apprentissage évolutionnaire d'une chaîne de Markov cachée pour la reconnaissance automatique de la parole audiovisuelle

Makhlouf Amina^{a, b,*}, Bouhadada Tahar^b, Benmachiche Abdelmadjid^{a, b}, Maatallah Majda^{a, b}

^a Département d'informatique, Université Chadli Bendjedid- El Tarf, Algérie.

^b Laboratoire de Recherche en Informatique (LRI), Université Badji Mokhtar, Annaba, Algérie.

Informations	Résumé
Mots clés : La reconnaissance automatique de la parole fusion audiovisuelle information visuelle DCT RASTA-PLP GA-HMM optimisation. *Correspondance : Makhlouf_Amina@yahoo.fr Reçu le 12 mars 2018, Révisé le 24 mai 2018, Accepté le 25 mai 2018.	La présence de bruit de fond et des conditions variables (environnement, réverbération, types de microphones) peut affecter significativement la qualité de la reconnaissance automatique de la parole. La reconnaissance audio-visuelle propose d'améliorer les performances des systèmes de reconnaissance de la parole, spécialement lorsque le canal audio est corrompu, en ajoutant de l'information provenant de la modalité visuelle, sous la forme d'images vidéo du locuteur. Dans cet article, nous présentons un système de reconnaissance de la parole combinant les données acoustiques et les données visuelles. Le système AVASR proposé utilise les HMM comme moteur de reconnaissance. Toutefois, en raison de méthodes Expectation-Maximisation (EM) comme l'algorithme de Baum-Welch tendent à converger vers l'optimum local qui est le plus proche des valeurs initiales de la procédure d'optimisation, l'estimation des paramètres de la reconnaissance nécessite une examination plus attentif. Un algorithme génétique combiné avec le HMM (GA/HMM) est présenté ici pour résoudre ce problème. L'idée est de combiner un algorithme génétique pour explorer rapidement tout l'espace des solutions avec un algorithme de Baum-Welch pour trouver les valeurs des paramètres exactes de l'optimum. Les résultats expérimentaux montrent que le GA/HMM peut atteindre un taux plus élevé de reconnaissance avec moins de calculs par rapport à un HMM classique.

1. Introduction

La reconnaissance automatique de la parole s'applique à ce jour sur de nombreux signaux de qualité différente (fréquence d'échantillonnage, quantification, codage, conditions d'enregistrement).

Nous rappelons que la parole est l'un des moyens les plus naturels par lequel des personnes communiquent. La RAP a pour objet la transformation du signal acoustique en une séquence de mots qui, idéalement, correspond à la phase prononcée par un locuteur. Les systèmes de reconnaissance qui utilisent comme entrée uniquement le signal acoustique atteignent leurs limites surtout dans des cas de situations environnementales bruitées donc réelles. Dans ces cas, l'intégration de l'information visuelle dans le système de reconnaissance peut constituer une voie de solution. A cet effet nous nous intéressons à la mise en œuvre d'un système de reconnaissance intégrant conjointement les deux informations acoustique et visuelle des la parole

se sont focalisés sur une interaction sensorielle de type fusion ou intégration. A ce niveau, reste posée la question du ou 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 en vue de leur intégration.

La reconnaissance automatique de la parole audiovisuelle est née de l'idée que si l'homme exploite les informations provenant du visage du locuteur pour améliorer l'intelligibilité, la machine peut en faire autant, si d'une part le principe d'intégration des deux modalités est suffisamment bien connu, et si d'autre part les informations visuelles sont exploitées d'une façon optimale.

Notre travail de recherche tente de vérifier ces deux conditions, c'est-à-dire la modélisation de la perception audiovisuelle de la parole en vue d'une implémentation logicielle, et de l'extraction des informations les plus pertinentes. Notre étude a été au centre d'une

recherche pluridisciplinaire : de la psychologie cognitive aux techniques de traitement d'images couleurs, nous nous sommes investis dans le domaine de la paramétrisation des lèvres, le traitement du signal et la reconnaissance automatique des formes.

Ce papier est organisé comme suit. Dans la section I, une introduction au domaine de l'audiovisuel. Nous suivons cela avec une description du système proposé. Dans la section 2, la section 3, nous allons présenter toutes les méthodes utilisées dans ce travail. Les résultats expérimentaux sont illustrés dans la section 4. finalement, les conclusions et les travaux futurs sont formulés à la section 5.

2. Description du système

AVASR (Audio-Visual Automatic Speech recognition) comprend trois modules qui sont : le module de reconnaissance acoustique, le module de reconnaissance visuelle et le module de fusion [17, 19].

Le module de reconnaissance acoustique utilise l'approche stochastique basée sur les modèles des HMM qui sont un type particulier des réseaux bayésiens. On processus générique est basé sur trois phases qui sont : la paramétrisation du signal acoustique utilisant dans notre cas l'analyse log RASTA-PLP (*Relative Spectral Analysis-Perceptual Linear Predictive*), l'apprentissage des modèles repose sur une recherche génétique d'un bon modèle parmi une population hétérogène des HMM (contenant différentes architectures) et une optimisation par un algorithme de gradient (Baum-Welch) et leur décodage sur l'algorithme de viterbi. Le module de reconnaissance visuelle utilise la même approche stochastique, il diffère uniquement par la phase de paramétrisation basée elle sur la DCT (Discrete Cosine Transform).

La figure 1 présente les différentes étapes dans les processus d'apprentissage et de reconnaissance du système proposé. Chacun des éléments présents sur cette figure sera décrit dans ce papier.

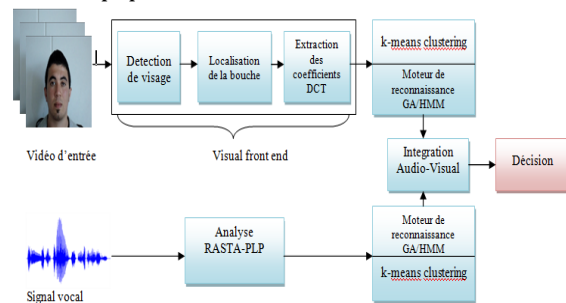


Fig. 1 : Système d'un AVASR mis en œuvre.

3. Traitement des données audiovisuelles

3.1. Traitement de la partie acoustique

En traitement de la parole, il est courant d'utiliser le domaine spectral plutôt que le domaine temporel directement. Dans le cadre de notre étude, ceci est renforcé par le fait que la bimodalité de la parole lie l'information visuelle du visage du locuteur au spectre d'amplitude des sons émis [10].

La phase de paramétrisation prend un signal en entrée et retourne un vecteur de paramètres (appelé vecteur acoustique ou encore vecteur d'observations). Les vecteurs de paramètres doivent être pertinents (précis, de taille restreinte et sans redondance), discriminants (pour faciliter la reconnaissance) et robustes (aux différents bruits et/ou locuteurs).

Le traitement des données acoustique se fait en utilisant la méthode RASTA-PLP [7] qui est une amélioration de la méthode de prédiction linéaire perceptuelle (PLP). Celle-ci consiste à ajouter un filtre passe-bande afin de réduire l'influence des bruits convolutifs dûs au canal de transmission, voir [8] pour plus de détails.

A la limite des méthodes de rehaussement de la parole, les représentations obtenues à l'aide des filtres RASTA visent à supprimer le bruit en utilisant les variations temporelles pour différencier le bruit du signal de parole. Le plus souvent, le signal clair est perturbé par un bruit ambiant de variation lente. Les filtres sont donc définis pour éliminer les signaux stationnaires. Ils consistent à calculer un banc de filtre représentant l'énergie du signal à différentes bandes de fréquence sur une échelle logarithmique, puis à filtrer chaque canal par un filtre passe haut avec une pente raide éliminant les très basses fréquences (1Hertz), éliminant ainsi les variations lentes dues aux modifications de l'environnement. [7] Différentes études réalisées avec cette méthode ont permis de confirmer ces bonnes qualités relativement aux distorsions et ses moindres qualités face aux bruits qualifiés d'additifs, signe de la présence de plusieurs sources sonores dans un même environnement.

Le processus de calcul des coefficients RASTA-PLP est décrit par la figure suivante :

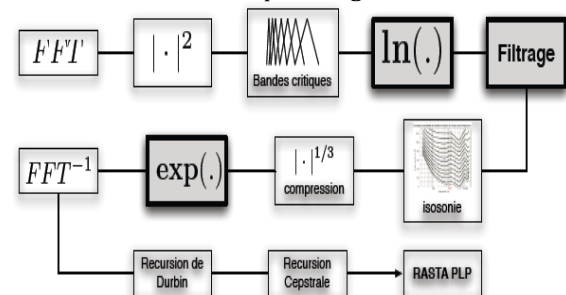


Fig. 2 : Analyse RASTA-PLP.

Les vecteurs acoustiques utilisés dans cette expérience sont les coefficients de 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 dérivée des paramètres, on obtient des matrices de 27 paramètres.

3.2. Traitement de la partie visuelle

Nous savons que les lèvres véhiculent la majeure partie de l'information visuelle utile pour la compréhension de la parole. Les êtres humains emploient l'information visuelle de façon subconsciente afin de comprendre les paroles, particulièrement dans des environnements bruyants, mais également quand les conditions acoustiques sont bonnes. Le mouvement des lèvres du locuteur apporte une série d'information importante, par exemple au sujet des articulations, ce qui sont automatiquement intégrés par le cerveau. L'effet McGurk [9] apporte la preuve en montrant que le cerveau, soumis à des stimuli auditifs et visuels inconsistants, perçoit un son différent de celui qui a été dit. Tandis que pour la reconnaissance audio de la parole, les types de dispositifs employés sont relativement bien établis, les coefficients dits mel-cepstraux constituant la base de la majorité des approches, la situation n'est pas identique pour les dispositifs visuels. On peut en distinguer deux types principaux. Le premier est basé sur l'image de la bouche en mouvement ou sur certaines transformations appliquées sur cette image, comme celles utilisées aussi en compression d'image, se fondant sur l'hypothèse qu'une image comprimée maintiendra également l'information nécessaire pour la reconnaissance visuelle de la parole. Le deuxième type de dispositifs visuels est basé sur la forme de la bouche, sa découpe ou certain attribut géométrique comme sa taille, sa largeur, etc.

3.2.1. Détection de visage et Localisation des lèvres

Comme commenté dans l'introduction, tous les systèmes de reconnaissance audio-visuelle de la parole nécessitent l'identification et le suivi de la région d'intérêt (ROI), qui peut être soit seulement la bouche, ou une région plus vaste, comme tout le visage. Cela commence généralement par localisation de visage du locuteur, en utilisant un algorithme de détection de visage. La deuxième étape est la localisation de la bouche du locuteur et l'extraction de la région d'intérêt. Cette région peut être mise à l'échelle et pivotée de telle sorte que la bouche est centrée et alignée.

La détection des visages pose le problème de la localisation des visages présents

dans une image d'entrée. Idéalement, la détection fournit aussi leurs dimensions pour un éventuel traitement ultérieur.

Son utilité se manifeste dans des domaines variés allant de la vidéosurveillance au jeu interactif. Les premières difficultés rencontrées par les méthodes s'attendant à détecter les visages sont les variations de pose (vue de profil, de face), d'expression, de rotation du visage, d'âge et d'illumination. Pour le reste la difficulté est d'autant plus grande que la plupart des applications ayant recours à cette technologie requièrent une exécution en temps réel, limitant les marges de manœuvre de l'algorithme.

3.2.2. Algorithme de Viola & Jones

Une avancée majeure dans le domaine a été réalisée par Viola et Jones en 2001 [11]. Ces derniers ont proposé une méthode basée sur l'apparence ("Appearance-based methods") robuste et tournant à 15 fps pour des images de 384 x 288 pixels sur un pc Intel Pentium III 700Mhz. Ce fut la première méthode en temps réel présentée. La renommée de cette approche est faite sur trois concepts :

A. L'image intégrale

L'algorithme se base sur les descripteurs simples (Haar features) pour localiser les visages présents sur une image d'entrée. Dans le but d'extraire rapidement ces caractéristiques, l'image est représentée sous forme intégrale. En effet, sous cette forme, l'extraction d'une caractéristique à n'importe quel endroit et à n'importe quelle échelle est effectuée en un temps constant tandis que le temps de conversion vers la représentation intégrale ne remet pas en cause ce gain de temps offert par l'utilisation de la représentation en image intégrale. La définition des caractéristiques de Haar et la manière dont la représentation intégrale accélère considérablement leur extraction sont présentées ci-après pour une image en niveaux de gris.

Dans toute image, une zone rectangulaire peut être délimitée et la somme des valeurs de ses pixels calculée. Une caractéristique de Haar est une simple combinaison linéaire de sommes ainsi obtenues.

Plusieurs caractéristiques de Haar peuvent être définies selon le nombre, les échelles, les positions et les dimensions des zones rectangulaires considérées. 4 exemples sont présentés à la figure 3.

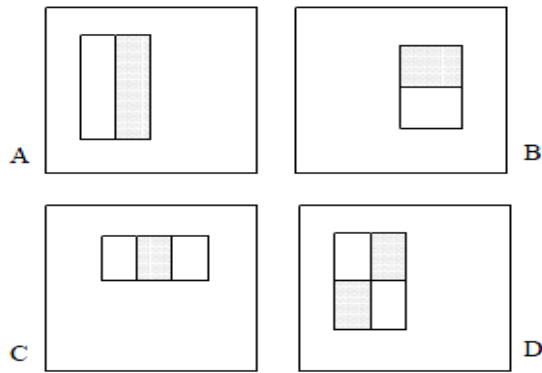


Fig. 3 : Exemple de 4 caractéristiques de Haar. La somme des valeurs des pixels appartenant aux zones encadrées claires est soustraite à la somme des valeurs des pixels appartenant aux zones encadrées sombres pour obtenir la caractéristique de Haar. Chacune des quatre caractéristiques de Haar est représentée avec son cadre de détection respectif.

Présenté comme tel, le calcul d'une caractéristique de Haar demande à chaque fois l'accès aux valeurs de tous les pixels contenus dans les zones rectangulaires considérées. Cela devient vite contraignant temporellement dès que les caractéristiques de Haar sont définies par des zones rectangulaires de grandes dimensions. L'image intégrale permet de surmonter ce problème en rendant constant le temps de calcul d'une caractéristique de Haar à n'importe quelle échelle.

Le calcul de la somme des valeurs des pixels appartenant à une zone rectangulaire s'effectue donc en accédant seulement à quatre pixel de l'image intégrale : Soit un rectangle ABCD dont les sommets sont nommés dans le sens des aiguilles d'une montre en commençant par le sommet supérieur gauche et soit x la valeur sous la représentation intégrale d'un sommet X du rectangle ($X \in \{A, B, C, D\}$). La somme des valeurs des pixels appartenant à ABCD est, quelle que soit sa taille, donnée par $c - b - d + a$. Une caractéristique de Haar étant une combinaison linéaire de tels rectangles ABCD, son calcul se fait alors en un temps indépendant sa taille.

B. Algorithme d'apprentissage basé sur Adaboost

Pour localiser les visages sur l'image d'entrée, cette dernière est scannée par une fenêtre de dimension déterminée. La fenêtre parcourt l'image et son contenu est analysé pour savoir s'il s'agit d'un visage ou non. Comme dit plus haut, les caractéristiques de Haar sont extraites pour effectuer la classification et de ce fait la représentation intégrale de l'image accélère l'analyse. Mais, pour une fenêtre de 24×24 pixels il y a 45 396 caractéristiques de Haar, les traiter toutes prendrait beaucoup trop

de temps pour une application en temps réel. Pour surmonter ce problème, une variante de la méthode de boosting Adaboost est utilisée.

C. Cascade

L'idée de base derrière le concept de Cascade est que parmi l'ensemble des candidats, c'est à dire l'ensemble des états de la fenêtre de recherche, une partie peut être éliminée sur base de l'évaluation de seulement quelques caractéristiques de Haar. Une fois cette élimination effectuée, les candidats restants sont analysés par des classifieurs forts plus complexes (utilisant plus de caractéristiques de Haar) demandant un plus grand temps de traitement. En utilisant plusieurs 'étages' de ce type, le processeur évite d'effectuer des analyses lourdes en temps de calcul sur des échantillons pour lesquels il est rapidement possible de se rendre compte qu'ils sont négatifs. Le processus de classification apparaît alors comme une cascade de classifieurs forts de plus en plus complexes ou à chaque étage les échantillons classifiés négatifs sont sortis tandis que les échantillons classifiés positifs sont envoyés aux classifieurs suivants.

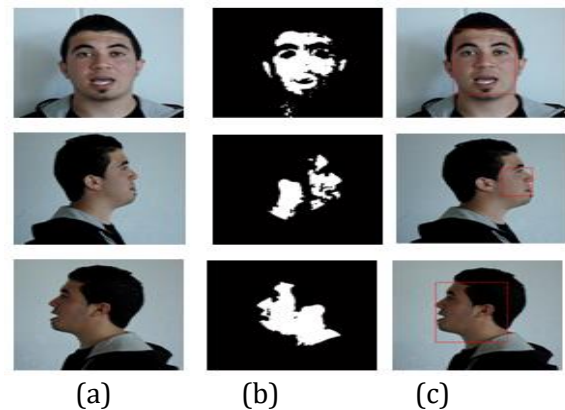


Fig. 4 : Un exemple de détection de visage : (a) image originale (b) détection de peau avec suppression de bruit (c) résultat de détection de visage.

3.2.3. Localisation de la bouche

Détecter le milieu de la bouche n'est pas aussi simple qu'on le pense. Il ya beaucoup de possibilités, allant sur gradient horizontal et / ou vertical décent, la teinte ou la saturation [12]. Au moment où il est exécuté en utilisant la teinte distincte des lèvres. La lumière se reflète sur les lèvres et ce point est récupéré par une valeur de teinte définie. Contrairement aux autres méthodes, cette méthode n'est pas indépendante de lumière, ainsi l'intensité et la direction de la lumière peut influencer les résultats.

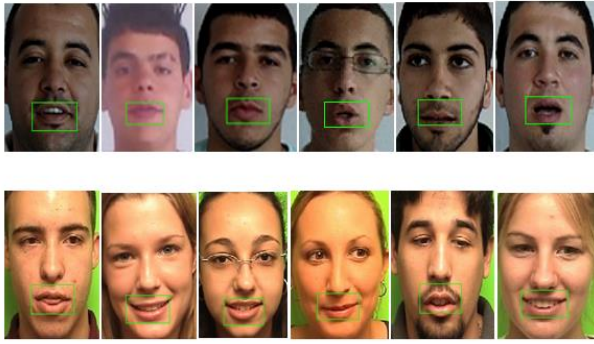


Fig.5 : Exemples des zones d'intérêt extraites des bases de donnée audiovisuelle.

3.2.4. Extraction des paramètres

L'extraction des caractéristiques vidéo est effectuée avec la Transformée en cosinus discrète (DCT) [16]. Les raisons de l'utilisation généralisée de la DCT aussi bien dans la compression d'image comme extraction des caractéristiques [1] sont la compaction élevée de l'énergie du signal d'entrée sur quelques coefficients de DCT et la disponibilité d'une mise en œuvre rapide de la transformée, similaire à la Transformée de Fourier (FFT). Par contre, contrairement à la FFT, elle comporte seulement les coefficients réels. Depuis la DCT n'est pas invariant passer la performance dépend d'un suivi précis du ROI. DCT transforme chaque composante de couleur en coefficients DCT en utilisant l'équation suivante [13]:

$$F(u, v) = \frac{1}{\sqrt{MN}} \alpha(u) \alpha(v) \sum_{x=1}^M \sum_{y=1}^N f(x, y) \cos \left[\frac{(2x-1)u\pi}{2M} \right] \cos \left[\frac{(2y-1)v\pi}{2N} \right] \quad (1)$$

avec,

- u est la fréquence spatiale horizontale,
- v est la fréquence spatiale verticale,
- $f(x, y)$ est la valeur de pixel aux coordonnées (x, y),
- $F(u, v)$ est le coefficient de DCT au point de coordonnées (u, v), elle est dimensionnée de $M \times N$, et $\alpha(\bullet)$ est définis comme suit:

$$\alpha(w) = \begin{cases} \frac{1}{\sqrt{2}}, & w=1 \\ 1, & \text{otherwise;} \end{cases} \quad (2)$$

L'implémentation de la transformée en cosinus discrète peut être basée sur la transformée de FOURIER. Pour une ligne (ou colonne), les N coefficients de la DCT peuvent calculés en utilisant une FFT sur $2N$ points. On montre en effet aisément que la DCT d'une séquence est équivalente à la transformée de FOURIER de cette même séquence préalablement reproduite par symétrie autour de l'origine.

Dans cet article nous avons appliqué la DCT sur chaque bloc 8×8 de luminosité et de chrominance, chacun des 64 coefficients DCT sont quantifiés. Après quantification, le coefficient le plus petit (coefficient DC) est traité différemment par rapport aux autres (coefficients AC). Le coefficient DC correspond à l'intensité moyenne du bloc et il est encodé avec une méthode de prédiction différentielle DC. Le reste des coefficients non nuls sont ensuite codés en zigzag comme il est montré dans la figure 6. Afin de limiter la taille mémoire et les calculs nécessaires pour l'apprentissage et la reconnaissance nous avons gardé que les 100 premiers coefficients pour représenter l'image [15].

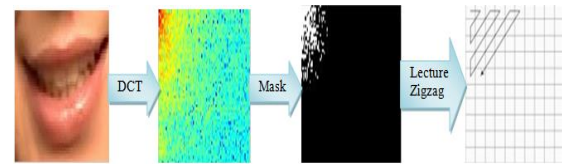


Fig.6 : Processus de sélection des coefficients DCT.

3.3. Reconnaissance

Après la phase de paramétrisation et l'obtention des vecteurs de paramètres acoustiques ainsi visuelles, un moteur de reconnaissance basé sur les réseaux bayésiens de type HMM utilisés pour la modélisation des phonèmes arabes qui sont de nature discrète, ainsi que leurs densités de probabilités d'observations, ce qui justifie la nécessité d'utiliser un quantifieur vectoriel pour faire correspondre chaque vecteur continu (représentant une trame) à un indice discret d'un dictionnaire de référence (CodeBook), ainsi on peut assimiler le rôle de la quantification vectorielle à la conception du CodeBook on utilisant pour cette étape la méthode de K-means [18].

Cette procédure consiste en premier lieu à partitionner les vecteurs caractéristiques issus de l'étape de paramétrisation en K ensembles disjoints (où K est la taille du CodeBook à concevoir). En deuxième lieu, à représenter chaque ensemble par un vecteur unique (point typique) qui est généralement le centroïde des vecteurs caractéristiques de l'ensemble d'apprentissage affectés à la même région, ensuite elle optimise itérativement la partition du CodeBook. Le centroïde est le point qui minimise la distance pour l'ensemble de tous les points d'une classe. Idéalement, tous les éléments d'une classe doivent être plus proches de leur centroïde que de tout autre centroïde.

Ainsi, un vecteur inconnu X, est classé comme élément de la classe c si la distance de X, au

centroïde de c est plus petite que la distance entre X, et tous les autres centroïdes.

La taille K du CodeBook est un paramètre crucial dont la valeur affecte en grande partie les performances des HMMs utilisés pour la reconnaissance, car on le considère que c'est le nombre des mixtures.

Nous avons utilisé trois HMM de type gauche-droite, chaque HMM est caractérisé par les éléments suivants :

- N : est le nombre d'états dans le HMM
- O = $\langle o_1 \dots o_T \rangle$ est la séquence d'observations
- M : est le nombre de mixtures
- A : Matrice de probabilités de transition
- B : Matrice de probabilités d'observation
- Π : est Le vecteur des densités initiales

Dans la phase d'apprentissage en utilisant comme il est mentionné avant les HMM mais cette fois combinés avec les GA [13], cet algorithme (inspiré de l'article de Jonathan Goh [3]) va chercher à obtenir des HMM optimales [6, 7, 20], Le processus de formation d'un modèle $\lambda = (A, B, \pi)$ Pour les données de référence en utilisant une méthode hybride GA/HMM.

Premièrement on commence par créer une population de taille S, aléatoirement, d'une façon que chaque individu contient un chromosome pour le nombre d'états, n chromosomes pour les probabilités initiales, $n \times n$ chromosomes contiennent les probabilités de transition et $n \times m$ chromosomes chacune contient la probabilité d'émission. Aucun individu n'est marqué « parent ». Le codage de chaque individu est comme suit :

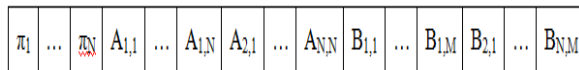


Fig.7 : Méthode de représentation des chromosomes dans l'apprentissage des GA-HMMs.

Après nous appliquons sur chaque HMM de la population non marqué « parent » l'algorithme de Baum-Welch à partir de l'observation O.

L'étape suivante est de calculer pour chaque individu de la population qui ne porte pas la marque « parent » la valeur de fitness (noté aussi la fonction objective) on utilisant l'algorithme de Baum-Welch, et noter la valeur de probabilité de l'observation. Pour tous ceux qui portaient cette marque, l'enlever. Mathématiquement la fitness de n model est exprimé comme suit [5]:

$$f(\lambda_i) = \frac{P_n}{\sum_{i=1}^N P_i} \quad (3)$$

$$P_n =$$

$$\frac{\sum_{i=1}^M \log(P(o_i|\lambda_i))}{M} \quad (4)$$

Où $P(o|\lambda)$ est la probabilité de vraisemblance.

La troisième étape est de sélectionner parmi tous les individus de la population, un certain nombre $S' < S$, qui seront utilisés comme parents pour régénérer les $S-S'$ autres individus non retenus. La sélection se réalise suivant les meilleurs scores calculés à la phase 3. Chaque individu sélectionné est marqué « parent ».

Les opérations génétiques peuvent inclure croisement et par mutation. L'opération génétique est réalisée pour améliorer la technique de ré-estimation de Baum-Welch pour que les populations génétiques de cette opération résultent un modèle optimal [4, 5]. chaque opérateur est expliqué comme suit :

Croisement :

Pour chaque individu non marqué « parent », sélectionner au hasard deux individus parmi la population de ceux marqués « parent » et les croiser. Le croisement est à un point de coupure, et se réalise entre deux lignes des matrices des HMM. Ceci permet d'obtenir en retour deux fils, HMM également. On ne conserve qu'un des deux fils, au hasard.

Mutation / normalisation :

Sur chaque individu non marqué « parent » on applique l'opérateur de mutation. Celui-ci consiste à modifier d'une faible quantité aléatoire chaque coefficient des matrices de la HMM. Chaque coefficient est modifié, en fonction de la valeur de la probabilité de mutation.

Après avoir traité un individu, on lui applique un opérateur de normalisation, afin que cet individu soit une HMM (matrices stochastiques).

À la fin on termine par l'évaluation de la condition d'arrêt, Si le nombre d'itérations maximum n'est pas atteint, alors retourner à la deuxième étape, sinon aller à la dernière étape qui va renvoyer la meilleure HMM parmi la population en cours

Les techniques stochastiques, tel que les HMMs, sont actuellement les plus utilisées pour la modélisation acoustique de la parole permettant de déterminer la probabilité d'une suite d'observations.

Un tel classifieur est basé sur un critère de maximum de vraisemblance, il prend le mot à reconnaître comme étant une séquence d'observations discrètes (codes) produites par analyse et quantification vectorielle de la séquence de vecteurs de caractéristiques. Ce classifieur calcule la probabilité qui correspond à la probabilité d'obtenir la séquence par le modèle. Ces probabilités sont évaluées par la

version logarithmique de l'algorithme de Viterbi. Finalement, le mot testé est affecté à la classe du mot K du lexique L pour laquelle le modèle maximise la probabilité d'émission.

3.4. La fusion audiovisuelle

Il a longtemps été un rêve humain de créer des machines capables de parler et / ou à comprendre la parole. Dans la période de la Grèce antique, prêtres utilisés statues parlantes que les oracles de tricher sur et mieux convaincre leur auditoire. Plus récemment, en parlant de plusieurs machines ont été conçues, d'abord en tant que dispositifs mécaniques, alors qu'appareil électronique, et aujourd'hui, la plupart du temps avec les ordinateurs. En effet, la nouvelle ère du multimédia renouvelle le défi de la multimodal de la communication. Bien que le clavier et la souris restent les modes les plus utilisés de la communication homme-machine, il est évident que la communication orale est plus naturel et nécessite moins pratique. La communication multimodale Homme-machine parlée est un grand défi d'aujourd'hui, et des retombées technologiques sont naturellement attendues d'une meilleure compréhension de la façon dont les humains produisent et comprendre la parole multimodale. En retour, des visages parlants synthétiques et les reconnaissances automatiques de la parole multimodales sont des outils qui permettent i) l'évaluation quantitative de la production théoriques de la parole humaine et des modèles de perception, et ii) la génération de stimuli de parole très contrôlé indisponibles pour les parleurs humains.

Aujourd'hui, nous voyons l'émergence d'intérêt dans AVASR, en raison de plusieurs facteurs. Tout d'abord, la performance de RAP acoustique est le plafonnement, alors que l'usage est plus répandu, en particulier dans des environnements bruyants, dans lequel l'information optique pourrait considérablement compenser la mauvaise qualité des signaux acoustiques. Parallèlement, le traitement d'image en temps réel est maintenant accessible sur de nombreux ordinateurs : petit poids-léger de micro-caméras sont de plus en plus commune, de sorte qu'il est maintenant possible de concevoir la lecture labiale automatique.

Pour expliquer les phénomènes de fusion observés dans le processus d'intégration audiovisuelle de la parole, de multiples modèles ont été proposés. On peut classer tous les modèles d'intégration audiovisuelle en trois principales catégories [9, 15] : *Modèle d'Identification Directe*, *Modèle d'Identification Séparée*, et le *Modèle hybride*.

Dans ce travail nous avons utilisé le modèle de fusion séparée, c.à.d. la fusion des scores issus de chaque reconnaisseur GA/HMM

(reconnaisseur acoustique ainsi visuel). Leurs ensembles des log-vraisemblance peuvent être combinés en utilisant les poids qui reflètent la fiabilité de chaque flux particulier [21].

Les probabilités d'observation visuelle et audio sont pondérés de façon exponentielle par flux des poids μ_A et μ_V , où A et V est le flux audio et vidéo, nous contraignons μ_A et μ_V pour s'étendre entre zéro et un et la somme à un [14], à savoir que:

$$\mu_A + \mu_V = 1 \quad \text{for} \quad 0 < \mu_A, \mu_V < 1 \quad (5)$$

4. Expérimentations

4.1. Bases de données utilisées

Dans notre travail nous avons utilisé deux bases de données :

- A. La base de donnée 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 [6].



Fig.8 : Exemples de trames de la base CUAVE.

- B. 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), les vidéos sont capturées à une distance moyenne égale à 14.5 cm avec une résolution de 690x340 pixel et à 30 trames/sec et avec des variations de pose (vue de profil, de face) de 11 locuteurs sauvegardées avec l'extension « .avi », alors que les fichiers audio sont sauvegardés avec l'extension « .wav ».

l'échantillonnage standard après des tests 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.

Ces locuteurs sont de différentes régions dialectes, et chaque locuteur prononce chaque mot 5 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.



Fig.9 : quelques exemples de trames de notre base audiovisuelle arabe.

4.1. Résultats et discussions

Nous avons pris pour le test des performances de notre système 2 bases de données audiovisuelles la première est la base CUAVE et la deuxième une base de données arabe. Pour réaliser notre système proposé, nous avons utilisé les $\frac{2}{3}$ pour la phase d'apprentissage et le $\frac{1}{3}$ restant pour tester l'efficacité de notre système.

Un système de reconnaissance de la parole audiovisuelle a été implémenté, et qui utilise la méthode RASTA-PLP pour l'extraction des paramètres acoustiques, et pour caractériser les signaux vidéo nous utilisons la DCT, et comme moteurs de reconnaissance les modèle de Markov cachés avec un algorithme hybride (les algorithmes génétiques) pour leur apprentissage, Cet algorithme optimise à la fois les paramètres (probabilités) de la HMM. Il repose sur une recherche génétique d'un bon modèle parmi une population hétérogène de HMM (contenant différentes architectures) et une optimisation par un algorithme de gradient (Baum-Welch) comme les résultats montrent dans les figures 10 et 11.

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

Après des diverses tests nous avons trouvé que et pour les paramètres de l'algorithme génétique sont comme suit :

• taille de la population de l'AG = 80
• utiliser un seul type de sélection qui est la sélection ordonnée (rank selection).
• probabilité de croisement = 0.5
• probabilité de mutation = 0.2
• nombre d'itérations de l'AG = 20
• nombre d'itérations de BW = 5

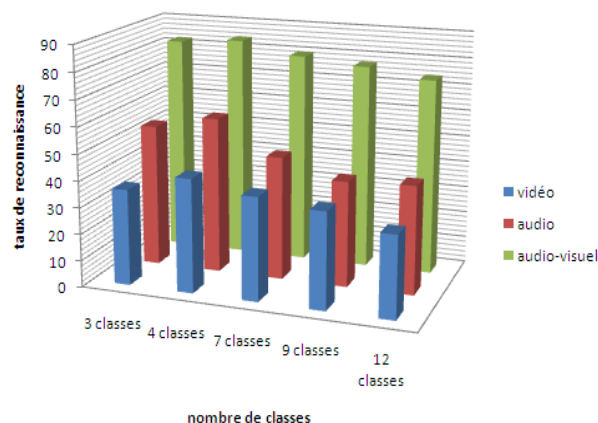


Fig.10 : Comparaison entre les taux de reconnaissances audio, vidéo, et audiovisuel, on utilisant la méthode HMM classique.

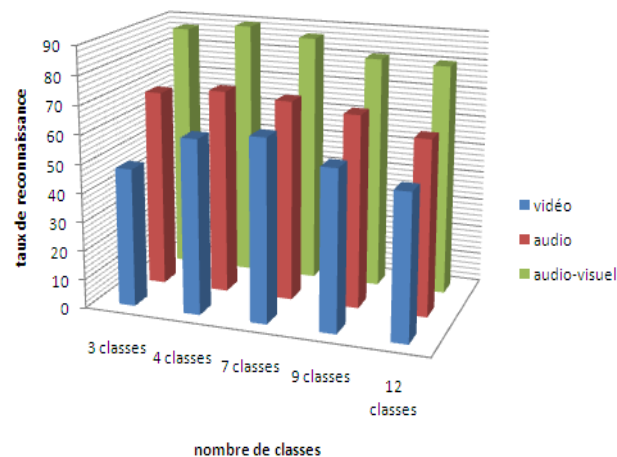


Fig. 11 : Comparaison entre les taux de reconnaissances audio, vidéo, et audiovisuel, on utilisant la méthode GA-HMM.

La phase de test est une étape très importante dans l'évaluation de notre système. À partir les résultats, il peut être facilement constaté que GA/HMM peut atteindre un taux de reconnaissance plus élevé par rapport aux HMM classique avec un nombre de clusters égale à 9, et avec une augmentation de pourcentage entre 5% à 38%.

5. Conclusion et perspectives

Dans cet article, nous avons présenté, un système AVASR qui utilise RASTA-PLP comme méthode d'extraction des caractéristiques acoustiques, les coefficients DCT pour les caractéristiques visuelles, et une méthode hybride GA/HMM pour la modélisation audio-visuelle de la parole.

Les tests tendent à démontrer les divers avantages de cet algorithme : résolution du problème de détermination de l'architecture d'une CMC, meilleurs résultats d'apprentissage par rapport aux algorithmes traditionnels et aux algorithmes hybrides,

Ces caractéristiques permettent en outre d'apprendre de plus grandes quantités de données. Nous projetons également d'ajouter le nombre des locuteurs dans notre corpus audiovisuel, Nous avons aussi l'intention de tester notre système avec d'autres méthodes alternatives de fusion telles que la fusion hybride. Finalement essayer d'utiliser un autre type de HMM (comme le HMM couplée) afin de l'optimiser par les algorithmes génétiques pour la fusion audiovisuelle.

Référence :

- [1] Nyoman Rizkha, E., Suyanto, Maharani, W. (2012) Isolated Word Recognition Using Ergodic Hidden Markov Models and Genetic Algorithm, TELKOMNIKA, 10(1): 129-136.
- [2] Brouard, T., Slimane, M., Venturini, G., Asselin de Beauville, J.P. (1997) Apprentissage du nombre d'états d'une chaîne de Markov cachée pour la reconnaissance d'images, SEIZIÈME COLLOQUE GRETSI, 15-19, GRENOBLE.
- [3] Goh, J., Tang, L. Al turk, L. (2010) Evolving the Structure of Hidden Markov Models for Micro aneurysms Detection, Computational Intelligence (UKCI), UK Workshop on:1-6.
- [4] Xueying, Z., Yiping, W., Zhefeng, Z. (2007) A Hybrid Speech Recognition Training Method for HMM Based on Genetic Algorithm and Baum Welch Algorithm. *IEEE Transactions on Neural Network*.
- [5] Mourad O, Raja NA. (2010) HMM Parameters Estimation Using Hybrid. IEEE.
- [6] Patterson, E.K., Gurbuz, S., Tufekci, Z. Gowdy, J.N. (2002) Cuave: A new audio-visual database for multimodal human-computer interface research, [Online]. Available: file:///home/helge/medialab/project/papers/2002-cuave.pdf
- [7] Hermansky, H., Morgan, N., Bayya, A, Kohn, P. RASTA-PLP Speech Analysis, ICSI Technical Report TR-91-069, Berkeley, California.
- [8] Brian K. and Morgan, N. (1997) Recognizing reverberant speech with RASTA-PLP. In ICASSP, Munich, Germany, 2: 1259-1262.
- [9] McGurk H., MacDonald J. (1976) Hearing lips and seeing voices. *Nature* 264(5588): 746-748.
- [10] Potamianos G., Neti C., Luttin J., Matthews I., (2004) 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.
- [11] Viola, P.A. and Jones, Michael J. (2001) Rapid Object Detection using a Boosted Cascade of Simple Features. CVPR (1): 511-518
- [12] Pai, Y., Ruan, S., Shie, M., Liu, Y. (2006) A Simple and Accurate Color Face Detection Algorithm in Complex Background.; In ICME: 1545-1548.
- [13] Benmachiche, A., Bouhadada, T., Laskri, M-T., Zendi, A. (2016) A Dynamic Navigation for autonomous mobile robots, iospress: Intelligent Decision Technologies; ISSN: 1875-8843 (E), 10(1).
- [14] Reikeras, H., Herbst, BM. Johan du Preez, Engelbrecht, H. (2010) Audio-Visual Automatic Speech Recognition using Dynamic Bayesian Networks. Proceedings of the Twenty-First Annual Symposium of the Pattern Recognition Association of South Africa: 233-238.
- [15] Bakir, N., Debyeche, M., Chibani, Y. (2010) Reconnaissance automatique des chiffres arabes en milieu réel par fusion audiovisuelle, 10ème Congrès Français d'Acoustique, Lyon.
- [16] Ahmed, N., Datta, S., Mulvaney, D., Farooq, O. (2008) A Comparison of Visual Features for Audio-Visual Automatic Speech Recognition, Acoustics'08.
- [17] Rogozan A. (1999) Étude de la fusion des données hétérogènes pour la reconnaissance automatique de la parole audiovisuelle, Thèse présentée et soutenue publiquement le 9 juillet 1999 pour l'obtention du Doctorat de l'Université d'Orsay - Paris XI.
- [18] Lazli, L. (2007) Rapport interne I²S , Univ Paris 12.
- [19] Makhlouf, A., Lazli, L., Bensaker, B. (2016) Structure Evolution of Hidden Markov Models for Audiovisual Arabic Speech Recognition, International Journal of Signal and Imaging Systems Engineering, IJSISE, 9(1):55-66.
- [20] Maatallah, M., Seridi-Bouchelaghem, H. (2015) A Fuzzy hybrid approach to enhance diversity in top-N recommendations, International journal of business information systems, 19(4):505-530.
- [21] RODOMAGOU LAKIS, I. (2008) Feature Extraction Optimization and Stream Weight Estimation in Audio-Visual Speech Recognition, a Phd thesis from Technical University of Crete.