10ème **C**ongrès **F**rançais d'**A**coustique
Lyon, 12-16 Avril 2010

Reconnaissance Automatique des chiffres arabes en milieu réel par fusion audiovisuelle

Nadia BAKIR[1], Mohammed DEBYECHE[2] et Youcef CHIBANI[2]

Université des sciences et de la technologie Houari Boumediene (USTHB)
Laboratoire de communication parlée et de traitement du signal
Faculté d'électronique et d'informatique
bak_nad03@yahoo.fr

Résumé Dans cet article, nous présentons un système de Reconnaissance Automatique de la Parole (RAP) combinant les données acoustiques et les données visuelles. Ce système de reconnaissance audiovisuelle utilise comme moteur de reconnaissance les modèles de Markov cachés continus (Continuous Hidden Markov Model, CHMM) et comme méthode de fusion l'Identification Séparée (IS) basée sur les réseaux de neurones de type Perceptron Multi Couches (PMC). L'information visuelle utilisée conjointement avec les données acoustiques est basée sur la forme et les mouvements des lèvres lors de l'élocution. Les expériences réalisées pour la reconnaissance des chiffres arabes indiquent que l'utilisation conjointe de la modalité acoustique et de la modalité visuelle augmente la performance du système de RAP en milieu réel (fortement bruité), une augmentation du Taux ce Reconnaissance (TR) de l'ordre de 15% a été constatée.

1 Introduction

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 phrase 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 bruités 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 et à travers ce travail nous nous intéressons à la mise en œuvre d'un système de reconnaissance intégrant conjointement les deux informations acoustique et visuelle. 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 [1, 2]. Dans ce premier travail, nous avons opté pour la fusion à identification séparée étant donné que notre intérêt s'est porté sur la reconnaissance de mots isolés donc le problème de synchronisation entre les deux modalités acoustique et visuelle ne se pose pas. Ce papier est organisé comme suit : après cette brève introduction, la section 2 donne un aperçu sur le système de reconnaissance proposé avec une présentation détaillée de ces différents modules. La section 3 présente les expériences réalisées, les résultats comparatifs obtenus et les commentaires y afférent. La conclusion générale et les perspectives sont données dans la section 4.

2 Système de Reconnaissance par Fusion Audio-Visuelle

La structure du Système de Reconnaissance par Fusion Audio-Visuelle (SRFAV) mis en œuvre est donnée par la figure 1 ci-dessous. Ce système comprend trois modules qui sont : le module de reconnaissance acoustique, le module de reconnaissance visuelle et le module de fusion.

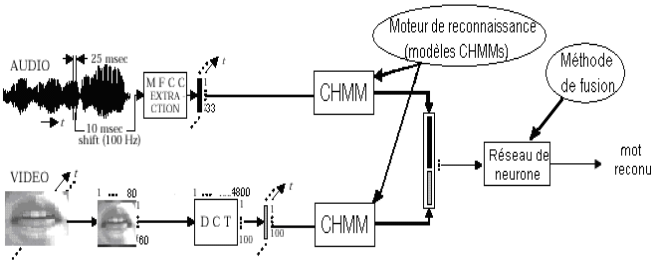

Figure 1 : Structure du SRFAV mis en œuvre.

Le module de reconnaissance acoustique utilise l'approche stochastique basée sur les modèles de Markov cachés continus [3, 8]. Son processus générique est basé sur trois phases qui sont : la paramétrisation du signal acoustique utilisant dans notre cas les coefficients cepstraux MFCCs (Mel Frequencies Cepstral Coefficients), l'apprentissage des modèles et la phase de décodage basée sur l'algorithme de Viterbi [3, 8]. Le module de reconnaissance visuelle utilise la même approche stochastique, il diffère uniquement par sa phase de paramétrisation basée elle sur la DCT [4]. Le module de fusion repose sur la méthode de fusion à identification séparée utilisant les réseaux de neurones artificiels de type PMC (Perceptron Multi-Couches).

2.1 Traitement des données audiovisuelles

2.1.1 Séparation audiovisuelle

Une fois l'enregistrement des séquences vidéo du locuteur est réalisé à l'aide du logiciel *Windows Movie Maker* de l'extension *''.wmv''*, la première opération consiste à convertir les séquences vidéo de l'extension *''.wmv''* vers l'extension *''.avi''* ce que nous avons fait en utilisant le logiciel *BPS (Vidéo Converter & Decompiler)*. Puis, on passe 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 *''.wave''*, 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.

2.1.2 Données acoustiques

Le traitement des donnés acoustiques est basé sur le calcul de l'enveloppe spectrale à travers les coefficients cepstraux dans l'échelle Mel à savoir les coefficients *MFCCs* [5]. Le signal acoustique est ainsi d'abord filtré à travers un filtre de fonction de transfert $H(z)=1-0.9\times z^{-1}$, il est ensuite fragmenté en trames et pondéré par une fenêtre de Hamming de 25ms avec un déplacement de 10ms. Aux coefficients MFCC est ajoutée l'information dynamique portée par la vitesse (*ΔMFCC*) et l'accélération (*ΔΔMFCC*). Chaque trame est ainsi représentée par un vecteur x_t de la forme : $xt=\{MFCC(m),\Delta MFCC(m),\Delta\Delta MFCC(m)\}$

Les formules de calcul des coefficients *ΔMFCC* et *ΔΔMFCC* inspirées des travaux réalisés par [6] sont données respectivement par l'équation 1 et l'équation 2 suivantes :

$$\Delta MFCC_l(m)=\left[\sum_{k=-K}^{K}k(MFCC_{l-k}(m))\right] \quad (1)$$

$$\Delta\Delta MFCC_t(m)=\left[\Delta MFCC_{t+1}(m)-\Delta MFCC_{t-1}(m)\right] \quad (2)$$

Où l est l'index de la trame courante et k le nombre de trames de part et d'autre de la trame courante (dans notre cas k est pris égal à 2).

2.1.3 Données visuelles

Pour caractériser les signaux vidéo nous utilisons la DCT (Discrete Cosine Transform), ou transformée en cosinus, qui permet de représenter une image de dimension $m\times n$ en une image de dimension égale, dont les coefficients sont classés dans l'ordre croissant des basses fréquences. Cette technique est très utilisée dans les codeurs d'image ou vidéo de type JPEG et MPEG [4]. 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 (DCT) à deux dimensions et la transformée inverse (IDCT) sont définie par les relations 3 et 4 suivantes :

$$H(u,v)=\frac{2}{\sqrt{MN}}C(u)C(v)\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}h(x,y)\cos\left[\frac{(2x+1)u\pi}{2M}\right]\cos\left[\frac{(2y+1)v\pi}{2N}\right] \quad (3)$$

$$h(x,y)=\frac{2}{\sqrt{MN}}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1}C(u)C(v)H(u,v)\cos\left[\frac{(2x+1)u\pi}{2N}\right]\cos\left[\frac{(2y+1)v\pi}{2N}\right] \quad (4)$$

Avec :

$$C(u)=\begin{cases}\frac{1}{\sqrt{2}} & pour\ :u=0\\ 1 & pour\ :u>0\end{cases} \quad (5)$$

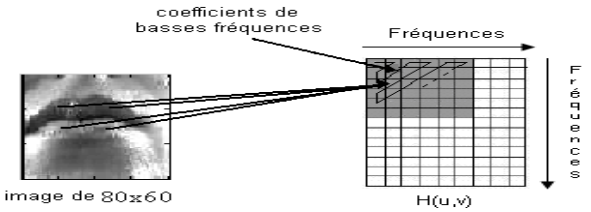

Figure 2 : Représentation graphique de la transformée en cosinus *(DCT)*

Les vecteurs d'entrées sont formés des coefficients basses fréquences qui se trouvent dans le coin supérieur gauche de la matrice résultante comme montré par la *figure 2*. Dans cette figure, nous avons conservé uniquement les 100 premiers coefficients de hautes amplitudes d'une image de dimension 80x60, 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. L'énergie totale E de l'image est calculée (théorème de Parseval, à partir des coefficients de la DCT, par la relation 6 suivante :

$$E=\sum_{u=0}^{M-1}\sum_{v=0}^{N-1}\left|H(u,v)\right|^2 \quad (6)$$

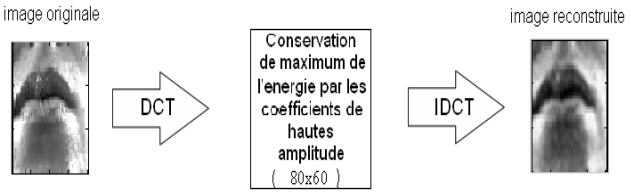

Figure 3 : Reconstitution d'une image à partir de 100 coefficients de hautes amplitudes de dimension 80x60

L'idée principale de l'algorithme pour encoder l'image par la DCT est de ne pas utiliser la totalité des coefficients (4800 coefficients), afin de limiter la taille mémoire et les calculs nécessaires pour l'entraînement et la reconnaissance par les modèles de Markov cachés continus *CHMM*. Dans notre travail nous avons gardé les cent (100) premiers coefficients pour représenter l'image.

2.2 Méthode de reconnaissance : Modèles de Markov Cachés Continus

Le moteur de reconnaissance basé sur les modèles de Markov cachés continus (*CHMM*) est utilisé aussi bien pour reconnaitre la modalité acoustique que la modalité visuelle. Les modèles continus utilisent des fonctions continues de densité de probabilité pour évaluer les probabilités d'observation directement dans l'espace des primitives. Chaque état modélise ses observations indépendamment des autres états du modèle, par une somme pondérée de fonctions élémentaires. La fonction densités de probabilités, associées aux états, est une somme pondérée de gaussiennes

multidimensionnelles (GMM : Gaussian Mixture Model pour l'anglais), elle est donnée par l'équation 7 suivante :

$$b_i(x_t) = \sum_{m=1}^{M} c_{im} \; N(x_t ; \mu_{im} , \Sigma_{im}) \qquad (7)$$

Où μ_{im} et Σ_{im} sont respectivement le vecteur moyen et la matrice de covariance de la $m^{ième}$ gaussienne de l'état i, et c_{im} le coefficient de pondération qui lui est affecté. La topologie du modèle de Markov utilisé est un modèle de type gauche-droite à trois états émetteurs (*figure 4)*.

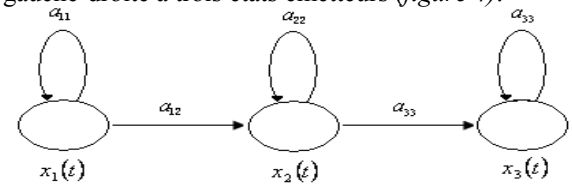

Figure 4 : Modèle gauche-droite à trois états émetteurs.

2.3 Fusion audiovisuelle

L'intégration des informations auditives et visuelles peut se faire de différentes manières [1, 2, 7, 9]. Les modèles de fusion sont classés en trois principales catégories : le modèle de fusion directe, le modèle de fusion séparée et le modèle de fusion hybride [6]. Dans ce travail nous avons utilisé le modèle de fusion séparé ou plus précisément la fusion des scores issus de chaque reconnaisseur (reconnaisseur acoustique et reconnaisseur visuel). Cette fusion est réalisée grâce à un réseau de neurones de type PMC (Perceptron Multi Couches) dont l'architecture est donnée par la *figure 5* ci-dessous. C'est un réseau à trois couches, composé d'une couche d'entrée contenant 20 cellules, une couche cachée de 100 cellules et une couche de sortie à 10 cellules.

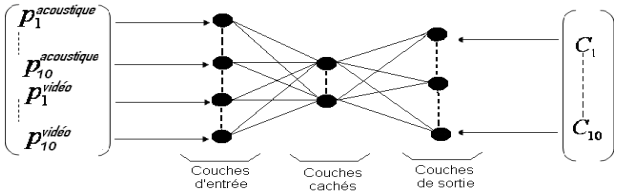

Figure 5 : Architecture du réseau de fusion.

Avec : Pi sont les scores acoustiques plus visuels.

Et : Ci sont les classes ou modèle.

3 Expériences et résultats

3.1 Base de données

Dans ce premier travail, une base de données en mode mono locuteur a été construite. Cette base de données a été enregistrée dans un milieu réel (salle de cours très bruyante). Elle comporte des prononciations des chiffres arabes isolés prélevés à une fréquence d'échantillonnage de 16 KHz. Elle est constituée de 25 répétitions de chaque chiffre arabe de zéro (0) à neuf (9) (*siffer (0), wahed (1), ithnani (2), thalatha (3), arbaa (4), khamssa (5), sitta (6), sabaa (7), thamania (8), tissaa (9)*).

3.2 Influence du nombre de mixtures sur le TR

L'objectif ici est de trouver la configuration optimale pour les modules de reconnaissance acoustique et visuelle à savoir donc le nombre de mixture donnant le meilleur Taux Reconnaissance (TR).

L'apprentissage des modèles acoustiques et visuels se fait par estimation de leurs paramètres sur un corpus dit *''apprentissage''* disjoint du corpus de *''test''*. Nous avons utilisé 40% (10 répétitions) de notre base de données pour l'apprentissage et 60% (15 répétitions) restant pour le test.

Les résultats de la *figure 6* montrent l'influence du nombre de mixtures de gaussiennes M sur les TR audio et vidéo. On remarque que les meilleurs taux de reconnaissance correspondent à un nombre de mixture égale à cinq ($M = 5$) aussi bien pour la modalité acoustique que pour la modalité visuelle.

Ces résultats indiquent aussi que le TR acoustique est en général meilleur que le TR vidéo.

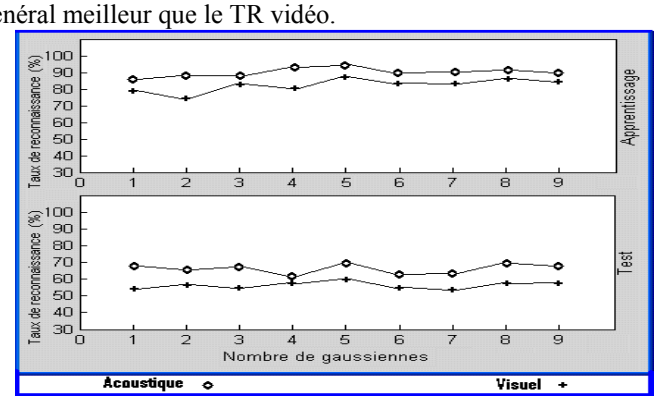

Figure 6 : Influence du nombre de mixtures sur le TR acoustique et le TR visuel.

3.3 Résultats audiovisuels

Les résultats obtenus sont résumés dans deux tableaux. Le tableau 1 indique pour chaque chiffre le TR acoustique, visuel et audiovisuel pour la base de test, le tableau 2 regroupe les performances exprimées pour les systèmes acoustique, visuel et audiovisuel en termes de TR Global (TRG) pour la base d'apprentissage et pour la base de test.

Chiffres	Résultats Audio (%)	Résultats Vidéo (%)	Résultats Audiovisuels (%)
siffer	86,66	73,33	90
wahed	66,66	60,00	85
ithnani	53,33	46,66	87
Thalatha	73,33	60,00	89
Arbaa	53,33	46,66	83
Khamsa	60,00	53,33	89
Sitta	66,66	60,00	88
Sabaa	80,00	53,33	85
Thamania	80,00	66,66	89
Tissaa	73,33	73,33	85
TRG	**69,33**	**59,33**	**87**

Tableau 1 : TR en % audio, visuel et audiovisuel.

Résultats	*TRG Audio (%)*	*TRG Vidéo (%)*	*TRG Audiovisuel (%)*
B.A	94	87	99
B.T	69,33	59,33	87

Tableaux 2 : TRG comparatifs audio, visuel et audiovisuel.

Avec : B.A. : base d'apprentissage.
B.T. : base de test.

4 Conclusion

L'objectif principal de ce premier travail est la mise en œuvre d'un système de reconnaissance de la parole audiovisuelle. Ainsi, nous avons abordé la fusion d'informations acoustiques et visuelles pour la RAP. Un modèle d'intégration audiovisuelle à identification séparée a été développé. Le système mis en œuvre a été testé sur un corpus audiovisuel, constitué de séquences des chiffres arabes mono-locuteur, dans un milieu réel fortement bruité.

Les tests réalisés ont montré que l'intégration de la modalité visuelle fondée sur le modèle de fusion séparée améliore les performances du système en environnement bruité (TR = 69.33% pour le système audio seul), (TR = 59.33%, pour le système visuel), et (TR = 87%, pour le système audiovisuel).

Nos efforts restent concentrés sur l'utilisation d'une base de données de plus grande taille prononcée par plusieurs locuteurs afin de tester la fiabilité réelle de notre système étant donné que le module de reconnaissance (CHMMs) et la méthode de fusion par réseaux de neurones sont des techniques qui nécessitent une large base de données afin de réaliser un bon apprentissage garant d'une meilleure performance globale.

Nous projetons également de tester notre système dans le cas des signaux de parole contaminés par des bruits spécifiques à différents niveaux RSB (rapport Signal/Bruit) en utilisant des méthodes de fusion alternatives telle que la fusion directe et la fusion hybride.

Références

[1] Adjoudani Ali, "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).

[2] Rogozan A., "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)

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

[4] Bovik A., "Handbouk of Image and Video Processing", Academic Press, p891 (2000).

[5] Hermansky H., "Perceptual linear predictive analysis of speech", Journal of the Acoustical Society of America, Vol. 87, N° 4, pp. 1738-1752 (1990).

[6] Wilpon J.C., Lee C.H., Rabiner L.R., "Connected digit recognition based on improved acoustic resolution", Computer Speech and Language, 7, pp. 15-26 (1993).

[7] Potamianos G., Neti C., Luttin J., Matthews I., "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).

[8] Boite R., Bourlard H., Dutoit T., Hancq J., Leich H., "Traitement de la parole", Collection Electricité, presses polytechniques et universitaires remandes, EPLF-Centre Midi, CH-1015 Lausanne (1999).

[9] Potamianos G., Neti C., Deligne S., "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. Jorion France, (2003).

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