Fusion de données acoustiques et articulatoires en reconnaissance automatique de la parole

Bruno Jacob, Régine André-Obrecht, Nathalie Parlangeau, Christine Sénac

Institut de Recherche en Informatique de Toulouse CNRS URA 1399

Université Paul Sabatier

118, Route de narbonne, 31062 Toulouse Cédex

RESUME

Dans le cadre du projet Applications Multimodales pour Interfaces et Bornes Evoluées (projet AMIBE soutenu par les PRC Informatique 1993 - 1995), est étudiée la bi-modalité naturelle auditive et visuelle de la communication orale. La reconnaissance automatique de la parole s'opère en synchronisant une "lecture labiale" avec un module de reconnaissance des formes acoustiques, par Modèles de Markov Cachés (MMC).

Pour fusionner les données acoustiques et articulatoires, nous proposons deux alternatives :

- un modèle de Markov caché classique où les données acoustiques et articulatoires sont supposées indépendantes,
- une liaison maître esclave entre deux modèles de Markov cachés, le modèle articulatoire pilote le modèle acoustique.

Des expériences de reconnaissance automatique sont réalisées sur un corpus de chiffres et lettres épelées connectés, et conduisent à une évaluation comparative des deux approches.

ABSTRACT

In the project AMIBE (Applications Multimodales pour Interfaces et Bornes Evoluées), we study the natural visual and auditive bi modality of the speech communication. The automatic speech recognition is performed by synchronizing the labial lecture and the acoustic pattern recognition based on Hidden Markov Models (HMM).

To merge acoustic and labial observations, we propose two alternatives :

- a classical HMM where the acoustic observations and the labial ones are assumed independent,
- a master-slave relation between two HMM, the articulatory HMM enslaves the labial one.

Automatic recognition experiments are performed on connected digit and spelled letter databases. We compare the two approaches and we show the labial lecture interest.

1 . Introduction

Dans le cadre du projet Applications Multimodales pour Interfaces et Bornes Evoluées (projet AMIBE soutenu par les PRC Informatique 1993 - 1995), est étudiée la bi-modalité naturelle auditive et visuelle de la communication orale. La reconnaissance automatique de la parole s'opère en synchronisant une "lecture labiale" avec un module de reconnaissance des formes acoustiques, par Modèles de Markov Cachés (MMC).

Pour fusionner les données acoustiques et articulatoires, plusieurs alternatives se présentent. Les informations peuvent être traitées sans discernement, par un MMC classique ; le vecteur d'observations est la concaténation des deux familles de paramètres labiaux et acoustiques ; ils sont considérés comme indépendants et l'utilisation de pondérations permet de réduire l'importance de l'une par rapport l'autre (nous parlerons, dans la suite de cet article d'approche globale et de MMC global). Une deuxième alternative consiste à modéliser chaque famille de paramètres par un modèle de type MMC, et corréler les deux modèles par une dépendance entre les lois. Nous avons étudié plus particulièrement cette approche appelée par la suite approche maître-esclave ; nous nous sommes inspirés des travaux de Brugnara et De Mori qui ont appliqué cette liaison maître-esclave pour traiter la durée des sons [Brugnara, 92].

Au cours de cette présentation, le principe de l'approche maître-esclave est rappelé brièvement, et nous définissons un MMC équivalent moyennant certaines hypothèses simplificatrices appropriées à la reconnaissance de paramètres acoustiques et labiaux. Des résultats expérimentaux illustrent cette approche, dans les cas de la reconnaissance de suites de chiffres et lettres épelées. Une comparaison entre approche globale et approche maître-esclave est proposée.

2 . Principe du modèle théorique :

Le principe des modèles dits "maître-esclave" repose sur la modélisation d'une application non plus par un MMC unique, mais par deux MMC mis en parallèle et corrélés. L'idée générale est de parvenir à une adaptation dynamique des lois de probabilités d'un des modèles de Markov cachés, en fonction du contexte courant modélisé par l'autre MMC. Le contexte est une notion qui doit être prise au sens large, il peut s'agir d'un indice de voisement, de nasalisation,... d'un réel contexte phonétique, d'un indice suprasegmental comme la durée des sons... tandis que le MMC piloté est traditionnellement lié à des paramètres acoustiques.

Un modèle maître esclave se compose de deux modèles, un modèle maître λ' et un modèle esclave λ'', définis comme suit :

- Le modèle λ' est un MMC classique, composé de deux processus stochastiques X' à valeurs dans $\chi' = \{ x' \}$, Y' à valeurs dans un ensemble mesurable $\mathcal{Y}'$, dont les lois respectives sont caractérisées par :
  — la matrice de transition $(a'_{z'x'})$,
  — les lois d'observation $(b'_{z'x'}(y'))$.

- Le modèle λ'' est un MMC dont les paramètres

dépendent à tout instant de l'état dans lequel se trouve le modèle maître ; le processus caché X" est à valeurs dans $\chi'' = \{x''\}$ ensemble d'états fini et le processus Y" est à valeurs dans un ensemble mesurable $\mathcal{Y}''$; leurs lois sont définies de la manière suivante :

$$P(X''_t = x''_t / X'^{T}_0 = x'^{T}_0 \ , \ X''^{t-1}_0 = x''^{t-1}_0 \ , \ Y'^{T}_1 = y'^{T}_1)$$
$$= P(X''_t = x''_t / X'_t = x'_t, X''_{t-1} = x''_{t-1})$$
$$= a''_{z''x''x'}$$

avec $x''_{t-1} = z''$ et $x'_t = x'$.

•

$$P(Y'' = y'' / X'^{T}_0 = x'^{T}_0, X''^{T}_0 = x''^{T}_0, Y'^{T}_1 = y'^{T}_1, Y''^{t-1}_1 = y''^{t-1}_1)$$
$$= P(Y''_t = y''_t / X'_t = x'_t, X''^{t}_{t-1} = x''^{t}_{t-1})$$
$$= b''_{z''x''x'}(y'').$$

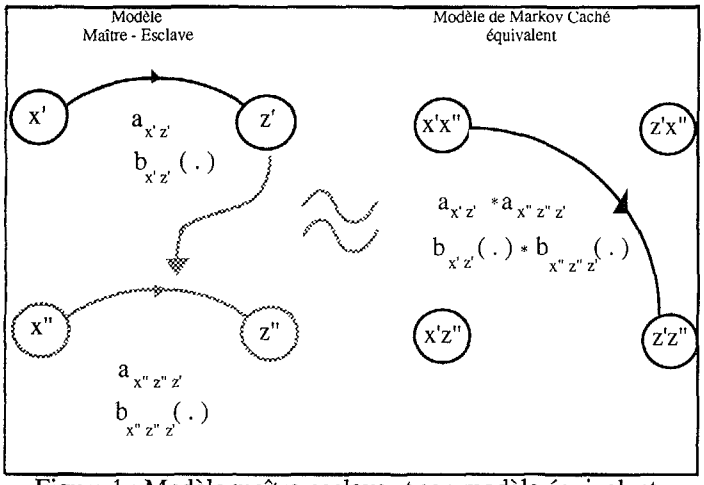


Figure 1 : Modèle maître-esclave et son modèle équivalent

3 . Modèle équivalent:

Un modèle maître-esclave est équivalent mathématiquement à un modèle classique de type MMC. En reprenant les notations du paragraphe précédent, le processus caché X est alors un double processus (X', X") à valeurs dans le produit cartésien $\chi' \times \chi''$ de cardinal $N' \times N''$ et le processus observable $Y = (Y', Y'')$ est à valeurs dans l'ensemble mesurable $\mathcal{Y}' \times \mathcal{Y}''$.

Les N' N" probabilités initiales et les $(N' \ N'')^2$ probabilités de transitions et lois d'observations du modèle de Markov caché $\lambda = (X, Y)$ sont caractérisées par les contraintes :

$$a_{xz} = a'_{x'z'} a''_{x''z''z'}$$
$$b_{xz(y)} = b'_{x'z'(y')} b''_{x''z''z'(y'')}$$
$$\Pi_x = \Pi'_{x'} \Pi''_{x''x'}$$

L'inconvénient de ce modèle réside dans son important nombre d'états et de lois (figure1). Ne pouvant raisonnablement implanté un tel modèle, nous avons émis des hypothèses simplicatrices.

Hypothèses simplificatrices

Dans le cadre de notre application, le modèle maître décrit le contexte labial. Nous supposons que toute configuration est équiprobable indépendamment de la précédente, ce qui s'exprime par :

— tout état est équiprobable et en particulier $a'_{x'z'} = 1 / N$, si N est le nombre d'états.

— les lois d'observations portées par les transitions (x'z') ne dépendent que de l'état d'arrivée z' $b'_{x'z'}(y') = b'_{t'z'}(y')$ pour tout t',x',y'.

Il s'en suit que l'on peut remplacer le modèle équivalent par un modèle simplifié dont le nombre d'états est le nombre d'états du modèle esclave, mais chaque transition du modèle esclave est dupliquée par le nombre d'états du modèle maître et chaque nouvelle transition est indexée par un état maître.

4 . Expérimentations :

Dans le cadre du projet AMIBE, nous disposons de deux types de signaux : le signal acoustique et le signal articulatoire synchronisé. Le signal acoustique est échantillonné à 16 kHz, tandis que pour le signal articulatoire (issu d'un traitement d'image [Lallouache 91]), nous disposons d'un vecteur d'observations toutes les 20ms. Ce signal se compose de la largeur A et la hauteur B internes du contour des lèvres, et la surface intérolabiale S (figure 2).

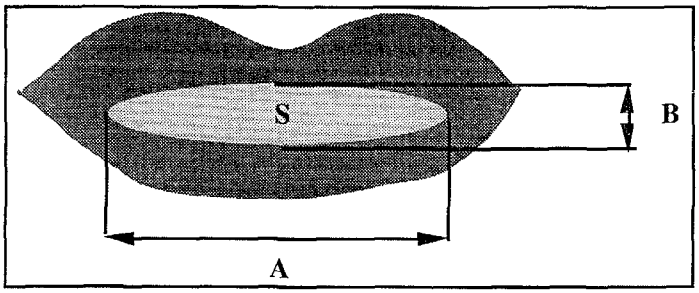


Figure 2 : Coefficients labiaux.

Pré traitement des données :

Le signal acoustique est segmenté automatiquement [André-Obrecht, 88] et une analyse spectrale est faite sur chaque segment : 8 coefficients cepstraux (MFCC) sont obtenus après recalage du spectre selon l'échelle Mel. Leur sont adjoints l'énergie (E) et la dérivée de ces coefficients (Δ MFCC, Δ E). Les frontières issues de la segmentation statistique sont projetées sur les signaux articulatoires. Pour chaque segment projeté, est calculée une valeur moyenne de chaque paramètre labial ainsi que les dérivées correspondantes. Le vecteur d'observations est finalement composé de 18 coefficients de nature acoustique, de 6 coefficients articulatoires, auxquels est ajoutée la longueur du segment correspondant (T).

Système de reconnaisance :

Pour fusionner les données acoustiques et articulatoires, nous avons envisagé un modèle équivalent simplifié correspondant au modèle maître esclave suivant :

— lemodèle Maître est un modèle ergodique à 3 états modélisant les configurations des lèvres : ouvertes, fermées et semi-ouvertes.

— le modèle Esclave est un modèle gauche-droit. dont les unités acoustiques élémentaires sont des pseudo-diphones [André-Obrecht 93].

• Données :

Cette application de reconnaissance est monolocuteur et le système est évalué sur deux corpus de phrases :

— Corpus des chiffres : chacune des phrases est

composée de 4 chiffres connectés ou des mots "oui" ou "non". L'ensemble d'apprentissage est formé de 84 prononciations de phrases (288 mots de base). L'ensemble de test est formé de 35 phrases (soit 125 mots).

— Corpus des lettres : chacune des phrases est composée de 4 lettres éppelées. L'apprentissage contient 158 phrases (soit 632 mots) et le test se compose de 48 phrases (soit 192 mots).

• Résultats :

Afin de valider ce type d'approche, nous avons comparé systématiquement les taux de reconnaissances à ceux obtenus à l'aide d'un modèle de Markov Caché global M_{glob} construit de manière classique en utilisant aussi le pseudo-diphone comme unité élémentaire. Un vecteur d'observations est traité globalement, à raison d'une loi gaussienne par transition (matrice de covariance diagonale).

R I / RESULTATS SUR LE CORPUS DES CHIFFRES :

Pour l'application des chiffres connectés , le modèle global M_{glob} est appris avec 8 coefficients cepstraux, l'Energie E, et la durée T. Sont adjoints les 3 coefficients labiaux dans un deuxième temps. Ce modèle est alors comparé au modèle Maître-Esclave.

Les meilleurs résultats sont obtenus avec le MMC M_{glob} avec 8 coefficients cepstraux, l'énergie et la durée du segment : 1 mot sur 125 est mal reconnu. L'ajout des paramètres labiaux n'entraine qu'une petite dégradation (2 erreurs de substitution supplémentaires) (figure 3). Pour des paramètres d'observations équivalents, le modèle global s'avère être meilleur que le modèle Maître-Esclave (3 erreurs contre 5), mais nous devons observer que l'intervalle de confiance ne permet pas d'en tirer une conclusion définitive. Etant donné la complexité du modèle Maître-Esclave et le faible ensemble d'apprentissage dont nous disposons, nous ne pouvons espérer obtenir une bonne estimation de la totalité des paramètres.

R II / RESULTATS SUR LE CORPUS DES LETTRES :

Cette deuxième évaluation sur le corpus des lettres épelées et connectées, nous permet d'examiner les performances du système afin de quantifier plus correctement l'apport des paramètres labiaux, dans les deux sortes de MMC.

Le MMC global est appris initialement avec 8 coefficients cepstraux, l'énergie et la durée. Nous avons ajouté successivement les paramètres labiaux et leurs dérivées. La même expérience a été répétée en initialisant le modèle global avec 8 coefficients cepstraux, leurs quatre premières dérivées, l'énergie ainsi que sa dérivée, et la durée du segment. Le meilleur taux de reconnaissance, à savoir **91,6** % (taux mots) est obtenu en utilisant la hauteur et la largeur des lèvres (figure 4). L'introduction de la surface des lèvres n'apporte pas d'information pertinente car elle est fortement corrélée aux paramètres A et B [Benoit 91]. Les dérivées des coefficients labiaux dégradent le taux de reconnaissance : une des causes principales peut être le manque de synchronisation entre les informations labiales et acoustiques, ou le manque de données d'apprentissage.

Le même protocole d'expérimentation est réalisé pour tester l'approche Maître-Esclave. Le coefficient labial A fait partie des paramètres initiaux. Le meilleur taux de reconnaissance est obtenu par le modèle MMC5, à savoir 91,7 % en terme de mots correctement reconnus. Lorsque le nombre de paramètres augmente, les performances décroissent, la cause est très certainement liée au relativement faible ensemble de données d'apprentissage par rapport au nombre de paramètres à apprendre.

5 . Conclusion

Nous avons présenté deux approches probabilistes pour traiter la fusion de données acoustiques et articulatoires dans un but de reconnaissance. L'approche classique consiste à supposer les informations issues des deux canaux indépendantes tandis que l'approche maître-esclave exploite une certaine corrélation par l'intermédiaire de liens entre les lois d'observation.

Les deux approches *modèle global* et *modèle maître-esclave* donnent des résultats très comparables dans le cadre de la reconnaissance mono locuteur de suites de chiffres connectés ou de lettres épelées (92 % de taux de reconnaissance en mots). L'avantage de la deuxième méthode est liée à une meilleure compréhension du phénomène labial et offre des perspectives intéressantes :

— Au niveau maître, nous augmenterons le nombre d'états de manière à se rapprocher des études statistiques qui ont montré l'émergence de visèmes [Benoit 91]
— Le modèle maître-esclave est, dans son actuelle implémentation, fort simplifié et certaines hypothèses sont trop fortes : le passage d'un état *ouvert* à celui de *fermé* ne se réalise pas de manière instantannée! En fonction du volume croissant de l'ensemble d'apprentissage qui nous sera ultérieurement fourni, nous complexifierons le modèle simplifié pour tendre vers le modèle exact et tester ses réelles possibilités.
—L'étude d'une désynchronisation entre le labial et l'acoustique est plus abordable par cette approche.

L'utilisation des paramètres labiaux a pour but de rendre plus robuste la reconnaissance automatique de parole en milieu bruité ; nous avons montré que cette information ne dégradait absolument pas les performances des systèmes actuels déjà très performants. Nous sommes actuellement en cours d'évaluation des deux approches sur la même application en milieu bruité.

REFERENCES

[André-Obrecht, 88] R. André-Obrecht : *A new statistical approach for the automatic segmentation of continuous speech signals*, IEEE Trans. on Acoustics, Speech, Signal Processing, vol. 36, n°1, janvier 1988.

[André-Obrecht, 93] R. André-Obrecht : *Segmentation et parole?* Habilitation à diriger des recherches, IRISA, Rennes, juin 1993.

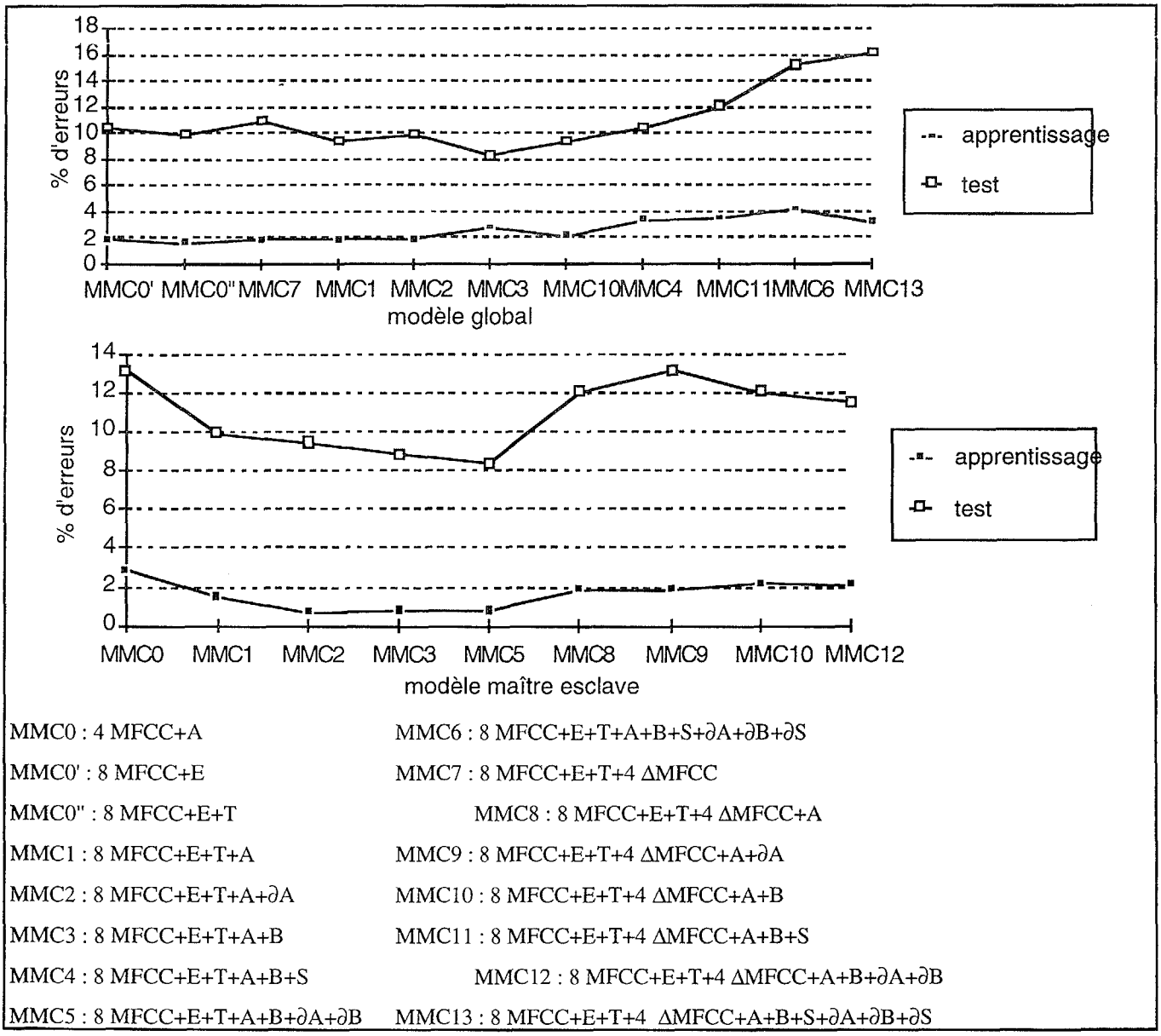
[Benoit 91] C. Benoit, C. Abry, L.J. Boë : *The effect of context on labiality in french*. Eurospeech 91, Genova.

[Brugnara, 92] F. Brugnara, R. De Mori, D. Guiliani, M. Omologo : *A family of Parallel Hidden Markov Models*, ICASSP 92, San Francisco, 1992.

[Lallouache 91] T. Lallouache : *Un poste "visage parole" couleur. Acquisition et traitement automatique des contours de lèvres.* Thèse de doctorat de l'Institut National Polytechnique de Grenoble, 1991.

modèle	coefficients	phrases / 35	mots / 125
M_{glob}	8 MFCC + E + T	1	1
	8 MFCC + E + T + A + B + S	3	3
$M_{m/e}$	8 MFCC + E + T + A + B + S	5	5

Figure 3: Nombre d'erreurs sur l'ensemble test en terme de phrases et mots incorrectement reconnus, en fonction des coefficients et de la modélisation utilisés.



MMC0 : 4 MFCC+A
MMC0' : 8 MFCC+E
MMC0" : 8 MFCC+E+T
MMC1 : 8 MFCC+E+T+A
MMC2 : 8 MFCC+E+T+A+∂A
MMC3 : 8 MFCC+E+T+A+B
MMC4 : 8 MFCC+E+T+A+B+S
MMC5 : 8 MFCC+E+T+A+B+∂A+∂B
MMC6 : 8 MFCC+E+T+A+B+S+∂A+∂B+∂S
MMC7 : 8 MFCC+E+T+4 ΔMFCC
MMC8 : 8 MFCC+E+T+4 ΔMFCC+A
MMC9 : 8 MFCC+E+T+4 ΔMFCC+A+∂A
MMC10 : 8 MFCC+E+T+4 ΔMFCC+A+B
MMC11 : 8 MFCC+E+T+4 ΔMFCC+A+B+S
MMC12 : 8 MFCC+E+T+4 ΔMFCC+A+B+∂A+∂B
MMC13 : 8 MFCC+E+T+4 ΔMFCC+A+B+S+∂A+∂B+∂S

Figure 4: Pourcentage d'erreurs en termes de mots mal reconnus