Évaluation de modèles d'extraction d'informations visuelles pour la reconnaissance automatique de parole audiovisuelle

Philippe Daubias, Paul Deléglise

LIUM (Laboratoire d'Informatique de l'Université du Maine)
Institut d'Informatique Claude Chappe, 72085 Le Mans Cedex 9 FRANCE
Tél. : ++33 (0)2 43 83 38 58 - Fax : ++33 (0)2 43 83 38 68
Mél1 : philippe.daubias@lium.univ-lemans.fr - http ://www-ic2.univ-lemans.fr/~daubias
Mél2 : paul.deleglise@lium.univ-lemans.fr - http ://www-ic2.univ-lemans.fr/lium/pagesperso/deleglise

ABSTRACT

In this article, we make a progress report of our research towards lipreading in close to "natural" conditions. More precisely, we describe first audio-visual speech recognition experiments carried using visual parameters extracted from "natural" images. Unlike many other experiments in the AV ASR field, these visual parameters are obtained without any hand-labelling phase and are naturally noisy, due to the extraction process. We evaluate our models through different ways of using them. These strategies include the use of shape model combined with the appearance model and the use of appearance model followed by the use of shape model. For the audio and visual parameters fusion, we used a basic DI architecture with a fixed weight and afterwards with an adaptative weighting scheme based on an energy criterion.

1. INTRODUCTION

La Reconnaissance Automatique de la Parole (RAP) est une tâche complexe en milieu bruité. Il est possible d'effectuer un ensemble de traitements sur le signal pour le débruiter avant de l'utiliser dans un système de reconnaissance ou de construire des systèmes de reconnaissance robustes à certains types de bruits en isolant différentes bandes de fréquence comme cela a été fait par Besacier [2]. Une autre possibilité réside dans l'utilisation d'informations complémentaires issues d'autres capteurs non affectés par ce bruit. Dans cette catégorie, la parole audiovisuelle propose d'étudier le mouvement des lèvres du locuteur. Le canal visuel présente l'intérêt de ne pas être soumis au bruit acoustique et il apporte une information complémentaire à celle véhiculée par l'acoustique. Les travaux en RAP audiovisuelle se sont intéressés au choix des paramètres pertinents et à leur apport par rapport à l'information acoustique seule. Dans ce cadre, il faut également étudier les modes de fusion des paramètres visuels avec les informations acoustiques. La plupart des travaux ont été réalisés à partir de paramètres "idéaux", extraits dans des conditions favorables, ce qui a permis de mesurer sans effet de bords la pertinence des différents paramètres et d'élaborer différents modèles de fusion et de les valider ou les invalider en les comparant sur des bases fiables. Pour notre part, nous nous intéressons à l'extraction des paramètres visuels sur des images que nous qualifions de "naturelles", c'est-à-dire acquises dans des conditions que nous pensons réalistes, sans utilisation d'artefacts (maquillage ou pastilles réfléchissantes), ni prise de vue ou utilisation de dispositif d'acquisition spécifique et sans contrôle particulier sur les conditions d'éclairement (utilisation de la lumière solaire ambiante). Nous cherchons tout d'abord à valider les modèles que nous avons construits pour extraire les paramètres, puis par la suite, à étudier les différents protocoles d'utilisation de ces modèles. Nous nous intéressons également aux paramètres obtenus à l'aide de ces modèles pour déterminer quel peut être leur apport dans le cadre de la RAP audiovisuelle. Les modèles labiaux seront présentés dans une première section, puis nous exposerons les différents modes d'utilisation que nous proposons en présentant les expériences de reconnaissance de parole que nous avons effectuées. Enfin, nous abordons l'évaluation au travers de cette expérience de RAP audiovisuelle dans la section résultats.

2. MODÈLES LABIAUX

Nous décrivons brièvement dans cette partie les modèles que nous avons développés et utilisés pour extraire les informations labiales des images de notre locuteur. L'intérêt majeur de ces modèles est, selon nous, qu'ils sont appris à partir de données, ce qui permet de les adapter en fonction de la tâche de reconnaissance de parole à effectuer. La phase d'apprentissage étant entièrement automatisée, cela permet de réduire d'autant le travail nécessaire pour apprendre le modèle dans de nouvelles conditions et assure la possibilité effective de l'adapter. L'objet de cet article étant davantage de présenter l'évaluation de nos modèles au travers des résultats obtenus lors d'expériences de RAP audiovisuelle que de présenter les modèles eux-mêmes, nous ne détaillerons pas ces derniers. Une présentation plus approfondie des modèles est disponible par ailleurs dans [5]. Nous allons donc brièvement présenter les deux modèles des lèvres : celui de la forme, puis celui de l'apparence des lèvres.

2.1. Forme

Les lèvres sont une entité déformable à 3 dimensions, évoluant dans l'espace pendant la production de parole. Pour obtenir une modélisation fidèle des lèvres et de leurs mouvements, il faut utiliser un modèle 3D, comme cela a été fait notamment à l'ICP [3] ou par Basu [1]. Une majeure partie de l'activité labiale peut cependant être observée sur une unique vue de face, mais, lorsque l'on ne dispose que d'une telle prise de vue, il est difficile d'extraire tous les paramètres nécessaires pour connaître la forme 3D des lèvres. Revéret [12] a montré que cela était réalisable à partir d'images où les lèvres du locuteur sont maquillées en bleu, mais ceci est beaucoup plus complexe avec des locuteurs non maquillés filmés dans des condi-

TAB. 1 – Pourcentage de la variance globale en fonction du nombre de vecteurs utilisé pour différents modèles de notre locuteur.

	nombre de vecteurs				
locuteur	1	2	3	4	5
BJ(lett)	94.51	96.60	97.62	98.33	98.80
BJ(ppeq)	86.27	90.10	93.08	94.97	96.05
BJ(mixte)	90.16	92.79	95.17	96.53	97.36

tions d'éclairage non-idéales. Dans notre cas, n'ayant pas rencontré dans la littérature de modèle *a priori* validé pour de nombreux locuteurs différents, nous avons préféré poser l'hypothèse que nous ne disposions pas d'un modèle des lèvres. Nous avons choisi d'utiliser une approximation polygonale 2D des lèvres, pour construire un modèle *a posteriori* en apprenant statistiquement la forme et les déformations 2D à partir de données.

Le modèle de la forme que nous avons utilisé est, comme dans nos travaux précédents ([4, 5]), un polygone inspiré des travaux de Luettin [10]. Contrairement au modèle présenté dans [5], qui était appris à partir de plusieurs locuteurs, le modèle utilisé pour les travaux présenté dans cet article est monolocuteur. Il a été appris sur le locuteur à reconnaître pour le même style de corpus de parole que celle étudiée dans l'expérience de reconnaissance, à savoir des séries de 4 lettres de l'alphabet français tirées aléatoirement, sans répétition, épelées en élocution continue. L'utilisation d'un modèle monolocuteur avait pour but de contraindre fortement le modèle en excluant des déformations liées au changement de locuteur ou à une élocution ne correspondant pas à la tâche de parole à étudier. Dans la pratique, le modèle monolocuteur obtenu est cependant très semblable à ceux conçus précédemment.

La table 1 présente le pourcentage de la variance globale obtenue en fonction du nombre de vecteurs utilisés pour différentes tâches de parole pour notre locuteur : l'épellation de lettres (lett), la production de phrases phonétiquement équilibrées (ppeq) ou les deux (mixte). En rapprochant ces résultats d'autres obtenus précédemment [5], on constate que la qualité de l'approximation faite avec les quelques premiers vecteurs propres dépend du locuteur et naturellement que plusieurs locuteurs engendrent des modes de variation plus nombreux qu'un seul. Ces résultats permettent également d'observer que différentes tâches de parole pour un même locuteur engendrent des modèles différents. Par ailleurs, nous avons amélioré la procédure de normalisation des contours, en les redressant par interpolation linéaire. Ce redressement permet d'éliminer un biais qui existait dans le cas où le locuteur penche la tête de façon importante (typiquement 10 degrés). En effet, dans ce cas, les points de contour qui sont extraits le long de colonnes des images, après normalisation en rotation, ne se trouvaient plus deux à deux sur des verticales. Cette différence de position bien que très légère et invisible dans le cas d'un tracé de contour pouvait avoir des répercussions lors de l'Analyse en Composante Principale qui permet de construire le modèle, en ajoutant un mode de variation "parasite" correspondant à un déplacement latéral des points.

Ces résultats montrent que l'on obtient pour la tâche étudiée, une approximation très fidèle de la forme des lèvres en utilisant les 4 premiers modes de déformation. Ceci va permettre d'effectuer une recherche dans un espace à 8 dimensions : 4 dimensions correspondent aux 4 modes de déformation retenus, les 4 autres dimensions correspondent aux déplacements dans l'espace suivant :

- x (déplacement horizontal et faible rotation autour de l'axe y),
- y (déplacement vertical et faible rotation autour de x),
- z (variation du facteur de zoom ou de la distance à la caméra)
- rot (rotation autour de l'axe z)

En raison du manque de robustesse aux conditions changeantes de l'environnement du modèle de l'apparence *a priori* constitué par l'utilisation de l'information de teinte centrée sur le rouge, que nous avions initialement choisi, nous avons également été amené à proposer une modélisation *a posteriori* de l'apparence.

2.2. Apparence

Le modèle de l'apparence que nous avons utilisé est également un modèle utilisant un apprentissage statistique. Il s'agit d'un réseau de neurones (plus précisément un perceptron multi-couches) ayant une couche d'entrée contenant 75 noeuds (un par valeur Rouge, Vert ou Bleu de chaque pixels d'un bloc d'image carré de 5 points de coté), une couche cachée de 10 ou 15 noeuds et 3 noeuds en sortie qui indiquent la probabilité d'appartenance, pour le bloc placé en entrée, aux classes "peau", "lèvres" ou "intérieur de bouche". De la même manière que pour le modèle de forme, nous avons souhaité entraîner le modèle spécifiquement à notre locuteur. Pour cela, nous lui avons fait prononcer 5 phrases phonétiquement équilibrées de BDSONS [6] avec et sans maquillage bleu sur les lèvres. L'entraînement de nos modèles d'apparence est supervisé, mais nous le réalisons sans étiquetage manuel avec des traitements automatiques : nous avons utilisé la programmation dynamique pour établir la correspondance entre les phrases prononcées avec maquillage et sans maquillage. Nous avons ainsi utilisé l'information acoustique et les formes extraites grâce au maquillage pour étiqueter automatiquement les blocs d'image à fournir au réseau de neurones pour apprentissage (pour plus de détails, voir [5]). Nous avons comparé par inspection visuelle sur une partie des images, les modèles monolocuteur avec des modèles pluri-locuteurs obtenus préalablement [4]. Les modèles spécifiques ont toujours été au moins aussi précis que les pluri-locuteurs et même meilleurs dans de nombreux cas. Lors de notre étude précédente [5], nous avions mis en évidence qu'il y avait un écart relativement faible entre différents modèles entraînés. Les modèles entraînés avec des blocs d'images "certifiés", obtenus en excluant les blocs proches des contours qui peuvent être sujets à des erreurs d'étiquetage, étaient cependant légèrement meilleurs et nous avons choisi, pour des raisons de temps de calcul, d'utiliser dans l'étude présentée ici, le réseau à 10 noeuds dans la couche cachée entraîné avec ces blocs : bjcert10.

On peut remarquer que, contrairement à ce qui a été fait pour le modèle de forme où l'entraînement a été effectué sur des images correspondant parfaitement à la tâche de reconnaissance de parole, l'alignement a été effectué sur des phrases phonétiquement équilibrées qui ne correspondent pas à la tâche de parole étudiée. Il se peut donc que la partie du modèle correspondant à l'identification de la classe "intérieur de bouche" soit mal entraînée, mais

les répercussions sur la détection des lèvres doivent être faibles. La figure 1 présente, pour trois des 80 séquences du corpus utilisé, la 10ème image ainsi que le résultat de la détection des lèvres par le modèle d'apparence bjcert10. L'image centrale présente la série que nous avons rejetée. Pour cette séquence, la caméra a retourné des images plus bleues qu'habituellement, ce qui a perturbé de façon importante la détection des lèvres par le modèle d'apparence qui n'avait pas été entraîné à ce type d'images.

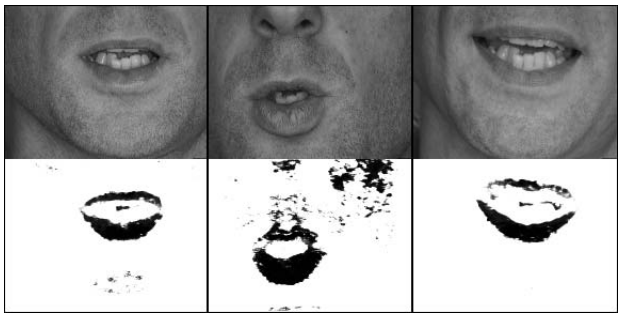

FIG. 1 – Une image de trois séquences du corpus (haut) et détection de lèvre par le modèle d'apparence bjcert10.

3. EXPÉRIENCE

3.1. Partie acoustique

Nous utilisons comme paramètres acoustiques 12 coefficients MFCC ainsi que leurs vitesses et accélérations. Pour entraîner nos modèles acoustiques, nous avons étiqueté manuellement chacune des 80 séries de 4 lettres prononcées par notre locuteur. Nous avons ensuite déterminé une transcription phonétique des 26 lettres en gérant d'éventuelles variations de prononciation. Nous sommes ainsi arrivé à une taille de vocabulaire de 29 phonèmes. Nous n'avons utilisé aucune grammaire pour tenir compte de la tâche que nous traitions, une séquence à analyser est considérée comme une suite de phonèmes et de silences. Nous avons rejeté une séquence du corpus à cause d'un problème avec la caméra qui rendait hasardeux l'extraction de paramètres visuels sur ces images. Les modèles sont alors entraînés sur 78 phrases et testés sur la 79ème. Nous effectuons ainsi 79 entraînements et tests différents en testant sur chacune des phrases un modèle entraîné sur toutes les autres.

3.2. Partie visuelle

Pour les expériences présentées ici, nous avons utilisé les paramètres A, B et S proposés par Fromkin [7] ainsi que leurs vitesses et accélérations. Ces paramètres mesurent : l'étirement labial (A), la séparation intéro-labiale (B) et l'aire intéro-labiale (S). Nous les calculons automatiquement à partir du modèle de la forme des lèvres en utilisant le polygone décrivant le contour interne. Pour obtenir les contours et calculer ces mesures, nous avons testé différents protocoles d'utilisation des modèles labiaux présentés dans la section précédente :

p1 - utilisation combinée des 2 modèles : recherche par minimisation des paramètres du modèle de forme en utilisant une carte des régions appartenant à la classe "lèvres" obtenue avec le modèle d'apparence,

p2 - utilisation du modèle d'apparence seul : le contour est extrait directement à l'aide de règles heuristiques en utilisant l'image issue du modèle d'apparence,

p3 - modèle d'apparence puis modèle de forme : recherche des paramètres du modèle de forme permettant d'obtenir le contour le plus proche du contour obtenu par le traitement précédent (p2).

Nous avons également sélectionné les meilleurs contours obtenus par les 3 approches précédentes selon le critère énergétique utilisé pour la recherche par minimisation (p1). Les contours résultants ont été utilisés pour calculer des paramètres que nous appellerons pm par la suite.

3.3. Fusion du visuel dans l'acoustique

Pour fusionner les paramètres acoustiques et visuels, nous avons utilisé l'identification directe (ID). Bien que nous sachions que ce modèle n'est pas le meilleur pour la fusion des informations acoustiques et visuelles (voir notamment Rogozan [13]), notre choix s'est porté sur lui principalement en raison de la simplicité de sa mise en oeuvre. De nombreux travaux ont étudié la fusion des canaux acoustiques et visuels et se sont en particulier intéressés à la pondération des deux sources d'informations. Parmi les plus récents, on peut citer Heckmann [9] ou Glotin [8]. Nous avons dans un premier temps choisi d'utiliser une pondération fixe des canaux acoustiques et visuels pour mesurer dans quelle mesure les informations visuelles potentiellement bruitées allaient perturber l'acoustique, mais nous avons également effectué des premières expérimentations de pondération adaptative basées sur un critère non encore étudié, lié à la fonction d'énergie utilisée pour minimiser le modèle de forme sur le modèle d'apparence. Il faut signaler que la plupart des expériences supposent des paramètres visuels de qualité constante, ce qui n'est pas le cas sur le corpus que nous utilisons, il est donc d'autant plus important de pouvoir mesurer la validité des paramètres visuels obtenus.

Différence de cadence Pour extraire les paramètres visuels, nous avons utilisé des images acquises à une cadence de 25 Hz dans un mode non-entrelacé. Le choix que nous avons effectué peut sembler aboutir à une cadence trop faible, mais les expériences de Potamianos [11], montrent qu'une diminution de la fréquence des mesures visuelles de 60 à 20 Hz ne dégrade pas de façon importante la qualité de la lecture labiale. La limite critique semblerait se situer autour de 15 Hz. Nous avons donc préféré privilégier la définition sur la cadence des mesures. Pour faire une fusion ID, il faut atteindre la même cadence pour les paramètres visuels que pour l'acoustique. Pour se faire, nous avons utilisé une interpolation par spline sous tension pour déterminer les paramètres visuels manquants.

4. RÉSULTATS ET ÉVALUATION

Pour mesurer la qualité de la reconnaissance et l'apport de l'information visuelle, nous utilisons comme indice la précision Acc, qui est calculé comme en reconnaissance de la parole "classique" avec :

$$Acc = \frac{N - D - S - I}{N} * 100\%$$

Où N est le nombre d'éléments à reconnaître, D le nombre d'éléments supprimés, S le nombre d'éléments substitués (reconnus à la place d'un autre) et I le nombre d'éléments insérés (reconnus mais non présents).

TAB. 2 – Précision pour différents paramètres visuels en fonction du poids visuel - parole non bruitée

paramètres	poids du visuel				
	0.10	0.15	0.20	0.25	0.30
Audio	92.77				
p1	92.55	92.84	92.84	92.84	92.98
p2	92.49	92.77	92.77	92.77	92.92
p3	91.76	91.62	91.04	90.61	90.90
pm	92.49	92.63	92.49	93.06	92.92
ref2	92.55	92.69	92.55	91.83	91.69

TAB. 3 – Précision pour différent paramètres visuels en fonction du poids visuel - parole bruitée à 0 dB

paramètres	poids du visuel				
	0.10	0.15	0.20	0.25	0.30
Audio	70.38				
p1	70.06	70.92	71.35	71.49	71.78
p2	69.94	70.81	71.24	71.39	71.68
p3	69.36	70.38	71.39	68.93	67.20
pm	73.27	74.71	74.71	73.84	72.54
ref2	72.06	72.92	72.49	71.33	71.78

Nous étudions le taux de reconnaissance au niveau phonétique non pas au niveau des “phrases” (séries de 4 lettres) et rappelons que nous n’utilisons ni grammaire, ni lexique. Nous avons utilisé des pondérations fixes entre les canaux acoustiques et visuels pour toutes les séries et avons fait différentes expérimentations en faisant varier ces poids. Nous avons effectué des expérimentations en utilisant directement le signal acoustique (table 2) et également après l’avoir dégradé avec du bruit de foule à 0 dB (table 3). Les paramètres ref2 correspondent aux p2, mais ont été obtenus grâce à un modèle d’apparence construit à partir d’un étiquetage manuel des images.

Les résultats suivants nous semblent importants : les taux de reconnaissance obtenus avec les modèles ref2 et p2 sont très proches, que ce soit en milieu bruité ou non. Ceci confirme que la méthode de construction automatique de modèle d’apparence que nous proposons donne des modèles très proches de ceux que l’on peut obtenir manuellement avec l’avantage de pouvoir construire de tels modèles dans un cadre multilocuteur, ce qui est difficilement envisageable avec l’étiquetage manuel. Les paramètres pm permettent d’atteindre les meilleurs scores de reconnaissance, ce qui semble montrer que la fonction d’énergie utilisée pour les sélectionner mesure effectivement la qualité des paramètres. Enfin, avec un poids de 0.25, les résultats obtenus en ajoutant l’information visuelle dépassent ceux obtenus avec le canal acoustique seul. Ceci est surtout significatif en présence de bruit.

Nous avons également effectué une expérimentation sur l’utilisation ou non de l’information visuelle en fonction du critère énergétique pour chaque séquence avec les paramètres qui avaient donné les meilleurs résultats (pm avec un poids de 0.20) et avons obtenus des résultats légèrement plus élevés, 75.00 % de précision. Ceci semble à nouveau montrer que la fonction d’énergie que nous utilisons pour localiser le modèle sur l’image permet également d’évaluer l’information visuelle obtenue.

5. CONCLUSIONS ET PERSPECTIVES

Les résultats montrent d’une part qu’il semble possible d’utiliser la procédure que nous proposons ne recourant à aucun étiquetage manuel, pour construire des modèles de forme et d’apparence des lèvres utilisables pour une lecture labiale sur des images “naturelles”. Les résultats ne sont pas spectaculaires, étant donnée la tâche de reconnaissance de parole assez simple, cependant, ils montrent nettement qu’une amélioration des taux de reconnaissance (précision) est possible en utilisant une information labiale potentiellement bruitée. D’autre part, nous avons pu constater la supériorité d’une sélection entre audiovisuel et acoustique seul en fonction d’un critère énergétique. Ce critère semble efficace pour mesurer la qualité d’un contour et donc des paramètres extraits. Ceci nous conforte sur la voie d’une pondération adaptative basée sur ce nouvel indice (l’énergie du modèle de l’apparence). Enfin, selon de premières expérimentations, l’utilisation des mesures A, B et S se révèle plus efficace que l’utilisation directe des coefficients correspondants aux vecteurs propres, les mesures ayant un effet régularisant sur des paramètres vraisemblablement bruités. Par la suite, nous allons continuer nos expérimentations pour déterminer le meilleur mode d’utilisation de nos modèles et travailler sur les indices permettant d’évaluer la qualité de la mesure labiale pour pondérer de façon adaptée avec le plus de finesse possible (paramètre par paramètre).

RÉFÉRENCES

[1] S. Basu, N. Oliver, and A. Pentland. 3D modeling and tracking of human lip motion. In *Proc. ICCV*, pages 337–343, Bombay, India, January 1998.

[2] L. Besacier. *Un Modèle Parallèle pour la Reconnaissance Automatique du Locuteur*. Ph.D. thesis, Université d’Avignon, 1998.

[3] P. Borel, P. Badin, L. Revéret, et G. Bailly. Modélisation articulatoire linéaire 3D d’un visage pour une tête parlante virtuelle. In *XXIIIèmes JEP*, pages 121–124, Aussois, June 2000.

[4] P. Daubias and P. Deléglise. Evaluation of an automatically obtained shape and appearance model for automatic audio visual speech recognition. In *Proc. Eurospeech*, vol 2, pages 1031–1034, Aalborg, Denmark, 2001.

[5] P. Daubias et P. Deléglise. Construction de modèles pour l’extraction des informations visuelles en vue de la reconnaissance de la parole audiovisuelle. In *Proc. RFIA’02*, Angers, January 2002.

[6] R. Descout, J.-F. Sérignat, O. Cervantes, and R. Carré. BDSONS : Une base de donnés des sons du français. In *Proc. 12th ICA*, Toronto, Canada, 1986.

[7] V. Fromkin. Lip positions in american-english vowels. *Language and Speech*, 7(3) :215–225, 1964.

[8] H. Glotin, D. Vergyri, C. Neti, G. Potamianos, and J. Luettin. Weighting schemes for audio-visual fusion in speech recognition. In *Proc. ICASSP*, vol 1, Salt Lake City, USA, May 2001.

[9] M. Heckmann, F. Berthommier, and K. Kroschel. Optimal weighting of posteriors for audio-visual speech recognition. In *Proc. ICASSP*, volume 1, Salt Lake City, USA, May 2001.

[10] J. Luettin, N. A. Thacker, and S. W. Beet. Active shape models for visual speech feature extraction. In D. G. Stork and M. E. Hennecke, editors, *Speechreading by Humans and Machines*, volume 150 of *NATO ASI*, pages 383–390. Springer, Berlin, 1996.

[11] G. Potamianos, H. P. Graf, and E. Cosatto. An image transform approach for HMM based automatic lipreading. In *Proc. ICIP*, volume III, pages 173–177, Chicago, USA, October 1998.

[12] L. Revéret and C. Benoît. A new 3D lip model for analysis and synthesis of lip motion in speech production. In *Proc. AVSP*, pages 207–212, Terrigal, Australia, December 1998.

[13] A. Rogozan and P. Deléglise. Adaptive fusion of acoustic and visual sources for automatic speech recognition. *SpeechCom*, 26 Iss. 1-2 :149–161, December 1998.