Construction d'un modèle des lèvres pour la reconnaissance de parole audiovisuelle

Philippe Daubias
Laboratoire d'Informatique de l'Université du Maine
F-72085 Le Mans Cedex 9, France
philippe.daubias@lium.univ-lemans.fr

Résumé

La parole visuelle a démontré son utilité, particulièrement dans les situations où le canal acoustique est bruité. L'utilisation de la parole audiovisuelle suppose néanmoins la transmission de volumineuses informations (images et son), ce qui limite souvent son utilisation. Nous avons construit des modèles des lèvres à partir de données et nous proposons d'utiliser ces modèles pour extraire de façon automatique des paramètres labiaux sur des images acquises dans des conditions "naturelles". Nous modélisons la forme des lèvres grâce à leurs contours internes et externes et obtenons une représentation compacte en utilisant les paramètres issus d'une ACP. Au travers d'un ensemble de mesures, nous montrons qu'il est possible de modéliser la forme des lèvres d'un locuteur avec un nombre très restreint de paramètres, tout en conservant l'essentiel de l'information labiale.

Mots Clef

Reconnaissance de parole audiovisuelle, étiquetage automatique de formes, modèle des lèvres.

1 Introduction

La reconnaissance automatique de parole a vu de considérables avancées au cours des 20 dernières années. Il existe maintenant des systèmes commerciaux de dictée vocale ayant un taux de reconnaissance relativement élevé. Cependant, les taux de reconnaissance diminuent très significativement en présence de bruit. Certains travaux en reconnaissance de la parole visent à améliorer la robustesse des systèmes, notamment en isolant la bande de fréquence où se situe le bruit pour que le système ne soit pas affecté par sa présence [1]. Mais un bruit réparti sur l'ensemble des fréquences ou un locuteur perturbateur ("cocktail-party effect") restent dévastateurs pour les performances de ces systèmes. Une des méthodes pour palier à cette perte de performance est l'utilisation d'une autre source d'informations en complément du signal acoustique bruité. La reconnaissance de parole audiovisuelle se propose d'utiliser le mouvement des lèvres du locuteur comme source complémentaire d'information. Pour faciliter l'extraction des paramètres labiaux, les recherches en reconnaissance de parole audiovisuelle utilisent généralement des dispositifs d'acquisition intrusifs [7] ou un maquillage spécifique du locuteur [2], ce qui est difficilement envisageable dans des applications réelles. Ces artefacts sont utilisés pour faciliter l'extraction automatique des paramètres labiaux, problème de traitement d'images complexe en raison de sa grande variabilité. Nous souhaitons travailler dans des conditions relativement naturelles, c'est-à-dire en imposant un minimum de contraintes au locuteur. Nous supposons possible une localisation de la bouche [9] et avons filmé nos locuteurs dans des conditions réalistes (lumière solaire ambiante, pas de maquillage), en cadrant la bouche approximativement au centre de l'image. Contrairement à d'autres recherches utilisant directement ce type d'images [9], nous souhaitons passer par une étape de modélisation qui permet d'étudier le phénomène de la production de parole et qui devrait également rendre le système plus robuste. Enfin, nous souhaitons construire nos modèles à partir de corpus, ce qui pousse à limiter les interventions manuelles, fastidieuses voire irréalisables pour de grands corpus [10].

2 Approche modèle ou image

En reconnaissance de parole audiovisuelle, deux questions principales se posent : "quels paramètres extraire des images et comment ?" et "comment combiner les informations acoustiques et visuelles ?". Des éléments de réponse à la seconde question ont été proposés dans [12]. Pour notre part, nous nous sommes principalement intéressés à la question de l'extraction automatique des paramètres visuels sur des images acquises dans des conditions "naturelles". Ce problème est complexe car la bouche est une entité fortement déformable dont l'aspect peut changer au cours de la production de parole (visibilité de la langue et des dents notamment). De plus, il existe une grande variabilité entre locuteurs. Dans notre cas, en raison des conditions d'éclairement non contraintes, à ces variabilités inter- et intra-locuteur viennent s'ajouter des problèmes d'ombres.

Deux types d'approches ont été utilisées ces dernières années pour l'extraction d'informations labiales à partir d'images du locuteur : des approches de bas niveau orientées image et d'autres de plus haut niveau fondées sur l'utilisation de modèles. L'approche de bas niveau consiste à localiser approximativement la bouche et à opérer des traitements sur cette partie de l'image à des fins de normalisation, pour limiter l'effet des variations de prise de vue (éclairement, rotation, zoom), ensuite, l'image entière est fournie au système de reconnaissance (généralement après une phase permettant de diminuer la dimension du vecteur d'observation). Il a été montré [11] que les approches bas niveau amènent des résultats supérieurs à ceux obtenus avec les méthodes orientées modèle dans de nombreuses conditions. Cette supériorité vis à vis des approches orientées modèle s'explique principalement, de notre point de vue, par le manque de robustesse des approches modèle utilisées. La mauvaise localisation ou le mauvais suivi des lèvres fournit des données erronées aux systèmes de reconnaissance et ce "bruit visuel" fait diminuer significativement les performances. Cependant, seuls les modèles peuvent permettre une étude du mouvement des lèvres, ainsi qu'une comparaison avec les résultats obtenus en production, c'est pourquoi nous pensons qu'il est utile de persévérer dans la voix de la modélisation. Nous avons donc choisi de construire un modèle de la forme et de l'apparence des lèvres à partir de données. Plus précisément, nous avons appris statistiquement un modèle de la forme des lèvres en utilisant des images où les lèvres sont maquillées en bleu, puis nous avons cherché à construire un modèle de l'apparence des lèvres sur des images où les locuteurs n'ont subi aucun maquillage.

3 Modélisation de la forme

Dans cette partie, nous décrivons le modèle de la forme que nous avons utilisé et la manière dont nous l'avons obtenu, puis nous fournissons des mesures permettant d'évaluer la qualité d'une représentation compacte du modèle, obtenue par Analyse en Composantes Principales (ACP), utilisable pour localiser le modèle sur des images.

3.1 Description du modèle

Notre modèle de forme des lèvres est un polygone à 44 points représentant les contours interne et externe de la bouche (inspiré de [8], développé dans [4]). La forme et les déformations de ce polygone sont apprises à partir de données. Pour construire cette partie du modèle, nous avons utilisé une craie de maquillage bleu (teinte absente des pigments de la peau), pour peindre les lèvres des locuteurs et faciliter la segmentation. Nous avons développé une méthode d'extraction automatique du contour des lèvres, robuste à différentes conditions d'éclairement, en utilisant cette teinte. Nous avons normalisé les contours obtenus en échelle, rotation et translation, en nous efforçant de ne pas perdre d'informations qui pourraient s'avérer utiles pour la reconnaissance de parole. Ensuite, nous avons effectué une ACP sur les contours résultants. Chaque contour c observé dans le corpus peut être obtenu à partir de la forme moyenne $\overline{c}$ des lèvres en lui appliquant les différents vecteurs propres de déformation v_i pondérés par les poids p_i correspondant aux nouvelles coordonnées du contour dans le repère obtenu par ACP. L'ACP donne en plus des vecteurs propres de déformation, leur ordre d'importance en terme de pourcentage de la variance totale, ce qui signifie que les quelques premiers vecteurs propres de déformation suffisent à représenter l'essentiel de la déformation du contour.
Dans tous les cas de figure que nous avons expérimentés (monolocuteur pour différents locuteurs et différentes tâches de parole et pluri-locuteurs), les quatre premiers vecteurs propres représentaient nettement plus de 90% de la déformation. Notre modèle est ainsi composé d'un polygone $\overline{c}$ représentant la forme moyenne des lèvres et de quatre vecteurs v_i permettant de déformer ce polygone pour obtenir une bonne approximation de tous les contours c observés dans le corpus.

$$c \approx \overline{c} + p_1 v_1 + p_2 v_2 + p_3 v_3 + p_4 v_4 \qquad (1)$$

La figure 1 présente le recouvrement entre des contours extraits sur les images après normalisation et ceux estimés par projection sur les quatre premiers vecteurs propres, pour trois locuteurs du corpus que nous avons utilisé. Les contours extraits et projetés sont remplis par coloriage en magenta et en cyan respectivement. Les zones qui apparaissent en noir sont celles pour lesquelles il y a recouvrement (intersection). Cette illustration permet de remarquer que le modèle arrive à se déformer pour s'adapter à la forme spécifique des lèvres des différents locuteurs, mais que la qualité de l'approximation de la forme est très variable.

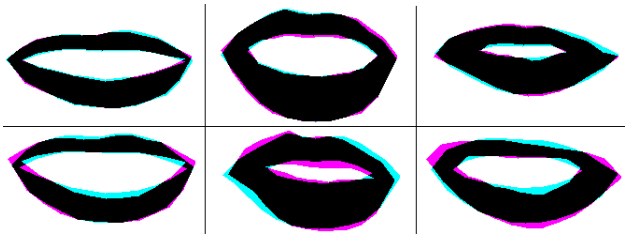

FIG. 1 – *Exemples de bonnes (en haut) et mauvaises (en bas) approximations de la forme des lèvres pour deux locuteurs et une locutrice.*

3.2 Évaluation du modèle de la forme

L'analyse en composantes principales donne une évaluation en terme de pourcentage qui peut se révéler

vecteurs	1	2	3	4	5	6
variance	78.4	86.8	90.0	92.1	94.0	95.5
pire	42.9	50.3	50.6	52.1	46.8	46.2
moyen	75.2	81.9	83.4	85.4	86.2	87.5
meilleur	93.1	93.1	93.2	95.6	95.4	95.9

TAB. 1 – *Pourcentages de recouvrement entre forme réelle et forme projetée en fonction du nombre de vecteurs propres.*

trompeuse, comme en témoigne la figure 1, c'est pourquoi nous avons introduit d'autres mesures pour évaluer la qualité de l'approximation. Nous avons rempli les polygones correspondant aux lèvres et mesuré le recouvrement entre les contours : (nb de pixels intersection) / (nb de pixels union). La table 1 montre l'évolution de la qualité de l'approximation faite en fonction du nombre de données utilisées. La première ligne indique, pour information, le pourcentage de la variance totale. Un point remarquable est la faiblesse des valeurs pour le pire des cas. Pour cet exemple, si l'on souhaite conserver, pour tous les contours, une forme proche de la forme extraite, il faudra utiliser une dizaine de valeurs.

Ne conserver que les quelques premiers vecteurs de déformation permet d'éliminer les déformations statistiquement les moins fréquentes (qui peuvent être des erreurs de localisation), cependant, pour pouvoir extraire correctement la forme des lèvres, il ne faut pas négliger des déformations valides des lèvres sous prétexte qu'elles sont statistiquement rares. Il faut également savoir que les déformations statistiquement les plus fréquentes ne sont pas nécessairement les plus discriminantes pour la reconnaissance de parole, comme cela a été montré par Kaucic [6]. Il est donc important de trouver un équilibre entre trop de vecteurs propres (qui donneraient un modèle inutilisable pour la recherche des lèvres sur une image) et la perte de déformations importantes pour la reconnaissance de parole audiovisuelle. Il serait également intéressant d'étudier dans quelle mesure l'information perdue correspond à une régularisation par suppression de valeurs aberrantes ou à une réelle perte d'information, mais ceci nécessite une évaluation visuelle individuelle de chacun des contours mal approximés que nous n'avons pas effectuée.

3.3 Utilisation du modèle de la forme

Nous avons utilisé ce modèle de la forme pour localiser le contour des lèvres sur des images "naturelles", il s'agit d'une recherche dans un espace à 8 dimensions : 4 pour la position du contour (x, y, z et rotation) et 4 pour sa forme (α, β, γ et δ). Nous avons utilisé l'information de teinte centrée sur le rouge comme attache aux données, mais les résultats ont été assez mauvais en raison de très nombreux minima locaux. Ce problème n'a pas disparu, même après des tentatives de filtrage de l'image en teinte rouge. Nous pensons que la mauvaise qualité des résultats est due à l'inadéquation de notre modèle d'apparence *a priori* (la teinte rouge) et avons décidé d'apprendre statistiquement l'apparence des lèvres "naturelles" à partir des données.

4 Modèle d'apparence

Nous avons choisi de construire un modèle de l'apparence de la bouche pour rendre plus robuste l'extraction du contour des lèvres. Notre intention est clairement de faciliter la localisation du contour (modèle de la forme) sur des images acquises dans des conditions "naturelles", et pas, dans cette phase du travail, l'étude de l'apparence de l'intérieur de la bouche (présence ou absence de la langue et des dents). Notre modèle est global et nous décomposons l'image en trois zones (comme dans [5]) : la peau, les lèvres et l'intérieur de la bouche. Nous n'avons pas d'a priori sur ce qui différencie les trois classes et avons entraîné un réseau de neurones à la tâche de classification de blocs d'image provenant des trois zones. Le nécessaire étiquetage des blocs (peau, lèvres ou intérieur de bouche) peut être effectué automatiquement en utilisant des images et leurs contours associés, mais nous ne disposons pas de ces contours. Une solution serait de les localiser manuellement, mais ceci est difficilement envisageable pour des corpus importants (comme cela a pu être remarqué dans [10]). Nous proposons plutôt d'utiliser une procédure automatique qui utilise l'information acoustique pour obtenir une estimation de la forme des lèvres avant de la rechercher sur l'image. Nous avons demandé à certain locuteurs de prononcer les mêmes phrases avec et sans maquillage sur les lèvres. En utilisant la programmation dynamique, nous avons aligné les vecteurs acoustiques correspondant aux deux séquences d'images (pour plus de détails sur cette partie du travail, voir [3]). L'idée est d'utiliser la forme des lèvres extraite des images avec maquillage pour la rechercher dans l'image sans maquillage correspondante. La localisation devient alors une recherche dans un espace à 4 dimensions (x, y, z, rotation). Nous avons effectué cette recherche, toujours en utilisant l'information de teinte, en utilisant plusieurs méthodes :

expl : exploration combinatoire de l'échelle (z), puis de l'angle de rotation et, pour chacune de ces valeurs, recherche du minimum de notre fonction d'énergie en la calculant en chaque point (x, y).

dsm : minimisation par "downhill simplex method", initialisée au centre de l'image.

best : nous avons sélectionné le meilleur des deux contours obtenus par les méthodes précédentes selon notre critère d'énergie.

Type	distance (pixels)			recouvrement (%)		
	-	moyen	+	-	moyen	+
expl	19.46	7.07	2.93	59.58	77.98	89.73
dsm	23.86	7.41	3.24	60.23	77.72	86.94
best	23.86	7.39	2.93	60.23	77.96	89.73

TAB. 2 – *Qualité des contours obtenus en utilisant différentes méthodes de minimisation.*

FIG. 2 – *Sortie du modèle d'apparence ref-10 obtenu à partir d'un étiquetage manuel des images (à gauche) et cert-10 grâce à un étiquetage automatique (droite).*

5 Résultats

L'évaluation de la qualité de la localisation des contours est un problème délicat. Il est possible d'effectuer une évaluation perceptive comme cela est proposé dans [8], mais cela nécessite une observation des contours par un opérateur qui peut être subjectif. Une autre manière de procéder consiste à comparer les contours obtenus avec ceux qu'un opérateur humain aurait placé directement sur les images. Le placement par l'humain sert alors de référence, bien qu'il puisse y avoir des variations dans ce placement. L'intérêt principal de ce second mode d'évaluation est qu'il permet de disposer de métriques pour l'évaluation et que l'humain n'est sollicité qu'une unique fois, même si l'on fait de nombreuses expérimentations. Nous avons mesuré la distance moyenne (en pixels) entre chaque point du contour de référence et les points correspondants dans le contour évalué et à nouveau le recouvrement entre les contours. Les résultats sont résumés dans la table 2. Ces résultats peuvent sembler faibles (recouvrement de 78% et distance de 7 pixels), ils sont cependant tout à fait satisfaisants visuellement. Il faut signaler qu'un des locuteurs a maquillé plus que ses lèvres. La forme extraite sur les images maquillées ne peut alors pas correspondre parfaitement avec la forme réelle et l'estimation de la forme à rechercher sur l'image "naturelle" correspondante est donc faussée. Enfin, il faut rapporter l'erreur de 7 pixels à la largeur du contour, soit environ 200 pixels.

6 Conclusions et perspectives

Nous avons obtenu une représentation de la forme des lèvres à la fois compacte et relativement fidèle. Ceci nous permet d'effectuer une recherche des lèvres sur une image en ayant un nombre restreint de paramètres à minimiser. Bien qu'il n'ait pas été créé dans cette optique, ce modèle de la forme peut également être utilisé pour transmettre des mouvements de lèvres avec un très faible débit. Plus on transmet de valeurs, meilleure sera l'approximation, mais même avec quelques paramètres, on obtient déjà une bonne approximation dans certain cas. Toutefois, nous n'avons fait aucune étude sur l'intelligibilité d'une parole visuelle ainsi approximée, ce qui serait nécessaire avant d'utiliser ce modèle dans un cadre applicatif.
En utilisant l'information acoustique et un maquillage, nous avons réussi à obtenir une localisation des lèvres proche de celle que nous aurions obtenu manuellement. Ceci nous a donné la possibilité de construire un modèle d'apparence très proche de celui que nous aurions obtenu en étiquetant manuellement les images comme en témoigne la figure 2 (ces résultats sont développés de façon plus approfondie dans [4])

Références

[1] L. Besacier. *Un Modèle Parallèle pour la Reconnaissance Automatique du Locuteur.* Ph.D. thesis, 1998.

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

[3] P. Daubias. Utilisation de l'information acoustique pour aligner deux séquences de parole audiovisuelle. In *Proc. RJC Parole*, pages 74–77, Belgique, 2001.

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

[5] R. Kaucic and A. Blake. Accurate, real-time, unadorned lip tracking. In *Proc. ICCV*, 1998.

[6] R. Kaucic, B. Dalton, and A. Blake. Real-time lip tracking for audio-visual speech recognition applications. In *ECCV* , pages 376–387, 1996.

[7] M. Liévin and F. Luthon. Unsupervised lip segmentation under natural conditions. In *Proc. ICASSP*, volume 6, pages 3065–3068, Phoenix, USA, March 1999.

[8] J. Luettin, N.A. Thacker, and S.W. Beet. Active shape models for visual speech feature extraction. In *Speechreading by Humans and Machines*, volume 150, pages 383–390. NATO ASI, 1996.

[9] U. Meier, R. Stiefelhagen, J. Yang, and A. Waibel. Towards unrestricted lip reading. In *Proc. ICMI*, 1999.

[10] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou. Audio-visual speech recognition. Tech. Report, ICSI, CLSP, Johns Hopkins University, Oct. 2000.

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

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