Définition et sélection d'attributs visuels pour la reconnaissance audio-visuelle de la parole

Jean-Philippe Thiran, Andrés Vallés, Thomas Drugman, et Mihai Gurban

Ecole Polytechnique Fédérale de Lausanne
Institut de Traitement des Signaux
1015 Lausanne, Suisse
JP.Thiran@epfl.ch, Andres.Valles@epfl.ch, Thomas.Drugman@epfl.ch, Mihai.Gurban@epfl.ch

Résumé La présence de bruit de fond et des conditions variables (environnement, réverbération, types de microphones) peuvent 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. Cependant, le nombre d'attributs visuels utilisés classiquement est élevé, pour des gains en performance assez limités. Nous présentons ici un aperçu général de nos travaux en définition et sélection d'attributs visuels pour la reconnaissance audio-visuelle de la parole. En particulier, nous proposons une méthode qui présente des gains de performance comparables à la méthode classique utilisant les attributs de DCT, mais avec beaucoup moins d'attributs visuels, issus des mouvements de la bouche du locuteur. L'avantage d'avoir un faible nombre d'attribut réside dans la possibilité d'obtenir de bonnes précisions en reconnaissance avec relativement peu de données d'entraînement, ainsi que dans des vitesses d'entraînement et de test accrues.

1 Introduction

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 [1] en 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.

La même intégration peut être effectuée par des ordinateurs pour améliorer les performances des systèmes de reconnaissance de la parole. La reconnaissance de la parole audio-visuelle (AVSR) peut en effet atteindre des taux de reconnaissance supérieurs à ce qui est possible avec le canal auditif seul. Le lecteur trouvera une vue d'ensemble des techniques de reconnaissance audio-visuelle de la parole dans [2].

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.

Nous proposons une méthode pour extraire les attributs visuels basés sur le mouvement des lèvres et nous montrons comment ces attributs (en nombre peu élevé) mènent à une reconnaissance de la parole aux performances comparables à celles obtenues par des méthodes de l'état de l'art, en particulier les attributs issus de la transformation en cosinus discret (DCT), employés dans beaucoup de systèmes d'AVSR.

Nous montrons ensuite comment l'ajout de la valeur d'un pixel bien choisi dans l'image améliore significativement les performances du système. Dans nos expériences, nous étudions et comparons deux types différents d'intégration multimodale, la fusion d'attributs et fusion de décision

Les attributs que nous proposons, de dimensionnalité réduite, sont constitués des différences entre les vecteurs de flux optique calculés sur différentes régions de la bouche du locuteur. Il faut noter que d'autres méthodes basées sur le flux optique ont été proposées dans la littérature, mais la dimensionnalité de l'espace des attributs était toujours très élevée. Par exemple Gray et collègues dans [3] construisent un vecteur d'attributs visuels de dimension 140, auquel ils ajoutent un vecteur de dimension 150 constitué des valeurs des pixels d'une imagette prise dans la région de la bouche. En comparaison, notre vecteur d'attribut a une dimension de 3. Il est composé du mouvement relatif vertical et horizontal de la bouche, et de la valeur de gris d'un pixel situé au centre de la bouche. L'avantage d'avoir un vecteur de basse dimensionnalité réside dans la possibilité de faire un entraînement du système de reconnaissance audio-visuelle de la parole nécessitant moins de données. De même, les performances en reconnaissances sont améliorées lorsque peu de données sont présentées au système, car les erreurs d'estimation dues à une haute dimensionnalité de l'espace des attributs sont évitées. De plus, l'entraînement et le test sont beaucoup plus rapides quand la dimensionnalité de l'espace des attributs est petite. Notre méthode inclut également une évaluation de la position de la ligne moyenne de la bouche, ce qui améliore l'exactitude des attributs de mouvement. Enfin, soulignons que comme nous utilisons une différence des vecteurs de mouvement, tout mouvement global de la tête est annulé, car il sera présent dans les deux composantes que nous soustrayons.

Notre article est organisé comme suit. Nous introduisons le contexte de la reconnaissance audio-visuelle de la parole dans la section 2. Au chapitre 3 nous décrivons les détails de notre méthode et la base de donnée de test que nous utilisons. La section 4 illustre nos résultats et nous les y discutons en comparaison avec ceux obtenus sur la même base de donnée en utilisant les attributs visuels bien connus obtenus par DCT. La section 5 conclut cet article et présente les directions de recherches futures pour ce travail.

2 La reconnaissance audio-visuelle de la parole

Dans cette section nous présentons brièvement la structure d'un système audio-visuel de reconnaissance de la parole. Même si tous ces systèmes partagent des traits communs, ils peuvent différer à trois égards principaux.

La première différence peut venir du "front end" visuel, c-à-d de la partie chargée de suivre la région de la bouche du locuteur et en extraire les attributs visuels. La deuxième est relative à la stratégie d'intégration audio-visuelle, à savoir la façon dont les attributs auditifs et visuels sont assemblés pour produire une décision quant au mot prononcé. Finalement, le type de système de reconnaissance de la parole peut varier, selon qu'il s'agisse par exemple de reconnaissance de mots isolés ou de parole continue. Notre système reconnaît des séquences de mots séparés par des silences, sur une base de donnée à vocabulaire restreint.

La grande majorité des systèmes de reconnaissance de la parole utilise des modèles de Markov cachés (HMM) [4] comme classificateur. Notre système repose également sur des HMM, et nous étudions deux types d'intégration multimodale.

2.1 Les attributs visuels

Une fois la région d'intérêt isolée, il s'agit d'en extraire l'information utile, en utilisant un nombre minimum d'attributs, car une dimensionnalité élevée pour l'espace des attributs en rend la modélisation statistique difficile. Il existe trois types principaux d'attributs utilisés en reconnaissance audio-visuelle de la parole [2] :

- Des attributs d'apparence, qui sont extraits directement des pixels de la région d'intérêt. Ils peuvent êtres très simples comme un sous-échantillonnage des valeurs des pixels, ou la différence entre les pixels de deux images successives (attributs Delta), ou plus compliqués, par exemple par l'emploi de transformations telles l'analyse en composantes principales (PCA), l'analyse discriminante linéaire (LDA) ou des transformées de l'image comme la transformée en cosinus discrets (DCT) ou les transformées en ondelettes (DWT). Des attributs de mouvement comme les vecteurs issus du flux optique peuvent aussi être inclus dans cette catégorie.
- Des attributs de forme, extraits des contours des lèvres du locuteur. Ceci inclut des attributs géométriques simples, tels la hauteur, la largeur ou la surface intérieure des lèvres. Des attributs plus complexes peuvent être déduits des contours des lèvres, par exemple par modèles de formes actifs (ACM).
- Des attributs conjoints d'apparence et de forme, résultant de la combinaison des deux types précédents, et encodant donc des informations différentes et complémentaires.

De manière générale, l'utilisation des attributs de forme exige un suivi très précis des lèvres, ce qui peut être difficile en pratique, par exemple si la résolution des images est basse. De plus, certaines informations sont perdues si on se contente d'informations sur la forme du contour des lèvres. En effet plusieurs articles rapportent que les méthodes basées sur la DCT surpassent les méthodes basées sur la forme des lèvres [5,6].

2.2 L'ntégration audio-visuelle

Considérant la classification par HMM, l'intégration des informations auditives et visuelles peut se faire de différentes manières [2]. La plus simple consiste à concaténer les vecteurs d'attributs [7] avant de les présenter au classificateur. Même si cette méthode présente des performances intéressantes, elle a le défaut de ne pas permettre d'intégrer l'information de fiabilité que l'on pourrait avoir sur chaque modalité, et qui pourrait varier selon les conditions de l'environnement audio-visuelle.

Dans la cas de la fusion de décision, des classificateurs différents sont entraînés pour les canaux audio et vidéo, et leurs sorties (log-vraisemblances) sont combinées linéairement par un poids approprié. On peut distinguer trois façons différentes de combiner les vraisemblances individuelles des deux modalités [2] :

- l'intégration précoce : où les vraisemblances sont combines au niveau des état, en forçant la synchronie des deux flux de données. Ceci conduit à des classificateurs HMM multi-flux [8]. En pratique ceci est réalisé par un HMM unique audio-visuel, mais où les probabilités d'émission sont estimées séparément pour les flux audio et vidéo.
- l'intégration tardive, qui demande deux HMM. La reconnaissance finale est obtenue en sélectionnant le mot selon la méthode des "n-meilleurs" parmi les sorties des HMM audio et vidéo.
- l'intégration intermédiaire, qui utilise des modèles qui forcent la synchronie aux limites des phonèmes ou des mots.

3 Notre méthode

3.1 Les données

Dans nos expériences, nous utilisons des données de la base de données CUAVE [9]. Les vidéos sont enregistrées à 30 images/s et l'audio à 44kHz. Pour extraire la région d'intérêt, la région de la bouche est détectée automatiquement, normalisée en taille et redressés, de sorte que toutes les bouches ont approximativement le même taille et orientation. Cependant, comme la détection de la bouche n'est pas parfaite, nous y ajoutons une étape de détection de la ligne centrale (voir ci-dessous).

3.2 Les attributs visuels

Dans ce travail, nous proposons des attributs visuels très simples : deux valeurs de mouvement de la bouche et, en option, la valeur de l'intensité d'un pixel supplémentaire, situé au centre de la bouche. Pour l'analyse du mouvement de la bouche, nous utilisons l'algorithme du flux optique de Lucas & Kanade [10], dont nous extrayons les mouvements relatifs moyens horizontaux et verticaux dans la région de la bouche, à savoir la différence entre le mouvement moyen de la partie supérieure et inférieure de la bouche, respectivement de la partie gauche et de la partie droite de la bouche. Ces deux valeurs, plus la valeur d'intensité du pixel central de la bouche, créent des attributs aussi informatifs que les attributs DCT de très haute dimension, comme nous le verrons dans la section consacrée aux résultats.

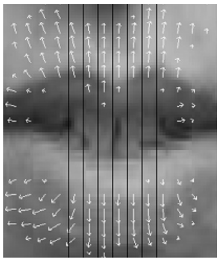

Fig. 1. Les colonnes considérées pour la détection de la ligne centrale de la bouche.

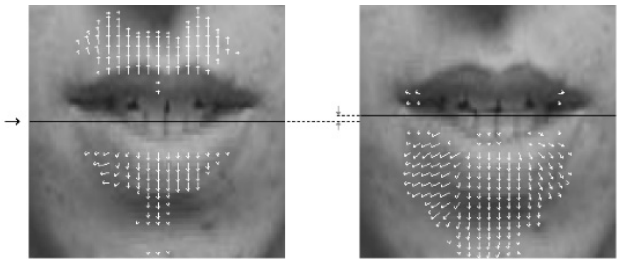

Fig. 2. Un exemple de suivi de la ligne centrale de la bouche.

Notre extraction d'attributs est extrêmement dépendante de la qualité de la localisation de la bouche, et en particulier de la détection du centre de la bouche. Ceci se fait par une étape de suivi de la bouche. Les images où les lèvres bougent dans des sens opposés sont sélectionnées et traitées comme indiqué dans la figure 1. Dans ces images, les flux optiques sur les parties supérieures et inférieures sont non nuls, et de sens opposés. Comme le montre la figure, chaque colonne dans une région centrale de l'image est analysée de la façon suivante. En partant du centre de l'image, le premier vecteur pointant vers le haut, situé dans la partie supérieure de la colonne, est détecté, et de même, en partant du centre, le premier vecteur pointant vers le bas, dans la partie inférieure de la colonne. Le point situé à mi-distance entre ces deux vecteurs est retenu. Ensuite, la moyenne de ces valeurs (une par colonne) est calculé. Une moyenne de cette dernière valeur sur les 15 dernières images donne finalement notre information de position verticale du centre de la bouche. La figure 2 montre le résultat de ce suivi de la bouche.

L'axe de symétrie vertical (gauche-droite) de la bouche quant à lui est simplement placé au centre de la région d'intérêt, la précision sur la position de cet axe étant moins importante que pour l'axe de symétrie horizontal.

Cette partie se termine par le calcul des attributs eux-mêmes. Nos attributs sont donc la différence entre le mouvement moyen vertical dans les zones supérieures et inférieures de la bouche, la différence entre le mouvement moyen horizontal dans les zones gauche et droite de la bouche, plus la valeur de l'intensité de l'image au centre de la bouche (à l'intersection des axes).

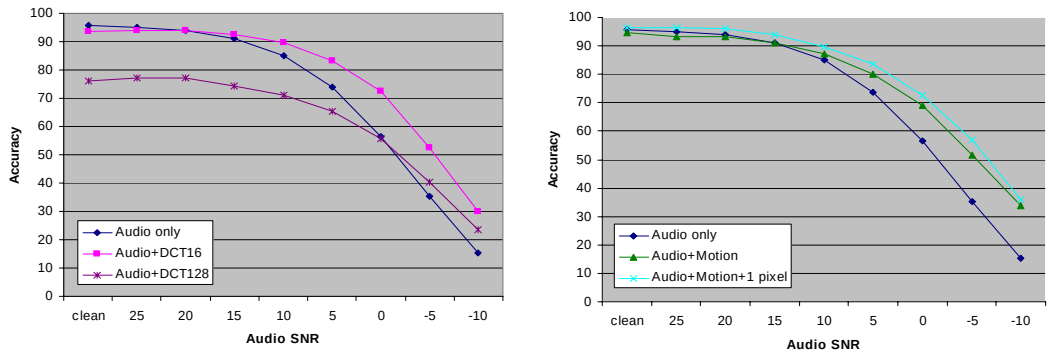

Fig. 3. Résultats avec fusion d'attributs.

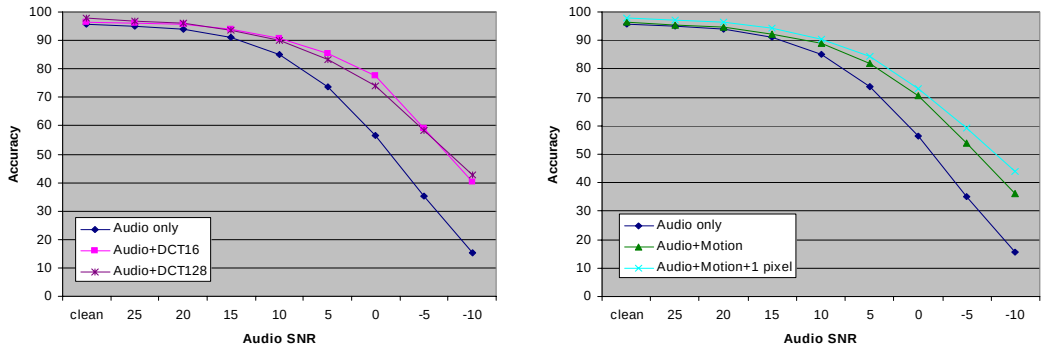

Fig. 4. Résultats avec fusion de décisions.

3.3 Notre système de reconnaissance de la parole

Nous utilisons la libraire HTK [11] pour construire notre système de reconnaissance vocale. Les modèles de mots ont 8 états par mot, et une gausienne par état. Nos attributs audio sont les coefficients mel-cepstraux classiques (MFCC), avec leurs valeurs Delta et Delta-Delta, formant en tout des vecteurs à 39 composantes.

Nous avons testé la fusion d'attributs, par simple concaténation des attributs audio et vidéo, ainsi que l'intégration précoce. Dans ce dernier cas, pour un état j la probabilité d'émission $b_j(o_t)$ pour une observation o_t est calculée comme suit [11] :

$$b_j(o_t) = \prod_{s=1}^{S} N(o_{st}; \mu_{js}, \Sigma_{js})^{\lambda_s} \tag{1}$$

Où $N(o; \mu, \Sigma)$ est une gausienne multivariée de moyenne μ et de matrice de covariance Σ. Le poids λ_s du flux s est choisi manuellement pour le moment.

Pour la combinaison des vraisemblances, la règle du produit est l'une des plus utilisées, mais d'autres existent, comme la règle de la somme, du maximum ou du minimum [12]. Ces règles sont comparées dans [13], dans le but de combiner les sorties de classificateurs

entraînés sur différents types d'attributs purement auditifs. Il y est montré que la règle du produit est la meilleure.

4 Résultats

Nous avons effectué deux types d'expériences sur la base de données CUAVE. Tout d'abord, nous avons utilisé l'intégration audio-visuelle la plus simple, la fusion d'attributs, pour obtenir les résultats présentés à la figure 3. Ensuite nous avons utilisé la fusion de décision, menant à des résultats meilleurs, comme on peut le voir sur la figure 4. Dans les deux cas, nous comparons nos attributs de mouvement aux attributs DCT, avec soit 16 soit 128 coefficients. La précision présentée est la précision de reconnaissance des mots.

Les expériences de fusion d'attributs montrent que, pour un espace d'attributs à haute dimensionnalité, l'entraînement est peu efficace, du fait du nombre relativement restreint de données d'entraînement. Ceci peut se voir par le fait que tous les tests avec 128 coefficients DCT ont des performances médiocres, alors que ceux avec 16 coefficients se comportent honorablement. De même, nos attributs de faible dimensionnalité surpassent les coefficients DCT de 2 à 3%

Quant à la fusion de décisions, les deux cas de coefficients DCT se comportent de façon similaire. Et à nouveau, nos attributs surpassent la DCT de quelques pourcents. De façon générale, la fusion de décisions surpasse la fusion d'attributs dans tous les cas.

Enfin, on peut voir que l'ajout de la valeur de l'intensité du pixel central de la bouche induit des performances améliorées. Ces 3 attributs ont des performances similaires à 128 coefficients DCT.

5 Conclusions

Nous avons présenté une méthode d'extraction d'attributs visuels qui crée des attributs de dimensionnalité réduite, sur la base du flux optique. La ligne centrale de la bouche est détectée et suivie, ce qui améliore significativement la qualité des résultats obtenus. Nos attributs surpassent ceux obtenus par DCT, couramment utilisés dans ce domaine, et ceci que ce soit en utilisant la fusion d'attributs ou la fusion de décisions.

Comme travail futur, nous aimerions améliorer la qualité de l'estimation du flux optique, ou le remplacer par une méthode de *block matching*. Nous aimerions aussi améliorer la méthode de fusion de décisions, en trouvant une façon automatique et dynamique de définir les poids.

Remerciements

Ce travail est soutenu par le Centre National Suisse de Compétences en Recherche IM2 sur le traitement de l'information multimodale (http ://www.im2.ch) ainsi que par le Réseau d'Excellence européen SIMILAR (http ://www.similar.cc).

Références

1. H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.
2. G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "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.
3. M. S. Gray, J. R. Movellan, and T. J. Sejnowski, "Dynamic features for visual speech-reading : A systematic comparison," in *Advances in Neural Information Processing Systems* (M. C. Mozer, M. I. Jordan, and T. Petsche, eds.), vol. 9, MIT Press, 1997.
4. L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77(2), 1989.
5. G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proceedings of the International Conference on Image Processing*, vol. 3, pp. 173–177, 1998.
6. R. Reilly and P. Scanlon, "Feature analysis for automatic speechreading," *Proc. Workshop on Multimedia Signal Processing*, pp. 625–630, 2001.
7. A. Adjoudani and C. Benoît, "On the integration of auditory and visual parameters in an HMM-based ASR," in *Speechreading by humans and machines* (D. G. Stork and M. E. Hennecke, eds.), pp. 461–471, Springer, 1996.
8. H. Bourlard and S. Dupont, "A new asr approach based on independent processing and recombination of partial frequency bands," *Proc. International Conference on Spoken Language Processing*, pp. 426–429, 1996.
9. E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Moving-talker, speaker-independent feature study and baseline results using the CUAVE multimodal speech corpus," *EURASIP JASP*, vol. 2002(11), pp. 1189–1201, 2002.
10. B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *Proc. International Joint Conference on Artificial Intelligence*, pp. 674–679, 1981.
11. S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge, Entropic Ltd., 1999.
12. J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.
13. K. Kirchhoff and J. Bilmes, "Dynamic classifier combination in hybrid speech recognition systems using utterance-level confidence values," *Proceedings ICASSP-99*, pp. 693–696, 1999.