Classification IN/OUT/OFF d'un intervenant dans un document audiovisuel

J. Philippeau J. Pinquier P. Joly

IRIT équipe Structuration Analyse MOdélisaion de la Vidéo et de l'Audio
Université Paul Sabatier
UMR 5505 CNRS - INPT, 118 Route de Narbonne, 31062 Toulouse Cedex 9
{philippe, pinquier, joly}@irit.fr

Résumé

Ce papier s'inscrit dans le cadre de l'indexation de documents audiovisuels. Il traite de la définition d'un nouveau descripteur : l'intervenant. Nos travaux ont porté sur la caractérisation de sa localisation, c'est-à-dire sa recherche dans une séquence audiovisuelle et sa classification en 3 catégories : IN, OUT ou OFF. A partir de l'étude de différents outils d'analyse des modes audio et vidéo, nous définissons un jeu de descripteurs qu'il est possible de renseigner automatiquement, potentiellement influents pour décider de la classe de la localisation de l'intervenant. Cette décision est effectuée à l'aide d'une modélisation des transitions d'une classe à une autre.

Mots clefs

intervenant, indexation audiovisuelle, descripteur multimodal, classification IN OUT OFF, modélisation multimédia.

1 Introduction

De nombreux travaux ont été entrepris dans le domaine de la caractérisation automatique de contenus audiovisuels grâce à des descripteurs à la fois audio et vidéo, mais les orientations choisies l'ont été dans le but d'améliorer par la vidéo les performances de systèmes se basant exclusivement sur l'audio (en Reconnaissance Automatique de la Parole [1] par exemple) ou inversement [2].

Nous avons mené notre étude dans le but de concevoir un nouveau descripteur pertinent pour caractériser un document audiovisuel en vue de son indexation. Considérons un intervenant, c'est-à-dire tout individu qui intervient par la parole et est localisable par celle-ci, dans une séquence audiovisuelle. Notre préoccupation est de savoir si, à un instant donné, sans connaissance *a priori* sur le type de document traité, un intervenant est visible ou s'il ne l'est pas. Jusqu'à présent, les travaux s'apparentant le plus à la classification d'intervenants considéraient le problème ainsi : un locuteur est de classe IN lorsqu'une personne est détectée à l'écran pendant la locution, sinon il est OUT. Toutefois, cette classification un peu arbitraire ne prenait pas en considération l'activité visible de parole à part entière : la personne détectée à l'écran n'est pas forcement celle qui parle.

Notre objectif est de préciser cette classification en considérant cet aspect visible de la locution et en définissant de nouvelles classes d'intervenants :

- la personne qui parle est visible, elle est de classe IN,
- la personne qui parle n'est pas visible, mais elle a déjà été filmée ou le sera durant son élocution, elle est de classe OUT.
- la personne qui parle n'est jamais visible pendant toute son intervention, elle est de classe OFF.

La pertinence du choix de l'intervenant comme contenu descriptif dans un document audiovisuel prend sens dès que l'on considère l'apport conjoint des modalités audio et vidéo.

Après avoir exhibé les descripteurs vidéo et audio que nous avons choisis pour caractériser un intervenant, nous présenterons les expériences menées sur ceux-ci et montrerons la manière dont nous les avons conjointement utilisés pour créer un descripteur audiovisuel à part entière. Nous ferons enfin une comparaison entre notre proposition et la classification traditionnelle IN/OUT.

2 Contexte applicatif

2.1 Choix du corpus

Dans un soucis de confort et de généricité, nous avons choisi d'étudier des séquences répertoriées pour la campagne d'évaluation TRECVID2004 [3]. Nous avons également porté notre attention sur une émission du jeu télévisuel français « Pyramide ». La résolution de ces vidéos (352*264 px à 29.97 fps) est jugée suffisante pour les traitements à opérer. De plus, la qualité relativement mauvaise des images due à l'encodage mpeg gage de la généricité de nos travaux sur la qualité des documents. Pour finir, la parole y est omniprésente et n'est pas interrompue, nous permettant une analyse mono-locuteur du signal de parole.

2.2 Segment audiovisuel

Nous définissons un segment audiovisuel comme une séquence pendant laquelle une classe d'intervenant reste stable. Un segment sera donc délimité par les frontières suivantes : un changement de locuteur, un changement de plan, une combinaison des deux ou un long silence. Afin de réaliser ces segmentations, nous nous sommes appuyés sur les travaux de [4] pour trouver les zones de parole, et sur ceux de [5] pour détecter les changements de plans. Les taux de reconnaissance (ou « accuracy ») listés dans ce papier ont été obtenus sur des segments extraits à la fois de TRECVID2004 et de Pyramide.

3 Point de vue vidéo

3.1 Détection du visage

De nombreux travaux ont été menés sur la détection automatique de visages (cités dans [6]) et s'appuient sur des techniques variées : sur des caractéristiques « bas-niveau » (comme la couleur, la forme ou la texture), sur la détection de caractéristiques faciales (comme les yeux, le nez ou la bouche), ou encore grâce à des approches statistiques. C'est un détecteur provenant de cette dernière catégorie que nous avons utilisé : le détecteur de visages de Violat et Jones [7]. L'analyse se fait donc image par image sur la totalité de la durée du segment vidéo considéré, pendant que de la parole est détectée.

Nous décidons qu'il y a effectivement présence d'un visage lorsqu'il a été détecté dans au moins 7 images sur une fenêtre temporelle de 11 images. Ces deux paramètres sont des valeurs optimales utilisées lors du processus de détection de personnes basé sur le détecteur de Viola et Jones et utilisé dans [6].

Pour savoir si un visage est le même d'une image à l'autre, nous avons construit une fenêtre de recherche autour de chaque visage détecté. Si 2 visages localisés dans la même fenêtre ont des dimensions suffisamment proches (plus ou moins 10%), nous considérons qu'il s'agit de la même personne.

Ce détecteur « oubliant » régulièrement un voir plusieurs visages sur la totalité du segment, nous avons du générer visuellement les visages manquants. Nous avons choisi de générer un visage V_0 non détecté par interpolation linéaire des coordonnées des deux visages temporellement les plus proches de V_0, à savoir V_1 (détecté avant V_0) et V_2 (détecté après V_0). Cette méthode nous a donné de bons résultats (non quantifiables mais visuellement corrects). En ce qui concerne les fausses détections, elles sont relativement rares et partiellement évincées grâce à l'algorithme du calcul du score d'activité explicité section 3.2.

La présence d'un visage apparaissant pendant tout un segment dans le champ de la caméra constitue notre premier descripteur fiable. Grâce à celui-ci, nous obtenons un taux de précision de détection des intervenants IN de 90,2%.

3.2 Analyse de l'activité labiale

Nous nous sommes ensuite penchés sur le problème de la localisation des lèvres pour pouvoir quantifier l'activité de celles-ci. Les divers travaux existant s'appuient sur des dispositifs intrusifs et/ou sur des images propres (bonne définition et fréquence élevée) dans des conditions de laboratoire (prises de vue frontale avec illumination constante) [8]. Ces méthodes étant impossibles à mettre en œuvre dans le cadre de nos travaux, nous nous sommes donc restreints à les localiser approximativement, c'est-à-dire dans le tiers bas du visage, entre les 2 et 4 cinquièmes de la largeur du visage (cf. Figure 1).
Outre le fait que cette localisation soit extrêmement facile et rapide à mettre en œuvre, elle permet de toujours cerner les lèvres, que le visage soit de face ou de profil.

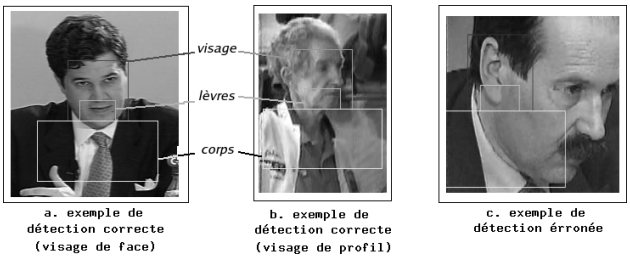

Figure 1 – *Exemple de résultats de détections du visage, du corps et des lèvres.*

Pour quantifier l'activité labiale, nous avons procédé par paires d'images pour ensuite obtenir un résultat global. Nous avons donc considéré deux images successives I1 et I2 contenant le visage d'une même personne. Après localisation des lèvres, représentées par les régions L(I1) et L(I2), nous avons construit une fenêtre autour de L(I1) et déplacé L(I2) dans cette zone de recherche. L'appariement ainsi que la valeur représentant la différence de pixels entre L(I1) et L(I2) ont tous deux été obtenus en minimisant l'Erreur Quadratique Moyenne (EQM), normalisée par la taille de L(I2), sur le canal de luminance de l'espace HLS. La moyenne des EQM ainsi calculées sur l'ensemble du segment vidéo considéré nous donne une valeur quantitative de l'activité labiale d'un personnage. Nous l'avons appelé Taux d'activité Labial (TAL).

Nous avons ensuite considéré que l'activité de parole concernait une zone plus large que celle des lèvres, car une personne bouge généralement corps et visage lorsqu'elle parle. Ainsi, pour le :

– Taux d'Activité du Visage (TAV), nous avons opéré l'appariement en privilégiant la ressemblance au niveau des lèvres, ceci en appliquant un masque de poids sur le visage, donnant une valeur deux fois plus élevée aux pixels de la région des lèvres que sur le reste du visage,

– Taux d'Activité du Corps (TAC), nous avons considéré un rectangle de largeur deux fois la taille du visage et de même hauteur que le visage, positionné sous celui-ci.

De ce fait, une hiérarchie entre les descripteurs s'organise d'elle même, plaçant dans l'ordre décroissant d'importance le TAL, le TAV, puis le TAC. Pour comparer l'activité labiale de deux personnages *i* et *j* filmés dans une même séquence ou dans deux séquences successives (afin de savoir lequel des deux est susceptible d'être le locuteur), nous avons défini un score d'activité basé sur une somme pondérée de ces trois taux.

Le taux de précision d'une décision basée sur ce score appliqué à l'identification du locuteur entre deux personnages au sein d'un même segment ou sur 2 segments consécutifs est de 95,7%.

4 Point de vue audio

4.1 Soustraction cepstrale

La soustraction cepstrale est couramment utilisée pour débruiter le signal de parole du bruit de la source d'enregistrement (micro, canal téléphonique...) [9]. C'est ce bruit du canal qui nous intéresse.

Nous décrivons ici le processus usuel d'analyse cepstrale ([10]) : Pour chaque trame de signal de parole, une préaccentuation des aigus et un fenêtrage de Hamming sont effectués. Les énergies sont calculées dans 24 filtres après application du module de la Transformée de Fourier. On répartit alors ces canaux selon l'échelle Mel pour tenir compte de la perception humaine. L'analyse de chaque trame donne un vecteur d'observations de 26 paramètres, comprenant l'énergie du signal et sa dérivée, ainsi que 12 coefficients cepstraux (ou MFCC) et les dérivées respectives.

Pour assurer une relative indépendance vis-à-vis du canal de transmission, on soustrait habituellement à chaque coefficient cepstral la moyenne des 12 MFCC. Nous avons décidé d'exploiter les informations contenues dans l'évolution des MFCC entre classes vocales au vu des résultats obtenus durant l'étude de la soustraction cepstrale.

4.2 Réflexions sur le comportement des descripteurs

Il est tout d'abord nécessaire d'énumérer les différentes configurations possibles de transitions entre classes et le comportement attendu des descripteurs dans chaque cas, comme l'illustre la figure 2 :

1. *Les descripteurs doivent caractériser une stabilité de l'environnement audio dans le cas d'une transition due à un changement de plan si le même intervenant continue de parler* (**groupe A**). Le cas particulier de la transition entre 2 intervenants en voix IN sur un changement de locuteur sans changement de plan est aussi à prendre en compte dans ce groupe ci.
2. *Ils doivent aussi témoigner d'un changement d'environnement audio dans le cas de transition entre locuteurs sensés évoluer dans des cadres d'enregistrements acoustiques différents* (**groupe B**).
3. Il est à noter que *des transitions particulières ne se produisent jamais* (**groupe C**). Il s'agit d'un passage de voix OFF à voix IN ou OUT (et réciproquement) sans changement de locuteur. Cela insinuerait que l'intervenant en voix OFF a été ou sera dans le champ de la camera, ce qui contredit sa définition.
4. En ce qui concerne le reste des cas envisageables (**groupe D**), *nous ne pouvons pas tirer d'informations en considérant uniquement ces descripteurs*. En effet, une transition entre deux locuteurs OFF, par exemple, peut se passer dans le même environnement sonore (deux commentateurs sportifs par exemple) ou non (une voix OFF introduit un reporter sans liaison vidéo avec celui-ci).

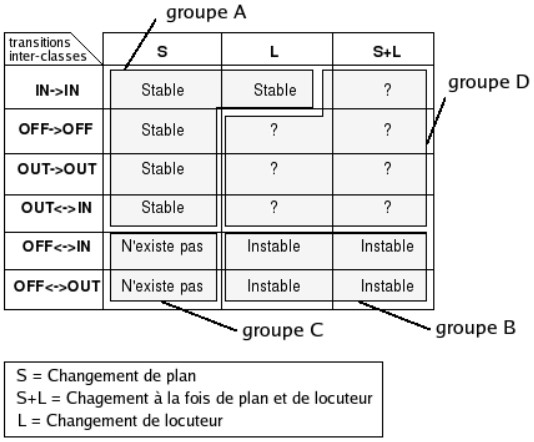

Figure 2 – *Comportement souhaité des MFCC utilisés pour caractériser les transitions entre classes d'intervenants.*

4.3 Expériences et résultats

Après avoir vainement tenté de caractériser dans quel environnement sonore évoluait le locuteur, nous nous sommes penchés sur la caractérisation des changements d'environnements entre deux segments adjacents s^1 et s^2.

Soit un segment s^k d'une seconde échantillonné à 16kHz, l'analyse cepstrale étant calculée sur des fenêtres de 256 points avec un recouvrement sur la moitié. Nous obtenons 125 vecteurs $y_i = (y_{i,1} \cdots y_{i,12})$ de dimension 12 (autant que de MFCC), avec $i \in \{0, \cdots, 125\}$ l'indice de vecteur. Si nous effectuons l'analyse sur les 2 dernières secondes du segment s^1 et sur les 2 premières du segment s^2, cela nous donne respectivement deux collections de vecteurs $(y_1 \dots y_{250})$ et $(y_{251} \dots y_{500})$ (cf. Figure 3).

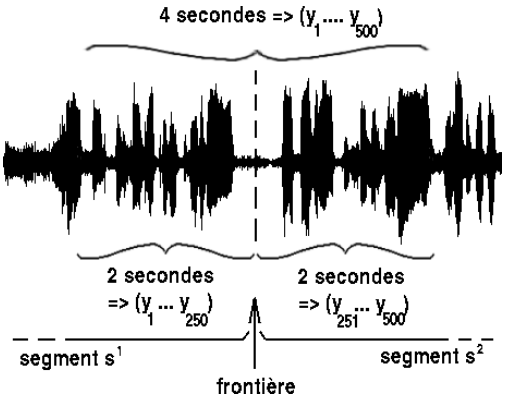

Figure 3 – *Récupération de la collection de vecteurs sur s^1 et s^2.*

Si nous voulons caractériser un changement de comportement des MFCC à la frontière des segments s^1 et s^2, nous allons supposer que les :
- $(y_1 \dots y_{250})$ suivent une loi Normale $N(M^1, \Sigma^1)$,
- $(y_{251} \dots y_{500})$ suivent une loi Normale $N(M^2, \Sigma^2)$,
- $(y_1 \dots y_{500})$ suivent une loi Normale $N(M^3, \Sigma^3)$.

Si nous considérons que toutes les composantes des vecteurs sont indépendantes [11], il est possible de faire les deux hypothèses suivantes :

- hypothèse (h_1) : *il y a un changement d'environnement sonore entre s^1 et s^2*. Ceci se traduit par :

$$P(y_1 \dots y_{500}/h_1) = P(y_{1\dots}y_{250}/N(M^1, \Sigma^1)) \cdot P(y_{251\dots}y_{500}/N(M^2, \Sigma^2)) \quad (1)$$

- hypothèse (h_2) : *l'environnement sonore de s^1 est le même que celui de s^2*. Ceci s'exprime ainsi :

$$P(y_1 \dots y_{500}/h_2) = \prod_{i=1}^{500} P(y_i/N(M^3, \Sigma^3)) \quad (2)$$

Le test d'hypothèses est basé sur le rapport de vraisemblance :

$$\Delta(s^1, s^2) = \frac{P(y_1 \cdots y_{500}/h_2)}{P(y_1 \cdots y_{500}/h_1)} \quad (3)$$

En fixant un seuil θ à la forme logarithmique de ce test, il est possible alors de prendre une décision en faveur d'une des deux hypothèses (h1) ou (h2).

Nous avons remarqué qu'un seuil expérimental de $-68.5 * 10^{-3}$ fonctionne pour 92.8% des cas étudiés.

5 Mise en œuvre conjointe des descripteurs

5.1 Présentation des résultats

Nous avons décidé d'utiliser les descripteurs audio et vidéo suivants :

- **Presence**$_t \in \{yes, no\}$: présence ou absence de personnage durant le segment t (section 3.1).
- $\Phi_{t,t+1}$: score d'activité comparé entre le deux personnages ayant le TAL le plus élevé, présents dans deux segments consécutifs t et t+1 (section 3.2).
- $\Delta_{t,t+1} \in \{yes, no\}$: stabilité ou instabilité de l'environnement acoustique au passage du segment t au segment t+1 (section 4.3).
- **Transition** $\in$ {S,L,S+L} : frontières audio et/ou vidéo. S pour une détection de changement de plan, L pour une détection de changement de locuteur et S+L pour une combinaison de ces deux détections (Figure 2, section 4.3).

Nous avons choisi de créer un automate à 4 états : **IN**, **OUT**, **OFF**, ainsi qu'un état de **doute** qui nous sert d'état initial ainsi que d'échappatoire lorsque les informations dont nous disposons ne sont pas suffisantes pour attribuer une classe à un intervenant (Figure 4). Nous avons considéré qu'un état restait stationnaire sur toute la durée du segment analysé, et défini les transitions de l'automate comme les possibilités à explorer à chaque nouvelle prise de décision, c'est-à-dire comme les conditions acoustiques et visuelles inhérentes à nos descripteur pour chaque nouvel état.

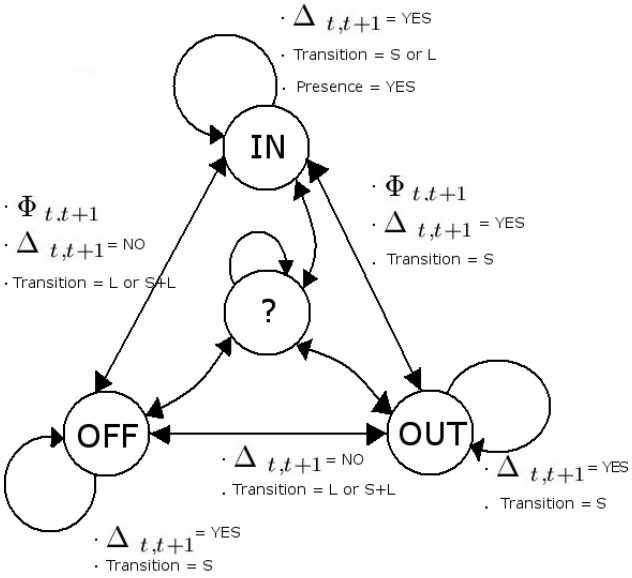

Figure 4 – *Présentation de l'automate à 4 états : IN, OUT, OFF et ? (doute).*

Par exemple, une transition de l'état IN à l'état OUT doit être considérée :
- si personne n'est détecté au segment $t + 1$ ou si le personnage détecté en $t + 1$ a un score d'activité plus faible que celui détecté en t,
- si l'environnement acoustique reste stable,
- et si un changement de plan a été détecté.

Comme il n'existe aucune vérité terrain prenant en compte une classification IN/OUT/OFF, nous avons développé notre propre corpus d'évaluation d'une durée totale de 21 minutes. Voici une présentation des résultats obtenus par notre automate :

- si nous considérons **DOUTE** comme une classification correcte, nous obtenons un taux de précision de 87,1%,
- si nous considérons **DOUTE** comme une erreur de classification, nous obtenons un taux de précision de 55,8%,
- si nous faisons abstraction du doute, c'est-à-dire si nous considérons uniquement les segments qui ne sont pas classés **DOUTE**, nous obtenons un taux de précision de 82,6%.
- l'automate rentre en état **DOUTE** dans 24,2% des cas.

Afin de pouvoir apprécier les performances de ce classifieur, nous allons comparer les résultats obtenus avec le type de classification qui existait avant notre étude : *un locuteur est de classe IN lorsqu'une personne est détectée à l'écran pendant la locution, sinon il est OUT*. Pour pouvoir faire la comparaison nous appellerons cette classification **old**. Nous nommerons la notre **new** et nous fusionnerons nos classes OUT et OFF en une classe **OUT** unique (les erreurs entre OUT et OFF ne sont plus considérées).

Le tableau 1 donne le pourcentage de segments correctement classifiés (de manière automatique) si nous ne tenons pas compte des segments classés **DOUTE** pour l'évaluation. Le tableau 2 présente les résultats si nous les prenons en compte. Dans les deux cas, une amélioration notable est obtenue grâce à l'utilisation de notre classifieur.

Tableau 1 – *Résultats obtenus en ne tenant pas compte du doute.*

	Doute = Void
old	55%
new	88.4%

Tableau 2 – *Résultats obtenus en tenant compte du doute.*

	Doute = False	Doute = True
old	48%	48%
new	70.5%	91.4%

5.2 Explications

La figure 5 présente une comparaison entre 3 séquences audiovisuelles extraites de notre corpus et les résultats obtenus par notre classifieur :

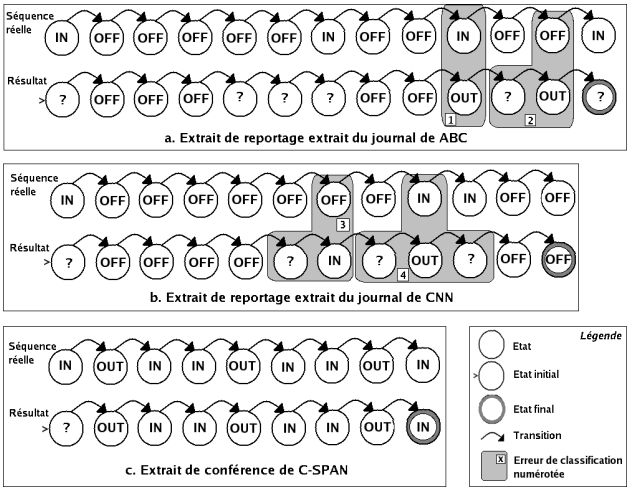

Figure 5 – *Schémas illustrant la comparaison entre des séquences audiovisuelles et les résultats obtenus par notre classifieur.*

– **analyse de la figure 5c :**

Ces valeurs proviennent d'un extrait (5 minutes 12 secondes) d'une conférence. Les segmentations sont des changements de plans : la caméra passe du protagoniste à des personnes du public.
Nous obtenons dans ce cas d'excellents résultats car nous nous trouvons dans une configuration claire, sans ambiguïté sonore ou visuelle :
- l'environnement sonore reste stable car le son émane du même micro par une même personne durant toute sa locution,
- le personnage n'est jamais interrompu et le flux de parole est constant,
- le protagoniste ainsi que les spectateurs sont filmés en plans américains, sont assis et ne bougent pas trop.

– **analyse des figures 5a et 5b :**

Ces valeurs sont issues d'extraits (1 minute 37 secondes pour la 5a et 1 minute 5 secondes pour la 5b) de reportages. La voix OFF d'un reporter couvre le document, qui est une succession de changements de plans, cédant ponctuellement la parole à des intervenants filmés qui sont interviewés au micro.

Les erreurs de classification sont dues :

- à la non-détection visuelle du locuteur, à cause de la trop faible qualité vidéo du document : un intervenant IN est pris pour un intervenant OUT (**erreur 1** figure 5a),

- à la prise de décision aléatoire de sortie de l'état **DOUTE** lorsque plusieurs possibilités sont offertes (**erreur 2** figure 5a). Une manière très simple de remédier à ce problème est d'attribuer des poids aux transitions suivant

leur fréquence d'apparition dans tel ou tel type de document (et rajouter ainsi de l'information *à priori*),

- à une mauvaise interprétation de la stabilité de l'environnement sonore : la voix OFF post-synchronisée du reporter enregistrée en régie, se fond à certains moments avec l'ambiance sonore propre à l'environnement visuel filmé. Le canal de la voix OFF est alors perturbé par un bruit acoustique différent de celui propre à l'enregistrement (**erreur 3** figure 5b),

- au manque de mémoire de l'automate (**erreur 4** figure 5b) : le prototype que nous avons construit ne traite que 2 états à la fois, ne gardant pas de trace du chemin déjà effectué ou à venir. Une manière de résorber ce dernier problème serait d'opérer une classification en deux passes.

6 CONCLUSION

Nous avons présenté des descripteurs vidéos qui nous ont permis de d'étudier l'activité de parole des intervenants d'un segment sur l'autre, de déterminer lequel des personnages parle au sein d'un même segment, et enfin de pallier une partie des déficiences du détecteur de visages de Viola et Jones lors de son utilisation pour du suivi de visage. Nous pensons que la fiabilité de ces descripteurs peut devenir meilleure si des mouvements de caméra, tels que le zoom ou le travelling, sont pris en compte.

Nous avons également montré que la variation des MFCC aux frontières des transitions entre classes sonores constitue un descripteur fiable lorsqu'il s'agit de caractériser un changement ou une stabilité entre deux environnements sonores. Malgré les mauvais résultats que nous avons eu lors de nos premières analyses concernant l'évolution des MFCC, analyses faites dans le but de caractériser un environnement acoustique plutôt que des changements entre eux, nous pensons toujours qu'il y a suffisamment d'information dans les coefficients cepstraux pour exhiber ce genre d'information.

Enfin, la mise en commun de ces informations au sein d'un automate nous a permis de créer un descripteur audiovisuel pertinent pour obtenir une classification IN, OUT ou OFF inédite d'un intervenant dans un document. Nous pouvons envisager, afin de perfectionner ce classifieur, de *rajouter de la mémoire à l'automate* en lui permettant de pratiquer une classification en deux passes. Nous pouvons également *adapter le classifieur au contexte* en attribuant des poids à chaque transition entre états selon leur fréquence d'apparition, suivant le type de document traité. De plus, *pondérer l'état* ***DOUTE*** de manière plus ou moins forte permettrait de pouvoir jouer entre précision (en augmentant son poids) et fiabilité (en diminuant son poids).

Nous espérons que ce type de classification pourra permettre, de part la nature bimodale de ces classes, de simplifier l'identification du locuteur en liant les medias audio et vidéo au sein même de la modélisation du problème.

Références

[1] G. POTAMIANOS, C. NETI, J. LUETTIN, et I. MATTHEWS. Audio-visual automatic speech recognition : An overview. Dans G. BAILLY, E. VATIKIOTIS-BATESON, et P. PERRIER, éditeurs, *Issues in Visual and Audio-Visual Speech Processing*. MIT Press, 2004.

[2] E. KIJAK. *Structuration multimodale des videos de sports par modeles stochastiques*. Thèse de doctorat, Universite de Rennes 1, Décembre 2003.

[3] Wessel Kraaij, Alan Smeaton, Paul Over, et Joaquim Arlandis. Trecvid 2004 - an introduction. Dans *Proceedings of the TRECVID 2004 Workshop*, pages 1–13, Gaithersburg, Maryland, USA, Novembre 2004.

[4] Julien Pinquier, Jean-Luc Rouas, et Régine André-Obrecht. Fusion de paramètres pour une classification automatique parole/musique robuste . Dans *Technique et science informatiques (TSI) : Fusion numérique/symbolique* , volume 22, pages 831–852. Hermès, 8, quai du marche neuf, F-75004 Paris, 2003.

[5] G. JAFFRE, P. JOLY, et S. HAIDAR. The SAMOVA Shot Boundary Detection for TRECVID Evaluation 2004. Dans *TRECVID 2004 Workshop, Gaithersburg, Maryland USA*, pages 179–183. NIST, 15-16 novembre 2004.

[6] G. JAFFRE et P. JOLY. Costume : A new feature for automatic video content indexing. Dans *RIAO 2004*, pages 314–325, Avignon, France, april 2004.

[7] P. VIOLA et M. JONES. Rapid object detection using a boosted cascade of simple features. Dans *IEEE CVPR*, 2001.

[8] G. POTAMIANOS, H.P. GRAF, et E. COSATTO. An image transform approch for hmm based automatic lipreading. Dans *Proceedings of the Internationnal Conference on Image Processing*, volume 3, pages 173–177, Chicago, 1998.

[9] C. MOKBEL, D. JOUVET, et MONNE J.. Blind equalization using adaptitive filtering for improving speech recognition over telephone. Dans *European Conference on Speech Communication and Technology*, pages 817–820, Madrid, Spain, 1995.

[10] Calliope. *La parole et son traitement automatique*. Masson, Paris, France, 1989.

[11] QIAN-JIE F. TIANHAO, L. Analyze perceptual adaptation to spectrally-shifted vowels with gmm technique. Dans *10th Annual Fred S. Grodins Graduate Research Symposium*, pages 120–121. USC School of Engineering, 04 2006.