Étiquetage Automatique de Programmes de Télévision

Xavier Naturel[1] Guillaume Gravier[1] Patrick Gros [1]

[1] IRISA - INRIA Rennes

Campus de Beaulieu
Rennes - France

{prenom.nom}@irisa.fr

Mots clés : Macro-segmentation, étiquetage automatique, télévision, EPG.
Concours jeune chercheur : oui

Étiquetage Automatique de Programmes de Télévision

Résumé

L'indexation automatique de grande quantité de flux de télévision est un sujet peu abordé par la communauté de l'indexation vidéo. L'objectif est la détection des ruptures sémantiques telles les changements de programme ou les débuts de plage de publicité. Nous montrons que des techniques basées sur des attributs très simples suffisent à créer une macro-segmentation pertinente qui est utilisée pour étiqueter automatiquement les programmes à partir du guide des programmes. Toutefois, nous montrons aussi les limites d'une telle approche et proposons de robustifier la méthode en intégrant des informations de reconnaissance de plans vidéo.

Mots clés

Macro-segmentation, étiquetage automatique, télévision, EPG.

1 Introduction

L'indexation de flux importants de télévision pose de nombreux problèmes pratiques rencontrés par les centres d'archivage, tel l'INA[1], ou les organismes de contrôle de la diffusion tel le CSA[2]. Dans le cas de l'archivage, un problème est par exemple de pouvoir extraire une collection d'anciennes émissions hebdomadaires. Il n'existe pas de moyen d'accéder à ces émissions autrement que par une fastidieuse recherche dans le flux à partir de l'horaire de diffusion supposé. Un autre exemple est de vérifier automatiquement qu'une chaine ne dépasse pas ses quotas de diffusion de publicité, ou de mesurer l'importance croissante du parrainage d'émissions. Les solutions existantes sont avant tout manuelles ou semi-automatique. De plus, la multiplication des chaines et l'apparition d'appareils grand public d'enregistrement de la télévision numérique poussent aussi à trouver des solutions de structuration automatique des flux télévisés.

Dans ce contexte, nous proposons une méthode d'étiquetage automatique de flux télévisés, basée sur des a priori de production et sur le guide des programmes. Un rapide état de l'art est tout d'abord proposé dans la section 2 sur les techniques proches de notre travail qui s'attachent essentiellement à la détection des publicités. Une approche de détection des séparations entre programmes est ensuite présentée dans la section 3. Cette méthode est basée sur la détection de segments de silence et d'images monochromes. Les détections sont tout d'abord présentées pour chaque modalité, puis combinées. Nous discutons ensuite dans la section 4 de méthodes d'alignement entre la macro-segmentation trouvée et le guide des programmes, ainsi que le moyen d'y intégrer des contraintes.

[1] Institut National de l'Audiovisuel

[2] Conseil Supérieur de l'Audiovisuel

2 Travaux Antérieurs

À notre connaissance, il n'existe pas de travaux cherchant à détecter le début des programmes dans un flux télévisé. Par contre, de nombreux travaux se sont intéressés à la détection des plages de publicité dans un flux vidéo de télévision, avec des techniques très proches des nôtres. Une approche classique et efficace est la détection des images noires qui marquent la séparation entre deux publicités [1, 2]. Cette détection est très simple à réaliser mais produit de nombreuses fausses alarmes, et est pour cette raison combinée avec d'autres attributs comme le silence [2], la fréquence du nombre de coupures, le niveau d'action [1]. La séparation des publicités par des images noires n'est cependant pas une constante universelle. Les chaines japonaises et américaines n'utilisent pas cette technique, et en France, les images de séparations sont blanches, bleues ou noires. D'autres approches utilisent l'absence de logo de chaine pendant les publicités [3], ce qui n'est pas non plus systématique, en particulier pour les chaines françaises. Une approche plus générique est utilisée par [4] qui détecte les publicités en tant que séquences répétitives. L'approche utilise malheureusement une classification basée sur des attributs couleur et audio spécifique au corpus utilisé, qui est limité aux journaux télévisés. Les résultats sont moins probants que les approches utilisant la détection d'images noires.

Des travaux de détection de publicités sur des chaines françaises ne semblent pas exister.

3 Macro-segmentation multimodale

Pour détecter le début et la fin des programmes, nous utilisons une constante de production, qui est l'insertion d'images monochromes lors d'un changement sémantique important (fin d'un programme, début d'une bande-annonce, séparation entre deux publicités, etc…). Ces images sont aussi accompagnées d'une forte chute du niveau sonore, ce qui produit des segments facilement identifiables. Ces segments sont appelés des **séparations**. Nous présentons tout d'abord une méthode pour détecter ces séparations à partir d'une seule modalité, puis en combinant deux détecteurs.

	Precision	Rappel
détection	0.82	0.9
localisation	0.84	0.79

Tableau 1 – *Détection des séparations - audio seul*

	Precision	Rappel
détection	0.41	0.89
localisation	0.96	0.79

Tableau 2 – *Détection des séparations - image seule*

3.1 Détection de segments de silence

La figure 1 montre l'énergie du signal audio sur une vidéo d'une durée de trois heures. Les séparations ont une très faible énergie et sont clairement visibles, ainsi que les plages de publicités, caractérisées par de nombreuses séparations très rapprochées.

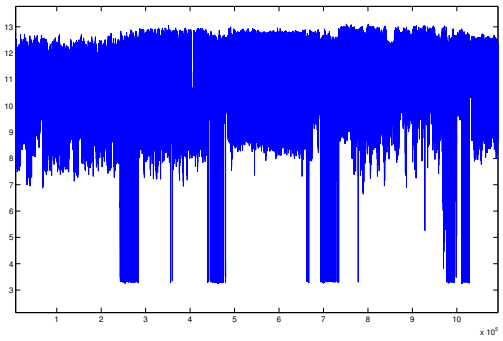

Figure 1 – *énergie du signal audio sur 3 heures de télévision.*

La méthode de détection de silence consiste à modéliser l'énergie du signal par un modèle bi-gaussien. La gaussienne de plus petite moyenne représente le silence tandis que la deuxième correspond à de l'activité audio. L'étiquetage d'un segment en tant que silence est réalisée en seuillant la moyenne de l'énergie du segment. Le seuil est fixé par $seuil = m_h - \alpha\sigma_h$, avec (m_h, σ_h) les paramètres de la gaussienne modélisant les hautes énergies, et α le paramètre règlant l'éloignement entre le seuil et la moyenne m_h.

Les résultats sont présentés dans le tableau 1 sur une séquence de télévision française d'une durée de 5h30. La ligne *détection* indique la précision et la rappel de la présence de la séparation, tandis que la ligne *localisation* indique la précision et le rappel du nombre d'images détectées comme appartenant à une séparation. Ils montrent que l'énergie du signal audio est un excellent indicateur de la présence d'une séparation.

3.2 Détection d'ensembles d'images monochromes

La deuxième piste est d'utiliser le fait que les segments de séparations sont constitués d'images monochromes. Malgré son apparente simplicité ce problème est assez délicat, pour de multiples raisons :

- la couleur des images insérées dépend de la chaine (noires, bleues, blanches), une même chaine pouvant utiliser plusieurs couleurs d'images suivant le type de coupure qu'elle souhaite effectuer.
- ces images sont très bruitées
- il y a des très nombreuses fausses alarmes (film très sombre, changement de scène dans les téléfilms...)
- un cadre peut entourer l'image : elle est alors bicolore.

Des méthodes simples pour détecter des images noires ont été proposées. Lienhart et al. [1] seuillent la variance de l'image dans le domaine pixelique, Sadlier et al. [2] seuillent la moyenne dans le domaine DCT. Ces méthodes ne sont pas assez robustes ou adaptées à notre problématique.

Nous proposons ici d'utiliser l'entropie de l'histogramme de l'image. Pour un histogramme h quantifié sur N niveaux, son entropie H est donnée par :

$$H = -\sum_{i=1}^{N} p_i \log p_i \text{ avec } p_i = \frac{h(i)}{\sum_k h(k)}$$

En pratique, l'histogramme est quantifié sur $N = 48$ niveaux afin de réduire l'influence du bruit.

H mesure la quantité d'information moyenne contenue dans l'histogramme. On s'affranchit de cette manière du problème de la couleur de l'image, et on conserve plus d'information qu'en calculant le maximum de l'histogramme. L'entropie est aussi moins sensible à la quantification de l'histogramme que le maximum, et donne de légèrement meilleurs résultats.

Toutefois l'approche image est sujette à de nombreuses fausses alarmes, car de nombreuses images monochromes ne font pas partie d'une séparation. Le Tableau 2 montre les résultats. On peut toutefois remarquer que la précision de la localisation est très bonne.

3.3 Méthode de fusion des segments audio et image

Il semble raisonable de penser que l'utilisation de plusieurs média peut améliorer les résultats de détection. L'intégration de différentes sources d'information pose cependant un problème de décision lorsque les résultats sont contradictoires. De plus, l'audio et la vidéo ont des fréquences d'échantillonnage différentes, il y a donc aussi un problème de synchronisation.

Kijak et al. [5] distinguent trois méthodes de fusion des informations multi-modales : l'analyse successive, l'intégration précoce, et l'intégration tardive. L'intégration précoce consiste à intégrer les attributs audio et vidéo au sein d'un même vecteur avant la classsification [5]. L'intégration tar-

Modalité	Détection		Localisation	
	Precision	Rappel	Precision	Rappel
Audio	0.82	0.9	0.84	0.79
Image	0.41	0.89	0.96	0.79
Fusion	1	0.9	0.94	0.74

Tableau 3 – *Détection des séparations - son et image*

dive consiste à classifier indépendemment les portions du flux pour chaque modalité, puis en combinant les résultats de ce ces classifications. L'approche par analyse successive consiste à analyser tout d'abord le flux avec une seule modalité, puis d'analyser les portions du flux détectées par la seconde modalité. Cette approche est la plus intéressante lorsqu'une modalité est clairement supérieure à une autre. C'est le cas ici, où le détecteur de silence exhibe de très bonnes performances.

La méthode consiste donc à tout d'abord repérer les segments de silence à l'aide de la technique exposée en 3.1, ce qui nous donne un intervalle de silence $I_1 = \{i_0 \dots i_n\}$. On utilise ensuite l'information image. Une décision basée sur un seuillage de la moyenne de l'entropie dans l'intervalle I_1 ne donnerait pas de très bon résultats à cause d'une relative mauvaise précision de la localisation audio. La localisation des séparations par l'attribut image possède au contraire une excellente précision que nous allons utiliser pour détecter précisement les limites de la séparation dans un intervalle élargi $I_2 = \{i_0 - \gamma, \dots, i_n + \gamma\}$

L'agorithme est le suivant. On parcourt l'intervalle I_2. Un début de segment est détecté dès que $H(i) \leq \alpha$. Le segment se termine lorsque $i = i_n + \gamma$ ou que le nombre maximal d'exceptions est atteint. On autorise en effet k exceptions $\{j_1 \dots j_k\}$ telles que $H(j_p) > \alpha, \forall p \in \{1 \dots k\}$.

Les résultats de la fusion des modalités est donné dans le tableau 3. Le principe même de la méthode fait que le rappel de la fusion ne peut excéder celui de l'audio. Par contre la précision est très nettement améliorée puisqu'il n'y a pas de fausses alarmes, et ceci sur 5h30 de télévision. Des fausses alarmes ont toutefois étés constatées sur des segments où il n'existe pas de vérité terrain. Ces fausses alarmes sont dues aux changements de scènes dans certains téléfilms. Il est néanmoins important de souligner que l'étiquetage manuel utilisé pour calculer les scores indique les changements sémantiques et non les séparations elles-mêmes. C'est le principal défaut de la méthode : les changements de programmes ne sont pas toujours signalés par une séparation, d'où des scores de rappel toujours inférieurs à 1 quelque soit la qualité de la détection des séparations. La section suivante montre que les résultats sont cependant suffisant pour créer une macro-segmentation pertinente.

De façon plus marginale, la vérité terrain indique des intervalles assez larges pour les changements, d'où les mauvais scores de rappel pour la localisation.

4 Étiquetage automatique

4.1 Problématique et première solution

La macro-segmentation est une première étape importante pour une navigation aisée dans les flux de télévision car elle permet un découpage du flux en segments pertinents pour un utilisateur. Toutefois ces segments ne sont pas étiquetés et ne peuvent donc pas être recherchés facilement, à moins d'être manuellement étiquetés. L'information concernant les programmes est pourtant disponible dans le guide des programmes, soit dans sa version papier classique, soit sa version électronique transmise avec le flux vidéo, l'EPG (Electronic Program Guide). Il est donc possible en TV numérique de récupérer automatiquement ces informations, d'où notre proposition d'étiquetage automatique. Dans DVB [6], l'information concernant les programmes est transmise dans une table de signalisation, plus précisément la table EIT, Event Information Table, qui décrit les programmes courant et à venir. Une autre possibilité est de recevoir ces informations sous la forme d'une description TV-anytime [7] qui permet de décrire de façon à la fois simple et très détaillée un ensemble de programmes de télévision.

Toutefois, l'utilisation de ces informations n'est pas immédiate, en raison de leur manque de précision : des retards de plusieurs dizaines de minutes peuvent apparaitre, certains programmes sont manquants ou erronés. Un recalage entre les informations extraites de l'EPG et la macro-segmentation est nécessaire. Un procédé simpliste consiste à tout d'abord classer comme programme les segments d'une durée supérieur à un certain seuil. Puis, pour un segment (x_1, x_2) classifié comme programme, on cherche dans l'EPG le premier segment (y_1, y_2) dans l'ordre chronologique qui vérifie :

$$(x_1 - y_1) > -\gamma \text{ et } (x_2 - y_2) < \gamma$$

La valeur de γ représente la valeur maximale de décalage (avance ou retard) autorisée. La figure 2 montre un exemple d'un tel recalage. Cette figure montre à la fois le manque de précision de la macro-segmentation et de l'EPG. La non-utilisation systématique des séparations introduit de faux étiquetages d'inter-programmes en tant que programmes, en particulier le parrainage (sponsoring), qui est accolé au programme, voir à l'intérieur. De plus, certains segments ne peuvent être étiquetés correctement du fait de l'absence de cette information dans le guide.

Des résultats plus quantitatif à partir d'une segmentation en plans donnent 65% de bonne classification des plans en inter-programme, ou en programme correctement étiqueté, sur 19h de vidéo. Ces résultats clairement insuffisants nous montrent la nécessité de chercher d'autres méthodes de recalage, à la fois pour extraire plus d'informations du flux, pour prendre en compte des a priori sur la construction du flux et du guide des programmes et pour être résistant aux retards, aux suppressions et aux segmentations des programmes par les publicités.

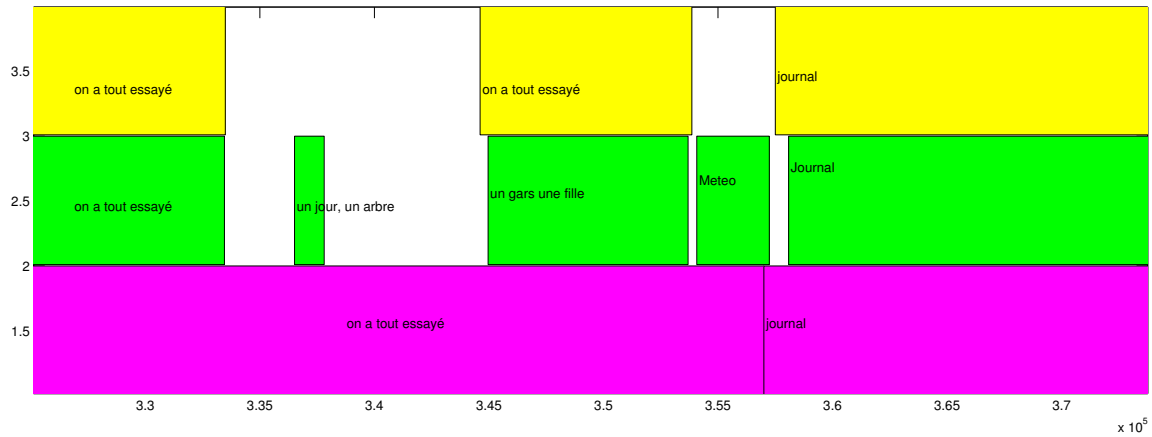

Figure 2 – *Exemple de recalage simple. En bas,violet, l'EPG, au milieu, vert, la vérité terrain, en haut, jaune, l'étiquetage automatique*

4.2 Méthode d'alignement et d'intégration de connaissances

Une technique intéressante pour aligner des séquences tout en intégrant des règles a priori est la distance d'édition utilisée classiquement pour aligner des chaines de caractères [8]. La distance d'édition est définie comme le coût minimal des transformations pour passer d'une séquence à une autre. Les transformations autorisées sont l'insertion, la suppression et la substitution. L'efficacité de l'algorithme est assurée par l'utilisation de la programmation dynamique.
Ainsi pour une macro segmentation $X_i = \{x_0 \dots x_i\}$ et le guide de programme associé $P_j = \{p_0 \dots p_j\}$ la distance d'édition est calculée par

$$D(X_i, P_j) = Min \begin{cases} D(X_{i-1}, P_{j-1}) + C_{sub}(x_i, p_j) \\ D(X_i, P_{j-1}) + C_{del}(p_j) \\ D(X_{i-1}, P_j) + C_{ins}(x_i) \end{cases}$$

Chaque élément de X_i et P_i est un couple de valeur indiquant le début et la fin du programme $x_i = (x_i^d, x_i^f)$. $C_{sub}, C_{del}, C_{ins}$ sont respectivement les coûts de substitution, de suppression et d'insertion. Ces fonctions de coût sont à définir et sont donc l'endroit idéal où définir des heuristiques permettant d'intégrer des informations a priori du domaine. Un exemple d'heuristique est de remarquer que la structure des émissions la nuit est complètement différente (peu ou pas de publicités, pas de coupure d'une émission par des publicités, suppression de programmes fortement probable, etc...) et de traduire ces remarques dans les fonctions de coût.
Nous donnons un exemple de définition de fonctions de coût, en reprenant les simplifications utilisées en traitement de la parole [9] et connu sous le nom de Dynamic Time Warping (DTW) :

$$\begin{array}{ll} C_{sub}(x_i, p_i) & = \gamma d(x_i, p_i) \\ C_{del}(p_i) & = d(x_i, p_i) \\ C_{ins}(x_i) & = d(x_i, p_i) \end{array}$$

où d est la distance locale entre composantes des vecteurs X_i et P_i. Classiquement, $\gamma = 2$. Cependant pour privilégier une substitution par rapport à une suppression puis une insertion, on doit vérifier $C_{sub}(x_i, p_j) < C_{ins}(x_i) + C_{del}(p_j)$ d'où $\gamma < 2$.
On définit ici la distance d par :

$$d(x_i, p_j) = \alpha|p_j^f - p_j^d - (x_i^f - x_i^d)| + \beta \left[|p_j^d - x_i^d| + |p_j^f - x_i^f| \right]$$

Cette distance est composée de 2 termes, le premier mesure la similarité de la longueur des segments x_i et p_j tandis que le deuxième terme mesure la similarité des horaires de diffusion. La seule information disponible pour effectuer l'alignement est en effet l'horaire de début et de fin du programme.
Cette méthode donne des résultats légèrement meilleurs en terme de classification (72% sur 19h) mais est surtout bien plus souple pour intégrer des contraintes. Il reste cependant à étudier si des fonctions intégrant des a priori permettent d'obtenir de meilleures performances. La figure 3 donne un exemple de recalage par DTW.

4.3 Intégration d'informations de reconnaissance

Un des moyens de s'affranchir du peu de précisison de la macro-segmentation et de corriger les erreurs de l'alignement est d'utiliser une méthode de reconnaissance de séquences vidéo de façon à détecter les répétitions. Les séquences répétées sont en général soit des inter-programmes ou soit des génériques d'émissions et sont donc utiles pour identifier les instants d'inter-programmes ainsi que les débuts et fins d´émissions. Nous utilisons ici la méthode de

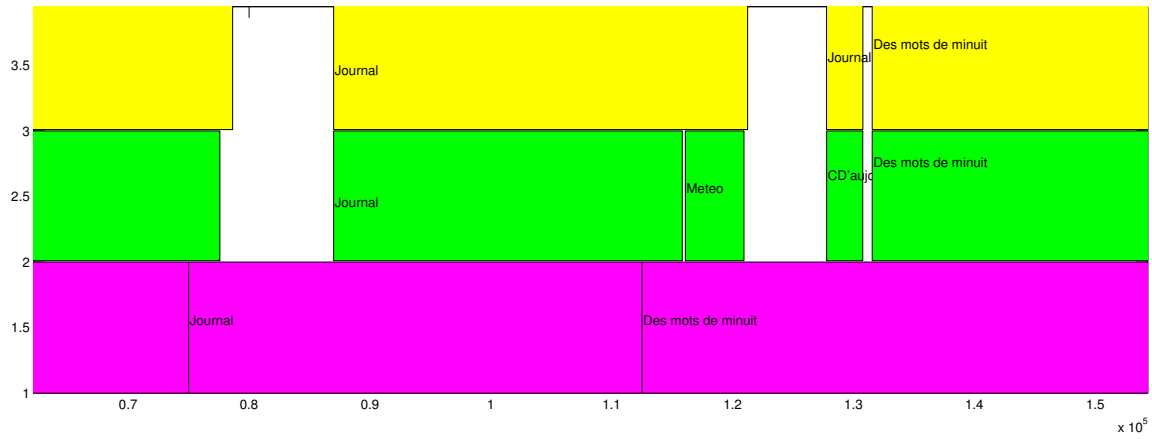

Figure 3 – *Exemple de recalage par DTW*

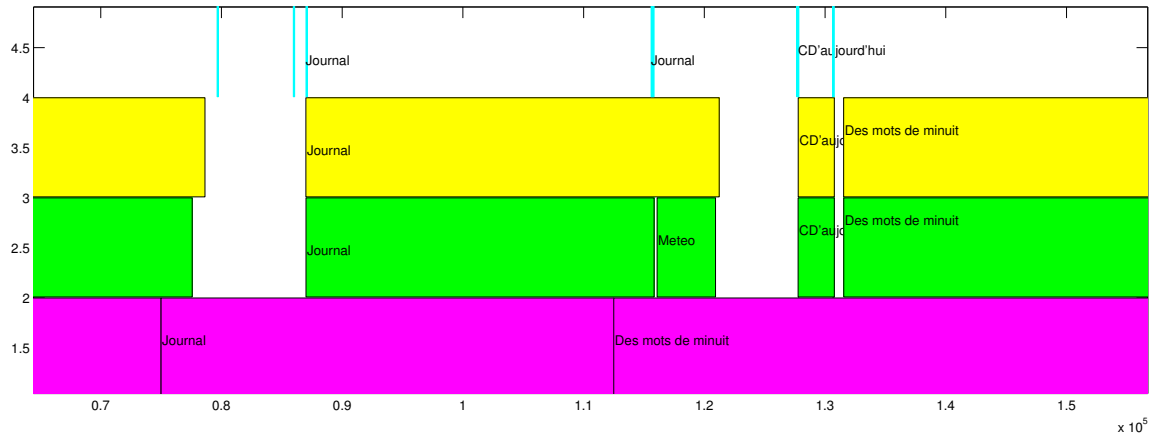

Figure 4 – *Recalage par DTW et intégration d'informations de reconnaissance. En haut en bleu, les segments ayant étés reconnus à l'aide d'une base de données partiellement labelisée.*

[10], qui calcule sur chaque image une signature extraite des coefficients DCT, et retrouve facilement un plan identique grâce à une table de hashage stockant les signatures.
Il s'agit maintenant d'intégrer trois couches d'information :
– une macro-segmentation basée sur la détection des séparations
– une information de reconnaissance par rapport à un historique de diffusion ou une base de donnée étiquetée
– le guide des programmes
La distance d'édition se prête bien à cette intégration. On peut supposer pour simplifier que l'on dispose d'une base de données de séquences vidéos étiquetées. Dans ce cas, la reconnaissance revient à étiqueter quelques parties du flux de façon certaine. Ceci se traduit dans l'algorithme d'alignement par des points de passage obligatoires. Ces points permettent aussi de contraindre l'espace de recherche sans avoir recours à une restriction arbitraire de l'espace.
Supposons qu'un segment ait été reconnu et que ce segment soit localisé dans une segmentation x_i (en jaune sur la figure). Ce segment possède une étiquette qui est cherchée dans l'EPG. On obtient ainsi un programme p_j si cette étiquette est présente, sinon le programme p_j est ajouté à l'EPG. L'intégration d'information de reconnaissance dans la distance d'édition se fait simplement en pré-remplissant la matrice des coûts par

$$\begin{cases} d(x_i, p_j) = \epsilon \\ d(x_l, p_k) = \infty \; \forall k < j \; , \forall l \geq i \end{cases}$$

La matrice des coûts est ensuite calculée de façon classique par la méthode du 4.2.
La figure 4 montre l'apport de la reconnaissance sur le même segment vidéo que la figure 3. L'étiquette "CD'aujourd'hui" qui n'existe pas dans l'EPG provient de la base de données utilisée pour la reconnaissance. L'information de reconnaissance est cependant encore mal utilisée puisque le générique de fin du journal a été reconnu, mais n'a pas permi de modifier la segmentation et d'isoler le programme "Météo". Une meilleure utilisation des résultats de la reconnaissance est à l'étude dans nos prochains travaux.

5 Conclusion

Le but de notre travail est de montrer la faisabilité d'une solution de structuration entièrement automatique des flux de télévision. Des méthodes très simples basées sur des a priori de production sont utilisées pour segmenter le flux en émissions et les étiqueter grâce au guide de programme. Une méthode basée sur la fusion des résultats d'un détecteur de silence et d'un détecteur d'image monochrome est présentée pour détecter les séparations entre programmes. Ces séparations ne sont cependant pas utilisées de façon systématique par les chaines, et de plus, le guide de programme est incomplet. L'étiquetage obtenu ne peut donc pas être parfait.
Afin d'améliorer la méthode, nous proposons d'une part d'utiliser un algorithme d'alignement utilisant la programmation dynamique résistant aux insertions et suppressions de programmes. D'autre part, l'utilisation de cet algorithme permet de prendre en compte les résultats d'une méthode de reconnaissance de séquences vidéo, qui permet d'étiqueter avec une confiance élevée certains segments, et d'améliorer et de compléter l'étiquetage.

Références

[1] Rainer Lienhart, Christoph Kuhmunch, et Wolfgang Effelsberg. On the detection and recognition of television commercials. Dans *International Conference on Multimedia Computing and Systems*, pages 509–516, 1997.

[2] D. Sadlier, Sean Marlow, Noel OConnor, et Noel Murphy. Automatic tv advertisement detection from mpeg bitstream. 2002.

[3] Alberto Albiol, Maria José Ch. Fulla, Antonio Albiol, et Luis Torres. Commercials detection using hmm's. Dans *Image Analysis for Multimedia Interactive Services, Wiamis'2004, Lisboa, Portugal*, April 2004.

[4] Pinar Duygulu, Ming yu Chen, et Alex Hauptmann. Comparison and combination of two novel commercial detection methods. *Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME 2004)*, june 2004.

[5] E. Kijak, G. Gravier, L. Oisel, et P. Gros. Audiovisual integration for sport broadcast structuring. *Multimedia Tools and Applications*, 2005.

[6] Etsi, es201-812, digital video broadcasting, multimedia home platform specifications 1.0.3.

[7] Tv-anytime forum, specification series on metadata. parta : Metadata schemas. http ://www.tv-anytime.org/.

[8] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR*, 163(4) :845–848, 1965.

[9] René Boite et Murat Kunt. *Traitement de la parole*. Presses polytechniques Romandes, 1987.

[10] Xavier Naturel et Patrick Gros. A fast shot matching strategy for detecting duplicate sequences in a television stream. Dans *CVDB'05 : Proceedings of the 2nd ACM SIGMOD international workshop on Computer Vision meets DataBases*, 2005.