Segmentation du son en se basant sur la distance de KullBack-Leibler

Hadi HARB , Liming CHEN, Jean-Yves AULOGE

Dept. Mathématiques - Informatique, ECOLE CENTRALE DE LYON. 36, avenue Guy de Collongue B.P. 163, 69131 ECULLY Cedex, France

{hadi.harb, liming.chen, jean-yves.auloge}@ec-lyon.fr

Résumé

L'indexation par le contenu d'une vidéo nécessite une analyse combinée du flux visuel et du flux sonore. Nous présentons ici un travail sur la segmentation du son qui a pour but de détecter les changements significatifs du flux audio. Ces changements peuvent correspondre à un changement de classe, entre musique, parole, silence, et autre, ou changement de type de musique, de locuteurs. Notre algorithme de segmentation du son est basé sur l'utilisation de la distance de Kullback-Leibler KL. L'application de cet algorithme n'a pas besoin d'une phase d'apprentissage comme cela est nécessaire pour les systèmes basés sur les Modèles de Markov Cachés ou sur les mélanges de lois Gaussiennes. Elle se fait en temps réel, avec une précision d'une seconde et les résultats d'expérimentation sur 14 minutes du film d'action « Eraser» sont encourageantes, donnant un taux d'insertion de 39.5% et un taux de suppression de 0%.

1- Introduction

La convergence rapide de l'informatique et de l'audiovisuel conduit à une profusion d'images et de vidéos numériques. Cependant, s'il existe des outils automatiques permettant de résumer ou d'indexer par des mots-clés un document textuel afin de faciliter l'accès à son contenu, il est difficile de faire de même avec un document audiovisuel.

Un document audiovisuel est composé d'un flux d'image, et d'un ou plusieurs flux audio en synchronisation avec les images. Si une analyse du contenu du flux visuel d'une vidéo est importante pour son indexation [Mahd 00] [Ardeb 01], l'analyse de la bande sonore, qui véhicule aussi beaucoup d'informations sémantiques est aussi indispensable [Gauc 99],[Wang 00].

Dans ce travail, nous proposons d'étudier le problème de la segmentation de la bande sonore afin d'en détecter les changements acoustiques significatifs d'un flux audio, entre les classes musique, parole, silence ou autre. La segmentation de la bande sonore est une phase importante avant la classification de celle-ci en des classes de base, comme musique/silence/parole. La norme MPEG7 en cours de définition, visant à être un standard de description multimédia, exige une telle segmentation et classification [Mpeg7] car un découpage précis d'une bande sonore en segments de musique, de parole, de silence ou autre permet d'envisager par la suite de nombreuses applications [Spin 00][Eide 00]. Par exemple, l'apparition de musique après un long silence peut être indicatrice d'un changement de scènes dans un film. En plus, les segments de parole détectés avec précision peuvent être l'objet d'une analyse approfondie pour une classification automatique en locuteurs masculins ou féminins puis un suivi de ceux-ci. Cette classification peut aider ensuite les outils de dictée vocale en vue d'une transcription automatique de paroles. Le schéma suivant illustre les différentes applications possibles d'une telle segmentation.

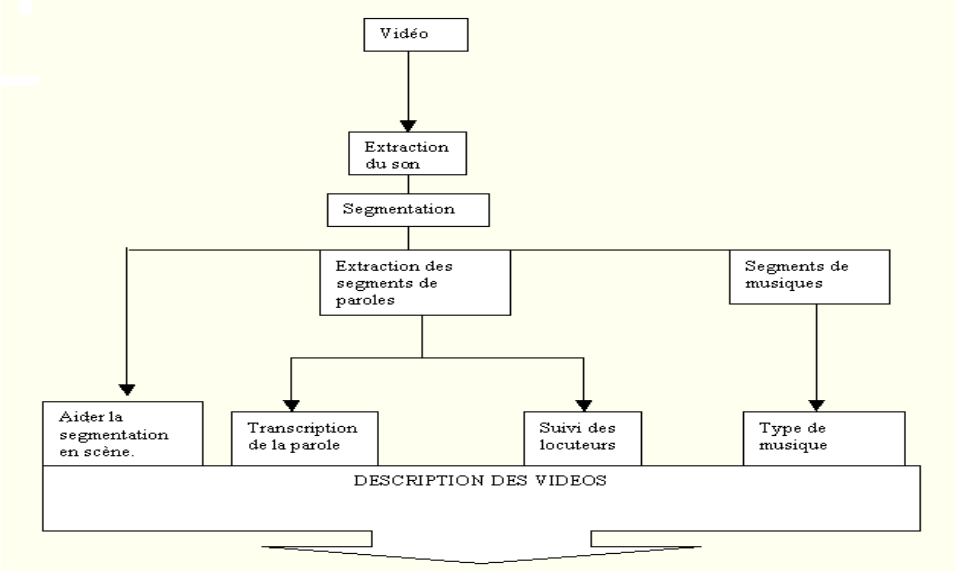

Fig. 1 - l'analyse du flux audio pour la description de vidéos

Le papier est organisé comme suit. Un état de l'art rapide est décrit dans la section 2. Nous décrivons notre approche dans la section 3. La section 4 présente les résultats d'expérimentation. Nous concluons dans la section 5 en donnant quelques indications sur nos travaux en cours.

2- Segmentation audio

La première approche pour une segmentation audio consiste à extraire des paramètres du signal de son dans des fenêtres consécutives. Une distance ou mesure de similarité appropriée entre les paramètres de deux fenêtres consécutives permet de marquer les bordures de segments [Siegler97]. Une autre approche que l'on peut trouver dans la littérature modélise d'abord les différentes classes de son, musique, parole, silence, etc., par des lois Gaussiennes, et marque les bordures de segments quand il y a changement de classe [Woodl98].
Compte tenu d'une très importante variabilité d'une classe sonore – la classe musique inclut la musique classique, rock, jazz, etc. -, l'approche basée sur les mélanges des lois Gaussiennes nécessite des larges bases d'apprentissage pour chaque classe. Ce qui n'est pas pratique pour notre domaine d'indexation par le contenu de la vidéo où chaque classe conduit à une très grande diversité existent (plusieurs locuteur, plusieurs types de musiques, plusieurs types de bruits …). Un autre inconvénient de cette approche est aussi le temps de calculs qui est généralement très coûteux. Néanmoins son avantage théorique est qu'elle réalise la segmentation et la classification en une seule étape.
Une comparaison de ces approches a été faite dans [Kemp00], où l'auteur montre qu'une approche métrique semble plus avantageuse.

3- Distance

Notre travail suit cette approche métrique et propose d'extraire de chaque seconde du signal un vecteur de variance et un vecteur moyen. Puis la distance de KullBack-Leibler est appliquée sur deux fenêtres (secondes) consécutives. Un seuillage de la courbe ainsi obtenue nous permet de marquer les bordures de segments.

3.1 Mesure de similarité

L'approche métrique pour la segmentation audio nécessite d'abord de définir une mesure de similarité entre les paramètres de deux fenêtres du signal sonore. Une mesure de similarité idéale est celle qui doit distinguer deux segments de nature différente tout en regroupant des segments similaires.
Plusieurs types de mesure de similarité peuvent être utilisés pour le son. La mesure la plus simple peut consister à produire une différence absolue entre deux vecteurs de caractéristique, ou encore utiliser la distance euclidienne entre deux vecteurs. Il s'agit des distances directes entre les paramètres extraits du signal sonore.
Des mesures de similarité basées sur la statistique peuvent être aussi utilisées. Dans ces distances, la similarité se fait entre la distribution statistique des données (dans des fenêtres de quelques secondes) non pas entre les données directement.
Dans le domaine de l'indexation, l'important est de détecter des changements significatifs, comme par exemple changement de classe (musique, parole, silence…) ou changement de type de musique ou de locuteurs. Ces changements importants peuvent aider à la compréhension de la sémantique dans une vidéo.

Considérons par exemple des segments vidéo associés respectivement à une musique douce calme et une musique rapide rythmique. Compte tenu du langage cinématographique généralement pratiqué, on peut en déduire que les segments, où la musique est douce harmonique, ne peut pas être une scène violente du type poursuite policière. D'un autre côté, une scène où il y a une musique rapide et rythmique ne peut pas non plus être une scène de coucher de soleil. Ce sont des éléments importants pour une indexation par le contenu de la vidéo.

Enfin, un utilisateur qui cherche des scènes de conversation doit pouvoir obtenir toutes les scènes contenant de la parole même s'il existe quelques changements locaux, un silence d'une seconde par exemple.

Les mesures de similarité adéquates pour la segmentation-indexation des vidéos doivent donc être des mesures qui cachent des changements locaux non significatifs (silence court, un mot dans une musique…) et qui montrent des cas de changement plus généraux, comme changement de type de musique par exemple.

En conséquence, les distances basées sur les distributions statistiques conviennent mieux pour ce types d'application car, elles sont basées sur la distribution ou le comportement acoustique dans des larges fenêtres (de quelques secondes). D'où notre choix de la distance statistique de KullBack-Leibler.

3.2 Distance de KullBack-Leibler

La distance de KullBack-Leibler (KL) est une distance issue de la théorie de l'information. C'est une distance entre deux variables aléatoires [Cove 91]. Dans le cas des distributions Gaussiennes, cette mesure sera [Siegler97]:

$$KL2(X,Y) = \frac{\sigma_X^2}{\sigma_Y^2} + \frac{\sigma_Y^2}{\sigma_X^2} + (\mu_X - \mu_Y)^2 (\frac{1}{\sigma_X^2} + \frac{1}{\sigma_Y^2})$$

avec $\sigma_X, \sigma_Y, \mu_X, et \mu_Y$ sont respectivement les variances de X et Y et les moyens de X et Y.

l'application de la distance KL nécessite donc la définition des variables X et Y et le calcul des gaussiennes, moyennes et variances sur les paramètres caractéristiques. Dans notre travail, nous extrayons comme paramètres caractéristiques les coefficients de la Transformée de Fourrier Discrète (TFD) du signal sonore pour une bande de fréquence entre 100 Hz et 4KHz. L'application de la TFD se fait chaque 10ms, donc le pas de la discrétisation de fréquence est de 100Hz[1] . Le choix de la taille de la fenêtre temporelle dans laquelle les paramètres des gaussiennes sont calculés est de 2s (nous avons testé la taille de 1s aussi). Dans chacune de ces fenêtres 20 vecteurs de TFD sont extraits[2].
Les variables X et Y sont dans notre cas les fenêtres consécutives sur lesquelles la distance KL sera appliquée.
Pour chaque fenêtre temporelle de 2s l'estimation des paramètres des Gaussiennes (moyennes et variances) se fait suivant les règles :

$$\vec{\mu}_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{iN} \end{pmatrix} \quad \vec{v}_i = \begin{pmatrix} v_{i1} \\ v_{i2} \\ \vdots \\ v_{iN} \end{pmatrix} \quad \vec{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iN} \end{pmatrix}$$

$\vec{\mu}_i$ le vecteur moyen, et $\vec{v}_i$ le vecteur de variance, $\vec{x}_i$ le vecteur de caractéristiques (dans notre cas c'est le vecteur de TFD)

$$\mu_{ij} = \frac{1}{M_i} \sum_{l=1}^{M_i} x_{lj} \qquad j = 1, \cdots, N$$

$$v_{ij} = \frac{1}{M_i} \sum_{l=1}^{M_i} (x_{lj} - \mu_{ij})^2 \qquad j = 1, \cdots, N$$

4- Résultats d'expérimentation

Les expérimentations de la segmentation ont été menées sur deux séquences de 7 min extraites du film «Eraser ». Ce film est un film d'action très riche en changement acoustique comme bruit, parole, musique. La première séquence choisie est une séquence « calme » juste avant que les combats commencent. La deuxième séquence de 7 min est une partie du film où il y a des actions avec des combats, des bruits de pistolet et des cris.
Deux indicateurs classiques, le taux d'insertion et le taux de suppression, sont utilisés pour évaluer la qualité de notre méthode de segmentation. Remarquons que, dans notre application d'indexation par le contenu de la vidéo, le taux de suppression est un critère plus significatif car

[1] Pas de fréquence = 1/Fenêtre sur laquelle la TFD est appliquée
[2] 1 vecteur TFD chaque 10ms donc 20 vecteurs chaque 2s

chaque segment supprimé sera par la suite classifié par erreur.
Dans nos expérimentations nous avons défini le taux de suppression comme étant le temps des segments supprimé par rapport au temps total. Cela indique les données perdues destinées principalement aux segments de parole. Le taux d'insertion ici est le nombre de segments ajoutés par rapport au nombre total des segments manuellement détectés.
Une segmentation manuelle précise en écoutant le son est difficile. Dans nos expérimentations, on s'est appuyé sur le graphe temporel du signal pour réaliser cette segmentation manuelle de référence.

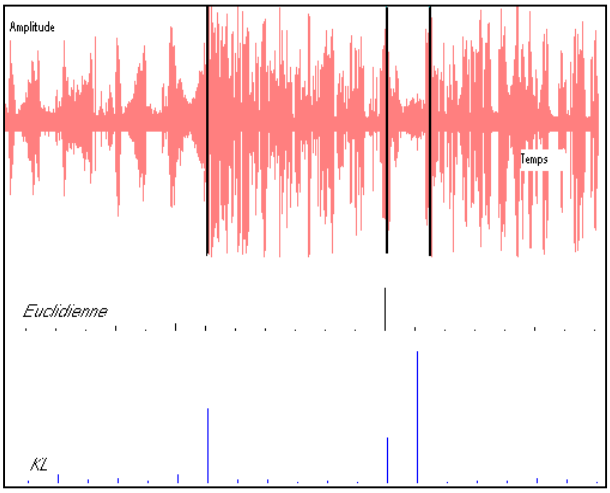

Fig.2 -la segmentation basée sur le graphe temporel du signal

Dans un but de comparaison, nous avons aussi implémenté la simple distance euclidienne. Les résultats pour une taille de fenêtre de 2s sont résumés par les tables suivantes :

4.1 Film « Eraser », séquence 1 :

	KL	Euclidienne
Temps supprime en seconde	0	140
Durée totale en seconde	420	420
Taux de suppression	0%	33%

	KL	Euclidienne
Nbre de segments ajoutés	24	4
Nbre de segments détectés	68	26
Taux d'insertion	35%	15%

4.2 Film « Eraser », séquences 2 :

	KL	Euclidienne
Temps supprimé en seconde	0	36
Durée totale en seconde	600	600
Taux de suppression	0%	6%

	KL	Euclidienne
Nbre de segments ajoutés	31	11
Nbre de segments détectés	70	45
Taux d'insertion	44%	24%

Ces résultats nous montrent que la distance de KulBack Leibler est plus avantageuse que la distance Euclidienne classique pour ce problème de segmentation du son car le taux de suppression par la distance KL est faible (0% pour 14 min). Cependant le taux d'insertion est important, et il est lié au type de scène.

Les tests ont aussi montré que pour une taille de fenêtre de 1s le taux d'insertion augmente considérablement sans obtenir une amélioration pour le taux de suppression.

Fig.3 -un exemple avec une taille de fenêtre de 1s (en haut) et 2s (en bas), on peut voir les segments ajoutés pour une taille de fenêtre de 1s et qui ne sont pas détectés pour une taille de fenêtre de 2s

5- Conclusion

Dans cet article nous avons présenté notre méthode de segmentation du flux audio dans un but d'indexation par le contenu de vidéos. Après une

analyse des besoins de l'indexation de la vidéo, nous sommes arrivé à la conclusion qu'une approche métrique statistique convient le mieux pour notre problématique. Notre méthode utilise la distance KullBack-Leibler (KL) afin de marquer les bordures des changements significatifs de la bande sonore. Les expérimentations que nous avons menées montrent que l'utilisation de la distance de KL pour une taille de fenêtre de 2s est la plus avantageuse. Notre méthode peut être appliquée en temps réel. Ces travaux ont été à la base de notre technique pour une segmentation et classification de la bande sonore menée sur un très grand corpus qui fait l'objet d'un dépôt de brevet [Harb01]. A l'heure actuelle nous poursuivons notre travail pour une reconnaissance et suivi de locuteurs.

6- Références

[Ardeb01] M. Ardebilian, « une contribution à la segmentation par le contenu de la vidéo », thèse doctorat de l'université de technologie de Compiègne, 2001.

[Gauch99] John M. Gauch, Susan Gauch, Sylvain Bouix, Xiaolan Zhu, *Real Time Video Scene Detection and Classification*(1999), Information Processing and Managment 35 pp 401-420, 1999.

[Kemp00] T. Kemp, M. Schmidt, M. Westphal, A. Waibel, *Strategies for automatic segmentation of audio data,* Proc. icassp2000 ISTANBUL pp 1423-6, 2000

[Mahdi01] W. Mahdi, M. Ardebilian, L. Chen, *Automatic Scene Segmentation Based on Spatial-Temporal Clues and Rhythm*, to appear in International Journal of Networking and Information Systems, Vol. 5 Septembre 2001.

[Siegler97] M. Siegler, U. Jain, B. Ray and R. Stern, *Automatic segmentation, classification and clustering of broadcast news audio*, Proceedings of the Speech Recognition Workshop, pp 97-99, 1997.

[Wood98] P. C. Woodland, T. Hain, S. Johnson, T. Neisler, A. Tuerk, S. Young, *Experiments in Broadcast news transcription,* Proc. ICASSP 1998, Seattle, May 1998.

[Wang00] Yao Wang, Zhu Liu, Jin-Cheng Huang *Multimedia Content Analysis Using Both Audio and Visual Cues*(2000), IEEE Signal Processing Magazine, PP 12-36, novembre 2000.

[Harb01] Hadi Harb, Liming Chen, Jean-Yves Auloge, *Segmentation et classification du son* (2001), Rapport technique, Juillet 2001, Dépt. MI, Ecole Centrale de Lyon.