MajecSTIC 2009
Avignon, France, du 16 au 18 novembre 2009

Classification de Genre Vidéo basé sur l'audio

Mickael Rouvier[1], Georges Linarès[1] et Driss Matrouf[1] *

1 : Laboratoire Informatique d'Avignon, Université d'Avignon, 84000 Avignon - France.

Contact : mickael.rouvier@univ-avignon.fr,
georges.linares@univ-avignon.fr, driss.matrouf@univ-avignon.fr

Résumé
Dans un contexte mondial de croissance rapide des collections vidéos accessibles sur Internet, la classification de genre vidéo devient une tâche difficile. Dans ce papier, nous présentons une nouvelle méthode pour l'indentification de genre vidéo basée sur l'analyse du contenu audio. Notre approche repose sur la combinaison de bas et haut niveau de feature audio. Nous étudierons la capacité discriminative des paramètres liée à l'instabilité acoustique, l'interactivité du locuteur, la qualité de la parole et la caractérisation de l'espace acoustique. L'indentification de genre est effectuée sur ces paramètres en utilisant un classifieur SVM. Les expérimentations sont conduites sur un corpus composé de cartoons, films, actualités, publicités et musiques sur lequel nous obtenons, pour la meilleure configuration, un taux de classification de 91%.

Abstract
Video genre classification is a challenging task in a global context of fast growing video collections available on the Internet. In this paper, we present a new method for video genre identification by audio contents analysis. Our approach relies on the combination of low and high level audio features. We investigate the discriminative capacity of features related to acoustic instability, speaker interactivity, speech quality and acoustic space characterization. The genre identification is performed on these features by using a SVM classifier. Experiments are conducted on a corpus composed from cartoons, movies, news, commercials and musics on which we obtain an identification rate of 91%.

Mots-clés : identification de genre vidéo, facteur analysis, classification automatique

Keywords: video genre identification, factor analysis, automatic classification

1. Introduction

Ces dernières années, le nombre de vidéos accessibles sur Internet et la télévision a considérablement augmenté, rendant ainsi la recherche dans ses larges bases de données, difficile pour les utilisateurs, sans outil efficace. Ce point a suscité beaucoup de travail pour structurer les bases de données audio-visuelles par l'analyse de contenu. Dans la littérature, la plupart des approches trouvées sont basées sur la vidéo [2]. Quelques auteurs ont cherché à faire de la catégorisation de genre en se basant sur le texte [10] issu des sous-titres ou de la transcription automatique de la parole. Le web est une base très variable et le taux de reconnaissance est généralement trop élevé pour faire une analyse correcte de la transcription automatique. Une approche bas niveau, uniquement basée sur l'audio, pourrait être plus efficace. Ainsi, quelques auteurs proposent des méthodes de classification de genre vidéo basées uniquement sur l'audio.
Plutôt que de se focaliser sur les stratégies de classification, nos efforts se sont portés sur les paramètres utilisés pour la caractérisation du genre vidéo. L'approche conventionnelle consiste, dans la caractérisation de l'espace acoustique, à utiliser un classifieur statistique comme par exemple un Gaussian Mixture Model (GMM), Réseau de neurones ou Support Vector Machines (SVM)

*. Cette recherche est supporté par l'ANR (Agence Nationale de la Recherche), projet RPM2 (ANR-07-AM-008)

sur le domaine cepstral [4, 8, 9]. [7–9] proposent de classer le genre vidéo avec des paramètres temporels comme le Zero Crossing Rates (ZCR) ou l'énergie.
Dans ce papier, nous nous focalisons sur la classification de genre vidéo en utilisant uniquement l'audio. Nous proposons de combiner des descripteurs de haut et bas niveau basés non seulement sur l'analyse cepstral, mais aussi sur la structure audio de la vidéo cible. Finalement, nous comparons diverses méthodes de combinaison basées sur des classifieurs statistiques.
Ce papier est organisé comme suit : la prochaine section décrit précisément la tâche ainsi que le corpus utilisé pour cette étude. La section 3 présente le principe et l'architecture du système proposé pour la catégorisation. Dans la section 4, nous discutons des performances obtenues ainsi que de la complémentarité des différents niveaux de discripteurs.
Nous évaluons tout d'abord la capacité discriminative de chacun des descripteurs et estimer leur complémentarité. La section 5 se focalise sur les performances de classification. Nous comparerons différents systèmes de combinaisons : une combinaison dans l'espace des paramètres et un autre dans l'espace des probabilités. La section 6 conclura et proposera quelques perspectives.

2. Tâche et corpus

Nous avons sélectionné 5 catégories qui sont communément utilisées dans la classification de genre vidéo : actualité, film, cartoon, musique et publicité. Le corpus est construit avec 1200 vidéos indexées et ayant toutes une durée comprise entre 2 et 5 minutes. 1000 sont utilisées pour l'entraînement de notre système et 200 composent le test. Les 5 genres sont équitablement représentés dans la base de données (environ 200 vidéos par genre pour l'entraînement et 40 pour le test). La langue contenue dans les vidéos est systématiquement le franccais, cependant la catégorie musique contient quelques fois des morceaux en anglais.

3. Architecture du système

Le système proposé est une architecture à 2 niveaux. Le premier niveau consiste à extraire sur une vidéo les paramètres acoustiques qui sont ensuite combinés au second niveau. Nous identifions 4 groupes de descripteurs qui sont décrits plus en détail dans les prochains paragraphes :
– espace acoustique : c'est le descripteur le plus fréquemment utilisé pour la catégorisation vidéo basée sur l'audio. L'idée générale est de distinguer le genre par des classifieurs statistiques sur les coefficients cepstraux de notre signal.
– intéractivité du locuteur : le genre vidéo peut être différent selon le profil d'intéractivié, par exemple, le nombre de locuteur et le temps accordé à chaque locuteur sont probablement différents entre un cartoon et une actualité.
– qualité de la parole : nous évaluons la qualité de la parole par une analyse acoustique, mais aussi le contenu de la parole par une évaluation à posteriori de la reconnaissance de la parole.
– instabilité : ces paramètres représentent la régularité acoustique dans le domaine du temps.

Afin d'évaluer la capacité discriminative de chacun de ces groupes de paramètres, nous entraînons un jeu de GMM pour chaque genre. Le processus de décision suit le schéma classique de classificiation bayesien. Le genre identifié est celui qui maximise la condition de probabilité suivante : $P(X^k|G_i^k)$ où X est le vecteur de paramètres, i le genre et k le groupe de paramètres.
Le second niveau combine les paramètres extraits du niveau-1 pour chaque document. Les probabilités sont groupées dans un vecteur. Un classifieur SVM est entraîné sur l'ensemble des vecteurs estimés sur la base d'apprentissage. Nous utilisons pour le SVM un noyau linéaire qui a été entraîné par une stratégie de *leave-one-out*. Le SVM combine les résultats sortis par les GMM ; nous ne pouvons donc pas utiliser pour son entraînement, les données utilisées pour la classficiation GMM. Par conséquent, nous découpons notre corpus en 2 parties. Le premier est utilisé pour entraîné les classifieurs GMM, les probabilités résultant du GMM deviennent les vecteurs d'entrée pour l'entraînement du SVM.
La complémentarité des paramètres est estimée en suivant un protocole qui consiste à ajouter, à chaque étape, un nouveau groupe de paramètres au précédent jeu. Ensuite, nous entraînons un SVM et nous observons si l'ajout de ce nouveau groupe apporte un complément d'informations au précédent groupe. Finalement, les performances sont évaluées sur le vecteur total qui intégre les 4 descripteurs.

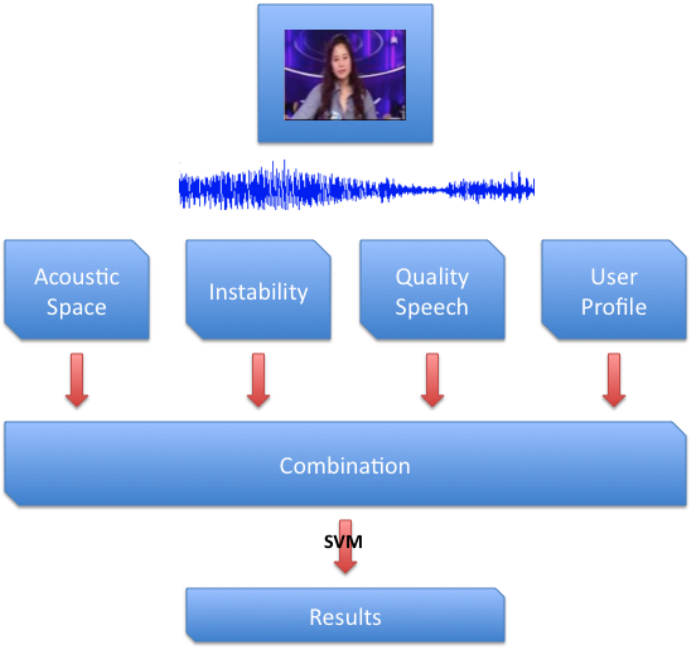

FIGURE 1 – *Principle scheme of genre classifier : features related to acoustic space characterization, speaker interactivity, speech quality and Acoustic Instability are extracted and combined by a SVM classifier.*

4. Bas et Haut niveau de paramètre pour l'indentification de genre

4.1. Caractérisation de l'espace acoustique

Une des approches les plus populaires pour l'identification de genre est la caractérisation de l'espace acoustique à travers une analyse MFCC et un classifieur GMM ou SVM [7–9]. [7] ont démontré l'efficacité une paramètrisation du signal MFCC, avec des vecteurs de 14 coefficients MFCC auquel sont ajoutés les dérivés premières et secondes. En utilisant un classifieur GMM, ce système relativement simple obtient une identification correcte de 52% sur 5 genres (film, cartoon, publicité, actualité et musique). Cette approche a été motivée par les performances du MFCC/GMM sur la tâche d'identification du locuteur.

Ici, la caractérisation de l'espace acoustique est effectuée de manière similaire, mais nous proposons d'explorer différentes voies. Les coefficients PLP sont connus pour être plus robustes que les MFCC sur la tâche de reconnaissance du locuteur. Tout d'abord, nous allons comparer ces 2 jeux de paramètres.

Ensuite, une des principales difficultés de la catégorisation de genre repose sur la diversité des vidéos qui sont étiquetées similairement. Quelques techniques pour réduire la variabilité intra-class ont été proposées dans l'indentification du locuteur, plus particulièrement le facteur analysis a démontré une efficacité améliorée ([6]). Ici, nous évaluons ces méthodes pour réduire la varabilité intra-genre.

4.1.1. Analyse factorielle pour la classification de genre

L'utilisation de GMM dans un cadre GMM-UBM est une approche standard dans la vérification de locuteur. Un modèle du monde (UBM-GMM) modélise tous les genres : pour chaque genre un GMM est obtenu en adaptant le modèle du monde. Seule la moyenne des vecteurs est adaptée, le poids et les variances restent inchangés.

L'approche de l'analyse factorielle (Factor Analysis, FA) est de décomposer le modèle genre-spécifique en 3 différentes composantes : une composante genre-session-indépendant, une composante genre-dépendant et une composante session-dépendant. Un supervecteur GMM est défini comme la concaténation des composantes des moyennes des gaussiennes du GMM. D est la

dimension de l'espace du paramètre, la dimension du supervecteur est MD où M est le nombre de gaussienne dans le GMM. Pour un genre GE appartenant à la session h, le modèle du FA peut être écrit comme :

$$\mathbf{m}_{(h,GE)} = \mathbf{m} + \mathbf{D}\mathbf{y}_{GE} + \mathbf{U}\mathbf{x}_{(h,GE)}, \quad (1)$$

où $\mathbf{m}_{(h,GE)}$ est la session-genre dépendant du supervecteur. $\mathbf{D}$ est MD $\times$ MD la matrice diagonale, $\mathbf{y}_{GE}$ le vecteur de genre (a MD vecteur), $\mathbf{U}$ est la variable de session de ranq R (une matrice MDxR) et $x(h, ge)$ sont la composante du canal, un vecteur R. Tous les paramètres du modèle FA sont estimés en utilisant le critére de maximum de vraisemblance et l'algorithme EM. L'ensemble des sessions correspondant à chaque genre a été utilisé pour une meilleure estimation des paramètres FA. La composante session du modèle est simplement obtenue 1, comme :

$$\mathbf{m}_{GE} = \mathbf{m} + \mathbf{D}\mathbf{y}_{GE}, \quad (2)$$

4.1.2. Expérience

Les paramétrisations acoustiques en PLP et MFCC sont tout d'abord comparées à une baseline (classification GMM). Pour tous les deux, nous utilisons une mixture de modèles de 512 gaussiennes (un GMM par genre) entrainés à chaque itération par maximum de vraisemblance (ML) et l'algorithme d'espérance-maximisation (EM). Les résultats sont reportés dans le tableau 1. Les performances obtenues par une analyse MFCC sont proches de celles rapportées dans [7] sur une tâche similaire. Les PLP obtiennent de meilleurs résultats, avec une amélioration relative de 42%, en comparant à la baseline MFCC/GMM.

Les performances du système sont, quant à elles, fortement améliorées par le FA. Sur le systéme basé sur les PLP, le taux d'erreurs relatif est réduit d'environ 29%, le taux d'indentification augmente de 76% à 86%.

TABLE 1 – *Taux correct d'identification de genre par classification GMM sur l'espace acoustique : les paramètres PLP et MFCC sont comparés sans et avec la réduction de variabilité du Factor Analysis (FA-MFCC and FA-PLP).*

	Musique	Actualité	Publicité	Cartoon	Film	Total
MFCC	0.58	0.84	0.17	0.31	0.73	0.52
PLP	0.95	0.92	0.46	0.78	0.70	0.76
FA-MFCC	0.95	0.84	0.58	0.85	0.92	0.83
FA-PLP	0.97	0.97	0.56	0.87	0.95	0.86

Finalement, le meilleur système obtient 86% d'indentification. Tous les genres (excepté publicité) sont reconnus avec plus de 86%. Les publicités obtiennent le plus mauvais résultat pour toutes les méthodes. Ce dernier résultat est vraisemblablement obtenu par sa nature similaire aux autres genres : nous avons observé par les taux de confusions avec la musique, actualité, cartoon et film respectivement sont de 22%, 10%, 10% et seulement 2% pour le cartoon.

4.1.3. Paramètre d'interactivité

Le nombre de locuteurs et la faccon dont ils communiquent entre eux peuvent être différents selon le genre. Par exemple, il y a généralement un principal locuteur dans l'actualité, contrairement aux cartoons et films qui contiennent beaucoup de locuteurs avec un nombre de changements de locuteur important.

Le vecteur est construit avec 3 paramètres : le nombre de changements de locuteur, le nombre de locuteurs et le rapport entre le temps de parole du principal locuteur sur le temps de la vidéo.

Ces données sont extraites en utilisant un système de segmentation et regroupement de locuteur basé sur une segmentation à 3 étages et un processus de clusterisation. Le premier étage effectue une passe de segmentation Viterbi basée sur les 3 classes : parole, parole sur de la musique et musique. Chacune de ces classes est un modèle de GMM de 64 mixtures. Les vecteurs acoustiques

sont composés de 12 coefficients MFCC de l'énergie, ainsi que les dérivées premières et secondes. Ce système est complètement décrit ici ([3]). Les 2 derniers étages permettent d'améliorer la détection de changement des locuteurs. Nous utilisons le système décrit en [11] basé sur un Critére d'Information Bayesien (BIC) et une stratégie d'agglomération de *cluster*. Cette technique nous permet d'estimer le nombre de locuteurs et le nombre de changements de locuteur pour chaque document.

Nous nous intéressons d'abord à une classification basée sur les GMM avec les 3 descripteurs d'interactivité. Ensuite, les sorties de GMM sont combinées à un classifieur acoustique décrit dans les précédentes sections : toutes les probabilités sont groupées dans un vecteur de probabilité sur lequel on procède à une catégorisation SVM. Comme décrit en section 3, le SVM est entraîné par une stratégie de *leave-one-out* ce qui permet d'estimer les paramètres SVM sans aucun corpus ni donnée additionnelle.

TABLE 2 – *Les paramètres d'intéractivité pour la classification de genre : taux correcte d'identification par genre en utilisant un GMM sur l'espace acoustique (AS), interactivité (Int), et la combinaison de l'interactivité et de l'espace acoustique (AS+Int) dans un classifieur SVM.*

	Musique	Actualité	Publicité	Cartoon	Film	Total
AS	0.97	0.97	0.56	0.87	0.95	0.86
Int.	0.29	0.52	0.90	0.95	0.56	0.64
AS+Int	0.95	0.84	0.85	0.85	0.92	0.88

Les résultats montrent globalement que l'interactivité est moins discriminante que les paramètres cepstraux. Néanmoins, celle-ci permet d'apporter un complément d'informations afin d'augmenter de faccon significative le rappel sur les publicités et les cartoons qui était très mal reconnu par les méthodes d'espace acoustique.

4.1.4. Qualité de la parole

L'idée est que la qualité de la parole permet d'apporter une information pertinente sur le genre. Par exemple, l'acoustique est normalement de bonne qualité dans l'actualité. Son domaine linguistique est bien couvert dans par le modèle de langage du système de reconnaissance, tandis que le domaine linguistique de la publicité peut être inattendu lié à des produits spécifiques et à un style de parole propre.

Nous utilisons 3 paramètres dans ce groupe, tous basés sur le moteur de reconnaissance vocale Speeral du LIA ([5]). Celui-ci est utilisé sur un n-gram modèle de langage et des états partagés context-dépendant HMM pour le modèle acoustique.

Le premier groupe de paramètres est la probabilité à posteriori de la meilleure hypothèse donnè par le système de transcription automatique. Le dernier paramètre est basé sur l'entropie phonétique. Ce paramètre a été proposé par [1] pour la classification parole/musique. Il est calculé comme l'entropie de la probabilité acoustique :

$$H(n) = -\frac{1}{N}\sum_{m=1}^{N}\sum_{k=1}^{K} P(q_k|x_m)\log_2 P(q_k|x_m) \tag{3}$$

où les valeurs des trames sont calculées sur une fenêtre glissante de taille N, K représente un modèle phonétique et $P(q_k|x_m)$ représente la probabilité d'avoir un phonéme sachant une trame. Cette mesure est supposée être grande sur une qualité de parole moyenne, et décroître sur une qualité de parole propre.

La qualité de la parole semble être moins précise que le précédent descripteur, excepté pour l'actualité qui et reconnue à 94%. En combinant tous les descripteurs, aucun gain supplémentaire n'est observé, la reconnaissance de la musique est parfaite tandis que le taux de reconnaissance des publicités et des cartoons décroît. Ces résultats indiquent une faible complémentarité avec les autres descripteurs sur le genre. De plus, la combinaison dans l'espace des scores n'est peut être

TABLE 3 – *La qualité de la parole pour la classification de genre : taux correcte d'identification par genre en utilisant un GMM sur la qualité de la parole (Q), et combiné avec les autres paramètres (AS+Int+Q).*

	Musique	Actualité	Publicité	Cartoon	Film	Total
AS+Int	0.95	0.84	0.85	0.85	0.92	0.88
Q	0.63	0.94	0.46	0.41	0.39	0.56
As+Int+Q	1.0	0.84	0.82	0.78	0.95	0.88

pas adaptée. Nous discuterons de ce point dans la section 5, où nous nous focaliserons sur les stratégies de classification.

4.1.5. Instabilité acoustique

Ce groupe de paramètres permet d'extraire la variabilité acoustique dans le domaine du temps. Nous implémentons ici 5 descripteurs : l'énergie à court terme (Short Time Energy, STE), le taux de passage à zéro (Zero Crossing Rates, ZCR) et le nombre de cassures de modèle acoustique (Acoustic Model Breaking, AMB).

STE est calculé classiquement dans une fenêtre glissante. Considérant que le niveau moyen de l'énergie ne permet pas d'apporter une information significative, nous normalisons l'énergie par le niveau d'énergie maximum observé dans le document. Ensuite, la moyenne STE et les variances sont extraites du document et utilisées comme descripteur d'instabilité.

ZCR est le nombre de fois où le signal audio passe l'axe des zéros chaque seconde. ZCR est plus fréquemment utilisé dans les algorithmes de classification parole/musique, mais plus généralement, il permet de représenter la variabilité d'une forme qui est trouvé dans un document. Ici, nous utilisons la moyenne et la variance de ZCR comme un indice d'instabilité.

Le dernier descripteur est la densité de ruptures acoustiques. Cette valeur est calculée par un détecteur de rupture similaire au BIC, qui est celui utilisé dans le système de détection de locuteur. Un modèle mono-gaussienne unique est estimé sur chaque fenêtre de 30 ms. Ensuite, chaque fenêtre de signal est partagée en 2 parties sur lesquelles 2 modèles uniques sont entraînés. Le gain de vraissemblance obtenu peut être divisé du modèle et utilisé comme indice de rupture. Si cet indice dépasse une certaine valeur, nous considérons que cette rupture est détectée. Le nombre de ruptures détectées (normalisé par la durée du document) est utilisé comme un indice d'instabilité.

TABLE 4 – *L'instabilité acoustique pour la classification de genre : taux correcte d'identification de genre en utilisant un GMM pour l'instabilité (Un), et combiné avec les autres paramètres (AS+Int+Q+Un).*

	Musique	Actualité	Publicité	Cartoon	Film	Total
As+Int+Q	1.0	0.84	0.82	0.78	0.95	0.88
Un.	0.60	0.68	0.36	0.73	0.41	0.55
AS+Int+Q+Un	0.97	0.86	0.85	0.92	0.92	0.91

Les résultats montrent que l'instabilité seule n'est globalement pas pertinente, mais tous les descripteurs apportent une information complémentaire qui permet d'obtenir une réduction relative du taux d'erreurs d'environ 25% (de 88% à 91% d'identifications correctes).

5. Combinaison de modèles contre combinaison de paramètres

Dans les précédentes expériences, nous choisissons de combiner les résultats des classifieurs GMM tandis que l'approche classique serait de grouper tous les paramètres dans un vecteur de paramètres multi-dimensionnel directement intégré dans le classifieur. Cette approche est motivée par le fait que la difficulté de la classification peut être améliorée par le fort taux d'hétérogénéité des données. La projection des observations dans l'espace des probabilités permet de

travailler dans un cadre unifié, où les stratégies de combinaisons peuvent être appliquées. D'un autre côté, quelques informations peuvent être perdues par une classification intermédiaire.
Nous comparons les 2 stratégies en utilisant une classification basée sur le SVM.
Suite à cette classification intermédiaire, nous allons aussi utiliser un entraînement de SVM par la stratégie de *leave-one-out* décrite précédemment. Les résultats de ces 2 méthodes sont notées dans les tableaux 5 et 6. Ceux-ci montrent que les 2 approches obtiennent des performances globalement similaires, même si : cartoon et film sont mieux détectés dans la combinaison de paramètres tandis que la musique l'est mieux dans la combinaison de modèles. La qualité de la parole permet d'obtenir un meilleur gain, lequel n'est pas observé avec la combinaison de modèles. L'interactivié des paramètres n'améliore pas les performances globales. Comme dit précédemment, les changements sont probablement dûs au fait que les corrélations de paramètres sont lissées par la combinaison de modèles. Cela permet aussi de penser que la de combinaison par genre permet d'améliorer le taux d'identification.

TABLE 5 – *Les performances des* **paramètres** *combiné aux paramètres cepstraux (AS), interactivié (I), qualité de la parole (Q) et l'instabilité (U). La combinaison est effectué par un SVM chaque vecteur contient successivement chaqu'un des descripteurs. La baseline consiste a une approche classique : MFCC/GMM.*

	Musique	Actualité	Publicité	Cartoon	Film	Total
Baseline	0.58	0.84	0.17	0.31	0.73	0.52
AS	0.97	0.97	0.56	0.87	0.95	0.86
AS+I	0.95	0.84	0.85	0.85	0.92	0.88
AS+I+Q	1.0	0.84	0.82	0.78	0.95	0.88
AS+I+Q+U	0.97	0.86	0.85	0.92	0.92	0.91

TABLE 6 – *Les performances du* **modèle** *combiné aux paramètres cepstraux (AS), interactivié (I), qualité de la parole (Q) et l'instabilité (U). La combinaison est effetué par un SVM chaque vecteur contient les probabilités des GMM du niveau-1.*

	Musique	Actualité	Publicité	Cartoon	Film	Total
Baseline	0.58	0.84	0.17	0.31	0.73	0.52
AS	0.97	0.97	0.56	0.87	0.95	0.86
AS+I	0.82	0.78	0.90	0.90	0.90	0.86
AS+I+Q	0.80	0.92	0.87	0.97	0.87	0.89
AS+I+Q+U	0.87	0.84	0.85	0.97	1.0	0.91

6. Conclusion et Perspective

Nous avons étudié les paramètres haut et bas niveau pour la classification de vidéos par genre basée sur l'audio. La classification sur chaque groupe de paramètre montre clairement que la caractérisation de l'espace acoustique reste le descripteur de genre le plus discriminant, spécialement avec une paramétrisation PLP et une réduction de la variabilité par une analysise facorielle. Cependant, les descripteurs de haut niveau comme l'interactivité et la qualité de la parole permettent quelques compléments d'informations : nous obtenons finalement une identification d'environ 91%. Ce résultat est significativement meilleur que les résultats basés sur la vidéo, et est similaire à ceux obtenu dans la littérature par une combinaison audio-vidéo. Nous envisageons de généraliser cette méthode en augmentant le nombre de classe et d'évaluer les différentes combinaisons de stratégies.

Bibliographie

1. Jitendra Ajmera, Iain A. McCowan, et Herve Bourlard. Robust hmm-based speech/music segmentation. In *ICASSP 2002*, 2002.
2. Darin Brezeale et Diane J. Cook. Automatic video classification : A survey of the literature. In *Systems, Man, and Cybernetics*, 2008.
3. Dan Istrate, Nicolas Scheffer, Corinne Fredouille, et Jean-FranÃ§ois Bonastre. Broadcast news speaker tracking for ester 2005 campaign. In *Interspeech 2005*, 2005.
4. R.S. Jasinschi et J. Louie. Automatic tv program genre classification based on audio patterns. In *Euromicro Conference, 2001*, 2001.
5. Georges Linarès, Pascal Nocéra, Dominique Massonie, et Driss Matrouf. The lia speech recognition system : from 10xrt to 1xrt. In *Lecture Notes in Computer Science*, 2007.
6. Driss Matrouf, Nicolas Scheffer, BenoÃ®t Fauve, et Jean-FranÃ§ois Bonastre. A straightforward and efficient implementation of the factor analysis model for speaker verification. In *InterSpeech 2007*, 2007.
7. Matthew Roach et John Mason. Classification of video genre using audio. In *European Conference on Speech Communication and Technology*, 2001.
8. Matthew Roach, Li-Qun Xu, et John Mason. Classification of non-edited broadcast video using holistic low-level features. In *IWDC'2002*, 2002.
9. Li-Qun Xu et Yongmin Li. Video classification using spatial-temporal features and pca. In *Multimedia and Expo, 2003. ICME '03*, 2003.
10. Weiyu Zhu, Candemir Toklu, et Shih-Ping Liou. Automatic news video segmentation and categorization based on closed-captioned text. In *Multimedia and Expo, ICME*, 2001.
11. Xuan Zhu, Claude Barras, Sylvain Meignier, et Jean-Luc Gauvain. Combining speaker identification and bic for speaker diarization. In *Interspeech 2005*, 2005.