TALN 2010, Montréal, 19–23 juillet 2010

Classification du genre vidéo reposant sur des transcriptions automatiques

Stanislas Oger, Mickael Rouvier, Georges Linarès *
LIA, Université d'Avignon, France
{stanislas.oger, mickael.rouvier, georges.linares}@univ-avignon.fr

Résumé. Dans cet article nous proposons une nouvelle méthode pour l'identification du genre vidéo qui repose sur une analyse de leur contenu linguistique. Cette approche consiste en l'analyse des mots apparaissant dans les transcriptions des pistes audio des vidéos, obtenues à l'aide d'un système de reconnaissance automatique de la parole. Les expériences sont réalisées sur un corpus composé de dessins animés, de films, de journaux télévisés, de publicités, de documentaires, d'émissions de sport et de clips de musique. L'approche proposée permet d'obtenir un taux de bonne classification de 74% sur cette tâche. En combinant cette approche avec des méthodes reposant sur des paramètres acoustiques bas-niveau, nous obtenons un taux de bonne classification de 95%.

Abstract. In this paper, we present a new method for video genre identification based on the linguistic content analysis. This approach relies on the analysis of the words in the video transcriptions provided by an automatic speech recognition system. Experiments are conducted on a corpus composed of cartoons, movies, news, commercials, documentary, sport and music. On this 7-genre identification task, the proposed transcription-based method obtains up to 74% of correct identification. Finally, this rate is increased to 95% by combining the proposed linguistic-level features with low-level acoustic features.

Mots-clés : classification de genre vidéo, traitement audio de la vidéo, extraction de paramètres linguistiques.

Keywords: video genre classification, audio-based video processing, linguistic feature extraction.

∗. Ces travaux ont été en partie financés par l'Agence Nationale de la Recherche (ANR), par l'intermédiaire du projet RPM2 (ANR-07-AM-008).

1 Introduction

L'indexation vidéo est un domaine porteur dans le contexte actuel où l'on voit se développer les chaînes de télévisions numériques et les collections de vidéos sur internet. L'un des critères les plus utiles pour la recherche de vidéos dans une base de données est sans doute le genre (dessin animé, documentaire, etc.). Son identification automatique a suscité récemment beaucoup d'intérêt de la part de la communauté scientifique. La grande majorité des travaux publiés repose sur l'extraction de paramètres vidéos, tel que la couleur, la luminosité, ou des informations sur les changements de prise de vue (Brezeale & Cook, 2008). Ces descripteurs bas-niveaux sont en général combinés en utilisant des classifieurs statistiques, tel que les machines à vecteurs de support (MVS) ou les modèles de mélanges gaussiens (MMG). L'autre approche classique consiste à extraire des paramètres cepstraux de la piste audio des vidéos (Roach & Mason, 2001). C'est ce que nous avons proposé dans Rouvier *et al.* (2009), en plus d'autres paramètres acoustiques haut niveau.

L'identification automatique du genre est une tâche de catégorisation de texte étudiée depuis longtemps (Karlgren & Cutting, 1994). Ce qui caractérise le genre dans un texte est principalement le style éditorial. Pour les vidéos, le contenu linguistique est de nature parlée mais contient la même information pour caractériser le genre que le texte (Biber, 1988). Malgré cela, très peu de travaux proposent d'utiliser cette modalité pour la classification en genre vidéo. La raison principale est que le contenu linguistique des vidéos est rarement disponible et que l'obtenir revient souvent à faire de la transcription manuelle, qui est très onéreuse. On trouve cependant quelques études qui utilisent les sous-titres associés aux vidéos (Lin & Hauptmann, 2002; Brezeale & Cook, 2006), mais ces approches sont inutilisables lorsque ceux-ci ne sont pas disponibles, par exemple sur les plateformes internet d'échange de contenus vidéos.

Une manière peu coûteuse d'obtenir la transcription des vidéos est d'utiliser un système de reconnaissance automatique de la parole (RAP). La plupart des travaux proposant cette approche concernent la classification thématique et généralement dans un domaine où le système de RAP est performant, ce qui permet d'aborder le problème comme s'ils s'agissait de données textuelles non erronées. Par exemple, Wang *et al.* (2003) proposent d'effectuer la classification automatique de nouvelles issues de journaux radiodiffusés à partir de transcriptions automatiques avec un taux d'erreur mot de l'ordre de 10%. L'investissement nécessaire pour obtenir de telles performances de RAP sur des genres vidéos variés, tel que les dessins animés, documentaires, etc., serait très coûteux.

Dans cet article nous proposons une méthode de classification automatique du genre vidéo reposant sur la caractérisation du style éditorial dans des transcriptions issues de RAP avec un taux d'erreur variable. Nous proposons plusieurs métriques pour sélectionner les termes qui seront fournis au classifieur. Le modèle que nous proposons s'inscrit dans l'architecture de classification multimodale à deux niveaux proposée dans Rouvier *et al.* (2009) : (I) des descripteurs de contenu homogènes (acoustique, linguistique, etc.) qui peuvent être eux-mêmes des classifieurs bas-niveau qui pré-classifient le genre des vidéos et (II) un classifieur global qui prend la décision finale en fonction des descripteurs et des sorties des classifieurs du niveau inférieur.

Dans la première section, nous présentons en détails notre approche pour extraire des paramètres linguistiques des transcriptions automatiques. Ensuite, nous analysons les résultats expérimentaux obtenus avec l'approche linguistique proposée. Finalement, nous présentons et discutons de la complémentarité de ces paramètres avec ceux que l'on trouve dans l'état de l'art.

2 Identification du genre basée sur la linguistique

L'analyse du contenu linguistique des vidéos que nous proposons repose sur l'utilisation d'un système de RAP pour obtenir les transcriptions des vidéos. Ce système utilise un lexique fermé et un modèle de langage qui est estimé sur un corpus textuel de grande taille. Entraîner un tel modèle pour chaque genre vidéo n'est pas réalisable car nous ne disposons pas du volume de données textuelles nécessaires. Nous proposons donc d'utiliser un modèle de langage standard avec un lexique peu adapté à certains genres, ce qui causera la plupart du temps un fort taux d'erreur dans les transcriptions.

Dans la description des méthodes et formules qui suivront, le mot *terme* est employé dans un sens large, proche du sens mathématique, et peut désigner un n-gramme de mots ou d'étiquettes morphosyntaxiques, avec n variant de 1 a 3.

Le principe du *sac de termes* est utilisé pour la modélisation des documents. Selon ce modèle, chaque dimension de l'espace des paramètres représente un terme et chaque document est représenté par un vecteur de fréquences de termes dans cet espace. Dans le cas où le terme est un unigramme de mots, alors ce modèle est un *sac de mots*. La taille du n-gramme utilisé comme terme est appelé l'ordre du modèle.

2.1 Approche classique pour le genre vidéo

Pour les problèmes de catégorisation automatique de texte, les approches généralement proposées reposent sur l'extraction de mots porteurs de sens des documents à classer. Pour la classification du genre vidéo, les études qui s'appuyent sur la modalité textuelle utilisent en général cette approche. Soit les mots-outils de la langue sont filtrés, soit une métrique de type *Term Frequency-Inverse Document Frequency* (TF.IDF) est utilisée pour ne sélectionner que les mots porteurs de sens des documents (Brezeale & Cook, 2006; Lin & Hauptmann, 2002). Cette approche sera notre modèle de référence dans cet article.

Pour un terme t et un document d, TF.IDF est défini comme suit :

$$\text{TF.IDF}(t,d) = \text{TF}(t,d) \times \text{IDF}(t) \tag{1}$$

avec $\text{TF}(t,d)$ la fréquence normalisée du terme t dans le document d et $\text{IDF}(t)$ une métrique représentant le pouvoir discriminant du terme t.

$\text{TF}(t,d)$ est défini comme suit :

$$\text{TF}(t,d) = \frac{n_{t,d}}{\sum_{k \in d} n_{k,d}} \tag{2}$$

avec $n_{t,d}$ la fréquence du terme t dans le document d.

$\text{IDF}(t)$ est défini comme suit :

$$\text{IDF}(t) = \log\left(\frac{N^d}{\text{DF}(t)}\right) \tag{3}$$

avec N^d le nombre total de documents dans le corpus et $\text{DF}(t)$ le nombre de documents du corpus qui contiennent le terme t. Plus la valeur de TF.IDF d'un mot est élevée, plus le mot considéré est représentatif du document et porteur de la thématique qu'il aborde.

Pour chaque genre, le vecteur de paramètres contient la liste de termes résultant de la fusion des n termes ayant les meilleurs TF.IDF de chaque document du genre. Un seul exemplaire de chaque terme est conservé lors de la fusion. Ces vecteurs sont ensuite regroupés dans un supervecteur qui est fourni au classifieur bas-niveau.

2.2 Termes les plus discriminants pour le genre

L'approche classique basée sur TF.IDF permet de donner plus de poids aux termes discriminants pour les documents, qui sont souvent porteurs des thématiques des vidéos. Pourtant, le genre vidéo fait référence au style éditorial et n'est pas systématiquement corrélé avec les sujets abordés et, plus généralement, avec le contenu sémantique.

Nous proposons ici de concevoir une métrique permettant d'identifier les termes discriminants pour le genre et non pour le document. Pour cela, nous proposons d'adapter le calcul de TF.IDF en *Genre Term Frequency-Inverse Genre Frequency* (GTF.IGF), défini comme suit pour un terme t et un genre g :

$$\text{GTF.IGF}(t, g) = \text{GTF}(t, g) \times \text{IGF}(t) \tag{4}$$

avec $\text{GTF}(t, g)$ la somme normalisée des fréquences normalisées du terme t dans les documents du genre g et $\text{IGF}(t)$ une métrique représentant le pouvoir discriminant du terme t.

$\text{GTF}(t, g)$ est définit comme suit :

$$\text{GTF}(t, g) = \frac{\sum_{d \in g} \text{TF}(t, g)}{|g|} \tag{5}$$

avec $|g|$ le nombre de documents dans le genre g et $\text{TF}(t, g)$ la fréquence normalisée du terme t dans le genre g, défini à l'équation 2.

$\text{IGF}(t)$ est défini comme suit :

$$\text{IGF}(t) = \log\left(\frac{N^g}{\text{GF}(t)}\right) \tag{6}$$

avec N^g le nombre total de genres dans le corpus et $\text{GF}(t)$ le nombre de genres du corpus qui contiennent le terme t. Plus la valeur de GTF.IGF d'un mot est élevée, plus le mot considéré est discriminant pour l'identification du genre.

Pour chaque genre, un vecteur est construit avec les n termes ayant le meilleurs GTF.IGF et ces vecteurs sont ensuite concaténés dans un supervecteur, de taille $n \times N^g$, qui est fourni au classifieur bas-niveau.

2.3 Termes les plus fréquentes

Les méthodes précédentes permettent d'identifier des termes discriminants pour un document ou un genre. Ces termes sont souvent porteurs de sens et plutôt rares en général, ils auront donc une probabilité élevée d'être victimes du décalage entre le lexique du système de RAP et celui du document. A l'opposé de ces approches, Stamatatos *et al.* (2000) ont démontré l'efficacité de l'utilisation des fréquences des mots-outils pour identifier le genre écrit. Nous pensons que ces paramètres peuvent tout aussi bien être porteurs d'information pour détecter le genre vidéo. Contrairement à l'approche TF-IDF, celle-ci est indépendante des thématiques des documents, et est donc plus robuste pour classifier des genres comme les news, les documentaires ou les cartoons, qui abordent des thématiques très variées. De plus, les mots outils sont caractérisés par leurs fréquences très élevées et sont donc plus robustes aux erreurs lexicales du système de RAP.

Les n termes les plus fréquents de l'ensemble des transcriptions automatiques des documents du corpus d'entraînement servent ainsi de paramètres au classifieur bas-niveau.

2.4 Généralisation des constructions caractéristiques du genre

Le but des précédentes approches est de capturer des séquences de un ou plusieurs mots qui soient caractéristiques du genre vidéo. Nous proposons de les généraliser en capturant des patrons de construction de phrases. Pour cela, nous pouvons utiliser des séquences d'étiquettes morphosyntaxiques au lieu des séquences de mots comme termes dans les méthodes précédentes.

Une limite à l'utilisation d'étiquettes morphosyntaxiques de cette manière est que les mots sont systématiquement généralisés. Il y a peut-être certains mots au sein d'une classe morphosyntaxique pour lesquels il faudrait conserver la distinction. Nous proposons donc de ne pas étiqueter certains mots et d'utiliser les techniques précédentes pour identifier les séquences intéressantes pour la classification. Ainsi nous obtiendrons des séquences hybrides, contenant des mots et des étiquettes morphosyntaxiques. Nous proposons que les n mots les plus fréquents du corpus d'apprentissage ne soient pas étiquetés. De cette manière, nous capturerons des patrons d'utilisation des mots-outils plus robustes que des séquences de mots. La valeur de n est définie dans la Section 3.3.

3 Dispositif expérimental

Nous allons évaluer notre approche sur deux corpora : un corpus de vidéos appartenant à sept genres pour lesquelles le contenu linguistique est obtenu avec un système de RAP, et un corpus de sous-titres de vidéos appartenant à quatre genres.

3.1 Corpus de vidéos issu de RAP

Ce corpus est composé de 1150 vidéos appartenant à sept genres : *clip de musique*, *publicité*, *dessin animé*, *documentaire*, *journal télévisé*, *sport* et *film*. Elles ont été extraites d'une plateforme d'échange de vidéos sur internet et elle durent de 2 à 5 minutes. Le genre de ces vidéos a été manuellement renseigné. 870 de ces vidéos sont utilisées pour l'entraînement et 280 pour le test. Les sept genres sont également représentés dans le corpus (environ 125 vidéos de chaque genre pour l'entraînement et 40 pour le test). Les vidéos sont en français et nous ne disposons ni des transcriptions de référence ni des sous-titres.

L'approche classique de classification en genre, reposant sur des paramètres cepstraux et des classifieurs de type MMG, obtient 52% de bonne classification sur ce corpus, ce qui correspond aux résultats obtenus par Roach & Mason (2001) sur une tâche similaire.

3.2 Corpus de sous-titres

Ce corpus n'est pas la cible principale de nos travaux, mais il va nous servir a de référence pour identifier les spécificités de l'identification de genre sur les sorties de la RAP. Il est composé de sous-titres issus de vidéos appartenant à quatre genres : *dessin animé*, *documentaire*, *journal télévisé* et *film*. Il contient 1960 documents dont 1400 servent pour l'entraînement et 560 pour le test. Les quatre genres sont également représentés dans le corpus. Les vidéos auxquelles sont associés ces sous-titres durent de 25 minutes à 2h00.

3.3 Systèmes utilisés

La transcription des vidéos est réalisée avec le système de RAP grand vocabulaire du LIA, SPEERAL (Nocéra *et al.*, 2004). Ce système utilise un algorithme A* pour le décodage et des modèles de Markov cachés pour la modélisation acoustique. Le lexique contient 65k mots et le modèle de langage est un 3-gramme estimé sur 200M de mots du journal *Le Monde* et sur environ 1M de mots du corpus d'entraînement de la campagne d'évaluation ESTER. Ce système est destiné à transcrire des journaux radiodiffusés francophones pour lesquels il obtient un taux d'erreur mot d'environ 20%. Il est fort probable que ce taux soit très élevé pour les autres genres (probablement entre 40% et 80%).

L'étiquetage morphosyntaxique des corpus est obtenu automatiquement avec l'outil LIA_TAGG [1], qui fournit un jeu d'étiquettes très détaillé. Il permet notamment de distinguer le genres et le nombre sur les pronoms et les verbes, ce qui va permettre de capturer facilement le point de vue narratif, qui est très différent suivant le genre considéré.

Concernant les modèles hybrides, nous avons fixé à 90 le nombre de mots les plus fréquents qui ne seront pas remplacés par leurs étiquettes morphosyntaxiques.

4 Résultats et discussion

Tous les résultats présentés dans cette section sont obtenus en faisant de la validation croisée sur la réunion du corpus d'entraînement et de test. Le corpus est découpé aléatoirement en 50 parties. A chaque tour de validation une partie est utilisée pour le test, 5 pour le développement et 44 pour l'entraînement. Il faut donc faire 50 cycles pour faire une évaluation. Le classifieur bas-niveau utilisé est de type *Multi-Layer Perceptron* à trois couches (Cybenko, 1989).

Les méthodes que nous avons proposées possèdent quatre axes de variabilité : le type de termes (mots, étiquettes ou hybride), la taille des n-grammes utilisés (l'ordre du modèle), la manière dont les termes sont sélectionnés (TF.IDF, GTF.IGF ou les plus fréquents, TF) et enfin le nombre de paramètres utilisés. Les combinaisons offrant les meilleurs résultats sur le corpus de vidéos issu de la RAP sont présentés dans le tableau 1. Le système de référence sur cette tâche est le modèle TF.IDF d'ordre 1(première colonne de la dernière ligne du tableau 1). Le tableau 2 contient les résultats des quatre meilleures combinaisons des configurations précédentes.

Les résultats présentés dans le tableau 1 montrent que dans tous les cas, augmenter l'ordre des modèles à tendance à dégrader les résultats, les modèles d'ordre 1 étant les meilleurs. De plus, les résultats du tableau 2 ne montrent quasiment pas de gain à combiner des modèles d'ordre différent, ce qui indique que l'information que contiennent les fréquences de séquences de mots ou d'étiquettes est moins importante et surtout redondante avec celle contenue dans les fréquences de ces mots ou étiquettes.

En regardant le nombre de paramètres optimal pour chaque configuration, on observe que l'approche TF.IDF en nécessite un nombre considérable, ce qui indique que le classifieur a besoin des termes les moins bien notés par cette métrique pour fonctionner. Il semble que TF.IDF ne soit pas adapté pour identifier les termes pertinents pour la classification en genre à partir de transcriptions issues de RAP.

Concernant la nature des termes, les résultats reportés dans le tableau 1 montrent que l'utilisation de classes morphosyntaxiques seules ou de séquences hybrides n'améliore globalement pas les performances de classification par rapport à l'utilisation des mots pour les approches TF et GTF.IGF, alors que les modèles TF.IDF d'ordre supérieur à 1 profitent d'une amélioration (pour $n = 2$, 39.5% avec les mots et 59.8% avec

1. http ://pageperso.lif.univ-mrs.fr/∼frederic.bechet

les séquences hybrides). Etant donné le grand nombre de paramètres nécessaires pour l'approche TF.IDF, la généralisation introduite par les classes morphosyntaxiques élimine certainement une variabilité lexical lié à la thématique des documents, au profit d'une meilleure modélisation des expressions caractéristiques du genre vidéo, sous la forme de séquences d'étiquettes morphosyntaxiques récurrentes. De plus, les résultats du tableau 2 ne montrent qu'une amélioration de 0.6% à combiner des modèles qui utilisent des métriques différentes pour sélectionner les termes. On peut en conclure encore une fois que l'information capturée par les différentes métriques est globalement redondante.

On remarque que la méthode GTF.IGF représente une alternative à la méthode classique TF.IDF puisqu'elle offre des performances toujours meilleurs, sauf avec des unigrames de mots où l'on constate une légère perte (71.7% contre 71.9%), mais permet une réduction considérable de l'espace de représentation.

Les résultats présentés dans le tableau 1 montrent que dans le cadre de notre tâche, le meilleur moyen de sélectionner les paramètres pertinents reste de prendre les mots les plus fréquents du corpus. On constate une amélioration d'environ 2% absolus des performances par rapport au système de référence TF.IDF avec uniquement les 90 mots les plus fréquents, tout en réduisant l'espace de représentation de 99.7% (de 34k à 90 mots).

TABLE 1 – Taux de bonne classification (%c) et nombre de paramètres optimaux (#p) obtenus sur le corpus de vidéos issu de la RAP en fonction de la métrique utilisée (TF, GTF.IGF ou TF.IDF), du type de terme (mots, étiquettes morphosyntaxiques ou hybride) et de l'ordre du modèle (n).

		mots			étiquettes morpho.			hybride		
		$n=1$	$n=2$	$n=3$	$n=1$	$n=2$	$n=3$	$n=1$	$n=2$	$n=3$
TF	%c	**73.9**	63.6	55.0	**65.6**	61.1	53.8	**70.8**	61.5	52.9
	#p	**90**	300	700	**50**	200	150	**100**	400	400
GTF.IGF	%c	71.7	60.2	48.9	64.3	62.8	54.3	70.4	60.1	43.3
	#p	200	60	500	40	500	150	60	700	200
TF.IDF	%c	71.9	39.5	29.1	61.2	55.9	52.0	66.1	59.8	42.7
	#p	34k	50k	50k	70	10k	50k	200	50k	50k

TABLE 2 – Taux de bonne classification (%c) obtenus sur le corpus de vidéos issu de la RAP pour les quatre combinaisons de modèles qui offrent les meilleures performances.

combinaison	%c
mots TF n=1 + étiquettes morpho. GTF.IGF n=2	74.5
mots TF n=1 et n=3 + étiquettes morpho. TF n=1 + etiquettes morpho. GTF.IGF n=2	74.1
mots TF n=1 + étiquettes morpho. TF n=1	73.8
mots TF n=1 + étiquettes morpho. TF n=1 + etiquettes morpho. GTF.IGF n=2,3	73.4

Afin de déterminer si ces résultats sont liés à la nature du corpus issu de la RAP, nous avons mené les mêmes expériences sur le corpus de sous-titres qui contient les transcriptions exactes de vidéos de quatre des sept genres. Les résultats sont présentés dans le tableau 3 (colonnes intitulées ST), aux côtés de ceux obtenus sur le corpus issu de la RAP (colonnes intitulées RAP) en ne prenant en compte que les quatre genres considérés. Ces résultats ne peuvent pas être directement comparés étant donné que les corpus utilisés sont différents, mais peuvent fournir une information indicative. On constate que la méthode TF.IDF est la meilleure approche sur les sous-titres suivie de près par TF. Sur le corpus issu de RAP ces résultats sont inversés. On pourrait penser que TF est plus robuste aux erreurs de RAP que TF.IDF. Le résultat le

plus important est que la méthode TF fonctionne très bien sur le corpus de sous-titres, ce qui indique que les fréquences d'utilisation des mots-outils de la langue contiennent une information caractéristique du genre vidéo et qui n'est pas liée au système de RAP, mais bien un phénomène linguistique.

TABLE 3 – Taux de bonne classification (%c) et nombre de paramètres optimaux (#p) obtenus sur le corpus de vidéos issu de la RAP (RAP) et sur le corpus de sous-titres (ST) en fonction de la métrique utilisée (TF, GTF.IGF ou TF.IDF) et du type de terme (mots, étiquettes morphosyntaxiques ou hybride) avec des modèles d'ordre 1.

		mots		étiquettes morpho.		hybride	
		ST	RAP	ST	RAP	ST	RAP
TF	%c	79.0	**89.7**	70.0	82.0	73.0	85.5
	#p	700	**80**	80	40	200	70
GTF.IGF	%c	80.8	87.0	69.7	**82.1**	73.1	**87.1**
	#p	800	80	20	**40**	90	**80**
TF.IDF	%c	**83.1**	88.9	**77.7**	81.7	**79.1**	83.0
	#p	**50k**	34k	**100**	100	**200**	200

Ces résultats valident notre hypothèse initiale : les fréquences des mots-outils contiennent une information permettant de caractériser le genre vidéo. De plus, le modèle TF proposé d'ordre 1 permet un gain de bonne classification absolu d'environ 2% en comparaison du modèle de référence TF.IDF sur le corpus de RAP, alors que l'espace de représentation est réduit de 99.7%. La section suivante présente les résultats obtenus en combinant le modèle de classification que nous proposons avec d'autres utilisant des paramètres audio de plus bas niveau.

5 Complémentarité avec les paramètres acoustiques

Les paramètres linguistiques proposés dans cet article sont basés sur le contenu parlé des vidéos. Dans un précédent article nous avons proposé l'utilisation de paramètres qui ne sont pas directement dépendants du contenu linguistique (Rouvier *et al.*, 2009) et qui permettent d'attendre un taux de bonne classification de 94% sur ce corpus. Dans cette section nous allons évaluer à quel point ces jeux de paramètres sont complémentaires.

La combinaison de descripteurs est réalisée en utilisant un classifieur de type MVS avec pour chaque document un vecteur d'entrée contenant les sorties des classifieurs bas-niveaux de chaque descripteur et les performances sont mesurées sur le corpus de test issu de la RAP. Le modèle de contenu linguistique est TF d'ordre 1, donc les 90 mots les plus fréquents du corpus d'entraînement. Les résultats de ce modèle seul sont présentés dans la colonne notés L du tableau 4.

5.1 Paramètres cepstraux à court-terme

L'espace acoustique représenté par des paramètres cepstraux est le descripteur le plus couremment utilisé dans la classification de genre vidé. Dans Rouvier *et al.* (2009), nous avons montré que 12 coefficients de prédiction linéaire perceptuelle (PLP) et l'énergie du signal avec leurs dérivées première et seconde permettent de bien représenter cet espace et offrent de bonnes performances pour la classification en genre. La variabilité intra-genre est réduite grâce à l'utilisation de l'analyse factorielle. Les performances de ce descripteur de l'espace acoustique sont présentées dans la colonne intitulée EA du tableau 4.

5.2 Interactivité

Le nombre d'intervenants et la manière dont ils interagissent sont souvent dépendants du genre vidéo considéré. Par exemple dans les journaux télévisés il y a généralement un présentateur qui occupe la majorité du temps de parole, contrairement à la plupart des dessins animés et des films. Dans Rouvier *et al.* (2009), nous avons proposé d'extraire des vidéos un descripteur de la manière dont les intervenants communiquent. Pour chaque document, un système de suivi de locuteur [2] est utilisé pour estimer le nombre d'intervenants ainsi que les tours de parole. Le vecteur de paramètres contient trois éléments : la densité des tours de parole, le nombre de locuteurs et le temps de parole de l'intervenant principal. Les performances de ce descripteur d'interactivité sont présentées dans la colonne notée I du tableau 4.

5.3 Qualité de la parole

Cette métrique consiste à mesurer l'adéquation du système de RAP avec les documents à transcrire. Par exemple, les journaux télévisés sont en général bien reconnus par le système de RAP que nous utilisons puisqu'il est destiné à cet usage. Nous proposons d'extraire du système plusieurs informations qui forment le vecteur de paramètres fourni au classifieur : la probabilité *a posteriori* des mots de l'hypothèse retenue, la probabilité *a posteriori* globale de l'hypothèse retenue et l'entropie phonétique. Les performances de ce descripteur sont présentées dans la colonne notée Q du tableau 4.

5.4 Combinaisons

Les résultats de la combinaison des descripteurs précédents avec le descripteur linguistique que nous avons proposé sont présentés dans le tableau 4 (colonnes EA+L, I+L et Q+L). On constate que dans tous les cas, les performances de la combinaison sont meilleures que celles du meilleur des descripteurs individuellement. On peut en déduire que le descripteur linguistique contient une information complémentaire par rapport à l'information apportée par les autres descripteurs.

La colonne notée EA+I+Q du tableau 4 contient les résultats de la combinaison de tous ces descripteurs sans le descripteur linguistique proposé. On observe un gain de performances de 3% absolu par rapport au meilleur descripteur et un gain de 2% absolu par rapport à la meilleure des combinaisons précédentes. La colonne intitulée EA+I+Q+L contient les résultats de la combinaison précédente en ajoutant le descripteur linguistique. Ce descripteur apporte un gain absolu de 1%, soit une diminution relative du taux d'erreur de classification de 16% (de 94% à 95%). Ces résultats montrent que le nouveau descripteur linguistique proposé et les descripteurs acoustiques sont pertinents et surtout complémentaires pour la classification en genre vidéo.

TABLE 4 – Taux de bonne classification [%] obtenus avec les descripteurs proposés et leurs combinaisons.

	L	EA	I	Q	EA+L	I+L	Q+L	EA+I+Q	EA+I+Q+L
%c	74	91	56	53	93	81	81	94	95

2. Nous avons utilisé le système de suivi de locuteur du Laboratoire Informatique de l'Université du Maine, disponible à l'adresse http ://liumtools.univ-lemans.fr

6 Conclusion

Dans cet article nous avons proposé une nouvelle approche pour tirer parti du contenu linguistique de vidéos dans le contexte de leur classification en genre lorsqu'aucune méta-donnée n'est disponible. Le contenu linguistique est obtenu par transcription automatique du canal audio. Différentes méthodes d'analyse des transcriptions ainsi obtenues sont comparées.

Les résultats montrent que, contrairement à l'approche classique en catégorisation de texte qui consiste à se concentrer sur les mots porteurs de sens, l'analyse des fréquences des mots outils permet une caractérisation du style éditorial qui est robuste aux erreurs de RAP. La même approche, appliquée à un corpus de sous-titres montre les bonnes performances de ces descripteurs qui restent, sur du texte correct, légèrement moins bonnes que la méthode classique s'appuyant sur TF.IDF, mais permettent de réduire considérablement l'espace de représentation des documents.

Enfin, les paramètres linguistiques proposés sont complémentaires avec les paramètres acoustiques déjà utilisés dans des systèmes antérieurs. Ils apportent une information que ces derniers ne peuvent probablement pas capturer : en combinant les niveaux acoustiques et linguistiques, on obtient un diminution du taux d'erreur de classification de 16% relatif par rapport à l'acoustique seul. Finalement, cette combinaison permet d'atteindre un taux de bonne classification de 95%, ce qui correspond aux meilleurs résultats publiés sur ce type de tâches avec des systèmes mélangeant descripteurs audio et vidéo.

Références

BIBER D. (1988). *Variation across speech and writing*. Cambridge University Press.

BREZEALE D. & COOK D. (2006). Using closed captions and visual features to classify movies by genre. In *Proceedings of Multimedia Data Mining / Knowledge Discovery and Data Mining*.

BREZEALE D. & COOK D. J. (2008). Automatic video classification : A survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics*, **38**, 416–430.

CYBENKO G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, **2**(4), 303–314.

KARLGREN J. & CUTTING D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the international conference on computational linguistics*, p. 1071–1075.

LIN W. & HAUPTMANN A. (2002). News video classification using svm-based multimodal classifiers and combination strategies. In *Proceedings of the International Conference on Multimedia*, p. 323–326.

NOCÉRA P., FREDOUILLE C., LINARÈS G., MATROUF D., MEIGNIER S., BONASTRE J., MASSONIÉ D. & BÉCHET F. (2004). The LIA's french broadcast news transcription system. In *SWIM : Lectures by Masters in Speech Processing*.

ROACH M. & MASON J. (2001). Classification of video genre using audio. In *Proceedings of EUROSPEECH*, p. 2693–2696.

ROUVIER M., LINARÈS G. & MATROUF D. (2009). Robust audio-based classification of video genre. In *Proceedings of INTERSPEECH*, p. 1159–1162.

STAMATATOS E., FAKOTAKIS N. & KOKKINAKIS G. (2000). Text genre detection using common word frequencies. In *Proceedings of the international conference on computational linguistics*, p. 808–814.

WANG P., CAI R. & YANG S. (2003). A hybrid approach to news video classification multimodal features. In *Proceedings of the International Conference on Information, Communications and Signal Processing (ICICS)*, volume 2, p. 787–791.