Représentation paramétrique des relations temporelles appliquée à l'analyse de données audio pour la mise en évidence de zones de parole conversationnelle

Zein Al Abidin IBRAHIM, Isabelle FERRANÉ, Philippe JOLY

Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier, 31062 Toulouse cedex
{ibrahim, ferrane, joly}@irit.fr

ABSTRACT

The general aim of our work is the automatic analysis of audiovisual document to caraterize their structure by studying the temporal relations between the events occurring in it. For this purpose, we have proposed a parametric representation of temporal relations. From this representation, a TRM (Temporal Relation Matrix) can be computed and analyzed to identify relevant relation class. In this paper, we applied our method on audio data, mainly on speaker and applause segmentations from a TV game program. Our purpose is to analyze these basic audio events, to see if the observations automatically highlighted could reveal information of a higher level like speaker exchanges or conversation, which may be relevant in a structuring or indexing process.

1. INTRODUCTION

Analyser automatiquement des documents audiovisuels en vue d'en dégager la structure ou de les caractériser de façon à pouvoir les rattacher à des catégories de documents est l'une des principales motivations de nos travaux. L'indexation automatique de documents audiovisuels passe d'abord par le développement d'outils dont le rôle est de détecter des caractéristiques primaires sonores ou visuelles (parole, musique, bruit, couleur dominante, quantité de mouvement, …). Les résultats produits sont généralement des séquences de segments dans lesquels la caractéristique primaire recherchée est ou non présente. Ces informations bas niveau doivent ensuite servir de base à la recherche d'événements plus pertinents, utilisés pour la génération de résumé ou l'indexation en vue d'effectuer des traitements plus évolués.

Dans les documents audiovisuels, le temps est une composante essentielle. Notre approche est basée sur l'analyse des relations temporelles qui peuvent être observées entre événements présents dans un document. Compte tenu de l'aspect multimédia des données que nous manipulons, nous nous sommes fixé pour objectif de proposer une méthode générique, qui puisse utiliser tout type d'informations, primaire ou de niveau un peu plus élevé, sonore ou visuelle. Cette méthode d'analyse sera décrite dans la section 2.

Des travaux récents sur les documents audiovisuels ont porté sur la détection et l'analyse de scènes dites conversationnelles. Certains de ces travaux présentés dans [6] sont basés sur des caractéristiques audio seules, voix et parole [2], ou vidéo comme la couleur et la durée des plans, voire même multimodales. Comme dans ces travaux, la parole conversationnelle, dans sa forme plutôt que dans la teneur des échanges entre individus, nous intéresse particulièrement. En effet, la notion de conversation en tant que successions de tour de parole est un type d'événements qui est indissociable de l'aspect temporel. Notre motivation, dans la présentation de cet article, est donc d'étudier comment et en quoi notre méthode peut aider à la détection de zones de parole conversationnelles dans un document audiovisuel. Bien que notre méthode soit définie pour être indépendante du type de caractéristique utilisée, pour cette étude, nous nous focaliserons sur des données uniquement audio et présenterons les résultats de notre analyse dans la section 3. La section 4 présentera la conclusion de ce travail et les perspectives de travaux futurs.

2. REPRÉSENTATION PARAMÉTRIQUE DES RELATIONS TEMPORELLES

L'analyse temporelle du contenu de documents audiovisuels, élargie à tout type de relations temporelles observables entre tout type d'événements, nécessite de disposer d'outils pour représenter et raisonner sur des informations temporelles. Deux classes de modèles temporels existent : ceux utilisant le point comme unité temporelle [5] et ceux basés sur la notion d'intervalles temporels [1]. Cette seconde représentation est la plus adaptée à nos travaux qui reposent sur l'utilisation de segmentations élémentaires d'un même document.

2.1. Segmentation élémentaire d'un document

On définit une segmentation élémentaire comme un ensemble de N intervalles temporels disjoints correspondant à un seul et même type d'événements survenant dans le document traité : présence d'une caractéristique audio (parole, locuteur donné, applaudissement) ou visuelle (personne donnée visible à l'écran). Dans la mesure où des outils de segmentation automatique sont disponibles, nous utilisons les résultats produits par ces outils, sinon, les segmentations sont produites manuellement. Soit **S** une telle segmentation qu'on notera $\mathbf{S} = \{ \mathbf{s_i} \}$ avec $i \in$

[1,N] , $\mathbf{s_i}$ un segment ou intervalle temporel représenté par ses deux extrémités : début ($\mathbf{s_{id}}$) et fin ($\mathbf{s_{if}}$) et noté $\mathbf{s_i} = [\mathbf{s_{id}}, \mathbf{s_{if}}]$.

2.2. Relation temporelle entre deux segments

Soient deux segmentations élémentaires S1 et S2 réalisées sur un même document et telles que :

$$\mathbf{S1} = \{ s_{1i} \}\ i \in [1,N_1] \text{ et } \mathbf{S2} = \{ s_{2j} \}\ j \in [1,N_2]$$

$$\text{avec } \mathbf{s_{1i}} = [s_{1id}, s_{1if}] \text{ et } \mathbf{s_{2j}} = [s_{2jd}, s_{2jf}].$$

La relation temporelle observable entre un couple de segment ($\mathbf{s_{1i}}$, $\mathbf{s_{2j}}$) est alors représentée par trois variables [4] :

$$\mathbf{Lap} = s_{2jd} - s_{1if} \quad \mathbf{DB} = s_{1id} - s_{2jd} \quad \mathbf{DE} = s_{2jf} - s_{1if}$$

Ceci peut également être formulé de la façon suivante : $\mathbf{s_{1i}}$ **R (DE, DB, Lap)** $\mathbf{s_{2j}}$ où **R** est la relation observée entre les deux segments $\mathbf{s_{1i}}$ et $\mathbf{s_{2j}}$ avec pour paramètres avec DE, DB et Lap. En considérant chaque triplet comme les coordonnées d'un point dans un espace tridimensionnel, on obtient une représentation graphique des relations temporelles comme cela sera illustré plus loin.

2.3. Matrice des relations temporelles (TRM)

Si on généralise cela à tous les couples de segments des deux segmentations considérées, et que l'on associe chaque paramètre à une dimension, on peut alors créer une matrice tridimensionnelle dans laquelle chaque élément sera considéré comme un compteur. On pourra ainsi comptabiliser toutes les occurrences d'une même relation. Toutes les relations observables seront déduites de la comparaison deux à deux des segments de S1 avec les segments de S2. Cet histogramme tridimensionnel, pourra permettre d'étudier les fréquences des relations temporelles observables. La matrice pourra également être utilisée comme matrice de vote dont la distribution pourra permettre d'identifier des règles générales relatives au comportement temporel des événements présents dans le document. Avant d'étudier le contenu de la TRM, quelques petites transformations sont nécessaires. Elles concernent l'uniformisation des unités de temps qui peuvent ne pas être les mêmes d'une segmentation à l'autre lorsque celles-ci ne concernent pas le même média. Elles concernent également les valeurs utilisées comme index pour accéder à un élément de la matrice. Au départ réelle, leur valeur doit être ramenée à une valeur entière ce qui permet de réduire la taille de la matrice.

2.4. Classification des données de la TRM

Contrairement aux techniques de vote classiques, identifier une relation temporelle pertinente ne peut pas se limiter à la recherche d'un maximum local, mais plutôt à celle d'une zone dans laquelle les votes seraient majoritairement distribués. Cela revient à localiser des nuages de points dans la représentation graphique ou de votes dans la TRM. Cela nous conduit soit à procéder à une étape de classification du contenu de la TRM ou bien à décomposer la matrice en régions différentes en fonction de connaissances a priori sur la nature des relations que l'on souhaite observer, comme les relations de Allen par exemple. En effet, la définition même des relations de Allen introduit des contraintes qui limitent directement la zone représentative de la relation. A titre d'illustration, la table 1 représente les contraintes imposées par la prise en compte des relations temporelles de Allen *'avant'* et *'après'*.

Table 1: Représentation paramétrique des relations 'avant' et 'après'

Relation	Lap	DB	DE
avant	$0 < Lap \leq \alpha$	$DB < -Lap$	$DE > Lap$
après	$DE-DB < Lap < 0$ et $0<DB -DE+Lap \leq \alpha$	$DB > 0$	$DE < 0$

Lap mesure l'écart entre les deux segments. Si $\mathbf{s_{1i}}$ est avant $\mathbf{s_{2j}}$ alors **Lap** doit être strictement positif. De plus si on veut que les relations observées restent significatives, alors comparer deux segments lorsqu'ils sont trop éloignés l'un de l'autre peut ne pas avoir de sens. Une seconde limite nommée α est alors introduite. Dans la figure 1 ci-dessous on constate que la zone relative à la relation *'après'* est limitée dans l'espace tridimensionnel considéré. Les contraintes relatives aux autres relations de Allen sont données dans [3].

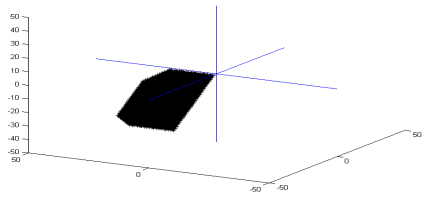

Figure 1 : Représentation graphique de la relation *'après'* avec x = DE, y = DB , z = Lap

Une fois les zones correspondant aux classes de relations identifiées, on leur associe un nombre d'occurrence correspondant à la somme des votes qu'elles contiennent. Ainsi la taille de la matrice est réduite à une matrice cubique, chaque dimension correspondant au nombre de classes ou de relations prédéfinies considérées.

2.5. Combinaison de relations temporelles

Chaque classe de relations temporelles peut correspondre à un type d'événements. Combiner les relations temporelles appartenant à des classes différentes peut être un moyen de mettre en évidence des événements sémantiquement plus significatifs quant au contenu du document.

La conjonction de relations temporelles a déjà été étudiée dans [1]. Notre approche permet de procéder à la conjonction de relations temporelles et de produire un résultat sous la même forme paramétrique. Soient trois intervalles temporels, $\mathbf{s_{1i}}$, $\mathbf{s_{2j}}$ et $\mathbf{s_{3k}}$ appartenant respectivement à trois segmentations **S1, S2 S3**, effectuées sur le même document. Deux relations temporelles **R1** et **R2** peuvent être définies par :

$\mathbf{s_{1i}}$ **R1** (a_1, b_1, c_1) $\mathbf{s_{2j}}$ et $\mathbf{s_{2j}}$ **R2** (a_2, b_2, c_2) $\mathbf{s_{3k}}$

Une nouvelle relation temporelle **R3** résultant de la conjonction des relations **R1** et **R2** pourra être exprimée par : $\mathbf{s_{1i}}$ **R3** (a_3, b_3, c_3) $\mathbf{s_{3k}}$

avec $a_3 = \mathbf{a_1} + \mathbf{a_2}$; $\mathbf{b_3} = \mathbf{b_1} + \mathbf{b_2}$; $\mathbf{c_3} = \mathbf{c_1} - \mathbf{b_2}$

Si **R1** appartient à la classe de relation **C1**, **R2** à la classe de relation **C2** alors une nouvelle classe de relations temporelles **C3** pourra être mise en évidence par l'ensemble des conjonctions pouvant être calculées entre les relations de la première classe et celles de la seconde. Ainsi en procédant hiérarchiquement on peut espérer faire remonter des informations pertinentes et de plus haut niveau sémantique que celles utilisées en entrée du traitement et mieux caractériser le contenu du document traité. Nous avons mené une étude en ce sens pour voir si l'information que l'on peut récupérer est significative et exploitable. Cette étude est présentée dans la section suivante.

3. ETUDE POUR LA MISE EN EVIDENCE DE ZONES DE PAROLE CONVERSASIONNELLE

3.1. Contexte expérimental de l'étude

Notre méthode se veut avant tout la plus générique possible, que ce soit vis à vis du type du document traité, des relations temporelles observées ou d'événements mis en évidence. Pour cet article nous avons utilisé des données audio et notamment des segmentations élémentaires en locuteur effectuées manuellement sur une émission de jeu télévisée d'une durée de trente et une minutes. Huit locuteurs y sont présents et par conséquent, huit segmentations élémentaires, numérotées de 1 à 8 ici, servent de base à cette étude.

3.2. Calcul des TRM par couple de locuteurs

Suivant le principe de construction présenté dans la section 2.3, une TRM a été calculée pour chaque couple de locuteurs distincts, soit 28 TRM. Ces TRM vont être traitées pour voir si l'aspect conversationnel peut être mis en évidence par l'analyse des relations temporelles entre segments. Comme spécifié dans la section 2.4, il est nécessaire de définir des contraintes pour limiter le champ des observations à réaliser. La contrainte α notamment, introduit une limite dans l'écart des couples de segments à comparer. Initialement fixée à 1, soit une seconde d'écart maximum entre deux segments, nous avons obtenu des résultats très pauvres. La limite α a été augmentée (α = 10) pour tenir compte en réalité, des dix secondes de réflexion qui peuvent séparer deux échanges. Cette contrainte est donc, malheureusement, influencée par la nature du document. Pour illustrer le type de résultats que nous avons obtenu considérons les segmentations numéro 2, 3, 4 et 5 ainsi que les TRM associées. La représentation graphique des $TRM_{2,3}$ et $TRM_{4,5}$ sont données ci-dessous figure 2.a et 2.b.

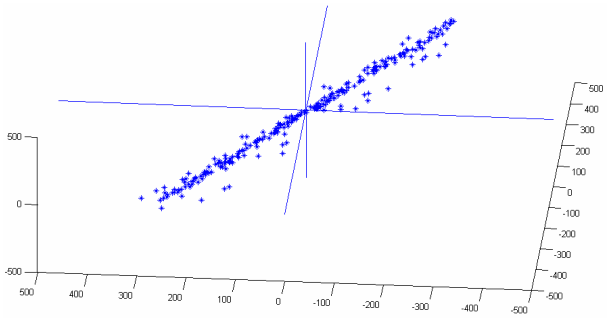

Figure 2.a : Représentation graphique de la $TRM_{2,3}$

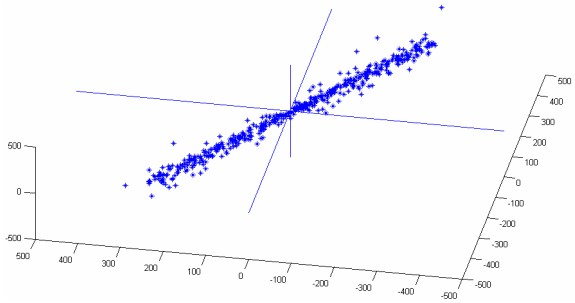

Figure 2.b : Représentation graphique de la $TRM_{4,5}$

Les $TRM_{2,4}$, $TRM_{2,5}$, $TRM_{3,4}$, $TRM_{3,5}$ sont quasiment vides car comparativement, le nombre global de votes pour les relations temporelles observées dans chaque cas est très faible, alors qu'il atteint 247 pour la $TRM_{2,3}$ et 450 pour la $TRM_{4,5}$. En regardant rétrospectivement le contenu du document étudié, et plus particulièrement les locuteurs concernés, on constate que les couples (2,3) et (4,5) correspondent respectivement aux deux équipes impliquées dans le jeu. La nature du jeu veut qu'il y ait plus d'interaction intra équipes qu'inter équipe. Ceci se reflète dans les TRM.

Trois autres observations peuvent être faites d'après l'étude du contenu des TRM. Le locuteur 1 est le seul locuteur à interagir avec chacun des autres locuteurs. Il s'avère que ce locuteur est l'animateur du jeu et son rôle veut qu'il s'adresse à tous. La deuxième observation montre qu'au contraire le locuteur 6 n'est en relation qu'avec le locuteur 1 (cf. observation précédente) et le locuteur numéro 4. On constate effectivement que dans le jeu, le locuteur numéro 6 correspond en réalité à une personne du public avec lequel un des joueurs (le 4), joue une manche spéciale. Enfin, les données des TRM indiquent un nombre plus élevé de votes entre les locuteurs de la seconde équipe (4,5) que ceux de le première (2,3) (cf. plus haut)En effet, une manche consiste pour le joueur d'une équipe

à faire deviner à son équipier autant de mots ou d'expression qu'il le peut dans le temps imparti. La seconde équipe est effectivement la meilleure et la plus rapide, ce qui peut justifier le nombre plus élevé d'échanges détectés.

3.3. Analyse des TRM et classification en deux classes.

Pour aller plus loin dans l'analyse des TRM, nous avons opéré une classification de chacune d'elles en deux classes en utilisant la méthode des k-means. La table 2 présente pour chaque $TRM_{A,B}$ les nombres de votes répartis sur les deux classes C1 et C2.

Table 2 : Classification des TRM en 2 classes

A B	C1	C2	A B	C1	C2	A B	C1	C2
1 2	65	60	**1 3**	49	49	**1 4**	84	71
1 5	106	97	**1 6**	6	5	**1 7**	89	79
1 8	3	5	**2 3**	*123*	*124*	**2 4**	4	7
2 5	6	6	**2 6**	0	0	**2 7**	6	7
2 8	0	0	**3 4**	6	5	**3 5**	10	5
3 6	0	0	**3 7**	7	4	**3 8**	0	0
4 5	*245*	*205*	**4 6**	4	8	**4 7**	15	19
4 8	0	0	**5 6**	0	0	**5 7**	39	26
5 8	0	0	**6 7**	1	0	**6 8**	0	3
7 8	4	3						

Une interprétation possible de la répartition en deux classes, peut être faite en considérant que lorsqu'on examine les locuteurs deux à deux, on a les cas de figure suivants : soit ils ne se parlent pas, et d'un point de vue audio, ils ne seront pas en relation (pas de vote enregistré) soit A parle à B (C1) ou B parle à A (C2). La répartition des votes donnée ici, ne donne qu'une vision globale des échanges entre locuteurs au cours de l'émission. Le recours à l'opération de conjonction doit peut être permettre de détecter les zones de parole conversationnelle.

3.3. Conjonction des relations temporelles

Une conversation peut être considérée ici comme une suite d'échanges entre deux locuteurs où *A parle à B* (A/B) puis *B parle à A* (B/A), ainsi de suite. Faire la conjonction de relations temporelles de la classe C1 avec celles de la classe C2 (resp. C2 et C1) peut-être un moyen de faire apparaître le schéma : (A/B/A) (resp. (B/A/B). Par ce biais, nous avons pu identifier la plupart des séquences d'échanges entre deux locuteurs, les plus longues étant pour les couples (2,3) et (4,5). Nous avons ensuite introduit une nouvelle segmentation, manuelle également, indiquant la présence d'applaudissements. Des TRM associant chaque locuteur (loc_X) à ce nouvel événement (app) ont été calculées et classifiées en 2 classes correspondant aux schémas (loc_X/app) et (app/loc_X). En généralisant l'application de l'opération de conjonction entre les classes appropriées, on peut faire apparaître le schéma (A/B/app) ou (A/B/A/B/app) etc. En effet, un coup réussi dans le jeu correspond à une séquence d'échange entre joueurs de la même équipe close par des applaudissements. Ces motifs, répété à nouveau pourront s'ils sont détectés, correspondre à une manche du jeu. Par contre lorsqu'un coup échoue, aucun applaudissement ne suit, par conséquent notre méthode enchaîne directement avec le coup suivant. La durée des applaudissements pourrait permettre de distinguer entre le gain d'un coup lors d'une manche et le gain de la manche elle même.

4. CONCLUSION ET PERSPECTIVES

Nous avons étudié à travers un exemple comment la représentation paramétrique des relations temporelles entre événements peut conduire après plusieurs phases d'analyse à mettre en évidence des événements d'une teneur plus sémantique que les événements utilisés à la base. A partir d'informations sur les locuteurs présents dans un même document, l'analyse et la classification des TRM obtenues ont fait remonter des informations qui peuvent caractériser les zones de parole conversationnelle, et être indicateurs de la structure du document. Par cette étude nous avons voulu montrer quelle était la nature des résultats que nous pouvions obtenir et quelles étaient les pistes intéressantes à exploiter pour nos travaux futurs. La détermination du nombre optimal de classes dans l'étape de classification et l'application de notre méthode à d'autres types de documents où d'autres types d'échanges sont présents (débats politiques, interviews, émissions de variétés) restent encore à évaluer.

BIBLIOGRAPHIE

[1] J. F. Allen. Maintaining Knowledge about Temporal Intervals. *In Communication of ACM*, Volume 26, number 11, pp. 832 – 843, 1983.

[2] S. Basu. Conversational Scene Analysis. *Ph.D. Thesis. MIT Department of EECS*. September, 2002.

[3] Z. Ibrahim, I. Ferrané, P. Joly. Temporal Relation Analysis in Audiovisual Documents for Complementary Descriptive Information. *In proc. of AMR 2005*, Glasgow, UK, July 2005.

[4] B. Moulin, Conceptual graph approach for the representation of temporal information in discourse. *Knowledge based systems*, volume 5, n°3, pp 183–192, 1992.

[5] M. Vilain, and H. A. Kautz, "Constraint propagation algorithms for temporal reasoning," In AAAI-86, pp. 132-144, 1986.

[6] Y. Zhai, Z. Rasheed, M. Shah. Conversation Detection in Feature Films Using Finite State Machines. *In 17th International Conference on Pattern Recognition ICPR'04*, Volume 4, pp. 458-461, 2004.