Caractérisation des zones d'interactivité entre locuteurs : vers la détection de zones de parole conversationnelle.

Benjamin Bigot *Isabelle Ferrané* *Zein Al Abidin Ibrahim*

IRIT - Université Paul Sabatier
118 Route de Narbonne, F-31062 Toulouse Cedex 9, France
{bigot, ferrane , ibrahim}@irit.fr

ABSTRACT

Our work focuses on the detection and the characterization of conversational speech zones in audio documents. We want to provide enriched annotations of large sets of audio data to people working on tools for conversational speech processing. To adress this issue we adopt a data mining point of view and use a method based on temporal relation analysis. This method enables the detection of zones in which two characteristics are both active and to better characterize them we propose to compute a set of additional descriptors. These descriptors put to the fore some interesting information about speaker profile and give indications on their potential role in audio documents. This method is applied in the scope of the EPAC[1] project.

Keywords: parole conversationnelle, fouille de données, analyse temporelle, annotation enrichie, structuration de contenu audio

1. Introduction.

La parole conversationnelle est une problématique commune à plusieurs domaines de recherche. Il est d'un grand intérêt pour ces communautés de se voir proposer des méthodes et des outils permettant d'explorer des masses de documents audio afin d'extraire puis de traiter ce type de parole. Ce travail préliminaire est nécessaire pour permettre à d'autres spécialités, par exemple, d'améliorer les performances des systèmes de transcription automatique sur ce type de parole difficile à traiter [2], de réaliser une analyse linguistique des phénomènes liés à la spontanéité du langage (répétitions, révisions, ...) comme dans [4], d'extraire et étudier les opinions [5], ou bien encore d'indexer le contenu audio (ou vidéo) sur la base de détections d'éléments structurants.

Dans le cadre de notre étude, nous définissons la parole conversationnelle comme étant *une parole énoncée spontanément au fil d'une conversation, sans préparation.*

Nos travaux se placent dans une optique de structuration de contenu audio, et sont basés sur une approche de type fouille de données puisque nous essayons de faire émerger des connaissances ou des évènements de haut niveau à partir d'une masse de données.

Les corpus sur lequel nous travaillons est celui constitué pour la campagne d'évaluation ESTER [1]. Il s'agit d'un volume important d'émissions de radio (100 heures complètement annotées et transcrites et de 1700 heures brutes) où la parole conversationnelle est présente à travers des interviews, des débats, des réponses aux questions téléphoniques des auditeurs ou éventuellement des questions du public. Explorer les données audio brutes mises à disposition pour en extraire les zones d'interactivité entre locuteurs est une première étape vers l'extraction de zones de parole conversationnelle.

Nous utilisons la méthode [3] qui se veut générique car elle ne dépend pas a priori du type de media (audio ou vidéo), du type de document traité (journaux, magazine, variété,...) ni du type d'évènement recherché. Nous décrivons brièvement cette méthode dans la section 2, et afin d'aller plus loin dans la caractérisation de ces zones, nous définissons un ensemble de descripteurs complémentaires. En section 3 nous appliquons cette méthode pour la détection des zones contenant potentiellement de la parole conversationnelle. Des exemples sont alors donnés dans le cadre de son application aux données du projet EPAC[1].

2. Relations temporelles et zones d'activité.

2.1. Représentation des relations temporelles.

La méthode [3] définit et exploite une représentation paramétrique des relations temporelles. A partir de deux segmentations élémentaires S_1 et S_2 réalisées sur un même document, on calcule, pour chaque couple de segments ($s_{1_i} = [s_{1iD}\, s_{1iF}]$, $s_{2_j} = [s_{2jD}\, s_{2jF}]$, avec $(s_{1i},\, s_{2j}) \in S_1 \times S_2$), trois distances entre les extrémités : DE ($s_{2jF} - s_{1iF}$), DB ($s_{1iD} - s_{2jD}$), Lap ($s_{2jD} - s_{1iF}$). Ces trois paramètres ne sont calculés que si les deux segments sont présents dans la même fenêtre temporelle (telle que $Lap < \alpha = 1$ seconde par exemple). Ainsi, chaque relation temporelle identifiée, $R(DE, DB, Lap)$, met en relation un événement E_1 avec un événement E_2 notés E_1/E_2. Cette relation peut également être représentée par un point dans un espace à trois dimensions (cf. Figure 1) ou comme un élément d'une matrice de vote (appelée Matrice de Relations Temporelles ou MRT). Une MRT peut être calculée pour chaque couple de segmentations disponibles sur un même document.

Considérer une relation temporelle en soi, n'étant pas significatif, nous cherchons ensuite à mettre en évidence des classes de relations temporelles présentes

[1]Les travaux decrits dans cet article ont été partiellement menés dans le cadre du projet EPAC ANR-06-CIS6-MDCA-006

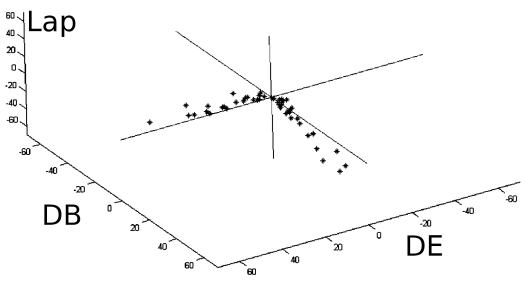

Fig. 1: MRT : exemple d'un dialogue.

dans une MRT. Pour cela, nous appliquons, soit une méthode de classification, soit un jeu de contraintes sur les paramètres.

2.2. Analyse et détection des zones d'activité.

Soient deux classes de relations temporelles (CL_1, CL_2) mises en évidence dans une même MRT, l'application d'une opération de conjonction ($CL_1 \wedge CL_2$) permet d'obtenir un motif du type $E_1/E_2/E_1$ (ou $E_2/E_1/E_2$). Une conjonction suppose la mise en relation de 3 segments s_{1_i}, s_{2_j}, s_{1k} tels que : il existe dans CL_1 une relation temporelle entre (s_{1i}, s_{2j}) et dans CL_2 une relation temporelle entre (s_{2j}, s_{1k}) où s_{1i} et s_{1k} appartiennent à S_1 et $i < k$. En réitérant l'application de l'opérateur de conjonction on peut mettre en évidence des motifs plus longs :

$$CL_1 \wedge CL_2 \wedge CL_1 \wedge CL_2 \wedge \ldots \Rightarrow E_1/E_2/E_1/E_2/\ldots$$

Les zones détectées constituent de nouveaux segments dans lesquels les deux caractéristiques initiales sont actives. Un premier descripteur primaire (*niveau d'activité*) lui est associé et correspond au nombre de conjonctions successivement appliquées. Ainsi, on considère qu'un nouveau type d'événement E_P a été détecté et qu'il faut le caractériser plus finement. Pour cela nous définissons un ensemble de descripteurs complémentaires.

2.3. Caractérisation des zones d'activité.

Une MRT donne une vision globale des co-relations entre deux caractéristiques d'un même document. Aucune information n'est mise en évidence concernant la répartition temporelle de cette activité, son intensité, ni la prépondérance d'une caractéristique par rapport à une autre ou à un sous-ensemble d'autres. Ce sont autant d'indices qui peuvent aider à caractériser le contenu des zones détectées. Soit D la durée d'un document ramenée au nombre d'unités de base (*milliseconde* par exemple) et K le nombre de segmentations élémentaires disponibles $S_1 \ldots S_K$. Une segmentation S_i est liée à une caractéristique C_i et peut être considérée comme une suite de D unités binaires ($u_{d=1\ldots D}$) valant 1 si la caractéristique est active et 0 sinon. Soit T un sous-ensemble de t caractéristiques parmi les K disponibles ($T \subseteq \{C_1, \ldots, C_K\}, C_j \in T$), E_P le type d'événement associé à une zone d'activité dans laquelle un motif P a été identifié, C_p la nouvelle caractéristique active dans cette zone. Plusieurs familles de descripteurs peuvent être définies.

Descripteurs liés à l'activité globale de C_i :
- nombre d'unités actives pour la caractéristique C_i

$$w_{C_i} = \sum_{d=1}^{D} u_{d_{C_i}}$$

- nombre d'unités où au moins une caractéristique de T est active :

$$w_T = \sum_{d=1}^{D} \{ \bigcup_{j \in T} u_{d_{C_j}} \}$$

- nombre d'unités où au moins une des caractéristiques disponibles est active, on a

$$W = \sum_{d=1}^{D} \{ \bigcup_{j=1}^{K} u_{d_{C_j}} \}$$

- nombre d'unités où la caractéristique C_p associée à un motif P détecté est active.

$$w_{C_p} = \sum_{d=1}^{D} u_{d_{C_P}}$$

Ceci permet de préciser quelle est la couverture des segmentations dont on dispose.

Descripteurs liés à la contribution de C_i :
Ce descripteur est utilisé pour mesurer la contribution d'une caractéristique, d'un sous-ensemble T, ou d'un motif P globalement dans le document ou relativement les unes aux autres.
- contribution de la caractéristique C_i dans le document :

$$contrib_{C_i} = \frac{w_{C_i}}{D}$$

- contribution de T, un sous-ensemble de caractéristiques :

$$contrib_T = \frac{w_T}{D}$$

- contribution d'une caractéristique C_p associée à un motif P :

$$contrib_{C_P} = \frac{w_{C_p}}{D}$$

Descripteurs liés à la comparaison de la contribution d'une caractéristique C_i avec d'autres :
- comparaison de la contribution d'une caractéristique C_i par rapport à une autre C_j quelconque :

$$comp_{C_iC_j} = \frac{w_{C_i}}{w_{C_j}} \quad i \neq j$$

- comparaison de la contribution d'une caractéristique C_i par rapport à l'activité d'un sous ensemble T :

$$comp_{C_iT} = \frac{w_{C_i}}{w_T}$$

Descripteurs liés à la répartition de l'activité de C_i :
Afin d'affiner la description d'un caractéristique, et d'analyser la répartition de son activité dans le document, nous divisons un document en parties de tailles égales. Il peut s'agir dans un premier temps d'un découpage arbitraire en trois sections ($Nb_{part} = 3$) et d'étudier ainsi plus précisément ce qu'il se passe dans chaque tiers du document. Par la suite cela peut être généralisé à un nombre de sections fonction de la durée du document analysé

($Nb_{part} = \mathbf{E}(\sqrt{D})$). Soient $d_{\{1,\ldots,Nb_{part}\}}$, les sections considérées, la contribution d'une caractéristique peut être calculée section par section. On obtient alors un vecteur de contribution dans $\mathbf{R}^{Nb_{part}}$. $V_{C_i} = [contrib_{C_{i_{d_1}}}, contrib_{C_{i_{d_2}}}, \ldots, contrib_{C_{i_{d_{Nb_{part}}}}}]$
Cela indique si une caractéristique est présente de manière équilibrée ou si elle est présente ponctuellement et plus localement. Une classification en fonction de cette répartition peut donner des précisons sur le type d'activité.

Descripteurs liés à la répartition des longueurs de segments où C_i est active : On peut également considérer trois catégories de segments (longs, moyens ou courts) ou bien affiner les catégories en fonctions des longueurs maximales de segments observées. Pour chaque caractéristique on visualise comment se répartissent ces longueurs au fil des événements qui se produisent. Ceci peut également être évalué sur le document entier, sur les sections de documents ou bien sur des zones d'activités détectées.
L'ensemble de ces descripteurs peut être défini pour n'importe quelle caractéristique. Dans [3] nous avions montré que l'analyse des MRT donnait des indications sur les rôles des intervenants, mais nous n'avions pas proposé de règles permettant de déduire automatiquement les observations faites alors. Grâce à ces descripteurs, nous avons fait un pas vers une automatisation de l'extraction d'informations de plus haut niveau.

3. Application : Zones d'interactivité entre locuteurs.

Dans EPAC, la parole conversationnelle est définie comme un cas particulier d'interactivité entre locuteurs : *suite d'échanges énoncés spontanément*. Afin de se focaliser sur les zones susceptibles de comporter ce type de parole, nous proposons d'utiliser la méthode générique décrite précédemment en l'appliquant aux segmentations en locuteur disponibles sur les documents qui constituent le corpus d'EPAC. Il s'agit ici de segmentations manuelles de référence et directement disponibles, mais à terme, il s'agira également des résultats de segmentations et regroupements automatiques en locuteur, produits par les autres acteurs du projet.

3.1. Extraction de zones d'interactivité.

L'extraction de ces zones est basée sur l'analyse de chacune des MRT calculées entre deux segmentations en locuteur (et événements associés $E_1 = Loc_1$ et $E_2 = Loc_2$). Un jeu de contraintes est appliqué pour diviser l'espace de représentation des paramètres en trois classes (α étant l'écart maximum entre s_{1i} et s_{2j}) :
- CL_1 : $DE > 0$ et $DB < 0$ et ($Lap \leqslant 0$ ou $0 < Lap < \alpha$) correspondent aux relations temporelles calculées dans le cas où le premier segment commence (et se termine) avant le second segment ;
- CL_2 : $DE < 0$ et $DB > 0$ et ($Lap \leqslant 0$ ou $DB - DE + Lap < \alpha$) correspondant aux cas inverse ;
- CL_3 correspond à un recouvrement complet d'un segment par l'autre.
En procédant par conjonctions successives des classes

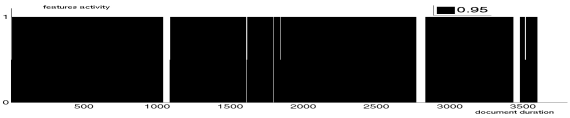

Fig. 2: Activité globale des descripteurs disponibles

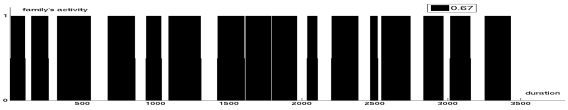

Fig. 3: La caractéristique en zones d'interactivité

CL_1 et CL_2 (CL_3 étant ignorée), on obtient des motifs $Loc_1/Loc_2/Loc_1/Loc_2 \ldots$ La zone d'activité est conservée et annotée comme zone d'interactivité entre Loc_1 et Loc_2. Comme pour le cas général, un nouveau type d'événement a été détecté qu'il faut caractériser plus finement par un ensemble de descripteurs complémentaires.

3.2. Caractérisation des zones d'interactivité.

Nous illustrons chacun des points décrits de manière générale dans la section 2 en prenant comme exemple un fichier représentatif d'une majeure partie du corpus (fichier France Inter Tranche 7/8).

Mesure de l'activité globale des locuteurs : A partir de l'ensemble des segmentations en locuteur disponibles sur le document, on constate que l'activité globale portée sur la figure 2 donne $W = 0.95$, ce qui est cohérent avec le type de document où la parole est naturellement omniprésente.

Mesures sur les zones d'interactivité : La caractéristique C_P est représentée sur la figure 3. Elle correspond ici aux cas particuliers des zones d'interactivité détectées. La contribution globale des zones d'interactivité calculée par $contrib_{C_P} = 0.67$ est relative à la durée du document. Si elle est comparée à l'activité en parole globale, on obtient $contrib_{C_P W} = 0.67 \times 0.95 = 0.63$. On s'intéresse maintenant à la répartition de l'activité des locuteurs. On scinde le document en trois sections et on calcule le vecteur de contribution de chaque locuteur. Afin d'illustrer le type de résultats obtenus, nous représentons ici deux locuteurs caractéristiques.

Répartition de l'activité d'un locuteur Loc_1 : L'activité assez singulière du premier locuteur est représentée sur la figure 5 et correspond au vecteur $V_{Loc_1} = [0.26 \quad 0.31 \quad 0.42]$.

Répartition de l'activité d'un locuteur Loc_2 : Celui-ci présente un profil d'activité totalement différent (figure 6). La totalité de son intervention est localisée temporellement dans la première section du document. Le vecteur associé est $V_{Loc_2} = [1 \quad 0 \quad 0]$
Cela semble un constat simple à faire, mais nous disposons ainsi de mesures qui permettent de comparer automatiquement les interventions des différents locuteurs et de définir des profils différents. En effet le Loc_1 est en fait l'animateur de la tranche matinale,

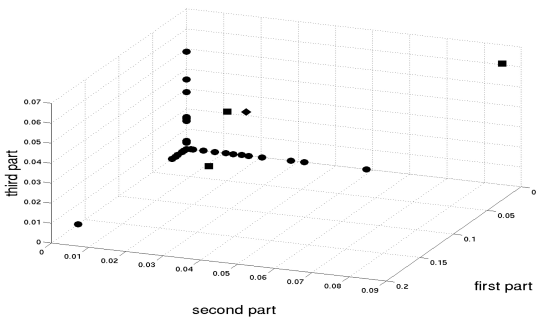

Fig. 4: Répartition des locuteurs.

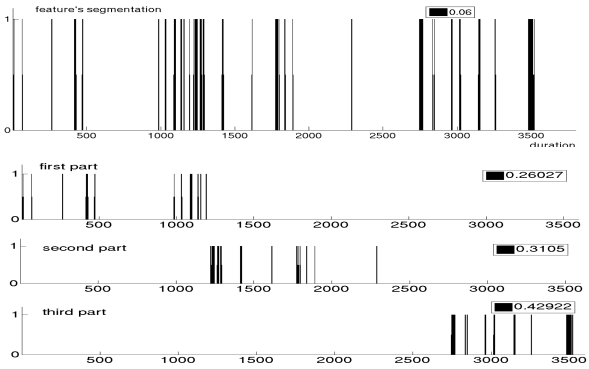

Fig. 5: Répartition pour le locuteur Loc_1

tandis que le second Loc_2 est un chroniqueur ponctuel.

Espace des répartitions : Nous calculons un deuxième jeu de vecteurs définissant un nouvel espace de répartition. Ce nouveau vecteur $(w_{Loc_i} \times {}^tV_{Loc_i})$ correspond à la répartition de chaque locuteur pondérée par son activité globale sur le document.(cf. figure 4). On constate que les interventions localisées sur une seule partie du document (exemple Loc_2) se trouvent sur les axes (•), les interventions localisées sur deux parties se trouvent sur les plans (□) et que l'intervention qui se trouve uniformément répartie sur le document (Loc_1 ⋄)) est le seul point qui se démarque.

Evolution des longueurs de segment : On peut projeter chacun des événements de façon à visualiser l'évolution de leur durée(cf fig 7). Une alternance de segments courts du Loc_1 et de segments longs du Loc_2, pourraient marquer une interaction très préparée donc peu conversationnelle. Une analyse plus approfondie doit permettre d'exploiter cette information et de l'utiliser comme critère discriminant de zones potentiellement conversationnelles.

4. Conclusion et Perspectives

Dans cet article nous avons présenté une méthode pour la détection et la caractérisation des zones particulières entre locuteurs appliquée aux données du projet EPAC. L'approche fouille de données permet de faire émerger des informations qui peuvent servir de base à la définition de profils d'interactivité entre locuteurs. Ceux-ci permettent d'indiquer si la zone détectée est potentiellement conversationnelle. C'est un

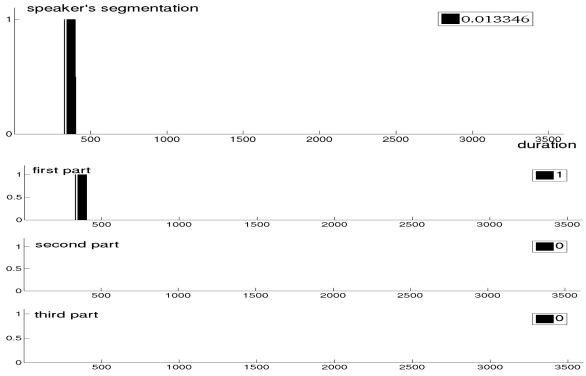

Fig. 6: Répartition pour le locuteur Loc_2

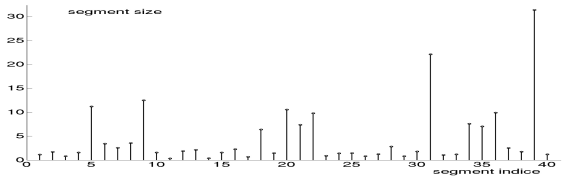

Fig. 7: projections des longueurs de segments pour Loc_1

premier pas vers la structuration du contenu du document, même s'il reste à appliquer ces traitements à des résultats de segmentation automatique et à mesurer l'impact des erreurs de segmentations sur les résultats obtenus. Ce travail d'évaluation fait partie de la suite du projet EPAC. Un autre intérêt que nous voyons également dans la mise en évidence de profils est qu'ils sont aussi "transversaux" dans le sens où des profils similaires peuvent se retrouver d'un document à l'autre. Explorer cette voie nous permettra d'avancer sur la structuration en collection de documents.

Références

[1] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.F. Bonastre, and G. Gravier. The ESTER phase II evaluation campaign for the rich transcription of french broadcast news. In *Interspeech/Eurospeech*, September 2005.

[2] J-L. Gauvain, G. Adda, L. Lamel, F. Lefèvre, and H. Schwenk. Transcription de la parole conversationnelle. In *TAL.*, volume 45. Association pour le traitement automatique des langues, Paris, France,(revue), 2004.

[3] Zein Al Abidin Ibrahim. *Characterization of audiovisual structures by statistical analysis of temporal relations*. Phd thesis, Paul Sabatier University, Toulouse, France, 2007.

[4] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, M. Ostendorf D. Hillard, M. Tomalin, P.Woodland, , and M. Harper. Structural metadata research in the ears program. In *ICASSP*, 2005.

[5] A-M. Popescu and O. Etzioni. Extracting product features and opinions from reviews. In *HLT '05*, pages 339–346, 2005.