

Recherche d'information multimédia

17 octobre 2014
Grenoble

Georges Linarès, LIA
Université d'Avignon et des Pays de Vaucluse



L'audio pour la RIM

Préambule

- *Un film est un mélange de cinq formes d'expressions* *:
 - *L'image photographique mouvante, la parole, le bruit, la musique, les mentions écrites*
- **Objectif de la présentation :**
 - Audio, parole
 - Présenter quelques problèmes clefs
 - Survol des bases théoriques et technologiques
 - Evaluation

* AUMONT Jacques et al. *Esthétique du film*. Paris : Nathan, 1983

L'audio pour la RIM

Plan

- La RIM *par* l'audio : introduction
- (1) Les problèmes clefs :
 - Segmentation, caractérisation, indexation, ré-éditorialisation
- (2) Les bases scientifiques & technologiques
 - Indexation des contenus parlés
 - Indexation audio « bas-niveau »
- (3) Une évaluation : le défi REPERE



L'audio pour la RIM

Introduction

- L'audio dans la RIM et la RIM *par* l'audio
 - Parole :
 - contenus linguistiques
 - Extra-linguistiques : locuteur, émotions, sémantique etc.
 - Audio : événements acoustiques, environnements, etc.
 - Musique : *Music Information Retrieval (ISMIR)*
- *Documents audio*, **documents multimédias**
 - Il en s'agit pas de structurer la partie audio mais de structurer **par l'audio** ou **avec l'audio**

L'audio pour la RIM

Partie I : les problèmes clefs



L'audio pour la RIM

- Les tâches de base
 - Segmenter : partitionner le flux en segments homogènes
 - Caractériser, indexer : produire des représentations expressives, utilisables, compactes des contenus
 - Détecter : dans un flux, dans une collection
 - Générer : présenter les résultats de la recherche



L'audio pour la RIM

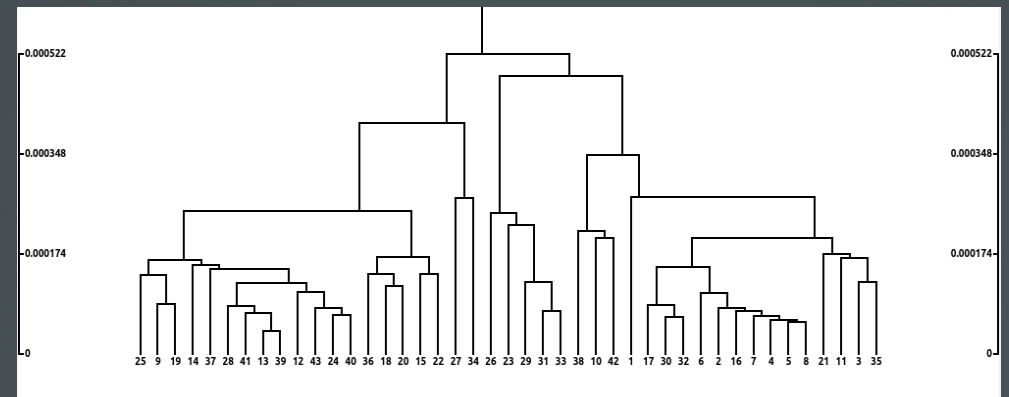
- Segmenter :
 - En macro-classes acoustiques : parole/musique
 - **En locuteurs**
 - **En thèmes**
 - En genre
 - Scènes, *Histoires* ?
 - ...



- Locuteur
 - Diarization : nombre et identité des locuteurs inconnus



- Les techniques de base :
 - Détection de changement de locuteur+regroupements
 - Clustering hiérarchique, Programmation linéaire en nombre entiers



Segmentation

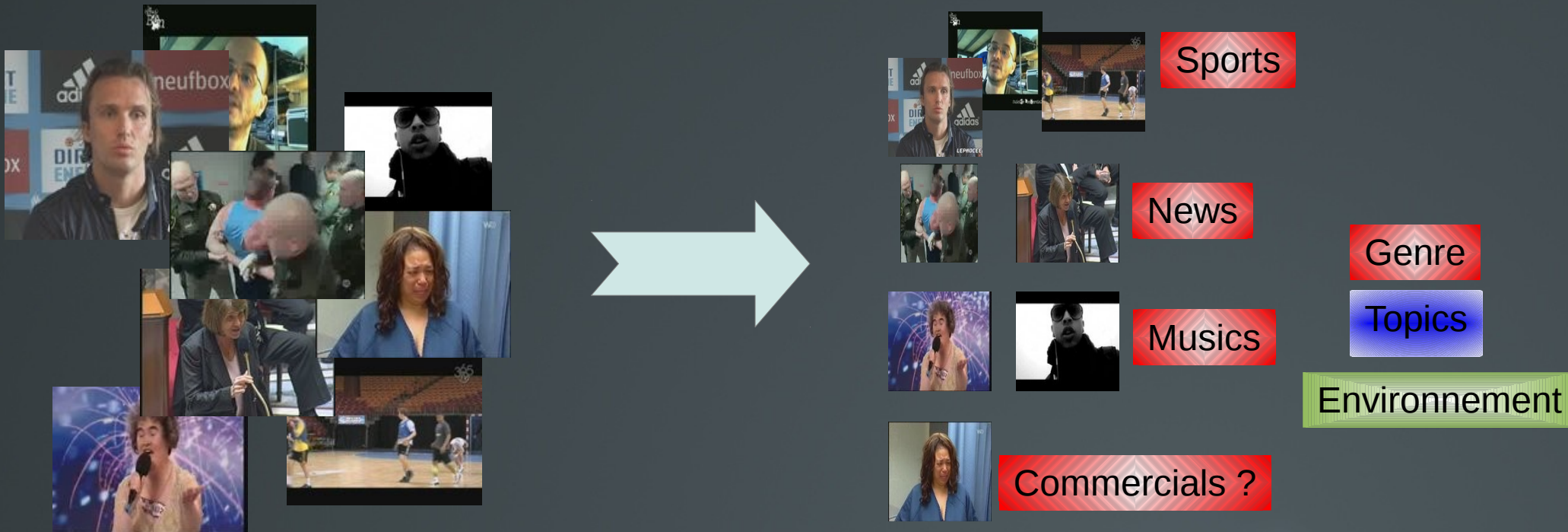
- Segmentation thématique
 - Objectif : segmenter en fonction du contenu sémantique
 - Verrous : problème lié à la parole et à sa transcription
 - Approches classiques :
 - Cohésion lexicale, modèles thématiques (LDA/LSA)
 - Utilisation de graphes de reconnaissance, de mesures de confiance, etc.

Caractérisation, indexation

- Caractérisation des segments
 - Qui parle ? Quel type d'environnement ? Quelle langue ? Quel genre ?
- Extraire la structure des flux et des collections
 - Les graphes d'interactions entre locuteurs
 - Intérêt, saillance, etc...



Caractérisation, indexation Genre vidéo



Caractérisation, indexation

Genre vidéo

- Approches standards :
 - Apprentissage automatique, MFCC/GMM
 - Combinaison audio/vidéo
- Plus récemment :
 - Combinaison contenus parlés (texte), interactions, qualité de la voix, acoustique bas-niveau



- Détection de termes parlés (*Spoken Term Detection*)
 - NIST 2006, recherche dans les treillis/réseaux, de mots, de phonèmes
- De thèmes (*Topic Spotting*)
- Suivi de locuteurs : (*Speaker Tracking*)
- De formes (*Pattern spotting*)
 - Recherche sans connaissance a priori
 - Mediaeval 2012, 2013 : Spoken Web Search

Détection de termes

- Détection de termes parlés (*Spoken Term Detection, STD*)
 - Ça n'est pas du Word-spotting :
 - fouille de grandes bases transcrites
 - Difficultés :
 - erreurs de transcription, mots hors vocabulaire, indexation
 - Approches standards :
 - recherche de séquence de phonèmes, indexation de graphes de phonèmes, réseaux de confusion

Résumé multimédia

- Résumé audio
 - Résumé texte en audio
 - Techniques issues du résumé texte
 - Maximal Marginal Relevance
 - Programmation Linéaire en nombres entiers
- Résumé Multimédia
 - Projet RPM2
 - Résumé vidéo par extraction
- Problème central : l'évaluation

L'audio pour la RIM

Partie II : la boîte à outils



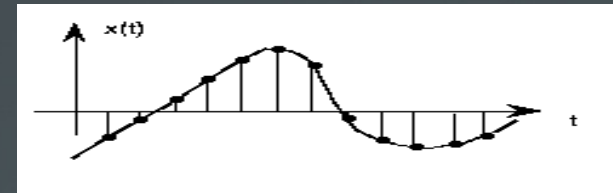
La boîte à outils

- Paramétrisation

Signal analogique (acquisition)

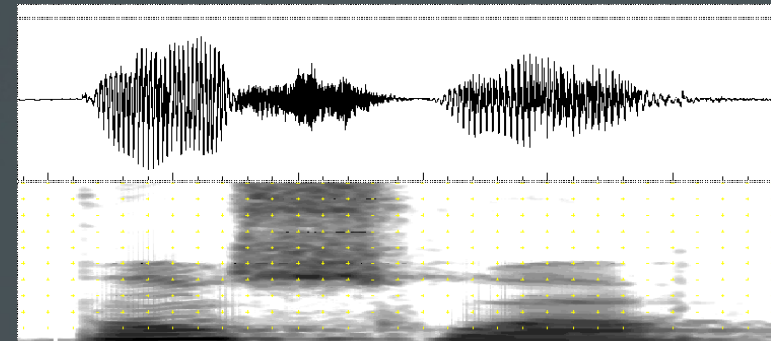


Digitalisation, quantification



Représentation 2D :

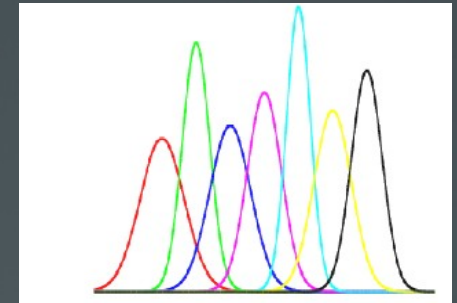
- Analyse dans une fenêtre glissante
- Paramètres basés sur des modèles perceptifs
- MFCC, PLP, LPC



- Gaussian Mixture Models
 - Approxime n'importe quelle fonction de densité de probabilité (de formes cepstrales)

$$l(x|\mu, \Sigma) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

$$l(x|Gmm_k) = \sum_{i=0}^N w_k \cdot l(x|\mu_k, \Sigma_k) \quad \lambda_k = (\mu_k, \Sigma_k)$$

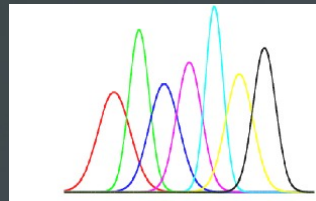


- Méthode standard pour la caractérisation audio
 - Souvent associée avec la MFCC



La boîte à outils

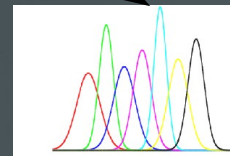
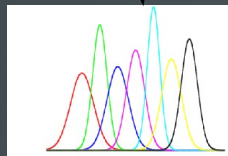
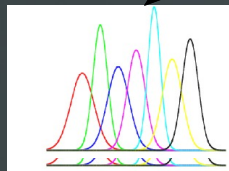
- Reconnaissance du locuteur
 - Différentes tâches : segmentation, suivi, identification, vérification...



Modèle du monde :
Universal Background Model

Adaptation des paramètres de l'UBM

Modèles dépendants des locuteurs

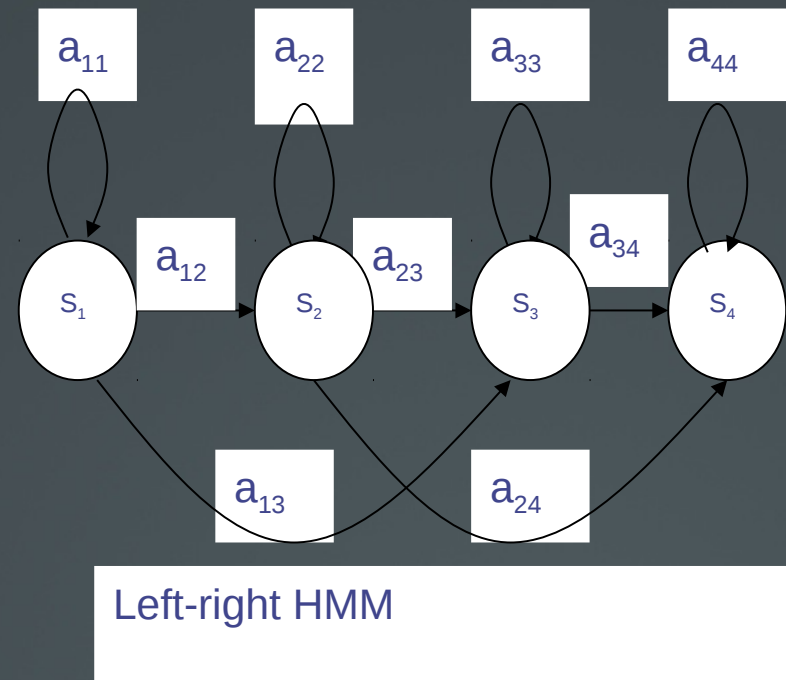
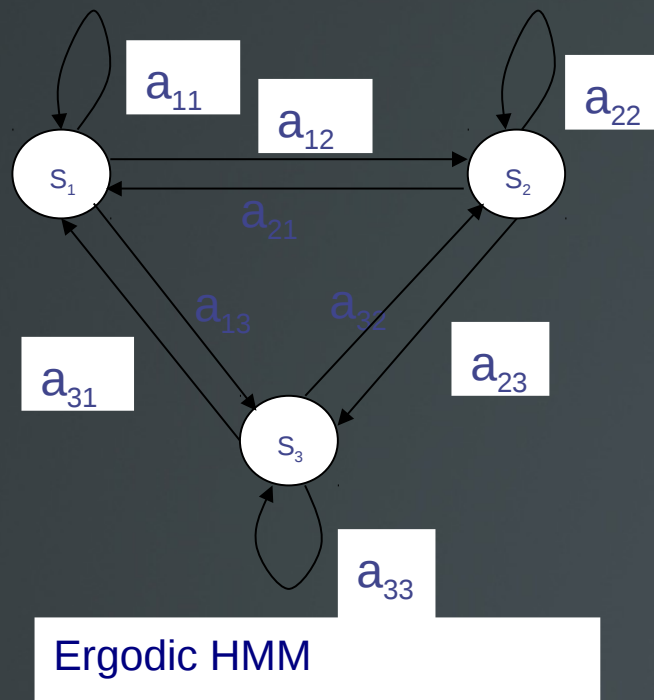


La boîte à outils

- Identification du locuteur : évolutions récentes
 - Analyse factorielle :
 - Hypothèse : la variabilité « nuisible » se concentre dans des sous-espaces de dimension réduite
 - Principe : estimation (supervisée) de la décomposition optimale
 - *I-Vector* :
 - Super-vecteur réduit par SVD+FA

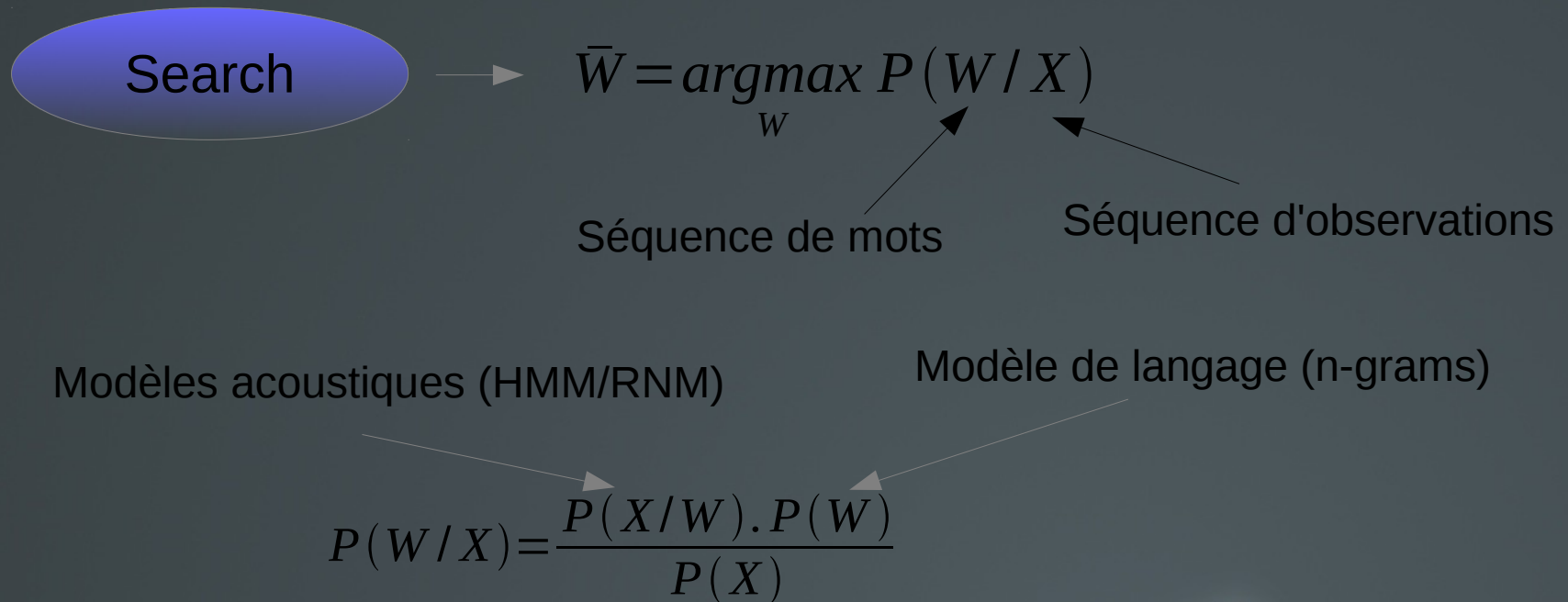


Modèles de Markov cachés (HMM)



La boîte à outils

Indexation de contenus parlés : Reconnaissance de la parole



La boîte à outils

- Reconnaissance de la parole
 - Apprentissage acoustique :
 - Centaines d'heures de parole annotée, apprentissage automatique (estimation des GMM ou Réseaux de neurones)
 - Estimation des modèles de langage :
 - Plusieurs milliards de mots, modèles n-grammes ou réseaux de neurones
 - Evolutions : Deep Learning



La boîte à outils

- Démonstrateur EDD (*Européenne des Données*)
 - Objectifs :
 - *Audio search* dans une grande quantité de flux audiovisuels (100 max)
 - Latence inférieure à 2 minutes
 - Technologie : Segmentation en ligne, Reconnaissance de la parole, système parallèle
 - Recherche par mots clefs !



Partie III : le défi REPERE



Indexation de personnes

Le défi REPERE

- Organisé par l'ANR et la DGA
 - 2010-2014
- Recherche de personnes dans les flux audiovisuels
- 3 consortiums :
 - QCOMPERE : LIMSI, Vocapia Research, KIT, LIG, INRIA Grenoble, YACAST, GREYC
 - SODA : LIUM , IDIAP
 - PERCOL : LIF, LIA, Orange, LIFL

Indexation de personnes

Le défi REPERE

- Tâche : recherche de personnes dans les flux audiovisuels
 - On cherche la présence i.e. les personnes visibles et/ou audibles
 - Sources d'informations possibles : voix (locuteur), contenus parlés, cartouches, visage, costumes, type de scènes, camera
- Verrous : quelle information, quelles stratégies de fusion multimodale



Indexation de personnes

Le défi REPERE

Jean-Remi Baudot : ... Minister of Ecology **Nathalie Kosciusko-Morizet**, has ... **Philippe Salvador** you are ...

Philippe Salvador : ... come, **Nathalie Kosciusko-Morizet** ... **Roland Vierende**, who is the ...

Roland Vierende : the incident is closed at this ...

Jean-Remi Baudot : Thank you **Philippe** from ... who was with **Romain Garrigues** ...

ALERTE INFO FONDOS AFRICAINS Omar Bongo "a financé la campagne électorale de Jean-Marie Le Pen en 1988", a affirmé Robert Bourgi sur BFMTV.

PHILIPPE SALVADOR CIVILITE SPECIALE A GARD (GARD)

ROLAND VIERNE REPRESENTANT DE LA DIRECTION DE SOCODEI

ALERTE INFO FONDOS AFRICAINS Omar Bongo "a financé la campagne électorale de Jean-Marie Le Pen en 1988", a affirmé Robert Bourgi sur BFMTV.

Le CESE form adresses à P pouvoirs public de quatre orien poursuivre le loration de la accentuer l'effi

Le CESE form adresses à P pouvoirs public de quatre orien poursuivre le loration de la accentuer l'effi

Le CESE form adresses à P pouvoirs public de quatre orien poursuivre le loration de la accentuer l'effi

Le CESE form adresses à P pouvoirs public de quatre orien poursuivre le loration de la accentuer l'effi

Identification non-supervisée de personnes dans les flux télévisés,
Thèse de Yoann Poignant, UJF, Grenoble, 2013

Indexation de personnes

Le défi REPERE

- 2 modes
 - non supervisé : pas de modèles a priori de personnes
 - Supervisé : modèles pré-estimés
- Tâches secondaires
 - Détection visages, reconnaissance de la parole, identification du locuteur, détection de noms de personnes dans les transcriptions



Indexation de personnes

Le défi REPERE

- Résultats
 - Corpus : 47 heures annotées
 - TV (BFM TV, LCP Info, Culture et vous,, TOP Questions...)

Equipe	EGER-Supervised	EGER-Unsupervised
PERCOL	28.1	34.3
QCOMP	35.3	43.9
SODA	39.4	48.1

G. Bernard, O. Galibert, J. Khan : *The second official REPERE evaluation*
SLAM 2014, Penang, Malaisie.

Indexation de personnes

Le défi REPERE

- Les conclusions de REPERE :
 - Les modalités locuteur et visage sont cruciales
 - Intégrer des informations relatives à la structure des émissions est décisif
 - Il faut produire des hypothèses de présence de personnes "en situation"
 - Costume, camera, type d'émission, rôle, etc.



Références



Références

- *Identification non-supervisée de personnes dans les flux télévisés*, Thèse, Yoann Poignant, UJF, Grenoble, 2013
- *Extraction multimodale de la structure narrative des épisodes de séries télévisées*, Thèse Philippe Ercolerssi, UPS, Toulouse, 2013
- *Comparaison et combinaison de rendus visuels et sonores pour la conception d'interfaces homme-machine : des facteurs humains aux stratégies de présentation à base de distorsion*, Thèse Tifanie Bouchara, UPS, Paris, 2013
- *Structuration automatique de flux télévisuels*. Thèse Camille Guinaudeau INSA Rennes, France, Décembre 2012.
- *Structuration de contenus audio-visuel pour le résumé automatique*, PhD, Mickael Rouvier, UAPV, Avignon, 2011
- *Approches semi-automatiques pour la recherche d'information dans les documents audio*, Thèse G. Senay, UAPV, Avignon, 2011
- *Navigation dans les documents audio par le résumé automatique*, PhD, Benoit Favre, UAPV, Avignon, 2007

