Interfaces de navigation dans des contenus audio et vidéo

Géraldine Damnati
(1)France Telecom, Orange Labs, Lannion
geraldine.damnati@orange.com

RESUME

Deux types de démonstrateurs sont présentés. Une première interface à visée didactique permet d'observer des traitements automatiques sur des documents vidéo. Plusieurs niveaux de représentation peuvent être montrés simultanément, ce qui facilite l'analyse d'approches multi-vues. La seconde interface est une interface opérationnelle de "consommation" de documents audio. Elle offre une expérience de navigation enrichie dans des documents audio grâce à une visualisation de métadonnées extraites automatiquement.

ABSTRACT

Navigation interfaces through audio and video contents

Two types of demonstrators are shown. A first interface, with didactic purposes, allows automatic processing of video documents to be observed. Several representation levels can be viewed simultaneously, which is particularly helpful to analyse the behaviour of multi-view approaches. The second interface is an operational audio document "consumption" interface. It offers an enriched navigation experience through the visualisation of automatically extracted metadata.

MOTS-CLES : Traitements multi-vues, navigation enrichie.
KEYWORDS : Multi-view processing, enriched navigation.

1 Interface didactique

Il s'agit d'un démonstrateur qui permet d'illustrer les traitements automatiques réalisés sur des contenus vidéo. Le principe est de visualiser sous forme de timeline des informations de structuration extraites automatiquement. Pour chaque segment, un onglet permet de visualiser des informations issues du canal audio (typiquement la transcription automatique synchronisée avec le player) et un onglet permet de visualiser des informations liées au canal vidéo (typiquement des images clé ou *key frames*). L'interface offre des fonctionnalités de navigation d'un segment à l'autre. Au-delà de ces fonctionnalités de base, l'intérêt de l'outil est de pouvoir cumuler plusieurs timeline et observer ainsi l'apport de traitement multi-niveaux. Plusieurs résultats de travaux de recherche seront montrés via cette interface.

Reconnaissance du rôle du locuteur

La capture d'écran ci-contre représente une analyse en rôle des tours de parole dans des Journaux Télévisés. Elle illustre une approche multi-vue qui consiste à fusionner une analyse purement acoustique modélisant l'intonation

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 5: Démonstrations, pages 3–4,
Grenoble, 4 au 8 juin 2012.

des locuteurs en fonction de leur rôle et une analyse purement linguistique basée sur une analyse de la transcription automatique du contenu parlé (Damnati et Charlet, 2011). L'interface permet de visualiser les résultats de chacune des analyses ainsi que de leur fusion, afin de mieux analyser leur complémentarité.

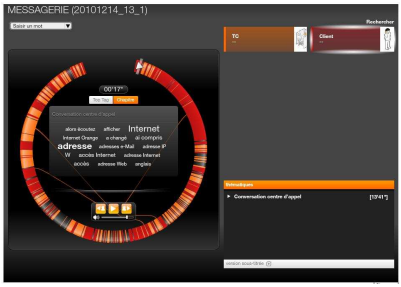
Détection de personnes dans des documents vidéo

Les travaux réalisés dans le cadre du défi REPERE (Béchet *et al.*, 2012) seront également montrés via cette interface. Ce projet a pour but d'identifier les personnes dans des contenus télévisés en exploitant conjointement le canal audio (contenu parlé et analyse en locuteurs) et le canal vidéo (texte incrusté et analyse de visages). L'interface permet de visualiser les informations extraites dans les différentes modalités ainsi que le résultat de la fusion.

2 Interface de navigation enrichie

Cette interface à pour vocation de proposer aux utilisateurs une expérience de navigation enrichie dans des contenus purement audio, en s'appuyant sur des métadonnées produites automatiquement. Elle propose en quelque sorte de "visualiser" des contenus audio. Elle est déclinée à Orange Labs dans différents domaines, allant de la consommation de podcast de radio à l'écoute de conversations issues des centres d'appels.

La capture d'écran ci-contre illustre une interface de visualisation conversations client/téleconseiller, et s'inscrit dans le domaine plus large du *Speech Analytics*. Elle permet d'avoir une vue synthétique du déroulé de la conversation, structurée en locuteurs, une visualisation d'expressions clés extraites des transcriptions automatiques, un filtrage des conversations par motif d'appel, etc…



Références

DAMNATI, G., CHARLET, D. (2011). Multi-view approach for speaker turn role labeling in TV Broadcast News shows, *Proc. Interspeech'11,* Florence, 2011.

BECHET, F., AUGUSTE, R., AYACHE, S., CHARLET, D., DAMNATI, G., FAVRE, B., FREDOUILLE, C., LEVY, C. (2012). Percol0 - un système multimodal de détection de personnes dans des documents vidéo. *Proc. JEP'12*, Grenoble, 2012.