E-Inclusion research network Theme 2 : Speech and Audio Context Recognition

Presented by Gilles Boulianne

CRIM

Presented at CRIM
March 31th, 2008

© CRIM 2008

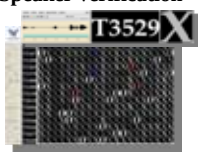
E-Inclusion Theme 2 Project Map

THEME 2
Speech and audio context recognition (Boulianne, CRIM)

PROJECT 2.1
Core speech technology improvements (Boulianne, CRIM)

- Integrated into live captioning product and services
- Real-time correction

- Segmentation and tracking
- Software tools and GUI
- French broadcast news corpus
- Speaker verification



PROJECT 2.2
Speech and audio context recognition (Boulianne, CRIM)

Cultural speech and audio context recognition (Boulianne, CRIM)

Emotional content (Dumouchel, ÉTS)

- Feature extraction, emotion identification

Sync with actor Discourse (Dumouchel, ÉTS)

- Levenstein distance PID controller
- Proof of concept, demo

Lecture speech (Rose, McGill)

- Recognition and indexing
- Web site demo

PROJECT 2.3
Anything, anywhere Live captioning (Boulianne, CRIM)

- Live captioning of large conferences and events
- Mobile captioning apparatus

PROJECT 2.4
Course lecture E-accessibilty (Rose, McGill)

- Prototype of e-Learning site
- COOL collection of lectures

PROJECT 2.5
Collaborative Captionning (Boulianne, CRIM)

Prototype collaborative Web site

- Tools for creation/involvement

PROJECT 2.6
Assisted indexation (Gupta, CRIM)

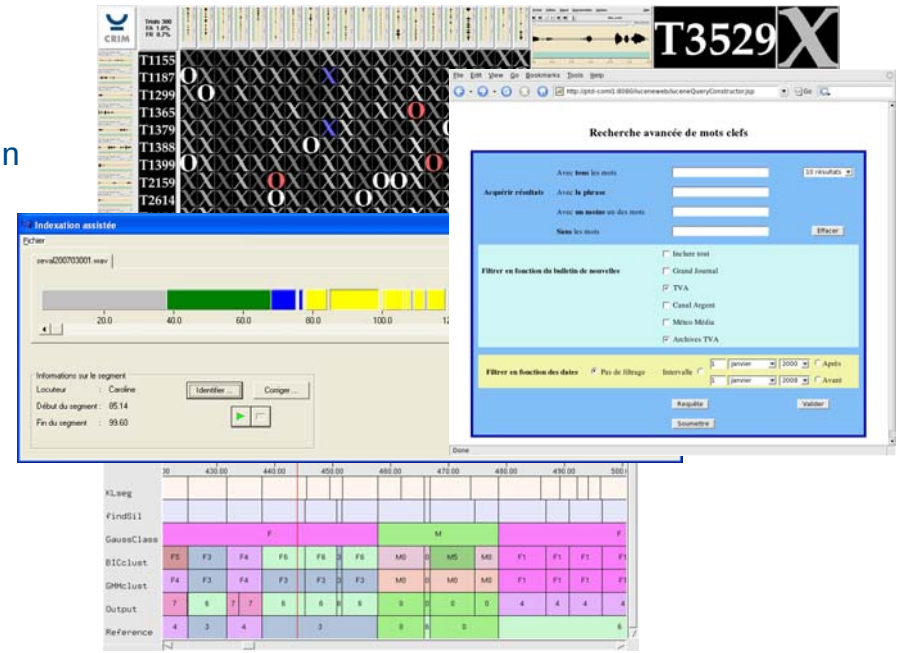
- Prototype of assisted indexation system
- Canadian multimedia searchable archive

Started
Not started
Completed

2.6 Assisted indexation

- Combine speech technologies into a prototype to amplify the human ability to identify spoken content, in order to dramatically reduce the time-intensive manual task of indexing large audiovisual archives.

  - Audio content indexation
    - Advertisement detection
    - Speech / music / noise classification
    - Automatic text alignment
    - Speaker segmentation
    - Speaker recognition
    - Emotion detection
    - Word and phoneme recognition
  - Audio content retrieval
    - Approximate matches:
      vector space vs statistical models

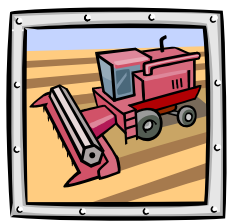


2.6 Assisted indexation

- Previous work (E-Inclusion phase 1)
  - Speech/music/noise
  - Speaker detection, identification and tracking
  - French television broadcast corpus
- New work (E-Inclusion phase 2)
  - Speaker detection, identification and tracking on French TV corpus
  - Assisted speaker labeling
  - Advertisement detection
  - Approximate match incorporated into Lucene text search engine

2.6 Assisted speaker labeling

- Assisted speaker labeling
  - Automatic segmentation into speaker turns and ids
  - Human labeling and corrections
  - Reference clusters propagated to segment new data



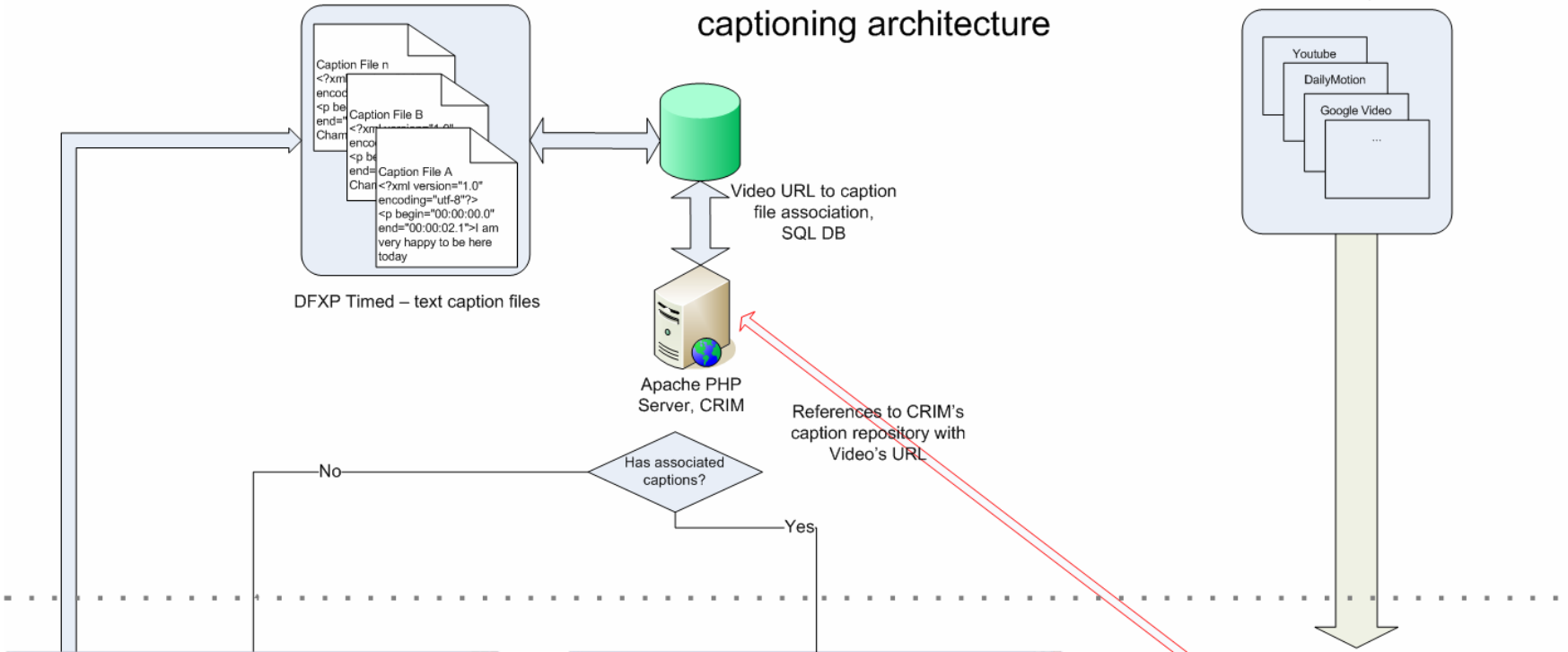
2.6 Indexation and search engine

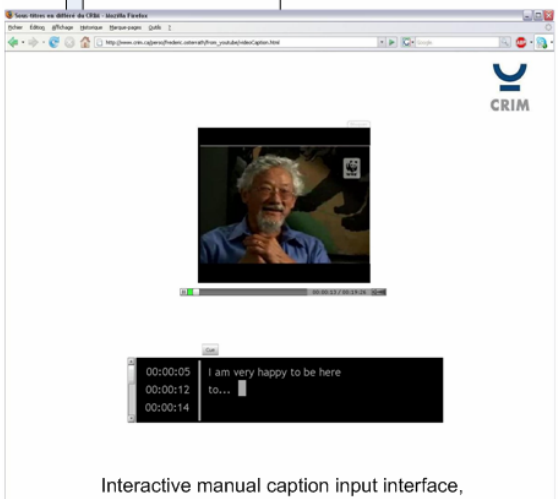
- Text-based search engines
  - Fast and efficient
  - Query processing time: independent of database size
  - Vector-space model (TF-IDF, inverse index)
- Approximate match searching
  - Recognition errors: statistical models
  - Query processing time: proportional to database size (e.g. Nexidia limited to 4000 hours)
  - Integrate TF-IDF and statistical scoring
  - Incorporated into Lucene indexation and search engine

2.5 Collaborative captioning

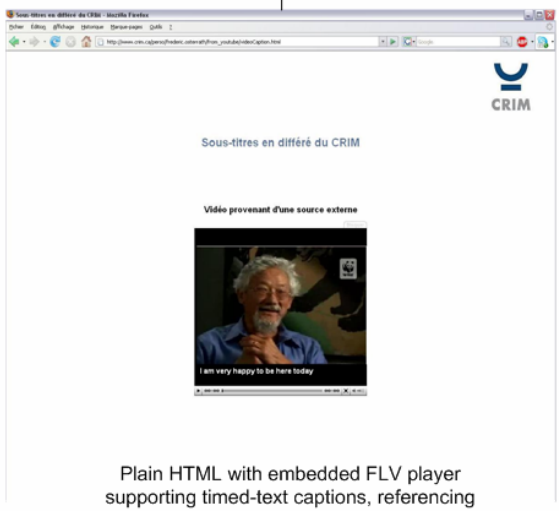
- Populate a central repository of captions in a collaborative fashion - Wiki-style - by Internet users, and display synchronized captions in a browser applet for any video on the Web which has captions in the repository.
- Distinct from « provider-hosted captions » approach
  - Collaborative (Web 2.0, under viewer control, etc.)
  - Aims to support any video content provider
  - Tools for viewing / creating captions (WYSIWYCM)
- Status: design phase

E-Inclusion collaborative captioning architecture

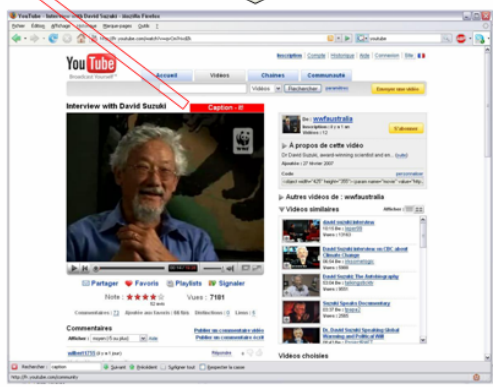




Interactive manual caption input interface, built with OpenLaszlo



Plain HTML with embedded FLV player supporting timed-text captions, referencing external Video URL and internal caption file



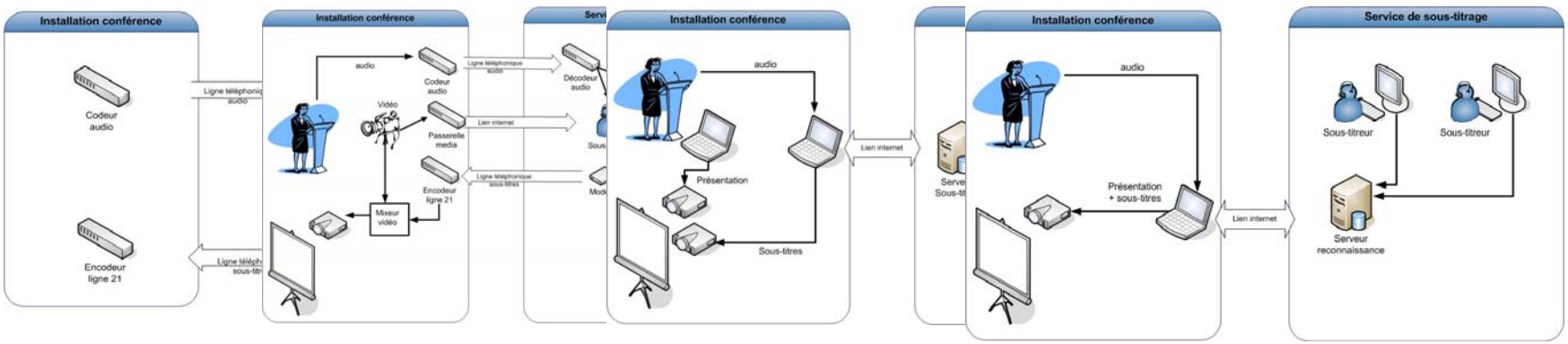
Firefox with Video-Content aware "GreaseMonkey" javascript extension, dynamically monitoring HTML content

2.3 Anything, anywhere: live captioning

- Extend ASR live captioning to live plenary sessions of conferences or large meetings : one-time events, in remote locations, very little data in advance.
- Why? Need expressed by D&H community
  - Live captioning in the work place
- Why is it difficult?
  - ASR problem: generic LM gets 80%, need 90% accuracy
    - Written, formal, planned → spoken, familiar, spontaneous
    - Domain or corporate specific, proper names, etc.
  - Captioning personnel and booth in meeting room?
- Approach
  - Adaptation algorithms requiring much less data
  - Sentence extraction from large corpora using small text
  - Mobile captioning apparatus for remote captioning

2.3 Anything, anywhere: live captioning

- Starting point: live TV broadcast captioning
  - video feed from broadcaster
  - captioning at CRIM
  - captions sent to broadcast encoder over phone line
- Mobile captioning
  - laptop with camera and audio capture at event location
  - captioning done anywhere on the Internet (CRIM)
  - captions sent by webcast
  - projected from laptop at event location on large screen
  - or viewable on any laptop with internet access



2.3 Anything, anywhere: live captioning

- 11th International Conference on Mobility and Transport for Elderly and Disabled Persons (TRANSED 2007) 06/2007
  - PGA President's cup 08/2007
    - IRD Annual general meeting 10/2007
      - IRD User committee meeting 12/2007
        - La Boule de Cristal du CRIM 2008 02/2008
          - IRD User committee meeting 03/2008

Publications

- **Périodiques (1)**
- Dehak, N., Dumouchel, P. and P. Kenny. “Modeling Prosodic Features with Joint Factor Analysis for Speaker Verification” IEEE Transactions on Audio, Speech and Language Processing, 15 (7) September 2007 : 2095-2103.

- **Actes de congrès, colloques, conférences (5)**
- Dehak, N., Dehak, R., Kenny, P. and P. Dumouchel. “Comparison Between Factor Analysis and GMM Support Vector Machines for Speaker Verification” In Proceedings of IEEE Odyssey 2008 - The Speaker and Language Recognition Workshop, pp. *. Stellenbosch, South Africa, January 21-25, 2008.

- Dehak, R., Dehak, N., Kenny, P. and P. Dumouchel. “Kernel Combination for SVM Speaker Verification” In Proceedings of IEEE Odyssey 2008 - The Speaker and Language Recognition Workshop, pp. *. Stellenbosch, South Africa, January 21-25, 2008.

- Dehak, N., Kenny, P. and P. Dumouchel. “Continuous Prosodic Features and Formant Modeling with Joint Factor Analysis for Speaker Verification” In Proceedings of the 8th International Conference of Interspeech 2007, pp. 1234-1237. Antwerp, Belgium, August 27-31, 2007.

- Dehak, R., Dehak, N., Kenny, P. and P. Dumouchel. “Linear and Non Linear Kernel GMM Support Vector Machines for Speaker Verification” In *Proceedings of the 8th International Conference of Interspeech 2007*, pp. 302-305. Antwerp, Belgium, August 27-31, 2007.
- Cardinal, P., Boulianne, G., Comeau, M., and M. Boisvert, "Real-time correction of Closed-Captions", Proceedings of ACL 2007, 2007.