**VoxaleadNews: Robust Automatic Segmentation of Video into Browsable Content**

Julien Law-To  
Exalead  
10 place Madeleine  
Paris, France, 75008  
lawto@exalead.com

Gregory Grefenstette  
Exalead  
10 place Madeleine  
Paris, France, 75008  
ggrefens@exalead.com

Jean-Luc Gauvain  
LIMSI CNRS  
Université Paris-Sud  
Orsay, France, 91403  
gauvain@limsi.fr

**ABSTRACT**

We present an interface to video and audio podcasts that extracts semantics from the speech content, and packages the extracted information in a variety of navigation tools. The user can jump to the relevant sections and browse from relevant section to relevant section. This interface is related to the Yahoo! Challenge: Robust Automatic Segmentation of Video According to Narrative Themes.

**Categories and Subject Descriptors**

H.5.2 [Information Interfaces and Presentation]: User Interfaces—Prototyping

**General Terms**

Algorithms, Experimentation, Human Factors, Languages

1. **INTRODUCTION**

Most video search today relies primarily on textual metadata associated with the video (title, tags, surrounding page-text). We present an alternative approach for searching videos and audio newcasts based on content derived from an automatic speech recognition technology. Most of the linguistic information is encoded in the audio channel of video data, which, once transcribed, can be accessed using text-based tools. Our interface, called Voxalead News\(^1\), allows users to search for sections within a news broadcast, via a navigation look and feel that is familiar to search engine users.

2. **BACK END**

Our online system is fed free RSS emissions, available as podcasts. For the moment all the selected sources deal with News.

**ASR.** The state-of-the-art automatic speech transcription systems for the French, English, Mandarin and Arabic languages are based on statistical modeling techniques. In Voxalead News, we exploit the modeling and decoding techniques described in [1], which concerns the LIMSI\(^2\) English broadcast news system, although the acoustic and language models and pronunciation dictionaries are language dependent [2]. Speech decoding is carried out in a single pass with a statistical n-gram language model. This LIMSI technology has been frequently demonstrated to obtain top performance in international benchmarks.

**Index and semantic filters.** Once speech is transcribed into text, words, named entities, and terms are indexed with their time codes using the Exalead Cloudview search engine. Named entities are organized in three broad categories: people, location, organization. Each category is exposed and color coded in the interface. The full range of Exalead’s semantic processing is available for the transcribed text.

**Processing time.** Currently on the demo website, about 110 audio and video podcasts are processed daily from about 20 television stations and a dozen radios in four languages, providing an additional 13 hours of newly indexed content per day. Video processing, including speech transcription, takes about 5 minutes for a 5 minute video, with an additional minute for indexing and storage corresponding to a processing in about 1.2 times real-time. A four processor server (like the one we underuse) can process 84 hours of new content per day. Servers can be replicated immediately in the system to handle any load.

3. **FRONT END**

Once the podcasts have been indexed, the Voxalead front-end presents the content in an interface familiar to search engine users.

---

\(^1\)http://voxaleadnews.labs.exalead.com

\(^2\)http://www.limsi.fr/

---

**Copyright is held by the author/owner(s).**

**MM’09, October 19–24, 2009, Beijing, China.**

**ACM 978-1-60558-608-3/09/10.**

---

**Figure 1:** Trends which suggest a query.

**3.1 The welcome page: trends of the News**

The welcome page of Voxalead News gives a tagcloud overview of current news, underneath a familiar search box as illustrated by Figure 1. It presents color coded named entities extracted from all podcast news within a specified
time period: previous day, previous week or previous month. Figure 1 shows Omar Bongo due to his recent death, and Airbus and AirFrance from the recent plane tragedy. Clicking on the word in the tagcloud executes a query over the broadcasts. The query can also be run as in any search engine, with the user typing keywords in the search box. An auto-completion service is provided, helping to spell named entities known to the system. All advanced search functions are available, as with classic web search engines: matching exact phrases, logical and regular expressions.

3.2 Result page thumbnails

If you enter the query Roger Federer you see an automatic selection of video shots from different sources with their relevant thumbnail snippets as illustrated in Figure 2.

3.3 Segment browsing

Once a hit has been selected, the video is streamed, starting directly from the segment relevant to the query. If more than one segment is relevant, the user can jump or scroll through relevant segments easily. A snippet of 30 words containing the query-relevant keywords is overlayed on the subscreen video timeline, providing further semantic clues for efficient browsing as illustrated in Figure 4. The tagcloud provides another browsing tool since the named entities of the video are clickable for advancing through the video.

4. CONCLUSION AND EVOLUTIONS

In Voxalead News, the query-specific tagclouds, scrollable relevant thumbnails, and transcribed speech overlays on the thumbnails provide the user with a type of query-specific narrative linked into the video. Video content is not presented as a single block, but is segmented by its content, and accessible in query-dependent segments. Podcast content is currently derived essentially from Speech-to-Text technology however, the planned next steps are to augment this content via deeper image processing tools like OCR and face recognition. The current interface will be also enhanced to improve the user experience based on user testing.

5. ACKNOWLEDGMENTS

This work has been partially financed by OSEO under the Quaero3 program.

6. REFERENCES


3http://www.quaero.org/