# Introducing topic segmentation and segmented-based browsing tools into a content based video retrieval system

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
Universite Paris-Sud
Orsay, France, 91403
gauvain@limsi.fr

Guillaume Gravier
IRISA CNRS
Campus de Beaulieu
Rennes, France, 35042
guig@irisa.fr

Lori Lamel
LIMSI CNRS
Universite Paris-Sud
Orsay, France, 91403
lamel@limsi.fr

Julien Despres
Vecsys Research
ZA de Courtaboeuf
Les Ulis, France, 91952
despres@vecsysresearch.com

## ABSTRACT

One important class of online videos are news broadcasts. Most news organisations provide immediate access to topical news broadcasts over the Internet, through RSS streams or podcasts. Until lately, technology has not made it possible for a user to automatically find, within a longer broadcast, the smaller parts that might interest them. Recent advances in both speech recognition systems and natural language processing have led to a number of robust tools that allow us to provide users with quicker and more focussed access to relevant segments of one or more news broadcast videos. Here we present our new interface for browsing or searching news broadcasts (video or audio) that exploits these new tools to provide immediate access to topical passages within news broadcasts; to browse news broadcasts by events as well as by people, places and organisations; to perform cross lingual search of news broadcasts; to search for news through a map interface; to browse news by trending topics; and to have a brief automatically generated textual characterisation of news segments before listening. Our publicly searchable VoxaleadNews demonstrator currently indexes daily broadcast news content from 50 sources in English, French, Chinese, Arabic, Spanish, Dutch and Russian.

## 1. INTRODUCTION

Many of the major news organisations provide immediate access to topical news broadcasts over the internet, through RSS streams or podcasts. Many users rely on third-party sites[1] to describe topical extracts of longer news broadcasts. Until lately, technology has not made it possible for a user to automatically find, within a longer broadcast, the smaller segments that might interest him, so users have depended on human editors to edit longer videos into smaller topic focussed units. This dependence has limited the range of topics that a user can search, since a human must be motivated to perform the work. But there is hope. Recent advances in both speech recognition systems and natural language processing have led to a number of robust tools that allow us to provide the user with quicker and more focussed access to relevant segments of one or more news broadcast videos.

Most video search today relies in a large part on indexing the textual metadata associated with the video (title, tags, surrounding page-text). Videos that are returned for a search over common search engines are those which contain the search terms in this metadata.

We present an alternative approach for browsing and searching videos and audio newscasts 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 processed using natural language processing and other semantic processing. Our interface, called VoxaleadNews[2], integrates many of these technologies to provide topical access to automatically identified broadcast segments. The biggest novelty in VoxaleadNews over last year's version is the topical segmentation of news broadcasts. All the search and browsing tools are generated from automatically detected topical segments, using methods described below. Search is constrained to each segment, and these segments are the result of the search, though the entire broadcast is still accessible if desired. If a user searches for *Ron Paul AND Barack Obama* then only those segments in which both men are mentioned are returned. An elaborate interface provides additional, optional annotations for each segment: these annotations include named entities, timestamps of mentions of each query term, a pincushion timeline bar showing mentions, and for each segment a list of corpus derived important terms mentioned in that segment. Querying can be performed in a language different from broadcast language, exploiting commonly available translation tools.

In what follows, we describe the back end processing of the

[1] For example, reddit.com, snowspotmedia.com, huffingtonpost.com, crooksliars.com, ...

[2] http://vox.labs.exalead.com/voxalead

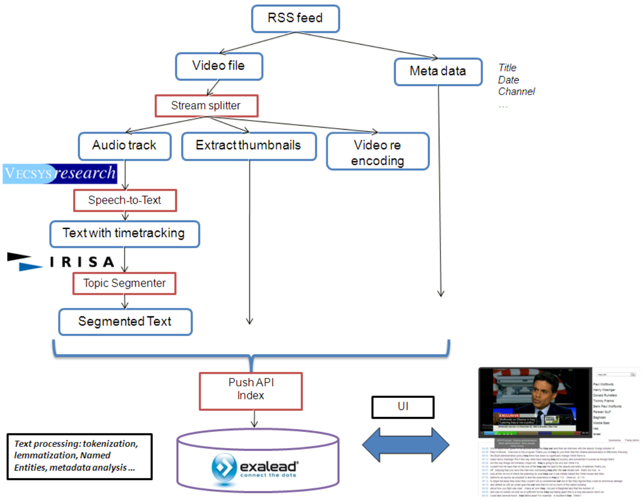


**Figure 1: VoxaleadNews Architecture Overview.**

video and audio sources in Section 2. Section 3 describes the user interface. Section 4 talks about processing time, and this is followed by a conclusion and future evolutions segment.

# 2. NEWS SOURCES, BACK END

The VoxaleadNews system indexes free podcast sources, available via RSS streams. Figure 1 presents the main steps of the processing of these podcasts.

## 2.1 ASR

In the context of the Quaero program, LIMSI and VECSYS RESEARCH[3] have provided state-of-the-art automatic speech transcription systems for 7 languages: French, English, Spanish, Mandarin, Dutch, Russian and Arabic. The transcription systems make use of statistical modelling techniques described in [1], which gives details for the English broadcast news system. The acoustic and language models and pronunciation dictionaries are language dependent [2], and trained on large audio and text corpora. Speech decoding is carried out in a single pass with a statistical n-gram language model, and takes less time than the signal duration. Proper case is provided for all languages, and a post processing converts numerical quantities for amounts, dates, telephone numbers to Arabic form for the English, French, and Spanish languages. The system outputs an xml file containing the words identified in the audio document, along with their time codes and a confidence measure.

The first processing step partitions the data in to speech segments, and, after determining the gender, clusters segments from the same speaker. In future versions this information will be combined with the content in the automatic transcription to associate true names to parts of the data.

The systems have word lists containing from 50k to 300k words and generally have a good coverage of the language. However, breaking news may have repeated occurrences of words that are unknown to the system. A new functionality has recently been incorporated which allows users to update the recognition word list and is currently undergoing experimentation.

This technology has been frequently demonstrated to obtain top performance in international benchmarks.

## 2.2 Topic segmentation

In this latest version of Voxleadnews, we address the problem of searching for relevant segments of a video in a news broadcast. A news broadcast is often divided into stories, which may have no relation with each other. If the broadcast is transcribed into one textual document a complex search, such as *Barack Obama in China* may return videos in which *China* is mentioned in one story and *Barack Obama* in another, contrary to what the user intended to find. In this newest version of VoxaleadNews, we treat the uninterrupted textual output of the automatic speech transcription by applying topic segmentation to break the transcript of a show into topically homogeneous segments. These segments ideally would correspond to individual reports in classical news.

Topic segmentation has been studied in natural language processing since the early 1990s [3]. Most approaches use vocabulary differences over the document to detect subject changes and topic shifts.

VoxaleadNews relies on the IRISA news topic segmentation (Irints) system which itself extends the linear segmentation methods described in [4]. The general idea of this lexical cohesion based method is to search for the best possible segmentation among all the possible ones. A generalized probability criterion is used to measure thematic cohesion of a segment, exploiting repetitions in the vocabulary: A unigram language model, estimated from the word counts in the segment, is used to compute the probability of the word sequence corresponding to the segment. In the current version of Irints, language model estimation has been improved with respect to [4]. In addition, features were added to account for the peculiarities of broadcast news transcripts, namely transcription errors and the limited number of repeated words due to stylistic reasons. In particular, word level confidence measures are used to deal with transcription errors while semantic relations are introduced to counteract the limited number of repetitions [5].

In practice, each word in the transcripts is labelled with part-of-speech tags and lemmatized. The computation of the generalized probability is limited to nouns, adjectives and non-modal verbs. The output of the segmentation process is a set of segments. As a by-product, that we exploit, each segment is characterized by the few keywords which most significantly contributed to the lexical cohesion of the segment.

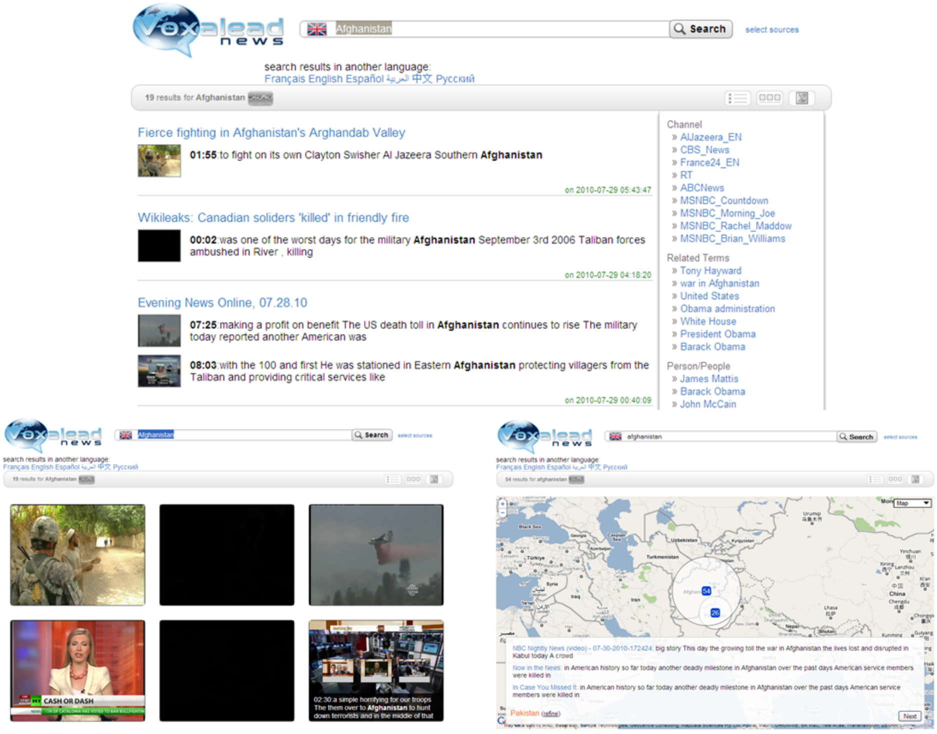
## 2.3 Index and semantic filters

Once speech is transcribed into text and segmented, standard natural language processing techniques are applied to each segment. Extracted words, named entities, and multiword terms are then indexed, along with their time codes from the original automatic speech recognition, using the Exalead Cloudview search engine. The named entities that are retained for VoxaleadNews are organized in four broad categories: people, location, organization and events.

# 3. USER INTERFACE, FRONT END

A user is presented with a welcome screen that now shows trending people, places, locations, or events (over the past day, week, month or over a user defined period). This allows users to kill time during lunch breaks in profitable ways,

[3] http://www.limsi.fr/, http://www.vecsysresearch.com



Afghanistan

search results in another language:
Français English Español العربية 中文 Русский

19 results for Afghanistan

Fierce fighting in Afghanistan's Arghandab Valley

**01:55**:to fight on its own Clayton Swisher Al Jazeera Southern **Afghanistan**

on 2010-07-29 05:43:47

Wikileaks: Canadian soliders 'killed' in friendly fire

**00:02**:was one of the worst days for the military **Afghanistan** September 3rd 2006 Taliban forces ambushed in River , killing

on 2010-07-29 04:18:20

Evening News Online, 07.28.10

**07:25**:making a profit on benefit The US death toll in **Afghanistan** continues to rise The military today reported another American was

**08:03**:with the 100 and first He was stationed in Eastern **Afghanistan** protecting villagers from the Taliban and providing critical services like

on 2010-07-29 00:40:09

Channel
» AlJazeera_EN
» CBS_News
» France24_EN
» RT
» ABCNews
» MSNBC_Countdown
» MSNBC_Morning_Joe
» MSNBC_Rachel_Maddow
» MSNBC_Brian_Williams

Related Terms
» Tony Hayward
» war in Afghanistan
» United States
» Obama administration
» White House
» President Obama
» Barack Obama

Person/People
» James Mattis
» Barack Obama
» John McCain

CASH OR DASH

02:30:a simple horrifying for our troops The them over to **Afghanistan** to hunt down terrorists and in the middle of that

afghanistan

54 results for afghanistan

NBC Nightly News (video) - 07-30-2010-172424: big story This day the growing toll the war in Afghanistan the lives lost and disrupted in Kabul today A crowd

Now in the News: in American history so far today another deadly milestone in Afghanistan over the past days American service members were killed in

In Case You Missed It: in American history so far today another deadly milestone in Afghanistan over the past days American service members were killed in

**Figure 2: Different visualizations of the results**

keeping abreast of the latest news. The user can also search in the unique search box, as in a classical search engine. Lemmatization and the use of stop words are specific to the language selected by the user.

## 3.1 Result page thumbnails

In response to a query, we have provided a richer interface than the one described last year [6], with multiple views on the results. In addition to thumbnails, timelines, and tag clouds views, the user now has access to a map view showing locations mentioned in fetched segments, as well as automatically determined trends in these segments. These maps and trends are all calculated from the ASR transcription of the audio streams of the broadcasts. Search is performed using a navigation look and feel that is familiar to search engine users.

Figure 2 illustrates these different views. The list view is the classic one, it presents a list of hits with text snippets related to the query. Thumbnails are contextual, this mean that they are time related to the text snippet. The user can launch the video at the direct time code by clicking on the corresponding thumbnail. On the right of this screen, facet search provides another mechanism for fast search and refine in set of data that has associated (like the source of the podcasts) or extracted (like named entities) typed metadata.

Another view presents the results with large thumbnails with text snippet overlaid. Smaller thumbnails inside present the different part of the video when the query is mentioned and here again, the thumbnails are related to the moments when the words are mentioned. The map view offers the possibility to see the location of the documents. These location are automatically extracted from the transcribed text. Trends view propose a better visualization of the facets related to the query using charts.

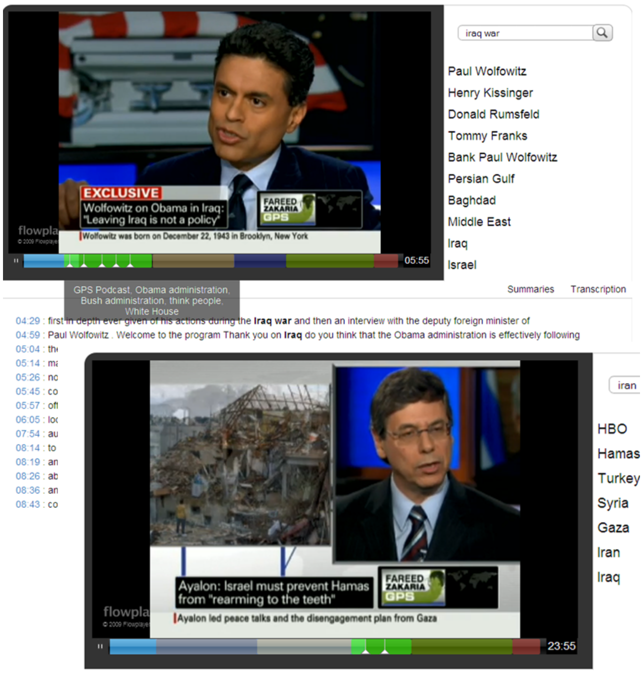
## 3.2 Segment browsing

Once a hit has been selected, the video is streamed, starting directly from the segment relevant to the query. The timeline of the video player presents markers with snippets of 30 words containing the query-relevant keywords corresponding to moments where the query words are mentioned. It also presents segments computed by the topic segmenter as shown in figure 3 which present the play after a search on the query "Iraq war"; the second play page presents the play of another segment of the same video which deals about another topic. When the user positions the cursor on a segment bar, we show keywords that are mentioned in this segment, reminiscent of [7]. We calculate these keywords by first processing a large number of news documents in a first pass. This processing generates a dictionary of common terms, which is updated from time to time. These keywords are displayed during the hover. Associated to these keywords, the named entities that correspond only to the segment are displayed on the right of the video player. This allows to clearly highlight the topics that are discussed in the segment.

## 3.3 Multilanguage search

Another new feature in VoxaleadNews is to be able to make queries in a language and see the results from other languages. For example in figure 4, a query ("Afghanistan") is done in English (as illustrated by the flag). By clicking on the search in other languages, the application uses tools[4]

[4] currently Google translation

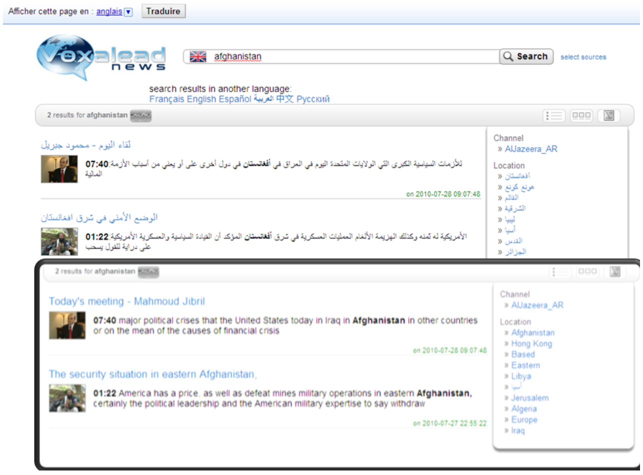


04:29 : first in depth ever given of his actions during the **Iraq war** and then an interview with the deputy foreign minister of

04:59 : Paul Wolfowitz . Welcome to the program Thank you on **Iraq** do you think that the Obama administration is effectively following

**Figure 3: Play page with segmentation**

to translate the query and display the results in the selected language (Arabic in the example). The results are displayed in Arabic but can also be translated in English: the snippets and the facets are translated and can be clicked to play the podcast. The play page can also be translated therefore the search can be in other languages and displayed in any language in an easy and intuitive way for the user.



لقاء اليوم - محمود جبريل

**07:40**:للأزمات السياسية الكبرى التي الولايات المتحدة اليوم في العراق في **أفغانستان** في دول أخرى على أو يعني من أسباب الأزمة المالية

on 2010-07-28 09:07:48

الوضع الأمني في شرق افغانستان

**01:22**:الأمريكية له ثمنه وكذلك الهزيمة الألغام العمليات العسكرية في شرق **أفغانستان** المؤكد أن القيادة السياسية والعسكرية الأمريكية على دراية للقول يسحب

Today's meeting - Mahmoud Jibril

**07:40** major political crises that the United States today in Iraq in **Afghanistan** in other countries or on the mean of the causes of financial crisis

on 2010-07-28 09:07:48

The security situation in eastern Afghanistan,

**01:22** America has a price, as well as defeat mines military operations in eastern **Afghanistan**, certainly the political leadership and the American military expertise to say withdraw

on 2010-07-27 22:55:22

**Figure 4: Search in multi languages.**

## 4. PROCESSING TIME AND PERFORMANCE.

Two servers are used in the VoxaleadNews demonstrator. The first one is for the back-end which is in form of a distributed task scheduler. The system is built around a highly available and distributed task queue, and workers. Each task is scheduled on a worker (multiple workers per machine) by the queue and executed. Workers can be removed or added at will without any task being lost. The system processes more than 150 new items each day. Amounting to roughly 3.5gb or 15 hours of new content each day, on a single 6 core server. Potentially, this server can absorb and process about 100 hours of videos per days. As this backend is distributed, servers can be added immediately in the system to handle any load.

Another server is dedicated to handle the users queries. This part is highly scalable and is very close to the one used for the Exalead.com website [5] (which contains 16 billions of web pages). This part is therefore highly scalable. The previous VoxaleadNews demonstrator handle easily 500 unique visitors per day.

## 5. CONCLUSION AND EVOLUTIONS

VoxaleadNews propose tools for searching and browsing news videos by their enriched contents. It 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 the audio part however, the planned next steps are to augment this content via deeper image processing tools like OCR and face recognition. Speaker recognition is also to be explored for the next version.

## 6. ACKNOWLEDGMENTS

This work has been partially financed by OSEO under the QUAERO[6] program and by the EU under the GLOCAL project (FP7 - 248984).

## 7. REFERENCES

[1] J.L Gauvain, L. Lamel. G. Adda, "The LIMSI Broadcast News Transcription System," *S*peech Communication, **37**(1-2):89-108, 2002.

[2] L. Lamel and J.L. Gauvain, "Speech processing for audio indexing," *GoTAL 2008 - Advances in NLP, no.5221/2008 LNCS,* pp. 4-15. Springer Verlag, 2008.

[3] M.A. Hearst, "Multi-paragraph segmentation of expository text," *3*2nd Annual Meeting of the Association for Computational Linguistics, 1994.

[4] M. Utiyama and H. Isahara, "A statistical model for domain-independent text segmentation," Proc. Annual Meeting of the Association for Computational Linguistics, 2001.

[5] C. Guinaudeau, G.Gravier and P. Sébillot, "Improving ASR-based topic segmentation of TV programs with confidence measures and semantic relations," *P*roc. of the Intl. Speech Communication Association Conference, 2010.

[6] J. Law-To, G. Grefenstette and J.L Gauvain, "VoxaleadNews: robust automatic segmentation of video into browsable content,". *A*CM Multimedia 2009, 2009.

[7] F. Bourdoncle, "LiveTopics: recherche visuelle d'information sur l'Internet", in *R*IAO, 1997.

[5] http://www.exalead.com/search/

[6] http://www.quaero.org/