ABSTRACT
One important class of online videos is news broadcasts. Most news organizations 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 focused access to relevant segments of one or more news broadcast videos. Here we present an interface for browsing or searching news broadcasts (video or audio) that exploits these new language processing tools to (i) provide immediate access to topical passages within news broadcasts, (ii) browse news broadcasts by events as well as by people, places and organizations, (iii) search for news through a map interface, (iv) browse news by trending topics, and (v) have a brief automatically generated textual characterization of news segments before listening. Our publicly searchable demonstrator currently indexes daily broadcast news content from sources in English, French, Chinese, Arabic, Spanish, Dutch, Russian and Italian. This demonstrator can be consulted online at: http://vox.labs.exalead.com/voxalead/ and a video that describe the interface is available at: http://www.youtube.com/watch?v=1RhWGu9ElUU

This demonstrator has an optimized version for tablet (Ipad).

Keywords
Audio/Video indexing and retrieval, Video browsing, Video tagging

INTRODUCTION

This is the space reserved for copyright notices.


Many of the major news organizations provide immediate access to topical news broadcasts over the internet, through RSS streams or podcasts. In parallel, many users rely on third-party sites 1 to describe topical extracts of longer news broadcasts. However, in spite of early attempts on broadcast news retrieval and browsing from speech at the turn of the century (Maybury 2003, Hauptmann 2003), technology has not made it possible for a user to efficiently find small segments of interest from longer broadcasts within a large collection spanning multiple languages. In particular, work on topic segmentation of broadcast news, e.g. Allan 2002, was limited in the number of shows and languages that could be handled. Most current video search engines rely, 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 those search terms in their metadata. Recent progress in spoken language processing—in particular transcription, topic segmentation, keyword extraction—has led to a number of robust tools that allow us to provide users with quicker and more focused access to relevant segments of one or more news broadcast videos. Researchers have focused on either one or the other aspect of this entire processing chain. Still, few systems integrate all of these components in a complete and comprehensive large scale demonstration enable to return portions of videos relevant to a query, while also providing query-free browsing capabilities.

The demonstrator described here is an alternative approach for browsing and searching videos and audio newscasts based on robust spoken document processing in multiple languages. In broadcast news, 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 semantic processing techniques. The interface presented here integrates many of these technologies to provide topical access to automatically identified broadcast segments. 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 returned as the result of a search, though the entire

1 For example, reddit.com, huffingtonpost.com, crooksliars.com, ...
broadcast is still accessible if desired. Our interface also provides additional, more elaborate optional annotations for each segment: named entities, timestamps of mentions of each query term, a pincushion timeline bar showing mentions, and for each segment a label of corpus-derived important terms mentioned in that segment.

This document gives first some explanation of how it works: back-end processing of the video and audio sources; then it describes the user interface of the demonstrator.

**BACK END: OFF LINE PROCESSING**

The demonstrator index everyday new podcast of news in different languages using RSS feeds. There are currently more than 100,000 podcasts in the system and about 200 new podcasts are added everyday.

**Automatic Speech Recognition**

State-of-the art speech transcription systems for 8 languages (French, English, Spanish, Mandarin, Dutch, Russian, Italian and Arabic) are the core of the demonstration. A description can be seen in Gauvain (2002). The acoustic and language models and pronunciation dictionaries are language dependent and trained on large audio and text corpora. The system outputs an xml file containing the words identified in the audio document, along with their time codes and a confidence measure. Each language has recognition word lists containing from 50k to 300k words and generally has a good coverage of the language. This technology has been frequently demonstrated to obtain top performance in international benchmarks.

**Topic segmentation**

We address here 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. To remedy this problem, we process 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. Our system relies on an extension of the linear segmentation methods described in Utiyama (2001). 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. Our language model estimation has been improved with respect to Utiyama (2001). We added features 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 with the same methods as in Guinaudeau (2010). In practice, each word in the transcript is labeled with part-of-speech tags and lemmatized. Computation of thematic coherence probability is limited to nouns, adjectives and non-modal verbs. The output of the topic segmentation process is a set of segments. Each segment can be labeled by the few keywords which most significantly contributed to the lexical cohesion of the segment.

**Index and semantic filters**

Once speech is transcribed into text and segmented, standard natural language processing techniques are applied to each segment considered as a document in our index. The index is based on Cloudview, the search platform developed by Exalead.

The indexed documents are enriched and tagged using several semantic processors. The tags added are:

- named entities: people, location, organization and events.
- multi-word terms (n-gram)

The documents are then indexed by:

- their metadata (most of the time only the title and a date)
- the automatic transcription
- the tags generated from the transcription

The next section presents the interface of the demonstrator.

**USER INTERFACE, FRONT END**

Once video and audio streams have been segmented and processed, the user is presented with a welcome screen that immediately shows trending people, places, locations, or events over a user definable period (day, week, months). This allows users to kill time during lunch breaks in profitable ways, keeping abreast of the latest news. The user can also search in the unique search box, as in classical search. All advanced search functions are available, as with classic web search engines: matching exact phrases, logical and regular expressions. Lemmatization and the use of stop words are specific to the language selected by the user.

**Result page thumbnail**

In response to a query, we have provided a rich interface 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 enriched by the indexing process. Search is performed using a navigation look and feel that is
familiar to search engine users. Figure 1 illustrates these different views with the query "Afghanistan".

The list view is the classic one; it presents a list of hits with text snippets related to the query. Thumbnails are contextual: 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 overlaid text snippets. By clicking on the next icon, the user can see the different part of the video when the query is mentioned with thumbnails related to the moments when the words are mentioned. The map view offers the possibility to see the location automatically extracted from the transcribed text of the documents. When clicking on the map on Iraq, for example, we can see the contextual snippets that correspond to the query Afghanistan with a refine on Iraq: both words are highlighted giving a quick overview of the context to the user. Trends view sometimes provides a better visualization of the facets related to the query using charts or tag cloud.

For all the views, a histogram time-line allows users to restrict the search to a specific time period. It also allows the user to visualize the number of times that the query term has been mentioned over time.

**Segment browsing**

Once a query result has been selected, the corresponding 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. As illustrated in Figure 2, the query was Tony Curtis. We can see that the last segment on the time line is the targeted segment that deals with Tony Curtis. The named entities are all related to the query: Bernie Schwartz is his real name, Jamie Lee Curtis is his daughter and Marilyn Monroe and Jack Lemmon were in the famous movie *Some Like It Hot* with him.

On the same page, we can see that other segments are presented. A thin progression line shows the position of the play within the whole video. The named entities that correspond only to the segment are displayed on the right of
the video player. By clicking on these named entities, a new marker in the video time line appears to show exactly where these entities were said. The user can then easily browse the video to go directly to the targeted part.

When the user positions the cursor on a segment bar, we show keywords that are mentioned in this segment. These labels and named entities highlight the topics that are discussed in the segment: in the second case the keywords were "foreclosure", "home" and "year"; "year" is obviously not very relevant but "foreclosure" would really be one of the best keywords for this report.

![Figure 2. Playing a content: segmentation and browsing](image)

**CONCLUSION**

Here we described our demonstrator for searching and browsing news videos by their language content. It provides the user with a type of query-specific narrative linked into the video. In our system, the video content is not presented as a single block, but is segmented by its content, and accessible in query-dependent segments.

This system can be adapted to other use case and we are currently working on a version for searching and browsing into lectures from university. A first prototype is in progress and we hope that we will be able to present it at the conference.

**ACKNOWLEDGMENTS**

The ASR system has been provided by LIMSI/CNRS and Vocapia. The topic segmenter has been provided by IRISA/CNRS. These works have been funded by OSEO within the QUÆRO project.

**REFERENCES**


