VOXALEAD: A Scalable Video Search Engine Based On Content

Julien Law-To
Exalead
75010 Paris
Julien.Lawto@3ds.com

Gregory Grefenstette
Exalead
75010 Paris
Gregory.Grefenstette@3ds.com

INTRODUCTION TO QUAUERO PROJECT
Quaero, a research and innovation program addressing automatic processing of multimedia and multilingual content, fosters the development of new tools for navigation in large volumes of audiovisual content. Quaero projects (automatic information retrieval, analysis, segmentation and classification of text, speech, music, image and video) are supported by nearly 200 M€ of funding from Oséo Innovation, and involve 30 French and German partners. Data-centric application projects are led by industrial partners who share a research structure with universities. Each application area includes systematic use of technology evaluation. Entering its fourth year, Quaero very active eco-system has produced innovative applications, such as Voxalead presented here, that have been launched in large scale deployments. A separate report can be found for each of the applications, but all under the banner of the Quaero program.

ABSTRACT
Most news organizations provide immediate access to topical news broadcasts through RSS streams or podcasts. Until recently, applications have not permitted a user to perform content based search within a longer spoken broadcast to find the segment that might interest them. Recent progress in both automatic speech recognition (ASR) and natural language processing (NLP) has produced robust tools that allow us to now provide users with quicker and more focused access to relevant segments of news broadcast videos. Our public online demonstrator of the Voxalead application currently indexes daily broadcast news content from 50 sources in English, French, Chinese, Arabic, Spanish, Dutch, Italian and Russian.

1. INTRODUCTION
Our new interface for browsing or searching news broadcasts (video or audio) 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) perform cross-lingual search of news broadcasts, (iv) search for news through a geographic map interface, (v) browse news by trending topics, and (vi) see a brief, automatically generated textual summary of news segments before listening.

2. TECHNICAL DESCRIPTION
2.1 Automatic Speech Recognition
The core of the Voxalead application is its state-of-the-art speech transcription system\(^1\) for 8 languages: French, English, Spanish, Mandarin, Dutch, Russian, Italian and Arabic. 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, providing and adequate to good coverage of the language. This technology has frequently achieved top performance in international benchmarks\(^2\).

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

Documents are richly annotated using several semantic processors. Added tags used in this Voxalead application are:
- named entities: people, location, organization and events
- multi-word terms found in the documents

The segmented broadcast documents are then indexed by:
- video metadata (usually only title and date)
- the automatic textual transcription
- the semantically tagged entities extracted

2.3 Scalability issue
The system processes more than 150 new video and audio items each day on a single 6 core server, corresponding to roughly 3.5 Gb or 15 hours of new content each day. Potentially, this same underused server could absorb and process about 100 hours of videos per day. A separate dedicated server, which could also be parallelized, handles user queries. It can currently easily handle its 500 unique visits per day. Both sides are highly scalable, and similar to our public web search engine website (www.exalead.com/search which contains 16 billions of web pages). Voxaleadnews currently holds more than 70,000 podcasts.

3. USER INTERFACE
3.1 Search for relevant video
In response to a query, our interface provides multiple views on the results: thumbnails, timelines, and tag clouds views, a map view showing locations mentioned in the retrieved audio segments, and topic trends in retrieved segments. Maps and trends are calculated from NLP over the ASR transcription of the audio streams of the broadcasts. Search is performed using a navigational look-and-feel familiar to search engine users. For all views, a histogram time-line allows users to restrict the search to a

---

\(^1\) \text{http://www.vocapia.com/voxsigma.html}

specific time period, which also allows the user to visualize the number of mentions of the query term over time.

### 3.2 Browsing videos

When a query result is clicked, the corresponding video is streamed starting directly from the segment relevant to the query. The timeline of the video player also displays markers with mouse-over snippets of 30 words containing the query-relevant keywords at every moment where the query words are mentioned. Topic segmentation allows us to display (on the right) mentioned named entities corresponding to each segment. Clicking on these named entities adds additional markers on the time line. Or the user can perform a further search in the displayed video and other markers will be added to the time line corresponding to the additional query words, along with mouse-over snippets that give the context of the word.

### 4. INDUSTRIAL USE CASES

The VoxaleadNews application focusses on one industrial use case: to make video and audio news podcast searchable by users with subsequent browsing, refinement on named entities and visualization of the transcription. Another use case is Education, for example, for the large sets of lectures now produced by schools and university. Figure 3 presents a prototype using lectures publicly available on the web. Here, the goal is first to find the lectures which deal with the targeted subject and then to navigate into the specific segments rather than listening to the entire hour long talk.

#### 4.1 Browsing videos

When a query result is clicked, the corresponding video is streamed starting directly from the segment relevant to the query. The timeline of the video player also displays markers with mouse-over snippets of 30 words containing the query-relevant keywords at every moment where the query words are mentioned. Topic segmentation allows us to display (on the right) mentioned named entities corresponding to each segment. Clicking on these named entities adds additional markers on the time line. Or the user can perform a further search in the displayed video and other markers will be added to the time line corresponding to the additional query words, along with mouse-over snippets that give the context of the word.

#### 5. CONCLUSION

VoxaleadNews is an industrial application that incorporates latest robust techniques from ASR and NLP to provide content based, focused access to mentioned topics in video and audio content. Semantic tools included in Cloudview, the indexing system, can then be used to categorize these contents automatically following rules defined by the user.

Figure 3: Voxalead Education

Some industrial use cases do not required the full transcription and in this case the system can be seen as a way to extract automatically keywords, tags or named entities for indexing audio and video content. Semantic tools included in Cloudview, the indexing system, can then be used to categorize these contents automatically following rules defined by the user.

Another further use case is the analysis of the trends based on the corpus indexed. For example, using the podcast news indexed over the last 3 months, the query “riot” displays the locations London, Tottenham, Syria (see Figure 4). The system can be used to monitor and explore trends in any given kind of field: videos for security, call centers, etc …

#### 4.2 Browsing videos

When a query result is clicked, the corresponding video is streamed starting directly from the segment relevant to the query. The timeline of the video player also displays markers with mouse-over snippets of 30 words containing the query-relevant keywords at every moment where the query words are mentioned. Topic segmentation allows us to display (on the right) mentioned named entities corresponding to each segment. Clicking on these named entities adds additional markers on the time line. Or the user can perform a further search in the displayed video and other markers will be added to the time line corresponding to the additional query words, along with mouse-over snippets that give the context of the word.

#### 5. CONCLUSION

VoxaleadNews is an industrial application that incorporates latest robust techniques from ASR and NLP to provide content based, focused access to mentioned topics in video and audio content. Semantic tools included in Cloudview, the indexing system, can then be used to categorize these contents automatically following rules defined by the user.

Another further use case is the analysis of the trends based on the corpus indexed. For example, using the podcast news indexed over the last 3 months, the query “riot” displays the locations London, Tottenham, Syria (see Figure 4). The system can be used to monitor and explore trends in any given kind of field: videos for security, call centers, etc …

#### 5. CONCLUSION

VoxaleadNews is an industrial application that incorporates latest robust techniques from ASR and NLP to provide content based, focused access to mentioned topics in video and audio content. Semantic tools included in Cloudview, the indexing system, can then be used to categorize these contents automatically following rules defined by the user.

Another further use case is the analysis of the trends based on the corpus indexed. For example, using the podcast news indexed over the last 3 months, the query “riot” displays the locations London, Tottenham, Syria (see Figure 4). The system can be used to monitor and explore trends in any given kind of field: videos for security, call centers, etc …