Search Engine for Multilingual Audiovisual Contents

José David Pérez[1], Antonio Bonafonte[1], Antonio Cardenal[2], Marta Ruiz[3], José A. R. Fonollosa[1], Asunción Moreno[1], Eva Navas[4], Eduardo R. Banga[2]

[1] *Universitat Politècnica de Catalunya*, [2] *Universidad de Vigo*
[3] *Barcelona Media*, [4] *University of the Basque Country*

Abstract. This paper describes the BUCEADOR search engine, a web server that allows retrieving. multimedia documents (text, audio, video) in different languages. All the documents are translated into the user language and are presented either as text (for instance, subtitles in video documents) or dubbed audio. The user query consist in a sequence of keywords and can be typed or spoken. Multiple Spoken Language Technologies (SLT) servers have been implemented, such as speech recognition, speech machine translation and text-to-speech conversion. The platform can be used in the four Spanish official (Spanish, Basque, Catalan and Galician) and in English.

Keywords: multimedia search, multilingual search, speech recognition, machine translation, speech synthesis, speech to speech translation

1 Introduction

The volume of available information is growing everyday. Information is present in different media, as text, audio, video and images. Furthermore, the information is produced in many languages. In regions with language diversity, as Europe or Spain, even local information exists in several languages.

BUCEADOR is a three years project focused on advanced research in the core Spoken Language Technologies (SLT) such as speech recognition, machine translation, and text-to-speech conversion, and the successful joint integration of all of them in a multilingual and multimodal information retrieval system to access audiovisual contents. The languages considered in the project are Spanish, Catalan, Galician, Basque and English.

This paper describes a showcase developed in BUCEADOR to show the achievements of the project in the above mentioned technologies and their successful joint integration. The platform includes several speech and language engines in different languages. A *multimedia digital library* has been created from sources in all the official languages in Spain. The *documents* include News and debates TV/radio programs, newspapers and magazines.

The remainder of the paper is structured as follows: section 2 presents an overview of the search engine. Section 3 describes briefly the technologies used. Section 4 describes the documents which have been included in the *digital library*.

The user interface and functionality are presented in Section 5. Some examples of use are also presented. The paper ends with a brief summary.

2 Overview of the BUCEADOR search engine

The BUCEADOR search platform retrieves information found in text and audiovisual documents. Many processes are performed off-line, as soon as a new document is introduced in the library. For each system language, a monolingual database is created and indexed. Each database includes the transcription of all the documents in the library. When a new document is added to the library, in case it is either audio or video, the spoken content is transcribed using the speech recognition engines. Then, the original text or the transcribed text is translated into all the languages managed by the system by means of the the translation engines. Each *monolingual* database is indexed so that it can be used by the information retrieval engine. Figure 1 shows a general diagram of the BUCEADOR search platform.

The search engine is a distributed system which includes several web services. The platform interacts with them through Internet. Each partner of the project is responsible for several technologies and languages and provides REST web services to the platform.

In the operative phase, the user selects its own language and introduces a query consisting of a sentence or a set of keywords using either the keyboard (text) or the microphone (speech). If the speech modality is used, a speech recognition service transcribes the input speech into text before doing the search. Then, the information retrieval engine gets the information from the database associated to the user language. The results are presented to the user using either text or speech. If the original information is a text document, the result is also a text document. The user can either read the information or listen to the synthesized speech using the TTS services. If the document is audio or video, the user can also read it or listen to its content. Note that if the language of the audiovisual document is the user language, the user listens to the original audio or reads the transcriptions generated by the speech recognition system. On the other hand, if the language of the document is not the language of the user, then the document has been generated off-line using speech recognition, automatic translation and, optionally, speech synthesis.

As explained above, speech recognition and translation are done off-line, when the documents are introduced, so that the information can be indexed by the information retrieval system. Furthermore, the dubbed version of the audiovisual documents are also generated off-line, when the documents are added to the library. In this way, video files can be produced without any additional delay.

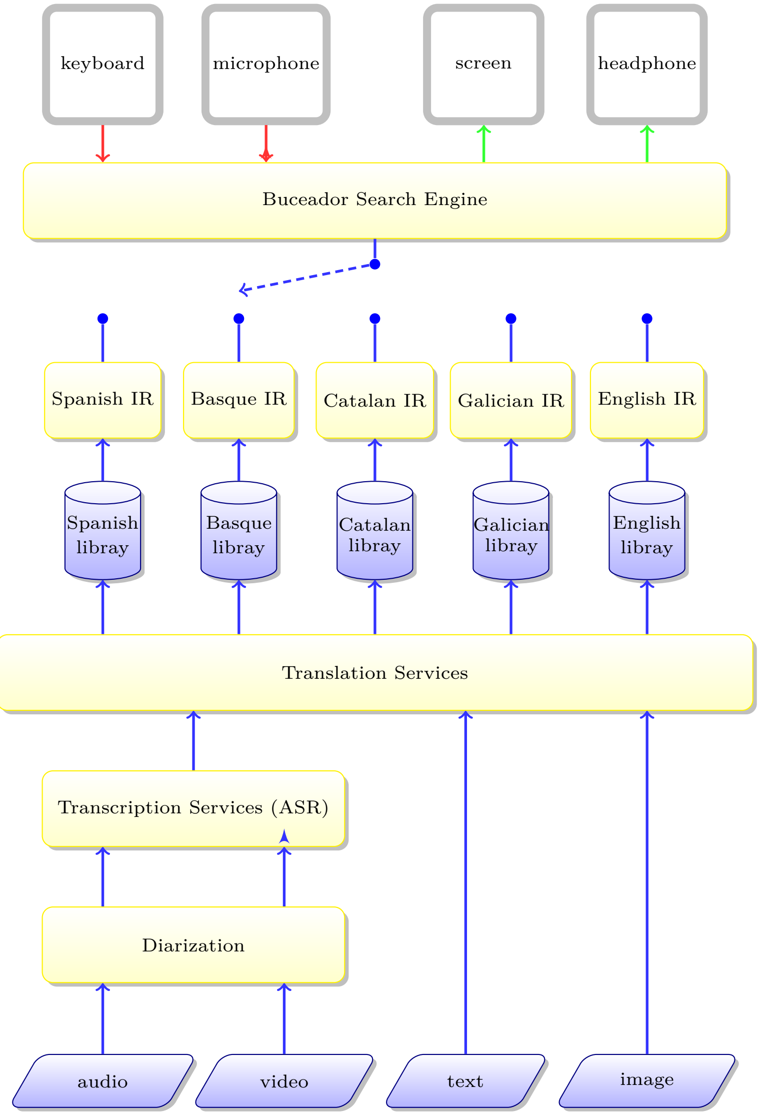


Fig. 1: Overview of the BUCEADOR search engine

3 Technology

This section presents a brief overview about the different technologies involved in this project. For more information, see [1]:

- Speech acquisition: an API programmed in java for the spoken queries. It uses a PHP script on the server side to call that API. The API will send the audio to the recognition servers using a POST method and then the servers will return the query recognized. This recognized query will be passed to the search engine.
- Speech recognition: The speech recognition engine is used in two processes. First, it is used after the diarization process, to transcribe the audio channel in the multimedia documents. It is also used to get the user keywords during the spoken queries.
- Statistical translation: there are translation engines services for Spanish ↔ Catalan, Spanish ↔ Galician and Spanish ↔ English. Spanish is used as the *pivot* language when a direct translation service is not available (for instance, to translate Galician documents into English). Translation from Basque is a very challenging task which is out of the scope of the project.
- Speech synthesis: Once the speech recognition and the translation are done, the speech synthesis engines are used for dubbing the multimedia documents to each possible output language (Spanish, Catalan, English, Galician). Then the dubbed audios are stored along with the original documents. The speech synthesis services are also required in case the user prefers to listen the content of text documents.
- Information retrieval: SOLR [2] has been used for indexation and query retrieval.

4 Multimedia library

The digital library is composed of several resources in all the Spanish official languages (Spanish, Galician, Basque and Catalan).

4.1 The resources

- In the audiovisual part we have:
  The Spanish TC-STAR database [3], is composed of plenary sessions of the European and Spanish parliaments. The speeches from the European Parliament Plenary Sessions (EPPS) have been obtained via Europe by Satellite. It comprises recordings of members and interpreters of the European Parliament speaking in the parliamentary plenary sessions. We have focused only on the Spanish speaking members fo the Parliament. Although in the recordings interventions in Spanish due to the interpreters' work can also be found those parts are not covered by our system. Additionally, the database includes recordings from the Spanish Parliament and the Spanish Congress.

For Galician language we used the Transcrigal-DB which is a database compiled on the University of Vigo during the last years. Transcrigal-DB [4] is composed of recordings of broadcast news programs emitted by the local Galician television (TVG) during years 2002, 2003, 2004, 2010 and 2011. The database contains recordings of 65 news programs of nearly one hour of duration each.

The ETB database is a broadcast news audio database that includes the audio of the news programs of the Basque TV Network (ETB) corresponding to years 2009 and 2010. It is formed by the documents that the journalists prepare and record for a news clip. The audio files contain also interviews and dubbed speech overlapped with the original voice. The files that have associated the corresponding orthographic transcription of the speech and include only one speaker have been selected to be included in the demo. In total there are 1302 files in Spanish and 1276 in standard Basque.

Agora-DB [5] is a Catalan broadcast conversational speech database of the Agora program, that are debates on selected topics from politics, economy or society. Each broadcast follows a repeating format: initially the anchorman presents the current topic, followed by an introduction of invited participants featuring background music. The main part features the debate between the invited participants, usually public figures. During the debate, public opinions are added, either as e-mails or faxes read by the anchorman, or telephone recordings played back featuring background music again. The debate generally comprises spontaneous speech, whereas the introduction of topic and participants features planned speech. Although the main language is Catalan, the recorded broadcasts contain a high proportion of Spanish speaking participants.

– In the text part we have:

  The documents from the bilingual newspaper *El Periódico de Catalunya*, included in ELDA catalog, have been added to the BUCEADOR library. This documents are presented in the written form (either in original language or translated). The original sources are in Spanish and Catalan.

  Finally the database includes a selection of different sections (*sentencias, consultorio*) of the *EROSKI CONSUMER* magazine. The original sources are in all the official languages of Spain (Spanish, Catalan, Galician, Basque). In these cases, the on-line TTS engines can be used to listen to the content.

4.2 Information organization: XML Documents

When the multimedia database is built, the speech recognition engines [1] are used to transcribe the audio files. The transcriptions are scattered in fragments corresponding to 1 minute of the audio session and stored in XML documents. We will consider that a session corresponds with an audiovisual file (a news program, a TV show, etc.); a session will consist of a set of XML documents.

Usually, speech recognition systems do not produce punctuation. Therefore, one of the problems is the construction of the sentences to be translated and synthesized. From the ASR system we can get the pauses made by the speakers.

In the case of Transcrigal-DB and TC-STAR the solution adopted was to place a punctuation mark when a pause is found. But that solution is not appropriate for Agora-DB spontaneous speech. The sentences produced using this approach would be too short and meaningless. At this moment, this is being investigated: a classifier is being trained using both prosodic and morpho-syntactic features.

When the transcriptions are done, the translation engines [1] translate every transcription to any of the following languages: Spanish, Catalan, Galician, English. As it has been mentioned above, translation from and to Basque is not considered in this project. The translations are stored along with the transcription in the same XML document.

Every XML keeps the most relevant information of the document, see listing 1.1:

- The database and the session to which the document belongs.
- The media type of the session.
- The start/end/duration of every sentence.
- The gender and identification of the speaker.
- The sentence translated into the 4 languages.
- Information about the language and transcription method of every sentence.

Listing 1.1: Xml document

```
<document xmlns:xsi="http://www.w3.org/2001/XMLSchema–instance"
   xsi:noNamespaceSchemaLocation=
   'http://www.buceador.org/demo/buceador_document.xsd'>
<document_ID> TCSTAR+20040720_1000_1230_ES_SAT+142 </document_ID>
<session_ID> TCSTAR+20040720_1000_1230_ES_SAT </session_ID>
<database_ID> TCSTAR </database_ID>
<URL_media>
    http://www.webs.uvigo.es/gtm_voz/Buceador/TCSTAR/20040720_1000_ES_SAT.mp3
</URL_media>
<media_type>audio/mpeg</media_type>
<segment segment_ID='0' start='8528727' end='8530667'
   duration='1940' start_turn_time='8528727'>
   <speaker id='spk2' gender='unknown'/>
   <original_language>ES_ES</original_language>
   <orth language='ES_ES' method='asr' version='1.0'>
       Muchas gracias señora frassoni.</orth>
   <orth language='GL_ES' method='asr_es2gl' version='1.0'>
       Moitas grazas señora frassoni.</orth>
   <orth language='CA_ES' method='asr_es2ca' version='1.0'>
       Moltes gràcies senyora frassoni.</orth>
   <orth language='EN_UK' method='asr_es2en' version='1.0'>
       Thank you very much, Mrs Frassoni.</orth>
</segment>
<segment segment_ID='1' start='8530667' end='8533307'
   duration='2640' start_turn_time='8528727'>
   <speaker id='spk2' gender='unknown'/>
   <original_language>ES_ES</original_language>
   <orth language='ES_ES' method='asr' version='1.0'>
     Tiene la palabra señor bösch.</orth>
   <orth language='GL_ES' method='asr_es2gl' version='1.0'>
     Ten a palabra señor bösch.</orth>
   <orth language='CA_ES' method='asr_es2ca' version='1.0'>
     Té la paraula senyor bösch.</orth>
   <orth language='EN_UK' method='asr_es2en' version='1.0'>
     You have the floor , Mr bösch.</orth>
</segment>
</document>
```

For each language, the information retrieval indexes are created. According to the user query in a given language the information retrieval system returns the set of documents that match the query. Each result includes the identification of the XML document. This allows presenting the original transcription, the audio or video resource, etc.

5 User Interface

A web page has been designed to integrate the above mentioned technologies. The web page has been programmed in PHP, Javascript and Java.

- The PHP will attend the different web services (query search, subtitles server and synthesizer servers). It also provides access to the information stored in the XML documents.
- The Javascript provides the controls while playing videos, the highlight of the subtitles, the navigation through the results list, the setup of the media player, the selection of the language, the activation/deactivation of the speech recognition. Most of the Javascript code relies on the jQuery library.
- An applet programmed in Java is used to record the speech of a query and send the audio file to a server which will return the keywords of the search to the web server using speech recognition.

Other technologies used in the web page are:

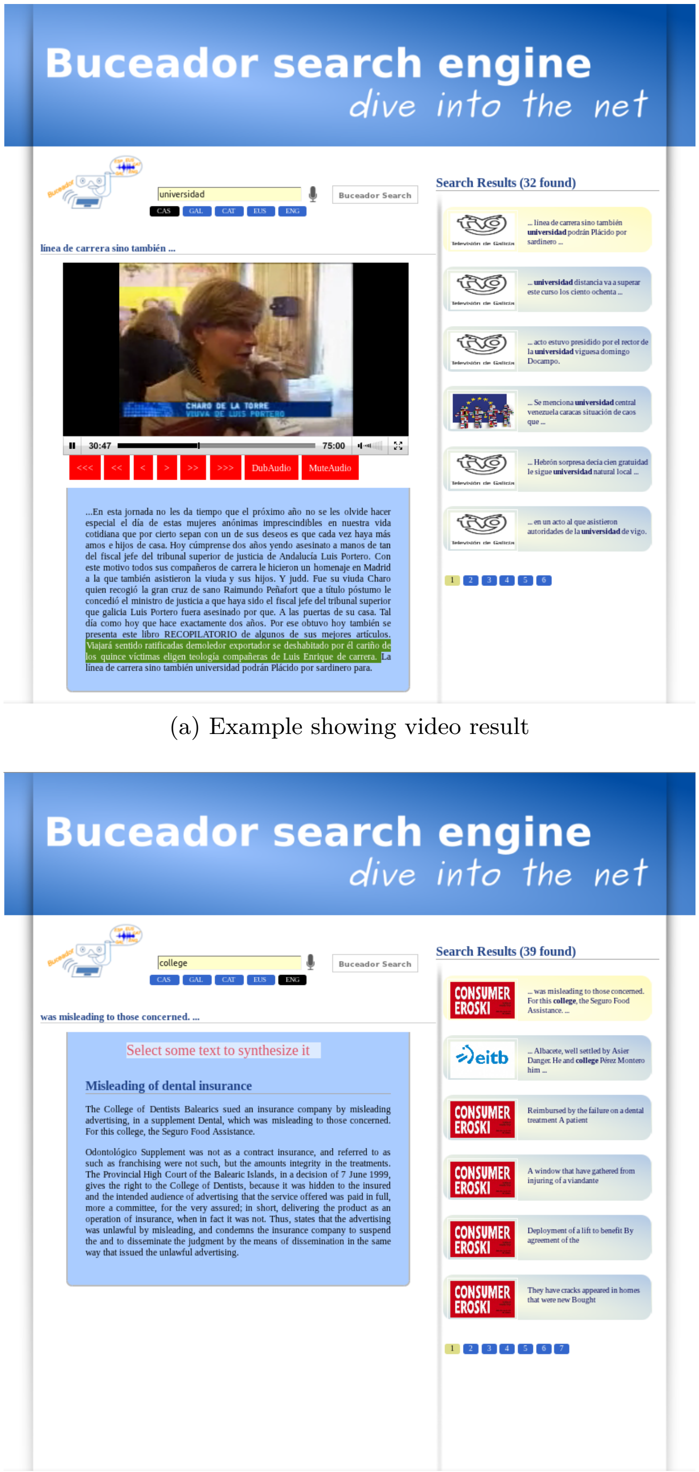
- JW Player: A player that is loaded on the user computer machine responsible of the audio/video playing.
- H264 Streaming module: an apache module for the transmission of H264 video files that enables the user to jump to any part of the video, no matter the length or whether that part has been downloaded yet.
- Rtmp server: the pseudostreaming module is only for H264 video files, so that server is used to transmit audio files.

5.1 Examples

Figure 2a shows a screen capture of the search of the keyword *universidad* (university) in Spanish. There are four main parts:

- The text/audio search field which can be filled either with the keyboard or with the microphone.
- The search results column on the right. Each result can be selected by clicking on the icon.
- The jwplayer part, in this case there is a video playing.
- The subtitles section, with the transcription of the audio. The text highlights are synchronized with the corresponding audio.

Figure 2b is quite similar to the previous figure, but in this case the result is a text document. The user can optionally select part of the text to synthesize it.



línea de carrera sino también ...

...En esta jornada no les da tiempo que el próximo año no se les olvide hacer especial el día de estas mujeres anónimas imprescindibles en nuestra vida cotidiana que por cierto sepan con un de sus deseos es que cada vez haya más amos e hijos de casa. Hoy cúmprense dos años yendo asesinato a manos de tan del fiscal jefe del tribunal superior de justicia de Andalucía Luis Portero. Con este motivo todos sus compañeros de carrera le hicieron un homenaje en Madrid a la que también asistieron la viuda y sus hijos. Y judd. Fue su viuda Charo quien recogió la gran cruz de sano Raimundo Peñafort que a título póstumo le concedió el ministro de justicia a que haya sido el fiscal jefe del tribunal superior que galicia Luis Portero fuera asesinado por que. A las puertas de su casa. Tal día como hoy que hace exactamente dos años. Por ese obtuvo hoy también se presenta este libro RECOPILATORIO de algunos de sus mejores artículos. Viajará sentido ratificadas demoledor exportador se deshabitado por él cariño de los quince víctimas eligen teología compañeras de Luis Enrique de carrera. La línea de carrera sino también universidad podrán Plácido por sardinero para.

Search Results (32 found)

... línea de carrera sino también **universidad** podrán Plácido por sardinero ...

... **universidad** distancia va a superar este curso los ciento ochenta ...

... acto estuvo presidido por el rector de la **universidad** viguesa domingo Docampo.

... Se menciona **universidad** central venezuela caracas situación de caos que ...

... Hebrón sorpresa decia cien gratuidad le sigue **universidad** natural local ...

... en un acto al que asistieron autoridades de la **universidad** de vigo.

(a) Example showing video result

was misleading to those concerned. ...

Select some text to synthesize it

Misleading of dental insurance

The College of Dentists Balearics sued an insurance company by misleading advertising, in a supplement Dental, which was misleading to those concerned. For this college, the Seguro Food Assistance.

Odontológico Supplement was not as a contract insurance, and referred to as such as franchising were not such, but the amounts integrity in the treatments. The Provincial High Court of the Balearic Islands, in a decision of 7 June 1999, gives the right to the College of Dentists, because it was hidden to the insured and the intended audience of advertising that the service offered was paid in full, more a committee, for the very assured; in short, delivering the product as an operation of insurance, when in fact it was not. Thus, states that the advertising was unlawful by misleading, and condemns the insurance company to suspend the and to disseminate the judgment by the means of dissemination in the same way that issued the unlawful advertising.

Search Results (39 found)

... was misleading to those concerned. For this **college**, the Seguro Food Assistance. ...

... Albacete, well settled by Asier Danger. He and **college** Pérez Montero him ...

Reimbursed by the failure on a dental treatment A patient

A window that have gathered from injuring of a viandante

Deployment of a lift to benefit By agreement of the

They have cracks appeared in homes that were new Bought

(b) Example showing text result

Fig. 2: Screenshot captures showing the BUCEADOR Search Engine

6 Summary

In this paper, we have presented the showcase of the BUCEADOR project. The goal of this platform is to provide a tool for search on multimedia and multilingual language resources, and present the information in the user language and in the preferred modality (text or audio/video).

One crucial problem is the punctuation of the transcribed audio, in particular in spontaneous (or non-formal) speech. This is essential to produce good translated sentences and intelligible synthetic speech.

7 Acknowledgments

The research leading to these results has been partially funded by the Spanish Government under grant TEC2009-14094-C00. (BUCEADOR project). It has also been supported from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 247762 (FAUST) and has received funding from the Spanish Ministry of Economy and Competivity through the Juan de la Cierva fellowship program.

References

1. J. Adell, A. Bonafonte, A. Cardenal, M. R. Costa-Jussà, J. A. R. Fonollosa, A. Moreno, E. Navas, and E. R. Banga, "Buceador, a multi-language search engine for digital libraries," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), may 2012.
2. R. Ku, Apache Solr 3.1 Cookbook. Packt Publishing, 1st ed. edition, July, 2011.
3. H. Van den Heuvel, K. Choukri, C. Gollan, A. Moreno, and D. Mostefa, "Tc-star: New language resources for asr and slt purposes," in *Proceedings LREC*, vol. 2006, 2006, pp. 2570–2573.
4. C. García-Mateo, J. Dieguez, L. Docío, and A. Cardenal, "Transcrigal: A bilingual system for automatic indexing of broadcast news." in *LREC*. European Language Resources Association, 2004.
5. H. Schulz and J. A. R. Fonollosa, "A catalan broadcast conversational speech database," in *I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages*, Porto Salvo, Portugal, Sep. 2009.