

# The BUCEADOR multi-language search engine for digital libraries

Jordi Adell<sup>1</sup>, Antonio Bonafonte<sup>1</sup>, Antonio Cardenal<sup>2</sup>, Marta Ruiz<sup>3</sup>,  
José A.R. Fonollosa<sup>1</sup>, Asunción Moreno<sup>1</sup>, Eva Navas<sup>4</sup>, Eduardo R. Banga<sup>2</sup>

<sup>1</sup>Universitat Politècnica de Catalunya, <sup>2</sup>Universidad de Vigo  
<sup>3</sup>Barcelona Media, <sup>4</sup>University of the Basque Country

## Abstract

This paper presents a web-based multimedia search engine built within the Buceador ([www.buceador.org](http://www.buceador.org)) research project. A proof-of-concept tool has been implemented which is able to retrieve information from a digital library made of multimedia documents in the 4 official languages in Spain (Spanish, Basque, Catalan and Galician). The retrieved documents are presented in the user language after translation and dubbing (the four previous languages + English). The paper presents the tool functionality, the architecture, the library and provides some information about the technology involved in the fields of automatic speech recognition, statistical machine translation, text-to-speech synthesis and information retrieval. Each technology has been adapted to the purposes of the presented tool as well as to interact with the rest of the technologies involved.

**Keywords:** multimedia search, multilingual search, speech recognition; machine translation, speech synthesis

## 1. Introduction

There is a growing interest on digital libraries. There exist a lot of databases containing multimedia content that could be useful for many people. For example: newspapers, book collections, TV documentaries, TV shows or news, etc. However, we cannot be aware of the existence of all of them. It is, most of the time, difficult to find relevant information within a library. In addition, the information we are interested in might be in a foreign language that we do not understand.

The presented tool is aimed at solving these problems by making digital libraries accessible to everyone easily and in the user's own language. We expect the search engine to find the information more relevant to the user, given the information supplied. If the information found is in a foreign language, then it is desired that the system translates the media content into the user's language, even if the items found are in the video format.

BUCEADOR is a project focused on advanced research in all core Spoken Language Technologies (SLT), (diarization, speech recognition, speech machine translation, and text-to-speech conversion), and the successful joint integration of all of them in a multilingual and multimodal information retrieval system. The project is supported by the Spanish Government. The research team is composed of four groups: Universitat Politècnica de Catalunya, that coordinates the project, AhoLab group from the Basque Country University, University of Vigo and Barcelona Media.

The goal of the project is to achieve improvements in all the SLT components and voice search applications to improve human-machine and human-to-human communication among all the official languages spoken in Spain (Basque, Catalan, Galician, and Spanish), as well as between these languages and English. As a result, the project provides research advances in each mentioned technology. Examples of such approaches are exploring new techniques for the diarization of speeches, incorporating confidence

measures for unconstrained conversational speech in automatic speech recognition, integrating linguistic knowledge in the statistical approach to spoken language translation, new acoustic and prosodic models for generating expressive speech in synthesis, and implementing new strategies for voice search information retrieval.

In order to show the achievements in the above mentioned technologies and their successful joint integration, the project implements a show case consisting of a search engine for multilingual audiovisual contents. Specifically, broadcast news and debate programs from several TV channels in all the official Spanish languages are utilized.

This paper describes the technologies developed in the project and focuses on the description of the demonstration platform. A web based interface that lets a user enter a query as either text or in spoken form, as well as define the targeted modality of the results and the presentation language, has been designed and developed. Search results are presented in either their original form (text, audio, video) or converted to a textual or audio representation as requested. Search results can also be translated to a different language according to user preferences.

In section 2. a functional description of the overall system is presented. It includes some information about the global architecture. Section 3. describes the digital library which has been include in this proof-of-concept. Section 4. describes the main speech and language engines integrated in the platform. The paper ends with a short summary.

## 2. Overview of the Buceador Platform

The design of the system comprises two different parts. First, there is a subsystem that is used to index new content. Databases can be uploaded to the system for them to be indexed and stored in the library. Databases have to be previously segmented into documents. A document is an item of the database or a chunk of an item, depending on their size. Documents should be small enough to be presented clearly to the user and large enough to be meaningful. Documents

can be long paragraphs or short news in a newspaper, or scenes in movies, or speaker turns in conversations.

Documents can be of four different types: text, audio, image or video. Each document of the database is analysed when uploaded to the platform. For all of them an XML descriptor is generated. It contains relevant information: document ID, database ID, text field, translation, gender, etc.

If a document is of *audio* or *video* media types, an automatic speech recognition (ASR) system is used to create textual information from it. If the document is of *text* or *image* types, this step is skipped. Then, once the description file, has the textual information, a statistical machine translation (SMT) engine is used to translate the textual information into all possible output languages and generates all possible translations. If the document is of media type *text*, then the description file is completed with all possible translations. However, for video and audio media types, the translated text is converted into speech by a text-to-speech (TTS) system. In the case of video, the audio is inserted back into the video as a new stream. Therefore, videos are, in fact, dubbed into the user language.

Language technologies are used in distributed systems. The platform interact with them through Internet. Each partner of the project is responsible for several technologies and languages and are offered to the platform as REST web services.

Search results are presented as a web page (see figure1). Audio, videos, texts and images are presented all together in order of relevance. Each document presented to the user contains information related to the database it comes from. The user can choose to access the content in both his or the original language. The platform also has the functionality of showing subtitles.



Figure 1: Buceador demonstration platform

For accessibility reasons the query can be performed by typing a query in a web text box but also by means of voice. The platform offers the possibility of choosing the language in which to make the query and record our own voice with the query. Then speech is sent to the available recognition systems via a REST web service. The answer of the ASR system is used to make the search in the library as it is done

when the query is typed by the user. The ASR system allows keywords queries. Keywords can be separated by a special words. The possible keywords allowed by the ASR system and their grammar are learnt from the databases uploaded. So the system adapts itself to the content of the library.

The speech recognition API for the spoken queries is composed of a Java applet that runs inside a web page and of a PHP script located on the server side. This Java applet implements a push-on button for recording the audio and sending it to the recognition server. The PHP script is responsible for calling the local recognition engine and sending back the transcription to the applet. The Java applet sends the audio file to the recognition server using POST method and waits for the response of the PHP script. Once the response file is received, the Java applet processes it and pass the results to the search engine.

The platform has been programmed in PHP, *Javascript* and Java. PHP modules are used to access other web services and access information contained in the description files. *Javascript* is used to allow the user to change the presentation of the results: Switch subtitles on and off, navigate through the list of results, choose which token to see and choose its language. Most of the *Javascript* code relies on the jQuery library. A Java applet is used to record the speech for the query and a PHP module interacts with the service uploading the recorded speech and downloading the recognition results. The software produced to implement the platform will be distributed open-source to the community. Therefore anyone will be able to create their own web services and make the platform interact with them.

Figure 1 show a screen capture of the BUCEADOR platform. At the beginning, only the upper part appears, showing the input interface. The user selects the desired language (in the picture, Castilian Spanish) and enter the keywords either using the keyboard, in the text box, or clicking in the microphone symbol. If speech is used, the spoken words are shown in the text box. The multilingual library is search and all the relevant documents are shown in the right column. The most relevant or the selected one is shown in the center. If the documents are found in other language (in this case, in Catalan), the translated transcription and the dubbed audio are presented.

### 3. Multimedia library

The multimedia library contain resources in official languages spoken in Spain (Spanish, Basque, Catalan, Galician).

The Spanish TC-STAR database (Van den Heuvel et al., 2006), is composed of plenary sessions of the European and Spanish parliaments. The speeches from the European Parliament Plenary Sessions (EPPS), obtained via Europe by Satellite. It comprises both, recordings of members and interpreters of the European Parliament speaking in the parliamentary plenary sessions 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 (García et al., 2004) 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 year 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 (Schulz and Fonollosa, 2009) 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: the anchorman is initially presenting 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 again featuring background music. 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.

Also 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). In this case, the on-line TTS engines can be used to listen to the content.

## 4. Speech and Language Engines

This section describes the different engines used for speech acquisition and speech and language processing.

### 4.1. Speech Acquisition

The speech recognition API for the spoken queries is composed of a Java applet that runs inside a web page and of a PHP script located on the server side. This Java applet implements a push-on button for recording the audio and sending it to the recognition server. The PHP script is responsible for calling the local recognition engine and sending back the transcription to the applet. The Java applet sends the audio file to the recognition server using POST method and waits for the response of the PHP script. Once the response file is received, the Java applet processes it and pass the results to the search engine.

### 4.2. Speech Recognition

The speech recognition engines are used in two processes. First, it is used, after the diarization process, to recognize the audio channel in the multimedia sessions. This is done offline, every time that a new session is added to the digital library. Speech recognition is also used to get the user keywords in case the user selects the spoken modality for entering the query.

For Spanish and Galician language BUCEADOR uses the decoder developed on the University of Vigo (Docio-Fernandez et al., 2006), which is based on the standard token-passing algorithm and includes LM look-ahead and a N-best re-scoring stage. The recognition is performed in five consecutive passes. First pass includes gender detection, acoustic segmentation, and a decoding pass using a trigram language model (LM) with 60K words and two-state demiphones with standard MFCC\_E\_D\_A parameterization. The second pass uses acoustic models obtained from unsupervised adaptation (MLLR and MAP). The N-Best lists obtained from this pass are re-scored using a 5-gram LM. Finally, a new unsupervised adaptation and 5-gram re-scoring passes are performed. While the second pass gives relative improvements in WER between a 2% and 6%, (between one and five absolute points), the improvements obtained by the remainder passes together are usually under a 1%.

The UPC recognition system used to transcribe the Catalan audiovisual database has been designed to operate in real time with vocabularies of over 250,000 words and n-gram language models of order 4. The acoustic analysis uses the standard MFCC parameters and the usual first and second time derivatives, while the acoustic HMM models are context dependent demiphones (Mariño et al., 2000) with a optimized pool of tied Gaussian mixtures. The architecture of the decoder is based on the standard weighted finite state transducer (WFST) framework (Mohri et al., 2002) with additional network optimizations for the selected acoustic modeling.

Both systems are also used for the recognition of the queries, but in this case the language model is replaced by a dictionary of keywords. The multi-word keywords are selected automatically from the most frequent ones found in the transcription filtered with a POS-based templates. For instance, keywords with only non-content words are not included.

For Basque, the transcriptions are already part of the digital documents. The AhoSR speech recognition engine is used for recognizing the speech query. AhoSR is a standard speech recognition algorithm based in a Viterbi decoder that uses HMM and has been written in C++ for Windows. The speech signal can be sampled at 8 or 16 kHz using 8 or 16 bit per sample and is parameterized using MFCC coefficients. These MFCC parameters can then be used to train speech models or to be decoded in the recognition process. AhoSR includes a VAD (Luengo et al., 2010) module and has three working configurations: phonetic recognition, speech recognition based in finite grammars and speech recognition based in word loops which al-

lows implementing a simple n-gram language model. The evaluation of AhoSR shows that its results are comparable and in some cases slightly better than the ones achieved using HTK.

### 4.3. Speech Synthesis

There are several speech synthesis engines which are used for dubbing the multimedia documents to the user language. To avoid additional delays, this process is also done offline. Therefore, for each audiovisual document five audio channels are stored: original sound track and the other 4 languages from English, Spanish, Basque, Catalan and Galician.

For Spanish and Basque the AhoTTS is applied. AhoTTS is a multilingual TTS (Hernández et al., 2001) system developed by the Aholab Signal Processing Group of the University of the Basque Country for commercial and research purposes. The system has a modular architecture allowing developers to work on different modules of the system at the same time. It is composed of three main processing modules: the text processor, the linguistic module and the synthesis engine. In addition, two databases are used: one dictionary containing morphological and phonetic information about the words and the synthesis database that contains the recordings manipulated to generate the synthetic speech. The language dependent linguistic module extracts linguistic features from the input text. The acoustic engine uses them to select previously trained statistical models and generate a sequence of acoustic parameters (log-f0, Mel-cepstral coefficients, and maximum voiced frequency). Finally, a high quality vocoder, ahoCoder (Erro et al., 2011), constructs the synthetic speech signal from the aforementioned parameters. Currently the system works in Basque (Erro et al., 2010), Spanish (Sainz et al., 2010) and English. It is written in C/C++ and is fully functional both in Unix and Windows operating systems.

The Galician channel is produced by the TTS of the University of Vigo. It is based on the classical unit selection technique but with some improvements in intonation modeling and the Viterbi search. Regarding intonation modeling, several candidate intonation contours are generated by means of a unit-selection procedure and a combined search with the selection of the acoustic units takes place (Campillo et al., 2009). A recent implementation of this technique avoids the exponential growth of the computational load as the number of candidate intonation contours is increased (Campillo and Banga, 2011). The computational load is also reduced with some additional considerations about the cost functions (Campillo et al., 2011). The principles of the linguistic module for Galician are described in (González et al., 2008).

Ogmios (Bonafonte et al., 2006), the UPC TTS system produces the Catalan audio channel. Ogmios is a multilingual research system with voices in several languages. It includes many techniques either for text processing, prosody modeling and voice generation, including a unit selection back-end and an interface to HTS.

The English track is produced using Ogmios in the front-

end and AhoTTS for the backend (Sainz et al., 2011).

### 4.4. Statistical Translation

The UPC *n-ii* translation engine provides translation between Spanish, Catalan and English. *n-ii* is an N-gram-based approach which appeared in (Marino et al., 2006). It stems from the Finite-State Transducers paradigm, and is extended to the log-linear modeling framework. A system following this approach deals with bilingual units, called tuples, which are composed of one or more words from the source language and zero or more words from the target one. The N-gram-based systems allow for linguistically motivated word reordering by implementing word order monotonicity. For the pairs Spanish ↔ Catalan and Spanish ↔ English the translation models are estimated from parallel corpora. The translation from Catalan to English uses Spanish as a pivot language.

With reference to the translator for the Galician-Spanish pair, the implementation of University of Vigo uses Moses in its phrase-based mode. It was difficult to get a parallel corpus of reasonable size due to the scarcity of bilingual resources for this pair of languages, but at present we are working with a parallel corpus of about twenty six million words in each language. At this time the BLEU for Spanish to Galician translation (the most difficult direction) is about 83%. Our present work is focused on improving this result by using additional linguistic information in the translation and language models.

Translation from and to Basque is a very challenging task which is out of scope for this project.

### 4.5. Information Retrieval

We have used SOLR (Ku, 2011), which is a high performance search server built using Lucene Core (Gospodnetic et al., 2009). For indexation and query retrieval, we have used standard preprocessing techniques such as tokenizing and lowercasing available in the same SOLR tool. Particularly, we have used:

1. The *StandardTokenizerFactory* which is a good general purpose tokenizer that strips many extraneous characters and sets token types to meaningful values.
2. The *LowerCaseFilterFactory* which creates tokens by lowercasing all letters and dropping non-letters.
3. The *StopFilterFactory* which discards common words specified in the stopwords.txt list. In our case, this list which contains Catalan prepositions and articles.
4. The *WordDelimiterFactory* which splits words into subwords and performs optional transformations on subword groups.
5. And, the *RemoveDuplicatesTokenFilterFactory* which filters out any tokens that are at the same logical position in the token stream as a previous token with the same text. This situation can arise from a number of situations depending on what the "up stream" token

spk_name	agora_2007-03_05_a.spk
text	Gabriel Masfurroll empresari fundador i president de USP Hospitals i exvicepresident del Futbol Club Barcelona Bona nit
init_time	143123
fin_time	148632

Table 1: Example of one file’s segment

filters are – notably when stemming synonyms with similar roots. It is useful to remove the duplicates to prevent idf (inverse document frequency) inflation at index time, or tf (term frequency) inflation at query time.

Table 1 shows an example of the file’s segment that we are indexing and its fields. The indexation fields contain speaker identification, file identification, starting and ending time in the file and content text in the different languages. In the final application, the search is always done by content, which corresponds to the speaker discourse, in the user language.

## 5. Summary

In this paper we have presented the demonstration platform of the Buceador project. The goal of the tool is to provide access to digital libraries that include audio, video, text and images breaking the language barrier. The user can get the information of the documents in the desired modality and language independently of the original format. Therefore, advanced speech technologies, machine translation and information retrieval technologies are required. The paper also presents briefly the speech technologies developed in the Buceador project.

## 6. Acknowledgment

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 Competitiveness through the Juan de la Cierva fellowship program.

## 7. References

Antonio Bonafonte, Pablo D. Agüero, Jordi Adell, Javier Pérez, and Asunción Moreno. 2006. Ogmios: The UPC text-to-speech synthesis system for spoken translation. In *Proc. of TC-Star Workshop*, Barcelona, Spain, June.

F. Campillo and E.R. Banga. 2011. Multiple f0 contour parallel viterbi search for unit selection speech synthesis. *Electronic Letters*, 47(16):937–938.

F. Campillo, J. van Santen, and E.R. Banga. 2009. Integrating phrasing and intonation modelling using syntactic and morphosyntactic information. *Speech Communication*, 51(5):452–465.

F. Campillo, I. Nozhov, and E.R. Banga. 2011. Segmentwise unit selection. *Electronic Letters*, 47(9):569–570.

L. Docio-Fernandez, A. Cardenal-Lopez, and C. Garcia-Mateo. 2006. TC-STAR 2006 automatic speech recognition evaluation: The uvigo system. In *TC-STAR Workshop on Speech-to-Speech Translation*, pages 145–150.

Daniel Erro, Iñaki Sainz, Eva Navas, and Inma Hernández. 2010. HMM-based speech synthesis in basque language using HTS. In *Actas de las VI Jornadas en Tecnologías del Habla*, pages 67–70, Vigo, Spain, November.

Daniel Erro, Iñaki Sainz, Eva Navas, and Inma Hernández. 2011. Hnm-based mfcc+f0 extractor applied to statistical speech synthesis. In *Proc. of ICASSP*, pages 4728–4731. IEEE.

C. García, J. Dieguez, L. Docío, and A. Cardenal. 2004. Trancrigal: A bilingual system for automatic indexing of broadcast news. In *LREC2004: IV International Conference on Language Resources and Evaluation*.

M González, E. R. Banga, F. Campillo, F Méndez, L. Rodríguez Liñares, and G. Iglesias. 2008. Specific features of the galician language and implications for speech technology development. *Speech Communication*, 50:874–887.

Otis Gospodnetic, Hatcher Erik, and Michael McCandless. 2009. *Lucene in Action*. Packt Publishing, 2nd ed. edition, June.

Inma Hernández, Eva Navas, J.L. Murugarren, and Extebarria B. 2001. Description of the ahotts system for the basque language. In *Proc. of the 4th ISCA Speech Synthesis Workshop*, Pitlochry, Scotland, August.

Rafal Ku. 2011. *Apache Solr 3.1 Cookbook*. Packt Publishing, 1st ed. edition, July.

I. Luengo, E. Navas, I. Odriozola, I. Saratxaga, I. Hernández, I. Sainz, and D. Erro. 2010. Modified LTSE-VAD algorithm for applications requiring reduced silence frame misclassification. In *Proc. of LREC Conf.*, pages 1539–1544, Valletta, Malta, May.

J.B. Marino, R.E. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa, and M.R. Costa-jussà. 2006. N-gram-based machine translation. *Computational Linguistics*, 32(4):527–549.

José B Mariño, Albino Nogueiras, Pau Pachès-Leal, and Antonio Bonafonte. 2000. The demiphone: An efficient contextual subword unit for continuous speech recognition. *Speech Communication*, 32(3):187–197, October.

Mehryar Mohri, Fernando Pereira, and Michael Riley. 2002. Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16, January.

I. Sainz, D. Erro, E. Navas, I. Hernández, J. Sánchez, I. Saratxaga, I. Odriozola, and I. Luengo. 2010. Aholab speech synthesizers for albayzin2010. In *Actas de las VI Jornadas en Tecnologías del Habla*, pages 343–348, Vigo, Spain, November.

Iñaki Sainz, Daniel Erro, Eva Navas, Jordi Adella, and Antonio Bonafonte. 2011. BUCEADOR hybrid TTS for blizzard challenge 2011. In *Proc. of The Blizzard Challenge 2011*, Turin, Italy, September.

Henrik Schulz and José A. R. Fonollosa. 2009. A catalan broadcast conversational speech database. In *I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages*, Porto Salvo, Portugal, September.

H. Van den Heuvel, K. Choukri, C. Gollan, A. Moreno, and D. Mostefa. 2006. Tc-star: New language resources for asr and slt purposes. In *Proceedings LREC*, volume 2006, pages 2570–2573.