Information search engine for multilingual audiovisual contents: BUCEADOR

Asunción Moreno

TALP Research Center, Technical University of Catalonia, Spain
asuncion.moreno@upc.edu

Abstract

BUCEADOR is a three years project focused on advanced research in the core Spoken Language Technologies (SLT) such as 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.

In order to show the achievements of the project in the above mentioned technologies and their successful joint integration, a show case, consisting of a search engine for multilingual audiovisual contents, will be created. Specifically, the audiovisual content includes broadcast news from several TV, radio, and internet channels in all the official Spanish languages (Spanish, Catalan, Basque and Galician) plus English.

Index Terms: Diarization, speech recognition, speech synthesis, speech to speech translation.

1. Introduction

BUCEADOR is a three years project funded by the Spanish Government. The project is driven by three public Spanish Universities with the collaboration of Barcelona Media – Centre d'Innovaciò (BMCI) research center. The project focuses on advanced research in all core Spoken Language Technologies (SLT), (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 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 and between them and English. The aim of the project is to obtain research advances in each SLT technology. New techniques will be explored for the diarization of speeches; in speech recognition, unconstrained conversational speech recognition systems will be implemented in several languages; in machine translation, new machine learning algorithms and linguistic knowledge will be incorporated; in speech synthesis, new acoustic and prosodic models will be implemented..

In order to show the achievements in the above mentioned technologies and their successful joint integration capability, the project will create a show case consisting of a search engine for multilingual audiovisual contents. Specifically, broadcast news from several TV, radio, and internet channels in all the official Spanish languages (Spanish, Catalan, Basque and Galician) plus English will be utilized. The system query and the audiovisual files, oral and textual, can be in any of these languages. By means of diarization, speech recognition and machine translation techniques, the search engine and the information retrieval system can explore the contents of all the files independently of the language in which they were created. Finally, the user of the system can choose the output format, either text or speech, in the language he/she chooses.

This paper is structured as follows: Section 2 describes the current state of the art in the mentioned technologies, Section 3 describes the general objectives of the project and the specific objectives to be achieved in each technology; Section 4 briefly describes the groups involved in the project. The paper ends with acknowledgments in section 5 and the references list in section 6.

2. Involved Technologies

The speech technologies involved in the BUCEADOR project show different research achievements and challenges

2.1. Speech Recognition

The objective of achieving truly human-level automatic transcription of speech, has received a vast amount of research efforts during past decades. As a result, the automatic speech recognition (ASR) technology has converged to an almost universal standard based in HMMs (Hidden Markov Models) and statistical N-Gram language models (LM). With this baseline technology, systems today are capable of achieving quite impressive performance levels in complex tasks, but with extreme sensitivity to mismatches between training and real conditions. Current challenge in ASR technology is to achieve universal systems, robust enough to resist changing acoustic conditions, and adaptable to different speakers and language styles. Most of the research conducted on projects like TC-STAR [1], CHIL [2], AMIDA [3] and EMIME [4] or in public evaluations like NIST [5] is directed to this objective.

Modern recognition systems are very complex programs, comprised of several interconnected modules which perform a variety of tasks: audio diarization, baseline decoding, acoustic model adaptation, language model adaptation, N-Best rescoring, etc. The task of audio diarization is to segment the audio stream into voice and non-voice regions, and to divide an audio stream into speaker homogeneous segments, determining which segments come from the same speaker. The state-of-art speaker diarization systems can de divided into three categories according to the approach used for speaker segmentation: metric-based methods, model-based techniques and hybrid one. The segmentation process can be done in a single pass or in multiple passes through the acoustic data. The most widely applied approach of clustering used in

diarization systems is the hierarchical agglomerative clustering with a BIC based stopping criterion.

In the decoding process, the main problem to solve is to achieve an adequate treatment for the high variability of the acoustics, language and vocabulary. Acoustic mismatches appear due to variations in the environmental noise or because of the inherent inter-speaker variability. Modern recognizers include parameterizations that partially incorporate characteristics of the human ear, as well as speech enhancement techniques that increase the robustness against noise (MFCC or PLP front-ends). Regarding the inter-speaker variability, normalization of the length of vocal tract (VTLN) and the adaptive training of speaker are common approaches. Most modern recognition systems also employ a multi-pass strategy, in which the acoustic models are adapted using MLLR,or MAP algorithms, in an unsupervised way. A confidence measure may be used to exclude the audio segments that are more likely to be incorrectly transcribed.

Linguistic mismatches, mainly caused by changes in style and topic, are also commonly tackled using unsupervised adaptation strategies, while vocabulary mismatches (a particularly serious problem with agglutinative languages like Basque, or inflecting languages like Spanish and Galician) are palliated using statistical methods to compose and select the vocabulary. The recognition of spontaneous speech shows a specific and complex problematic due to the presence of "filled pauses", hesitations, repetitions, false starts, etc. These phenomena are very difficult to model, both from the lexical point of view, but also from the linguistic one. Current research includes the use of prosodic information and specialized language modelling.

2.2. Machine Translation Technologies.

Machine Translation (MT) constitutes a research area that has gained much attention worldwide during the last years. Specifically in the European Community, where the language diversity still represents an important drawback for the integration process, a large amount of resources has been invested in R&D in this technology. As a representative of such an effort, the following projects can be mentioned: C-Star [6], Eutrans [6], Verbmobil [8], LC-Star [9], Nespole! [10], Fame [11], TC-Star [1], SMART [12] and EURO-MATRIX [13]. However, no matter the recent progress, MT technology is still far from achieving satisfactory performance and quality levels.

From the point of view of practical applications, MT can be categorized into two specific problems: written language machine translation (WLMT) and spoken language machine translation (SLMT). The problem of spoken language translation (SLMT), in addition to the specific MT complexities, involves two additional problems: first, those related to the nature of spoken language, such as spontaneity and poor structured (or even lack of) syntax; and second, those related to the state-of-the-art in automatic speech recognition technologies (ASR), such as recognition errors. These problems explain why statistical methods outperform rule-based methods in speech translation. Additionally, in recent years, the second problem had a relevant impact in the integration of ASR and MT systems. In the case of WLMT written language is much more controlled with respect to sentences' grammar content. This allows MT algorithms to explore richer linguistic information starting from lexical level up to the syntactical and semantic levels. However, during the last years the results of international evaluation campaigns carried out by NIST [5] or promoted by specialized conferences (IWSLT09, ACL08) showed that the performance of statistical machine translation systems is comparable to the performance of ruled-based systems, and can be even better when working on restricted domains.

This project considers the MT problem for both, WLMT and SLMT. It focuses on translation tasks among three languages spoken in Spain (Spanish, Galician and Catalan), as well as on translation tasks between them and English, as far as bilingual resources and methodologies are available to the project. Continuing the previous work of the consortium on MT, the selected approach is Statistical Machine Translation. Further effort will be invested in incorporating, using, and adapting available knowledge-based tools and linguistic resources. In particular, special attention will be paid to the reordering problem and to the integration of morpho syntactic knowledge and bilingual dictionaries.

2.3. Speech Synthesis

The synthetic speech produced by state-of-the-art TTS systems does not always fit the actual speech task. Considering speaking styles in synthetic speech produces an improvement in its quality and naturalness. If the desired style deviates only slightly from the standard speech of the TTS system, the style can be simulated with adapted prosody or generating the correct prosody using corpus based techniques. In other cases deeper adaptations are needed and spectral features as well as prosodic features must be controlled. Some works propose speaking style interpolation and adaptation for HMM-based speech synthesis. In most cases flexible and high quality speech synthesis techniques like HMMs are used.

In the building of voice transformation systems different speech models and analysis/reconstruction techniques allowing spectral transformations have been successfully applied: LP-PSOLA or FD-PSOLA, sinusoidal harmonic models, hybrid models, and STRAIGHT. In general, the voice conversion functions are trained from a set of aligned acoustic features that captures the source-target phonetic correspondence. Several alignment strategies can be adopted, depending on the requirements of the spectral envelope transformation method applied by the system. These strategies involve mapping between acoustic classes, or frame-level alignment when a parallel training corpus is available, or alignment techniques for non-parallel training corpora, although some systems do not require an explicit source-target alignment. There is still a trade-off between the quality of the converted speech and the similarity between converted and target voices. Despite the recent appearance of new methods with very good similarity-versus-quality balance, a higher quality is still desirable in some applications.

There is a lack of general agreed criteria for the evaluation of TTS systems. One attempt to establish a common framework for TTS evaluation is the Blizzard Challenge organized since 2006 [14] where the participants must build a TTS system using the same database. In a similar way, Albayzin 2008 TTS evaluation campaign was organized to evaluate Spanish TTS systems. This kind of evaluation campaigns is costly in time and effort and some objective evaluation method that complements them is desired.

3. Objectives

The objectives of the project can be sumarized as making advanced research in all the spoken language techologies necessary to create a multilingual and multimodal information retrieval system. This ambitious objective implies the improvement of all the involved technologies and the specific problems created due to their interaction. Specifically:

- In speech recognition it is necessary to work in unconstrained domains. It is necessary to improve the robustness of the system against speakers, noise, and unknown vocabulary. Diarization techniques should be improved and closely interact with the ASR system.
- In speech translation is necessary to improve the current performance when the SMT system works in unconstrained domains and with automatic transcriptions from oral discourses. Lack of punctuation, false starts, hesitations and recognition errors reduce the performance of the SMT systems. New algorithms and knowledge integration should be incorporated.
- In speech synthesis, expresive speech and high quality text to speech conversion are the main goals to be achieved. Machine learning techniques will be applied to improve prosody and speaking styles. In a speech to speech translation system, the purpose is that the translated voice be able to imitate the style, and even the voice of the source speaker.

In the field of information retrieval, a state of the art system will be implemented. New techniques will be studied to improve the rank of the retrieved information.

3.1. Specific objectives in Diarization, Segmentation and Speech Recognition

The objectives in this area are to improve the performance of the University of Vigo (UVIGO) ASR system for Spanish and Galician languages, to develop a new system for Basque language and to implement a spoken query recognizer. These main objectives may be split in the following sub-objectives

- Development of a robust audio pre-processing module for audio classification and speaker segmentation and classification
- Improvement of the UVIGO LVCSR performance for spontaneous speech by means of automatic detection of speech disfluencies.
- Development of confidence measures for continuous speech recognition.
- Incorporation of unsupervised acoustic adaptation mechanisms using confidence measures in the LVCSR.
- Development of statistical methods for optimizing the recognition vocabulary in flexible and agglomerative languages.

3.2. Specific objectives in Speech translation

The main objective in Speech Translation is to develop high quality statistical machine translation (SMT) systems for all the pairs among the Spanish, Galician, Catalan and English languages. This main objective can be divided into the following sub-objectives:

- Development of algorithms concerning both, word reordering and discriminative training to improve the translation system performance.
- Integration of morpho-syntactic knowledge into SMT.
- Development of an efficient methodology to use Spanish as a pivot language.
- Develop a speech to speech translation system. For this purpose, the integration of all the speech technologies involved in the project is needed

3.3. Specific objectives in Speech Synthesis

The main objective is to develop high quality expressive text-to-speech synthesis systems for the five languages considered in this project. Sub-objectives include the following,

- Characterization and generation of speaking styles, mainly by the use of the correct modelling of the prosody.
- Development of a voice transformation system allowing the modification of any aspect of the synthetic speech signal that conveys expressivity and speaker identity.
- Development of flexible and high quality TTS systems, considering harmonic/stochastic models and HMM techniques.
- Organization and participation in TTS evaluation campaigns with the goal of defining objective criteria that complement subjective evaluation results.

3.4. Technologies integration

In this task, speech recognition, translation, synthesis technologies and the information retrieval engine are integrated in order to build a complete search system. The achievement of his general objective implies the fulfilment of the following sub-objectives:

- Development of bilingual text resources for Spanish – English/Catalan/Basque/Galician language pairs, including parallel texts and language processing tools.
- Specification and development audiovisual signal processing tools to generate metadata for an audiovisual repository. Metadata includes diarization, speech recognition and translation of the audiovisual material.
- Implementation of a web based architecture to put into communication the different subsystems.
- Definition and implementation of efficient interfaces among technologies that improve the performance of the overall system.

- Implementation of tools for building a voice and text driven engine for audiovisual information retrieval search among the official languages in Spain (Spanish, Catalan, Galician and Basque) and English. The engine will be a demonstrator of the developed technologies. This demonstrator will be part of the project dissemination.

4. The partners of the project

The objectives described in this project are multilingual, multimodal and interdisciplinary. They require players and organizations with varying skills and expertise in a wide variety of domains; e.g. speech recognition, translation, synthesis, speech and text processing, implementation of search engines, information retrieval, production of language resources, etc. The requirements to carry out such a project are therefore considerable and the feasibility of such a project relies on the close cooperation of different institutions and the support of external companies. In this project each group will contribute with its specific and complementary technological expertise as well as with its previous experience with the various languages involved.

The involved groups are the TALP research center from the Technical University of Catalonia (TALP-UPC), the Aholab Signal Processing Laboratory from the Universidad del País Vasco (EHU-Aholab) and, the Signal Theory Group from the Universidad de Vigo (UVIGO). The Information Retrieval Group from Barcelona Media – Centre d'Innovaciò (BMCI) will collaborate in the project as an external partner and will contribute to the objectives of the project with their experience in Information Retrieval tools.

TALP-UPC [15] provides expertise in three areas: speech recognition, with competitive systems for telephony applications, and a continuous speech recognition system, both in Catalan, Castilian and English; speech translation in Catalan, Castilian and English; and speech synthesis in Castilian and Catalan. TALP-UPC coordinates the translation activities and integration technologies.

UVIGO [16] brings experience in continuous speech recognition in a complex task such as broadcast news in Galician and Castilian, supplementing TALP-UPC system perfectly. The group has sound experience in speech synthesis in both languages Galician and Spanish. UVIGO coordinates the work of speech recognition.

EHU-Aholab [17] is an expert group in speech synthesis and specifically, in research on prosody. Their synthesis system works in Euskera and Castilian. The group has also developed a speech recognition system in Euskera. EHU-Aholab coordinates the activities of speech synthesis.

The Information Retrieval Group of Barcelona Media – Centre d'Innovaciò (BMCI) [18] is composed of a multidisciplinary team of advanced and starting researches that allows for concentrating, into a single research group, first class expertise and experience in the areas of: information storage and retrieval, natural language processing, machine translation, artificial intelligence, signal processing and data mining technologies. They will set up the search engine in the show case.

5. Acknowledgements

This Project is supported by the Spanish Government grant number TEC2009-14094-C04.

The text of this paper has been extracted from the Technical Annex of the BUCEADOR project. Consequently, all the research team has contributed to it

6. References

[1] http://www.tc-star.org (2004/2007)

[2] http://chil.server.de

[3] http://www.amidaproject.org (EC IST Integrated Project., 2006-2009)

[4] http://www.emime.org (ICT-2007.2.1)

[5] NIST http://www.nist.gov/speech/tests/mt/

[6] http://www.c-star.org/, (1991/2004)

[7] http://cordis.europa.eu/esprit/src/30268.htm (SPRIT-20268-96; SPRIT-30268, 1997/2000)

[8] http://verbmovil.dfki.de/overview-us.html, (1993/2000)

[9] http://www.lc-star.com (IST-2001-32216, 2002/2004)

[10] http://nespole.itc.it/, (IST-1999-11562, 2001/2004)

[11] http://isl.ira.uka.de/fame (IST-2000-28323, 2001/2004)

[12] http://www.smart-project.eu/node/1

[13] http://www.euromatrix.net/ (P6-IST-5-034291-STP)

[14] A.W: Black,.K. Tokuda "The blizzard challenge - 2005: evaluating corpus-based speech synthesis on common datasets", In Proc. INTERSPEECH, pp. 77-80, 2005.

[15] http://www.talp.cat/talp/

[16] http://www.gts.tsc.uvigo.es/web/index.php

[17] https://bips.bi.ehu.es/aholab/

[18] http://*www.barcelonamedia.org*