Search and access to information contained in the speech of multimedia resources *

Búsqueda y acceso a la información contenida en el habla de recursos multimedia

A. Varona, L.J. Rodriguez-Fuentes, M. Penagarikano, S. Nieto, M. Diez, G. Bordel
University of the Basque Country (UPV/EHU)
GTTS, Department of Electricity and Electronics
amparo.varona@ehu.es

Resumen: El proyecto tiene como objetivo hacer aportaciones científicas e introducir mejoras de tipo tecnológico en el sistema de indexado y búsqueda de contenidos multimedia (*Hearch*) desarrollado por el Grupo de Trabajo en Tecnologías Software de la UPV/EHU. *Hearch* es un buscador de aspecto convencional (como Google, Bing, etc) pero con la capacidad de obtener como resultado segmentos de vídeo gracias a la transcripción automática de sus contenidos de voz. El sistema consta de un *back-end* que capta, procesa e indexa los recursos, y de un *front-end* que permite realizar búsquedas, configurar los distintos módulos y monitorizar el funcionamiento, mediante una interfaz web. Actualmente se encuentra operativa una primera versión de la herramienta que trabaja frente a repositorios de noticias en castellano y euskera (*http://gtts.ehu.es/Hearch/*), aunque está preparada también para tratar con recursos en inglés.

Palabras clave: indexado y búsqueda de información, reconocimiento automático del habla, identificación de la lengua, identificación del locutor

Abstract: The main goal of this project is to make scientific contributions and technological improvements related to the spoken document retrieval system (*Hearch*) developed by the Working Group on Software Technologies of the University of the Basque Country. *Hearch* looks like a conventional search tool (such as Google, Bing, etc) but it is designed to retrieve audio/video segments based on the automatic transcription of speech contents. The system consists of a *back-end* that captures, processes and indexes audio/video resources, and a *front-end* that allows to search contents, configure various modules and display performance statistics through a web interface. An early version of this tool is available (*http://gtts.ehu.es/Hearch/*), which searches and retrieves segments on broadcast news repositories in Spanish and Basque, through it can also deal with resources in English.

Keywords: Information Retrieval, Automatic Speech Recognition, Language Recognition, Speaker Recognition

1. General description

The project spans from January 2010 to December 2012. It continues the developments of two previous projects, leaded by the Working Group on Software Technologies (GTTS) of the University of the Basque Country (UPV/EHU). In those projects, GTTS developed *Hearch*: a spoken document retrieval system which looks like a conventional search tool (such as Google, Bing, etc) but it is actually designed to retrieve audio/video segments based on the automatic transcription of speech contents.

The system is based on a simple and efficient architecture, which allows to replace or integrate new modules in a easy and elegant way (see Figure 1). The architecture consists of four key elements: (1) the crawler/downloader; (2) the audio processing module; (3) the information retrieval module; and (4) the user interface. The crawler/downloader fetches audio and video resources from internet or from local repositories. In the case of video resources, only the audio signal is processed. For the speech

[*] This project has been supported by the Spanish MICINN, under Plan Nacional de I+D+i (project TIN2009-07446, partially financed by FEDER funds)

recognizer to work properly, the audio input is segmented and classified as speech and non-speech and the language in speech segments is identified. The information about segment boundaries, language, word transcription, morphosyntactic analysis, etc is stored in an XML resource descriptor.

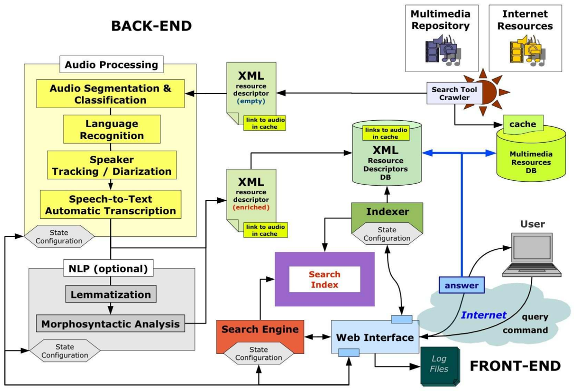

Figure 1: Hearch Architecture

The collection of XML resource descriptors is taken as input by the indexer (which is part of the information retrieval module) to build an index database. The search engine traverses this structure and returns a list of audio and video resources related to any given query. A web interface allows the user to formulate queries and process the answer of the spoken document retrieval system. The system has been developed to deal with spoken documents in Spanish, Basque and English.

2. *Objectives*

From a **scientific point of view**, we aim to make contributions in theoretical issues that eventually may lead to significant improvements in system performance:
1.- Objective and subjective evaluation platforms must be designed and developed. For objective evaluation, we will use a closed list of questions with a list of correct answers determined by a committee of experts. For subjective evaluation we will make a survey of satisfaction on a representative population of users.
2.- Then, we will try to improve system performance, especially regarding the indexing backend, which includes, among other modules, voice/non-voice segment discrimination, language verification and speaker verification.
3.- We will also study and develop techniques to increase the robustness to environmental and channel conditions, and investigate ways of using the information provided by confidence measures in the information retrieval module.

From a **technological point of view**, we have planned the following tasks:
1.- Improving the structure and ergonomics of the user interface, in order to make it easier to migrate to new applications. In particular we will develop an information retrieval system dealing with meeting data (multichannel audio, with close-talk and far-field microphone).
2.- Acquiring a database of broadcast news in Spanish and Basque. It will be done available to the scientific community at the end of the project.
3.- Participating in the international competitive evaluation campaigns organized by the National Institute of Standards and Technology (NIST), especially in the areas of speaker recognition and language recognition.
4.- Integrating the scientific and technological contributions achieved in this project in our spoken document retrieval system, *Hearch*, which will be publicly available through a web interface (http://gtts.ehu.es/Hearch/).

3. *Current state of the project*

Updated information on the progress of the project is made periodically available at the following website: *http://gtts.ehu.es/TWiki/bin/view/Main/SabuesoProject*.
These are the most significant accomplishments by now:
1.- A large database of broadcast news in Spanish and Basque has been acquired from the Basque public television (EITB).
2.- An early version of our spoken document retrieval system on broadcast news repositories in Spanish and Basque is now available (*http://gtts.ehu.es/Hearch/*)
3.- Remarkable progress has been made in language recognition, in part due to the efforts devoted to build various systems for the NIST 2009 LRE. Advances will be incorporated soon into the prototype.
4.- State-of-the-art speaker recognition technology has been developed for the NIST 2010 SRE. The most efficient system will be integrated into the prototype.