Hearch: a Multilingual Spoken Document Retrieval System

Contributors in alphabetical order
G. Bordel, M. Díez, I. Landera, S. Nieto, M. Peñagarikano, L.J. Rodríguez-Fuentes, A. Varona, M. Zamalloa

GTTS - Departamento de Electricidad y Electrónica - Universidad del País Vasco - Spain
german.bordel@ehu.es

Finding audio and video resources in internet has become a highly demanded application. However, search engines are usually limited to adjacent texts (hand supplied transcripts or close captions) to index and classify multimedia documents. Using automatic speech recognition and natural language processing technologies allow transcribing and enriching spoken documents, thus leading to more accurate indexes and more focused search results. Here, a multilingual (Basque, Spanish, English) spoken document retrieval system is presented. The system, organized around a collection of XML resource descriptors, consists of four main elements: (1) a crawler/downloader; (2) an audio processing module; (3) an information retrieval module and (4) a user interface (see Figure 1).

The crawler/downloader fetches audio and video resources from internet or from local repositories and builds an initial XML resource descriptor including info about URLs, size, format, etc.

The audio processing module enriches the XML with information extracted from the audio signals. An interesting issue at this point is the architecture of Hearch: a piping system allows processing the input signal through an arbitrary number of steps, making use of external commands and coping with the parsing and regeneration of the XML. In particular, for the automatic speech-to-text transcription task, HTK as well as Sautrela packages have been used.

For the speech recognizer to work properly the audio input is previously segmented and classified as speech or non-speech and the language in speech segments is identified. All the information about segment boundaries, language, word transcription, morphosyntactic analysis, etc. is obtained by different specialized stages and is incrementally stored in the XML resource descriptor.

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. A search engine, based on Lucene, 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 answers of the SDR system, which are ordered according to their relevance. The user can select one of the segments matching the query and watch the video, whose transcription is presented on its side (see Figure 2).

This work has been jointly funded by the University of the Basque Country under project INF09/29, and the Spanish MICINN under project TIN2009-07556.

1 Sautrela is a highly modular and pluggable open source development package for speech processing applications. It unifies in a single framework almost all the tasks related to pattern recognition such as signal processing, model training and decoding.