Automatic Speech Transcription

Overview

Automatic transcription of audio and video recordings is a complex process that includes: audio segmentation, segment classification and clustering, and speech transcription. During the TOSCA-MP project we addressed automatic transcription of videos of two different genres: television news and talk-show. Videos were from several content providers and were in in four different languages: Dutch, English, German and Italian. During the project automatic transcription systems for the German and Dutch (Flemish variant) languages have been developed. As a result, the capability of the FBK transcription technology to cope with different languages has been extended and currently it covers the following languages: Arabic, Dutch, English, French, German, Italian, Portuguese, Russian, Spanish and Turkish.

In depth description

The FBK internal Hidden Markov Model (HMM) toolkit was employed for developing the transcription systems. For n-gram Language Models (LMs) training, the IRSTLM toolkit was used. The IRSTLM toolkit features algorithms and data structures suitable to estimate, store, and access very large n-gram LMs (http://sourceforge.net/projects/irstlm).
The FBK transcription system is based on several processing stages. The input of the transcription process is assumed to be a file containing an audio recording, e.g. the audio track of a video. The flow of processing consists of several steps:

1. *Segmentation, classification and clustering*
2. *Acoustic features extraction*
3. *Acoustic features normalization*
4. *HLDA projection*
5. *First decoding step*
6. *Unsupervised adaptation/normalization*
7. *Second decoding step*

The generated output, in addition to the recognized words and the corresponding time markers, includes the results of audio segmentation, segment classification and clustering. Furthermore, optionally, the system provides for each speech segment a word graph representing multiple recognition hypotheses. Possibly, the system exploits parallelization with a load-balanced dispatching of segments across several decoder instances.

Potential fields of Application

The technology for automatic transcription of audio and video files has a great potential for application in many application areas such as information retrieval (by enabling automatic indexing), media monitoring (of TV channels and radio stations), video subtitling and captioning (by supporting the manual workflow), media intelligence, etc.

Possibilities for exploitation

The exploitation of results achieved in automatic transcription will go through the FBK spin-off company Pervoice (http://www.pervoice.it).

Further Information

Further technical information is available in TOSCA-MP Deliverables D2.1, D2.2 and D2.3 “Automatic Metadata Extraction and Enrichment”.

© TOSCA-MP consortium: all rights reserved

Contact Person

Diego Giuliani
Fondazione Bruno Kessler
Human Language Technology research unit
Via Sommarive 18, I-38123
Trento – ITALY
giuliani@fbk.eu

© TOSCA-MP consortium: all rights reserved