Audio Analysis

OVERVIEW
The main objective of this module is to analyse the audio component of video items. It provides the video content structure by identifying segments with homogenous audio characteristics/features and by supplying a descriptive label for the content (such as speech, male, speaker A, noise, etc). It finally performs automatic speech recognition (ASR) on those segments that have been identified as speech segments, thus providing a transcript of the spoken words.

The functionality of this module includes extracting the audio track from the video item and classifying the audio stream into different features. The module integrates speaker diarisation (speaker segmentation of the input signal followed by speaker clustering of the created segments into homogenous groups), speech/non-speech classification and gender classification. The results of speaker diarisation are used as input for ASR and for the multimodal analysis.

The ASR component of the module relies on an Acoustic Model, a Vocabulary and a customised Language Model. Its output is used as input for the Textual Analysis modules. The module allows manual corrections, if necessary, of the transcriptions to enhance the reliability of the Papyrus system whenever needed, by using third party open source software (such as Transcriber).

INNOVATION
The module integrates state of the art methods in speaker diarisation allowing the user to select different combinations of them according to the sound characteristics and to the segmentation target. Default settings are tuned on documentary-like videos to provide suitable results for content structuring purposes and multimodal analysis, on the one side, as well as to optimise performance on continuous, spontaneous, large-vocabulary
speech recognition of different speakers, over different channels, in a noisy environment.

The speech recognition component takes advantage of a specifically built domain Language Model that improves the Word Recognition Rate.

BUSINESS IMPACT
This module can be used at different levels: from audio segments boundaries identification and labelling, to speech recognition and transcription. It offers the possibility to provide additional semantic descriptors/metadata, which are hidden in the audio characteristics of news items. By doing so, the end user performing a query request can take advantage of these metadata fields in order to gain access to the most relevant content.

INTEROPERABILITY
The module is implemented as a web service and can be accessed through a RESTful interface. The web service is implemented using Tomcat 6 and the jax-ws/jersey libraries. If the module is applied to a new domain, it would benefit of a domain Language Model provision (that can be derived from a collection of documents related to the domain).

Contact: Roberta Turra - r.turra@cineca.it

PAPYRUS is a project partly funded by the EC FP7 ICT Programme