NoTube

Networks and ontologies for the transformation and unification of broadcasting and the Internet

FP7 – 231761

D4.3 Audio/Video content analysis component

Coordinator Anne-Lore Mevel (TVN)

With contributions from:
L Vignaroli (RAI), F Negro (RAI), R Del Pero (RAI), G Spikofski (IRT), J Groh (IRT), M Elser (IRT), P Altendorf (IRT)

Quality Assessor: Dong Liu (OU)
Quality Controller: Pavel Mihaylov (OT)

<table>
<thead>
<tr>
<th>Document Identifier:</th>
<th>NoTube/2011/D4.3 v3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version:</td>
<td>2.3</td>
</tr>
<tr>
<td>Date:</td>
<td>30/10/2011</td>
</tr>
<tr>
<td>State:</td>
<td>Final</td>
</tr>
</tbody>
</table>
EXECUTIVE SUMMARY

This deliverable describes the functional description of audio/video content analysis component.

The initial version of this document (delivered in M13) provided a general overview of audio/video content analysis components. The second version (delivered in M23) aimed at giving updates of previous contributions on the Ad insertion and on the loudness harmonisation. The current document (due in M33) aims at giving final updates on the audio and video content analysis component.

These updates address more specifically the following subjects:

- Integration of ROI technology in the Personalized Semantic News use case
- Enhancements of Ad insertion
- Description of loudness normalisation, User evaluation of loudness harmonisation on the web
### DOCUMENT INFORMATION

<table>
<thead>
<tr>
<th>IST Project Number</th>
<th>FP7 - 231761</th>
<th>Acronym</th>
<th>NoTube</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Title</td>
<td>Networks and ontologies for the transformation and unification of broadcasting and the Internet</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Project URL</td>
<td><a href="http://www.notube.eu/">http://www.notube.eu/</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document URL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EU Project Officer</td>
<td>Leonhard MAQUA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Deliverable Number</th>
<th>4.3 v3</th>
<th>Title</th>
<th>Audio/Video content analysis component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work Package Number</td>
<td>4</td>
<td>Title</td>
<td>Metadata Acquisition and Enrichment</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date of Delivery</th>
<th>Contractual</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status</td>
<td>version 2.4</td>
<td>final</td>
</tr>
<tr>
<td>Nature</td>
<td>prototype☑</td>
<td>report□</td>
</tr>
<tr>
<td>Dissemination level</td>
<td>public☑</td>
<td>consortium□</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Authors (Partner)</th>
<th>TVN, RAI, IRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Responsible Author</td>
<td>Name</td>
</tr>
<tr>
<td></td>
<td>MEVEL</td>
</tr>
<tr>
<td></td>
<td>Partner</td>
</tr>
<tr>
<td></td>
<td>TVN</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Abstract (for dissemination)</th>
</tr>
</thead>
<tbody>
<tr>
<td>This deliverable describes the functional description of audio/video content analysis component.</td>
</tr>
<tr>
<td>The initial version of this document (delivered in M13) provided a general overview of audio/video content analysis components. The second version (delivered in M23) aimed at giving updates of previous contributions on the Ad insertion and on the loudness harmonisation. The current document (due in M33) aims at giving final updates on the audio and video content analysis component.</td>
</tr>
<tr>
<td>These updates address more specifically the following subjects:</td>
</tr>
<tr>
<td>- Integration of ROI technology in the Personalized Semantic News use case</td>
</tr>
<tr>
<td>- Enhancements of Ad insertion</td>
</tr>
<tr>
<td>- Description of loudness normalisation, User evaluation of loudness harmonisation on the web</td>
</tr>
</tbody>
</table>
## Keywords

ROI, SOI, Segmentation, Classification, Loudness normalisation, Loudness harmonisation

<table>
<thead>
<tr>
<th>Version Log</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Issue Date</strong></td>
</tr>
<tr>
<td>13.01.2010</td>
</tr>
<tr>
<td>15.01.2010</td>
</tr>
<tr>
<td>28.01.2010</td>
</tr>
<tr>
<td>29.01.2010</td>
</tr>
<tr>
<td>30.01.2010</td>
</tr>
<tr>
<td>02.02.2010</td>
</tr>
<tr>
<td>19.02.2010</td>
</tr>
<tr>
<td>22.02.2010</td>
</tr>
<tr>
<td>23.02.2010</td>
</tr>
<tr>
<td>15.11.2010</td>
</tr>
<tr>
<td>30.11.2010</td>
</tr>
<tr>
<td>03.12.2010</td>
</tr>
<tr>
<td>18.12.2010</td>
</tr>
<tr>
<td>20.12.2010</td>
</tr>
<tr>
<td>15.09.2011</td>
</tr>
<tr>
<td>30.09.2011</td>
</tr>
<tr>
<td>07.10.2011</td>
</tr>
<tr>
<td>20.10.2011</td>
</tr>
<tr>
<td>25.10.2011</td>
</tr>
<tr>
<td>Participants</td>
</tr>
<tr>
<td>--------------</td>
</tr>
</tbody>
</table>
| Vrije Universiteit Amsterdam | Guus Schreiber  
Phone: +31 20 598 7739/7718  
Email: schreiber@cs.vu.nl |
| British Broadcasting Corporation | Libby Miller  
Phone: +44 787 65 65 561  
Email: Libby.Miller@bbc.co.uk |
| Pro-netics | Marco Bruni  
Phone: +39 06 45472503  
Email: marco.bruni@pro-netics.com |
| Engin Medya Hizmetleri A.S. | Ron van der Heiden  
Phone: +31 6 2003 2006  
Email: ron@engin.tv |
| Institut fuer Rundfunktechnik GmbH | Christoph Dosch  
Phone: +49 89 32399 349  
Email: dosch@irt.de |
| Ontotext AD | Atanas Kiryakov  
Phone: +35 928 091 565  
Email: naso@sigma.bg |
| Open University | Stefan Dietze  
Phone: +44 1908 858 217  
Email: s.dietze@open.ac.uk |
| RAI Radiotelevisione Italiana SPA | Alberto Morello  
Phone: +39 011 810 31 07  
Email: a.morello@rai.it |
| Semantic Technology Institute International | Lyndon Nixon  
Phone: +43 1 23 64 002  
Email: lyndon.nixon@sti2.org |
| Stoneroos B.V. | Annelies Kaptein  
Phone: +31 35 628 47 22  
Email: annelies.kaptein@stoneroos |
| Thomson Video Networks | Raoul Monnier  
Phone: +33 2 99 27 30 57  
Email: raoul.monnier@thomson-networks.com |
| Polymedia, SpA | Tullio Pirovano  
Phone: +39 02 257711  
Email: tullio.pirovano@polymedia.it |
| KT Corporation | Myoung-Wan Koo  
Phone: +82 2 526 6347  
Email: mskim@kt.co.kr |
TABLE OF CONTENTS

LIST OF FIGURES .................................................................................................................. 8
LIST OF TABLES ....................................................................................................................... 8
LIST OF ACRONYMS ................................................................................................................. 9

1 INTRODUCTION .................................................................................................................. 10
2 SCOPE OF THIS DELIVERABLE ......................................................................................... 10
3 AUDIO/VIDEO CONTENT ANNOTATION ........................................................................... 11
   3.1 AUDIO ANNOTATION ................................................................................................. 11
   3.2 VIDEO ANNOTATION ............................................................................................... 11
4 VIDEO REFRAMING BASED ON ROI TECHNOLOGY ....................................................... 13
   4.1 AUTOMATIC REFRAMING ....................................................................................... 13
      4.1.1 Overview ........................................................................................................... 13
      4.1.2 Visual attention model (VAM) ........................................................................... 14
         4.1.2.1 Existing VAM models ............................................................................... 14
         4.1.2.2 Model chosen for re-framing applications ................................................. 14
         4.1.2.3 Computation of the spatial salience ........................................................... 16
         4.1.2.4 Computation of the temporal salience ....................................................... 16
         4.1.2.5 Computation of the rarity map .................................................................... 17
      4.2 CROPPING WINDOW EXTRACTION ...................................................................... 17
         4.2.1 From the saliency map to the first cropping box ............................................. 17
         4.2.2 Temporal consistency ..................................................................................... 18
         4.2.3 Aspect ratio ..................................................................................................... 18
   4.3 INTEGRATION WITHIN THE WP7A USE CASE ....................................................... 18
5 AUTOMATIC AD INSERTION ............................................................................................... 20
   5.1 OVERVIEW ................................................................................................................ 20
   5.2 DESCRIPTION OF THE ALGORITHMS ..................................................................... 20
      5.2.1 General description of the 1st pass (metadata generation) .................................. 20
         5.2.1.1 Description of the 1st step: Analysis of each image of a video ...................... 20
         5.2.1.2 Description of the 2nd step: Computation of the average saliency value for each corner of each image ....................................................................... 20
         5.2.1.3 Description of the 3rd step: Selection of several sequences allowing the insertion of the Ad ................................................................. 21
      5.2.2 General description of the 2nd pass (Ad insertion) ............................................. 21
   5.3 AUTOMATIC AD INSERTION FIRST EVALUATION RESULTS .............................. 21
   5.4 IMPROVEMENT OF THE AUTOMATIC AD INSERTION ALGORITHM .................. 22
      5.4.1 Ranking improvement thanks to Audio level analysis ........................................ 23
         5.4.1.1 Specification of the loudness descriptors to be analyzed .............................. 23
         5.4.1.2 Basis of the algorithm improvement ......................................................... 23
      5.4.2 Ranking improvement thanks to scene cuts analysis .......................................... 23
      5.4.3 Ranking improvement thanks to global saliency maps analysis ......................... 23
      5.4.4 Improvement of the Ad visibility ..................................................................... 24
   5.5 INTEGRATION WITHIN THE OVERALL NOTUBE ARCHITECTURE ........................ 24
6 AUTOMATIC PROGRAM SEGMENTATION ......................................................................... 27
   6.1 ANTS FOR RAW DATA ............................................................................................. 27
   6.2 ANTS ARCHITECTURE ............................................................................................. 27
   6.3 SEGMENTATION OF NEWS STORIES ..................................................................... 28
   6.4 SUBJECT CLASSIFICATION AND ENRICHMENTS .............................................. 29
   6.5 INTEGRATION WITHIN THE OVERALL NOTUBE ARCHITECTURE ........................ 29
7 LOUDNESS HARMONISATION FOR THE AUDIO PART OF VARIOUS A/V-STREAMS ENTERING THE NOTUBE TERMINAL ................................................................. 30
   7.1 LOUDNESS OF AUDIO SERVICES: THE ACTUAL STATUS .................................. 30
      7.1.1 Specification of ITU-R BS.1770 ........................................................................... 31
7.1.2 Specification of EBU-R128 and related documents ................................................. 33
7.2 LOUDNESS MODULE IN NOTUBE ........................................................................... 34
  7.2.1 Analysis stage ........................................................................................................ 34
    7.2.1.1 Specification of loudness descriptors to be analyzed ..................................... 34
    7.2.1.2 Description of analysis algorithm .................................................................. 35
    7.2.1.3 LoudnessAnalysis WebService ...................................................................... 39
  7.2.2 Synthesis stage ....................................................................................................... 39
    7.2.2.1 Description of normalisation algorithm ......................................................... 41
    7.2.2.2 Loudness Normalisation Web Service ............................................................ 41
7.3 EVALUATION OF LOUDNESS HARMONISATION BY APPLYING LOUDNESS DESCRIPTORS DERIVED IN THE ANALYSIS STAGE ........................................................................... 42
  7.3.1 Introduction ............................................................................................................ 42
  7.3.2 Concept .................................................................................................................. 42
  7.3.3 Test items .............................................................................................................. 43
  7.3.4 Test set-up ............................................................................................................. 44
  7.3.5 Results ................................................................................................................... 45
7.4 USER EVALUATION OF LOUDNESS HARMONISATION ON THE WEB ................................................................................................................................. 47
  7.4.1 Introduction ............................................................................................................ 47
  7.4.2 Concept .................................................................................................................. 48
  7.4.3 Selection and preprocessing of test items ............................................................. 48
    7.4.3.1 Evaluation of loudness adaptation ................................................................. 48
    7.4.3.2 Evaluation of loudness range adaptation ....................................................... 49
    7.4.3.3 Listening level ............................................................................................... 52
    7.4.3.4 Organisation of loudness web evaluation ...................................................... 53
  7.4.4 Results ................................................................................................................... 53
  7.4.5 Conclusions .......................................................................................................... 57
8 CONCLUSION .................................................................................................................. 58
9 REFERENCES .................................................................................................................. 58
List of Figures

FIGURE 1: GENERAL DESCRIPTION OF AUTOMATIC REFRAMING PROCESS ................................................. 14
FIGURE 2: SALIENCY MAPS ................................................................................................................ 14
FIGURE 3: FLOW CHART OF THE PROPOSED SPATIO-TEMPORAL MODEL ............................................. 15
FIGURE 4: FROM THE SALIENCY MAP TO THE FIRST CROPPING BOX .................................................. 18
FIGURE 5: NEWS ITEM DETAILS WITH PLAY OUT FEATURE ON MOBILE DEVICE ................................ 19
FIGURE 6: REFRAMED VIDEO PLAYED ON A MOBILE DEVICE ............................................................... 19
FIGURE 7: TECHNOLOGY EVALUATION - RESULTS BY TESTER ............................................................ 22
FIGURE 8: TECHNOLOGY EVALUATION - RESULTS BY FILM .................................................................... 22
TABLE 1: LOUDNESS DESCRIPTORS ........................................................................................................ 23
FIGURE 9: NOTube PROTOTYPE DESCRIPTION .......................................................................................... 24
FIGURE 10: WORKFLOW GET LIST ........................................................................................................... 25
FIGURE 11: WORKFLOW JOB CREATE ....................................................................................................... 26
FIGURE 12: VIDEO ANALYSIS WORKFLOW ............................................................................................... 26
FIGURE 13: AD INSERTION WORKFLOW .................................................................................................... 26
FIGURE 14: THE ANTS ARCHITECTURE ..................................................................................................... 28
FIGURE 15: BLOCK DIAGRAM OF MULTICHANNEL LOUDNESS ALGORITHM ........................................... 32
FIGURE 16: RESPONSE OF THE PRE-FILTER USED TO ACCOUNT FOR THE ACOUSTIC EFFECTS OF THE HEAD .......................................................................................................................... 32
FIGURE 17: RLB WEIGHTING CURVE ......................................................................................................... 33
FIGURE 18: BLOCK DIAGRAM OF THE LOUDNESS MEASUREMENT PROCESS .......................................... 38
FIGURE 19: METHOD DESCRIPTION LOUENESS ANALYSIS .................................................................. 39
FIGURE 20: PRINCIPLE OF LOUDNESS RELEVANT METADATA APPLICATION ..................................... 40
FIGURE 21: WORKFLOW DIAGRAM OF THE LOUDNESS NORMALISATION WEB SERVICE .......................... 41
FIGURE 22: ADDED ATTRACTOR METADATA CONTAINING LOUDNESS GAIN VALUE ......................... 42
TABLE 2: EXAMPLE OF RANKING RESULTS FOR ONE SET OF SEQUENCES .......................................... 43
FIGURE 23: TEST SETUP .............................................................................................................................. 45
FIGURE 24: CALCULATED LOUDNESS DESCRIPTOR "INTEGRATED LOUDNESS" .................................... 46
FIGURE 25: RANKING OF LOUDNESS NORMALISATION ............................................................................. 47
TABLE 3: VIDEO CLIPS DESCRIPTION ....................................................................................................... 48
TABLE 4: AUDIO ADAPTATION TYPES ....................................................................................................... 49
FIGURE 26: COMPRESSOR: STATIC CHARACTERISTICS ............................................................................ 50
FIGURE 27: COMPRESSOR: DYNAMIC CHARACTERISTICS ........................................................................ 51
FIGURE 28: LOUDNESS RANGE OF ITEMS UNDER TEST (UNCOMPRESSED AND COMPRESSED c1_15 dB/c1_25 dB) .................................................................................................................. 51
FIGURE 29: AGE OF PARTICIPANTS ........................................................................................................... 53
FIGURE 30: LISTENING LEVEL ..................................................................................................................... 54
FIGURE 31: CATEGORY OF SPEAKER ............................................................................................................ 54
FIGURE 32: SIZE OF SPEAKER ...................................................................................................................... 55
FIGURE 33: DISTANCE TO SPEAKER (PC SPEAKER & STEREO SYSTEM) ................................................. 55
FIGURE 34: BACKGROUND NOISE ............................................................................................................. 56
FIGURE 35: EVALUATION OF LOUDNESS ADAPTATION ......................................................................... 56
FIGURE 36: EVALUATION OF LOUDNESS RANGE ADAPTATION ............................................................. 57

List of Tables

TABLE 1: LOUDNESS DESCRIPTORS ........................................................................................................ 23
TABLE 2: EXAMPLE OF RANKING RESULTS FOR ONE SET OF SEQUENCES .......................................... 43
TABLE 3: VIDEO CLIPS DESCRIPTION ....................................................................................................... 48
TABLE 4: AUDIO ADAPTATION TYPES ....................................................................................................... 49
# List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Achromatic component</td>
</tr>
<tr>
<td>ANTS</td>
<td>Automatic Newscast Transcription System</td>
</tr>
<tr>
<td>Cr1</td>
<td>Red and green antagonist components</td>
</tr>
<tr>
<td>Cr2</td>
<td>Blue and yellow antagonist components</td>
</tr>
<tr>
<td>CSF</td>
<td>Contrast Sensitivity Functions</td>
</tr>
<tr>
<td>Dialnorm</td>
<td>Audio meta-data parameter that controls playback gain within the Dolby Laboratories Dolby Digital (AC-3) audio compression system</td>
</tr>
<tr>
<td>DoG</td>
<td>Difference of Gaussian</td>
</tr>
<tr>
<td>IPTV</td>
<td>Internet Protocol TeleVision</td>
</tr>
<tr>
<td>ROI</td>
<td>Region of Interest</td>
</tr>
<tr>
<td>SOAP</td>
<td>Simple Object Access Protocol</td>
</tr>
<tr>
<td>SOI</td>
<td>Sequence Of Interest</td>
</tr>
<tr>
<td>SOI</td>
<td>Sequence Of non Interest</td>
</tr>
<tr>
<td>VAM</td>
<td>Visual Attention Model</td>
</tr>
<tr>
<td>WSDL</td>
<td>Web Services Description Language</td>
</tr>
</tbody>
</table>
1 Introduction

With the incredible growth of available Audio/Video contents, it is no longer possible to manually extract the interesting sequences of a video or the interesting parts of a picture.

Automatic tools for identification, segmentation and annotation [5][6][7][11] have to be studied and developed for video. For audio, automatic tools have been also developed to extract metadata based on loudness harmonisation of various audio programmes coming from different distribution media.

The work on Audio/Video metadata generation is split between:
- Metadata extraction on the video-part for the determination of the region of interest for intelligent cropping and scaling of video material, which is going to be used on terminals with various display sizes,
- Metadata extraction on the video-part for the determination of the best Ad insertion region which corresponds to the best moment in the video and the best corner of the screen,
- Automatic news segmentation and generation of metadata from newscast transcriptions,
- Metadata based loudness harmonisation of various audio programmes coming from different distribution media, such as terrestrial, satellite and cable broadcast, as well as from the Internet, such as Mediathek, YouTube, Audio/Video via Internet, IP-TV and others. Since NoTube wants to combine all these incoming streams or downloads in a single end-user terminal to enable a personalised Rich Media experience, it is very important that the loudness between all these audio streams is well balanced in order to avoid a continuous and annoying manual correction of the loudness level by the user.

2 Scope of This Deliverable

The goal of this deliverable is to describe the Audio/Video content analysis component with the generation of metadata.

The initial version of this document (delivered in M13) provided a general overview of audio/video content analysis components. The second version (delivered in M23) aimed at giving updates of previous contributions on the Ad insertion and on the loudness harmonisation.

The current version (due in M33) aims at giving the final updates on the NoTube Audio/Video content analysis components. More specifically, the following topics are updated and detailed:

- Integration of the video reframing based on Region of Interest (ROI) technology (section 4.3)
- Description and improvements realized on the Ad insertion algorithms (section 5)
- Description of loudness normalisation and User evaluation of loudness harmonisation (section 7.2.2.1, 7.2.2.2 and 7.4)
3 Audio/video content annotation

Several types of annotations of TV contents are available. In the context of NoTube project, we will focus on automatic Video and Audio annotation and metadata extraction. Tools that automatically generate metadata of AV content provide invaluable help on information retrieval, which is necessary due to the incredible growth of the available AV content.

3.1 Audio Annotation

For the first deliverable (M13), loudness annotation has been studied and implemented.

The loudness metadata which characterize the loudness level of each incoming stream or each download are required in order to normalise the loudness level before reproduction of various audio clips. These metadata have to be generated by loudness measurements. These metadata are essential for the normalisation of various audio programmes coming from different distribution media.

For the second version of the deliverable (M23), a loudness analysis component has been implemented as a web service. Furthermore, the harmonisation of the loudness by applying the metadata from this analysis (loudness descriptors) has been evaluated.

For this third version of the deliverable (M33), a loudness normalisation based on loudness metadata has been realised and implemented as a web service. In addition, loudness normalisation of multimedia content was evaluated in typical listening situations using a web application.

3.2 Video Annotation

Efficient and effective handling of video documents depends on the availability of metadata. Manual annotation is unfeasible for large video collections. The need to analyse the content has appeared to facilitate understanding and contribute to a better automatic video content indexing and retrieval.

We will present several methods aiming at automating this time and resource consuming process:

- The Regions of Interest (ROI) algorithm and its application in the Automatic Video Reframing
- The Program segmentation algorithm and the Subject classification

The detection of ROI is needed if cropping of the original video is required. To be able to have a cropping method that satisfies high quality demands, it is important, that the detection of the ROI is accurate in order to preserve the part of the video that holds the information that is important to the viewer. Automatic cropping is an important part of the NoTube system, as the viewing device of the user is unknown at production time. It can vary from large flat screen TVs to small handheld devices like mobile phones.

Program segmentation aims at temporally decomposing the video into coherent units. The program segmentation consists of identifying the parts, or sub-units, starting from the whole represented by an audiovisual content. News segmentation is the most common application of scene/story boundary detection. Automatic news segmentation is an important part of the NoTube system, because news item, which corresponds to user’s profile, can be provided to the end user. The resulting news items are the keys to provide the news to the end-user corresponding to his profile.

Subject classification (content metadata generation) is performed using an automatic speech recognition (speech-to-text algorithm) which is capable of transcribing both Italian and English audio contents accurately.
The detection of Sequence Of non Interest (SOI) is needed for the automatic Ad insertion algorithm. The aim of this algorithm is to insert advertising clips into a video sequence, at the best moment of the video, not only at the beginning or at the end of video sequence. In our application, the right moment is chosen as the combination of two criteria: The quietest sequence and the region of non-interest (with minimum ROI interference). The detection of the SOI requires the analysis of the whole video and the annotation of this video: SOI picture numbers, selected corner, position, Ad window size, saliency, scene-cuts.
4 Video reframing based on ROI technology

Due to the proliferation of new content distribution platforms, and especially in the case of mobile applications, new video viewing experiences on small screen devices are expected. It becomes mandatory to repurpose the video content, i.e. to adapt the image size to the screen size in a specific way, in order to enhance the viewing experience. Today this reframing is generally done manually, repurposing the video content is thus expensive and time consuming. An automated way, delivering the best viewing experience, would be a high economic differentiator. The specific problem of watching video on these new platforms is the size of the screen. As most of video contents are not produced to be viewed on a small-screen, the direct transfer of video contents would provide a video on which the main characters or other objects of interest may become indistinguishable from the rest of the image. For example, what is the interest of watching a soccer game on a mobile phone if the ball is not visible? The solution is to focus on the most visually interesting parts of the video. As simple as it appears, this solution brings a number of difficulties: the first concerns automatically detecting the regions of interest (RoI). The principle of the first studies [1], [2], [3], as it will be described in the next section, is based on the use of a visual attention model. This kind of model [4], [5], [6] is able to provide a map indicating the hot spots of a scene. The second issue concerns the quality of the reframing, in particular the temporal stability of the computed cropping window and the amount of zoom. For instance, how deep shall the algorithm zoom in? Although a low zoom factor does not improve the visual perception on small screen, a too high zoom in factor may be disturbing. The right zoom factor remains the real challenge for stabilizing the visual experience when reframing sequences.

4.1 Automatic Reframing

4.1.1 Overview

Many works applied on still picture have used saliency maps as a starting point for their reframing algorithm. In the same way, the proposed technology rests on both an efficient visual attention model and the cropping window extraction. Firstly, a visual attention model computes a saliency map per frame which indicates or classifies the frame regions according to their visual interest. Once these regions have been detected, a cropping window which encloses the most important part of frame is deduced from the saliency map. Both steps have a fundamental role in the final quality of reframed sequence (Figure 1). In the following section, the human visual model is briefly described. The main technical point deals with the cropping window extraction and its consistency over time.
Figure 1: General description of automatic reframing process

The main boxes (visual attention model and cropping window extraction) are presented in the following sections.

4.1.2 Visual attention model (VAM)

4.1.2.1 Existing VAM models

Human visual attention is likely the most important propriety of our visual system. As our visual environment contains much more information than we are able to perceive at once, it is necessary to select (or to detect) the most important parts of a scene. Then visual attention guides the movement of the eyes, allowing an accurate inspection by the fovea (retina area where the detailed vision is the most precise) of the chosen area.

Computational model of visual attention strives to detect the most salient parts of the video sequence. Most of the models provide a topographic saliency sequence which quantitatively predicts how conspicuous every location in the input sequence is. (Figure 2). These models, based on the use of the low level visual features, are commonly called bottom-up models. At the opposite top-down models take into account cognitive information as faces, text... or are dedicated to achieve a particular task.

Figure 2: Saliency maps

In Figure 2, Top row is a set of original frames. Bottom row is the corresponding saliency maps computed from our attention model. The whiter, the more attractive is the pixel.

4.1.2.2 Model chosen for reframing applications

The chosen model is described in [6], [7], which is a purely bottom-up model based on luminance, color and motion information. This model is based on the architecture of Koch and Ullman. The first step consists of the extraction of early visual features. The visual input is broken down into three separate feature channels (colour, intensity, and orientation). Each
channel is obtained from Gaussian pyramids. This allows the computation of different spatial scales by progressively applying a low-pass filter and subsampling the visual features. In order to take the organisation of the visual cells into account, a center-surround mechanism based on a Difference of Gaussian (DoG) is applied on each scale. The resulting maps are then linearly summed across feature channels to form the saliency map.

Figure 3: Flow chart of the proposed spatio-temporal model.
The model takes a video sequence as input and processes all the frames in three parallel channels using a range of spatial scales and orientation values. It yields a saliency map indicating the most salient region per image.

4.1.2.3 Computation of the spatial salience
First, the RGB picture is projected into the Krauskopf’s colour space (A, Cr1, Cr2) [8] simulating the three different pathways used by the brain to encode the visual information. The first pathway conveys the achromatic component (A), the second the red and green antagonist component (Cr1) and the third the blue and yellow antagonist component (Cr2). In order to express all data in the same unit (in term of visibility), three contrast sensitivity functions are used, one per component. If components (A, Cr1, Cr2) can be described in terms of their sinusoidal Fourier components, then the visibility of each spatial frequency can be measured by applying a contrast sensitivity function. Each spatial frequency is then compared to a threshold CT0. If the amplitude is above this threshold, the frequency is perceptible. This threshold is called the visibility threshold and its inverse defines the values of the contrast sensitivity functions (CSF) at this spatial frequency. According to electrophysiological measurements revealed that visual cells are tuned to certain types of visual information such as frequency, colour and orientation. A hierarchical decomposition is then conducted splitting the 2D spatial frequency domain both in spatial radial frequency and in orientation. This decomposition is applied to each of the three perceptual components. Each resulting sub band, or channel, may be regarded as the neural image corresponding to a particular population of cortical cells. These cells are tuned to a range of spatial frequencies and to a particular orientation. Finally, the masking effect alters the differential visibility threshold of each sub band. Three types of masking are considered in the proposed model: intra-channel intra-component masking, inter-channel intra-component masking and inter-component masking. To suppress redundant data, a centre-surround filter is applied. After applying the centre-surround filters, three saliency maps are derived: first, a two-dimensional achromatic saliency map, called SA, was computed from the direct sum of the outputs of the achromatic channels belonging to the crown III. Second, two chromatic saliency maps were computed by the direct summation of the outputs of the chromatic channels belonging to the crown II.

4.1.2.4 Computation of the temporal salience
The basic aim of the temporal saliency map computation [5] rests on the relative motion occurring in the retina. The relative motion is the difference between the local and the dominant motion.

The local motion at each point of an image (or the motion vector) is given by using a hierarchical block matching. It is computed through a series of levels (different resolution), each providing input for the next. In addition, on each level, the block matching is done for a certain neighbourhood size, that increases with the hierarchy level. In a way, these two points remind the properties of the motion processing in the monkey cortex.

As soon as the camera follows something in the scene, it is necessary to estimate the global transformation that two successive images undergo. This global transformation, or the dominant motion, is estimated from the previous estimated local motion. The dominant motion is represented by a 2D parametric model.
4.1.2.5 Computation of the rarity map

The eye is not attracted by particular features in an image but it is attracted by the features which are in minority in an image as it can be noticed by observing the following facts:

- Our vision can be attracted by homogeneous areas into a heterogeneous scene, but also by heterogeneous areas into a homogenous scene.
- Bright areas into a dark scene will be attractive for our eyes like dark areas into a bright scene.

Heterogeneous or homogeneous, dark or bright, symmetric or asymmetric, moving or static objects can all attract our visual attention. We also can note that the pairs of features we mentioned are opposite features describing the order and the disorder at several scales, in space and time.

The most basic operation to modelize rarity is to count similar areas in the image. The histogram is an adequate statistical tool which counts equivalent pixels.

Within the context of information theory, this approach based on the histogram is close to the so-called self-information. Let us note \( m \) a message containing an amount of information. A message self-information \( I(m) \) is defined as:

\[
I(m) = - \log p(m)
\]

where \( p(m) = \Pr(M=m) \) is the probability that a message \( m \) is chosen from all possible choices in the message space \( M \).

We obtain an attention map by replacing each message \( m \) by its corresponding self-information \( I(m) \). The self-information is also known to describe the amount of surprise of a message inside its message space: rare messages are surprising, hence they attract our attention.

4.2 Cropping window extraction

4.2.1 From the saliency map to the first cropping box

The goal of this step is to define a cropping window that encloses the most conspicuous parts of the scene. Based on the results coming from the attention model, a Winner-Take-All algorithm is applied. This algorithm allows the detection of the first \( N \) most important locations (having the highest saliency values). When the \( k \)th maximum location is selected and memorized, this location as well as its neighbourhood is inhibited. The size of the neighbourhood depends on the viewing distance. The further the viewing distance, the higher the size of the neighbourhood is. Due to the inhibition process, a new salience peak will dominate and will be selected at the next iteration. The iteration number parameter is a sensitive problem since it influences the size of the bounding box. If the iteration number is small (or in other words, the number of location to be selected is small), it is likely that the size of the bounding box will be small. On the contrary, a high number of iteration will ensure a bigger bounding box. Although this approach seems to be convenient for our needs, one aspect is clearly neglected: the distribution of salience in the map. To illustrate this point, two different pictures (the original picture depicted in Figure 4 and a landscape) can be considered. On the former, it is obvious that some parts of the picture pop-out due to their contrast to the background. In this case, it is reasonable to assume that visual attention is predominantly controlled by bottom-up stimulus properties. Therefore, the distribution of the salience (or the average observer variability) will likely be sparse. In other words, the separation between the salience of the peaks and the average background level will be important (see the second picture on Figure 4). Concerning a landscape, in which nothing clearly pops-out, the distribution of salience is more uniformly distributed.

In this case, selecting the \( N \) first locations can yield an erroneous result. To solve this problem, the idea is to stop the iteration loop when a significant amount of salience (enclosed by the bounding box) has been taken into account during the selection process.
4.2.2 Temporal consistency

The temporal filtering is likely one important issue, because it has to deal with two aspects:

- Objects (still or not) or texture may appear along a scene and could lead to a significant change in the distribution of salience. Significant changes of the salience distribution could alter the position and/or the size of the cropping window over time.

- Another fundamental idea is to be closed to camera motion. As soon as reframing application behaves as a camera, potential watcher will not make any distinction between a reframed sequence and the original one.

The smoother the temporal cropping process is, the better the visual impact for any users will be. The temporal stabilisation acts both on the position and size of the bounding box. Temporal consistency is composed by two sequential steps: a Kalman filter is first applied in order to better predict the current samples. Then, a temporal filtering allows avoiding unlikely samples. Regarding the position of the bounding box centre and its size, two discrete Kalman filters are used to ensure a good temporal consistency.

4.2.3 Aspect ratio

The aspect ratio gives the relationship that exits between the width and the height of the original sequence. The first extracted window cannot correspond to the final aspect ratio: a priori, there is no relation between the window extracted from saliency map and the user settings. In some way, one side has to be extended. The extension is either on width or on height to reach the targeted aspect ratio.

4.3 Integration within the WP7a use case

The video reframing functionality based on ROI technology was integrated in the Personalized Semantic News use case. The description of this integration is described in details in document “D7a.4 Personalized News Platform prototype, v3”. The WP7a 3rd prototype allows the final user to access the Personalized Newscast Program using the Smartphone as a secondary screen. In particular, the reframed video of News items is used in the 3rd prototype of WP7a for playing some particular news on the Smartphone. When the final user reaches particularly interesting news, he can choose between playing them on his TV screen or watching the news directly on the Smartphone screen. In case of Smartphone screen, the video content is reframed to enhance the viewing experience.
Figure 5: News Item details with play out feature on mobile device

Figure 6: Reframed video played on a mobile device
5 Automatic Ad insertion

5.1 Overview
The aim of this automatic Ad insertion module is the insertion of advertising clips into a video sequence, the objective being to insert it at the best moment of the video, not only at the beginning or at the end of video sequence. In our application, the right moment is chosen as the combination of two criteria: The quietest sequence and the region of non-interest (with minimum ROI interference).

The implementation of the Ad insertion algorithms on Thomson Video Networks equipment is based on the succession of 2 passes. The first pass processes the movie in order to extract the metadata describing the n “best sequences” available to insert the Ad and writes these metadata in an XML file. The second pass inserts the Ad in the video thanks to the metadata generated by the first pass.

5.2 Description of the algorithms

5.2.1 General description of the 1st pass (metadata generation)
The 1st pass is done in three steps. The first step consists of the analysis of each image of a video. When the last image is processed, the second step is launched. It consists in using the data produced by the 1st step to find the average saliency value for each corner of each image of the video. The third step consists of the selection of the n “best sequences” available to insert the ad and to produce the metadata file.

5.2.1.1 Description of the 1st step: Analysis of each image of a video
This analysis is based on several algorithms: Saliency map extraction for the four corners, motion vectors analysis and image differences computation. The result of these gives, for each image, a set of data.

Note: The black stripes (letter box or pillar box) are not processed during the analysis.

• Calculation of the saliency map for each corner
  The saliency map is extracted for each corner and for each image of the video. This saliency map is obtained by using the Visual Attention Model described in 4.1.2 and gives the zones of interests in each corner.

• Calculation of motion vectors
  The process includes an estimation of the “dominant motion” in a sequence of images, which corresponds to the motion of the maximum of pixels.

• Calculation of image differences
  This calculation is done as follows:
  1. For each image, get the difference between each pixel in the current image and the same pixel in the following image.
  2. For each image, calculate the average of the differences found in 1, for all the pixels of the image.

5.2.1.2 Description of the 2nd step: Computation of the average saliency value for each corner of each image
The video analysis done during the first step gives a set of data which describes each image. Based on this data, the second step determines the average saliency value for each corner of each image of the video.
For each corner of each image, a mean saliency value is computed based on the analysis of the saliency map produced during the previous step. This value is then corrected by taking into account two algorithms: motion vectors analysis and image differences computation:

- If the motion vector is high, this image is considered as less pertinent to insert an Ad because it contains motion, which means that the user may be interested by the sequence containing this image. The algorithm will then apply a correction coefficient that increases the saliency.
- If the correlation between images is high, this image is considered as more pertinent to insert an Ad, because it shows that not so much action is happening. The algorithm will then apply a correction coefficient that decreases the saliency value.

At the end of this 2nd step, the average saliency value is available for each corner of each image.

5.2.1.3 Description of the 3rd step : Selection of several sequences allowing the insertion of the Ad

The maximum number of sequences (N) to be inserted is given by the user parameters feeding the analysis algorithm.

The third step then consists in extracting those N sequences of images from the video:
- for which duration exceeds the duration of the advertising and
- which show, for one corner, the lowest average saliency values throughout the sequence.

The metadata describing these sequences is written in an XML file. The first sequence in the file corresponds to the best sequence, i.e. the sequence localised in one corner which shows the lowest average saliency values throughout the sequence.

5.2.2 General description of the 2nd pass (Ad insertion)

The metadata produced by the 1st pass is then used in the following way to actually insert the Ad at the right time and in the right corner:

- Movie processing:
  1. The Transport Stream input file is read and the Packet Elementary Streams are extracted.
  2. The video is decoded.
- Ad processing:
  1. The Transport Stream input file is read and the Packet Elementary Streams are extracted,
  2. The video is decoded,
  3. The Ad is resized to match the size of the window where it will be displayed.
- Ad insertion: The metadata is used to insert the resized Ad in the decoded video at the right time and in the right corner.
- Video re-encoding and file generation.

5.3 Automatic Ad insertion first evaluation results

An evaluation was made on 5 videos with 6 testers (colleagues & relatives). For each video, 5 sequences were produced with an automatic ranking.

Each tester was asked:
- For each sequence, to grade:
  - The time positioning (0: Bad, 1: Acceptable, 2: Good)
  - The space positioning (0: Bad, 1: Acceptable, 2: Good)
- To rank the 5 sequences (1 to 5, 1 being the best sequence and 5 being the worst)
The evaluation results (shown in Figure 7 and Figure 8) highlight the globally good results of the Ad insertion algorithm. Figure 7 shows that the best and the worst sequence given by the algorithm are also those which are given by the testers. However, in some movies (like “300” in Figure 8), we can see that the algorithm doesn’t give the best results.

![Figure 7: Technology evaluation - Results by tester](image)

![Figure 8: Technology evaluation - Results by film](image)

5.4 Improvement of the automatic Ad insertion algorithm

After the first Ad insertion technology evaluation, several solutions were studied to improve the efficiency of the Ad insertion algorithm:

- Use the audio level (See 7.2) to improve the choice of the best Sol,
- Use scene cuts analysis to find the quietest sequence,
- Use global saliency maps analysis (not only corners) to find the best sequence.

These three new features, detailed hereafter, are being implemented and will be tested on the sequences used for the 1st evaluation.

A second evaluation is scheduled in November 2011 with a larger panel (NoTube partners) in order to get a better confidence in the results.
The modification of the automatic ranking brought by each of these new features will be compared to the ranking done by the testers and the final automatic Ad insertion algorithm will use an optimal weighting of the sub-ranking produced by these three features.

The first evaluation also showed a lack of visibility of the commercials inserted in a film, especially when the luminance of the Ad is close to the luminance of the corner where it is inserted. In addition to the developments mentioned above, the visibility of the commercial will also be improved.

5.4.1 Ranking improvement thanks to Audio level analysis

In order to find the best moment to insert the ad, IRT’s work on audio level has been taken into account in order to improve the current algorithm. This work is based on the calculation on loudness descriptors.

5.4.1.1 Specification of the loudness descriptors to be analyzed

The specification of the corresponding loudness descriptors are fully detailed in EBU-R128. Two parameters (Programme Loudness and Loudness Range) (see Table 1) are calculated on the whole video with a short term window (3 seconds).

<table>
<thead>
<tr>
<th>Programme Loudness</th>
<th>The integrated loudness over the duration of a programme - Programme Loudness Level is the value (in LUFS) of Programme Loudness.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loudness Range (LRA)</td>
<td>This describes the distribution of loudness within a programme.</td>
</tr>
</tbody>
</table>

Table 1: Loudness descriptors

5.4.1.2 Basis of the algorithm improvement

The improved algorithm will work on the two descriptors: Programme Loudness and Loudness range. Among the sequences previously found by the first version of Ad insertion algorithm, this algorithm will penalize those with a high audio level.

5.4.2 Ranking improvement thanks to scene cuts analysis

In a film, action scenes have generally a lot of scene cuts and, naturally, are those parts that the viewer would not like “disturbed” by a commercial. For this reason, a scene cuts detection algorithm will be used to count the number of cuts during the sequences already identified. The higher the number of scene cuts, the lower the sequence ranking.

5.4.3 Ranking improvement thanks to global saliency maps analysis

The first version of the algorithm was only taking into account the saliency of the four corners of each picture of the video. However, in some cases, one corner can be “quiet” when the rest of the picture shows a lot of action. Processing the saliency of the full image should improve the ranking in these cases. For the sequences already found by the first version of Ad insertion algorithm, this new algorithm will process the global saliency map of each picture, penalizing the sequences having the highest saliency value.
5.4.4 Improvement of the Ad visibility

The Ad visibility will also be improved by the insertion of a coloured frame around the commercial so that it can be made more noticeable. The frame colour can be chosen by the user before the Ad insertion process.

5.5 Integration within the overall NoTube Architecture

The Ad insertion functionality based on ROI technology was integrated in the Personalized EPG use case. The Ad insertion WebService is used to insert advertising clips into a video sequence at the right moment and right part in the video.

In use case 7b, an end user accepts to watch a movie with an inserted Ad. The iFanzy demonstrator selects an Ad which corresponds to the end user’s profile. The demonstrator calls the Ad insertion WebService which analyses the best place and the best time of insertion and inserts the advertising in the initial video. The iFanzy demonstrator can then deliver this resulting video to the end user.

NoTube prototype description

![Diagram](image)

Figure 9: NoTube prototype description

Service prototype implementation

Web Services Description Language (WSDL) is used to describe Web services. It provides a model and an XML format. It represents a contract that governs the interaction between the Web service and a client application.
WSDL 1.1 shall be used in combination with SOAP 1.1 or 1.2 and XML schema to provide web services over the network. A client program connecting to a web service can read the WSDL to determine what functions are available on the server. Any special data types used are embedded in the WSDL file in the form of XML Schema. The client shall use SOAP over HTTPS to call one of the functions listed in the WSDL.

A workflow is a model that describes a processing like Video analysis (see Figure 12) and Ad insertion (see Figure 13). A Job is an instantiation of a workflow that corresponds to a user request. (For example, insert the “car” Ad in the “Star Wars” movie)

For example, to configure the Ad insertion on the platform with web services, the following steps have to be done:

- Define the Ad insertion workflow which is based on the different basic workflows
- Instantiate this workflow with a job creation

These two operations can be done with web services. For an easier management of the platform, the Ad insertion workflow is already created on the platform and the user needs only to create a job on this workflow.

The following web services can be called after the workflow configuration of the platform.

- **Workflow get list** (see Figure 10)

This web service operation is used to know the different configured workflows and obtain the workflow id and workflow name.

```plaintext
WorkflowGetList

Source code

Style: document

Operation type: Request-response. The endpoint receives a message, and sends a correlated message.

SOAP action: WorkflowGetList

Input: WorkflowGetListReq [soap body, use = literal] Source code

parameter types WorkflowGetListReq

Output: WorkflowGetListRsp [soap body, use = literal] Source code

parameter types WorkflowGetListRsp
```

**Figure 10 : Workflow Get List**

- **Job create** (see Figure 11)

This web service operation is used to create a job corresponding to a workflow (for example, to create the job "insertion of ad" and start this job)
Service interface
As mentioned in 5.1, the Automatic Ad insertion functionality is implemented in 2 passes, leading to a service interface divided into 2 workflows which are described by figures below.

Figure 11: Workflow Job Create

Figure 12: Video analysis workflow

Figure 13: Ad insertion workflow
6 Automatic program segmentation

Automatic program segmentation is one of the most challenging and complex subjects of research. Although being able to produce a correct segmentation is a key factor for improving the accessibility and precision of search and retrieval, we cannot count on an established approach at solving the problem in general. The common base of the approaches for news is constituted by the use of a combination of visual, audio and speech features.

As stated in D7a.2 Personalized News Platform prototype v1 deliverable, the RAI content and metadata, provided as NoTube system input, consists in newscast transcriptions (i.e. automatic speech recognition outputs) which are made available to the system by mean of xml files adherent to the PrestoSpace [9] metadata schema [10]. These files, coming out from the ANTS (Automatic Newscast Transcription System) environment, contains all the known information for the single editorial object each of them represents: identification and basic description, publication date time and channel, as well as the text derived from the speech and other automatically extracted enrichment [11] [12].

6.1 ANTS for raw data

The “Personalized semantic news” use case foresees automatic metadata and content enrichment on single news items provided as input by an external feed laying at the beginning of the use case chain.

Each news item should be enriched by linking to a set of related contents (i.e. having a specific relationship based on the main concept of the news item itself). Other activities that could be done include named entities recognition and categorisation, content summarisation etc.

ANTS is an integrated system for the automatic acquisition and automatic metadata extraction of radio and television programmes, it is made up by several components, integrated within a unified architecture.

ANTS can
• Identify story boundaries¹,
• Extract texts from spoken content,
• Classify stories by subject,
• Link external relevant information coming from the web, all within a flexible and extensible architecture.

6.2 ANTS architecture

The system has been designed to be highly distributed and scalable. Its main components are a centralised workflow engine and a collection of generic clients, each configured to carry out a specific task of the overall process, such as the speech-to-text activity or the news story segmentation.

The clients communicate with the workflow engine via web protocols, get the waiting jobs, operate them and notify their success/failure. Such an approach, described in Figure 14, allows the good scaling of the system as required for the needed throughput.

¹ Good results are achieved with Italian newscast programmes
All the metadata produced are collected within a centralised repository until the final stage where a delivery package is produced including audiovisual material and the metadata expressed as well known xml formats.

**Figure 14**: The ANTS architecture

### 6.3 Segmentation of news stories

One of the most important added value given by the services we foresee in the use case is the possibility given to the user, to get in his home environment programme segments (instead of entire programmes) of his own interest, giving the possibility to enjoy all the tracks together or separately (audio, video, text, metadata and links to related external resources).

The “Personalized semantic news” use case makes use of some of the ANTS outputs in order to achieve the above mentioned purpose. In particular ANTS will furnish the newscast programmes segmentation (i.e. newsitems boundaries) and the newsitem classification.

To reach this goal both aural and visual aspects are taken into consideration, with the help of a three-layered heuristic framework, deduced by the observation of editorial styles of a statistically significant set of Italian newscast programmes, spanning approximately 40 hours (~80 programmes).

The first heuristics, widely adopted in literature – e.g. by [13] – is that being able to detect boundaries of shots containing the anchorman is equivalent to detecting news story boundaries.

To detect anchorman shots we use another heuristics, namely that the most frequent speaker is the anchorman and that he/she speaks for periods of time spreading right along
the programme timeline. This allows us to select the most probable candidate speaker among the ones identified by a speaker-clustering process.

But this is not enough, in fact with this approach it is not still possible to recognize situations in which the anchorman introduces several brief stories in sequence without external contributions (e.g. reportages).

To overcome this limitation we use the third heuristic, consisting of the knowledge that, in the great majority of observed cases, the introduction of a new brief story is accompanied by a camera shot change (e.g. from a close-up shot to a wider one).

Therefore a videoshot-clustering is also performed in order to optimize the accuracy in selecting the camera shot changes, so that it is possible to detect and classify shot clusters as pertaining to studio shots containing the anchorman, following the same frequency/extension heuristic used for detecting the candidate speaker.

This double clustering process (both on audio and on video) enables a very simple and effective segmentation algorithm based on mutual coverage percentage of clusters.

6.4 Subject classification and enrichments

In ANTS, automatic speech recognition is performed using a speech-to-text engine based on [14], which is capable of transcribing with a quite good accuracy both Italian and English. Subject classification of segmented stories is done using open source software performing naive Bayesian classification whose model has been trained on a corpus made up of items of extracted text and annotated with standard subject taxonomy of 28 classes. The ANTS environment provides also a set of information derived from the transcribed text as contained named entities, related external links, related DBPedia URIs and related named concepts operating in the Italian language.

6.5 Integration within the overall NoTube Architecture

In the D7a.2 Personalized News Platform prototype v1 deliverable the ANTS environment integration issues related to the M13 prototype are presented. The use of ANTS in the semantic enrichment phase in the M13 prototype is due to its real availability and to its Italian language oriented capabilities. In next prototypes the semantic enrichment services provided by WP4 will be integrated.
7 Loudness harmonisation for the audio part of various A/V-streams entering the NoTube terminal

7.1 Loudness of audio services: the actual status

Audio levels in Broadcasting have become increasingly diverse and different over the last decades. Despite clear guidelines and recommended practices, the general use of peak measurement in audio metering and the development of more and more sophisticated level processors have led to over-compression of audio signals with the questionable aim of being louder than the competitor. Sometimes this situation has actually been specified as “Loudness War” [16].

The issue revolves around program and commercial loudness and a TV audio function called dialog normalisation or “Dialnorm”. This is a parameter of digital audio metadata that accompanies the transmitted sound and sets the volume in all TV home receivers. Prerequisite for programs and commercials to transition smoothly, all content must be encoded with the proper Dialnorm value.

Viewers continue to complain that some TV commercials are louder than others and in most cases are presented at a much higher volume than the program dialogue they precede or follow. As per the Advanced Television System Committee’s (ATSC) TV standard adopted by the FCC and made law by reference in 1996, all digital TV transmissions in the United States must broadcast properly matched loudness and Dialnorm.

Unfortunately, that requirement was never clearly explained, and viewers continue to complain about unbalanced loudness. Some have threatened to get their programming elsewhere (such as the Internet or DVD), which is a threat for stations in today’s highly competitive environment.

Presently, the audio playback of various streams from different sources causes quite often an annoying experience for the listeners. The reason behind that is that the loudness level of these different streams can be completely different and do not match a comfortable listening experience on the user terminal.

IRT has made measurements in the past, which show that even similar audio streams, which are going to be broadcasted on the broadcast network and simultaneously on the Web, i.e. their Internet-Radio service, might have very different loudness levels.

Even more different levels can be expected, if various streams from completely different sources will be collected to form a Rich Media Service.

In this context a special aspect of loudness harmonisation has to be addressed considering different reproduction environments. Usually broadcast programs, even if thinking about balanced programs, are not adjusted to different listening environments. The reason is that it is impossible to adapt the loudness range, the ratio between loud and quiet passages, to the wide variation of listening environments, such as home cinema, car environment or walkman. This necessary adaptation of the loudness range to different listening environments can only be achieved by audio processing on the receiving end, among others controlled by the device that plays the AV data which uses the corresponding metadata to harmonize the loudness level.

Considering audio levelling it has to be pointed out that despite the existing guidelines and recommendations with respect to audio metering, which actually means adaptation of peak levels, loudness harmonisation can in principle not be achieved. For this purpose a standardized loudness meter is indispensable. Actually an international loudness algorithm standard ITU-R BS.1770 [18] has been established in 2006. This international standard is the requirement for loudness harmonisation issues. On the other hand such standard could completely overthrow the still valid level metering guidelines.
In 2008 the European Broadcasting Union (EBU) has established a project group P-LOUD with more than 100 members, including the IRT working on refinements of the ITU-R standard, such as definition of target loudness level, short time/long time loudness or loudness range, and setting up EBU guidelines on loudness measurements.

Considering Europe now, two years after inception of P-LOUD, the new leveling recommendations based on ITU-R BS.1770 have been established. They are replacing the existing recommendations. The work of P-LOUD has almost been finished and accompanied by publication of the following bundle of recommendations [19]:

- EBU Technical Recommendation R128: Loudness normalisation and permitted maximum level of audio signals
- EBU Technical Document 3341: Loudness Metering. ‘EBU Mode’ metering to supplement Loudness normalisation according to EBU Technical Document R128
- EBU Technical Document 3342: Loudness Range. A descriptor to supplement Loudness normalisation according to EBU Technical Recommendation R128
- EBU Technical Document 3343: Practical guidelines in accordance with EBU Technical Recommendation R128
- Technical Recommendation R128

The arbitrative impact of these documents is up to the fact that concrete specifications are defined with respect to measurement and compliance of normalized loudness in broadcasting. This refers to target loudness plus measuring instruments as well as corresponding recommendations among others addressing production, archiving and programme exchange. The essential difference to the valid European leveling recommendations [20, 21] is that audio leveling does no longer take place with respect to peak levels (in detail to the - sometimes single - maximum quasi peak within an audio piece – measured by a level meter with 10 ms integration time QPPM (Quasi Peak Programme Meter) [22], but to the recommended target loudness. If future audio productions consequently meet these new audio leveling recommendations, this means the audio material in archives and on servers will be loudness harmonized and does not have to be adapted before broadcasting in order to avoid loudness jumps.

7.1.1 Specification of ITU-R BS.1770

In the following the actual loudness algorithm is described, which is the basis of the specific loudness parameters used in the loudness module. Figure 15 shows a block diagram of the various components of the algorithm. Labels are provided at different points along the signal flow path to aid in the description of the algorithm. The block diagram shows inputs for five main channels (left, centre, right, left surround and right surround); this allows monitoring of programmes containing from one to five channels. For a programme that has less than five channels some inputs would not be used. The low frequency effects (LFE) channel is not included in the measurement. Specifications of the objective multichannel loudness measurement algorithm are given in Annex 10f ITU-R BS.1770.
The first stage of the algorithm applies a pre-filtering of the signal prior to the $\text{Leq}(\text{RLB})$ measure as shown in Figure 16. The pre-filtering accounts for the acoustic effects of the human head, where the head is modelled as a rigid sphere.

Figure 15: Block diagram of multichannel loudness algorithm

Figure 16: Response of the pre-filter used to account for the acoustic effects of the head
The second stage of the algorithm applies the RLB weighting curve, which consists of a simple high-pass filter as shown in Figure 17.

![RLB weighting curve](image)

**Figure 17 : RLB weighting curve**

The frequency weighting in this measure, which is the concatenation of the pre-filter and the RLB weighting, is designated K weighting. The numerical result for the value of loudness should be followed by the designation LKFS. This designation signifies: Loudness, K weighted, relative to nominal full scale. The LKFS unit is equivalent to a decibel in that an increase in the level of a signal by 1 dB will cause the loudness reading to increase by 1 LKFS.

The algorithm can be used to accurately measure the loudness of mono, stereo and multichannel signals. A key benefit of the proposed algorithm is its simplicity, allowing it to be implemented at very low cost.

### 7.1.2 Specification of EBU-R128 and related documents

In the following the loudness definitions referring to EBU-R128 and related documents are listed. These definitions are the basis of the loudness analyser tool designed by IRT which will analyse the A/V-streams entering the NoTube platform. In principle all definitions are related to ITU-R BS.1770, i.e. the corresponding loudness algorithm. The increments compared to ITU are indicated as “EBU Mode”.

**Units**
Relative measurement: LU (Loudness Unit)
Absolute measurement: LUFS (Loudness Unit/Full Scale) (equates LKFS in ITU-R BS.1770)
Integration times
Momentary Loudness “M”: Integration time 0.4 s
Short-term Loudness “S”: Integration time 3 s
Integrated Loudness “I”: Average over extract between “start” and “stop”, usually over the complete audio piece or rather “audio file”

Measurement gate
Threshold for calculation of “absolutely gated” loudness level: -70 LUFS
Threshold for the calculation of “relatively gated” loudness level: 8 LU below “absolutely gated” loudness level

Target loudness
Normalisation to -23 LUFS including a relative gate below “absolutely gated” programme loudness

Loudness Range LRA
LRA refers to “Short-term Loudness”. LRA is defined referring to loudness level statistics. LRA is specified as the range between the 10th and 95th percentile.

Maximum True Peak
The permitted maximum level of audio signals shall be -1dB True Peak (dBTP) measured with a meter compliant with ITU-R BS.1770

7.2 Loudness Module in NoTube

In principle the designed loudness module in NoTube contains two stages, the “analysis stage” and the “synthesis stage”. In the first phase of the NoTube project, the focus was to only focus on the “analysis”.

7.2.1 Analysis stage

7.2.1.1 Specification of loudness descriptors to be analyzed

The specification of the corresponding loudness descriptors are fully related to ITU-R BS.1770 and EBU-R128. In detail the following loudness descriptors are considered to be relevant fulfilling loudness harmonisation in the NoTube audio module:

Integration times
Short-term Loudness “S”: Integration time 3 s
Integrated Loudness “I”: Average over extract between “start” and “stop”, usually over the complete audio piece or rather “audio file”

Measurement gate
Threshold for calculation of “absolutely gated” loudness level: -70 LUFS
Threshold for the calculation of “relatively gated” loudness level: 8 LU below “absolutely gated” loudness level

Target loudness
Normalisation to -23 LUFS

Loudness Range LRA
Calculation of “long-term loudness” level statistics with respect to percentiles
7.2.1.2 Description of analysis algorithm

This section describes the design of the software component. For the concrete implementation details, the source code and its formal documentation should be consulted.

Design criteria

The software source code was written with the objective that it should not only work, but ensure clarity & maintainability. Even if it is only a component used inside the loudness analysis web service for now, we consider it important to act as if the component were to be released as a toolkit to other developers. This facilitates re-use in future research phases. These goals imply a clean object-oriented design with strict encapsulation, self-explanatory identifiers, as well as structural preparations for code re-use and possible optimisation requirements.

The loudness measurement is implemented in Java language. Although performance optimisations are limited in Java, it is favoured over a native implementation, e.g in C, because it enables easy integration and portability is maintained. If necessary, optimisations could be added at a later stage, but are specific to all platforms the system is intended to be run on.

Software documentation is done in the standard "JavaDoc" form.

Program structure

We chose a layered design, consisting of a class "LoudnessAnalyzer" representing the business logic, i.e. the measurement purpose itself, and a web service wrapper class for adaption. This makes it possible to re-use the measurement class in different applications.

For example, "LoudnessAnalyzer" can be used manually, even in a non-networked context. For this purpose, LoudnessAnalyzer is equipped with a "main" method, which accepts audio file names in URL form as input from the command line, performs the analysis on these files and prints the results.

This command-line access enables us to perform parts of our auditory research on loudness normalisation without the need for separate software.

Web service layer and business logic layer are separated by an intermediate API that we defined.

The program is capable of analysing multi-channel audio files. In order to do so a binding to the library "gervill.jar" is required. The functionality of the Gervill library is part of OpenJDK, so that the library is not needed if the platform is based upon OpenJDK.

Web service wrapper layer

A wrapper class "LoudnessAnalysisWebService" only needs to read its "fileLocator" URL parameter and create a file input stream from it, construct a new instance of LoudnessAnalyzer with this input stream, and finally set the response data fields with the values retrieved from the LoudnessAnalyzer object. The LoudnessAnalyzer object is discarded afterwards. The "fileLocator" URL must point to a valid audio file in WAV format.

Business logic layer

In the LoudnessAnalyzer constructor, the audio file is referenced by using the parameter "inputStream", and all object attributes are set.

There is one constructor with the full loudness measurement parameters set and a second that sets those parameters to their default values.

The measurement parameters are:

2 http://java.net/projects/gervill/pages/Home
3 http://openjdk.java.net/
float targetLoudnessFSIntegrated, 
float peakSampleHeadroomFS, 
float silenceThresholdLoudnessFS, 
float loudnessGatingOffset, 
float loudnessIntegrationTime, 
boolean useTruePeak, 
boolean includeLFE.

Their default values are:

targetLoudnessFSIntegrated = –23.0 LUFS, 
peakSampleHeadroomFS = –1.0 dBFS, 
silenceThresholdLoudnessFS = –70.0 LUFS, 
loudnessGatingOffset = –8.0 LU, 
loudnessIntegrationTime = 0.4 s, 
useTruePeak = false, 
includeLFE = false.

All but the first measurement parameter, "targetLoudnessFSIntegrated", are considered "expert-level". Their purpose is to verify the algorithm in acoustics research and do not need to be exposed to the web service interface. The parameter "targetLoudnessFSIntegrated" designates a target value in LUFS units that the audio file is intended to be normalised to. Its default value, –23.0, conforms to EBU Tech. R128.

In the constructor, the loudness measurement process is immediately executed and its results are stored. This makes sense, because there would be nothing meaningful that could be done with a measurement object that is configured completely but not yet executed. After construction, the results can be retrieved from the object. The LoudnessAnalyzer object cannot be re-used.

The only public methods of LoudnessAnalyzer are the "getter" methods to retrieve the calculation results. These are:

float safeUncompressedGainForTargetLoudness, 
float uncompressedGainForTargetLoudness, 
float loudnessFSIntegrated, 
float peakLoudnessFSMomentary, 
float peakSampleLevelFS, 
float truePeakLevelFS, 
float loudnessFSIntegratedUngated, 
Vigintiles statLoudnessFSMomentary.

Similar to the parameters, all but the first result, "safeUncompressedGainForTargetLoudness", are considered "expert-level". Thus, with the method "getSafeUncompressedGainForTargetLoudness", a client can retrieve a measurement result value that specifies with how many decibel units the audio level should be increased to normalise the loudness to the desired level. However, the target loudness may not always be reached because the gain value guarantees that clipping of signal peaks is excluded.

Alternatively, if the loudness normaliser includes a level limiter, a client can retrieve a gain value that does not include the clipping safety measure by using the method "getUncompressedGainForTargetLoudness".
The audio data are read sample-by-sample (for all audio channels) from the input stream and are processed in this form by a sequence of operations whose structure follows the loudness measurement definition in ITU-R BS.1770-1.

This sequence is called once for each audio sample. The algorithm has been carefully designed so it only needs a single pass to process the audio samples and does not need to read the stream a second time.

Parts of the algorithm where the same operations are performed with different stored variables have been realised by encapsulating the variables in auxiliary classes, implemented as a private inner classes. Additionally, some more algorithmic elements were programmed as classes, in order to be usable as modules, namely the auxiliary classes "Integrator" and "Statistics".

All variables that need to be allocated once per audio channel, thus the state variables of the pre-filter and the RLB filter, are encapsulated in an auxiliary class "Channel". For example, for a stereo audio file, two instances of "Channel" are created. Any operations on these data are realised inside the class, so that the variables can be kept hidden. The only client access needed is the method "filterAndCalcPower", which calculates the power (squared amplitude) of the filtered audio. The power weighting factor for the appropriate channel (a value that is different for rear and front channels) has been made an immutable attribute of the "Channel" object and is thus set in the constructor.

The auxiliary class "Integrator" represents a sliding-window short-time averager. This means, it calculates a value for each sample, not just block values, and is in accordance with EBU Tech. R128, which only states a minimum block overlap, while we chose the maximum possible. "Integrator" is configured with an integration time constant at construction.

Calculation of the accumulated power of channels and the peak absolute value are performed procedurally, thus not in separate objects. Same goes for the result calculation, which is the following: After the execution of the per-sample sequence of operations, the statistics results are evaluated to calculate the gated loudness value as described. Then, the gain is determined from the difference between that value and the target value. Finally, the gain value is limited by the maximum possible gain that can be applied without letting the peak absolute signal value exceed the headroom limit.
Figure 18: Block diagram of the loudness measurement process
7.2.1.3 LoudnessAnalysis WebService

The LoudnessAnalysis WebService is used to analyse audio content regarding its loudness. After measuring the audio loudness, the service delivers the gain value of the target loudness. This value can be used as enrichment, for example in a corresponding News Item Container (NIC) used in the NoTube Use Case 7a. On the Home Ambient side, suitable audio hardware components will be able to run a loudness harmonisation on the audio content of the NIC.

Service interface

The LoudnessAnalysis WebService interface has been developed based on the Loudness Analyser component interface, delivering the analysis result. The service offers a single method named loudnessAnalysis. This method accepts a valid URI as file locator for the audio content and provides the analysis result data as output after successful processing. The result is a single element with the value of the actual measured uncompressed gain for the target loudness.

The web service provides a network interface for the underlying Loudness Analyser component, responsible for the actual analysis. A detailed method description of the LoudnessAnalysis WebService can be found in the following figure.

Figure 19: Method description loudnessAnalysis

Service prototype implementation

The LoudnessAnalysis WebService implementation is based on Axis2 [17], a WebService engine provided by the Apache Software Foundation. The service runs as Java-Servlet on Apache Tomcat Server. A WSDL has been defined that describes the API. With the service oriented approach and this WSDL, the NoTube Service Broker provided by WP5 can easily integrate the service and thus the underlying Loudness Analyser component and offer the results as possible enrichment to any application using the broker.

7.2.2 Synthesis stage

If using a metadata schema in the “synthesis stage”, the parameters which have been measured in the “analysis stage”, have to be extracted and formatted in the designed metadata schema for each audio sample to be transmitted. These loudness metadata which characterize the loudness level of each incoming stream or each download are required in order to normalise the loudness level before reproduction of various audio clips. These
metadata have to be generated by loudness measurements at e.g. the same stage in the NoTube chain as the metadata generation for the individual cropping of the video for different displays. These metadata have to be provided with the various audio clips to the rendering engine, i.e. after decoding of the audio in the NoTube terminal. In this case, it is up to the end-user, if he wants to use loudness normalisation or if he wants to listen to the original levels, provided by the program providers.

Another approach is the normalisation of the loudness at the stage of generation of the video and audio metadata. In this case, the end-user gets always a harmonized loudness experience of all audio streams or clips, which have passed the loudness measurement stage of the NoTube chain.

To provide a coherent experience, all audio streams (also from outside sources) have to be analysed and enriched with loudness metadata.

Figure 20 shows the principle of loudness relevant metadata application.

Figure 20: Principle of loudness relevant metadata application

On the transmitting side an audio tool including a loudness meter which meets ITU [18] and EBU [19] loudness recommendations is needed to measure and extract the loudness metadata.

It is proposed that the loudness metadata meet the designated video metadata transmitting format.

On the receiving end audio processing is needed to read and interpret the metadata set with respect to loudness harmonisation. The main features of the audio processor as a component of the NoTube terminal are:

Amplifier (Loudness normalisation meeting the target loudness level)

Limiter (Prevention of override 0 dBFS "clipping" accompanied by nasty distortions)

Compressor (Loudness range adaptation of reproduced audio clip to various listening environments; by means loudness range metadata scalable loudness and loudness range applicable)
### 7.2.2.1 Description of normalisation algorithm

Since the last deliverable, the API of the class “LoudnessAnalyzer” has undergone slight changes: the URL file locator reference was replaced by a file reference. Furthermore, those parts of the results which are not actual results of a measurement, but of applying the measurement to a target value, thus the gain values, were separated and encapsulated in a new class “LoudnessTarget”. Therefore, the creation of audio gain values has become a two-step process.

A newly introduced class "LoudnessNormalizer", which encapsulates the loudness normalisation, was structured following exactly the same model as the existing "LoudnessAnalyzer". It receives both an input and an output file reference in the constructor, where the input file reference must point to a valid audio file in WAV format, as well as the gain values that were stored in a "LoudnessTarget" object.

The normalisation process itself is basically a simple multiplication of the audio sample values with a constant factor $10^{\text{gain} / 20}$. It applies the "safe uncompressed gain" value, which includes consideration of peak absolute sample values.

Since the standard JavaSound capabilities for writing, rather than reading, audio file formats turned out to be extremely poor, it was necessary to bind the program code to another external library, namely the PureJava parts (and only these in order to keep the code portable) of the “Tritonus” project [http://www.tritonus.org/plugins.html].

### 7.2.2.2 Loudness Normalisation Web Service

As a proof of the concept described above, in which the loudness is normalised during the metadata generation stage, a web service was developed which covers the complete process of audio normalisation and metadata enrichment in the context of the use case 7a (see Figure 21).

![Figure 21: Workflow diagram of the Loudness Normalisation Web Service](image)

The service takes as input a metadata set associated to one news program which contains the name of the corresponding program essence file (program.asf). In a first step, an individual news item (item.asf) is extracted from the news program based on the time
interval given in the program metadata. Then the audio is extracted from the news item in WAV format which is required as input format for the loudness analysis (see Figure 22). Based on the loudness analysis results, the audio is levelled to -23 dBFS by the loudness normalisation and then transcoded to the initial audio format WMA2 and multiplexed with the original video. The output is a loudness normalised news item to be used in the Home Ambient of 7a.

The metadata enrichment is realised using the WP2 CRUD service (see D2.3 for details), i.e. the metadata concerning the loudness normalisation is added to the NIC Attractors in TV-Anytime format and can be retrieved from the service in the corresponding CRUD format.

```xml
<crud:AttractorsSetStructure controlledNameID="irt:LoudnessNormalisationSet">
  <crud:unstructuredSet>
    <crud:simpleAttractor controlledNameID="irt:LoudnessNormalisationSet:TargetLoudness">
      <crud:name xml:lang="en">Uncompressed Gain For Target Loudness</crud:name>
      <crud:value baseUnit="dB">-1.3824543</crud:value>
    </crud:simpleAttractor>
  </crud:unstructuredSet>
</crud:AttractorsSetStructure>
```

Figure 22: Added Attractor metadata containing loudness gain value

### 7.3 Evaluation of loudness harmonisation by applying loudness descriptors derived in the analysis stage

#### 7.3.1 Introduction

The problem of loudness jumps when viewing various video clips or listening to pure audio from the internet or zapping between TV or radio stations/programs are very well known for decades. Loudness differences or jumps up to 10 dB can be observed, which means doubling of loudness, which is very disturbing to the end-user. Since the introduction of appropriate loudness algorithms and meters in the past years and mainly since the ITU standardisation ITU-R BS.1770 in 2006 [18] and publication of EBU Technical Recommendation R128 including related documents [19] in 2010 the problem of loudness differences between various sources can principally be solved. It has to be mentioned that the standardisation process has necessarily been accompanied by psychoacoustic testing because loudness is a psycho acoustic parameter whereas loudness measurement by means of a meter is a pure physical measurement. The goal of such a standardisation process is to achieve as high as possible correlations between both psychoacoustics and physics. With respect to the mentioned recommendations we may expect correlation factors >95 % [23]. Nevertheless there are few investigations published that measure or quantify the benefit from loudness normalisation based on ITU-R BS.1770 and EBU-R128 and related. Therefore it seems useful to evaluate the here considered loudness harmonisation for the audio part of various A/V-streams entering the NoTube terminal. Such evaluation implies to carry some kinds of psycho acoustical experiments resulting in the quantification of the subjective assessment of the loudness normalisation under test.

#### 7.3.2 Concept

The concept of the carried out evaluation was to start with the selection of appropriate items with respect to the approach of NoTube, among others video clips with incorporated audio. The selection of the test items was done in two stages, a pre-selection and final selection including editing. A number of 10 - 15 test items of various genres was considered to be sufficient. The duration of the single test items was about 1 min, which is regarded to be a
good compromise with respect to loudness impression and global test run. The finally selected test items were presented in one block (one item after the other without pauses). That means sequences with identical content but different loudness normalisations were to be assessed block-wise. The loudness algorithms under test were as follows:

- Type w, original, audio from source without any processing
- Type x, normalised referring to EBU-R128 and related without gating
- Type y, normalised referring to EBU-R128 and related with gating
- Type z, normalised referring to valid European peak normalisation (-9 dBQPPM, 10 ms integration time)

In a controlled listening experiment test subjects were asked to assess the global loudness impression of the sequences under test (Type w, x, y, z anonymously indicated as A, B, C, D) to acquire a ranking of the test sequences by means of a ranking scale, comparable to German school marks (1-4, see Table 2).

<table>
<thead>
<tr>
<th>ranking</th>
<th>rank indication</th>
<th>rank number</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>excellent</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>good</td>
<td>2</td>
</tr>
<tr>
<td>D</td>
<td>fair</td>
<td>3</td>
</tr>
<tr>
<td>A</td>
<td>poor</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 2: Example of ranking results for one set of sequences

In order to simplify the assessment the following questionnaire was presented to the test subjects where they were asked to mark the reproduction of the corresponding test item if the loudness was subjectively not convenient, i.e. if they would like to use the remote to adjust the loudness manually. Thus the rank order could be easily read by the amount of marks, whereas the rank order of amount of marks represents the rank order based upon the given rank indications or rank numbers. An example is presented in Table 2 (ranking order: B=excellent, C=good, D=fair, A=poor).

**Questionnaire**

Evaluation of loudness correction module in NoTube

Dear test subject,

In the test 4 sequences (A, B, C, D) with 16 short (approx. 1 min) video clips of different genres are presented to you. Please assess the loudness of the single clips regarding the following aspect:

Does loudness appear to be convenient or would you like to use the remote control in order to change the loudness of the corresponding clip (softer/louder)? Please try to assess not only the transitions between the clips but also the loudness of the entire single clip. If you'd like to use the remote control please mark the considered clip.

Thank you!

Name/Age: ..................  Date: ..........................

<table>
<thead>
<tr>
<th>Sequence A</th>
<th>Sequence B</th>
<th>Sequence C</th>
<th>Sequence D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tagesschau</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audi</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Up</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wissen</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7.3.3 Test items
Initially, it was planned to compose the test items selection from the test material delivered by NoTube partners. But because of the relatively poor video quality it was decided to select test items from other sources. Except the Turkish advertisements (ads) all the test items were recorded from DTV or downloaded from internet.

The final selection of 16 test items is shown in the following table:

<table>
<thead>
<tr>
<th>indication</th>
<th>content</th>
<th>source</th>
<th>language</th>
<th>duration (min:sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tagesschau</td>
<td>news</td>
<td>broadcasting</td>
<td>German</td>
<td>0:55</td>
</tr>
<tr>
<td>Audi</td>
<td>car ad</td>
<td>internet</td>
<td>German</td>
<td>0:47</td>
</tr>
<tr>
<td>Avatar</td>
<td>movie trailer</td>
<td>internet</td>
<td>English</td>
<td>0:58</td>
</tr>
<tr>
<td>Basketball</td>
<td>sport reportage</td>
<td>internet</td>
<td>English</td>
<td>1:15</td>
</tr>
<tr>
<td>Bild</td>
<td>newspaper ad</td>
<td>broadcasting</td>
<td>German</td>
<td>0:29</td>
</tr>
<tr>
<td>Billard</td>
<td>sport reportage</td>
<td>broadcasting</td>
<td>German</td>
<td>0:24</td>
</tr>
<tr>
<td>French cheese</td>
<td>comestible ad</td>
<td>internet</td>
<td>German</td>
<td>0:14</td>
</tr>
<tr>
<td>Letterman</td>
<td>talk</td>
<td>internet</td>
<td>English</td>
<td>1:10</td>
</tr>
<tr>
<td>NJK</td>
<td>orchestra</td>
<td>internet</td>
<td></td>
<td>1:20</td>
</tr>
<tr>
<td>No country</td>
<td>movie trailer</td>
<td>internet</td>
<td>English</td>
<td>0:43</td>
</tr>
<tr>
<td>Sarah Connor</td>
<td>vocals</td>
<td>internet</td>
<td>English</td>
<td>0:59</td>
</tr>
<tr>
<td>Shampoo</td>
<td>cosmetic ad</td>
<td>broadcasting</td>
<td>German</td>
<td>0:24</td>
</tr>
<tr>
<td>Sibelius</td>
<td>violin concerto</td>
<td>internet</td>
<td></td>
<td>0:47</td>
</tr>
<tr>
<td>Turkish cheese</td>
<td>comestible ad</td>
<td>NoTube partner</td>
<td>Turkish</td>
<td>0:29</td>
</tr>
<tr>
<td>Up</td>
<td>movie trailer</td>
<td>internet</td>
<td>English</td>
<td>0:56</td>
</tr>
<tr>
<td>Wissen</td>
<td>health ad</td>
<td>broadcasting</td>
<td>English</td>
<td>0:35</td>
</tr>
</tbody>
</table>

7.3.4 Test set-up

The test was carried out simulating a stereo listening situation at a PC working place with PC desktop loudspeakers beside the monitor. But in order to minimize the influence of the wide range of PC speaker quality small studio monitor speakers, meeting EBU Tech. 3276 [24], were chosen (Klein&Hummel/Neumann O110) [25]. Furthermore, an extended stereo basis of 2.4 m was chosen in order to allow more than one test subject to take part in one test session. In addition to exclude the acoustic environment of typical bureaus the tests were carried out in a professional listening room meeting EBU Tech. 3276 [24].

The test setup is presented in the following photo (see Figure 23) with typically three test persons sitting in one line behind each other.
7.3.5 Results

To begin with, in Figure 24 the loudness descriptors “integrated loudness”, derived in the “analysis stage”, is presented for all test items and loudness algorithms under test. It can be indentified that there are remarkable differences with respect to LUFS between normalisation (type w_original) and (type x_norm_ungated/y_norm_gated/z_norm-9QPPM) as well as between (type z_norm-9QPPM) and (type x_norm_ungated/y_norm_gated) whereas there are only small differences between (type x_norm_ungated) and (type y_norm_gated).
The loudness assessments based on the above presented questionnaire filled in by 13 test persons were interpreted by means of the individual rank order of each test subject, using the rank numbers as indicated in the corresponding table above. The resulting database was analyzed by calculation of the means and 95 % confidence intervals of the means. The confidence intervals of the means give a range of values around the mean value in which it is expected that the “true” (population) mean is located with a given level of certainty. In the presented results the level of certainty (confidence) is 95 %. The range defined by the confidence interval is useful for estimating the significance of mean differences when comparing different results. If the confidence intervals overlap, it can be deduced that the means do not differ significantly (in the statistical sense). On the other hand, non-overlapping confidence intervals indicate that the means differ significantly. The corresponding results are presented in Figure 25 with respect to the average of all test items.
Concluding, the evaluation of loudness harmonisation by applying loudness descriptors derived in the analysis stage of the loudness module, shows a significant improvement compared to the original without any normalisation and compared to valid European QPPM normalisation.

7.4 User evaluation of loudness harmonisation on the web

7.4.1 Introduction

The carried out evaluation of different loudness harmonisation methods clearly showed the excellent performance of EBU-R 128 (see 7.3). However we have to keep in mind that this test had been carried out under quasi studio listening conditions including performance of monitors and listening room as well as geometry of the set-up (“sweet spot”) under test. Regarding listening set-ups in home environments there is a wide spread of listening conditions with respect to overall performance starting from very poor like “kitchen radio” to studio-quality like “home cinema”. In the carried out loudness web evaluation our intention was to include actual listening conditions of listeners. Based upon the tools applied by EBU-R 128 and corresponding technical documents we wanted to investigate the interdependence between loudness harmonisation, loudness range characteristics and listening condition. In the framework of NoTube, considering use of computer, absorbing streaming media data and web, it is self-evident that the user's listening environment typically is characterized as computer based listening environment among others built-in laptop speakers respectively small extern PC speakers. But considering the proceeding fusion from internet and TV (HbbTV) it can be expected that also studio-quality stereo systems and home cinema speaker system have to be included in such loudness harmonisation considerations.
7.4.2 Concept

The evaluation shall be done based upon a sufficient number of short video clips covering different genres like “Movie”, “Commentary”, “Concert”, “Sport”, “Show”, “Commercial”, “News” etc. The audio part of the selected clips to be evaluated shall be varied well-defined with respect to Programme Loudness and Loudness Range. In each case three variants shall be presented. The task of the participant is simply to indicate the preferred variant. Although EBU-R 128 defines the Target Programme Loudness it doesn't enclose any relation to the reproduced sound pressure level. This relation is among others dependent from individual taste, preference, custom and listening environment. This fact implies that such loudness evaluations are strongly influenced by the parameter “individual volume”. In order to include this important individual parameter a corresponding procedure was prepared to adjust and register the “individual volume” (see…). It shall be recommended that the finally adjusted volume defines the individual reference volume and shall be constant and not be change during the complete evaluation procedure. That means not a constant dictated listening level is to be recommended but each participant shall be able to choose his preferred listening level individually. Operating the loudness web evaluation it was decided to use the services of YouTube with respect to storage and playback of the video clips.

7.4.3 Selection and preprocessing of test items

7.4.3.1 Evaluation of loudness adaptation

For the evaluation of mere loudness adaptation, which is part of the evaluation questionnaire, the following short extracts from 10 video clips, representing different genres, were selected. Most of the video clips are downloaded videos provided by YouTube, few of them were provided by public German broadcasting corporations and the NoTube partner RAI.

<table>
<thead>
<tr>
<th>Identification</th>
<th>Genre</th>
<th>Duration/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBC-Dangerous Knowledge_D</td>
<td>Documentary</td>
<td>27</td>
</tr>
<tr>
<td>Canon Wildlife_D</td>
<td>Documentary</td>
<td>27</td>
</tr>
<tr>
<td>Cars 2_M</td>
<td>Movie</td>
<td>16</td>
</tr>
<tr>
<td>Esbjorn Svensson Trio_C</td>
<td>Concert</td>
<td>14</td>
</tr>
<tr>
<td>BBC world_D</td>
<td>Documentary</td>
<td>18</td>
</tr>
<tr>
<td>Joey Kelly_S</td>
<td>Show</td>
<td>16</td>
</tr>
<tr>
<td>Lena Meyer-Landrut_S</td>
<td>Show</td>
<td>24</td>
</tr>
<tr>
<td>Basket ball_S</td>
<td>Sport</td>
<td>14</td>
</tr>
<tr>
<td>Fiat_C</td>
<td>Commercial</td>
<td>18</td>
</tr>
<tr>
<td>Porsche_C</td>
<td>Commercial</td>
<td>36</td>
</tr>
</tbody>
</table>

Table 3: Video clips description

The compilation of the 3 variants of audio adaptation under test were carried out by measurement and adaptation of Programme Loudness, whereas measurement and adaptation were operated by the “Loudness Analysis Web Service” (see 7.2.1.3) and “Loudness Normalisation Service” (see 7.2.2.1) developed within the framework of NoTube.
Table 4: Audio adaptation types

<table>
<thead>
<tr>
<th>Adaptation</th>
<th>Resulting Programme Loudness/LUFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loudness harmonisation referring to EBU-R 128 (Target Level of Programme Loudness)</td>
<td>-23</td>
</tr>
<tr>
<td>Amplifying of Programme Loudness by + 4 dB</td>
<td>-19</td>
</tr>
<tr>
<td>Attenuating Programme Loudness by -7 dB</td>
<td>-30</td>
</tr>
</tbody>
</table>

7.4.3.2 Evaluation of loudness range adaptation

The following 5 extracts of video clips from different genres were selected for evaluation of loudness range adaptation.

<table>
<thead>
<tr>
<th>Identification</th>
<th>Genre</th>
<th>Duration / s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knight and Day_M</td>
<td>Movie</td>
<td>30</td>
</tr>
<tr>
<td>Offenbach_C</td>
<td>Concert</td>
<td>53</td>
</tr>
<tr>
<td>Beethoven_Lang Lang_C</td>
<td>Concert</td>
<td>38</td>
</tr>
<tr>
<td>Shostakovich_C</td>
<td>Concert</td>
<td>31</td>
</tr>
<tr>
<td>BMW M5_C</td>
<td>Commercial</td>
<td>29</td>
</tr>
</tbody>
</table>

Loudness range adaptation under test is strictly based on measurement of the loudness descriptor “Loudness Range (LRA)” as specified in EBU-TECH 3342. LRA and Programme Loudness of the video clips under test both can be measured by means of the “Loudness Analysis Web Service” (see 7.2.1.3). Whereas mere loudness adaptation of audio is simply achieved by analyzing and amplifying or rather attenuating Programme Loudness by the corresponding LU respectively dB valuation loudness range adaption is not such easy. Achieving loudness range adaptation you need to compress/expand the audio. There is a wide spread of compressors with different characteristics available normally to be intended for individual applications. Basically the performance of compressors is described by the static and dynamic characteristics. In this case a simple broadband compressor (VST Plug-in DirectX-Waves) with the following static and dynamic performance was adopted.
The dynamic characteristics of the compressor under test (c1_25 dB) are presented in Figure 27. Principally the design of a compressor without perceivably impairments resulting from the control processing requires short attack times (e. g. 2 ms) and long release times (e. g. 5 s). In order to display the dynamic characteristics in Figure 27, a sinus signal (400 Hz) is used with "stair-shaped" envelope.
Figure 27: Compressor: Dynamic characteristics

Figure 28 shows the resulting LRA of the loudness range items after compression under test.

Figure 28: Loudness Range of items under test (uncompressed and compressed c1_15 dB/c1_25 dB)
7.4.3.3 Listening level

In order to enable the participant to adjust and to indicate his individual listening level in the introduction part of the questionnaire a short extract from a news broadcast were selected as test item. Because our ears are especially familiar with the sound of human speech news anchors are ideally qualified adjusting the convenient (or even correct respectively original) listening level. The finally adjusted individual listening level is considered to be the reference volume and should not be changed during the following complete test procedure. Achieving to log the individually adjusted listening level the following procedure was disposed. A special test signal was recorded with announcements of different listening levels, whereas the announced listening level meets the corresponding loudness level. The relationship between announcement and loudness level is presented in the following table.

<table>
<thead>
<tr>
<th>Announcement</th>
<th>Programme Loudness/LUFS</th>
<th>Sound pressure level/dBA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 7</td>
<td>-72</td>
<td>32</td>
</tr>
<tr>
<td>Level 6</td>
<td>-65</td>
<td>39</td>
</tr>
<tr>
<td>Level 5</td>
<td>-58</td>
<td>46</td>
</tr>
<tr>
<td>Level 4</td>
<td>-51</td>
<td>53</td>
</tr>
<tr>
<td>Level 3</td>
<td>-44</td>
<td>60</td>
</tr>
<tr>
<td>Level 2</td>
<td>-37</td>
<td>67</td>
</tr>
<tr>
<td>Level 1</td>
<td>-30</td>
<td>74</td>
</tr>
</tbody>
</table>

These level announcements are presented after adjustment of the individual reference listening level. The participant is asked to indicate the first level announcement which he clearly can understand by clicking the corresponding button. This method is well-known as “Hearing Threshold Method” and presumed to be notably precise. The differences between the announced levels are 7 LU. From psychoacoustics point of view this difference can be indicates as “clearly distinguishable“. An estimation of the relationship to the corresponding sound pressure level can be achieved by measuring the resulting reproduced average sound pressure level of the anchorman on an individual listening condition.

The listening level adjustment and registration is part of the introduction page of the questionnaire. Additionally the introduction page contains a video sequence which is composed of 9 short clips with different loudness adaptations (0 : 6 : 6 : -9 : 0 : 8 : -9 : 8 : 0 LU). This sequence is presented to get familiar with the kind of loudness adaptation under test.

Finally the introduction page contains the questionnaire to register age and characteristics of the individual listening condition. In detail the following parameters are captured:

- age
- kind of speaker
- size of speaker
- distance to speaker
- background noise

After adjusting and indicating listening level and answering all questions about listening condition the participant is allowed to skip to the following page which presents to actual evaluation part. The arrangement of the loudness evaluation is each clip under test is presented in those 3 variants which have been described above. The participant is asked to listen to each of the 3 variants of the actual clip and then to indicate which loudness/loudness range variant he personally prefers considering his taste/custom respectively his individual listening environment. After indicating the preferred variant by clicking the corresponding button the next clip with its 3 versions is presented for evaluation. The order of the presentation of clip number respectively variant is done randomly. Naturally there is a possibility to abort and restart the evaluation part.
7.4.3.4 Organisation of loudness web evaluation

The action of the loudness web evaluation was performed by means of a corresponding html-script which was provided on the IRT server. The script included management of the evaluation procedure and register and storage of the user relevant data. The complete clip handling including storage and playback was organized by using the corresponding YouTube services. Activating the clips in this context is realized by means of the URL addresses which are automatically provided by YouTube when uploading the clips.

7.4.4 Results

The evaluation period took from 19\textsuperscript{th} of September to 12\textsuperscript{th} of October (3 week plus 2 days). Although we invited some other communities besides NoTube partners among others the EBU audio expert group “FAR-PLOUD” participation was not as strong as expected. Thus the below presented results are only based on 43 participants. In general the presented data in this stage of analysis are to be considered as pure descriptive; in detail the calculated percentage rate is presented in each case.

![Figure 29: Age of participants](image-url)
The results with respect to listening condition are presented in Fig. 31-34.

Figure 31: Category of speaker
Figure 32: Size of speaker

Figure 33: Distance to speaker (PC speaker & stereo system)
The results of evaluation of loudness and loudness range adaptation are presented in Fig. 35-36.

Figure 34: Background noise

Figure 35: Evaluation of loudness adaptation
7.4.5 Conclusions

First conclusions from the presented descriptive data shall be drawn with respect to evaluation of loudness and loudness range adaptation. Comparable to the former evaluation of different loudness harmonisation methods (see 7.3) the excellent performance of EBU-R 128 with respect to Target Programme Loudness is approved. This result answers the open question whether the clearly preferred EBU-R 128 harmonisation is dependent from individual listening conditions. Considering the covered listening conditions in this loudness web evaluation there is clearly no influence observable. With respect to the evaluation of different loudness range adaptation there is a tendency identifiable. The participants seem to prefer rather medium loudness range (compressed c1_15 dB/c1_25 dB) than uncompressed audio with high loudness range. In the next stage of analysis of the loudness web evaluation database being on hand among others we are interested in interdependences respectively correlations between the parameters under test.
8 Conclusion

This deliverable has provided the functional description of audio/video content analysis components used in the NoTube use cases. The automatic tools allowing the extraction of audio/video extraction metadata are described in details, and correspond more precisely of metadata on the video-part for the determination of the region of interest, metadata on the video-part for the determination of the best Ad insertion region, metadata on automatic news segmentation from newscast transcriptions, metadata based loudness harmonisation of various audio programmes coming from different distribution media.

9 References


[17] [http://axis.apache.org/axis2/java/core/](http://axis.apache.org/axis2/java/core/)


[19] [http://tech.ebu.ch/Jahia/site/tech/cache/offonce/loudness;jsessionid=EA0B4578EE80E85A722985425B87B5DB.jahia1](http://tech.ebu.ch/Jahia/site/tech/cache/offonce/loudness;jsessionid=EA0B4578EE80E85A722985425B87B5DB.jahia1)


[25] KH O110: Active Near-field Monitor, 5.5" + 1" drivers, magnetically shielded, 6.5 liters, 80 + 80 W, 107.7 dB SPL, 56 - 24k Hz.