

ListeningTV: Accessible Video using Interactive Audio Descriptions

Alex de Souza Vieira
alexvieira@unifesspa.edu.br
FACEEL/UNIFESSPA

Lucas Ribeiro Madeira
lucasrm148@gmail.com
TeleMídia-MA/UFMA

Álan Lívio V. Guedes
alan@telemidia.puc-rio.br
TeleMídia/PUC-Rio

Sérgio Colcher
colcher@inf.puc-rio.br
Department of Informatics/PUC-Rio

Daniel de Sousa Moraes
danielmoraes@telemidia.puc-rio.br
TeleMídia/PUC-Rio

Carlos de S. Soares Neto
csalles@deinf.ufma.br
TeleMídia-MA/UFMA

ABSTRACT

People with visual impairments suffer from the incapacity to understand contextual information in videos, such as the place where characters are, or any other non-spoken actions in general. Some content creators address this issue by providing a secondary audio to describe such information, called *Audio Descriptions* (ADs). However, some works in the literature have highlighted that people with visual impairment are usually not able to completely understand scene changes based only on characters' voices or traditional ADs. Moreover, traditional ADs do not completely describe some of the important visual information, such as the background scenery (e.g. colors, furniture) and characters' details (e.g. blond woman using a red dress). In this work, we propose incrementing the traditional AD techniques with the usage of interactive video features present in TV systems. More precisely, the proposed interactivity enables users to access specialized AD for different visual information (e.g., scene, scenario, character). To support the development of such interactive content, we present an application template, which helps to create the final interactive-enhanced video application. As a proof of concept for our approach, we created an interactive AD for an independent video mainly composed of visual information, with only a few talks.

KEYWORDS

datasets, neural networks, gaze detection, text tagging

1 INTRODUCTION

TV is an important device for delivering information, education, and entertainment. However, people with visual impairments suffer with their inability to understand contextual information in videos, such as the place where characters are, or any other non-spoken actions in general. Particularly in Brazil,¹ there are about 6.5 million persons with some kind of visual impairments (fully or partial blindness). Due to this disability, these people are not able to fully understand contextual information in videos [2, 5].

¹https://biblioteca.ibge.gov.br/visualizacao/periodicos/93/cd_2010_caracteristicas_populacao_domicilios.pdf

Currently, the main tool used to support accessible audiovisual content is the use of an additional audio track describing visual information [4]. This audio track, named *Audio Description* (AD), consists of a generic description of some scenes by a narrator. Its production should follow guidelines and recommendations about the best temporal spacing in which audio descriptions should be inserted and their adequate speed. However, these recommendations do not define what should be described in the scene. In particular, some work in the literature [8] has already highlighted that impaired viewers are usually not able to completely understand scene changes based only on characters' voices and traditional ADs. Moreover, traditional ADs do not completely describe some of the important visual information, such as the background scenery (e.g. colors, furniture) and characters' details (e.g. blond woman using a red dress).

In this sense, this work describes our ongoing research to provide what we call *Interactive Audio Descriptions* (iADs). As illustrated in Figure 1, we increment the traditional AD techniques with interactive video features present in TV systems. Such ADs aim at better supporting impaired viewers to understand not only the scene changes but also other visual information not supported by traditional ADs.

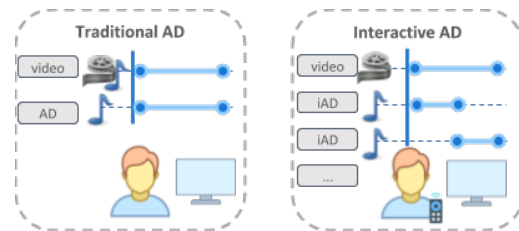


Figure 1: Interactive AD overview

We propose segmenting the traditional AD into six different types. Three are audio tracks, such as traditional ADs, but specialized with regard to (1) *scene*, (2) *scenery*, and (3) *character*. The other three, which can be accessed at any time, are complementary information regarding (4) *title*, (5) *synopsis* and (6) *interaction help*.

To support the authoring of such interactive content, we present an application template that helps to create a final interactive-enhanced video application. More precisely, we proposed a template for the NCL language [6], which support the development of interactive TV applications. A template [3] is an alternative that

makes it possible to optimize the process of developing interactive content, since the author of the document is relieved from many details and can focus mainly on the selection of the media that will compose the document. As a proof of concept, we present the result of creating an interactive AD for an independent video mainly composed of visual information, with only a few talks.

This paper is structured as follows. Section 2 introduces the traditional AD and its limitations. Then, Section 3 and Section 4 details our methodology and proposed application template, respectively. Finally, Section 6 presents final remarks and future work.

2 TRADITIONAL AUDIO DESCRIPTION

Visual impairment is a decreased ability to see to a degree that causes problems not fixable by usual means, such as glasses. Thus, audio description (AD) presents itself as an accessibility resource that gives people who have some kind of visual impairment access and enjoyment of content presented by audiovisual means. It consists of a narrated description of the main visual elements of a video, created scene by scene, from the information collected from the content. In the context of films, the audio description translates images, plot, scenery, actions, among other elements of history.

Some work in the literature has already highlighted that impaired viewers are usually not able to completely understand scene changes based only on characters' voices and traditional ADs. Vieira and Correia [8], for instance, present an experiment to verify which were the main characteristics perceived by visually impaired users when facing audiovisual content. In their experiment, YouTube videos were presented to the blind, low vision, and blindfolded subjects. The subjects were asked about the number of scenes, number of scenarios, number of characters and specific characteristics, perceived in the videos. The results indicated that, concerning the number of scenarios, most of the time the participants were wrong, informing a lower number than what existed. Regarding the number of characters, only two participants (one with low vision and the other with blindfold) got it right, despite the proximity of correct answers from the other participants. Regarding the description of the scenarios, it was noticed that there was a greater success in those videos with less variations; those presenting a greater diversity of scenarios caused higher rates of identification errors. Finally, none of the elements analyzed were 100% correct by the participants, and scenarios proved to be the most incomprehensible characteristic to the participants.

3 PROPOSED APPROACH

To better support impaired viewers, we revisited the traditional audio descriptions and segmented them into different categories. Each category is defined with a different level of detail that can be inserted by the author, and with a different kind of freedom regarding the moment of consumption by the end-user.

More precisely, we propose six different types of descriptions that can be interactively accessed. Three of them are audio tracks, such as traditional ADs, but specialized with regard to (1) *scene*, (2) *scenery*, and (3) *characters*. These three are accessible on a scene-by-scene basis and compose what we call *Specialized AD*. The other three are interactive content that can be accessed at any time by the user, and are complementary information regarding (4) *title*,

(5) *synopsis* and (6) *interactive help*. These last three are what we call the *Complementary AD*. To support the authoring of all of these content categories, we present an application template, which helps to create a final interactive-enhanced video application. When selecting one of them, the original audio of the content is enriched by the audio description corresponding only to the interactive AD selected by the user.

The first three types of descriptions, the so called *Specialized Audio Description*, have the following characteristics:

- *Scene AD*: a scene corresponds to the development of different acts that happen successively in a determined interval of the video. This AD plays along with the main audio of a scene, describing the visual interaction among its main elements and with the environment. These elements are objects movements, actions, expressions, or people signs that happen during a narrative. This kind of audio description should be paused (or its sound level should be strongly reduced) whenever there is a dialogue or when someone speaks. It is important to keep the user updated about the scene even when there is not any change in the scene. In other words, the same audio description should be enabled again a few seconds later, even if there was not any scene changes.
- *Scenario AD*: the scenario is a space (physical or not) of representation where a scene occurs. This AD activates an audio description of the main elements of the scenario like architecture, shapes, colors, dimensions, obstacles, and other objects present in these places. Similar to what happens in a scene, the audio description of a scenario should be paused (or its sound level should be strongly reduced) whenever occurring simultaneously with any speaking in the main audio so that visually impaired people can understand the linked information.
- *Character AD*: the character is an entity like a person, animal, or a cartoon object on a scenario and scene. Once activated, this AD shows the audio description about the character, one by one, and it is composed of the set of main elements like height, weight, skin color, hair color, wearing and the face details. As usual, this AD needs to be deactivated or to have its sound level reduced to prioritize the original audio of the characters' dialogue. Moreover, if there is no character present in the scene, the user should also be informed.

For the three *Complementary Audio Descriptions*, we define the following characteristics and requirements:

- *Title AD*: an audio with the video title, accessible at any time by the user, This audio should not have any descriptive information about the title.
- *Synopsis AD*: it consists of a clean presentation of the audiovisual content. This AD contains the summary of audiovisual content and the credits of the authoring process. In other words, it is a piece of audio information that summarizes the audiovisual narrative linked to it, and that names the production team.
- *Help AD*: an audio developed to guide the navigation and selection of the interactive AD.

4 NCL TEMPLATE

This section introduces our template, called *ListeningTV*, which enables authors to produce content in which users can navigate

and select interactive ADs using the standard TV remote control. A contribution of this template is the possibility of making existing audiovisual content accessible by adding new media with audio description to them. Figure 2 shows two main screens of the *ListeningTV* template. One screen is for iteration invitation (Figure 2a) while the other consist on interactive AD selection (Figure 2b).

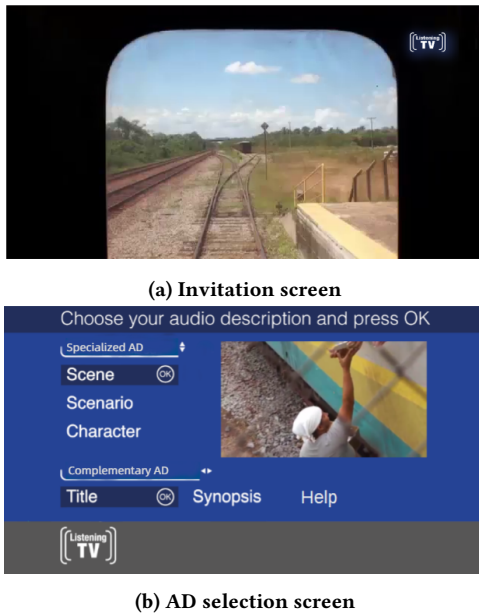


Figure 2: Template screens

The navigation and selection scheme uses the TV remote control arrows and the OK/ENTER button. The first interaction happens when an invitation icon appears on the upper right corner (Figure 2a), and audio alerts that the application is available. Once the invitation is selected, the main video is resized and a menu is shown (Figure 2b). On this menu, we present the AD options disposed beside the main video. The navigation is available using the TV remote control arrows in two flows, vertical and horizontal. The vertical flow allows navigation among the *specialized ADs* (i.e. scene, scenario, character), while the horizontal among the *complementary ADs*. There are audio when navigating to each AD option, that informs the impaired user which AD is currently highlighted. Once the user presses the button OK/ENTER, the audiovisual content returns to occupy the entire screen area again. Additionally, the original audio of the content is enriched by the audio description corresponding only to the AD selected by the user.

The *ListeningTV* template was developed in NCL language, which supports the development of interactive TV applications. We can see a code fragment from it in Listing 1, which for simplicity, focuses on the specialized ADs selection. It defines the main video and interactive ADs using the audio and button `<media>` elements (lines 5-17). The application starts with the main video and the scenario AD (lines 3-4). To enable the specialized AD selection, we use the sound level control (lines 18-29). When selecting an AD respective button, the application changed its sound level to 80%, and the other audios (scenario and character) are muted. This template

can be customized by the developer, which simply changing the source (`src`) attribute of the main video and ADs audio `<media>`.

```

1 <ncl>
2 ...
3 <port id="entry" component="mainVideo" />
4 <port id="entry" component="scenarioAD" />
5 <media id="main" src="main.mp4" descriptor="d1" />
6 <media id="sceneBtn" descriptor="d2" />
7 <media id="scenarioBtn" descriptor="d3" />
8 <media id="charactersBtn" descriptor="d4" />
9 <media id="sceneAD" src="sceneAD.mp3">
10   <property name="soundLevel" value="100%">
11 </media>
12 <media id="scenarioAD" src="scenarioAD.mp3" />
13   <property name="soundLevel" value="0">
14 </media>
15 <media id="charactersAD" src="charactersAD.mp3" />
16   <property name="soundLevel" value="0">
17 </media>
18 <link xconnector="onSectionSet">
19   <bind role="onSelection" component="scenarioBtn"
20     />
21   <bind role="set" component="scene" interface="
22     soundLevel">
23     <bindParam name="var" value="0" >
24 </bindParam>
25   <bind role="set" component="scenarioAD"
26     interface="soundLevel">
27     <bindParam name="var" value="80%" >
28 </bindParam>
29 </link>
30 ...
31 <ncl>

```

Listing 1: Code fragment from the *ListeningTV* template.

5 USE CASE



To validate our approach, we have created an interactive AD for an independent video.² This video mainly focus on visual information and has only a few talks. Its has a duration of 9 minutes and 29 seconds, and shows a train journey in an underprivileged region. It is has a strong visual dependence, and a dynamic variation of scenarios and characters, which appear for a few seconds in each scene. Thus, this video represents a challenge for impaired viewers users because it is a more “introspective” video.

Some examples of interactive ADs for this case are described in Table 1. During the video, the train passed through several cities and stopped at one station. In this moment, we selected two scenes: one showing the boarding of people, and the other in which a vendor

²<https://drive.google.com/file/d/1eaeVnDL-6Umm1b-Jz6uEmlAqZDLRB-kA/view?usp=sharing>

sells food from the outside the train. The table presents screenshots of these scenes and their respective ADs. To further demonstrate these ADs usage, we have created a video.³

Table 1: Interactive AD samples from our Use Case

Video scene sample	Interactive ADs
	<ul style="list-style-type: none"> – scene AD: People are boarding the train. – scenario AD: The train wagon has green and yellow colors. – character AD: A teenager wears a blue blouse, shorts, and black sandals.
	<ul style="list-style-type: none"> – scene AD: Saleswoman delivering food. – scenario AD: The ground has gravel bellow the train wagon. – character AD: A saleswoman wears a gray blouse and a cloth over her head.

Regarding the audio, we produced the ones containing audio descriptions about the scenario, characters, scene, title, summary, and user help. These audio descriptions were recorded with a simple Voice Recording application. Then, the NCL application template (discussed in Section 4) was filled with media files to create the final version. Such development, was done using the Eclipse IDE.⁴

6 FINAL REMARKS

In this paper we propose to increment the traditional AD techniques with the usage of interactive video features present in TV systems. More precisely, the proposed interactivity enables users to access: (1) specialized AD, regarding *scene*, *scenery* and *character* for each scene, and (2) complementary AD, regarding *title*, *synopsis* and *interaction help*, in any time. To support the authoring of such interactive content, we present an application template, which helps authors to create an interactive-enhanced video application. As a proof of concept for our approach, we present the result of creating an interactive AD for an independent video mainly composed of visual information, with only a few talks.

One difficulty encountered when creating the audio descriptions was to accommodate ADs when scene changes are very fast; in this case, the challenge was to create small chunks of audio that fitted into short intervals, but still being understandable to the impaired person.

Another problem was that, if the user decides to change the category in the middle of a scene, he/she may change to an AD that will not start playing from its beginning, and in this case, he/she might miss important information.

³https://drive.google.com/file/d/1562kbugyaqyD7uWvU8OA76_tViqv9p0u/view?usp=sharing

⁴<https://marketplace.eclipse.org/content/ncl-eclipse>

As we have previously mentioned, these are topics of our ongoing research in order to fully support *Interactive Audio Descriptions*. As our next steps, we first intend to perform evaluation tests to understand whether our approach is well accepted by impaired viewers. We also plan to investigate an important variation of our specialized ADs, in which they are split into different media files instead of one for the entire video. In this way, they might be started at the beginning of each scene and can be actively repeated by the user if not understood by any reason. It could also solve the problem of arriving in the middle of a scene and missing information from an AD. The producing of AD is a costly and complex process [7]. We can use technological advancements to help facilitate this process. In particular, this path may enable users to use Machine Learning techniques to dynamic create the ADs for each based on visual information scenes [1].

REFERENCES

- [1] V. P. Campos, L. M. G. Goncalves, and T. M. U. de Araujo. 2017. Applying audio description for context understanding of surveillance videos by people with visual impairments. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. 1–5.
- [2] Cosette Castro. 2014. Televisão digital e as possibilidades de acessibilidade audiovisual no Brasil. 0, 5 (2014). <https://doi.org/10.19174/esf.v0i5.5692> Number: 5.
- [3] Thacyla de Sousa Lima, Roberto Gerson de Albuquerque Azevedo, and Carlos de Salles Soares Neto. 2019. Increasing Reuse in Learning Objects Authoring: A Case Study with the Cacuriá Tool. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web (Rio de Janeiro, Brazil) (WebMedia '19)*. Association for Computing Machinery, New York, NY, USA, 193–200. <https://doi.org/10.1145/3323503.3349555>
- [4] Leonardo A. Domingues, Virgínia P. Campos, Tiago M.U. Araújo, and Guido L. de S. Filho. 2016. Accessibility in Digital Cinema: A Proposal for Generation and Distribution of Audio Description. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web (Teresina, Piauí State, Brazil) (Webmedia '16)*. Association for Computing Machinery, New York, NY, USA, 119–126. <https://doi.org/10.1145/2976796.2976867>
- [5] Marília Matos Gonçalves, Giorgio Gilwan Silva, and Robson Freire. 2015. Acessibilidade da TV digital interativa para deficientes visuais. *Human Factors in Design* 4, 8 (2015), 152–173.
- [6] ITU. 2014. *Recommendation H.761: Nested Context Language (NCL) and Ginga-NCL for IPTV Services*. ITU. <https://www.itu.int/rec/T-REC-H.761>
- [7] Rita Oliveira, Jorge Ferraz De Abreu, and Ana Margarida Almeida. 2016. Audio Description in Interactive Television (iTV): proposal of a collaborative and voluntary approach. *Procedia Computer Science* 100 (2016), 935–940.
- [8] Alex de Souza Vieira and Derek Oliveira Correia. 2016. Um Olhar Sobre Produção e Consumo de Conteúdos Audiovisuais Tradicionais Com Foco Nas Pessoas Com Deficiência Visual. In *7º Congresso Brasileiro De Educação Especial*. UFScar.