An Evaluation Framework for Automated Audio Description

Pacurar, Cristian

https://doi.org/10.34737/wzvw9

2024

An Automated Audio Description Evaluation Framework

PREPARING FOR THE FUTURE OF AUDIO DESCRIPTION AUTOMATION

CRISTIAN PACURAR

Abstract

The United Nations Convention on the Rights of Persons with Disabilities (CRPD, 2006) stipulates that all individuals have the right to access information and communicate through means of their choice. This underscores the fundamental right to be informed and access information. However, information is not always accessible for people with disabilities, particularly those with visual impairments. With recent advancements in AI models, such as GPT-4 with Vision by OpenAI and the Pegasus-1 model by Twelvelabs, the automation of the audio description process is becoming increasingly feasible.

The initial goal of this research was to create an automated system capable of generating audio description tracks automatically, thereby increasing accessibility for blind or partially sighted persons. The premise of the research was that the human audio description process could be split into smaller steps, each of which could be automated and then chained together.

Currently, there is no established framework for assessing the efficacy of algorithms in automating the audio description process. Although various algorithms can replace human audio describers in certain steps, there are no key performance indicators (KPIs) for analysis and comparison. Furthermore, there is no standardised method for evaluating and comparing multiple algorithms performing specific audio description tasks, which hinders objective decision-making.

To address this gap, the initial step involved analysing the stages of the human audio description process and breaking them down into self-contained actions suitable for automation. This led to the conceptualisation of an automated audio description system designed to replicate the entire human process.

To demonstrate the practicality of the evaluation framework, a partially automated system was developed, focusing on automating the creation of the audio description script. This system serves as a proof of concept for the usability and effectiveness of the proposed framework. Nevertheless, due to the absence of globally accepted guidelines, multiple guidelines were compared and synthesised to create a unified set of KPIs which could then be used in the evaluation framework.

Contents

1. Introduction

1. Importance of the topic

Vision represents a fundamental aspect of our day-to-day life, and one that we, most of the time, take for granted. From the moment we open our eyes, vision plays an important role in how we access information around us and how we interact with the world. Audio-visual experiences are everywhere around us and we rely on visual cues to help us interpret, filter and understand data quickly. Additionally, vision is tightly related to our communication capabilities, where non-verbal cues such as body language and facial expressions are key elements which allow us to transmit emotions and intentions that spoken words might be able to express.

According to the Pan America Health Organization (PAHO) and the World Health Organisation (WHO), globally there are at least 2.2 people who are having a type of vision impairment or blindness (PAHO, 2024, WHO, 2024). The UK National Health Service (NHS) estimated that in the UK at least 2 million people are leaving with sight loss, from which approximately 340.00 are registered blind or partially blind.

For blind or partially sighted persons, audio description(AD) represents the only method through which they can have access to information transmitted visually. Audio description is the auditory narration of the key visual elements presented any audio-visual representation: such as film, live performances and television/streaming programmes.

Nevertheless, although there are guidelines and legislation which stablish that a certain percentage of a broadcaster's content must be audio described, the choice of what is audio described is in the hands of the broadcaster and not the audience. The creation of an

automated audio description system would shift the power balance in the hands of the audience by offering the choice of what content they can enjoy with an audio description track.

According to Snyder “A picture is worth a thousand words? Maybe. But the audio describer might say that a few well-chosen words conjure vivid and lasting images (Snyder, 2005).”

2. Justification of the study

In 2019, when I first started the research that led to the creation of this paper, the idea of creating an automated system for creating audio descriptions seemed almost impossible. The main reason was the multimodal nature of the audio description creation process. Unlike translation, where we use the same medium (text), audio description is a cross-modal process which aims to transfer information from the visual (the video) into the acoustic (the audio description track) through a written intermediate (the audio description script).

Out of the three major steps of the process of creating an audio description track, the generation of the audio description script, which is the cornerstone of the audio description process, was the one deemed impossible to automate. One significant barrier is the sheer difference between the amount of information that can be transmitted visually and auditorily in the same amount of time. Not only would the system have to identify the right visual information, but it would also have to be able to condense it in such a way that it would fit the gap where the audio description could be inserted.

While it was feasible to conceptualize the functionality of such an automated audio description system, it became evident that the resources that I had available to develop a proof of concept were insufficient. As a result, the focus of the work was redirected to the creation of an evaluation framework. This framework was designed to be used as a tool that could be adapted

to assess various algorithms that could potentially automate specific steps in the audio description process.

Recent advancements in the field of Large Language Models (LLMs) and multi-modal solutions, exemplified by TwelveLabs and ChatGPT Vision will play an important part in the process of creating a system which can automatically generate audio description scripts. While the main goal of my research pivoted towards the framework creation, with the new available technologies, I was able to create a proof-of-concept end-to-end automated system which can take in a video and the then generate an Audio Description Script which can then be voiced using synthetic voices.

3. Research Aims Questions

The key question around which this research project revolves around is automation of the audio description process:

Initial research aim: Is it possible to automate the audio description process, especially the audio description script creation process?

Refocussed research aim: Is it possible to create an algorithm evaluation framework which could be used to evaluate various algorithms that could be used to automate the audio description process?

Can we envision what would be needed to evaluate such algorithms (which at the start of this research) that could describe the key visual elements in a video?

4. My research contributions

This research project attempted to create an automated audio description system during a time when extracting visual information was considered one of the most difficult tasks in

computer science. In my research, I attempted to envision an ideal automated audio description system which would use existing technologies (such as Active Speech Recognition, Voice Synthesis etc.) but also technologies, which at the time were not widely available to the public such as Vision Languages Models or Large Language Models. While at that time the lack of access to these technologies meant that I could not continue on the path of automating the audio description process, I was able to pivot towards envisioning an evaluation process/framework for when such technologies would be available and could be integrated in my proposed automated audio description system.

5. Thesis Organisation

The research thesis is divided into four chapters (excluding the introduction and conclusion) which reflect my research journey across the past several years:

Introduction: This chapter offers a short description of this research project.

Audio Description Analysis: In this chapter we look at audio description with the goal of understanding its various aspects such as: audio description as a process which aims to create an audio description track, audio description as the output and the ways it interacts with audio-visual media; and probably the most interesting aspect, audio description as an interface. Additionally, we dive into the audio description workflow and the flow of information during the audio description process. Furthermore, we will also attempt to map the human audio description workflow to an envisioned automated audio description workflow, with a focus on the creation of the audio description script.

Generating the Description Text: In this chapter we will focus on the process of creating the audio description script. We will look at how images/scenes and their description are related, with a focus on identifying types of algorithms which could be used to generate the audio description script. This chapter is split into two major subchapters, which focus on

the research that I have done on automatically generating image descriptions: the first subchapter will focus on the period before Large Language Models/Large Vision Models became available to the public, and the second subchapter will focus on the period since 2021-2022 when these LLMs/LVMs have been slowly being made available to the public.

Algorithm Evaluation Framework: In this chapter we will look at the process of creating the algorithm evaluation framework. The first subchapter describes the proposed creation process of the evaluation framework which will guide the rest of the chapter. We will look at the audio description script creation process and then move towards identifying and analysing guidelines and legislation pertaining to audio description: Audio description guidelines from several English-speaking countries (UK, USA, Ireland, Canada), the Information technology — User interface component accessibility Part 21 : Guidance on audio descriptions (ISO, 2015), and the ADLAB audio description guidelines(ADLAB, 2014). The goal was to extract and analysing the parts relevant to how a description is created and what visual elements are considered important when describing a scene. The goal was to start with the information provided in the ISO and then use the other documents to enhance it. Lastly, we will look at the proposed Evaluation Framework together with its structure and contents and how it can be adapted based on the needs of the entity who is doing the evaluation of the automatically generated audio description script.

Audio Description Automation Proof of Concept: While initially creating such a system was not possible with existing tools that I had access to at that time, with the recent developments in the field of LLMs, LVMs and multimodal models, I was able to create a proof of concept system based on my initial vision of an automated audio description system. In this chapter I leverage the newly available technologies and my professional knowledge in the field of automation acquired at my workplace to create a pipeline that takes in a video and

generates a draft audio description script. This subchapter is divided into the code explanation and the actual code of each script and the way in which they work together as a system.

Conclusion and Further Work: In this chapter, I will present my results and some of the further work that, I believe, would be interesting to be done.

2. Audio Description Analysis

Audio description is a type of verbal commentary which provides visual information for those who for various reasons, are unable to perceive it themselves. Audio Description (AD) offers blind or partially sighted persons the chance to access audio-visual media; or as Joel Snyder describes it: “the visual made verbal” (Snyder, 2005) Furthermore, the same audio description process is used in live settings such as but, not limited to, art galleries, museums, and sport events.

Audio description can be approached from two directions: as a practical process of creating an audio description track, and as an academic research discipline. The latter’s existence can be situated in the field of translation studies due to Jakobson’s (1959) distinction between interlingual, intralingual and intersemiotic translation. Multimodal or intersemiotic translation is a type of translation process in which context is constructed partially from information outside of the translated channel. Furthermore, in the field of Audio-visual Translation (AVT), AD is one of the two forms of AVT together with subtitling for the deaf and hard hearing (SDH) which systematically require inter-modal information transfers. The remaining AVT forms namely, subtitling, dubbing and voicers are focused with transferring/translating information from one language to another but in the same mode, the auditory, or from audio to written form. .(ADLAB, 2014; ISO, 2015; Fryer, 2016)

Since audio-visual content is highly complex from an information standpoint, audio and visual elements contribute simultaneously to the creation of context. The ‘modal’ aspect of this type of translation process can also be considered as related to sensory modes. In the case of AD, the visual information which would be decoded through sight is being converted into audio information that can be received through hearing.

The end result of the audio description process is represented by the audio description track, which contains a “translation” of the visual/audio elements, which is meant to create an accessible experience for blind or partially sighted persons.

From a more pragmatic approach, audio description can be defined as the process through which a trained person (audio describer) creates an audio description track. The production of an audio description track is split into three main steps:

- creating the audio description script

- voicing/recording the audio description script
- mixing the audio description track

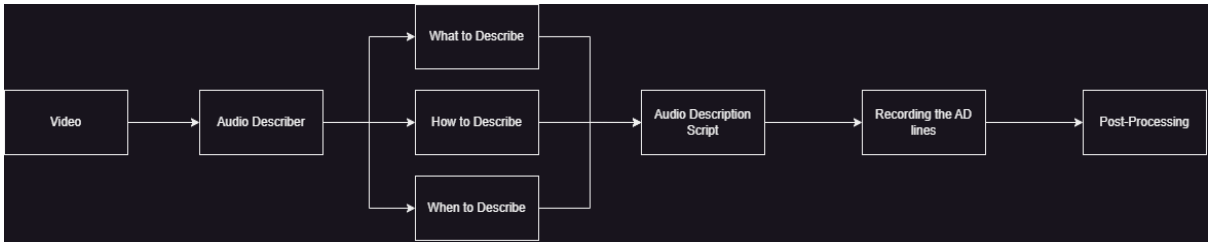


Figure 1 Audio description process diagram

1. The Role Audio Description

As previously mentioned, AD is a process through which information is being translated or transferred from a source medium to a target medium. In the case of textual translation, we can talk about Source Texts (ST) and Target Texts (TT), in the case of AD we have Source Material (SM) and Target Material (SM). Furthermore, if in the case of textual translation, the information is transferred from the ST to the TT, in the case of the AD process, the information is first translated into a written document, the Audio Description Script (ADS), which will then be voiced in order to produce the final Audio Description Track (ADT) (Orero, 2004).

For sighted persons, the information that is received when watching an audio-visual piece of content comes from a limited number of senses. While there are specialised environments which allow audiences to directly perceive information through more than just the visual and the auditory (e.g., 4D movie theatres which can add movement and other effects during movie experience), audio-visual content is generally low-immersive. As such, people who have no sensory impairments, perceive the information transmitted through the visual and the auditory channels. Since the two streams of information complement each other and create a 'valid co-occurrence' (Morgado, et al., 2020). As a result, if an actor performs an action which is accompanied by a complementary sound, the audience will fill in the remaining sensorial information in order to create the sensation of 'being there' (Biocca, 1997).

Additionally, according to Casile (2011), mirror neurons which are implicated in the neurocognitive functions (such as: social cognition, language) are responsible for 'mirroring' the actions and behaviours of others. Thus, when the audience receives complementary

information through multiple sensory channels, and 'valid co-occurrence' takes place, they will feel like they are the ones who are doing the actions the actors are performing.

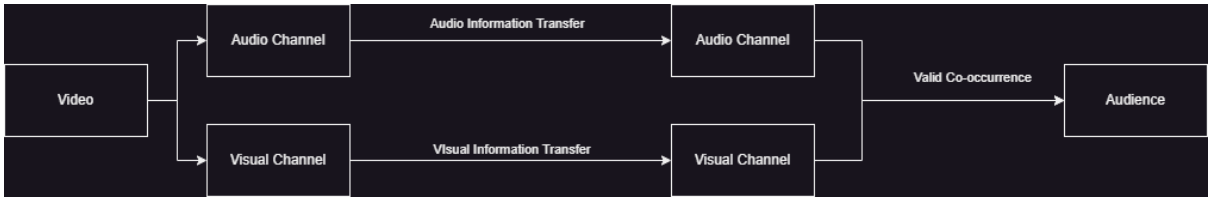


Figure 2 Cross-channel Valid Co-occurrence Diagram

In the case of AD for the blind or partially sighted (BPS), there is no visual channel to be used together with the audio channel in order to create this valid co-occurrence. Instead, the co-occurrence is being built in the same channel (the auditory channel) with the help of the original audio track and the AD track. The visual information is processed by the audio describer, who creates a new stream of information in the audio channel. This audio stream contains the audio description track which works together with the original audio track in order to create a valid co-occurrence event in the audio channel.

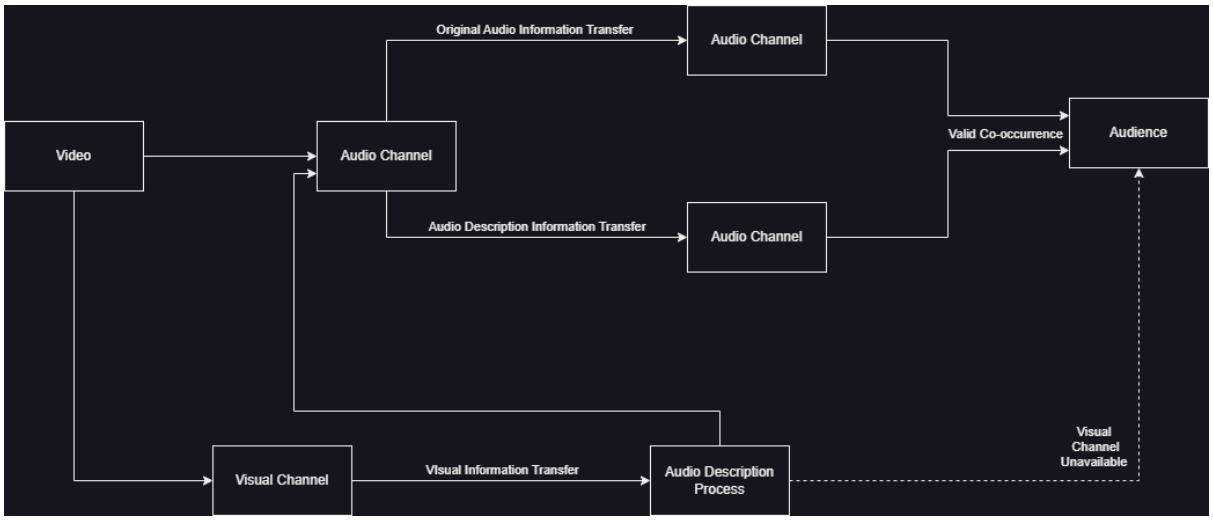


Figure 3 Intra-channel Valid Co-occurrence Diagram

This is why having an AD track that complements the existing audio track is paramount for offering BPS audiences an enjoyable experience. Should the AD track and the existing audio track not complement each other, the audience would have even greater difficulties at understanding/enjoying the content. (Fryer, 2016)

Thus, in order to create an AD track which comes to enrich the existing audio track, the audio describer has to first create an Audio Description Script (ADS). This represents the first and

the key step of the AD process. This document contains all the textual lines which are to be voiced, together with time markers which are meant to indicate where in the timeline of the audio-visual content it should be inserted. The location of the AD insertion has to be carefully chosen in order to complement the existing audio track, ideally it should be inserted in locations where there is no pre-existing dialogue or loud sounds. Furthermore, this timing element acts as a constraint on the amount of information and words that can be used to transform visual information into audio information. (ADLAB, 2014)

Consequently, the audio describer has to carefully select which of the visual elements and non-identifiable sounds/sound effects are paramount for the described scene. Multimedia content producers make use of audio elements (such as: dialogue, monologue or narration, sound effect, songs, background music) in order to enrich the visual elements and thus transmit their creative narrative intent to the viewer. Narrative is usually defined as involving chains of events in cause-effect relationships, occurring in space and time. Bordwell and Thompson (1997: 90-96), discussing narrative in films, append to this definition a statement about how the agents of cause and effect are characters with goals, beliefs and emotions. Thus, the audio describer has to be able to synthesize all these elements and transform them into an AD that can reproduce as much as possible of the original narrative intent (visual + audio) but only using the auditory channel.

Everything we have discussed until now points to the fact that AD texts could also be considered as specialized texts due to the nature of their purpose. Lehmann (1998) indicates that AD texts are neither spoken nor written but they stand somewhere between these two categories. This can further be backed up by a diagram describing the process of Audio Description, where the information passes through three mediums: first we have the visual and auditory (multimedia content), the written medium (Audio Description Script) and lastly the auditory (Audio Description Track).

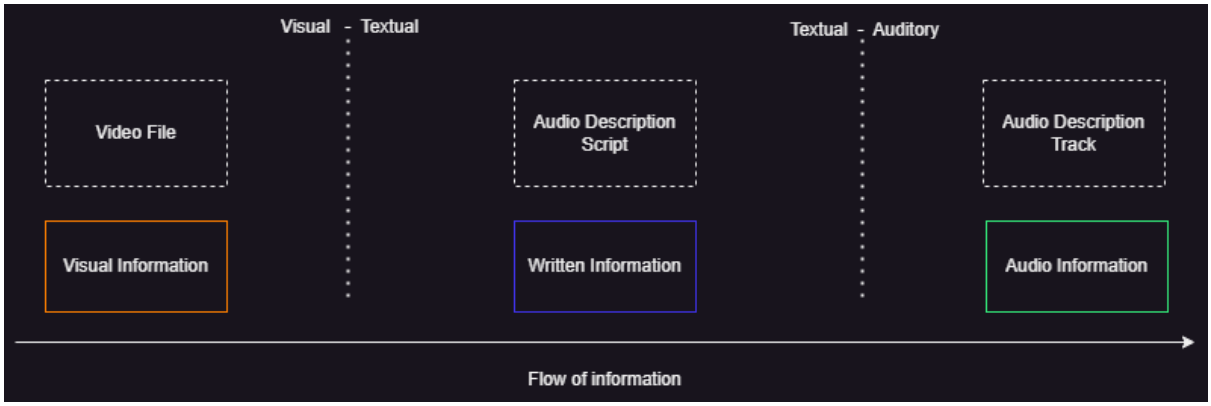


Figure 4 Audio Description Flow of Information

As a result, the AD texts retain characteristics of both written and spoken texts; one hand they display the communicative/conversational aspect of spoken texts, while making use of the structured nature of written texts. The latter being a direct result of the text length constraints imposed by the timing of the AD insertion points in the original audio track.

2. Audio Description as an Interface with the Visual

Unlike some other ISO standards which are standalone (e.g., Translation ISO 17100:2015), this ISO standard that deals with audio description, was published in the year 2015, ISO/IEC TS 20071-21:2015 or Information technology — User interface component accessibility; Part 21: Guidance on audio descriptions has been included as part of user interface component accessibility. While the discipline of audio description exists as a standalone entity, it finds itself at the confluence of two fields: on one hand,the field of multimedia translation, , on the other, we have the field of accessibility.

Additionally, if we take a broader view at what audio description can be, we can observe that there is another function that AD can have. Beyond being a process or a result of a process, at its core, AD acts as a cross-medium information channel which permits information to be transferred from the visual modality into the auditory. The ISO states that it "[…] provides audio description developers and practitioners with guidance in creating effective content describing audio-visual material in an auditory-only modality […]" (ISO/IEC, 2015, p. vi). As we can notice this excerpt indicates that audio description, broadly speaking, is employed in any situation which requires information to be described in an auditory-only modality. This means that the circumstances in which audio description can be used are highly diverse " The circumstances to which audio description applies include recorded video, broadcast and broadband television, cinema, live or recorded drama, museum and art gallery exhibits,

heritage tours, news, and comedies"(ISO/IEC, 2015, p. vi). Consequently, in its broad definition, audio description in itself facilitates information exchange between the transmitter (visual content) and the receiver (blind or partially sighted persons, or anyone listening to AD).

If something facilitates the information exchange between two components, it would appear that it serves as an intermediary between those components almost like an interface. In his 2014 book, Branden Hookway describes an interface as a shared boundary across two or more separate components of a computer through which an information exchange takes place (Hookway, 2014). He further goes to explain that this information exchange can happen between software, hardware, peripheral devices, humans and any combinations of the mentioned components (Hookway, 2014).

In the case of audio description, based on information provided by the ISO and the definition provided by Hookway, it can be concluded that audio description (as a product, not as a process) can be defined as an interface for the following reasons:

- In the case of persons who are able to see, the information acquisition from a video or through visual means occurs without the need of intermediaries.

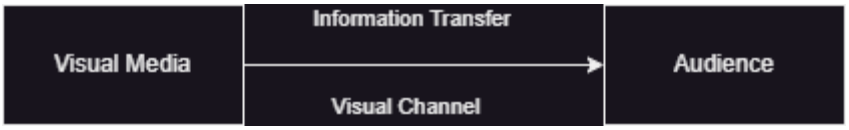


Figure 6 Visual information acquisition

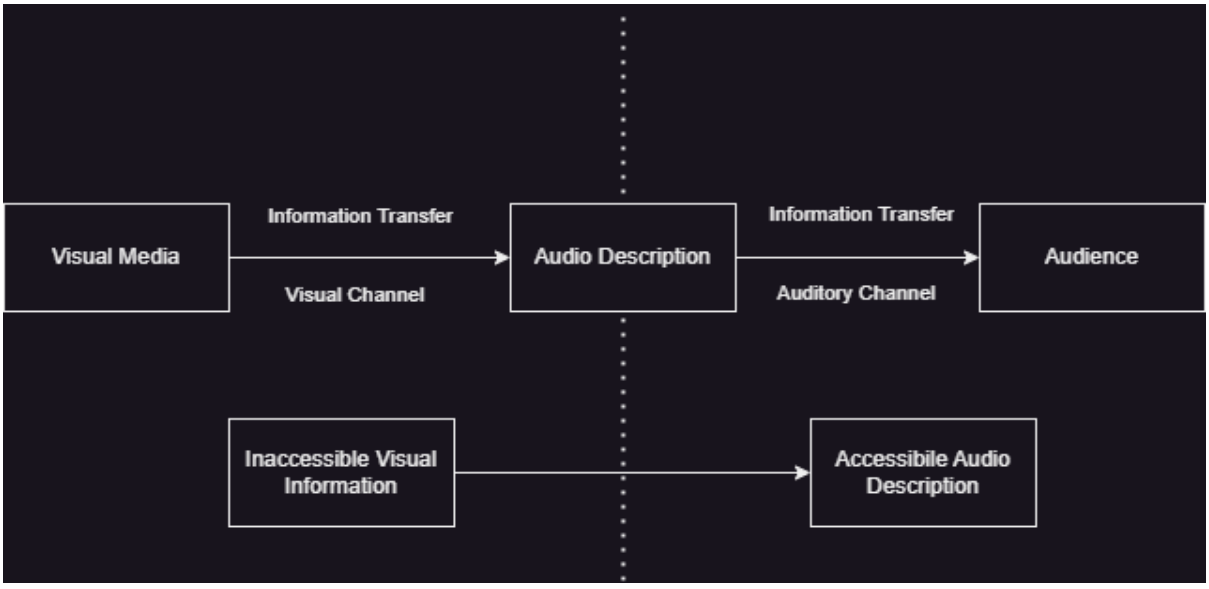


Figure 5 Audio Description Information Transfer

- However, in the case of blind or partially sighted persons, there is a disconnect between the information transmitter (in our case the video/ visual content) and the

receiver (the eye). In order to overcome this disconnection; a third component is introduced – the audio description track.

Since the audio description track carries the information across two mediums, from the visual to the auditory – and taking in consideration Hookway's definition of an interface, it can be concluded that an audio description track can be categorized as an interface.

3. Possible types of Automated Audio Description Systems

After analysing the human audio description process, we can identify areas suitable for automation, aiming to develop a fully automated end-to-end system that takes a video as input and produces an audio description track as output. Audio description, being a form of audio-visual translation, we can draw a parallel with the human translation process and existing systems/tools. For instance, we have fully automated translation systems that can operate without human input. Then we have human-in-the-loop translation systems, like LILT, which combine AI with human expertise. Finally, we have computer-assisted translation (CAT) tools, which facilitate the human translation process by offering access to translation memories, terminology management, alignment tools, and other productivity-enhancing features.

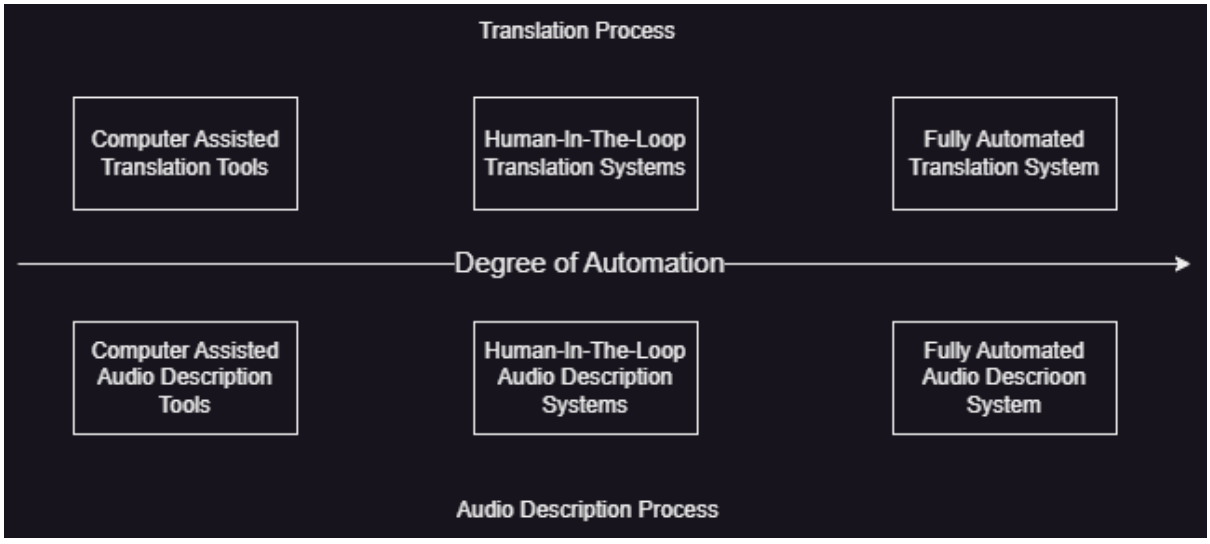


Figure 7 AD & MT Automation Comparison

As a result, we can envisage the same categories of system when discussing the audio description process: computer assisted audio description tools, which would be able to increase the productivity of the human audio description process. A human-in-the-loop audio description system, which would leverage AI solutions to automate most of the process, but

would still need the input of a human or several humans. Lastly, a fully automated end-to-end system which would not require the input of a human.

In order to determine how such audio description systems would function, we need to identify where in the human audio description process we introduce automation and to what degree.

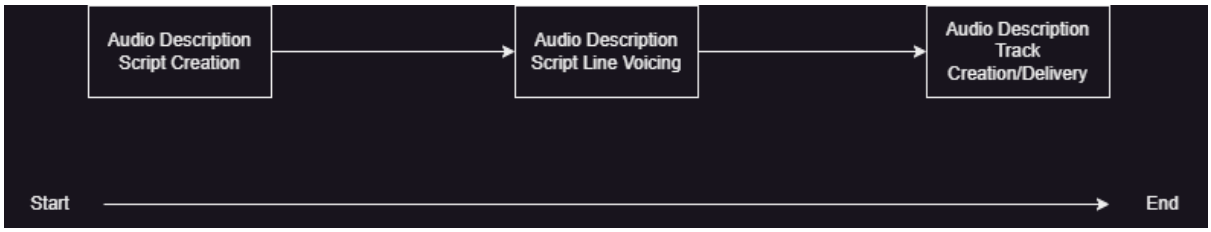


Figure 8 Human AD Process

For this purpose, we will divide the human audio description process into three broad steps: audio description script creation, audio description script line voicing, and audio description track creation/delivery. We will then examine at the series of actions that have to be performed in each step and how/what we could automate with the goal of obtaining a decision matrix which would help us with categorising the type of system we are dealing with.

1. Voicing the Audio Description

Once the audio description script has been created, the audio describer can start recording the script lines using specialised software which allows the recording of audio description. The audio description script is saved in a special type of file format which allows the software to associate each title box with an audio recording (.wav file). There are various types of audio description formats such ESF, TTAL or SRTAD. For our example we will use the SRTAD by YellaUmbrella format as this is an open file format which is human readable and allows us to better understand how it functions.

Since it's an extension of the original SubRip (SRT) format, it builds upon its structure and adds the required metadata needed to transform an format that was originally made for subtitles, into a format which works with audio description.

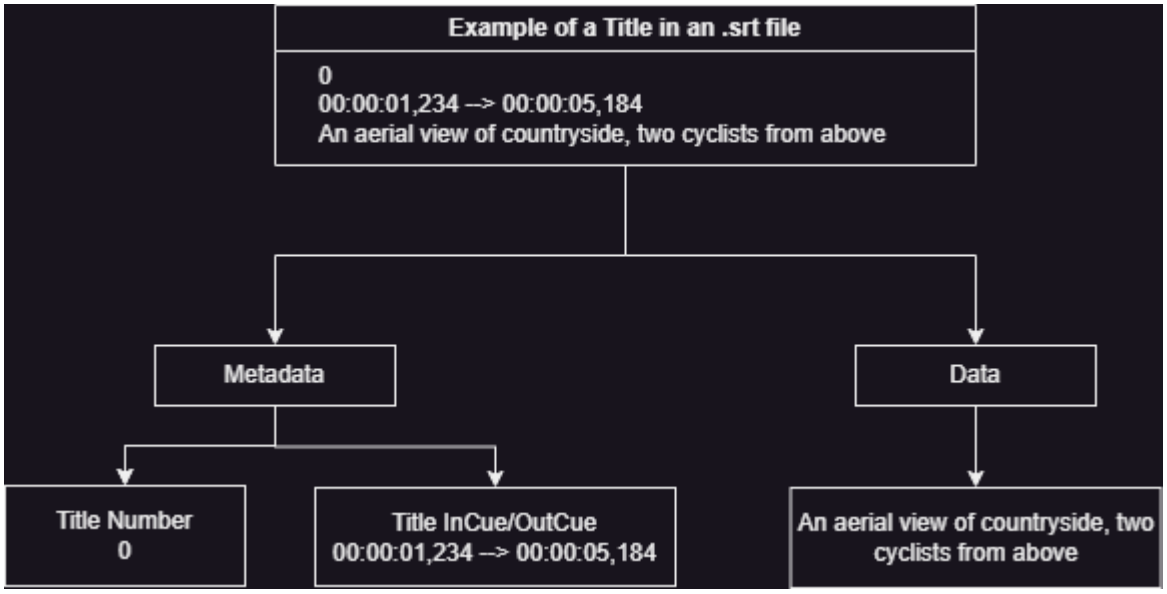


Figure 9 SRT File Format Structure

Each subrip title has three key elements:

title number : which allows the system to order the titles

timing : defined by the incue and outcue, tells the system when to show the title on the screen.

content: the actual subtitle text that will be shown on the screen

This combination of content and metadata will be represented in the timeline of a subtitling software, in our case we will look at the file using Stellar from YellaUmbrella. We can envision how the metadata is rendered in the software and how it influences the position and onscreen duration of the title.

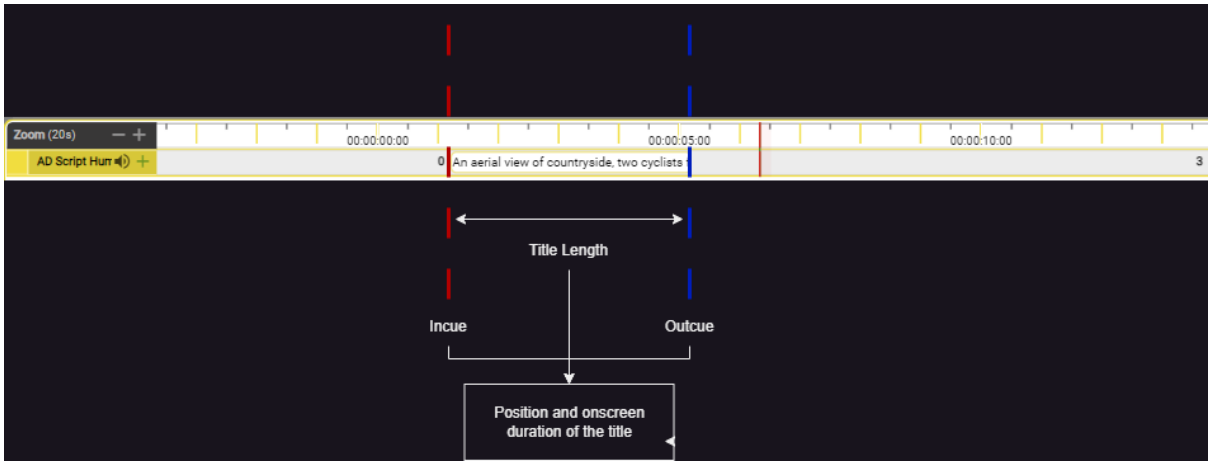


Figure 10 SRT Visual Representation

In the case of the SRTAD file, we need to add more metadata in order to make it compatible with the requirements needed to store audio description. As a result we end up with a superset of the SRT format which contains more metadata. Since we are also attaching/referencing a .wav file for each title, we will also have metadata which describes how the audio should be interpreted by the audio description system.

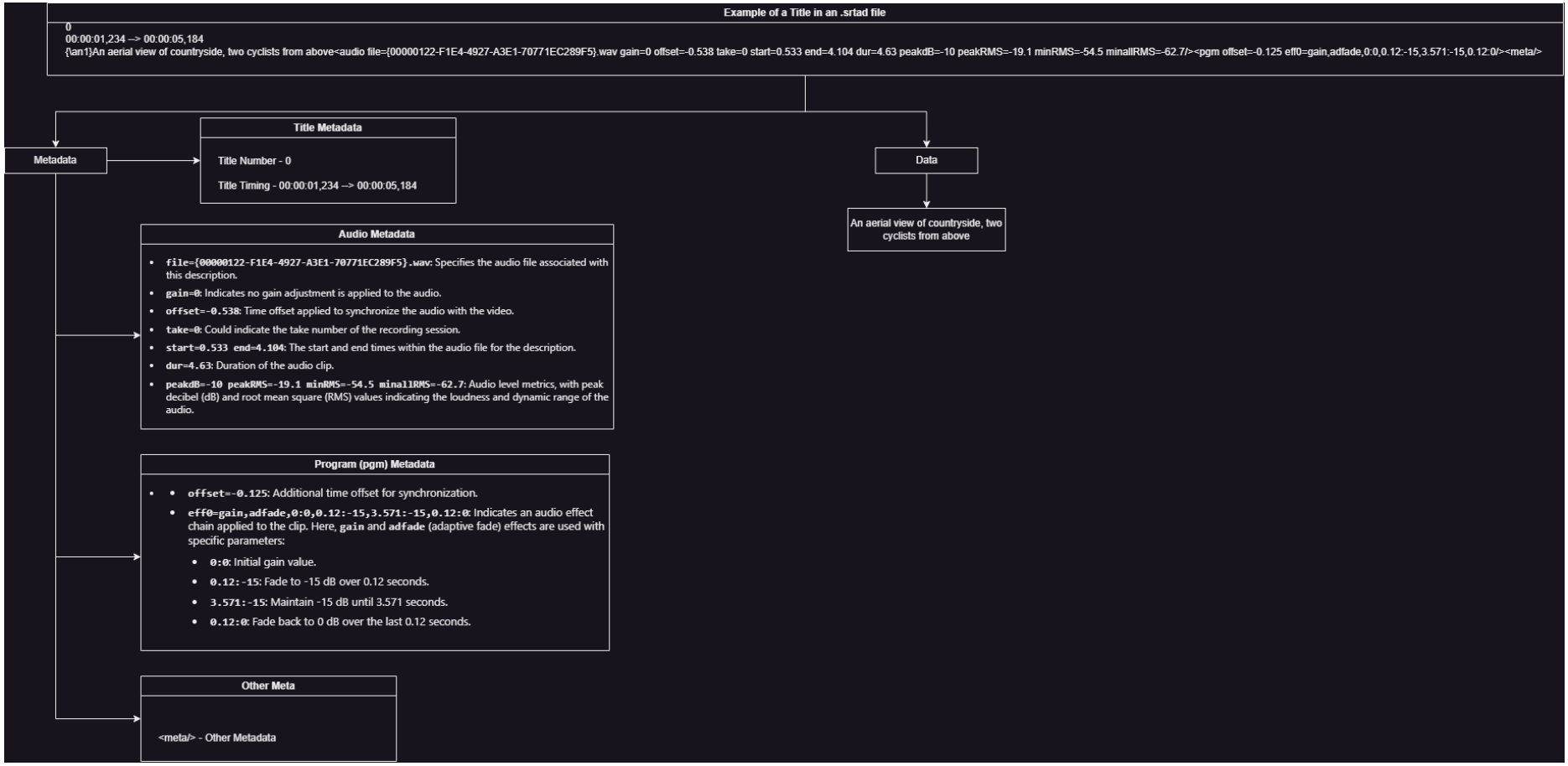


Figure 11 SRTAD Structure Example

The extra metadata present in the SRTAD file is extremely important as it both represents the link between the recorded audio file for each title and also, since we are only referencing the audio file, not actively modifying it, it controls how the audio is rendered by the system. Each time a title is recorded, the software will save the recorded WAV file and the metadata associated with it, and insert them into the corresponding title's field.

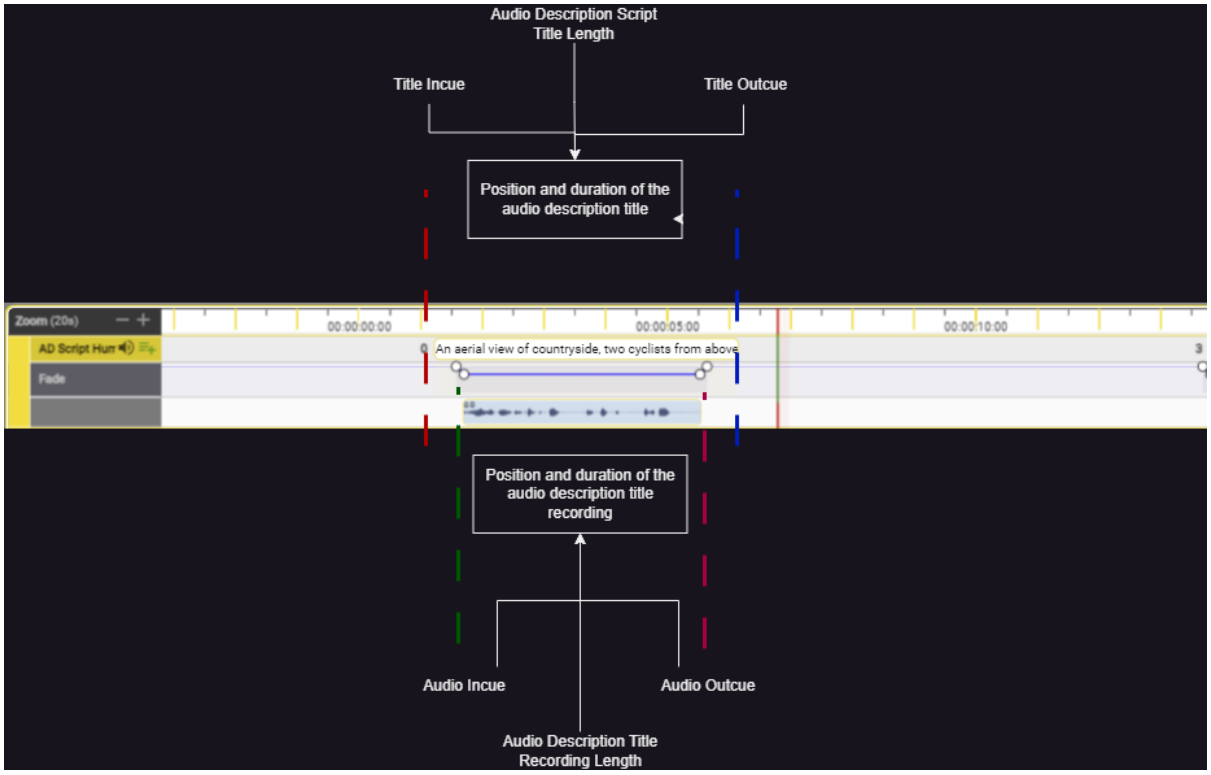


Figure 12 SRTAD Visual Representation

When the audio description title is inserted, it usually covers the entire area that it is described by its contents. For the purpose of our explanation we will consider that the recorded audio should be inserted in the area between the incue and outcue of the audio description script title. Nevertheless, the audio recording can be repositioned and adjusted as long as it is not placed outside the title incue or outcue. Unlike in the case of the subtitle file where we only work with the title entry, in the case of an audio description script file, we are working with the audio description title, but also with the audio recording. As a result we will have an additional incue/outcue pair which describes the position and length of the recording in the original media.

We can automate the rendering of each title by using any of the countless Speech Synthesis service providers such as Google, Amazon, Cereproc or Elevenlabs. These services take the

text of the title together with metadata such as the length of the title and generate the ‘spoken’ audio.

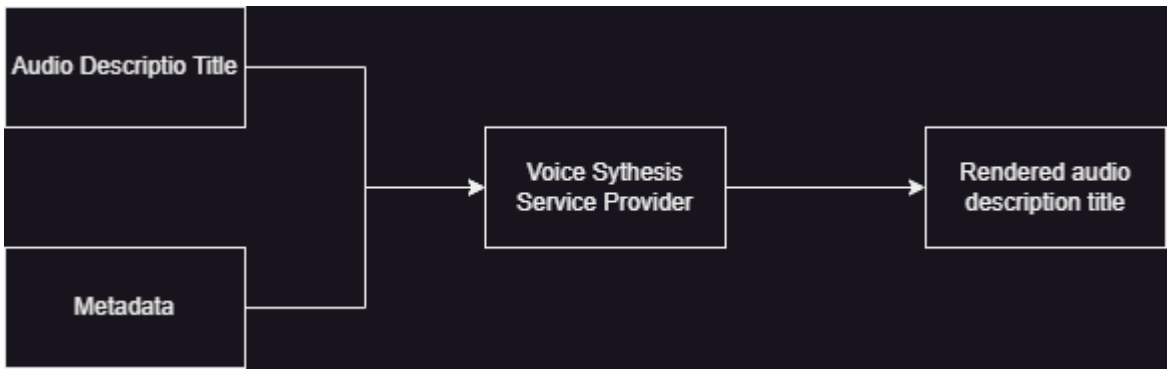


Figure 13 Voicing AD Title using Speech Synthesis

The machine-‘spoken’ title is then stored as a WAV file which is associated with the title from which it was created.

As a result, a finished audio description script which has had all the titles recorded will contain pairs of titles (which contain the description text) and audio recordings which represent the ‘spoken’ version of the description text.

2. Mixing the Audio Description Track

Since the resulting output of the audio description process is a separate audio description track, it needs to be mixed into the original media in order to be delivered. Mixing is a complex process which is performed by various audio/video editing pieces of software but also by specialised audio description software. Generally, this type of software uses FFmpeg, a powerful open-source multimedia framework, which facilitates these processes with its comprehensive suite of tools. To generate the mixes, the software calls FFmpeg using a series of parameters which describe the inputs and the desired output.

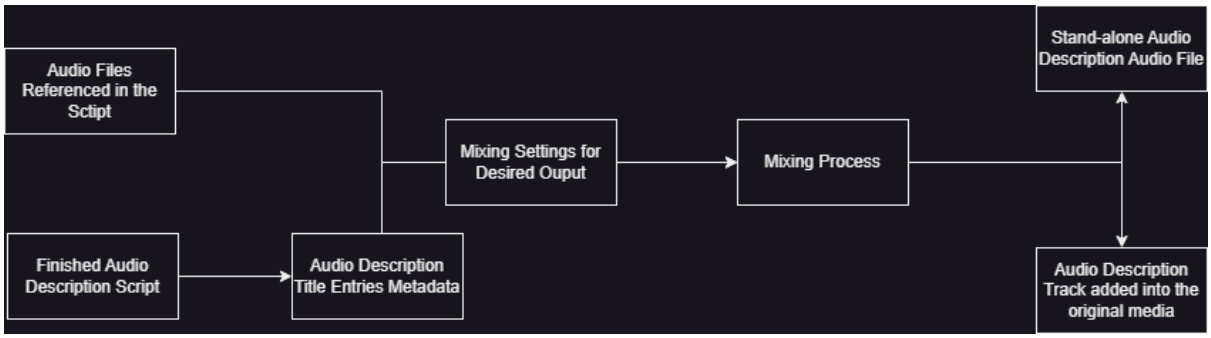


Figure 14 AD Mixing Workflow

The most common types of output obtained from the mixing process of an audio description track:

1. Creation of the separate audio description audio file: combining all the recorded titles into one continuous audio description track. This will generate a stand-alone audio file which can either contain only the audio description or it can contain a combination of the original audio and the audio description audio.

2. Adding the audio description track to the original media: the continuous audio track is mixed into the original media where it can be presented in various ways from having it replace the original audio to adding it as a separate audio track in a separate channel.

This step can be automated by creating profiles for each of the desired outputs of the mixing process. As long as all the inputs are present in a known location from where FFmpeg can access them, we can initiate the mixing process by creating an FFmpeg command which would reference the desired output profile and location of the input files and the location of the output result.

3. Generating the Audio Description Script

In the human audio description process, creating the script entails watching the media that is to be described, deciding where audio description can be inserted (insertion points), and then describe the scene by only conveying the most important key aspects

Figure 15 AD Script Creation Workflow

4. Analysing the Human Audio Description Workflow

In order to conceptualise an automated audio description system, we first need to understand each step of the audio description workflow. While, broadly speaking, we have four major steps in the workflow, as mentioned by Szarkowska (2011), when attempting to automate this process, our initial goal is to create an equivalent workflow that is fully automated. This automated workflow will only serve as a draft, a starting point that can be used as a base for creating an even more complex overview of the automated workflow.

When looking at each of the steps in the human AD workflow, some of the steps that Szarkowska (2011) mentioned, would have to be altered in order to better fit the automated environment. For example, instead of Material Selection and Material Viewing, we could have a unified step called Material Preparation. On the other hand, Recording and Mixing the AD Track will have to be split into Script Voicing and AD Track Mixing.

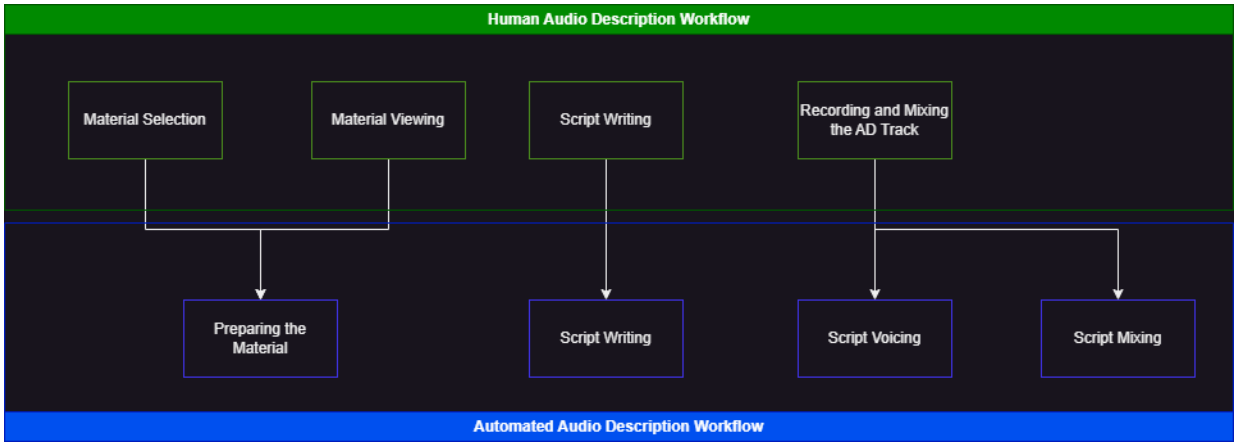


Figure 16 Human AD vs Envisioned Automated AD Workflow Comparison

1. *Flow of Information*

We can also approach the Audio Description Workflow from the point of view of the information highway, meaning that we can analyse the information and its path from the input material to the output audio description track. This will also entail the creation of a parallel process diagram that mirrors the Human Audio Description Workflow and provides more details by having an in-depth look at the transformative processes that the information undergoes. Additionally, it will offer a better understanding of the communication mediums and channels that are involved in transferring the visual information into the auditory.

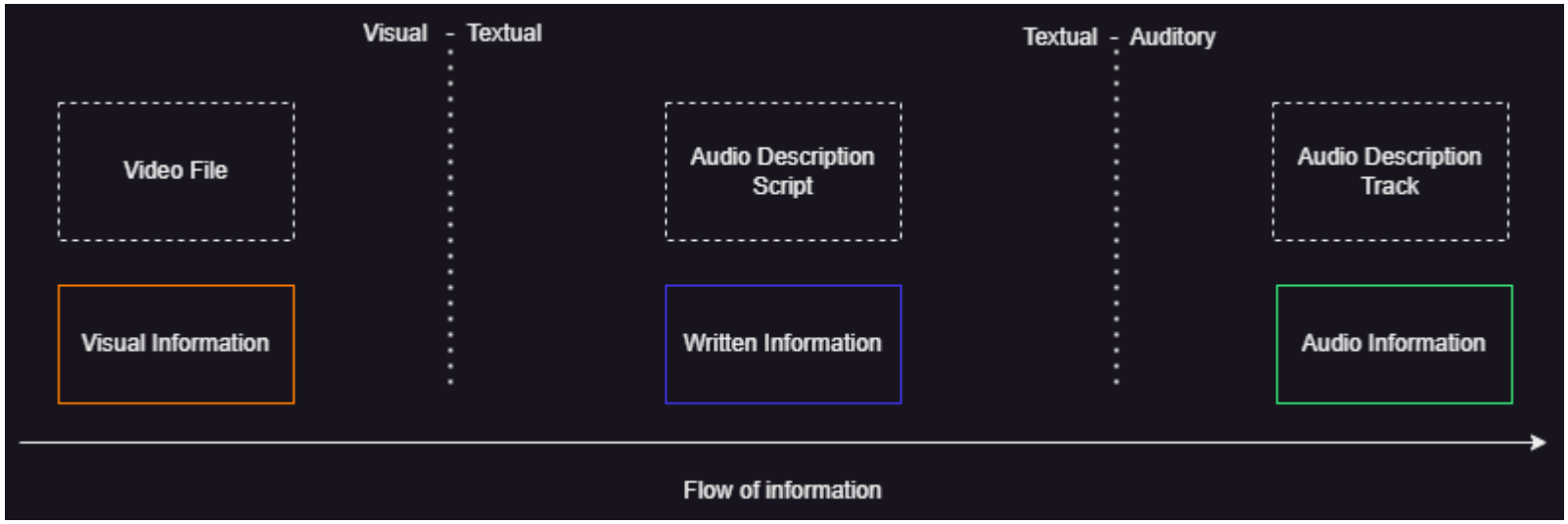


Figure 17 Audio Description Flow of Information

The goal of the audio description process is to generate an audio description track which acts as a complementary source of information for persons who do not have access to the visual channel. From the diagram we can observe that the audio description script acts as an interface between the video file and the audio description track. Thus, to get to our audio description track, we first have to analyse and curate the visual information through the means of the audio description script. Furthermore, the audio description script does not only contain the curated description, but also metadata which imposes constraints and boundaries on the generated audio description track. For example, the audio description script will also contain the timing of each audio description title, which will then influence the reading speed that the voice actor

would have to adhere to when recording the lines. This timing of the audio description title is based on the original audio track of the video file, which has been analysed in order to identify suitable areas in the media where audio description could be inserted.

2. Constraints Schema

Since we briefly touched on the constraints that are imposed during the audio description process, we can also create a diagram which focuses on how these constraints are created and where they originate. The multimedia file that we use as an input contains both audio and visual information which are being transmitted through their respective channels. While the audio description track is aimed at enhancing accessibility by describing visual information, due to its auditory nature, its creation process must take into account the existing original audio track. It is important that the audio description track does not disturb the complementary relation present in the original media between the visual and the auditory.

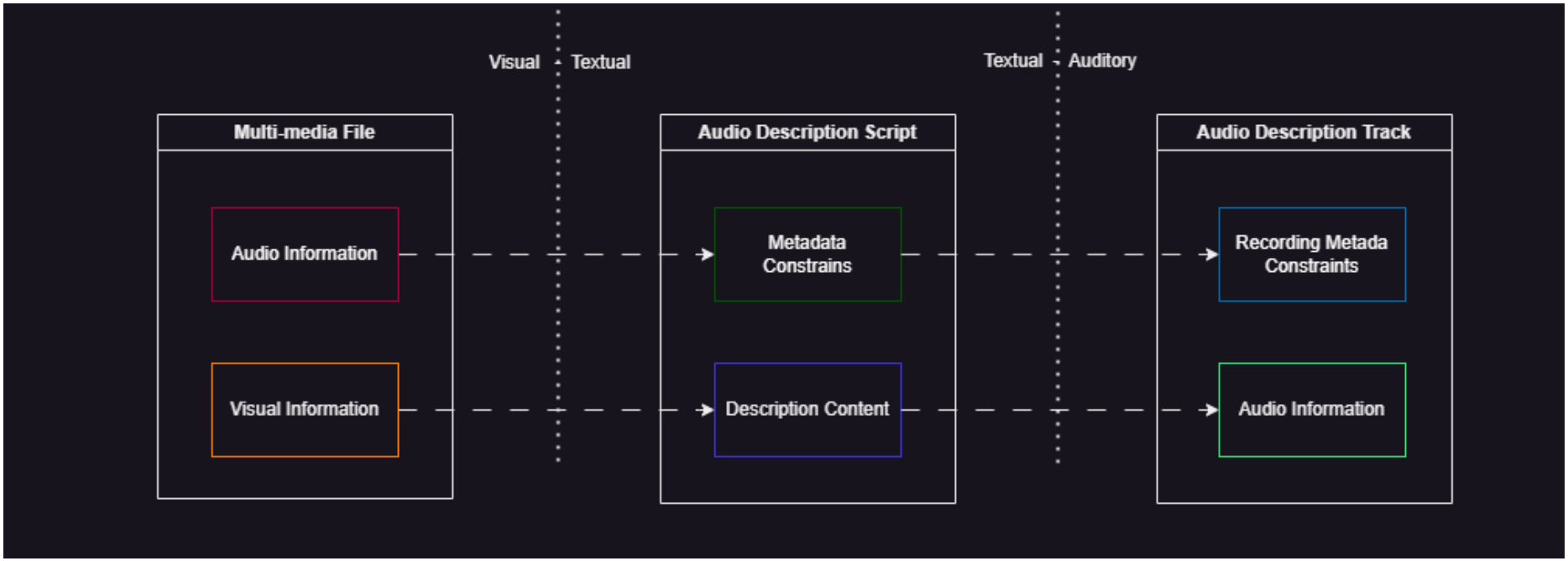


Figure 18 Audio Description Flow of Data and Metadata

To get a better understanding of how the original audio in the media imposes constraints on the creation of the audio description script and implicitly, on the resulting audio description track we can look at a simple piece of audio.

3. Audio Description Insertion Intervals

First we need to discuss about audio description insertion points or intervals. These points/intervals represent areas of the original media where an audio description title could be inserted, and the corresponding video section be described in the resulting audio description script title. The

original audio track contains several audio events which can be identified and used as guidelines for evaluating if a certain area of a video is a good candidate for adding an audio description title. These events can be split into three categories: dialog/speech, sound effects, and music. Additionally, the original audio can also be split into two subcategories, the voice track and the music & effects track.

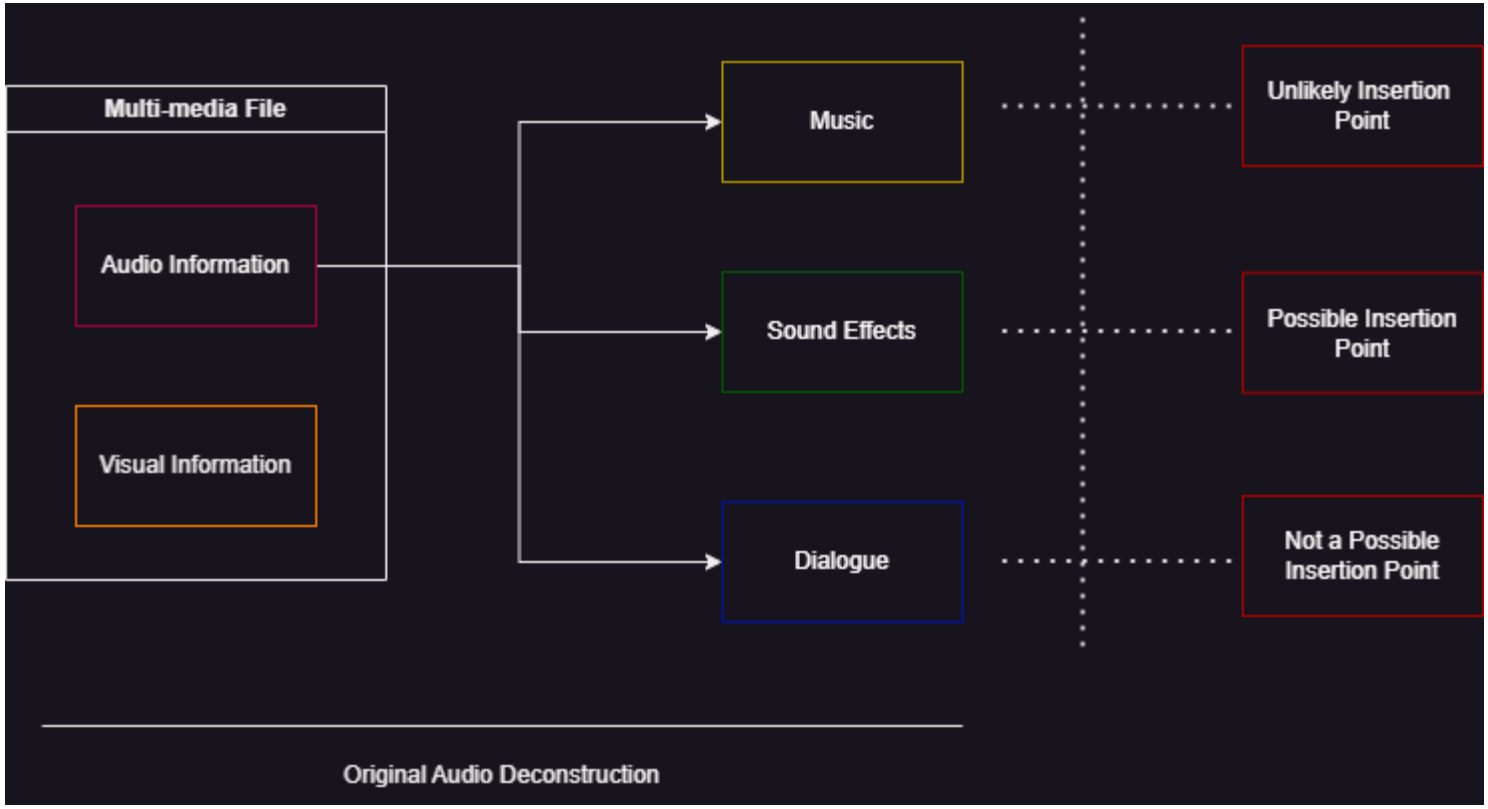


Figure 19 Audio Track Composition

In order to insert an audio description line at a certain point/interval:

- We first need to ensure that there is no existing dialog or speech. This is a hard stop when it comes to identifying possible insertion points/intervals. Since speech is already present, we cannot overlay another voice track on top of it.
- Next, we need to ensure that the existing sound effects permit us to insert an audio description line. The decision is usually made by the scriptwriter. Due to the nature of the audio-visual content, it might be that either the sound effect is too important, in which caseadding an audio description title would ruin the scene; or, if the sound effect is deemed not important, adding an audio description title would be acceptable.
- The same applies in the case of music.

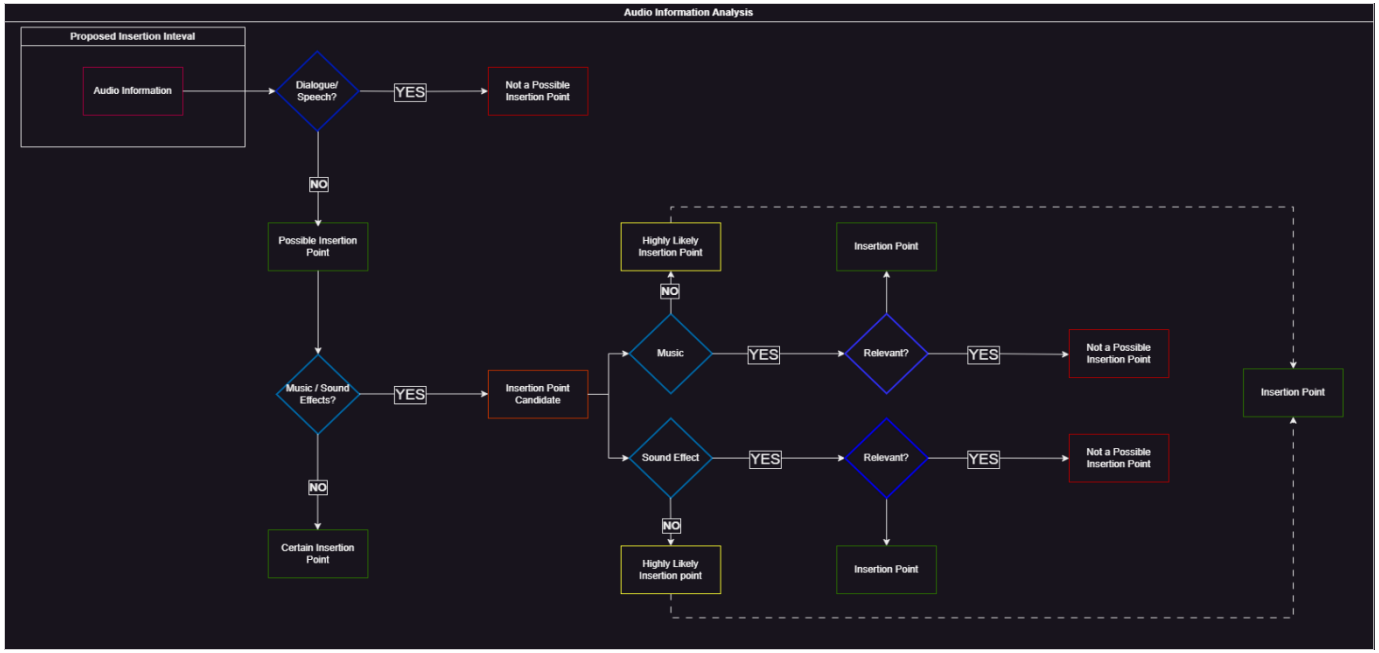


Figure 20 Audio Information Insertion Point Analysis Diagram

4. Insertion Point Decision Matrix

If we think about where exactly audio description could be inserted, most likely it will be a in point/interval which is “silent”. Nevertheless, if we take into consideration the audio events that we previously discussed, we can envision that there are different types of “silence”; each corresponding to the audio events, or rather, their lack of presence in the original track. As a result, we would end up with a classification of audio description insertion points based on the audio event characterising them. Additionally, we can systematise the identification of valid insertion points/intervals by looking at the audio events and their effects on the original audio track. The result can be easily represented by the following decision matrix:

Audio Event	Speech	Music	Sound Effects	Speech Silence	Music Silence	Sound Effects Silence
Speech	Not an Insertion Point	Not an Insertion Point	Not an Insertion Point	n/a	Not an Insertion Point	Not an Insertion Point
Music	Not an Insertion Point	Possible Insertion Point	Unlikely Insertion Point	Possible Insertion Point	n/a	Unlikely Insertion Point
Sound Effects	Not an Insertion Point	Unlikely Insertion Point	Possible Insertion Point	Possible Insertion Point	Unlikely Insertion Point	n/a
Speech Silence	n/a	Possible Insertion Point	Possible Insertion Point	Highly Likely Insertion Point	Highly Likely Insertion Point	Highly Likely Insertion Point
Music Silence	Not an Insertion Point	n/a	Possible Insertion Point	Highly Likely Insertion Point	Possible Insertion Point	Possible Insertion Point

Sound Effects Silence	Not an Insertion Point	Possible Insertion Point	n/a	Highly Likely Insertion Point	Possible Insertion Point	Possible Insertion Point

Table 1 Insertion Point Decision Matrix

We can observe that we end up with several types of insertion points/intervals based on the audio events that are present:

- Not an Insertion Point: Characterised by the existence of speech.
- Unlikely Insertion Points: Characterised by the existence of both music and sound effects but no speech.
- Possible Insertion Points: Characterised by the existence of at least one type of silence.
- Highly Likely Insertion Points: Characterised by the existence of speech silence and one other type of silence.
- Certain Insertion Points: Characterised by the complete absence of sound events.

To better illustrate this classification, we will use 65-seconds audio clip that contains all three types of audio events: speech, music, and effects. We used AudioShake, a service that allows us to split the original audio into stems. Each stem represents a track that contains one type of audio event.

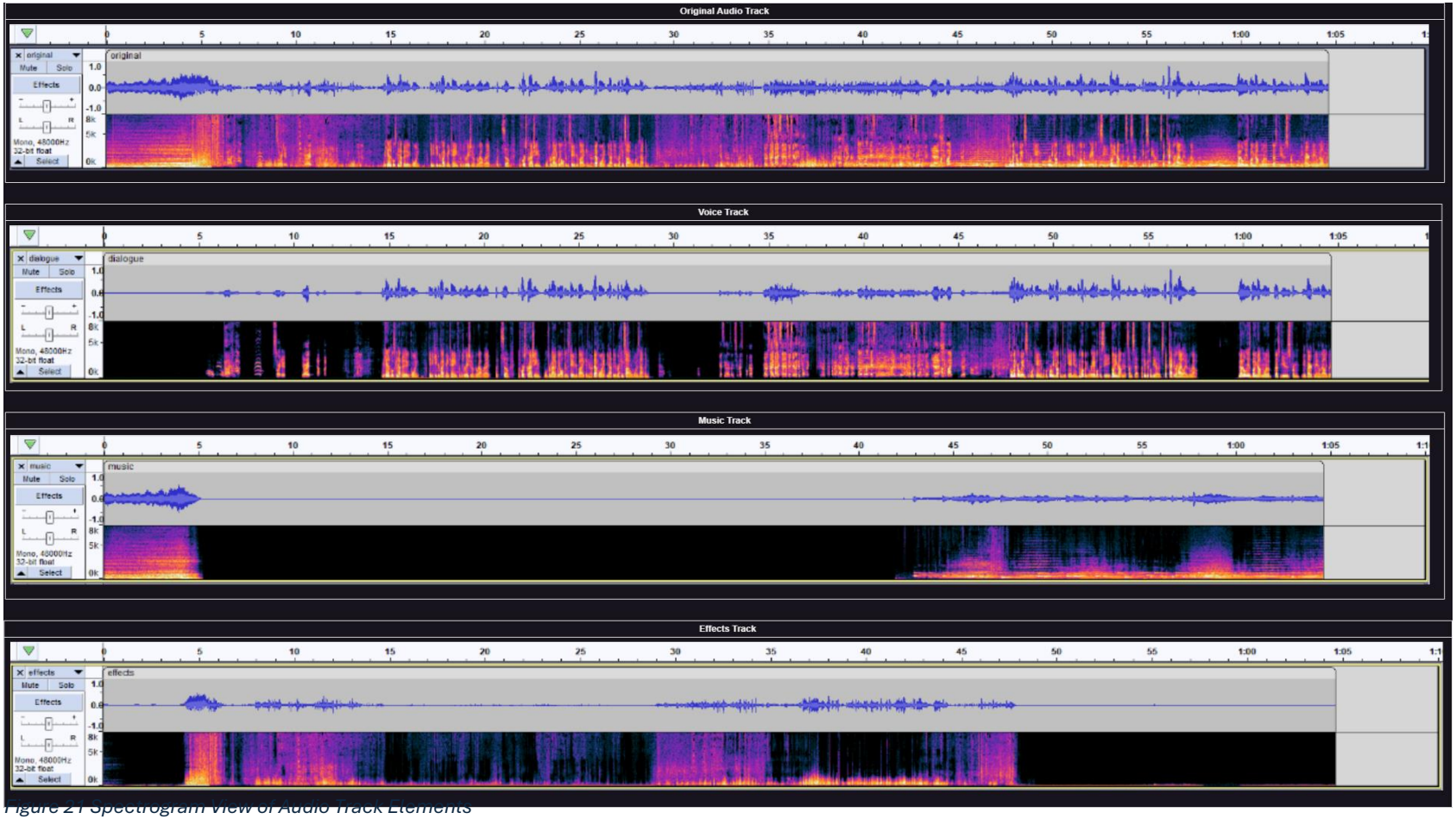


Figure 21 Spectrogram View of Audio Track Elements

The easiest way to understand the different audio events present in the original audio track is to visualise the sound using some type of audio editing software. In this case we used Audacity, which is a popular, free, open-source, cross-platform audio editing and recording software which can be used to accomplish a variety of audio tasks. Since our goal is to visually observe the structure of the original audio track, we can just import the original audio track into Audacity and then enable Multi-view from Track Control Panel. This is an advanced feature for expert users enabling Spectrogram and Waveform displays of the same audio shown simultaneously in the same track.

5. *Visualising the Insertion Points in the Audio Data*

Once we have imported the original track into Audacity and enabled Multi-view, the software will display the waveform and spectrogram that corresponds to the audio data in the original track. The waveform is a graphical representation of the audio signal's amplitude (volume) over time. The spectrogram is a visual representation of the spectrum of frequencies of the audio signal as they vary with time.

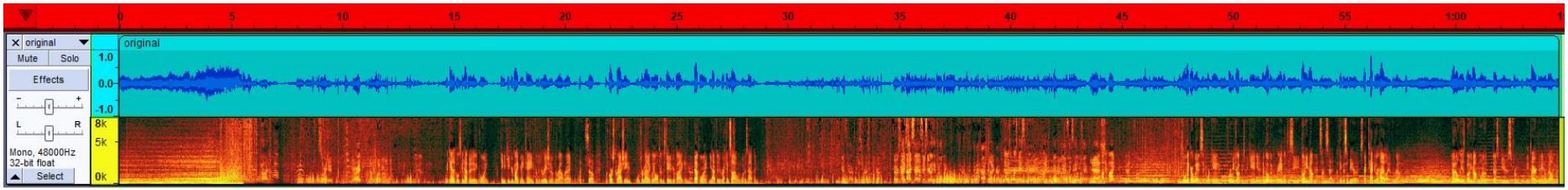


Figure 22 Due to using yellow to highlight the spectrogram area, the colours are shifted in this image. While yellow and red remained the same, the blue and purple were shifted towards green.

Horizontal Axis (Red Area) – Indicates the duration or the audio in seconds.

Waveform (Cyan Area) – A view with a linear vertical scale running from -1.0 (negative values) to +1.0 (positive values). It indicates the loudness of the audio signal. Peaks represent louder sounds, and troughs represent quieter sounds or silence.

Spectrogram (Yellow Area) - A visual indication of how the energy in different frequency bands changes over time. On the vertical axis we have the frequency range between 0Hz and 8kHz. The colour intensity represents the amplitude of the frequencies : brighter colours (yellow, red) indicate higher intensity, while darker colours (blue, purple) indicate lower intensity.

5. *Original Track*

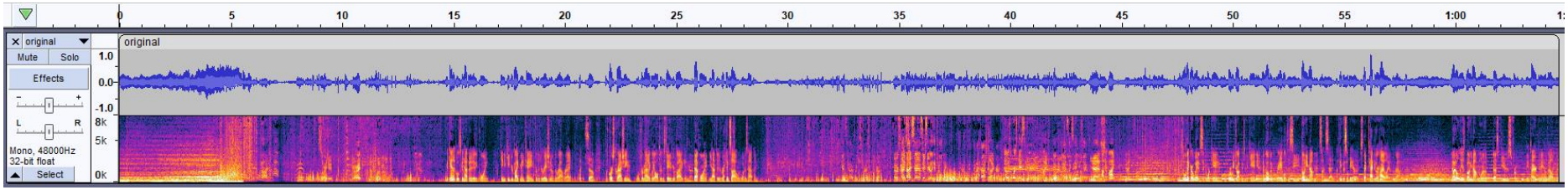


Figure 23 Original Track Combined View

The waveform presents a mix of louder and quieter sections throughout the track. This pattern of louder and quieter sections is indicative of the presence of a combination of speech, music and sound effects. Furthermore, we can observe that the amplitude varies and the peaks and troughs that suggest the dynamic range present in the audio.

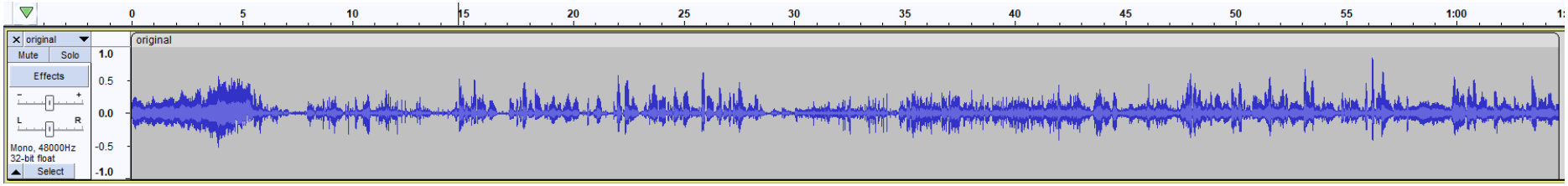


Figure 24 Original Track Waveform

The spectrogram presents a broad range of frequencies which can be identified by the bright bands at the bottom(low-frequencies bands) which signify music elements. The scattered bright spots in the middle and high frequencies areas are indicative of speech and effects. As in the waveform, we can notice that we are dealing with dynamic audio data indicated by the higher frequency sounds which are at the top, lower frequency sounds at the bottom, high-amplitude frequencies are brighter and low-amplitude frequencies are darker.

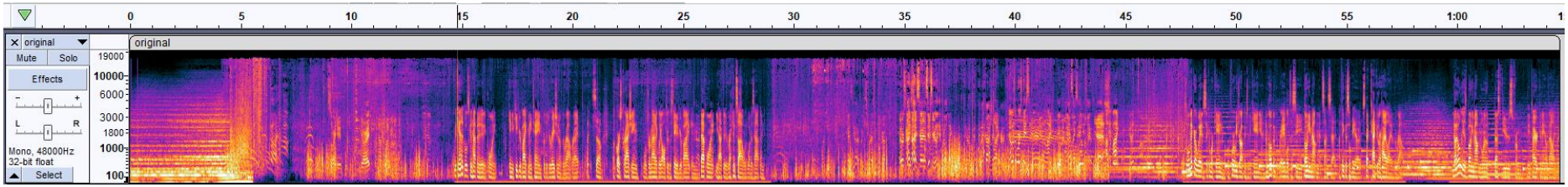


Figure 25 Original Track Spectrogram

1. Isolated Speech Track

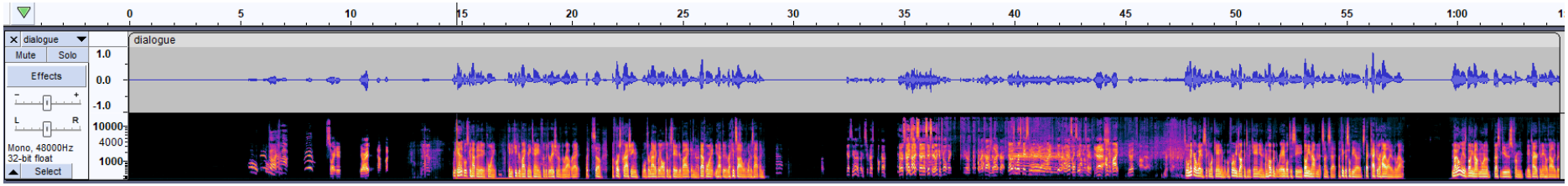


Figure 26 Isolated Speech Track Combined View

The waveform presents a mix of louder, quieter and completely silent sections throughout the track. We can easily observe the areas where the peaks and valleys are larger, indicating the louder sounds. The areas where the line is closer to 0.0-mark indicates periods of silence. Furthermore, since we are dealing with speech, we can observe rhythmic patterns with alternating loudness variations. It also provides a visualisation of the overall volume and timing of the speech.

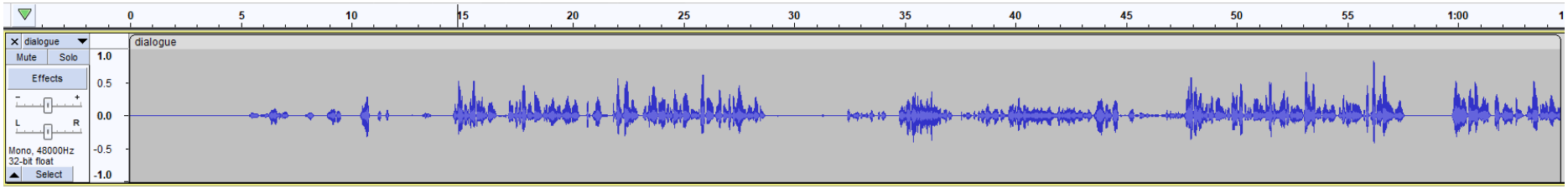


Figure 27 Isolated Speech Track Waveform

Unlike the spectrogram of the original audio, which shows dark areas of lower amplitude and bright areas of higher amplitude, we can see completely dark areas that indicate the absence of speech. We can also observe various speech features, such as: bands and clusters that represent

the fundamental frequencies and harmonics of speech. Formants, which are the bright, horizontal lines that are the resonant frequencies of speech, important for vowel sounds. We can even notice the difference between vowels and consonants, with unvoiced sounds such as “s” looking similar to noise, while the vowels have a more distinct shape due to their harmonic patterns.

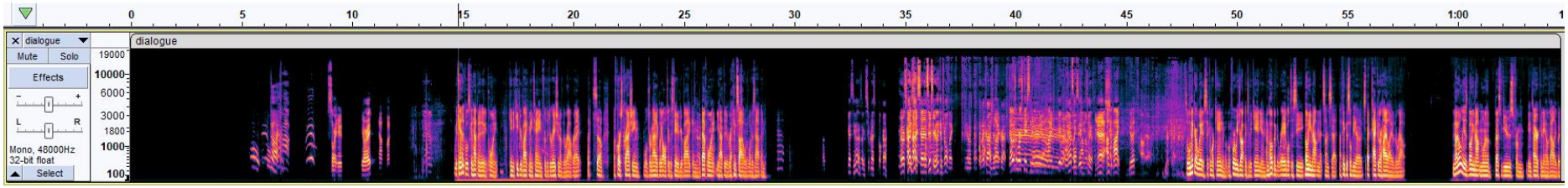


Figure 28 Isolated Speech Track Spectrogram

2. Music Track

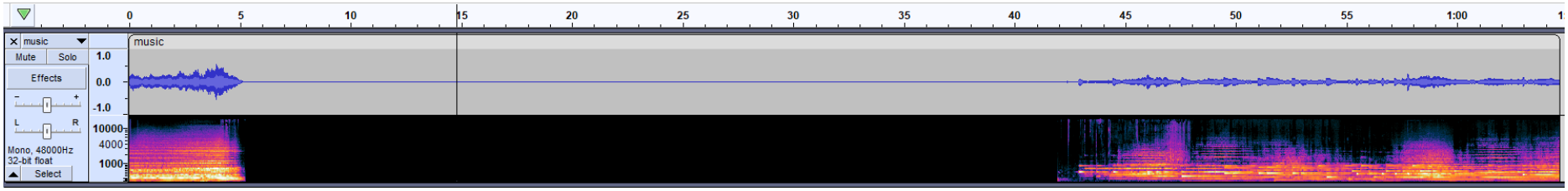


Figure 29 Music Track Combined View

From the waveform we can clearly see where there are segments that contain music and the areas where there is no music. Unlike the speech waveform, we can see that the areas that contain music show much more activity.

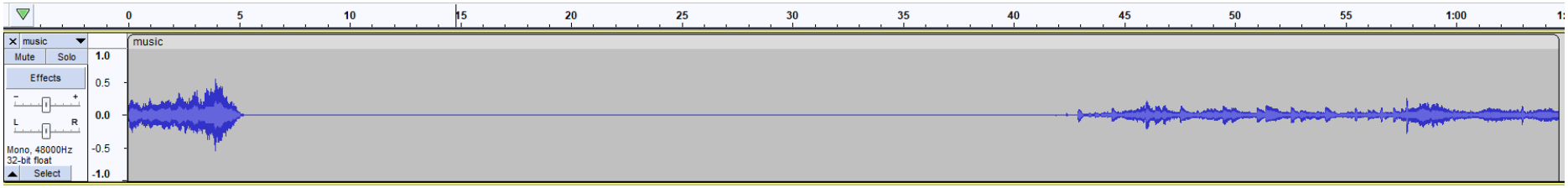


Figure 30 Music Track Waveform

In the spectrogram we can notice that there are areas that contain music appear to be much denser than in the speech track. The dense colour patterns indicate the presence of various instruments. Just like in the waveform, we can clearly notice the “silent” areas which appear completely black, indicating the complete absence of music.

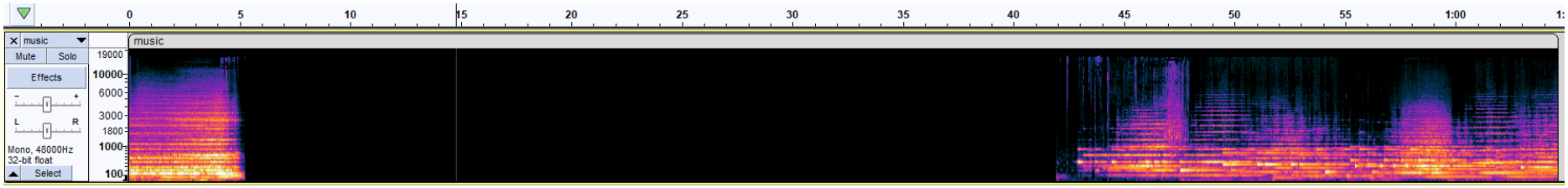


Figure 31 Music Track Spectrogram

3. Sound Effects Track

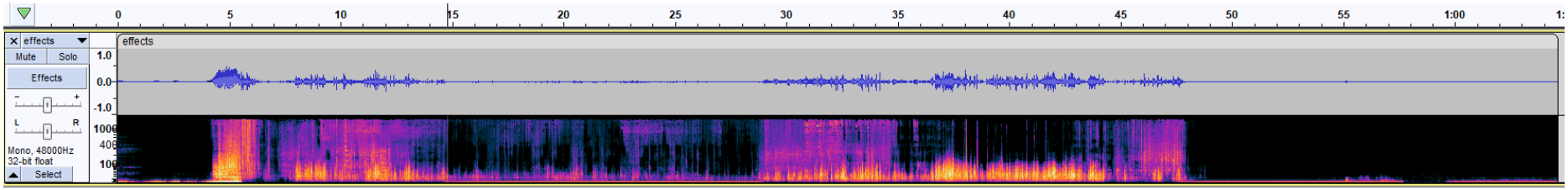


Figure 32 Sound Effects Track Combined View

From the waveform we can notice the same audio events as in our previous track, areas where there are loud sound effects, represented by the steep peaks and the short bursts of audio data. Completely silent areas are also present alongside areas that, although they contain sound effects scattered across most of the audio track, are not as loud as the music or the voice events. Nevertheless, there are areas where we can see steep peaks and valleys which indicate loud sounds.

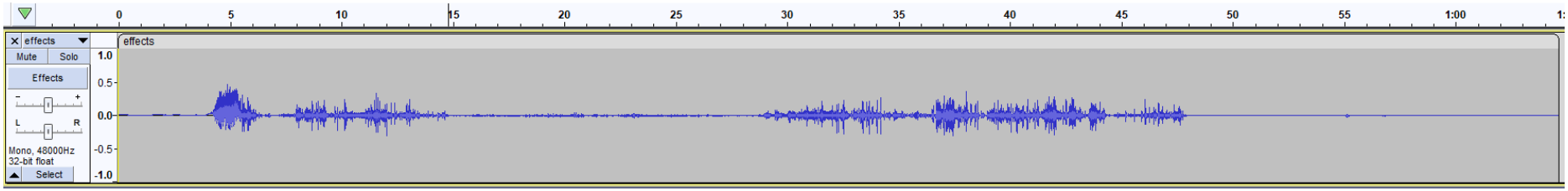


Figure 33 Sound Effects Track Waveform

From the spectrogram we can notice events that, although are not saturating all the frequencies, are distributed rather evenly throughout the track. Additionally, we can observe that there are different types of sound effects: we have loud and concentrated ones in the bright light-

coloured areas, but also quieter sound which saturate all frequencies. Lasty, we also have areas where the sound effects are extremely quiet or not present at all.

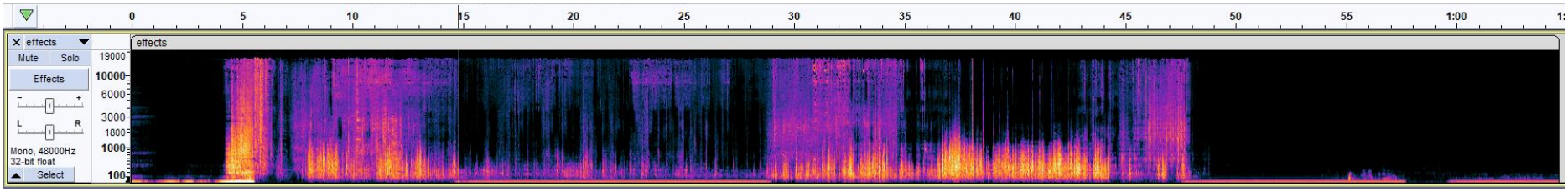


Figure 34 Sound Effects Track Spectrogram

4. Evaluating Insertion Points/Intervals

Now that we have a better understanding of how the various types of audio events are combined in order to create the original audio track, we can start using the Insertion Point Decision Matrix to analyse areas of the original audio track and score their suitability as an insertion point/interval.

For this example we will analyse short 10-second snippets from the same audio track example by looking at a cross section through all three tracks:

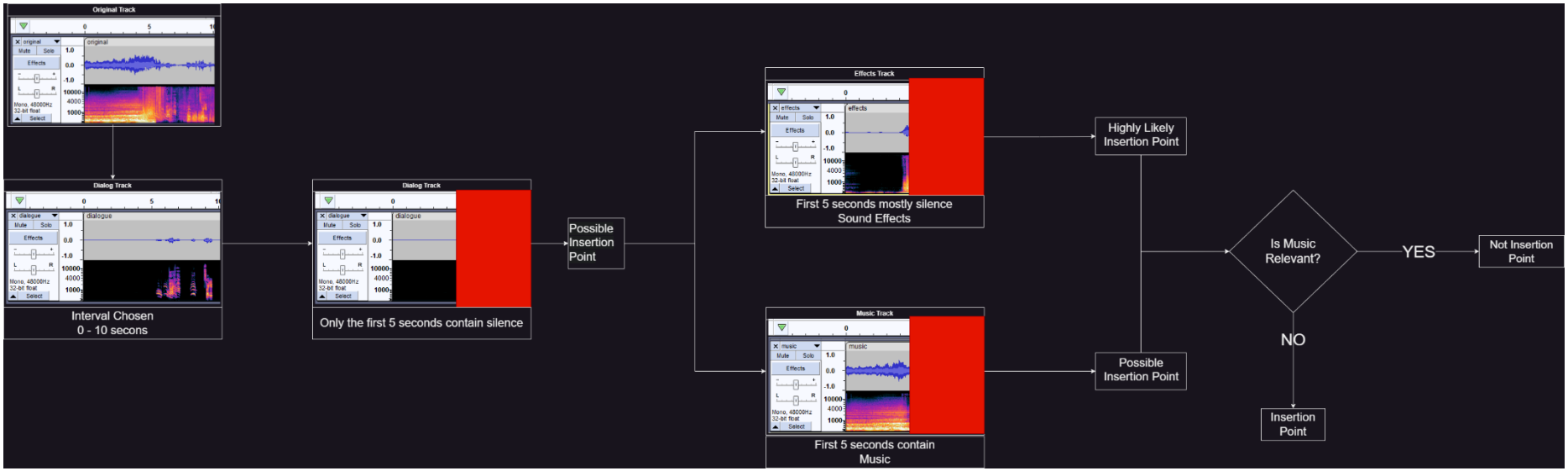


Figure 35 Example of Audio Track Analysis for Determining Insertion Points

We have selected a 10-seconds interval (from 0 to 10 seconds) in order to evaluate the feasibility of inserting an audio description title:

- By looking at the Speech Track, we can notice that only the first 5 seconds do not contain speech.
    - According to the insertion matrix, this would be a possible insertion point.
- We proceed with investigating only the first 5 seconds of our chosen interval.
- By looking at the Music Track, we can notice that music is present in the entire chosen interval
    - According to the insertion matrix, this remains a possible insertion point.

- By looking at the Sound Effects Track, we can notice that there are not major sound effects present.
    - According to the insertion matrix, this is a highly likely insertion point

Nevertheless, since we have music present across the entire 5 second interval, the final decision depends on the relevancy of the music element. If the music is paramount to the narration, then we will not be able to insert audio description. If the music is not paramount to the narration, we can insert audio description.

4. Audio Description System Type Decision Matrix

For the purpose of our investigation we will consider that the Voicing and Mixing step will always be automatic, and the difference between the system types will be defined by the manner in which the script creation process is performed.

System Type	Finding Insertion Points	Generating the Description	Voicing the AD track	Mixing the AD track
Fully Automated AD	Automated	Automated	Automated	Automated
Human-in-the-loop AD	Automated with QC	Automated with QC	Automated	Automated
Human-in-the-loop AD	Automated with QC	Automated	Automated	Automated
Human in the loop AD	Automated	Automated with QC	Automated	Automated
Computer Assisted AD	Automated	Manual	Automated	Automated
Computer Assisted AD	Manual	Automated	Automated	Automated
Traditional AD	Manual	Manual	Automated	Automated

Table 2 Audio Description System Decision Matrix

From the matrix we notice that we can have three types of audio description systems:

Fully Automated AD: this system does not need the input of a human. It would take in the media that has to be described, and will output the mixed result.

Human-in-the-loop AD: this type of system is less automated, but human intervention is only present during the quality check process. At least one of the main actions, the detection of insertion points or the description generation must be quality checked by a human.

Computer Assisted AD: this type of system is even less automated than the previous one, and it relies on at least of the main action, insertion point detection or description creation to be done by a human.

Traditional AD: this type of system relies on a human describer to identify the AD insertion points but also to create the description.

3. Generating the Description Text

The most difficult task in automating the AD process is to automatically generate the AD script, and this is due to the fact that AD is not an independent piece of text; its purpose is to further augment the existing meaning conveyed by the movie and complement the existing audio track that contains dialogue, narration, sound effects and musical scoring (Braun, 2011).

To automatically generate the audio description script, we need to identify the areas where audio description can be introduced and then generate the description that will be introduced in those audio description insertion points. In the previous chapter we have examined the analysis of the media in order to identify possible audio description insertion points, in this chapter we will focus on how we can extract that salient information from the scenes that span across those audio description insertion points.

The AD process makes use of the entire arsenal of knowledge that the audio describer has accumulated throughout their life; from mental models, anticipation, the ability to convert sensory inputs into text, inference and retrospective self-correction that enables one to understand and grasp the intended meaning of a scene. When individuals understand discourse, or perceive the world, or imagine a state of affairs, they, according to theory, construct mental models of the corresponding situations. In the case of verbal reasoning, they construct models from a representation of the meaning of the assertions and, where relevant, from general knowledge (Bell and Johnson-Laird, 1998).

Although the audio description contains the word *description*, it represents much more than a simple verbalised description of a scene. Unlike translation where the channel of transmission and modality of transmission are the same (i.e. visual, written text), in the case of AD we are dealing with a multi-modal type of translation(Bell and Johnson-Laird, 1998; Taylor, 2019). The audio describer has to "translate" what they see into words and thus must also consider the communication channel change and the amount of information that can be transmitted through the target channel. Moreover, they cannot simply describe everything that they see in an image, as that would miss the practical purpose of an AD and prove to be impossible. The visual content has to be processed, analysed and key elements have to be selected and actions condensed in order to fit the time allocated for each AD segment Furthermore, audio describers

have to go through training in order to be able to produce AD tracks that aghere to the necessary guidelines(Walczak and Fryer, 2017).

In order to describe an image, a human has to first extract the information presented in a visual form, process it (i.e. decode) and then express the same information in written form (re-encode the data). Image captioning is the process through which textual description are automatically generated from visual data. It stems from advances in the field of visual object recognition and detection, and it can be considered the next step describing the entire frame/scene/image once every object is identified and tagged (Xu et al., 2015). Furthermore, there have been researchers who have investigated selection and even the creation of corpora aimed at being used to train algorithms for this precise purpose, such as the dataset for visual storytelling (Huang et al., 2016).

1. Understanding Images and their Descriptions

An important aspect of the AD process is the correct identification of the key elements that have to be described. A human audio describer would rely on the experience that they have gathered through training in order to direct their attention to the most important elements of a scene. Xu et al., (2015) introduced an attention-based machine learning model that automatically identifies and describes key elements of an image. Their reasoning is that in order to mimic the human ability to acquire process and compress large amounts of salient visual information and convey the key elements using descriptive language; machines have to learn to direct their "attention". The proposed attention-over-time approach illustrates how as the model generates each word its attention shifts to relevant parts of the image that are being described (Xu et al., 2015).

If this concept of attention could be applied to a series of images, in order to lock on a key actor, it would help increase the narrative coherence of the generated description. Another important aspect of this framework is that it permits the user/researcher to visualise "where" and on "what" the attention is being directed. Such a feature allows a better understanding of the description generation process by offering insight into how machine attention is being directed in an image or a sequence of images (Xu et al., 2015). Xu et al.'s,(2015) proposed an attention-based approach that was benchmarked on three datasets by using the BLEU and METEOR metric. They demonstrated how learned attention can be exploited to give more interpretability into the model's generation process. Furthermore they demonstrate that the

learned alignments correspond very well to human intuition(Xu et al., 2015). Such types of research into trying to mimic the results of human intuition, a complex mental process, shall play an important role in developing ML models that can extract the salient information from a still video frame.

When describing an image, humans rely on their pattern recognition capabilities to identify key elements of a scene. Trained human audio describers use all their previous knowledge to extract information that is important from a narrative point of view (Walczak and Fryer, 2017). A description of a sequence of images needs to take into account the temporal dimension of the actions depicted across the image sequence. The connection and interaction between these elements is what creates the narration and conveys the true meaning of the information that was visually encoded. It is not enough to identify visual elements such as objects, entities and scenery to depict the narrative of the image sequence. (Fryer, 2016). The descriptions have to be connected and condensed in order to extract the actions that occur over a period of time during the sequence of images (Huang et al., 2016; Surikuchi, 2019; Braun, 2020).

Training a machine to have the capability to mimic human attention will greatly improve the chances of said machine "understanding" an image. But "when does a machine "understand" an image? One definition is when the machine can generate a novel caption that summarises the salient content within an image."(Mitchell et al., 2015).

2. Conceptualising Audio Description Generation

One way of conceptualising the AD process is in terms of an encode/decode system. This idea stems from the encoder-decoder system that is used in machine learning. One issue with this type of thinking is the fact that when viewed from a translation studies point of view, it requires the application of an interlingua theory (Vauquois, 1968). We could consider the type of meta-language through which the information is passed from one medium to another as interlingua. In the case of neural machine translation this interlingua approach works because the "analysis" component is represented by the encoding process, while the "transfer" component is represented by the decoding process. However, this type of reasoning does not work in the case of humans. Nevertheless such an approach might be helpful when creating machine learning models aimed at translating visual information into natural sounding sentences.

An important aspect of AD creation stems from the Relevance Theory (Grice, 1969)which asserts that explicating and implicating, two important mental processes in information

understanding, are guided by the human tendency to maximise relevance (Cognitive Principle of Relevance). An audio describer aims to maximise the amount of information that can be conveyed through the generated AD while also being aware of the limitations that AD entails such as sentence length and utterance speed.(Braun, 2016; Fryer, 2016)

Our society has developed relying on our senses, especially on our vision. Thus for most people, describing/transposing into words what they have seen in a short video represents an easy task (Venugopalan, Xu, et al., 2015). Naturally as most people are not trained to be video describers, the quantity and focus of the descriptions may vary and thus not be on par with those created by a professional audio describer. Audio description has to respect specific criteria in order for it to be of real use to a BPS person (Fryer, 2016).

Rohrbach et al. (2013) argue that computer vision is advanced enough to detect people, classify their actions, or distinguish between large numbers of objects and identify their attributes. Furthermore, these computer vision algorithms are able to output the information as semantically encoded activities and object categories that can easily be processed by automated systems (Rohrbach et al., 2013).

When taking a closer look at the reasons for which most video detection and classification algorithms have been built such as, but not limited to: object detection, collision detection, facial recognition; it is easy to understand why the generated output was not designed to take form of a natural sounding sentence. The outputs are simply unfit to create an audio description script (Surikuchi, 2019; Braun, 2020).

Although humans can make use of their entire arsenal of rich natural language to describe and communicate visual information, a person will never be able to fully convey the information and create the same impression on another human just by describing visual content (Rensink, 2000; Venugopalan, Xu, et al., 2015). Nevertheless, although it is unable to replace the visual experience, AD does convey information to which a BPS person would not have access due to their circumstances.

3. Before Large Language/Vision Models

The first step in automating this part of the AD process is analysing the content of the visual information by implementing deep learning in order to identify events and key elements. There

are two broad types of neural networks used in achieving these goals: convolutional neural networks (CNN) which are being used for object recognition such as: YOLO9000 (Redmon and Farhadi, 2017), YOLOv4(Bochkovskiy et al., 2020), Faster R-CNN (Ren et al., 2017), R-FCN (Dai et al., 2016) and DeepCaption (Aalto University).

Other research is being done by using recurrent neural networks (RNNs) to generate image description such as:

- Mind's Eye (Chen and Zitnick, 2015) - explores the bi-directional mapping between images and their sentence-based descriptions. The authors use a RNN that attempts to dynamically build a visual representation of a scene as a caption is being generated. Their model is able to automatically learn to remember long-term visual concepts; and also be able to generate a novel caption based on a given image.
- Long-term Recurrent Convolutional Networks for Visual Recognition and Description (Donahue et al., 2017) - Authors describe a class of recurrent convolutional architecture which is end-to-end trainable and suitable for large-scale visual understanding tasks such as image captioning and video description.
- From Captions to Visual Concepts and Back (Mitchell et al., 2015). This article showcases a novel approach for automatically generating image description by using visual detectors, language models, and multimodal similarity models learnt directly from a dataset of image captions.

With regard to video analysis and object detection and recognition, there are several approaches which combine both CNN and RNN models to describe videos:

- Jointly Modelling Embedding and Translation to Bridge Video and Language (Pan et al., 2016) This article describes a method that embeds rich temporal dynamics in visual features by hierarchically applying Short Fourier Transform to CNN features of the whole video.
- Sequence to Sequence – Video to Text (Venugopalan, Rohrbach, et al., 2015) The authors describe an LSTM model that is trained on video-sentence pairs and learns to associate a sequence of video frames to a sequence of words in order to generate a description of an event in a video clip.
- Translating Videos to Natural Language Using Deep Recurrent Neural Networks (Venugopalan, Xu, et al., 2015) The authors propose a to translate videos directly to

sentences using a unified deep neural network with both convolution and recurrent structure.

- Exploiting long-term temporal dynamics for video captioning (Guo, Zhang and Gao, 2019). The authors describe a novel approach named temporal and spatial LSTM (TS-LSTM) which is designed to incorporate both spatial and temporal information to extract long-term temporal dynamics within video sub-shots; and a stacked LSTM is introduced to generate a list of words to describe the video.

In order to create models that can produce novel image descriptions, such models have to be trained on datasets which contain pairs of images and their captions (i.e. Microsoft COCO). Mitchell et al. (2015) talk about three advantages that image captions have in training models: firstly, they mention that image captions contain only inherently salient information. Secondly, training a language model (LM) on image captions captures the common-sense knowledge about a scene. Thus, noisy visual detection is disambiguated by taking advantage of this knowledge. Thirdly, going back to the multimodal aspect of translating images to text, by learning a joint multimodal representation on images and their captions, global similarities between images and text can be calculated and then used to select the most appropriate description for the image (Mitchell et al., 2015).

Although AD is used for describing single images (such as paintings in an art museum), here we are concerned with AD that is used in describing movies and videos. While a movie is just a quick succession of images, merely describing each image would not suffice for creating a usable piece of AD. Description of a series of images, unlike stand-alone images, adds another dimension to the description process, namely narrative cohesion. The image succession creates a narrative story which evolves and changes with each new scene, character or action.

Visual storytelling aims at investigating the generation of narration through description of moving images or GIFs (Huang et al., 2016; Surikuchi, 2019). Huang et al. introduced in 2016 the first dataset for sequential vision-to-language (VIST). They further explored how this data can be used for the task of visual storytelling. Their approach was aimed at trying to move artificial intelligence from basic understanding of typical visual scenes towards a more human-like understand of grounded event structure and subjective expression (Huang et al., 2016). This type of approach plays an important part in producing machine descriptions of visual content which are more natural-sounding and thus easier to be accepted by the end users.

Although in the field of image recognition, object detection and description there is a large number of comparatively high-quality, annotated datasets, these captioned image datasets were not developed in order to serve linguistic studies. Therefore, the data is not optimized for training algorithms aimed at generating descriptions (Braun, 2020).

Datasets represent an important factor in the training of description models, a model can be only as good as the data on which it is trained allows. Unlike other computer vision tasks, generating a pertinent description that could be voiced into an AD track requires a higher precision and accuracy during the caption process for the dataset (Huang et al., 2016; Braun, 2020).

1. Training Datasets

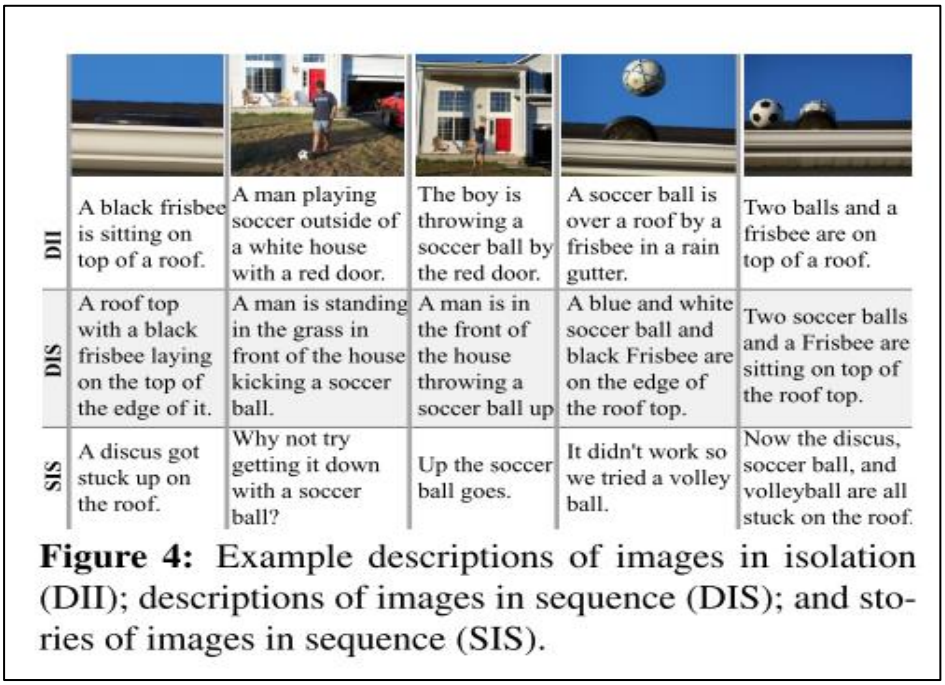
As mentioned above, generating captions/descriptions for a set of images introduces a new dimension in the description process. Unlike in the case of a single image description, when working with a series of images, the model has to be able to make connections between the images in order to establish a narrative. If the training datasets on which the model was trained lacked a thorough quality assurance process, the model will not perform as intended and thus not output quality results (Braun, 2020).

In 2015, Lin et al. introduced the Microsoft Common Objects In Context (MS COCO) dataset. This corpus was developed as a large-scale annotated dataset which could be used in visual object detection and caption tasks (Lin et al., 2014). Each image has been annotated with five captions, each caption was written by a different human operative tasked with describing the content of the image (X Chen et al., 2015). The goal was to extract visually pertinent data from the images which could then be used by the machines to learn connections between objects, actions and their corresponding semantic labels generated by the annotators (Chen et al., 2015).

Crowdsourcing through Amazon Mechanical Turk proved to be a good choice that greatly increased the speed with which the dataset was created. Although crowdsourcing is an accepted approach when dealing with such massive datasets, it introduces many factors that decrease the quality of the end result. Datasets such as this one contain large amounts of data which is needed in order to train models. Unfortunately, in order to create such a large dataset meant that quality was sacrificed in favour of quantity. While this is perfect for tasks such as object detection, lack of precision means that it cannot be successfully used to train models to generate

natural-sounding descriptions. Braun demonstrated how using MS COCO for creating other corpora can lead to "less reliable and, demonstrably low in quality" results (Braun, 2020).

The creation of the first sequential vision-to-language corpus brought new hope to the researchers that were investigating the generation based on sequences of images (Huang et al.,



DII	A black frisbee is sitting on top of a roof.	A man playing soccer outside of a white house with a red door.	The boy is throwing a soccer ball by the red door.	A soccer ball is over a roof by a frisbee in a rain gutter.	Two balls and a frisbee are on top of a roof.
DIS	A roof top with a black frisbee laying on the top of the edge of it.	A man is standing in the grass in front of the house kicking a soccer ball.	A man is in the front of the house throwing a soccer ball up	A blue and white soccer ball and black Frisbee are on the edge of the roof top.	Two soccer balls and a Frisbee are sitting on top of the roof top.
SIS	A discus got stuck up on the roof.	Why not try getting it down with a soccer ball?	Up the soccer ball goes.	It didn't work so we tried a volley ball.	Now the discus, soccer ball, and volleyball are all stuck on the roof.

Figure 4: Example descriptions of images in isolation (DII); descriptions of images in sequence (DIS); and stories of images in sequence (SIS).

Figure 36 Huang et al. (2016)

2016). Image description/captioning represents a problem that has to be approached from two sides: on one hand we have the computer vision side in which progress is made from single stand-alone images to sequences of images which depict events as they occur and change over a period of time. On the other hand, we have the linguistic side in which the progress is done from a literal description to several consecutive descriptions that generate a narrative (Huang et al., 2016). To be more exact, this is what makes a difference between "sitting next to each other" and "having a good time". Huang et al. (2016) explain how in order to obtain the second example, the computer has to be able to infer what exactly "a good time" represents or how exactly "sitting next to each other" can be interpreted as "having a good time" (Huang et al., 2016).The sequential vision-to-language dataset which contains sequential images with corresponding descriptions captures these subtle differences and aim to advance the task of visual storytelling (Huang et al., 2016).

One important aspect of the sequential vision-to-language dataset (VIST) is the way in which it has been structured: the authors released the data in three tiers of language for the same images: (1) Descriptions of images-in-isolation (DII); (2) Descriptions of images-in-sequence

(DIS); and (3) Stories for images-in- sequence (SIS) (Huang et al., 2016). This can be observed in Figure 2 where we have a table of several images each having three types of descriptions of various lexical complexities. A possible future research direction would be to investigate if such tiered approaches could be used to vary the lexical complexity of a description in order to offer users the chance to personalize the output to their needs.

2. Vision to Language Models

Vision to language work has expanded into several directions with researchers addressing image captioning (Elliott and Keller, 2013; Lin et al., 2014; Young et al., 2014; J Chen et al., 2015; Vinyals et al., 2015; Xu et al., 2015) visual phrases (Sadeghi and Farhadi, 2011), video understanding (Ramanathan, Liang and Fei-Fei, 2013) and visual concepts (Mitchell et al., 2015; Krishna et al., 2017).

The majority of these pieces of research are aimed at generating literal descriptions of the content of an image. Huang et al. consider this as an important step towards connecting vision and language, but also outline that this is far from the capabilities needed to generate natural interactions (Huang et al., 2016), and therefore descriptions.

They further stress the importance of a machine to be able to "understand" an image and distinguish between scene descriptions. The example used is that of a scene "sitting in a room" versus a "bonding" scene. Both descriptions are valid and come from the analysis of the same image, but the latter one can only be produced when alongside the literal visual information extracted (i.e. people sitting in a room), information about social relations and relations between entities present in the image can be inferred from the context of the image (Huang et al., 2016).

A human audio describer is trained to pay attention and extract only the most valuable information that can summarise the key aspects of a scene. Furthermore, as pointed out by Fryer (Fryer, 2016) audio description is neither a spoken, nor a written text – it lies somewhere in between. Moreover, the language used should be appropriate in order to fit the existing audio track and take into account the target audience. Mismatches between the tone of the audio description and the mood/atmosphere created by the audio track greatly reduce the quality of the AD experience for the end user (Fryer, 2016). Since I will be working with voice synthesis to voice the AD track, it will be extremely important to try and tune the voice in order to correctly match the atmosphere set by the original audio track.

Attention is what gives humans the ability to focus on a certain aspect of a situation and create a mind map of what relevant information in connection to given aspects. According to Rensink (Rensink, 2000), vision creates a powerful impression of a coherent, richly detailed world where everything is present simultaneously, but in reality, coherence appears only as long as attention is focused on visual structures that are considered stable (Rensink, 2000). Thus rather than observing and compressing the entire image or scene, our brain has the power to dynamically bring forward the salient features into forefront when needed by focusing attention on various parts of an image (Rensink, 2000; Xu et al., 2015).

Surikuchi (2019) argues it is the degree of inadequacy of previous automated image captioning methods that has sparked the need to achieve the generation of a narrative style text based on a given sequence of images. This is mostly due to the fact that sequence description, be it dense or sparse, lengthy or concise, is restrained under the cover of naïveté (Surikuchi, 2019). Such models lack imaginative power and inferential capabilities that would drastically increase the quality of the automatically generated descriptions (Surikuchi, 2019).

An audio description script is not just a literal description of a scene or a sequence of scenes, but the result of multimodal translation. Through this process, the audio describer decodes the information which is visually encoded, processes it, and filters it in order to re-encode the key parts needed to convey as much as possible in a written text that can then be verbalised. The aim of the audio describer is to offer BPS the opportunity to enjoy an experience as close as possible to that of a sighted person (Kiros and Zemel, 2013; Núñez, 2015; Fryer, 2016).

According to Núñez (2015), AD is an intersemiotic translation method which makes use of the human multimodal perception. The AD translation process involves a set of mental processes of multimodal and multi-sensorial nature. On one end of the process we have the perception of the audio and/orvisual content, while on the other we have the production of an equivalent text based on the information decoded and processed (Núñez, 2015).

From the perspective of image captioning, Surikuchi (2019) argues that the modules of encoder and decoder handle dependent but different objectives. The encoder generally is represented by a convolutional neural network and this is due to their viability in successful detection and summarization of visual semantics (Surikuchi, 2019). The decoder is typically represented by a standard recurrent neural network or one of its variants. These decoder networks are autonomously called language models (Surikuchi, 2019).

The dichotomy between the encoder and the decoder is in line with what would happen during the process of creating AD through multimodal visual-to-text translation (Núñez, 2015; Braun, 2016; Fryer, 2016).

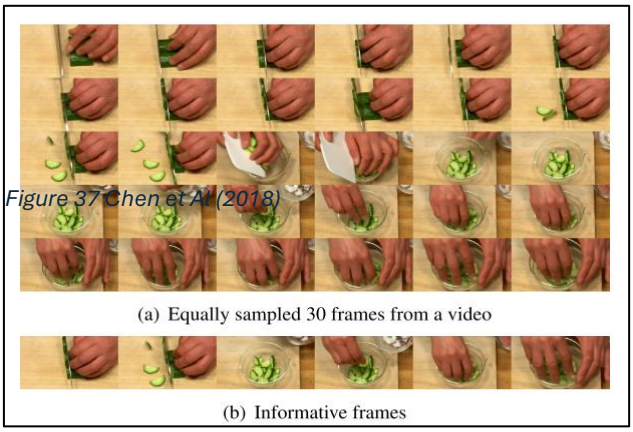


Figure 37 Chen et Al (2018)

Thus, in order to achieve better results when dealing with video captioning tasks, attention-based models such as the ones mentioned above (Fu, Jianlong and Zheng, Heliang and Mei, 2014; Hori et al., 2017) which extract and associate salient visual information with spoken sentences that have proven to be the most successful. Chen et al. (2018) argue that existing studies follow a common procedure "*which includes a frame-level appearance modelling and motion modelling on equal interval frame sampling".* Although sampling frames at a given fixed interval increases the chances of extracting relevant frames, it also increases the chances of picking duplicates and thus producing content noise which leads to increases in computational costs (Chen et al., 2018).

Chen et al. (2018) make reference to an article by Cromwell et al., (2008) which describes a biological mechanism called sensory gating. This describes the neurological processes that take place in humans in order to filter out unnecessary information caused by environmental stimuli and thus preventing an overload of redundant information in higher cortical centres of the brain (Cromwell et al., 2008; Chen et al., 2018).

As mentioned before, human audio describers are trained to identify the salient information from a video or scene and condense it in order to produce an AD track that can convey the key aspects presented in the video/scene (Fryer, 2016). From Cromwell et al., (2008) description of sensory gating and Rensink's (2000) explanation of how coherence only appears only as long as attention is focused on visual structures that are considered stable (Rensink, 2000) we can deduce that Chen et al.'s (2018) success in reducing the number of identical frames extracted not only decreases the amount of processing resources needed, but also brings the process of salient information extraction closer to how a human would identify key frames/scenes that are meaningful (Rensink, 2000; Cromwell et al., 2008; Fryer, 2016; Chen et al., 2018).

Chen et al. (2018) argue that despite the success that attention-based methods have with bridging vision and language, there still exist critical issues that have to be addressed such as frame selection and downstream video captioning (Chen et al., 2018).

4. After Large Language/Vision Models

In a previous chapter, we have looked at the possible concept for automating the generation of audio description, especially the difficulty of extracting the visual information and condensing it into a meaningful textual description. One of the major disadvantages of existing models was that although they could describe an image (video frame extracted from a scene) to a certain degree, they would not be able to describe a entire scene (a videoclip).

In recent years, models such as Chat-GPT4 Vision and Chat-GPTo from OpenAI,Pegasus1 from Twelvelabs, or Flamingo from Google Deepmind are capable of describing images and even scenes in a video. Nevertheless, these models are currently the most advanced models available for public use and also the culmination of a constant process of research which involves transfer learning, large language models, and lastly, large vision-language models. The last being the type of model which would enable the generation of audio description text automatically.

1. Foundation Models

A foundation model is a type of artificial intelligence model which is capable of performing a range of possible tasks with varied types of inputs and outputs such a text, image, video, and audio. As the name suggests, they can be used as a foundation for building more specialised models which would excel at performing specific downstream tasks. They can be used as a foundation because the foundation model has been pre-trained on large amounts of data such as large-scale datasets in order to learn a broad range of knowledge which can then be fine-tuned (adapted) for performing specific tasks (Ada Lovelace Institute, 2023.).

Imagine an audio describer at a sports event where they have to convey what is happening to a listener that cannot see the action. The audio describer should have a broad knowledge of the sport that is being played, its rules, the types of plays and their significance, and the context within the game. The easiest way to visualise the characteristic of a foundation model is to use an analogy with an audio describer.

Characteristic	Audio Describer	Foundation Model
Pre-Trained Knowledge	Have already studied the sport, are familiarised with the plays from watching countless games and know the terminology and the context of the various plays.	Is pre-trained on a large amount of data from various types of corpora, datasets and the internet from which it gains a broad understanding of language, facts, and general world knowledge.
Adaptation (contextual)	During the game they need to adapt their description to the events that are happening. They have prepared themselves with more knowledge about the teams that are playing and their players. They can predict possible outcomes based on the evolving context of the game.	Can be fine-tuned to perform translation tasks better by using their pre-trained knowledge.
Real-Time Processing of Information	Have to process the events and describe them in real-time, as the action unfolds,	Process input in real-time and generate the required responses based on their

	selecting and describing only the key elements of the action.	existing pre-trained knowledge.
Flexibility	By relying on their experience and knowledge about audio description, audio description practices and techniques, they are not limited to describing only sports, but can describe other events like concerts.	Can be fine-tuned to perform various specialised tasks.
Continuous Learning / Improvement	When encountering unfamiliar sports or events, they can learn from them and improve their description skills.	Can be updated and further fine-tuned using new data to improve their performance.

Table 3 AD Foundation Models Analogy

The vast knowledge base on which the foundation model was trained enables it to serve as a foundation for specific applications just like the vast knowledge and experience an audio describer has gathered can be applied to describe various types of performances or media.

2. Transfer Learning

Transfer learning is a machine learning approach in which a model that has been developed for a particular task is being repurposed as a base for a model that would perform a secondary task. It uses the knowledge that was gained from solving the initial problem and applies it to

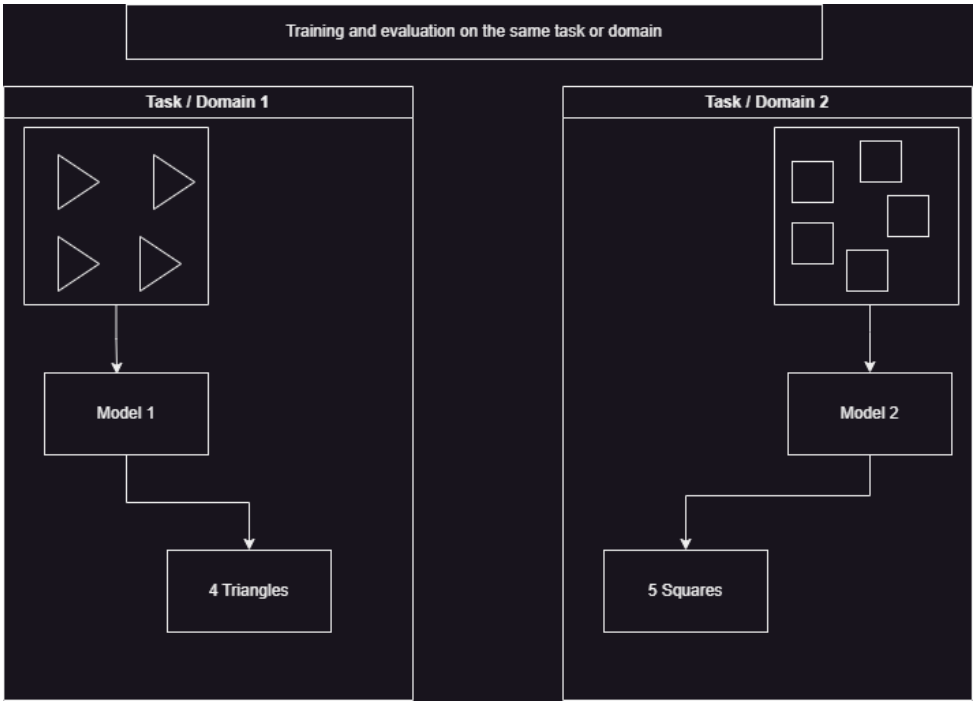


Figure 38 Training & Evaluation on the Same Domain/Task

solve a different but similar problem. As a result, instead of training a model from scratch, we can use an already trained model and fine-tune it to perform the new action. (Ruder, 2024).

In the diagram, we have two pairs of domains/tasks and one model that processes the information. For task 1, we have model that has to recognize and count the number of triangles, for task 2, we have a model that has to recognize and count the number of squares. Each of the models has been trained from scratch to recognize a specific geometric shape and count the number of geometric shapes. Nevertheless, since each of the models is trained to recognize specific geometric shapes, model 2 cannot recognize triangle and model 1 cannot recognize squares (Ruder, 2024).

With transfer learning, we can use the previously trained model 1 which detects triangles and fine-tune it to detect squares. This is done by using transfer learning, which allows the usage of existing labelled data of some related task and domain. The goal is to store the knowledge obtained from solving the source task (task 1) in the source domain (domain 1) and applies it to solve the problem from the target domain (domain 2) (Ruder, 2024)..

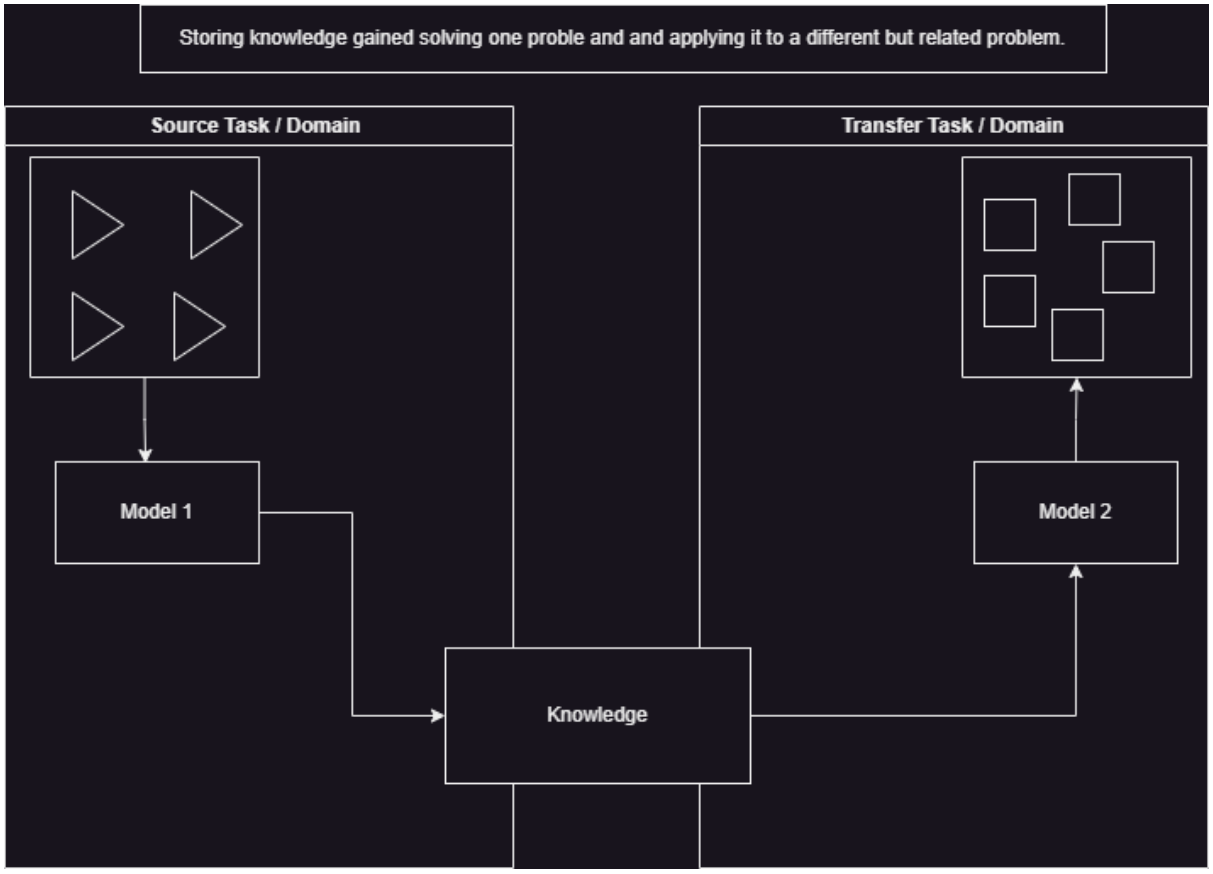


Figure 39 Knowledge Flow

The aim is to transfer as much knowledge from the source environment to the new target task or domain. The knowledge that is transferred can vary based on the type of data, from how the objects are defined, which would allow the identification of new objects, or identifying general words that people use to describe their sentiments etc. To better understand this concept, we will use our audio describer analogy.

Stage	Audio Describer	Pre-Trained Model
Initial Training (Initial Task)	They have spent years describing football games. They have adapted to describing fast-paced movements, player actions, and the whole flow of the game.	Pre-trained on large datasets composed of millions general-purpose texts, images and/or other data types which have general features of broad topics.
New Task	They are asked to describe another sport different from football, for example basketball. The sport is different, but the previously gained skills such as describing fast-paced game flow, player movement, player action etc., can be reused.	A pre-trained model is fine-tuned (adapted) to perform a new, more specific task. For example a model that is trained to classify images, could be fine-tuned to identify anomalies in weather satellite images.
Fine-Tuning (adaptation)	To improve the quality of their basketball description, the audio describer will have to learn new terminology related to basketball, understand the rules and the dynamics of a basketball game. But since they already have experience with sports, they will be able to pick this up faster than someone who has no sport background.	The pre-trained model is fine-tuned (the model's parameters are adjusted) using a small, more specific dataset. Since it has already understood general features, it will adapt much faster at recognizing new patterns and training times than it would take to train a model from scratch for the new task.

Figure 40 AD & Knowledge Transfer Analogy

3. Transformers

In 2017, the paper called “Attention is All You Need” (Vaswani et al., 2017), introduced the concept of transformers. This novel approach allowed language processing to be parallelised. Instead of using the previous Seq2Seq approach of processing data sequentially, one word at a time in the order in which the words appeared. The new approach enables the analysis of all the tokens in a given body of text to be analysed simultaneously. Transformers are relying on an AI mechanism known as attention in order to support this parallelisation process. Attention allows the model to take into consideration relationships between word, regardless of their distance in the text and also to determine which words and phrases in a section have the highest importance level and pay attention to them (Sutskever, 2014).

4. Vision Transformers

In computer vision the dominant approach was using Convolutional Neural Networks. With the success encountered by the Transformer architecture in Natural Language Processing, some researchers have tried to adapt this architecture to image data. In the paper “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale “, the Vision Transformer architecture was introduced, which applies the same encoder block of the Transformer architecture to the image classification problem. This new approach exceeded the results of the state-of-the-art image classification datasets while also being less expensive to train. (Dosovitskiy et al., 2020)

5. Large Language Models

The term Large Language Model has become a term, which most people are familiar with. This is all due to the success that OpenAI’s Chat GPT model has had with the public. We use language on a daily basis, expressing our thoughts, sharing our feelings and conveying information. Whether it is written or spoken, language is the one medium through which all

humans connect to each other. As a result, there was no surprise that a piece of software that can engage in conversations, either by answering questions, or engage in dialogues and even translate texts would become popular. Although it does not provide the same experience as talking to a human being, it still is a large language model which is considered to be a major advance in the field of artificial intelligence (Naveed et al., 2023).

What most people do not know is what GPT stands for: Generative Pre-trained Transformer. They are decoder-only models and use masked self-attention. Meaning that at a point in the output sequence, you can only attend to two input sequence vectors that came before that point in the sequence. The first GPT by OpenAI was made available in 2018 with the second version, GPT-2, in 2019; with GPT-2 was trained to predict the next word in all of the 8 million web pages it was trained on (Radford et al., 2019).

BERT (Bidirectional Encoder Representations for Transformers) by Google was published in 2019 as an encoder only Transformer and was designed to predict modelling tasks and also introduce the original concept of masked-language modelling. When training, BERT will mask out random words in the sequence and has to predict what the masked word was (Devlin et al., 2018).

In 2020, Google presented their T5 (Text-to-Text Transformer), an encoder-decoder model which takes in text strings as input and output (Roberts et al., 2020). It was trained on the C4 Dataset (Colossal Clean Crawled Corpus) (TensorFlow, 2024).

In 2020, OpenAI, presented GPT-3, their third generation GPT and one of their state-of-the-art models. The paper describes GPT-3 as having a capacity of over two magnitudes greater than the second generation GPT-2. It was also the first model that was able to write text pieces and articles which proved difficult to distinguish from human-written materials (Brown et al., 2020).

6. Large Vision Language Models

Just like in the case of generating an audio description script, where we are dealing with multi-modal manipulation of information which spans across the visual,textual, and auditory channels, as a result of the Vision Transformer architecture, the research interest in combining vision and language models has seen a steady growth.

Such hybrid vision-language models have demonstrated impressive capabilities at solving difficult tasks such as image captioning, visual question answering, and image generation. Since they are a combination of language and vision models, they consist of three elements: an image encoder, a text encoder, and a strategy to combine the information from both encoders

OpenAI introduced in 2021 CLIP (Contrastive Language-Image Pre-training) which was trained on 400 million pairs of text and images that were crawled from the internet. As a hybrid model, it encodes text using Transformers and encodes images using Vision Transformers. Then it applies contrastive learning to train the model, which matches the correct image and text pairs. While CLIP does not go directly from image to text or the other way around, it does use embeddings which are extremely useful when performing information search across modalities (Radford et al., 2021).

In 2022, Google introduced Contrastive Captioner (CoCa), a foundation model which combines contrastive learning (CLIP) and generative learning (SlimVLM) (Wang et al., 2021). This model uses a encoder-decoder architecture which was modified and trained with both contrastive loss and captioning loss. The result is a model that can learn global representations from unimodal images and text embeddings as well as fine-grained region-level features from a multimodal decoder (Yu et al., 2024).

Towards the end of 2022, DeepMind created a group of Visual Language Models called Flamingo. They have two parts: a vision model that can understand visual scenes, and a language model that helps with reasoning. Both models use their pre-trained knowledge to work together in order to handle sequences of mixed visual and textual data, as well as use images and videos as input (Alayrac et al., 2022).

In 2023, Google, Microsoft and OpenAI released their own large vision-language models:

PaLM-E, introduced by Google, is a proposed embodied language model which directly incorporates real-world continuous sensor modalities into language models. The inputs to this embodied language model are multi-modal sentences that interleave visual, continuous state estimation, and textual input encodings. These encodings are trained end-to-end, in conjunction with a pre-trained large language model, for multiple embodied tasks, including sequential robotic manipulation planning, visual question answering, and captioning (Driess et al., 2023).

Kosmos-1, introduced by Microsoft, is a Multimodal Large Language Model (MLLM) which was trained from scratch on web-scale multimodal corpora, including arbitrarily interleaved text and images, image-caption pairs, and text data. According to the paper, this model achieves impressive performance on (i) language understanding, generation, and even OCR-free NLP (directly fed with document images), (ii) perception-language tasks, including multimodal dialogue, image captioning, visual question answering, and (iii) vision tasks, such as image recognition with descriptions (specifying classification via text instructions) (Huang et al., 2023).

Chat GP-4 , introduced by OpenAI is a large multimodal model (accepting image and text inputs, emitting text outputs) that, while less capable than humans in many real-world scenarios, exhibits human-level performance on various professional and academic

benchmarks. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document (OpenAI, 2023).

Chat GPT-4o, introduced by OpenAI in May 2024, was trained as a single new model end-to-end across text, vision, and audio, meaning that all inputs and outputs are processed by the same neural network (OpenAI, 2024).

7. Large Video Models

When attempting to envision an automated system that would generate the description for a scene, the ideal situation was always to be able to analyse an entire scene, instead of just a frame from the scene. Videos add an extra temporal component which creates the narrative logic of a scene. Additionally, when analysing videos we also have to take in account the audio component, which, as previously discussed, has to work together with the visual in order to give the viewer the sense of being part of the action. Furthermore, the audio track contains audio queues and events which provide additional context to the scene and thus have to be taken in account when analysing a video scene. There are several emerging large video models such as VideoBERT (Sun et al.,2019), All in One (Lei et al., 2022) or VideoMAE (Tong et al., 2022) which would seem to be able to be good candidates for testing automatic generation of descriptions from videos.

Additionally, recently in 2024, Pegasus-1 by Twelvelabs and OpenAI's Chat GPT-4o represent the state-of-the-art multimodal models which could be used to generate audio description titles based on video scenes.

With the increasing interest in the development of large video models and large multimodal models, it is not impossible to imagine that soon enough, automating the creation of an audio description script will no longer be just a theory, but a reality. With this in mind, in the next chapter we will be looking at the research that I have conducted in order to create an

evaluation framework for algorithms and models which could be used to automate the audio description process.

4. Algorithm Evaluation Framework for Audio Description

Currently, there is no previously established framework for assessing how well an algorithm can be used to automate a part or step of the audio description process. While there are types and categories of algorithms that are fit to replace a human audio describer during certain steps of the audio description process, there are no established key performance indicators that can be analysed and logged for further analysis. Moreover, there is no established method or procedure for comparing how well multiple algorithms perform a certain audio description task. Consequently, there is no method of assessing and comparing the performance of these algorithms in a manner which would validate choosing an algorithm over another based on objective criteria.

Without access to such an important assessment tool, while the algorithms can be chosen and tested, the lack of a formal evaluation framework would decrease the credibility of the results. This is of paramount importance since any decision taken (by a human) during the process of creating an audio description track has to "[…]ensure that users of audio description have optimum access to audiovisual content thus minimizing the extent to which they are excluded from the experience"(ISO, 2015). Thus, any attempt at automating the audio description process must behave and perform as close as possible to the human driven counterpart in order to produce adequate outputs.

Assuming that human audio description is the "gold standard" when it comes to quality, we could envision the potential situation, however albeit unlikely, in which automated audio description could surpass human audio description and become the preferred option. Currently, the limiting factor is extracting the salient visual information and transposing it into words. Nevertheless, rapid advances in the fields of natural language processing and vision models are slowly getting closer to eliminating this limitation. Additionally, there are other factors which could lead to this potential shift towards automation:

- Consistency: Once an automated audio description system has been finely tuned, it would be able to produce uniform quality across high volumes of content. Such a thing could prove to be challenging when using multiple human audio describers.

- Scalability: Once an automated audio description system has been implemented, it would be easy to scale up by just running another instance of the automated system.
- Efficiency and Speed: An automated audio description system could generate audio description much faster than humans, thus could be used for live events/situations where quick turnaround time is paramount.
- Customizations and Adaptability: The automated audio description system could be adapted to different styles, preferences or requirements, and thus potentially offering users more personalised experiences.
- Cost-Effectiveness: In time, the cost of a developing and maintaining an automated audio description system could be lower than the cost of using human audio describers, especially in cases where solutions for large-scale operations are required.

Regardless of its advantages, for an automatically generated audio description to be preferred over human audio description, it must meet or exceed the quality of human-produced description; which means it would convey the same nuances, emotions and context of a human-created audio description.

Creating a comprehensive evaluation framework for audio description algorithms is vital for ensuring that the outputs of a possible automated audio description system are adequate. This framework must include clear KPIs and standardised methods for assessing algorithm performance. The goal of the framework is to ensure the algorithm output results are objective and trustworthy, and thus backing the selection of the best-performing algorithms when creating an automated audio description system.

1. Envisioned creation process for the evaluation framework

In order to create an evaluation framework for a given process, we need to fully understand the process and its outputs. An evaluation framework can be viewed as a model that will be used to assess/evaluate a given task or process. Thus, the first step is to analyse the reason for evaluating or the area where the evaluation framework will be applied and what exactly is to be evaluated. As mentioned before, the aim is to evaluate the algorithm fitness in the context of audio description automation, and in order to do that we need a more granular view of the human audio description process. This can be achieved by breaking down the human audio description process into smaller self-contained steps/processes. Such details can be extrapolated from analysing various documents that discuss the topic of human audio

description such as: national audio description guidelines, cross-national audio description guidelines, international standards, and audio description training courses.

The process of creating audio description for broadcast is broken down into three major steps, each comprised of smaller actions necessary to create an audio description track. These are creating the audio description script, voicing the audio description lines, and post-processing. Without an audio description script there is no audio description process, no voicing and no post-processing. With respect to automation, out of these previously mentioned steps, automating the creation of the audio description script is the most difficult one as it involves transcoding information from the visual into written form. Moreover, audio description is not just describing everything that is seen on the screen; as this would defeat the purpose of the audio description track. The audio description script contains all the audio description lines that were obtained from identifying key narrative elements together with the entry and exit time stamps which will produce the lengths of the voiced lines (Fryer, 2016; Walczak and Fryer, 2017). Furthermore, technologies such as voice synthesis and automated mixing are already being used in other domains, but there is no end-to-end automated audio description generator.

2. Audio Description Analysis

In this subchapter, we are going to continue analysing human audio description. In previous chapters, we have looked at the role of audio description and we analysed the human audio description workflow in order to better understand the flow of information and the transformative processes through which information is being transferred from the visual to the auditory. We will continue building on that analysis by analysing audio description from different angles: audio description as a process and what it entails, audio description as an output, and we will attempt to do a comparative analysis between the human audio description process and the proposed automated one. The goal is to ensure that we have a good understanding of what we are planning on evaluating and what we should focus on when creating the evaluation KPIs.

Audio description is not just a plain description of the visual content; it has to integrate and complement the content that is being described. As a consequence, the word choice and its delivery represent key factors in creating an enjoyable experience for the blind or partially sighted (ADLAB, 2014). Moreover, the audio description process is influenced by internal factors such as the quantity of visual information that has to be conveyed, but also external

factors such as the length and the number of audio track intervals where audio description can be inserted (Fryer, 2016).

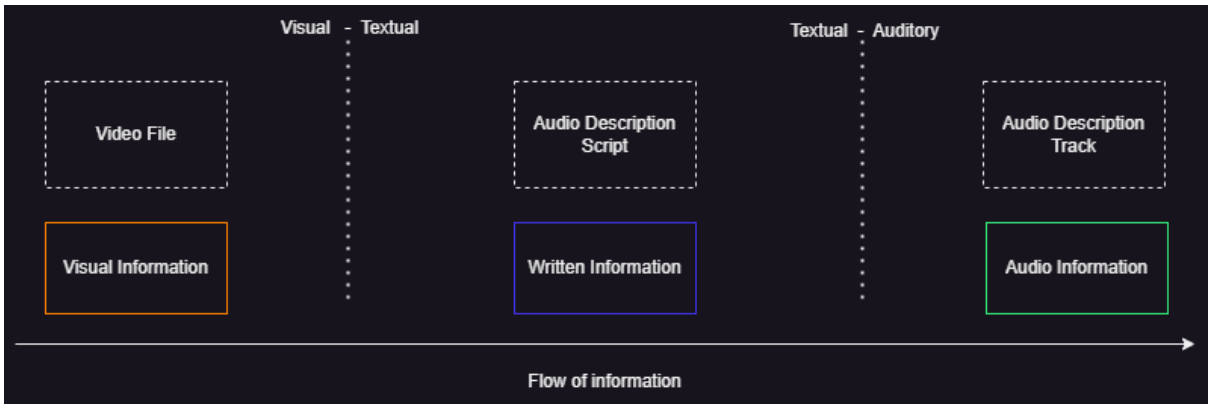


Figure 41 Human Audio Description Process Flow of Information

At its core, the audio description process aims to provide a secondary stream in of information in the audio channel. In order to make a piece of audio-visual content accessible, this stream of information is tasked with carrying the salient information from the inaccessible visual channel into the audio channel using the textual channel as an intermediary (Figure 1). There are important neurocognitive functions which are associated with audio-visual content and which are performed by specialised neurons. As we previously mentioned when addressing the role of audio description, mirror neurons play an important role in creating the 'valid co-occurrences' between the visual and the auditory channel, thus giving the audience the impression that they are the ones doing the actions that the actors are doing.

Additionally, we need to highlight the fact that the way in which the content of audio description is interpreted relies heavily on the previous knowledge and experience and expectations of the audience with the context in which it is presented. In case of sighted persons, in their research,(Marian et Al., 2021) has shown how experience with different types of auditory inputs, be it speech or environmental sounds will change the manner in which humans remember concurrently present visual objects. This means that audio inputs play an important role in how we build up our previous knowledge. A good example would be the nostalgia effect that we feel when we are listening to a song that we used to hear when we were younger. Research has shown that due to the multisensory manner in which we form memories, we do not form memories which purely rely on one sensory input, but that we use a mixture of senses to fixate in our brains how we felt at that moment. In our research, we are going to focus on the two main channels in a piece of audio-visual media: the visual and the auditory (Marian et al., 2021).

This multisensory aspect in which we experience the world, will impact not only the listener of the audio description track, but also the creator of the audio description track. This is the key reason for which the first step in creating audio description is to first watch and understand the piece of media that will be audio described. This represents a paramount step in the audio description process as it allows the human audio describer to contextualise and familiarise themselves with both the visual and the auditory information presented by the piece of media.

In the next step, the audio describer has to decide where audio description could be inserted in order to not disturb the original audio track. This is done by first creating an audio description script which would serve as the basis for the creation of the audio description track. Determining where and how audio description could be inserted is done by using guidelines either established by their clients/broadcasters or, where existing, guidelines established by the local broadcasting authority. While these documents can serve as guidance when creating an audio description script, the final decision will still rely on the audio describer's knowledge and experience.

1. Audio Description Script Creation

As previously mentioned, the audio description script represents a timed list of titles which contain written descriptions of the visual information presented in the scene they cover. An audio description script title has two components: the timing (the metadata) and the textual description (data). The most important type of metadata is the incue and outcue of the title as they set the timing of the title and thus define an audio description insertion point. As covered in a previous section of this paper, these insertion points represent areas in the original audio track where recorded audio description lines could be inserted.

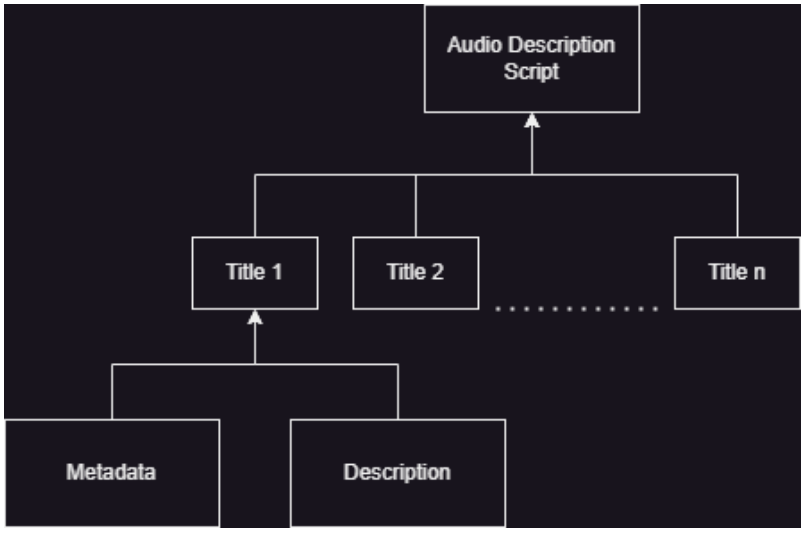


Figure 42 Audio Description Script Composition

1. What to describe

In his article (Snyder, 2005), Snyder explains that audio description is something that "[…]is a kind of literary art form in itself, to a great extent. It is a type of poetry—a haiku. It provides a verbal version of the visual—the visual is made verbal, and aural, and oral*"* (Snyder, 2005, pp. 936-937). Just as in the case of translation, where different translators will give different versions for a translation of the same text, audio description is heavily influenced by the audio describer's knowledge and experience. During the process of audio description, the audio describer has to observe and interpret the visual information based on their experience, thus the information is filtered based on their subjective point of view. It is important to remember that an audio describer should not insert their own subjective views in the audio description, they have to be as objective as possible when audio describing content (Snyder, 2005). As a result of this requirement for providing as much objectiveness as possible, a need for normative/guidance documentation has appeared. These documents are presented in the form of audio description guidelines which tackle the audio description process and its output. Nevertheless, more recently, there have been some audio description professionals and even academics who are encouraging audio description producers to explore subjectivity in descriptions, especially when dealing with different styles of audio description (Di Giovanni, et al., 2023).

In an article which addresses the issue of teaching relevance in audio description, Ibañez (2010, p. 143) puts emphasis on how important the topic of information relevance is when learning/teaching audio description. "Given its importance for professional practice, the acquisition of an appropriate yet flexible concept of relevance for audio descriptions is one of the main aims of any AD course[…]"(Ibañez, 2010, p. 143). Being able to identify the relevant visual elements which create the narrative force in a scene is one of the most important skills that an audio describer has to have. This is mainly due to the fact that, as mentioned in the previous paragraph, audio describers filter the visual information through the lens of their previously acquired knowledge. By being able to identify the key elements/aspects of a scene and describing them accordingly, the resulting audio description track will complement the existing context set up by the audio-visual content and thus provide valuable information to the blind or partially sighted audience. A crucial aspect of training an audio describer is teaching

them the relevance of the information conveyed by audio description. Additionally, information relevancy plays an important role in evaluating the audio description scripts.

Piety (2004), talks about the concept of appearance saying that *"*In some ways, appearance is the antecedent of all the other types of representation because all representations require an appearance of something in the original production to be realized in the description"(Piety, 2004, p. 459). Audio describers have to describe what they see/identify in the visual content; thus, it is the analysis of the appearance that allows them to provide, through verbal description, the properties of something present in the production. These characteristics can include: luminance, colour, size and shape (Piety, 2004).

Deciding what is relevant when describing a film, TV programme, or any other visually oriented content is crucial due to the fact that their creators *"want to offer their audiences an experience that is driven by a story or narrative"* (ADLAB Project, 2014, p.5). If creators of content which is inherently multimodal want to create an story-driven experience, the audio describer has to be able to analyse the original content and has "*to identify what story filmmakers want to tell and what principles and techniques they use to tell their story"* (ADLAB Project, 2014, pp.5-6).

The ADLAB guidelines highlight the fact that when watching a movie/piece of audio-visual content, the audience has to recreate the story that the producer/authors have initially created. As most of the multi-media content heavily relies on the visual aspect, the stories and narrative that are being told cannot be fully grasped by audiences who do not have access to that type of information. This is mainly caused by the fact that while authors/filmmakers have full knowledge of their characters, their actions, the temporal and spatial settings, whereas the audience has to learn all that information by watching the movie/content. Furthermore, the audience does not have immediate and instant access to all the information, they have to slowly piece it together throughout their watching experience (ADLAB Project, 2014).

2. When to describe

The audio describer has to analyse the video content in order to identify the key places where an audio description line can be inserted without interfering with elements from the existing audio track. Ideally, the audio described parts should be inserted in places where there is no dialogue or speech in the original audio track; this is done in order to not overwhelm the listener. Furthermore, there are also areas in which the volume of the original audio track is

loud or there are loud sound effects which would interfere with the clarity of the audio description track (Fryer, 2016).

With recent advances in the field of AI and large vision models, we are starting to see the emergence of several multimodal video understanding solutions, such as Pegasus-1 by Twelve labs or Chat GP-4o by OpenAI. While they are not exclusively aimed at automating the creation of an Audio Description Script, their state-of-the-art video-to-text generation capabilities can be combined with other technologies in order to automate the process. These new developments represent an important step towards automating the generation of the audio description script because movies and similar content rely on multiple information transmission channels, mainly the visual and auditory, to convey their message. Just like the audio describer makes use of the information received through both channels, the automated system has to be able to use information from both channels in order to determine where exactly audio description can be inserted. (ADLAB Project, 2014).

3. How to describe

The time constraints which are imposed on the audio description length are the major factors that influence the way in which audio descriptions are created. This directly affects the quantity of information that can be verbally transmitted. Thus, the word choices made while creating the audio description lines have to be adapted in such a way that the amount of information transmitted is maximized while keeping the sentence/utterance length as short as possible (ADLAB, 2014). The English language has a natural advantage over other languages due to the fact that it is a synthetic language and thus allows the user to convey more information while using fewer words. As a result, the English language allows the free creation of compound adjectives which can be used to condense multiple characteristics into one word (Fryer, 2016).

2. Subchapter conclusion

To further our understanding of the audio description script process, it is essential to analyse existing audio description guidelines. This involves a thorough analysis of some of the existing audio description guidelines and documentation with the goal of observing how they address key aspects such as: objectivity/subjectivity, information relevance together with timing and language/word choice.

3. Analysis of audio description documents

1. Overview of types of documents on audio description

Worldwide, there are several audio description guidelines, which, generally speaking, were developed to be used at a national level. While each of them was developed within a given broadcasting environment specific to each country, the process of audio description that they are referencing remains the same. The main goal of this part of my research is to analyse several audio description guidelines and compare them with Information Technology – User Interface Component Accessibility – Part 21: Guidance on audio descriptions (ISO / IEC TS 20071-21: 2015). While all the mentioned documents describe the audio description process in its entirety, the main focus of this research is to identify key visual elements that have to be described and presented in an audio description track in order for it to enhance the audiovisual experience of the listener. Identification of key visual elements is an important step of the audio description script creation process as these key visual elements lay the foundation upon which the rest of the description can be built. By having a defined set of categories of visual elements, audio describers are able to produce audio descriptions which accurately convey the emotional content and narrative context presented in a scene. Additionally, having a list of key visual elements will also permit the audio description to be more consistent across different audio description tracks.

As mentioned before, there are various audio description guidelines that are accepted at a national level. As cultural references and broadcasting culture varies from country to country, the documents chosen to be analysed in this piece of writing are all from major English-speaking countries: the United Kingdom, Ireland, Australia, New Zealand, the United States of America and Canada. The main reason for selecting this set of countries is that they have very similar broadcasting environments, broadcasting regulatory bodies and, to some extent, legislation, thus increasing the chances of compatibility between guidelines. Additionally, two more documents will be considered, the ISO / IEC TS 20071-21: 2015, published by the International Organization for Standardization which addresses guidance on audio descriptions and the ADLAB Project Audio Description Guidelines, which were the result of a project funded by the European Commission in order to create pan-European audio description guidelines.

Audio description guidelines are not only built at a national level, but they are also created by large audiovisual content delivery platforms such as Netflix, Hulu, HBO Max, Disney+ and many others. Depending on the company policy, these internal guidelines can be easily accessed by the public; for example, the Netflix Audio Description Style Guide is available on the Netflix website. Furthermore, since these platforms make content available at an international scale, while still being developed by the same entity, these commercial guidelines are implemented throughout each country where these platforms are available.

The goal of analysing these audio description guidelines and the audio description international standard is to gain a better understanding of the human audio description process. This in-depth understanding of this process is paramount in order to be able to automate the human AD process; but also, to be able to create a framework that aims to evaluate the envisioned automated audio description process. The foundation of the audio description algorithm evaluation framework and the cornerstone of this investigation into audio description guidance and standards is ISO/IEC TS 20071-21:2015.

2. Audio Description: Guidelines vs Laws/Regulation

In the United Kingdom, the Independent Television Commission (ITC) created in 2000 a code that was aimed at offering guidance on how AD should be produced in the UK. But there are other countries which have since then published their own national guidelines; countries such as Belgium, Greece, Germany, France, Spain, Sweden, United States of America and Canada. Furthermore, pan-national projects such as the ADLAB Project were developed at the European level in an attempt to create an AD guideline that would be used by multiple European countries (Rai, Greening and Petré, 2010; Mazur and Chmiel, 2012; ADLAB, 2014).

Audio description guidelines can be divided into two categories based on their purpose and enforcement level: on one hand, we have the guidelines which are part of the national broadcasting regulations or laws, while on the other hand, we have the audio description guidelines created by third parties. Thus, in the case of the first situation, the information presented will be as a legally binding document and in most cases, it involves details about the number of hours for which a broadcaster has to provide audio description, or the types of programmes for which audio description must be provided. A good example would be national broadcasting and licensing laws that a broadcaster must respect in order to legally offer their services.

In the second category, we have the audio description documents which are not legally binding and thus are considered only as guidelines which are there to help the audio describers with the audio description process. These guidelines are usually created by third-party organizations in collaboration with broadcasting companies or broadcasting regulators. A good example would be the Described Video Best Practices guidelines which were developed at the initiative of Accessible Media INC and The Canadian Association of Broadcasters (Pearson, 2013).

3. Laws and Regulatory Bodies

The laws and regulations which govern audio description and their implementation are generally overseen by each country's national broadcasting commission. In the case of the countries from which the AD guidelines were chosen, the regulatory bodies are as follows:

- United Kingdom – The Office of Communications [1]
- Ireland – Broadcasting Authority of Ireland [2]
- Australia – Australian Communications and Media Authority [3]
- Canada – Canadian Radio-television and Telecommunications Commission [4]
- USA – Federal Communication Commission [5]
- NZ – Broadcasting Standards Authority [6]

1. United Kingdom

In the United Kingdom, the broadcasting authority responsible for regulating broadcasting services is the Office of Communications (Ofcom). In 2003, Ofcom replaced the Independent Television Commission which developed the "ITC Guidance on Standards for Audio Description" (ITC, 2000).

Currently, accessibility (subtitling, audio description, sign language) to content provided by broadcasters in the United Kingdom is regulated through provisions which are part of The Broadcasting Act 1996 (Section 20) and the Communication Act 2003 (Section 368BC). Additionally, the Code on Television Access Services published by Ofcom (Ofcom, 2021) further details the regulations that the broadcasters have to adhere to. With respect to provision of audio description, the Code (Ofcom, 2021) provides broadcasters with a list of requirements such as: a 10-year target for the amount of audio described broadcasts provided, technical difficulties that might be encountered with providing AD for certain types/genres of broadcasts,

scheduling of programming with audio description at peak times and promoting audio description through various methods (Ofcom, 2021).

2. Ireland

The Broadcasting Authority of Ireland (BAI) is the regulatory body which oversees the broadcasting services in Ireland. The BAI was established in 2009 under the authority of the (Irish) Broadcasting Act of 2009 and formerly known as the Independent Radio and Television Commission (IRTC) which was created under the Radio and Television Act of 1988. The BAI effectively replaced the Broadcasting Commission of Ireland (BCI).

The (Irish) Broadcasting Act of 2009, Section 43(1)(c), requires "[…] each broadcaster of audiovisual material to take specific steps to provide access to that material by persons who are deaf or have a hearing impairment, persons who are blind or partially sighted, and persons who have a hearing impairment and are partially sighted […]" (BAI, 2009). As a result, the Broadcasting Authority of Ireland has developed a document detailing a set of rules for broadcasters, known as the BAI Access Rules. These Access Rules contain a section dedicated to regulation of broadcast accessibility, including those regulations which concern audio description. The audio description segment provides two types of information: firstly, it provides details on the programmes where AD should be provided, popular viewing times and genres and types of broadcast. Secondly, it provides guidance on what the audio description should describe, how it should describe and when it should describe (BAI, 2016).

3. Australia

Australian media and communications are regulated by the Australian Communication and Media Authority (ACMA). ACMA is an Australian government statutory authority and part of the Communications portfolio. It was formed in 2005 from the merger of the Australian Communication Authority (ACA) and the Australian Broadcasting Authority (ABA). Regulation concerning accessibility to broadcast content in Australia is established in the 1992 Broadcasting Services Act. While there are legal provisions for providing accessibility to audio-visual content, they are focused on closed captioning and not on audio description or audio captioning. After contacting the Australian Communications and Media Authority in order to attempt to find more details about the existence of any document that would serve as an audio description guideline, I was informed that there is no such document at the present time.

4. New Zealand

The standards of broadcasting for free-to-air, pay television and radio are being upheld by the Broadcasting Standards Authority (BSA). The BSA is a New Zealand Crown entity that was created as a result of the Broadcasting Act of 1989. After contacting the BSA, I was informed that there are no audio description guidelines or any similar document at a national level in New Zealand. Nevertheless, there was one inquiry into captioning conducted by a Government Administration select committee in 2017, but nothing related to audio description / audio captioning. Additionally, I was able to discover an NGO – The Access Alliance – which is campaigning for the introduction of an Accessibility for New Zealanders Act.

5. Canada

The Canadian Radio-television and Telecommunications Commission (CRTC) is a public organisation which acts as the regulatory agency for broadcasting and telecommunications. Starting from 2001, the CRTC required a minimum threshold of audio-described programs from certain broadcasters.

This was further expanded on in 2009 with the "Broadcasting and Telecom Regulatory Policy CRTC 2009-430" which *"intends to require television broadcasters to provide high-quality audio description through conditions of licence to be imposed at the time of their licence renewals" (CRTC, 2009).*

The Broadcasting Regulatory Policy CRTC 2015-104 deals with showcasing the findings of the Commission on identifying ways to create a better future Canadian television system. In the section named "Improved access and experience for Canadians with disabilities" the CRTC mentions that "Canadians with disabilities, should have more access to accessibility features and a seamless experience when accessing their content of choice." Consequently, this means that the broadcasters will have to increase the availability of audio-described video. To be more precise, by September 2019, certain programming services / broadcasters will have to audio describe for all programming aired in the prime-time hours between 7 p.m. and 11 p.m. (CRTC, 2015).

In 2013, the Described Video Best Practices Working Group (DVBP), a voluntary initiative that was led by Accessible Media INC. (AMI) and the Canadian Association of Broadcasters (CAB) with support from the CRTC, produced the Described Video Best Practices – Artistic

and Technical Guidelines. The aim of this document was to provide guidance to producers of audio-described programming throughout Canada in an effort to achieve AD uniformity at a national level (Pearson, 2013). Furthermore, in 2015 the same Described Video Best Practices (DVBP) Committee produced the Live Described Video Best Practices document through which it aimed at providing uniformity for all audio descriptions of live audio-visual content ((AMI) and (CAB), 2015).

6. United States of America

The Americans with Disabilities Act (ADA) prohibits discrimination against people with disabilities in several areas, including employment, transportation, public accommodations, communications and access to state and local government's programs and services. Furthermore, the Federal Communications Commission consumer guide on audio description, requires TV station affiliates of ABC, CBS, Fox, and NBC located in the top 60 TV markets to provide 87.5 hours per calendar quarter (about 7 hours per week) of audio-described programming, of which 50 hours must be prime time and/or children's programming and 37.5 hours may be any type of programming shown between 6:00 a.m. and midnight. Additionally, Subscription TV systems (offered over cable, satellite or the telephone network) with 50,000 or more subscribers must provide 87.5 hours per calendar quarter (about 7 hours per week) of audio-described programming on the top five most-watched non-broadcast networks, of which 50 hours must be prime time and/or children's programming and 37.5 hours may be any type of programming shown between 6:00 a.m. and midnight (Snyder, 2007).

The America Council of the Blind through their Audio Description Project drafted the Audio Description Guidelines and Best Practices in 2010. These guidelines are intended to have an overarching character, thus covering the topic of AD in general, regardless of the subject that is being described, the format or genre the audio description is used (American Council of the Blind, 2010).

4. Choosing the appropriate guidelines for my research

The reason for performing this research into audio description guidelines was prompted by the lack of an existing overarching and internationally accepted audio description guideline which would unify national AD guidelines. Such a document was needed for my audio description automation research in order for me to be able to formulate an evaluation framework that would facilitate an objective assessment of algorithms. The aim of this part of the research is to look

at various audio description guidelines and to identify the key visual elements that must be described in order to be able to create a valid audio description script. These key visual elements together with their descriptors/characteristics will be used as key performance indicators in the envisioned algorithm evaluation framework.

The comparative study published by the RNIB which looked at AD guidelines/standards/codes created in various national broadcasting cultures (the study addresses AD documents from the UK, Greece, France, Germany, Spain, and the American Council of the Blind's ADP project ADI standards) (RNIB, 2010), showcased that there are general similarities but also minor differences between the guidelines and standards. These minor differences arise from the variations in film/television formats, the viewing methods and the cultural differences between the national broadcasting environments in each country. For example Ofcom suggests that characters should be named as soon as possible unless withholding the name is plot-critical, while ADI and DE Guidelines state that characters should be named in the first 10 minutes unless the names are revealed later in the programme. Another example is the mention of colours. For example ADI and the French Guidelines suggest that they should be mentioned with descriptive associations or adjectives, while Ofcom does not impose any restrictions on colour use in descriptions. Nevertheless, all guidelines do agree that it is important to preserve certain audio events such as noise, music and silences to maintain the movie/video's narrative flow. Additionally, not every pause in speech has to be filled with audio description, and descriptions should be focused on actions that are not understandable or that are not audible to blind or partially sighted audiences.

My research is aimed at investigating the AD guidelines from several major English-speaking countries. The reason for choosing this set of countries is based on the similarities that exist between their national broadcasting cultures. Furthermore, these countries share English as their official language, and thus AD tracks will be produced in same language thus removing the variables that are introduced by comparing the AD process based on different languages.

5. Audio Description ISO Analysis

Information Technology – User interface component accessibility – Part 21: Guidance on audio description spans 133 pages and is divided into several parts as seen in Table 9. As an ISO standard, the document covers the entirety of the audio description process. Thus, providing guidance for audio description developers that work on different stages of the audio description process: scriptwriters, voice narrators, organisations, or groups responsible for creating or

delivering audio description in order to ensure that the visual content is represented faithfully and accurately (ISO, 2015).

The goal of Part 21: Guidance on audio descriptions of ISO/IEC 20071-21 is to provide audio description practitioners with guidance on describing visual content with the intention of creating an Audio Description track. It focuses on creating content that is solely describing audio-visual material in an auditory-only modality as opposed to Part 11 (Guidance for alternative text for images) which provides guiding for image description in text-only modality (ISO/IEC TS 20071-21:2015(E), ISO/IEC TS 20071-11:2012(E)).

There are several reasons why this ISO plays a crucial role in developing the algorithm evaluation framework. Firstly, it was developed with the purpose of being an internationally applicable document. Furthermore, ISO standards are documents which are internationally agreed on by experts in the field they aim to standardize. According to the International Organization for Standardisation an ISO standard represents "[…] distilled wisdom of people with expertise in their subject matter and who know the needs of the organisations they represent […]" (ISO.org, 2021). This means that unlike the other audio description guidelines which were designed around a single country's broadcasting services, the ISO document should have been developed based on researching the audio description practice in multiple countries. This conclusion is based on the description of how ISO standards are developed on the official ISO website (ISO, 2024).

As it can be seen from the excerpt above – the ISO aims to cover audio description use in multiple contexts and environments. Furthermore, it particularly addresses the international character of this document by indicating the recommendations are applicable regardless of the language or the transmission technology used to deliver the audio description. Moreover, the ISO addresses the fact that due to time limitations that have to be respected when creating audio description tracks, the amount of information transmitted by the audio description track is unable to "[…]provide an equivalent experience[…]" (ISO, 2015, p. 1).

1. The AD ISO (ISO,2015) Importance Levels:

1. Essential

Indispensable information without which the listener would not be able to understand the visual and auditory content. In order to be considered essential, information should present some or all the of the following properties (Chapter 4.4.3, ISO/IEC TS 20071-21:2015(E)).

1. **Target audience** – the information is of use to the target audience of the content.
2. **Enhance or detail** – the information identifies visual content that directly conflicts with the dialogue or background sound.
3. **Ease of understanding** – information that needs more than a quick glance at the content in order to be understood.
4. **User importance** – most audiences need or want this type of information
5. **Comprehension level** – listener would be confused without having access to this information.
6. **Content provider** – information which content providers consider the listeners must know.

2. Significant

Is built upon the essential information and aims at enriching the audio description experience through providing more details about essential elements.

1. **Target audience** – the information is of use to the target audience of the content.
2. **Enhance or detail** – the information is essential for a thorough understanding of the content.
3. **Ease of understanding** – information can be obtain by more than, for example a quick glance at the video.
4. **User importance** – information which is important for the user to know about as they are listening to the main audio content in order to understand the video and event.
5. **Comprehension level –** listener would have an idea of the topic of the video or event, but will have a detailed understanding.
6. **Content provider** – information which further explains or offers more details on what the AD track wants to transmit to the user.

3. Helpful

Information which might be of interest to some listeners in certain situations

1. **Target audience** – the information might be of interest to the target audience.
2. **Enhances or detail** – information which would further define cinematic and background details.

3. **Ease of understanding** – information that can be used to reassure the listener that they have not missed something of greater importance.
4. **User importance**: - information without which listeners still retain a fairly complete understanding of the audio-visual content.
   - information which is optional/extra that is seldom wanted/needed to reinforce what already was presented
5. **Comprehension level** – information which can include different or other possible interpretations of the visually expressed information.
6. **Content provider** – information which could clarify some things, for some groups of people.

4. Irrelevant

Described as information that does not provide any additional understanding for the visual content.

1. **Target audience** – very few users will want to know this information
2. **Enhances or detail** – information which does not provide any detail or enhance the AD experience
3. **Ease of understanding** – without this information the user knows everything they want/need to know to enjoy the content
4. **User importance**: – not important enough to mention
5. **Comprehension level** – information which might result in creating confusion or boredom among the users
6. **Content provider** – N/A

Importance Levels	Target Audience	Enhance or Detail	Ease of Understanding	User Importance	Comprehension Level	Content Provider
Essential	The information is of use to the target audience of the content.	Identifies visual content that directly conflicts with the dialog or background sound.	Needs more than a quick glance to be understood.	Most audiences need or want this type of information.	Listener would be confused without it.	Must-know information as deemed by the content provider.
Significant	The information is of use to the target audience of the content.	Essential for a thorough understanding of the content.	Can be obtained by more than a quick glance.	Important for understanding the video and event.	Listener will have a detailed understanding.	Further explains or offers more details.
Helpful	Might be of interest to some listeners.	Further defines cinematic and background details.	Used to reassure the listener they have not missed something important.	Listeners retain a fairly complete understanding without it.	Can include different interpretations.	Could clarify some things for some groups.
Irrelevant	Very few users will want to know this information.	Does not provide any detail or enhance the AD experience.	Users know everything they need to without it.	Not important enough to mention.	Might create confusion or boredom.	Not applicable.

Table 4 Information importance levels based on the ISO Description

6. Visual information and approaches of classification

Developing an evaluation framework for assessing the fitness of algorithms to perform various audio description tasks required an in-depth analysis and breakdown of the human audio description process. Thus, building on the research and knowledge I have acquired in my first year, I continued researching and understanding how the human audio description process takes place and how it could be divided into parts that could be automated. While a human can perform various audio description tasks simultaneously, a machine would need to perform the same tasks individually.

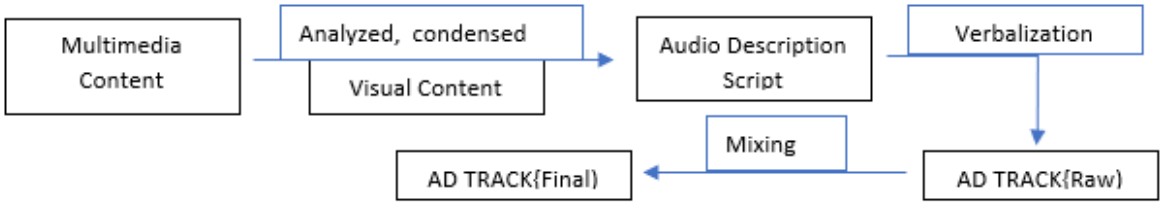


Figure 43 Human AD Process (ADLAB, 2014; Fryer, 2016)

The process of creating an audio description is similar to translating text from language A to language B. But instead of text in language A (Source Text) and the translated text in language B (Target Text), we have information that is encoded in the visual medium (Source Material) and the translated information encoded in the auditory medium (Target Material). The audio describer decodes the visual information, identifies the key information which facilitates the creation of narrative force, condenses it, and then re-encodes it in the auditory medium.

The audio description process is split into three stand-alone steps which take place in a chronological order: creating the AD script, voicing the AD script, and mixing the AD track. All these steps are valid candidates for automation, but the most difficult one is the creation of the audio description script. The reason is the fact that the information transfer takes places across two different communication channels. Furthermore, the amount of information that can be transferred through the visual channel is vastly larger than what the audio channel can transfer. It is precisely due to this reason that the audio describer has to identify and describe only those elements which are essential for transmitting the narrative intent of the audio-visual content (ADLAB, 2014; Fryer, 2016).

In order to be able to select the right elements and describe them, the audio describers rely on previously acquired information and knowledge. Furthermore, since the audio description track has to provide information in an unbiased and objective manner, audio description guidelines

are being followed in order to ensure that the information transmitted is adequate to the task. Additionally, the goal of the audio describer is not to describe everything they see in a scene, but to focus on certain key visual elements; the audio description guidelines act like a "recipe" or a list of such elements whose characteristics have to be described. Describing these key visual elements is necessary in order to create an AD track which complements the existing audio track and conveys visual information without which the user would not be able to fully understand the context of the content.

By cross-referencing several audio description guidelines with the audio description ISO/IEC TS 20071-21:2015(E), it is possible to create a list of the most important elements and characteristics that have to be included in an audio description track. Nevertheless, we always have to remember the presence of critical factors in deciding what needs to be described such as narrative cohesion and salience. Narrative salience represents the importance of certain elements or aspects of the story (narrative), while cohesion is there to ensure that these elements fit perfectly in the overall structure of the story. Identifying these elements is a skill which relies on a deep understanding of the context of the story, the presence of themes or underlying character motivations. Such a skill would be hard to automate as it relies on human judgements and complex interpretability mechanisms developed during the describer's life. Nevertheless, as mentioned in the Audio Description Script creation chapter, there are state-of-the-art technologies which are getting closer and closer to this goal.

1. Visual information categories

In subchapter three of the third chapter of ISO/IEC TS 20071-21:2015(E), we are provided with an overview of the whole audio description process, starting with the narrator preparation and ending with a short mention of evaluation. Unfortunately, the evaluation advice only revolves around including consumers in the process, but no frameworks for evaluation are provided (ISO, 2015, p. 1). During the audio description process, the creation of the audio description script represents the cornerstone that will heavily influence the quality of the final audio description track. This subjective nature of visual perception is what warrants the creation of audio description guidelines, courses, or ISO. The purpose of all these documents is to offer guidance through that audio describers can create audio description experiences which present information as objectively as possible.

In an article from 2004 (Piety, 2004), Piety investigates understanding audio description as a language system. In order to achieve this, they takes an in-depth look at what types of

information are presented in audio descriptions. Piety (2004) has based their categorisation on Halliday's (1987) description of functional grammar which included: participants, processes, and circumstances (Piety, 2004).

Appearance	The external appearance of a person, place, or thing.
Action	Something in motion or changing.
Position	The location of description or of characters
Reading	Written or understood information that is literally read, summarized, or paraphrased.
Indexical	An indication of who is speaking or what is making some sound.
Viewpoint	Relating to text-level information and the view as a viewer.
State	Not always visible information, but known to the describer and conveyed in response to visual information

Table 5 Piety, 2004, p.459

Such an approach can prove beneficial for both the development of the evaluation framework, but also for automating the creation of the audio description script. The following table contains non-exhaustive list of possible benefits which can enhance evaluation and/or automation of the audio description process.

Characteristic	Evaluation	Automation
Consistency / Completeness (Salience and Cohesion)	By using a standardised framework we can ensure that all important elements of a scene are covered and thus ensure that the salient information is consistent in the resulting audio description.	Such a standardisation creates a structure which can then be used in automation.
Systematisation / Structure	Breaking down the description in a initial overarching hierarchy based on Halliday's functional grammar, we can create a systematic approach when categorising and understanding / rationalising the different elements which compose a scene.	Since descriptions are initially approached from three sides: the participants (who or what), the processes (what is happening) and circumstances (context of the process/participants), they can serve as a blueprint for a automated system which should increase the coherence

		and contextual quality of the generated descriptions.
Key Elements/ Prioritization	The above categorization ensures a focus on identifying the key components without which the narrative intent or narrative context of a scene would be lost.	By focusing on these overarching categories, an automated system should be able to prioritize the information that is transmitted to the listener.
Aide Natural Language Processing		Well-defined grammatical structure benefits Natural Language Processing, thus it could be an aide in parsing and generating textual descriptions which are closely aligned to human-created descriptions.
Evaluation Criteria	This initial structure serves as a base which can be expanded and built upon in order to obtain more detailed sub-criteria, thus increasing the granularity of our evaluation framework.	

Table 6 Benefits of having a Standardised Approach of AD

Based on the information that Piety (Piety, 2004) provides in the article, it would appear that depending on the creator of the descriptor categorisation/taxonomy or the reason for its creation, there can be discrepancies in what categories of representations are included in a document/guideline. Furthermore, when analysing various AD guidelines, we can notice that although they are describing the same process and aim to convey the same information, the way in which this information is categorised can vary from one document to another.

Another aspect that can influence the taxonomy of the descriptor categories is represented by the period in which the document was published. For example, in the 2000 ITC Guidelines – chapter 3.10 discusses Colours / Ethnic origins in the same entry, unlike the 2015 AD ISO which provides a more granular description of these characteristics by addressing them in separate entries (ITC, 2000; ISO, 2015). I believe that such differences can signify an evaluation of the Audio Description standards, which could suggest that the practice of audio description is maturing and becomes more sophisticated and tailored to the needs of its audience. Additionally, discrepancies between guidelines allow comparative research and

integration of various approaches in order to cover more areas in one document. Furthermore, by being able to understand these variations, we might be able to develop audio description automation systems or evaluation frameworks which could be customised to needs/preferences of certain regions or audio description providers. Granularity can also increase the inclusivity and accuracy of the descriptions ensuring that the descriptions are more precise and inclusive.

7. Enhancing the Visual Information Categories using the ISO

As we have previously seen, the ISO 2015 guidelines highlight a structured approach with respect to determining and describing the important information from a scene. Each of the four levels (Essential, Significant, Helpful, Irrelevant) is defined by a specific set of criteria which we can map to the categories identified in the ISO, but also in Halliday's approach.

Categories	**Essential**	**Significant**	**Helpful**	**Irrelevant**
Participants (Characters and Objects)	Major characters and objects crucial to the understanding of the scene (e.g., the protagonist's actions, a key object that drives the plot).	Secondary characters and details that enrich the understanding (e.g., a character's facial expression, attire that signifies status or emotion).	Background characters or objects that add depth but are not crucial (e.g., a crowd in a marketplace).	Details that do not add to the understanding or enjoyment (e.g., minor background details with no narrative impact).
Processes (Actions and Events)	Actions and events that drive the main plot (e.g., a character escaping from danger).	Actions that add depth to the narrative but are not central (e.g., a character performing a routine task that adds to their development).	Minor actions that provide additional context (e.g., a character's casual movements).	Actions that do not affect the narrative (e.g., a character fidgeting without narrative significance).
Circumstances (Settings and Context)	Settings critical for understanding the scene (e.g., a battle taking place in a	Environmental details that enrich the scene (e.g., the time of day,	Background elements that provide additional atmosphere (e.g.,	Details that do not contribute to the narrative (e.g., minor architectural details).

	historical landmark).	weather conditions that affect the mood).	decorations in a room).	

Table 7 Mapping Importance Levels to Halliday's Approach

8. Taxonomy of visual descriptors

The algorithm evaluation framework will first address what categories of visual information have to be described. In the context of this taxonomy, categories refer to distinct classes / elements / groups of visual information that have to be described. Each of these categories would encompass a type of visual element that contributes to the overarching narrative for the audio-visual content.

As mentioned before, the ISO will be used as a starting point for developing this taxonomy. Starting from a standard published by the International Organisation for Standardisation will offer several key advantages such as:

Standardisation and Consistency: ISO standards are globally recognised and ,due to their nature, can provide a uniform structure which increases consistency in our taxonomy.

Comprehensive: Typically ISO standards cover various aspects of the subject/field in which they are developed. We can develop a robust taxonomy which will address all necessary components and considerations of the automated audio description process.

Quality: By starting from an internationally recognised standard we can improve the quality of our taxonomy by incorporating best practices and proven methodologies. Additionally, we increase the likelihood of the taxonomy being accepted by a wider audience.

Additionally, the taxonomy shall be enhanced by introducing information from other audio description documents/guidelines which are subject to this research.

There are two cases that can occur when comparing the ISO taxonomy with the information presented in the audio description guidelines:

- The compared documents reference the same information, and thus a correlation can be created and the choices be justified.
- The compared documents do not all reference the same information and thus a decision has to be taken regarding the information.

The ISO offers a list of entities that have to be described alongside their main features which play an important part in creating the narrative force of the audio-visual content. These are paramount to the audio description process because they are the main method through which the creator/author of the audio-visual content is transmitting the information and also creating an engaging viewing opportunity.

Levels of importance	Text / sound description	Identifying objects	Identifying persons	Guidance on Relationships
Essential	Sounds	Objects	Characters/places by name	Explicit content
Significant	Logos/Credits/Titles	Colour	Physical appearance	Relationships
Helpful	On-screen text	Visual Effects	Race/Ethnic origins	Place/setting/time of day
Irrelevant			Gender-related	Interactions animated characters/objects and real actors
			Disabilities	
			Age	

Table 8 ISO taxonomy for describing visual information

There are several types or categories of information, detailed in Table 3, that can be meaningful and thus must be conveyed through audio description. As we can observe, these are the elements carrying information which creates the story or narration of the movie including but not limited to, characters, places, and objects. Furthermore, there are meta elements which find themselves outside of the movie or content, for example: titles, credits, various types of navigation menus which allow the viewing of a certain scene of the movie.

Meta elements	Elements
Titles	Characters
Credits	Objects
On-screen text that conveys information about the production	Location / Scenes
Sounds	Time of the day
Production related Videos	Visual effects / Sound Effects

Table 9 Description elements – Based on ISO taxonomy

There are many methods in which the elements that have to be described can be grouped. One of these can be based on what Piety (2004) defines as appearance. To paraphrase him, appearance is essential to description; this is due to the fact that without the appearance of something, there can be no representation of that something to be conveyed through description. Furthermore, that representation of appearance is what provides defining characteristics which can be described, such as, luminance, colour, size, and shape (Piety, 2004, p. 457). As a direct consequence, there are categories such as objects, characters, time of the day, location, visual and sound effects which are defined by fixed characteristics. Furthermore, there are those elements which are harder to identify, as they do not solely rely on visual characteristics. One good example would be the case of relationships between characters, that rely on information which has already been presented and on the capacity of the viewer to recall that information.

Observable (which can be heard/seen)	Inferred
Objects Characters Visual effects	Character relationships Underlying meaning of actions

Table 10 Categories based on Piety's definition of appearance

Another way of categorizing what has to be described is by using the levels of importance of information presented in the AD ISO. One caveat of this approach, which is mentioned in the ISO, is the fact that information importance levels *"*[…] are determined by the persons responsible for developing the audio description (e.g. content provider, script writer, and narrator) and will vary between different contexts" (ISO, 2015, p. 14).

Such a classification allows creativity to play an important part in the audio description process; with the small disadvantage that audio describers rely on experiences and observations accumulated before their training in order to identify information. Thus, information importance levels are heavily influenced by the subjective lens through which the audio describer filters the visual information. This is the main reason for which audio description guidelines and training documentation are created. These documents offer the audio describer a list of elements that they should focus their attention on in order to ensure that the audio description track is adequate (Fryer, 2016).

9. National guidelines analysis

As national guidelines are being created by the broadcasting authority of each country, their presentation format is not standardised. Furthermore, depending on the history of audio description in a country, the guidelines can be more detailed containing a breakdown of the audio description process, or they can be less granular, combining several aspects of the audio description in one chapter (Mazur and Chmiel, 2012) As previously explained, although we are dealing with documents which come from English-speaking countries, there will still be differences between these documents. For example, some of the documents might have been published at a later date as the audio description practice in that country has matured. This can also impact the approach and granularity of the information presented in the documents, due to the changing requirements in audio description, which may come both from the audiences but also from the broadcasters. In this subchapter, we will look at several documents from the United Kingdom, Ireland, and the United States of America, as well as Canada. Additionally, although it is not a national guideline nor legislation, we will also look at the document produced by Audio Description: Lifelong Access for the Blind ADLAB. Although this document did not cover Anglophone countries, I believe the cross-national aspect of the document makes it indispensable for my research. The goal of this analysis is to attempt to identify information which could be used to further enhance our taxonomy.

1. United Kingdom

The ITC Guidance on Standards for Audio Description is a stand-alone document that was created by the Independent Television Commission (ITC) in the year 2000 in accordance with the (UK) Broadcasting Act of 1996. Its aim was to […] provide guidance on standards for the production and presentation of audio description […] (ITC, 2000, p. 2).

The ITC Guidelines document spans 38 pages and it is divided into 7 chapters, out of which 3 chapters (2,3 and 4) [Table 11] are focussed on an in-depth description of the audio description process:

- Chapter 2 - The principles of audio description using practical examples – covers the entire audio description process , starting from choosing the content that must be described to reviewing the finished audio description track. [Table 11, column 1]

- Chapter 3 - The principles of audio description using practical examples – addresses the description of the visual information. This chapter covers a vast number of issues that must

be addressed by the audio describer, such as but not limited to tense usage, information prioritisation, and the use of adjectives and adverbs, among others. [Table 11, column 2]

- Chapter 4 - Programme categories – tackles the issue of audio-visual genre by illustrating the various changes that the audio describer has to take into account when creating audio description for a certain genre type. Some of the given examples include musicals, feature films, sports and live events, comedy, children's programmes and several other. [Table 11, column 3]

ITC Guidance on Standards for Audio Description		
THE PREPARATION OF AN AUDIO DESCRIPTION	THE PRINCIPLES OF AUDIO DESCRIPTION USING PRACTICAL EXAMPLES	PROGRAMME CATEGORIES
Choosing Suitable Programmes for Description	Use of the Present Tense	Feature Films
Viewing the Programme	Prioritising Information	Musicals
Preparing a Draft Script	Giving Additional Information	Soap Opera .
Reviewing the Script	Signposting or Anticipating the Action.	Nature Documentaries
Adjusting the Programme Sound Level	Stating the Obvious	Current Affairs Documentaries
Recording the Description	Highlighting Sound Effects	Sport and Live Events
Reviewing the Recording	The Use of Proper Names and Pronouns	Foreign Language Drama
	Adjectival Descriptions	Foreign Language Material in Britain and Smaller European Countries.
	Use of Adverbs	Children's Programmes
	Colours / Ethnic Origins	Comedy
	Use of Verbs	Sexually Explicit or Violent Programmes
	Logos and Opening Titles	Advertisements / Programme Trailers / Product Placement
	Cast Lists / Credits.	

Table 11 ITC Guideline

2. Ireland

In contrast to the ITC Guidelines, the BAI Audio Description Guidelines are presented as part of a broader document (i.e., BAI Access Rules), span 7 pages (pp 52-59) and only offer a broad overview of the audio description process. The BAI further mentions that it acknowledges the work conducted by the former European Audetel (Audio Described Television) Consortium and the assistance of Ofcom's assistance on preparing this document (BAI, 2019, p. 52).

The BAI Guidelines on Audio Description are divided into 8 chapters which deal with issues ranging from what/when to describe to grammar & language and content genre. Due to the vast difference between the number of pages of the ITC guidelines and the BAI AD guidelines, the second document does not offer nearly as much information as its UK counterpart. Nevertheless, is does, in its eight chapters, tackle most of the issue that were present in the ITC guidelines but offers only minimum explanation for each of them (BAI, 2019; ITC, 2000).

A good example of this condensation of guidelines can be seen in the first three chapters which deal explicitly with the actual description of the video information. Unlike the ITC guidelines which offer an in-depth breakdown for each category, the BAI AD guidelines only enumerate and give a simple indication for each of the categories (BAI, 2019; ITC, 2000). Thus, the first three chapters presented as:

- Chapter 1 – What to describe – the introduction mentions that "the following chapter is a summary of the elements of a programme which should be described" (BAI, 2019, p. 52). The main goal is to introduce the main visual elements that should be described: characters, locations, time of day, on-screen action, sounds or sound effects, subtitles captions and opening titles and/or end credits (BAI, 2019, p. 53).

- Chapter 2 – When to describe – this chapter addresses matters concerning when in the original audio track should the audio description be inserted (BAI, 2019, p. 54).

- Chapter 3 – What not to describe – this chapterillustrates the audio description practices that have to be avoided, such as: introduction of the describes personal views/opinions (e.g., motivation/reasoning for an action) and that editing/continuity mistakes should not be replicated in the audio description (BAI, 2019, p. 54).

- Chapter 4 & 5 address the programme sound level of the audio description track and with the process of recording the track (BAI, 2019, p. 55).

- Chapter 6 discusses various aspects of grammar, but at a superficial level unlike the ITC guidelines (BAI, 2019, p. 56).

- Chapter 7 addresses the notion of information prioritisation; and while it does not offer a taxonomy for the levels of information importance like the ITC guidelines or AD ISO, it does provide practical examples that illustrate situations which can be equated with the ITC taxonomy (BAI, 2019, p. 57).

Broadcasting Authority of Ireland Access rules							
What?	When?	What not?	Sound Level	AD Recording	Grammar & Language	Information Prioritisations	Genres
Characters		describer personal opinion			present tense present continuous		Soap Opera
Locations		unseen actions/events			complete sentences		Current affairs documentaries
Time of the Day					proper names have to be used		Sporting and live events
On-screen Action					descriptive adjectives =/= describer personal view		Children's programming
Sounds/Sound effects							
Subtitled Captions							
Opening Titles and/or End Credits							

Table 12 BAI Access Rules

3. United States of America

As the ITC guidelines represent one of the most in-depth AD guidelines both from the point of view of its length and content, it comes as no surprise that there were other national guidelines that were based on it besides the BAI Guidelines. Another good example is the Audio Description Guidelines and Best Practices published by the American Council of the Blind in 2010 (American Council of the Blind, 2010). The document spans 98 pages and addresses not only AD for broadcasting, but also AD in performing arts such as theatre, dance, and opera. The key chapter which contains the information concerning the visual elements that have to be identified and which have to be described can be found in the Core Skill section.(American Council of the Blind, 2010).

Unlike the previously discussed guidelines, this guideline (American Council of the Blind, 2010) is not divided into chapters, but rather into sections; each addressing a different aspect of the AD process. As such, the document covers the following topics: (American Council of the Blind, 2010)

- it provides a definitions section that acts a glossary of terms and concepts related to AD
- a core skills section covers the key aspects of identifying and describing visual elements (American Council of the Blind, 2010, pp 9-20)
- the performing arts section addresses AD in the context of theatre, dance, and opera (American Council of the Blind, 2010, pp 21-42)
- the fifth section provides more details on movie description (American Council of the Blind, 2010, pp 43-50)
- the last section discusses AD in the context of visual arts / exhibitions (American Council of the Blind, 2010, pp 51-63)

The sections that address the information that is relevant to my research can be found in the core skills section and in the media section. These two sections of the AD guidelines (American Council of the Blind, 2010) are the ones which deal with key elements that an audio describer has to consider when creating an audio description: who to describe, what to describe, when/where to describe, and how to describe.

Quintessentially, each of these audio description guidelines describes the same fundamental audio description process. A good example is the overarching structure of the process: creation

of the AD script, recording, and mixing. These steps are fundamental to creating an audio description track and are present in all audio description guidelines. Nevertheless, these guidelines were developed to be used at an internal/national level, therefore, each contains information valid only within its specific national context. As a result, the underlying structure of the audio description guideline may vary from country to country based on these specificities. To achieve my goal of creating an algorithm evaluation framework, it was imperative to analyse the audio description guidelines and extract the information presented in them in a more structured manner.

Audio description guidelines and best practices - American Council of the Blind			
Who?	What?	When/Where?	How?
Age	General to Specific	Time of the day	Clear, concise, consistent, conversational
Hair/Build/Clothing	Colour	Passage of time	Point of view and Narrative tense
Relationships	Directional information	Location	Consider audience
Characters/people	Actions		"We see" – to be avoided
	Expressive gestures/movements		Vary verb choices
	Specificity		Definite/indefinite articles
	Less is More		Pronouns
	Logos/Credits		Multiple meanings
			Interpretive [TL]Adverbs/Gerunds: -ly words & -ing words
			Objectivity
			Metaphor/Simile – Shapes/Sizes/ other essential attributes
			Labels
			Censorship

Table 13 ACB - AD guidelines and best practices

4. Canada

In Canada, the major audio description document is presented under the title Described Video Best Practices: Artistic and Technical Guidelines (Pearson, 2013). This guideline was the result of a voluntary initiative led by Accessible Media INC (AMI) and the Canadian Association of Broadcasters (CAB) with support from the Canadian Radio-Television and Telecommunications Commission (CRTC). During the development of this document, audio describers and producers of audio description came together with representatives of the community group to attempt to standardise audio description/described video and bring context to a practice that combines both art and science. According to the authors the aim of this document was to provide guidance to producers of described programming across Canada in an effort to achieve uniformity (Pearson, 2013).

The scope of these Best Practices was limited and thus they are intended for use only in certain situations such as: English content only, Canadian-produced content, and be used in the post-production step. Furthermore, the document acknowledges that AD is a field that combines both the artistic and the technical. While both of these elements were discussed in the Best Practices, the main focus of the authors was on the artistic elements while also introducing some technical elements (Pearson, 2013).

The authors approach the visual elements that have to be included in an audio description script from two perspectives: a holistic overview of the visual information, and a detailed classification (artistic and technical guidelines) of the visual information based on identifying specific visual elements. Thus, in chapter 4 we are introduced to the idea of multi-level description (Pearson, 2013, p. 9); an idea that we have already encountered previously, but under a different name, when discussing the levels of importance presented in the ISO.

Multi-level description	
Primary descriptions	Descriptions that are absolutely crucial to the understanding of story development
Secondary descriptions	Descriptions that are defined as being important but not absolutely essential to the understanding of story development
Tertiary descriptions	Stylistic descriptions that are encouraged when time allows for them.

Table 14 Multi-level description according to the Described Video Best Practices

As shown from the table below, the categorization of information is extremely similar to that provided by the ISO; with the difference that it does not provide the same granular details as

described by the importance levels. Nevertheless, the same core idea of categorizing visual elements based on their narrative importance is present in both documents (Pearson, 2013; ISO, 2015).

After discussing the holistic approach in analysing and classifying visual information, the authors adopt a more granular approach to identifying the key visual elements which have to be described. One important aspect of the Best Practices is the fact that it provides instructions on how to use the visual information classification provided in the artistic guidelines subchapter (Pearson, 2013, p. 11). The above-mentioned artistic guidelines are divided into four sections, as detailed in the table below:

Topic	General grouping of multiple sub-topics of the same category.
Sub-topics	Specific elements of consideration for inclusion in all development.
Recommendations	Specific recommendations to facilitate the implementation of the sub-topics.
Techniques	Specific techniques to facilitate the application of the sub-topic recommendations.

Table 15 Artistic Guidelines format section details

The information that is essential to be described can be found in the topic and sub-topic sections. As the main goal of this chapter is to investigate audio description guidelines in order to establish the key visual elements which have to be described based on cross-referencing multiple AD guidelines, the focal point of this guideline analysis will be the topic and sub-topic elements of each entry of the guideline. There are six topics that comprisethe artistic guidelines (Pearson, 2013, p. 12).

Table 16 Topics according to the Artistic Guidelines

Individual/Physical Characteristics
Scene Transitions
Visual Effects
Non-Verbal sounds/Communication
Titles, Subtitles, Credits, Text on Screen, Signing
Style and Tone

Each of these entries contains a list of sub-entries [Table 16] together with situational information about when or how they can be used and techniques/methods of describing each topic by using its sub-topics (Pearson, 2013, p. 13). By comparing these topics with the

taxonomy used in the ISO, it can be observed that both follow a similar overarching method of grouping the information. (Pearson, 2013; ISO, 2015). For example, both discuss the importance of pinpointing the individual/physical characteristics of characters/persons present in a scene. Furthermore, since, as previously mentioned, both documents provide an information importance level/ranking, these physical characteristics can be further classified based on their importance.

Described Video Best Practices – Visual					
Individual/Physical Characteristics	**Scene Transitions Visual**	**Visual Effects Non-Verbal**	**Non-Verbal Sounds/Communications Titles,**	**Titles, Subtitles, Credits, Text on Screen, Signing**	**Delivery & Narration**
Race	Establishing Place/Time of Day	Colour	Identifying Relevant Objects, Information, Circumstances, Locations, Time and Action	Titles	Point of View and Tense
Ethnicity/Ethnic Origin	Passage of Time	Dancing and Choreography	Identifying Sounds and Sound Effects and the Placement of Descriptions	Subtitles	Descriptive Verbs, Types of Language
Identifying Characters/People by Name/Physical Appearance	Transitions and Time Changes	Lighting	Working with Music and Respecting the Soundtrack	Signing	Definite vs Indefinite Articles
Facial/Physical Expression	Scene Changes	Setting	Foreshadowing	Any Text on Screen	Repetition

Relationships		Respect the Content of the Program in its Usage and Placement of Branded Products	Real vs Non-Real	Signage Including Logos	Description of Foreign Languages
Attire				Captions/Captioning	
Age					
Accent					
Hair					
Height/ Weight					
Sexual Orientation and Gender					
Avoid Character Objectification					

Table 17 AMI/CAB Described Video Best Practices

5. *Audio Description: Lifelong Access for the Blind AD*LAB

The ADLAB project was financed by the European Union under their Lifelong Learning Programme with the aim of funding courses in higher education institutes (HEI) to train AD specialists and to develop reliable and consistent guidelines for the practice of AD. With participation of eight partners from six European countries (Belgium, Germany, Italy, Poland, Portugal and Spain), the project aimed to identify existing inconsistences in AD production methods and the provision of policies at a European level. (ADLAB, 2014) This cross-nation characteristic of the research, makes it a valuable tool in for my research into audio description.

The creators of the ADLAB guidelines have taken an approach that is different from the rest of the guidelines. Due to the nature of the project and its declared goal, this document presents itself more as a course rather than a set of defined, normative indications concerning the audio description process. Thus, the presentation of the information takes a more narrative approach and uses a combination of exemplifications and explanations. Nevertheless, the document presents almost the same information as the previously analysed documents. For example, the text discusses "[…] how the film medium guides the audience's attention to the core elements of the narrative[…]" (ADLAB, 2014, p. 10) and how understanding this process will help audio describers distinguish between "core elements" and "secondary elements" of narration. This type of importance-based classification is a concept that we already have encountered in other document analysis: such as the AD ISO and the DVBP multi-level description classification.

One disadvantage of the, mainly didactical, approach this guideline is taking is that the key information that I am interested in is scattered through descriptive paragraphs and lists of examples. Nevertheless, the information is there, but unlike the rest of the documents, it takes more time to be extracted and arranged in a more structured format that would match the other documents.

The ADLAB guidelines are structured in four chapters that span 66 pages. The main focus of this guideline analysis can be found in the second chapter titled AD Scriptwriting. To be more precise, the information will be extracted from the first two subchapters which discuss the main visual elements that have to be described (ADLAB, 2014, pp. 11-29).There are five major categories of visual elements mentioned by the ADLAB:

- characters
- spatial-temporal settings
- sound effects and music
- text on screen
- intertextual references (not always visual)

Table 18 ADLAB - Visual categories

These five broad categories form a taxonomy that is similar to those that were already encountered in previous documents such as: the DVBP topics taxonomy, ISO descriptor taxonomy or ISO element classification. Each of these five categories are further characterised by a set of descriptors which should be used to identify and describe the visual elements [Table 19], As mentioned before, due to the nature of the topic of this document, the descriptors presented by the ADLAB guideline, as illustrated in [Table 19], can be found, under different names in the rest of the documents that have been analysed: DVBP, ISO, ITC Guidance, BAI Access Rules, and ACB Guidelines.

Characters	**Spatial-temporal settings and their continuity**	**Sound effect and music**	**Text on screen**	**Intertextual references**
protagonist/antagonist	global spatial setting / local spatial setting	identify pauses	opening/end credits	aural
principal/secondary	scene settings with narrative/symbolic function	identifiable aurally/reference	title/intertitles/superimposed/inserts	visual
known/name or unknown	real/imagined	describe source of the sound	logos with text	aural-visual
authentic/fictional	new / known	diegetic / non-diegetic	text on other diegetic support	
real/unrealistic	explicit / implicit	function of sound effect	subtitles	
	well-known / familiar	simultaneous / non-simultaneous	other types	
focal / supporting	relations between settings and characters/action	lyrics vs visual info	diegetic / non-diegetic	
descriptive traits such as physical appearance,				

actions, reactions dialogues				
relationships				
new or already known				
identify unique character traits				

Table 19 ADLAB Classification of visual description

4. Proposed Audio Description Evaluation Framework

1. Creation Process

As there was no unified audio description guideline which would be globally accepted and applied when evaluating human audio description, I had to create such a tool for myself. As we discussed in this chapter, this framework is my attempt at creating a way of evaluating audio description outputs, generated audio description scripts to be more specific. We settled on evaluating the audio description script, as this is the most difficult step to automate. The other two steps of the audio description process (voicing the script and mixing the audio) can already be automated.

With this in mind, I have researched and narrowed down a list of documents that I believe are a good starting point for creating such a unified audio description guideline. Nevertheless, as I have mentioned several times, this is not a perfect solution, but it is one approach that I believe was worth investigating. By using an International Standard (AD ISO 2015) as a base for our guidelines and then building upon it by analysing several guidelines from Anglophone countries and one cross-national guideline, I do believe that I have produced a working, although far from perfect, guideline that I have used in creating my proposed audio description evaluation framework.

Another important aspect when trying to classify visual data is the fact that each of us has our own subjective view of the world. This view was shaped during our lifetime and is based on our experiences and interaction with the world around us. Furthermore, in audio description we have to extract only the information without which the scene that is described will lose its impact on the public. Nevertheless, that does not mean that we cannot try to create categories of visual information to guide the attention of the audio describer and encourage them to focus on certain aspects of the scene.

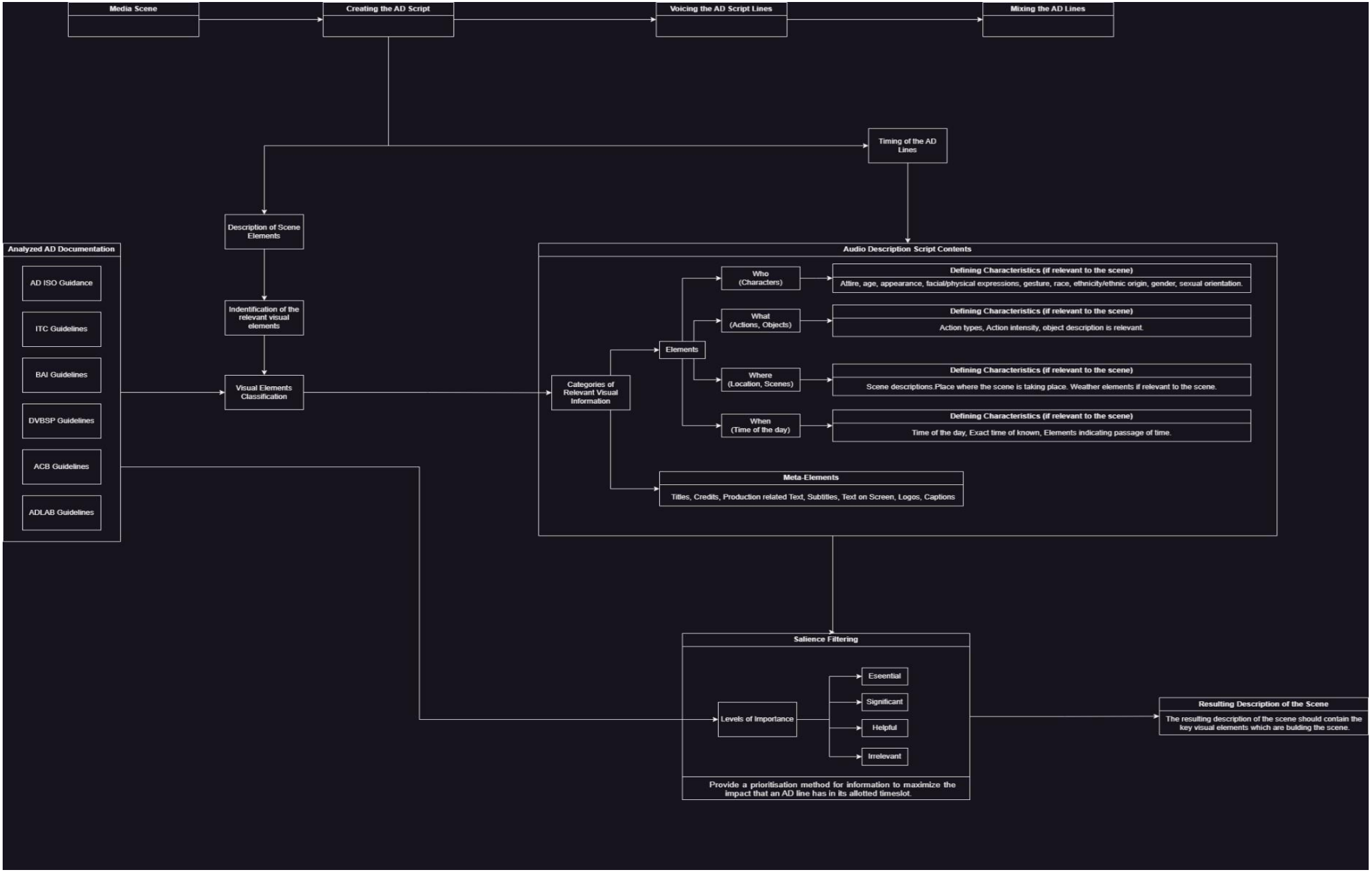


Figure 44 AD Documents Analysis Diagram

Even when investigating the selected documents, as they were developed in different countries and at different time periods, their approach on structuring visual information was different. I initially believed that, although different, these documents were aimed at describing the same process, the audio description process together with its components. As a result, the general knowledge and advice for writing audio description were broadly the same, with extremely small differences as previously highlighted. The difficulty but also the interesting part was attempting to extract and structure the information in each document in a more standardised way.

The envisioned result would be a list of visual characteristics and elements which could then be used, almost as a checklist to evaluate the description. I do want to state again that such a list is not exhaustive, I believe that such a list can always grow and become more granular based on the goals of the person who is using it. Fortunately, this also means that it is extremely flexible and adaptable to the varied usage requirements. To obtain our evaluation framework, we need to transform this list or taxonomy into Key Performance Indicators (KPIs) and then add a scoring mechanism to quantify the result.

There were two iterations for our descriptor list (list of visual characteristics and elements):

The first iteration was created around the concept of Primary and Secondary elements. This classification stems from the levels of information importance previously discussed, combined with corresponding elements found in our audio description guidelines and/or documents.

The second iteration aims to provide a more granular approach to the list by attempting to define more subcategories for each of the categories presented in the first iteration.

As mentioned before this list is extremely versatile and can be adapted to match the needs of the broadcaster/client that commissions the audio description.

2. First Iteration of Descriptor List

Primary Elements			
Character	Object	Location/Setting	Temporal/Spatial Relations
• Physical appearance • Age • Gender • Race/Ethnic origins • Disabilities • Actions/Behaviours • Emotions/Expressions	• Type/Name • Colour • Texture • Size • Shape • Function	• Location Type • Time of day • Weather conditions • Atmosphere • Mood	• Character positions • Object positions • Spatial relationships • Sequence of events • Scene transitions

Table 20 First Iteration - Primary Elements

Secondary Elements			
Abstract/Inferred Elements	Cinematic Elements	Audio Elements	Text Elements
• Character relationships • Implicit meanings/Symbolism • Tone/Mood of scenes • Cultural references	• Visual Effects • Camera Angles/Movements • Lighting • Colour Schemes/Palettes	• Sounds (non-speech) • Music • Sound effects	• On-screen text • Subtitles/Captions • Titles/Credits • Logos

Table 21 First Iteration - Secondary Elements

3. Second Iteration of Descriptor List

Characters	
Identification	Name Role / Impact in the Story
Physical / Visual Attributes	Age Race/Ethnicity Gender / Gender Expression Height and build Hair colour and style Eye colour Distinctive features Clothing etc.
Non-visual Attributes	Voice qualities Accent or dialect
Dynamic Attributes	Facial Expression

	Body Language Actions and Gestures
Character Development	Emotional States Personality Traits Motivations
Relationships and Interactions	With other characters With environment With specific objects

Table 22 Second Iteration - Characters

Settings and Environment	
Location	Indoor/Outdoor Specific location details
Time	Time of the Day Historical Period Season
Atmosphere / Environment	Weather Lighting Ambiance Mood
Cultural Context	Cultural Symbols Cultural and societal symbols

Table 23 Second Iteration - Settings and Environment

Objects	
Identification	Name and / or type Purpose and / or function
Physical properties	Shape and size Textures Colours Material Type
Significance / Symbolism	Cultural Significance or Symbolism Relevance to the story Relevance to the characters
Physical Condition	Age Dirty/clean etc.

Table 24 Second Iteration – Objects

Narrative or Thematic Elements	
Story/Plot Progression	Key event and / or turning points
Symbolism and Metaphors	Visual Metaphors Recurring motifs or symbols
Specific visual elements	Callback Foreshadowing Visual hints and references

Table 25 Second Iteration - Narrative or Thematic Elements

Actions and Movements	
Foreground Actions	Character Movement Interactions with Objects Interactions with Environment Interactions with other characters
Background Actions	Crowd movements Animas/ cars .etc passing by

	Natural Phenomena Mechanical movements

Table 26 Second Iteration - Actions and Movement

Visual Styles or Cinematography	
Camera Related	Movement such as panning, zooming etc. Angles and/or shots
Visual Effects	CGI elements Practical Effects Colour palette Lighting effects and effects Framing etc.

Table 27 Second Iteration - Visual Styles or Cinematography

Audio Components	Text and Graphic Elements
Sound Effects Ambient Noise Music Meaningful absence of sound	Captions Subtitles Signs and/or written information Graphical overlay or Logos Information Graphics / Charts Opening / Closing Elements Credits Title sequence

Table 28 Audio Components and Graphic Elements

4. Evaluation Framework

The goal of this chapter was to develop a structured approach for evaluating audio description scripts which can be used to assess the quality of any given audio description script or project. There are three components in the audio description script evaluation framework:

The Scoring Sheets – which contain a proposed scoring system for each descriptor category divided into subcategories.

The Importance Levels – which help determine what counts as important and thus has to be included.

Additional Metrics – I believe that these metrics can be used to further enhance the evaluation process.

1. Scoring Sheets

1. Character

In most narratives, the characters are the key element around which the story is revolves and progresses. As a result, I believe that they should be described in as much detail as possible.

Category	Criteria	Points
Character Portrayal (20 points)	a. Identification accuracy	0-4
	b. Physical attribute description	0-4
	c. Non-visual characteristic conveyance	0-3
	d. Dynamic element capture	0-3
	e. Character development portrayal	0-3
	f. Relationship and interaction description	0-3

Table 29

2. Environment / Setting

The environment where the action is taking place provides crucial context while also setting the atmosphere of the story, thus impacting the viewers ‘immersion and understanding of the current scene.

Category	Criteria	Points
Setting and Environment Depiction (20 points)	a. Location accuracy and detail	0-5
	b. Time element clarity	0-5
	c. Atmosphere conveyance	0-5
	d. Cultural context representation	0-5

Table 30

3. Objects

While objects can play important roles in advancing the plot or even foreshadowing certain aspects of the story, they usually require less description compared to the characters.

Category	Criteria	Points
Object and Prop Description (15 points)	a. Identification accuracy	0-4
	b. Physical property description	0-4
	c. Significance conveyance	0-4
	d. State and condition depiction	0-3

Table 31

4. Action/Movement

Category	Criteria	Points

Action and Movement Narration (15 points)	a. Character action description	0-5
	b. Background action inclusion	0-5
	c. Environmental motion depiction	0-5

Table 32

Action or lack of movement are key aspects of the narrative drive; nevertheless most of the time, only the key actions should be described.

5. Graphic Elements / Text

Category	Criteria	Points
Text and Graphic Element Inclusion (10 points)	a. On-screen text narration	0-4
	b. Graphical overlay description	0-4
	c. Opening/closing element inclusion	0-2

Table 33

Text and graphic elements can serve as important instruments in forwarding the narrative, but while they can contain essential information, their descriptions are usually brief and straightforward.

6. Audio Integration /Description

Category	Criteria	Points
Audio Component Integration (10 points)	a. Non-speech sound acknowledgment	0-4
	b. Music description	0-4

	c. Silence recognition	0-2

Table 34

Just like graphic elements or objects, audio can play an important role in transmitting the intention of the audio-visual material's producer. While audio description focuses on the visual elements of the scene, there are times when sounds have to be described, especially when they are not easily identifiable.

7. Narrative / Thematic Elements

Category	Criteria	Points
Narrative and Thematic Element (10 points)	a. Plot progression clarity	0-3
	b. Symbolism and metaphor explanation	0-3
	c. Tone and genre convention reflection	0-2
	d. Foreshadowing and callback indication	0-2

Table 35

Narrative and Thematic elements can play an important role in creating cohesion between scenes and providing clarity to the plot of the audio-visual piece. For example, symbolism and metaphors should be explained if critical to the plot.

2. *Importance levels*

Each importance level attempts to quantify the value of the information to the target audience, its role in enhancing the content, and the ease with which audience can understand it. Further information can be found in the AD ISO importance levels subchapter.

3. *Additional Metrics*

Metrics that could potentially be useful depending on use and scope of the evaluation framework.

5. Example of evaluating automated audio description script creation

To evaluate an algorithm performance in creating an audio description script we can make use of the previously mentioned Audio Description Evaluation Framework and build upon it. By adapting and extending this approach we can envision a more tailored approach.

For this example, we could, build it around some additional metrics such as: coherence, relevance, accuracy and coverage.

Coherence –(quality) – how well the described elements are inserted/respecting the logical and flowing narrative of the original piece of media. This ties directly to our discussion about narrative cohesion.

Relevance – (quality) - directly related to the importance levels that we have previously discussed. How well the algorithm described the key elements which are contributing to the narrative. We could also relate relevance to the narrative salience concept we have previously discussed.

Coverage – (quantity) - how well the algorithm covers and identifies all the relevant visual elements.

Accuracy – (quantity) - how correct is the description in depicting what is transmitted by the visual elements.

Timing – (quality) - how correct is the timing of the audio description when compared with the moments that it is describing.

We can even go a step further and include additional evaluation aspects, such as:

User Feedback – (quality) – feedback from actual users part who are part of the target audiences in order to evaluate usability and the overall satisfaction.

Word Choice / Writing Style – (quality) – the language and the writing style are appropriate for the intended audience and the type of content.

Depending on the direction to which we steer defining our metrics and criteria, we can create an evaluation framework which would can continuously be adapted in order to match our evaluation requirements.

5. Research conducted in this chapter

One key aspect of attempting to automate the audio description process is the method through which the algorithms that will be used in the automation process are chosen. In order to be able to perform this assessment, an evaluation methodology was needed. As there was no existent evaluation methodology for the purpose of assessing how well algorithms would perform in the context of audio description automation, a new methodology had to be developed to proceed with my research; as detailed in both the envisioned creation process section and the creation process section.

The focus of this research, as outlined in this section of the document, was the first step of the audio description process, namely the creation of the audio description script. This part of the process represents the starting point for the creation of an audio description track. A trained audio describer watches and analyses the footage to be audio described and decides which information is important from a narrative standpoint. There are several reasons for performing this importance classification of visual information:

- Condense/summarise the visual information to provide only the most important elements of the visual content. This ensures that the audience receives the necessary information for following the narrative.
- Ensures that the audio description track offers information which complements the existing audio track and provides value to the listener. Helps with avoiding overwhelming the audience with too many or useless details
- Ensure consistency and quality across production by using a standardised approach to cover multiple platform or post-production processes

- Efficient usage of the limited time in which audio description can be used to convey information without impacting the natural flow of the existing audio track.
- Quality control by instating a structured classification system which can aide with making sure that all critical elements are consistently covered, leading to a higher-quality audio description.
- Having information classified can help maintain the coherence and flow of the story by making sure that the key events, plot points, and character actions are clearly described and communicated.
- By including significant contextual details, audio description can enrich the understanding of the narrative without overshadowing the main plot

1. Visual elements

After analysing several audio description guidelines ([1], [2], [3], [4], [5], [6]) I can conclude that there are there are two major categories of visual information: visual elements, and visual meta-elements as can be observed in this document analysis diagram. Thus, we have visual information that is part of the narrative of the audio-visual content – the visual elements, and those elements which are present outside the narrative of the content – the visual meta-elements.

In the case of visual elements which are essential for the narrative of the audio described content, my analysis of the AD documents mentioned in the previous chapter revealed that there are three major categories crucial for creating an audio description track. These are: characters - what/who performs the action, location - where the action takes place , time - when is the action occurs. These three categories are needed in to create an audio description track which will fit in the existing narrative context and complement the existing audio track. From these categories, we can expand and add subcategories that can increase the granularity of our visual information classification as seen in the Descriptor List Iterations 1 and 2. On top of this structure, we have the overarching information importance classification used as a filter to select only the relevant descriptor from the list depending on the requirements of the scene that is being described.

Another important aspect that the document analysis has revealed is the fact that all these visual categories come alongside a series of descriptors that aid in identification and description, such as, in the case of characters, where we have: age, attire, appearance, facial expressions, meaningful gestures/motions, gender etc. There are as many ways of describing a scene as there

are describers, because everyone interprets information in a slightly different manner. Nevertheless, these descriptors presented in the guidelines serve to help and focus the attention of the audio describer on meaningful elements and their characteristics.

2. Quality and quantity

From my research I have concluded that there are two major aspects that have to be taken into account when analysing the audio description process: on one hand we have the quantity of visual information available, while on the other hand, we have the quality of it. While humans have the innate ability to make use of their previous knowledge to identify visual elements and establish relations between those elements, such inferential capabilities are still hard to automate. Nevertheless, it should be possible to evaluate an algorithm based on the quantity and the quality of its results.

The quantity of identified visual information is important because the more information we can extract, the more information is available to be used and transformed into a description of said visual information. Quality of identified visual information plays an important role in ensuring that what is described is, in fact, consistent with the visual information presented in the footage on which the algorithm is being tested.

The quality and quantity of the information extracted from the footage are two of the most important evaluation criteria for evaluating audio description scripts. Nevertheless, these are but of the dimensions that we have to take into consideration, we also have the narrative cohesion and salience of the descriptions. The goal is to use the quality and quantity of visual information to our advantage and analyse automatically generated audio description scripts based on their ability to highlight the most important and relevant elements that drive the story forward while also making sure that they contain a level of cohesion which ensure that the descriptions can seamlessly integrate into a coherent narrative.

5. Audio Description Automation Proof of Concept

The following chapter is split into two subchapters:

Chapter 1: Automated Audio Description Script Generation Workflow – provides more details on automated system from a workflow perspective. The goal of this chapter is to provide insight into how the visual information could be transformed into a textual representation and is split into three sections:

1. Processing the media
2. Extracting the Scene information
3. Generation and Insertion of descriptions

Chapter 2: System code Analysis – dives deeper into the code which allows the automation of the steps previously described. The goal of this chapter is to provide better understanding of the necessary processes and also the flow of information, from the visual to the textual. It provides insights into how each of the scripts interact with the information and with each other. Each section is divided into three parts: Information Diagram, Code Explanation, and Code.

1. Automated Audio Description Script Generation Workflow

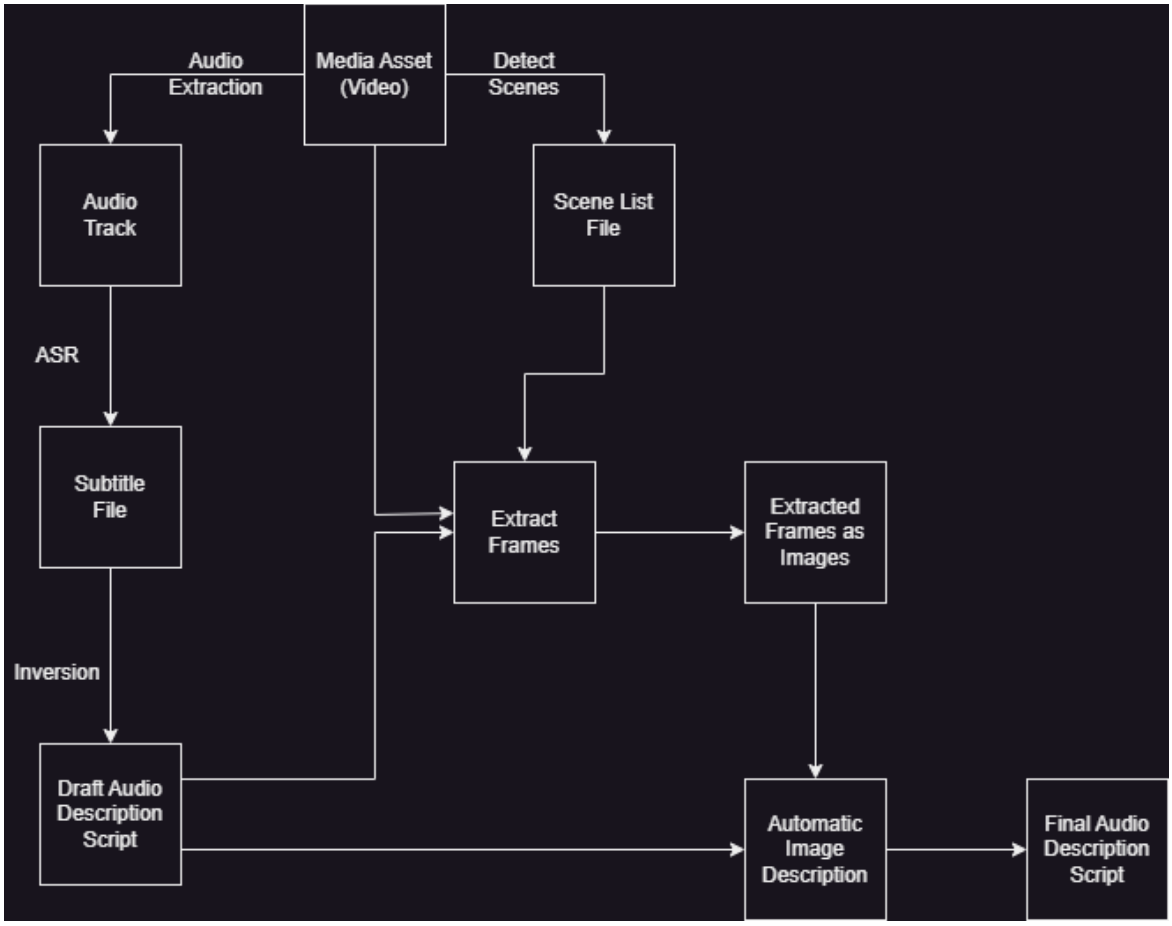


Figure 45 Script Creation Process Diagram

1. Process the media file:

The first step is to process the media file and extract the audio tracks from the video. This is done because the audio track is much smaller than the video file and thus easier to manipulate and process. While most Automatic Speech Recognition (ASR) systems can take in a video file, such files can take much longer to be transfered and processed compared to transferring just the audio file.

1. Identify areas where audio description can be introduced (no speech present)

ASR can be used to extract and convert speech into text. ASR providers such as, Whisper, Google ASR, Speechmatics, Amazon ASR etc. are able to generate a subtitle file that

matches the timing of the speech. Once we have the subtitle file, we can run a simple script to generate the audio description draft script with empty l (Figure 1):

i. Detect where the subtitles already exist.
ii. Analyse the in/out cues of each existing title.
iii. Use the in/out cues of the existing titles to identify where the gaps in speech are present.
iv. Generate new empty titles which correspond to each speech gap in the original subtitle file.
v. Remove the existing subtitles from the original subtitle file.

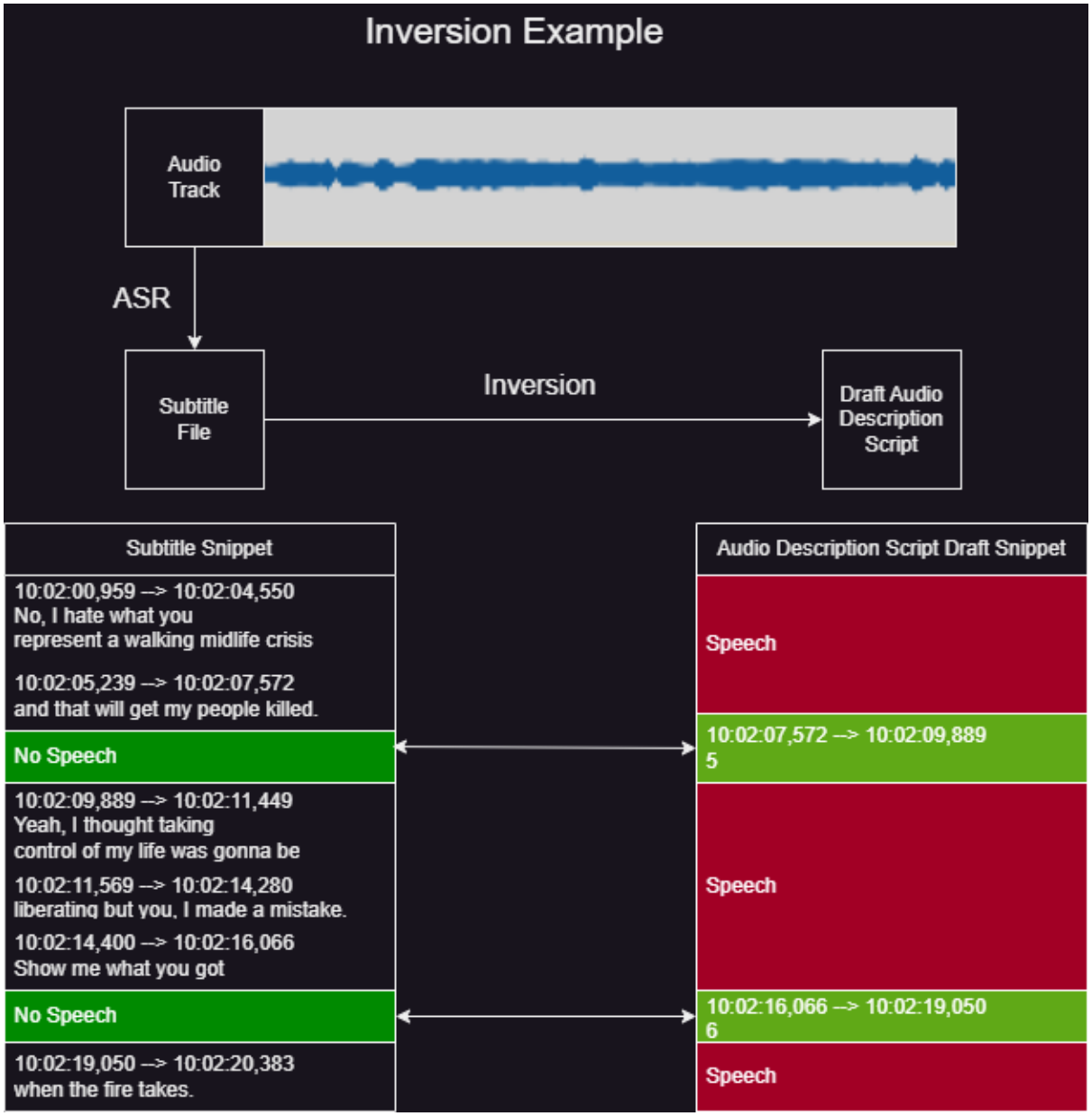


Figure 46 Inversion Example

The resulting file contains text/title boxes timed to the appropriate areas of the video where there is no speech. Each of these empty/draft title boxes represents an area in which one or more descriptions can be written. (Table 1)

Additionally, each title contains an ID number starting from 1, which will be used later in the process to match the image description with the corresponding title box.

In cases where the media is already paired with a subtitle file, instead of processing the audio using ASR, we can use the existing subtitle file to generate the Audio Description Script Draft by simply inverting the subtitle.

Subtitles – Speech	Audio Description Slot – No Speech
Incue (a) **Outcue (b)** 10:01:42,049 --> 10:01:44,049 and no idea what happens next.	
	Outcue (b) **Incue (c)** ↓ ↓ Incue Outcue 10:01:44,049 --> 10:01:49,199 4
Incue (c) **Outcue (d)** 10:01:49,199 --> 10:01:51,189 Unicorn. He's laughing at me.	

Table 36 Relation between subtitle timing and audio description line timing

Now that we have analysed the audio of the media asset and established where the audio description could be introduced, we will next analyse the visual information.

2. List the scenes in the video

In this step, we are going to generate a scene list which contains all the scenes from the media asset.

In order to generate this list, we are using a simple script which analyses each frame of the video, comparing the similarities between consecutive frames.

This similarity value is used as a threshold to determine when a scene is recorded. This function outputs a .csv file containing the in and out cues and length for each scene. (Figure 3)

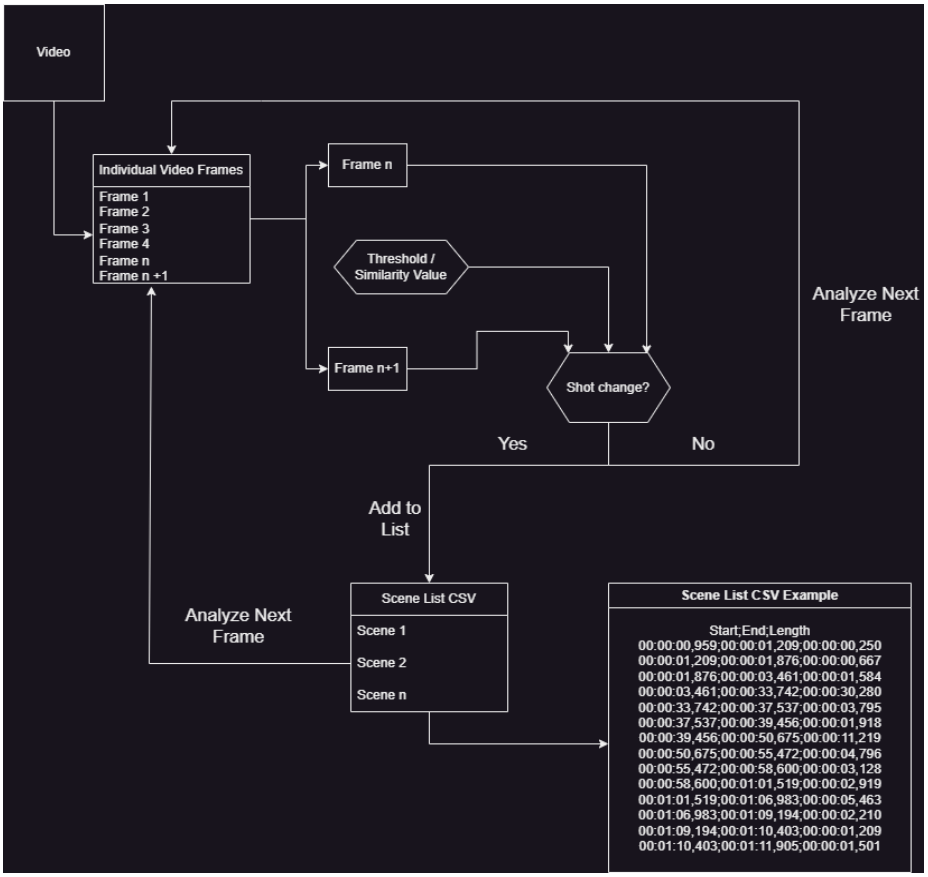


Figure 47 Scene Extraction Process

2. Extracting the scenes

Now that we have a list of scenes from the video, we can proceed to extract one frame from each scene. Since each scene is made up of similar frames (based on the threshold we have set in the previous step) which contain similar visual information, we will extract one frame as a standalone image. Each scene will have an associated frame/image, which will then be used to generate a description for that scene in the following steps.

1. *Basic Extraction*

This method will extract a frame from each of the scenes described in the scene list csv file. This was the first iteration of the extraction system, and it does not take into consideration the intervals where there is speech. As a result, it will output one image for each scene in the media asset.

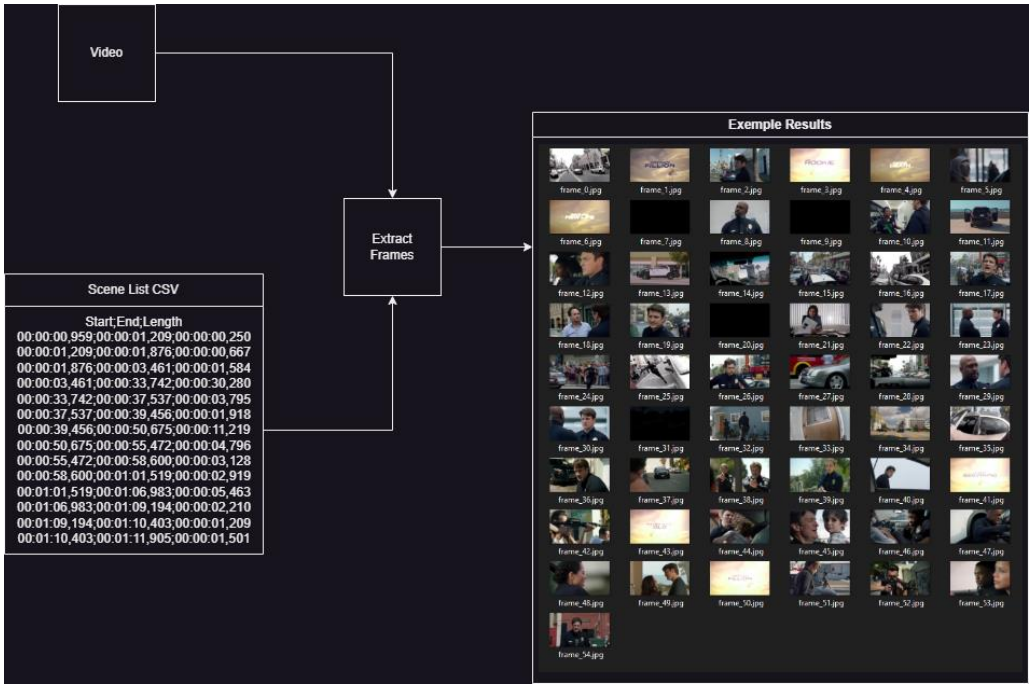


Figure 48 Simple Frame Extraction Process

2. *Advanced Extraction*

This approach uses the original video, the scene list file, and the draft script file to extract the frames. Unlike the first approach the introduction of the draft script file means that it will only extract frames if their start times fall within the intervals defined by the draft script file. While the first approach, can be used to test the description capabilities by extracting all scenes and describing them, the advanced extraction method is the process that is mapped on the human process of scene selection.

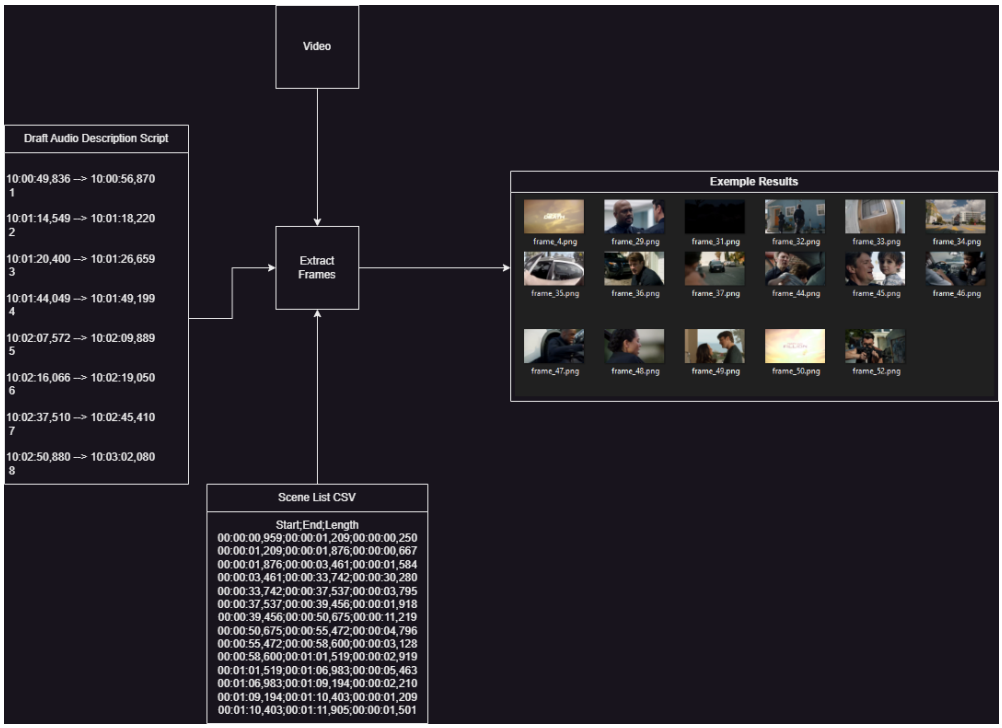


Figure 49 Advanced Frame Extraction Process

3. Generating and inserting the image descriptions in the Audio Description Script Draft

The last step in the automated creation process of the Audio Description Script is split into two parts:

b. Generating the image descriptions
c. Inserting the descriptions at the right points in the AD Script Draft

1. Generating descriptions

We use a script which iterates over the path where the images have been saved in the previous step. The script takes the images one by one and describes them using one of the two proposed description methods:

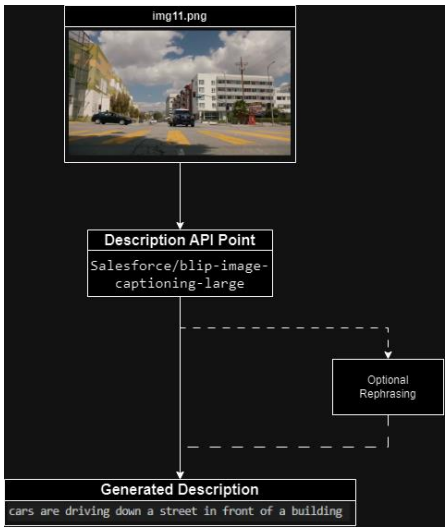


Figure 50 Example of generating an image description from a single source

a. Simple Description

The images are passed one by one to the chosen description API endpoint. Currently, the script is using APIs available on the HuggingFace platform which makes machine learning models available for the public to test. In our case, we use models that are available in this list, https://huggingface.co/models?other=image-captioning (for the full list, please check the Description chapter in the Code Analysis part). Each of these API's will return a response with the description of the image.

This method is best suited when the model used to describe the image is trustworthy. Unfortunately, it does mean that the model's bias and specificity might impact the quality of the resulting output.

Optional Rephrasing can increase the readability and accessibility, especially for situations where audiences may not be familiar with the sometimes mechanical-

sounding outputs of AI image captioning models. This rephrasing is done by using ChatGPT's API and instructing it to rephrase or check if there are any grammatical mistakes. Additionally, rephrasing can help adjust the length of the description offered by the description model in situations where the description is too long to fit in the audio description slot in the original audio.

2. *Advanced Description*

In the simplified approach presented above, we rely on a single description model, in order to increase accuracy, we can make use of several models and then attempt to build a consensus between the results.

Instead of using a single description/captioning model to produce a description of the extracted frame, we used several caption models, each of them returning a description. We use n-grams to determine the common phrases across each of the descriptions generated for a scene, with the goal of generating an accurate description by merging these frequent phrases.

The output will be comprised of the combined description based on common phrases, all descriptions from the specified API endpoints, and the common phrases themselves, which will then be rephrased using the OpenAI API. There are several advantages of using such a multi-model approach:

a. Multiple Interpretations: By accessing different AI models for image captioning, this approach is not confined to the interpretation or limitations of a single model's description. Each model may have been trained on different datasets or optimized for different aspects of image recognition, thus leading to varied perspectives when generating an image description.
b. Comprehensive Descriptions: This approach can capture a wider range of details, from basic elements to nuanced aspects of the image, offering a more comprehensive description than a single model might provide.
c. Consensus: By identifying the most common phrases across descriptions, we attempt to select the most agreed-upon elements depicted in the image, potentially enhancing the accuracy and reliability of the final description.
d. Rephrasing: As in the previous method, rephrasing can enhance the combined description by introducing a linguistic refinement step. This can transform a mechanically combined sentence into a more coherent and elegantly structured narrative, which can fit the allocated audio description slot in the original audio track.

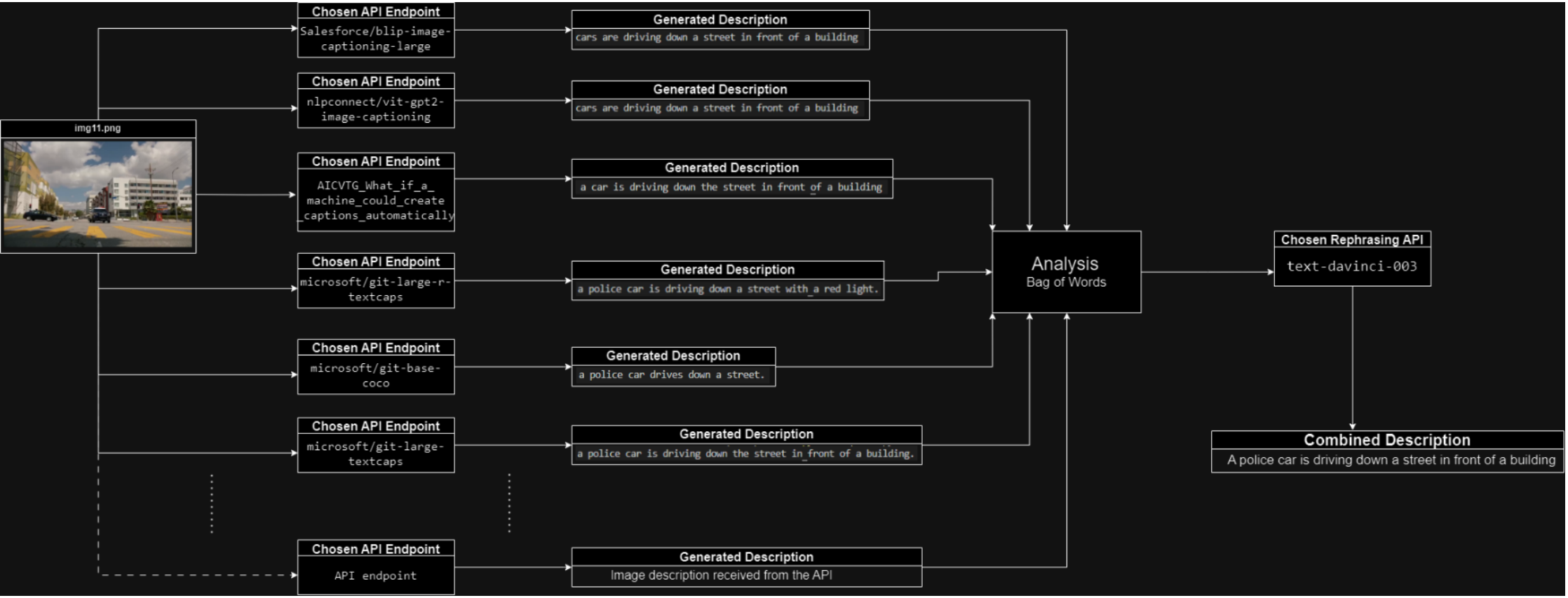


Figure 51 Example of generating an image description by using multiple sources

3. *Inserting the Description*

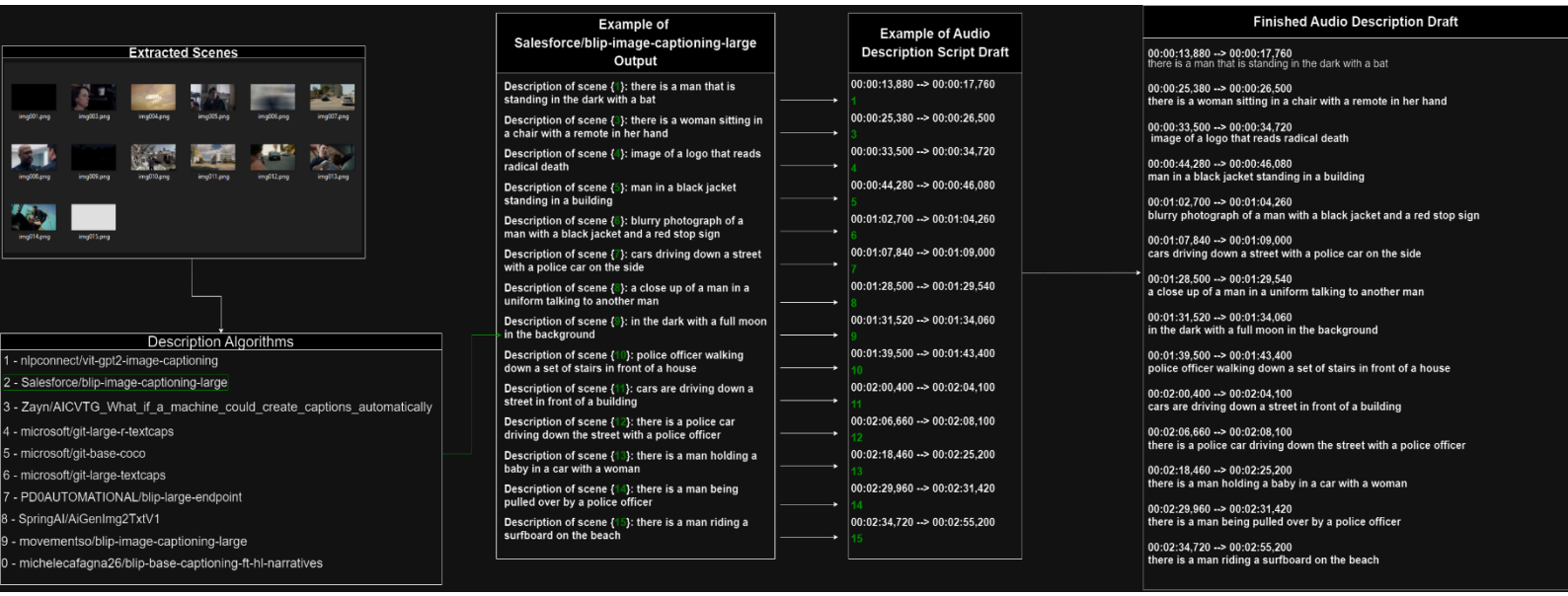


Figure 52 Example of inserting the descriptions into the AD Script Draft

To create the final draft of the audio description script, we use a Powershell script which inserts the descriptions in the appropriate title. This Powershell script iterates through each image present in the Extracted Scenes path, and for each image it:

a. Extracts the scene number from the image filename.
b. Runs the above-mentioned description method for each image.
c. The output description is captured and displayed in the console.
d. For each description obtained, it updates the corresponding line (marked in a specific format with a green font colour) from the Audio Description Draft with the new corresponding description.
e. Finally, it saves the Final Audio Description Draft file as a new subtitle file.

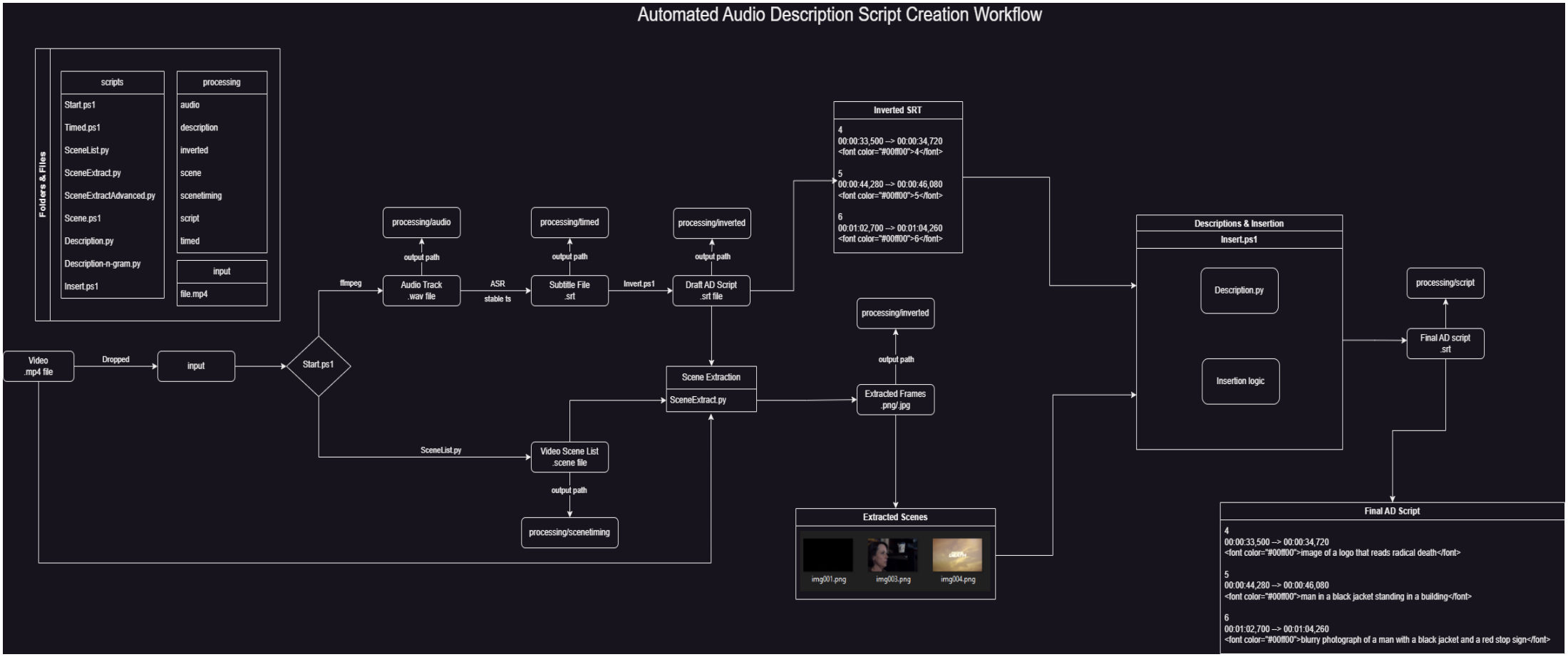


Figure 53 Automated Audio Description Script Creation Workflow

2. System Code Analysis and Explanation

1. Start.ps1

This part of the code represents the starting point of the system.

1. Code explanation

This script creates a system watcher using PowerShell which will monitor the /input path for any .mp4 files which are placed this path. Once a .mp4 file is detected, it will trigger a FFmpeg process which will extract the audio track from the video. The .wav audio track will then be placed in the /processing/audio path.

SET FOLDER TO WATCH + FILES TO WATCH + SUBFOLDERS YES/NO

1. $watcher = New-Object System.IO.FileSystemWatcher: Creates a new instance of the FileSystemWatcher class, which can monitor file system changes.
2. $watcher.Path = "input": Sets the directory to monitor to "input". This is where the script will look for changes.
3. $watcher.Filter = "*.mp4*": Sets the filter to only watch for changes to files that match the pattern *.mp4*, which means any file with .mp4 in its name.
4. $watcher.IncludeSubdirectories = $false: Configures the watcher to not monitor subdirectories of the specified path.
5. $watcher.EnableRaisingEvents = $true: Enables the watcher to begin raising events. This starts the monitoring process.

DEFINE ACTIONS AFTER AN EVENT IS DETECTED

1. Defines several script blocks (anonymous functions) to be executed in response to different file system events:
2. $log = { ... }: Defines actions to log information about deleted or renamed files. It captures the path, change type, and name of the file, and then writes this information to a log file named validatorlog.txt.

3. $validate = { ... }: Defines actions to validate and potentially rename the files. If a file name contains spaces, they are replaced with underscores. Otherwise, it simply renames the file to its current name, effectively making no change.
4. $extraction = { ... }: Defines actions to extract audio from the .mp4 files using ffmpeg. It changes the file extension to .wav and moves the extracted audio to a specified output folder.
5. $move = { ... }: Defines actions to move the .mp4 files to a specific directory named "processing\video" after certain events.

DECIDE WHICH EVENTS SHOULD BE WATCHED

Registers event handlers for the FileSystemWatcher to specify which actions to execute for different types of file system events:

1. Register-ObjectEvent $watcher "Created" -Action $extraction: Executes the $extraction script block when a file is created.
2. Register-ObjectEvent $watcher "Created" -Action $validate: Executes the $validate script block also when a file is created.
3. Register-ObjectEvent $watcher "Changed" -Action $move: Executes the $move script block when a file is changed.
4. Register-ObjectEvent $watcher "Deleted" -Action $log: Executes the $log script block when a file is deleted.
5. Register-ObjectEvent $watcher "Renamed" -Action $move: Executes the $move script block when a file is renamed.

2. *Code*

```
### # SET FOLDER TO WATCH + FILES TO WATCH + SUBFOLDERS YES/NO
$watcher = New-Object System.IO.FileSystemWatcher
$watcher.Path = "input"
$watcher.Filter = "*.mp4*"
$watcher.IncludeSubdirectories = $false
$watcher.EnableRaisingEvents = $true

### # DEFINE ACTIONS AFTER AN EVENT IS DETECTED
$log = {
    $path = $Event.SourceEventArgs.FullPath
    $splitp = Split-Path $Event.SourceEventArgs.FullPath -Leaf
    $changeType = $Event.SourceEventArgs.ChangeType
```

```
    $logline = "$(Get-Date), $changeType, $splitp $path"
    Add-Content "validatorlog.txt" -Value $logline
}

$validate = {
    $leaf = Split-Path $Event.SourceEventArgs.FullPath -Leaf
    if ($leaf -match '\s') {
        $splitp = $leaf.Replace(' ', '_')
        Rename-Item -Path $Event.SourceEventArgs.FullPath -NewName $splitp
    }
    else {
        Rename-Item -Path $Event.SourceEventArgs.FullPath -NewName $leaf
    }
}

$extraction = {
    $path = $Event.SourceEventArgs.FullPath
    $outputFolder = "processing\audio"
    $outputFile = Join-Path -Path $outputFolder -ChildPath
($Event.SourceEventArgs.Name -replace '\..*?$', '.wav')
    $ffmpegArgs = "-i `"$path`" -vn -acodec pcm_s16le -ar 44100 -ac 2
`"$outputFile`""
    Start-Process -FilePath "ffmpeg" -ArgumentList $ffmpegArgs -NoNewWindow -
Wait
}

$move = {
    $path = $Event.SourceEventArgs.FullPath
    Move-Item -Path $path -Destination "processing\video"
}

### DECIDE WHICH EVENTS SHOULD BE WATCHED
Register-ObjectEvent $watcher "Created" -Action $extraction
Register-ObjectEvent $watcher "Created" -Action $validate
Register-ObjectEvent $watcher "Changed" -Action $move
Register-ObjectEvent $watcher "Deleted" -Action $log
Register-ObjectEvent $watcher "Renamed" -Action $move


while ($true) {
    Start-Sleep 5
}
```

2. Timed.ps1

This part of the system will create a transcript of the extracted audio.

1. Code explanation

This script creates a system watcher using PowerShell which will monitor the /processing/audio path for any .wav files which are placed this path. Once a .wav file is detected, it will trigger as stable-ts command which will perform an ASR job on the .wav file. The resulting transcript will be moved into the /processing/timed path.

SET FOLDER TO WATCH + FILES TO WATCH + SUBFOLDERS YES/NO

1. $watcher = New-Object System.IO.FileSystemWatcher: Creates a new instance of the FileSystemWatcher class to monitor file system changes.
2. $watcher.Path = "C:\Users\CristianP\Desktop\compliance\processing\audio": Sets the directory to be monitored. This path points to a specific folder on the user's desktop.
3. $watcher.Filter = "*.wav": Configures the watcher to monitor only .wav audio files.
4. $watcher.IncludeSubdirectories = $false: Specifies that only the root directory should be monitored, not its subdirectories.
5. $watcher.EnableRaisingEvents = $true: Enables the watcher to start raising events, effectively starting the monitoring process.
6. # DEFINE ACTIONS AFTER AN EVENT IS DETECTED
7. $syncHash = @{}: Initializes a hashtable to track the completion of actions. This is used to synchronize subsequent steps.
8. $actionCompleted = { ... }: A script block that marks an event's action as completed by setting a flag in the $syncHash.
9. $log = { ... }: Defines actions to log information about changes to .wav files, including the event type and file path. This block marks itself as completed using $actionCompleted.
10. $transcribe = { ... }: Transcribes the audio file to a subtitle format (.srt) using stable-ts[1]. After transcription, it marks the transcription action as completed.

11. $move = { ... }: Waits for both the log and transcribe actions to complete before moving the processed audio file to a different directory. This script block uses the $syncHash to check for completion and then clears the hash for the next set of events.

DECIDE WHICH EVENTS SHOULD BE WATCHED

1. Register-ObjectEvent $watcher "Created" -Action $transcribe: Attaches the transcription action to the Created event. When a new .wav file is created in the directory, it will be transcribed.
2. Register-ObjectEvent $watcher "Changed" -Action $log: Attaches the logging action to the Changed event. Changes to .wav files trigger logging and content moderation.
3. Register-ObjectEvent $watcher "Deleted" -Action $log: Also attaches the logging action to the Deleted event. Deletion of a .wav file will be logged.
4. Register-ObjectEvent $watcher "Renamed" -Action $move: Attaches the move action to the Renamed event. Renaming a .wav file triggers its relocation after the necessary actions are completed.

2. Code

```
### # SET FOLDER TO WATCH + FILES TO WATCH + SUBFOLDERS YES/NO
$watcher = New-Object System.IO.FileSystemWatcher
$watcher.Path = "C:\Users\CristianP\Desktop\compliance\processing\audio"
$watcher.Filter = "*.wav"
$watcher.IncludeSubdirectories = $false
$watcher.EnableRaisingEvents = $true

### # DEFINE ACTIONS AFTER AN EVENT IS DETECTED
$syncHash = @{}
$actionCompleted = {
    $eventName = $Event.SourceEventArgs.Name
    $syncHash[$eventName] = $true
}


$log = {
    $path = $Event.SourceEventArgs.FullPath
    $splitp = Split-Path $Event.SourceEventArgs.FullPath -Leaf
    $changeType = $Event.SourceEventArgs.ChangeType
    $logline = "$(Get-Date), $changeType, $splitp $path"
    Add-Content "compliance\whisperlog.txt" -Value $logline
    $actionCompleted
}
```

```
$transcribe = {
    $path = $Event.SourceEventArgs.FullPath
    $splitp = Split-Path $Event.SourceEventArgs.FullPath -Leaf
    $outputFileName = $splitp -replace '\.wav$', '.srt'
    Start-Process cmd.exe -Wait "/k stable-ts $path -o
processing\timed\$outputFileName"
    $actionCompleted
}

$move = {
    $path = $Event.SourceEventArgs.FullPath
    $splitp = Split-Path $Event.SourceEventArgs.FullPath -Leaf
    while (-not ($syncHash.ContainsKey('transcribe') -and
$syncHash['transcribe'] -and $syncHash.ContainsKey('log') -and
$syncHash['log'])) {
        Start-Sleep -Milliseconds 500
    }
    Move-Item -Path $path -Destination "triggered\audio\$splitp"
    $syncHash.Clear()
}

### DECIDE WHICH EVENTS SHOULD BE WATCHED
Register-ObjectEvent $watcher "Created" -Action $transcribe
Register-ObjectEvent $watcher "Changed" -Action $log
Register-ObjectEvent $watcher "Deleted" -Action $log
Register-ObjectEvent $watcher "Renamed" -Action $move

while ($true) { Sleep 5 }
```

3. Inverted.ps1

This part of the system will invert the previously generated subtitle file.

1. Code explanation

This script will process .srt subtitle file which has been placed in the /processing/timed in order to generate a .srt file which contains only subtitles block which correspond to silent parts in the audio. Each silent part will have a corresponding subtitle which contain an index.

Parameters

The script starts by defining parameters that it requires to run:

1. $sourceSrtFile: The path to the original .srt subtitle file that needs to be processed. This parameter is mandatory.
2. $destinationSrtFile: The path where the modified subtitle file, which includes the added silent period markers, will be saved. This parameter is also mandatory.
3. $silenceThreshold: An optional parameter that specifies the minimum duration of silence required to insert a new subtitle entry. The default value is 1 second.
4. Loading the Original SRT File
5. $originalLines = Get-Content $sourceSrtFile: This line reads all the lines from the original .srt file into an array, $originalLines.

Initializing Variables

The script initializes a few variables to keep track of the previous subtitle's end time ($previousEnd), a counter for the new subtitle entries ($counter), and an array to store the lines of the new subtitle sections ($emptySubtitle).

Processing Each Line

1. The script iterates through each line of the original subtitle file with a for loop. For each line, it checks if the line contains a time interval using a regular expression match.

2. If a time interval is found, it parses the start and end times of the subtitle segment into [timespan] objects, replacing commas with periods to match the expected format.
3. It then calculates the gap between the current subtitle segment's start time and the end time of the previous subtitle. If this gap is greater than or equal to the specified silence threshold, it means there's a significant silence to mark.
4. For each identified silence, the script adds a new subtitle entry to the $emptySubtitle array. This entry includes a sequential number ($counter), the calculated time interval of the silence, and a placeholder text (in this case, the counter value within a green font tag) to mark the silence visually.
5. After processing each time interval, the script updates $previousEnd with the current segment's end time and increments the counter.
6. The loop also includes logic to skip over the actual subtitle text and only process the timecodes and gaps between them.
7. Outputting the Modified SRT File
8. Finally, the modified subtitle entries stored in $emptySubtitle are written to the specified destination file using Out-File $destinationSrtFile.

2. Code

```
param (
    [Parameter(Mandatory = $true)]
    [string]$sourceSrtFile, # Path to the original srt file

    [Parameter(Mandatory = $true)]
    [string]$destinationSrtFile, # Path to the inverted srt file

    [Parameter(Mandatory = $false)]
    [timespan]$silenceThreshold = [timespan]::FromSeconds(1) # Minimum silence
duration to be included in the output. Default is 1 second.
)

# Load original SRT file
$originalLines = Get-Content $sourceSrtFile

# Initialize variables
$previousEnd = [timespan]::FromHours(0)
$counter = 1
$emptySubtitle = @()
```

```
# Process each line
for ($i = 0; $i -lt $originalLines.Count; $i++) {
    # Check if the line contains the time interval
    if ($originalLines[$i] -match "(\d{2}:\d{2}:\d{2},\d{3}) -->
(\d{2}:\d{2}:\d{2},\d{3})") {
        # Parse the time interval
        $start = [timespan]::ParseExact($Matches[1].Replace(',', '.'), 'c',
$null)
        $end = [timespan]::ParseExact($Matches[2].Replace(',', '.'), 'c',
$null)

        # Check if there is a gap between the previous subtitle and this one
        $gap = $start - $previousEnd
        if ($gap -ge $silenceThreshold) {
            # There is a gap and it's longer than the silence threshold, add a
subtitle section
            $emptySubtitle += "$counter"
            $emptySubtitle += ($previousEnd.ToString('hh\:mm\:ss\,fff') + " --
> " + $start.ToString('hh\:mm\:ss\,fff'))
            $emptySubtitle += "<font color=`"#00ff00`">$counter</font>" #
Replace "Silence" with $counter
            $emptySubtitle += "" # Extra line to maintain the srt file format
            $counter++
        }

        # Update the end of the previous subtitle
        $previousEnd = $end

        # Skip the subtitle text lines
        while ($i + 1 -lt $originalLines.Count -and $originalLines[$i + 1] -
notmatch "^\s*\d+\s*$") {
            $i++
        }
    }
}

$emptySubtitle | Out-File $destinationSrtFile
```

4. SceneList.py

This script will generate a .scene file which is just a .csv file containing timing(start/end timecode) of all the scenes in the video. It does this be using OpenCV for video processing and NumPy for mathematical operations.

1. Code Explanation

Import Libraries

1. cv2: The OpenCV library for computer vision tasks, including video processing.
2. numpy (imported as np): A library for numerical operations, used here to calculate the mean of pixel value differences between frames.
3. os: A standard library module for interacting with the operating system, used to manipulate file paths.
4. datetime.timedelta: A class for representing differences between times, used to format scene change times.

Define a Helper Function

1. format_timedelta(td): Converts a timedelta object to a string formatted as "HH:MM:SS,mmm", where mmm is milliseconds. This is useful for timestamping the scenes in a human-readable format.

Initialize Video Processing

2. Specifies paths for the video file (video_path) and the output directory (output_dir).
3. Loads the video using OpenCV's VideoCapture method.
4. Initializes variables for processing: last_frame to store the previous frame for comparison, and scene_changes to record the times (in milliseconds) when scene changes occur.

Process Each Frame of the Video

1. The script enters a loop that reads each frame of the video one by one.
2. Converts the current frame to grayscale to simplify the scene change detection process.

3. If there is a last_frame to compare with, calculates the absolute difference between the current and last frames. A high average difference indicates a significant change in the scene.
4. If the difference exceeds a threshold (here, 70), records the current frame's timestamp as a scene change.
5. Updates last_frame with the current frame's grayscale version for the next iteration.

Output Scene Change Timing

1. Extracts the base name of the video file to use in naming the output file.
2. Constructs the output file path by combining the output_dir, base name of the video, and a .scene extension.
3. Writes the scene change data to the output file. For each scene change detected, it calculates and formats the start time, end time, and duration of the scene, then writes this information in a semicolon-separated format.

2. Code

```
import cv2
import numpy as np
import os
from datetime import timedelta


def format_timedelta(td):
    # Extract hours, minutes, seconds and milliseconds from the timedelta
object
    hours, remainder = divmod(td.seconds, 3600)
    minutes, seconds = divmod(remainder, 60)
    milliseconds = td.microseconds // 1000

    # Format the extracted values
    return "{:02}:{:02}:{:02},{:03}".format(hours, minutes, seconds,
milliseconds)


# The path to the video
video_path = 'C:/Users/CristianP/Desktop/compliance/Rookie.mp4'

# The output directory
output_dir = 'C:/Users/CristianP/Desktop/compliance/processing/scenetiming'

# Load the video
video = cv2.VideoCapture(video_path)
```

```
# Initialize variables
last_frame = None
scene_changes = []

while True:
    # Read the next frame
    ret, frame = video.read()

    # Break the loop if the video is finished
    if not ret:
        break

    # Convert the frame to grayscale
    frame_gray = cv2.cvtColor(frame, cv2.COLOR_BGR2GRAY)

    # If this is not the first frame
    if last_frame is not None:
        # Calculate the absolute difference between the current and last frame
        frame_diff = cv2.absdiff(frame_gray, last_frame)

        # If the difference is significant, record a scene change
        if np.mean(frame_diff) > 70:
            scene_changes.append(video.get(cv2.CAP_PROP_POS_MSEC))

    # Update the last frame
    last_frame = frame_gray

# Extract the base name of the video file (without the extension)
base_name = os.path.splitext(os.path.basename(video_path))[0]

# Create the output path
output_path = os.path.join(output_dir, f'{base_name}.scene')

# Save the scene change times to the output file
with open(output_path, 'w') as f:
    # Write the header of the CSV file
    f.write("Start;End;Length\n")
    for i in range(len(scene_changes) - 1):
        start_time =
format_timedelta(timedelta(milliseconds=scene_changes[i]))
        end_time =
format_timedelta(timedelta(milliseconds=scene_changes[i+1]))
        length = format_timedelta(
            timedelta(milliseconds=scene_changes[i+1] - scene_changes[i]))
        f.write("%s;%s;%s\n" % (start_time, end_time, length))
```

5. Scene Extraction Options

1. SceneExtract.py

This script makes use of previously generate files based on the start time listed in the .scene file (csv format) and then uses FFmpeg to extract the frames and save them as individual JPEG files.

1. Code Explanation

Import Libraries

1. os: Used for operating system-dependent functionality like path operations.
2. csv: For reading CSV files.
3. subprocess: To run the FFmpeg command as a subprocess, allowing the script to interact with the system.

Define Functions

1. time_to_seconds(time_str): Converts a time string formatted as "HH:MM:SS,mmm" (where "mmm" is milliseconds) to total seconds. This is useful for calculating the exact start time in seconds required by FFmpeg to seek to the correct position in the video.
2. extract_frames(start_times, filename, output_dir): Takes a list of start times in seconds, the path to the video file, and the output directory path. It iterates over the start times, constructing and executing an FFmpeg command for each to extract a frame and save it as a JPEG image.

Script Logic

1. output_dir: Specifies the directory where the extracted frames will be saved.
2. The script then opens and reads a CSV file containing the scene start times. The CSV is expected to use semicolons (;) as delimiters.
3. It uses the csv.reader to parse the file, skipping the header with next(reader).
4. It extracts the start times from the first column, converting them to seconds using the time_to_seconds function, and stores them in the starts list.

5. extract_frames(...): This function is called with the list of start times in seconds, the path to the video file, and the output directory. It uses FFmpeg to extract a frame at each start time and save it to the specified output directory.

Technical Details

1. The FFmpeg command constructed in extract_frames seeks to the specified start time (-ss) in the video file (-i), extracts exactly one frame (-vframes 1), and saves it as a JPEG image (output_file).
2. The subprocess.run(command) function is used to execute the FFmpeg command. It's crucial that FFmpeg is correctly installed and accessible in the system's PATH for this to work.
3. The output frames are named sequentially (frame_{i}.jpg), where {i} is the index of the start time in the list, ensuring unique filenames for each extracted frame.
4. This script is particularly useful for generating thumbnails or conducting frame-based analysis of video content, allowing for automated extraction of frames based on predefined timestamps.

2. Code

```
import os
import csv
import subprocess


def time_to_seconds(time_str):
    # Convert a HH:MM:SS,mmm string to total seconds
    hours, minutes, seconds = map(int, time_str.split(',')[0].split(':'))
    milliseconds = int(time_str.split(',')[1])
    return hours * 3600 + minutes * 60 + seconds + milliseconds / 1000.0


def extract_frames(start_times, filename, output_dir):
    for i, start in enumerate(start_times):
        output_file = os.path.join(output_dir, f"frame_{i}.jpg")
        command = ['ffmpeg', '-ss',
            str(start), '-i', filename, '-vframes', '1', output_file]
        subprocess.run(command)
```

```
# Specify your output directory here
output_dir = "C:/Users/CristianP/Desktop/compliance/processing/scene/RookiePy"

# Parse CSV
with
open('C:/Users/CristianP/Desktop/compliance/processing/scenetiming/Rookie.scen
e', 'r') as f:
    reader = csv.reader(f, delimiter=';')
    next(reader)  # Skip header
    starts = [time_to_seconds(row[0]) for row in reader]

# Extract frames from video
extract_frames(
    starts, 'C:/Users/CristianP/Desktop/compliance/Rookie.mp4', output_dir)
```

2. *SceneExtractAdvanced.py*

1. Code Explanation

This script makes use of previously generate files based on the start time listed in the .scene file (csv format) while also taking in consideration the interval defined by the .srt file present in the /processing/inverted path. The goal is to extract only the frames which occur during periods of silence in the audio denoted by the titles present in the .srt file.

Import Libraries

1. os: For interacting with the file system.
2. csv: To read the CSV file containing scene start times.
3. re: For regular expression operations, particularly parsing the SRT file.
4. subprocess: To execute FFmpeg commands for extracting frames.
5. codecs: For handling file encoding, useful for reading the SRT file which may be in a different encoding.
6. List, Tuple from typing: For type annotations, improving code readability and error detection.

Constants

Paths for the video file, the output directory, the CSV file with scene timings, and the SRT file are defined as constants.

Check/Create Output Directory

The script checks if the specified output directory exists and creates it if not.

Function: time_to_seconds

Converts a timestamp string ("HH:MM:SS,mmm") into total seconds as a float, aiding in precise frame extraction.

Function: extract_frames

1. Takes a list of start times, the video filename, the output directory path, and a list of subtitle intervals (start and end times).
2. Iterates through each start time, checking if it falls within any subtitle intervals before extracting the corresponding frame using ffmpeg.
3. This ensures frames are only extracted when subtitles are present, potentially reducing the number of irrelevant frames.

Function: parse_srt

1. Parses the SRT file to extract subtitle start and end times.
2. Uses regular expressions to find all matches for timestamps and converts them into a list of tuples containing start and end times in seconds.
3. Handles file encoding by opening the SRT file with utf-16 encoding, which is common for such files but might need adjustment based on the actual file encoding.

Main Script Flow

1. Parse CSV: Opens the CSV file to read scene start times, converting them to seconds. If the file doesn't exist, it handles the FileNotFoundError gracefully.
2. Parse SRT: Calls parse_srt to get a list of subtitle intervals (start and end times) from the SRT file.
3. Extract Frames: With the list of start times and subtitle intervals, it calls extract_frames to extract and save frames from the video. Only frames that coincide with subtitle intervals are extracted.

Execution

1. The script's main block (if __name__ == "__main__":) orchestrates the reading of start times from the CSV, parsing subtitle intervals from the SRT file, and then extracting frames based on these inputs.
2. It uses the FFmpeg tool, executed via subprocess.run, to extract frames. The command specifies the -ss option for seeking to the start time, -i for the input file, -vframes 1 to extract a single frame, and the output file path.

2. Code

```
import os
import csv
import re
import subprocess
import codecs
from typing import List, Tuple
```

```
# Constants
OUTPUT_DIR = "C:/Users/CristianP/Desktop/compliance/processing/scene/RookiePy"
CSV_PATH =
'C:/Users/CristianP/Desktop/compliance/processing/scenetiming/Rookie.scene'
SRT_PATH =
'C:/Users/CristianP/Desktop/compliance/processing/inverted/Rookie.srt'
VIDEO_PATH = 'C:/Users/CristianP/Desktop/compliance/Rookie.mp4'

# Check if output directory exists, and create it if it doesn't
if not os.path.exists(OUTPUT_DIR):
    print(f"The directory {OUTPUT_DIR} does not exist. Creating it...")
    os.makedirs(OUTPUT_DIR)

def time_to_seconds(time_str: str) -> float:
    """Convert a HH:MM:SS,mmm string to total seconds"""
    hours, minutes, seconds = map(int, time_str.split(',')[0].split(':'))
    milliseconds = int(time_str.split(',')[1])
    return hours * 3600 + minutes * 60 + seconds + milliseconds / 1000.0

def extract_frames(start_times: List[float], filename: str, output_dir: str,
subtitle_intervals: List[Tuple[float, float]]):
    """Extract frames at specific start times within subtitle intervals"""
    for i, start in enumerate(start_times):
        print(f"Checking scene {i} at time {start}")
        for interval in subtitle_intervals:
            if interval[0] <= start <= interval[1]:
                print(f"Scene {i} is within subtitle interval {interval}")
                output_file = os.path.join(output_dir, f"frame_{i}.png")
                command = ['ffmpeg', '-ss', str(start), '-i', filename, '-
vframes', '1', output_file]
                try:
                    subprocess.run(command, check=True)
                except subprocess.CalledProcessError:
                    print(f"Failed to extract frame {i} at time {start}")
                break

def parse_srt(filename: str) -> List[Tuple[float, float]]:
    """Parse an SRT file and return a list of subtitle start and end times"""
    try:
        with codecs.open(filename, 'r', encoding='utf-16') as f:
            content = f.read()
    except FileNotFoundError:
        print(f"File {filename} not found")
        return []
    pattern = re.compile(r"(\d{2}:\d{2}:\d{2},\d{3}) -->
(\d{2}:\d{2}:\d{2},\d{3})")
    matches = pattern.findall(content)
```

```
    return [(time_to_seconds(start), time_to_seconds(end)) for start, end in
matches]

if __name__ == "__main__":
    # Parse CSV
    try:
        with open(CSV_PATH, 'r') as f:
            reader = csv.reader(f, delimiter=';')
            next(reader)
            starts = [time_to_seconds(row[0]) for row in reader]
    except FileNotFoundError:
        print(f"File {CSV_PATH} not found")
        starts = []
    print(f"Scene start times: {starts}")

    # Parse SRT
    subtitle_intervals = parse_srt(SRT_PATH)
    print(f"Subtitle intervals: {subtitle_intervals}")

    # Extract frames from video
    extract_frames(starts, VIDEO_PATH, OUTPUT_DIR, subtitle_intervals)
```

3. Scene.ps1

1. Code Explanation

This script is designed to read the .srt file, extract the start and end times for each subtitle block, and use those times to capture frames from the corresponding video file using ffmpeg. This process generates images for each subtitle block, essentially creating snapshots of moments when title lines should be voiced by the audio describer.

Reading the SRT File

1. $srtPath: Defines the path to the SRT file.

2. $srtText = Get-Content -Path $srtPath -Raw: Reads the entire SRT file content as a single string, preserving the original line breaks (-Raw parameter ensures that the content is not split into an array but kept as one continuous string).

Extracting Video/Subtitle Name

1. $videoName = [System.IO.Path]::GetFileNameWithoutExtension($srtPath): Extracts the file name (without the file extension) from the SRT file path. This name is later used to create a directory specifically for storing the extracted images.

Creating a Directory for Storing Images

2. $directoryPath = "C:\Users\CristianP\Desktop\compliance\processing\scene\$videoName": Constructs a path for the directory where the images will be saved.
3. Test-Path and New-Item: Checks if the directory already exists; if not, it creates a new directory at the specified path.

Processing Each Subtitle Block

1. $blocks = $srtText -split "r?nr?n": Splits the SRT file content into individual subtitle blocks based on double newline characters, which typically separate subtitles in an SRT file.
2. The script then iterates over each block with a foreach loop.

Extracting Timestamps and Generating Images

1. For each subtitle block, the script uses regular expressions ([regex]::Match) to find the subtitle number and the time interval (start and end times).
2. The start and end times are extracted and formatted by replacing commas with periods to match ffmpeg's expected time format.
3. $imageNumber: A formatted string based on the subtitle number, padded with zeros for consistent naming.
4. ffmpeg: Invoked with parameters to seek to the start time (-ss $startTime), stop at the end time (-to $endTime), and select frames based on a scene change threshold (select='gt(scene,0.3)'). The -vsync vfr option ensures variable frame rate processing to handle frame selection accurately. The output is saved as an image in the previously created directory, with a name based on the subtitle number.

2. Code

```
# Read the SRT file
$srtPath =
"C:\Users\CristianP\Desktop\compliance\processing\inverted\Rookie.srt"
$srtText = Get-Content -Path $srtPath -Raw

# Extract video/subtitle name from srt file path
$videoName = [System.IO.Path]::GetFileNameWithoutExtension($srtPath)

# Create directory
$directoryPath =
"C:\Users\CristianP\Desktop\compliance\processing\scene\$videoName"
if (!(Test-Path -Path $directoryPath)) {
    New-Item -ItemType Directory -Force -Path $directoryPath
}

# Split by two newlines to get each subtitle block
$blocks = $srtText -split "`r?`n`r?`n"

foreach ($block in $blocks) {
    # Extract the subtitle number
    $subtitleNumber = [regex]::Match($block, '^\d+')
    if ($subtitleNumber.Success) {
        Write-Output "Block: $block"
        # Extract the start and end times
        $timeLine = [regex]::Match($block, '(\d{2}:\d{2}:\d{2},\d{3}) -->
(\d{2}:\d{2}:\d{2},\d{3})')
        if ($timeLine.Success) {
            $startTime = $timeLine.Groups[1].Value.Replace(",", ".")
            $endTime = $timeLine.Groups[2].Value.Replace(",", ".")
            Write-Output "Processing Start: $startTime, End: $endTime"
            $imageNumber = "{0:D3}" -f [int]$subtitleNumber.Value
            FFmpeg -i 'C:\Users\CristianP\Desktop\compliance\Rookie.mp4' -ss
$startTime -to $endTime -filter_complex "select='gt(scene,0.3)'" -vsync vfr
"$directoryPath\img$imageNumber.png"
        }
    }
}
```

6. Description

1. Description-n-gram.py

This script is designed to interact with API endpoints hosted on Hugging Face's servers. Each endpoint is associated to an image captioning model which returns a description if an image is sent to the correct endpoint. Additionally, it makes uses of OpenAI's GPT-3 or 4 endpoints in order to rephrase the description. Optionally, it can attempt to merge the most coming phrase into a cohesive sentence.

1. Code Explanation

Import Libraries

1. argparse for parsing command line arguments.
2. requests for making HTTP requests to the Hugging Face API.
3. Counter from collections for counting occurrences of elements.
4. CountVectorizer from sklearn.feature_extraction.text for converting text data into a matrix of token counts.
5. numpy for numerical operations on arrays.
6. openai for accessing OpenAI's GPT-3 or 4 API.

Constants

1. API_MODELS dictionary maps user input to specific model endpoints on the Hugging Face API for image captioning.
2. headers includes the authorization token for Hugging Face API requests.
3. openai.api_key sets the API key for OpenAI's GPT-3 or 4 API.

Functions

1. query(filename, api_model): Makes a POST request to the specified API model with an image file, returning the API's response as JSON.
2. get_common_phrases(sentences, topN=5): Identifies the top N most common 5-word phrases across a list of sentences.

3. merge_phrases(phrases): Attempts to merge a list of phrases into a single, coherent sentence by finding and aligning overlapping words.

Script Setup and Argument Parsing

Uses argparse.ArgumentParser to define expected command-line arguments, including input file path (--input), model(s) to use (--model), whether to rephrase captions (--rephrase), and whether to combine phrases (--combine).

Main Script Logic

1. Reads the input file path and models specified by the user.
2. Queries each selected model with the image file, collecting descriptions.
3. If rephrasing is enabled (--rephrase yes), it finds common phrases among the generated captions and uses OpenAI's GPT-3 or 4 API to rephrase them into a new description.
4. If combining phrases is enabled (--combine yes), it merges the most common phrases into a single, coherent sentence.
5. Otherwise, it prints out each generated description.

Usage Scenario

The script is invoked via the command line, where the user specifies an image file and chooses which models to query for captions. Optional rephrasing and phrase combining features can enhance the final output, especially when integrating captions from multiple models.

2. Code

```
import argparse
import requests
from collections import Counter
from sklearn.feature_extraction.text import CountVectorizer
import numpy as np
import openai
```

```
API_MODELS = {
    "1": "https://api-inference.huggingface.co/models/nlpconnect/vit-gpt2-
image-captioning",
    "2": "https://api-inference.huggingface.co/models/Salesforce/blip-image-
captioning-large",
    "3": "https://api-
inference.huggingface.co/models/Zayn/AICVTG_What_if_a_machine_could_create_cap
tions_automatically",
    "4": "https://api-inference.huggingface.co/models/microsoft/git-large-r-
textcaps",
    "5": "https://api-inference.huggingface.co/models/microsoft/git-base-
coco",  # best,
    # most accurate
    "6": "https://api-inference.huggingface.co/models/microsoft/git-large-
textcaps",
    # most accurate with background information
    "7": "https://api-inference.huggingface.co/models/PD0AUTOMATIONAL/blip-
large-endpoint",
    "8": "https://api-
inference.huggingface.co/models/SpringAI/AiGenImg2TxtV1",  # 2nd better
    # probably best one as a whole
    "9": "https://api-inference.huggingface.co/models/movementso/blip-image-
captioning-large",
    # fine tunned - great for a narration point of view, can add innacuracies
due to overinterpretation
    "0": "https://api-inference.huggingface.co/models/michelecafagna26/blip-
base-captioning-ft-hl-narratives"
}

headers = {"Authorization": "Bearer "}
# Set your OpenAI API key
openai.api_key = '' # OpenAI API key goes here
# Function to make a POST request to the API


def query(filename, api_model):
    with open(filename, "rb") as f:
        data = f.read()
   # print(f"Querying model: {api_model}")  # Debugging line
    response = requests.post(api_model, headers=headers, data=data)
   # print(f"Status Code: {response.status_code}")  # Debugging line

    if response.status_code == 200:
        response_json = response.json()
       # print(f"Response: {response_json}")  # Debugging line
        return response_json
    else:
        print("Error in query")  # Debugging line
```

```
        return None


def get_common_phrases(sentences, topN=5):
    vectorizer = CountVectorizer(ngram_range=(5, 5)).fit(sentences)
    bag_of_words = vectorizer.transform(sentences)
    sum_words = bag_of_words.sum(axis=0)
    words_freq = [(word, sum_words[0, idx])
        for word, idx in vectorizer.vocabulary_.items()]
    words_freq = sorted(words_freq, key=lambda x: x[1], reverse=True)
    return words_freq[:topN]


def merge_phrases(phrases):
    merged_sentence = ""
    for phrase in phrases:
        if not merged_sentence:
            merged_sentence = phrase
        else:
            # Split the current merged sentence and the new phrase into words
            merged_words = merged_sentence.split()
            phrase_words = phrase.split()

            # Find overlap and merge
            overlap = len(merged_words)
            for i in range(1, min(len(merged_words), len(phrase_words)) + 1):
                if merged_words[-i:] == phrase_words[:i]:
                    overlap = i
                    break

            if overlap == len(merged_words):
                # No overlap found; append the whole phrase
                merged_sentence += " " + phrase
            else:
                # Overlap found; append the non-overlapping part of the phrase
                merged_sentence += " " + " ".join(phrase_words[overlap:])

    return merged_sentence.strip()


parser = argparse.ArgumentParser(
    description="Query API models and optionally use OpenAI for rephrasing.")
parser.add_argument("-i", "--input", help="Input file path.", required=True)
parser.add_argument("-m", "--model", nargs='+', help="API models to use (1-
10).",
                    choices=API_MODELS.keys(), required=False, default="4")
parser.add_argument("-r", "--rephrase", help="Use OpenAI for rephrasing
(yes/no).",
                    choices=["yes", "no"], default="no")
```

```
parser.add_argument("-c", "--combine", help="Combine phrases into one sentence
(yes/no).",
                    choices=["yes", "no"], default="no")
args = parser.parse_args()

image_descriptions = []
for model in args.model:
    output = query(args.input, API_MODELS[model])
    if output and isinstance(output, list) and isinstance(output[0], dict):
        description = output[0].get('generated_text')
        if description:
            image_descriptions.append(description)

if args.rephrase.lower() == "yes":
    common_phrases = get_common_phrases(image_descriptions)
    combined_description = ' '.join([phrase[0] for phrase in common_phrases])
    response = openai.Completion.create(
        engine="text-davinci-003",
        prompt=f"Rephrase the following sentence: {combined_description}",
        temperature=0.6,
        max_tokens=90
    )
    rephrased_description = response.choices[0].text.strip()
    print(rephrased_description)
if args.combine.lower() == "yes":
    common_phrases = get_common_phrases(image_descriptions)
    phrases_to_merge = [phrase[0] for phrase in common_phrases]
    combined_description = merge_phrases(phrases_to_merge)
    print(combined_description)
else:
    for description in image_descriptions:
        print(description)
```

2. *Description.py*

This script is designed to interact with API endpoints hosted on huggingface.co. Each endpoint is associated to an image captioning model which returns a description if an image is sent to the correct endpoint. Additionally, it makes uses of OpenAI's GPT-3 or 4 endpoints in order to rephrase the description.

1. Code Explanation

Import Libraries

1. requests: For making HTTP requests to the Hugging Face API to query image captioning models.
2. CountVectorizer from sklearn.feature_extraction.text: Converts a collection of text documents to a matrix of token counts, used here to identify common phrases in captions.
3. numpy: For handling arrays and performing numerical operations, although its direct usage is not shown in the provided snippet.
4. openai: To access OpenAI's GPT-3 or 4 for rephrasing combined descriptions.

Constants and Configuration

1. API_MODELS: A list of URLs for different image captioning models hosted on Hugging Face's API.
2. headers: Authorization headers containing a bearer token for authenticating API requests.

Function: query

Performs a POST request to a specified image captioning model with an image file, returning the model's response as JSON. It prints the status code and response for debugging purposes.

Function: get_common_phrases

Identifies the top N most common bi-grams (two-word phrases) in a list of sentences using CountVectorizer with a ngram range of 2 to 2, highlighting frequent themes or concepts across different model outputs.

Function: generate_descriptions

Orchestrates the process of generating descriptions for a given image file by querying selected models, aggregating their descriptions, finding common phrases, and optionally rephrasing the aggregated content with OpenAI's GPT-3 or 4.

Accepts parameters to specify the image file path, which models to use, and whether to use OpenAI for rephrasing.

Collects descriptions from the specified models, extracts common phrases, and combines them. If use_openai is true, it sends the combined description to OpenAI's GPT-3 or 4 for rephrasing; otherwise, it uses the combined description as is.

Prints the rephrased description, original image descriptions, and the common phrases.

Example Usage

1. The script includes an example call to generate_descriptions, demonstrating how to specify an image file and select models for generating descriptions, with an option to disable OpenAI rephrasing.

Workflow Summary

1. The script reads an image file and queries specified image captioning models.
2. It collects and prints the generated descriptions from each model.
3. Identifies and prints common bi-grams across these descriptions.
4. Optionally rephrases the aggregated description using OpenAI's GPT-4, aiming for a more coherent or stylistically different summary.
5. Prints the final rephrased description (if OpenAI rephrasing is enabled) and the common phrases identified among the model outputs.

2. Code

```
import os
import csv
import re
import subprocess
import codecs
from typing import List, Tuple

# Constants
OUTPUT_DIR = "C:/Users/CristianP/Desktop/compliance/processing/scene/RookiePy"
CSV_PATH =
'C:/Users/CristianP/Desktop/compliance/processing/scenetiming/Rookie.scene'
SRT_PATH =
'C:/Users/CristianP/Desktop/compliance/processing/inverted/Rookie.srt'
VIDEO_PATH = 'C:/Users/CristianP/Desktop/compliance/Rookie.mp4'

# Check if output directory exists, and create it if it doesn't
if not os.path.exists(OUTPUT_DIR):
    print(f"The directory {OUTPUT_DIR} does not exist. Creating it...")
    os.makedirs(OUTPUT_DIR)

def time_to_seconds(time_str: str) -> float:
    """Convert a HH:MM:SS,mmm string to total seconds"""
    hours, minutes, seconds = map(int, time_str.split(',')[0].split(':'))
    milliseconds = int(time_str.split(',')[1])
    return hours * 3600 + minutes * 60 + seconds + milliseconds / 1000.0

def extract_frames(start_times: List[float], filename: str, output_dir: str,
subtitle_intervals: List[Tuple[float, float]]):
    """Extract frames at specific start times within subtitle intervals"""
    for i, start in enumerate(start_times):
        print(f"Checking scene {i} at time {start}")
        for interval in subtitle_intervals:
            if interval[0] <= start <= interval[1]:
                print(f"Scene {i} is within subtitle interval {interval}")
                output_file = os.path.join(output_dir, f"frame_{i}.png")
                command = ['ffmpeg', '-ss', str(start), '-i', filename, '-
vframes', '1', output_file]
                try:
                    subprocess.run(command, check=True)
                except subprocess.CalledProcessError:
                    print(f"Failed to extract frame {i} at time {start}")
                break

def parse_srt(filename: str) -> List[Tuple[float, float]]:
    """Parse an SRT file and return a list of subtitle start and end times"""
    try:
```

```
        with codecs.open(filename, 'r', encoding='utf-16') as f:
            content = f.read()
    except FileNotFoundError:
        print(f"File {filename} not found")
        return []
    pattern = re.compile(r"(\d{2}:\d{2}:\d{2},\d{3}) -->
(\d{2}:\d{2}:\d{2},\d{3})")
    matches = pattern.findall(content)
    return [(time_to_seconds(start), time_to_seconds(end)) for start, end in
matches]

if __name__ == "__main__":
    # Parse CSV
    try:
        with open(CSV_PATH, 'r') as f:
            reader = csv.reader(f, delimiter=';')
            next(reader)
            starts = [time_to_seconds(row[0]) for row in reader]
    except FileNotFoundError:
        print(f"File {CSV_PATH} not found")
        starts = []
    print(f"Scene start times: {starts}")

    # Parse SRT
    subtitle_intervals = parse_srt(SRT_PATH)
    print(f"Subtitle intervals: {subtitle_intervals}")

    # Extract frames from video
    extract_frames(starts, VIDEO_PATH, OUTPUT_DIR, subtitle_intervals)
```

7. Insert.ps1

This script can use one of the helper Python Scripts previously mentioned to generate description for scenes using the previously extracted images found in the /processing/scene/videoname path. After which it inserts the descriptions in the srt file present in /processing/inverted/ replacing the existing placeholder text.

1. *Code Explanation*

Variables Setup

Paths to the Python script, the folder containing scene images, the original SRT file, and the path for the new SRT file to be generated are defined.

Path Validation

The script checks if the specified paths for the Python script, scene images folder, and SRT file exist. If any path is not found, it outputs an error message and stops execution.

Loading SRT File Content

Reads the content of the original SRT file into an array of lines, preparing for text manipulation.

Processing Scene Images

1. Retrieves a list of all .png files in the scene images folder. If no images are found, it reports an absence of files and halts.
2. Iterates over each image file, extracting a scene number from the file's name (assuming a specific naming convention) and then running the Python script to generate a description for that image. The Python script is invoked with parameters specifying the image file, model to use for description generation (model 2 in this example), and disabling rephrasing.

Embedding Descriptions into SRT Content

1. For each image processed, it attempts to replace a placeholder line in the SRT content (identified by a specific scene number within a font tag) with the generated description.
2. The script tracks whether a match and replacement occurred for each scene. If no matching line is found in the SRT content for a scene, it reports this.

Outputting Updated SRT File

After processing all images and performing replacements in the SRT content, the script writes the updated SRT content to a new file, effectively creating an updated subtitle file with the generated descriptions embedded.

2. *Code*

```
# Variables
$descriptionPyPath = "C:\Users\CristianP\Desktop\compliance\Description.py"
$sceneFolderPath =
"C:\Users\CristianP\Desktop\compliance\processing\scene\Rookie"
$srtFilePath =
"C:\Users\CristianP\Desktop\compliance\processing\inverted\Rookie.srt"
$newSrtFilePath =
"C:\Users\CristianP\Desktop\compliance\processing\script\NewRookie.srt"

# Check if paths exist
foreach ($path in @($descriptionPyPath, $sceneFolderPath, $srtFilePath)) {
    if (!(Test-Path -Path $path)) {
        Write-Host "Path not found: $path"
        return
    }
}

# Load the srt file content into an array of lines
$srtContent = Get-Content -Path $srtFilePath

# Get all .png files
$pngFiles = Get-ChildItem -Path $sceneFolderPath -Filter "*.png"

if ($pngFiles.Count -eq 0) {
    Write-Host "No .png files found in the given directory: $sceneFolderPath"
    return
}

foreach ($file in $pngFiles) {
    $sceneNumber = [int]($file.BaseName.Substring(3, 3))

    # Capture both stdout and stderr
    $description = & py $descriptionPyPath -i $file.FullName -m 2 -r no 2>&1

    if ($description) {
        Write-Host "Description of scene {$sceneNumber}: $description"
    }
    else {
        Write-Host "Description script didn't return an output for the file:
$file.FullName"
        continue
    }

    # Replace the 'Silence' line in the srt file content
```

```
    $updatedSrtContent = @()
    $foundMatch = $false
    foreach ($line in $srtContent) {
        $newLine = $line -replace ("<font
color=`"#00ff00`">$sceneNumber</font>", "<font
color=`"#00ff00`">$description</font>")
        $updatedSrtContent += $newLine

        if ($line -ne $newLine) {
            $foundMatch = $true
        }
    }

    if (!$foundMatch) {
        Write-Host "No matching line found in the srt file for scene:
$sceneNumber"
    }

    $srtContent = $updatedSrtContent
}

# Write the updated srt file content back to the file
Set-Content -Path $newSrtFilePath -Value $srtContent
```

6. Conclusion and Further Works

1. Audio Description Evaluation Framework

The goal of my research project was to explore whether and to what degree could the process of creating audio descriptions be automated; with a heavy focus on investigating the possible automation of the creation of the audio description script. Audio descriptions are vital for making audio-visual content accessible to people who are blind or partially sighted as they enrich the experience by providing verbal explanations of the visual elements in a video, film, or live performance. By automating this process, we could dramatically increase accessibility and inclusivity in the multimedia space.

However, cross-modal transfer of information, especially between the visual channel and the auditory channel represents a challenging task. Thus, when I began this research, the technologies needed to make such automation possible such as advanced large language models (LLMs) and vision-language models (VLMs), have yet to be made available to the public. As a result of this limitation, instead of directly attempting to automate the process, the research pivoted to developing a framework for evaluating future algorithms that could handle audio description tasks. The goal was to create an evaluation framework that would serve as a tool to ensure that once the necessary technology is widely available, the quality of the automated descriptions could be properly assessed.

Despite its importance, audio description is not always available for every piece of multimedia content. Even with guidelines and legislation requiring broadcasters to provide AD for a certain percentage of their programming, unfortunately, the choice of what gets described is left up to the broadcaster, and not the audiences. This means people with visual impairments, persons with low vision and partially sighted persons do not always have access to the content they might enjoy the most.

At the core of my research, stand three guiding questions that can be boiled down to: Can the AD process be automated? Can the automated AD outputs be reliably evaluated? What would be the most important aspects to observe when evaluating an Automated AD system?

This research addressed these questions and strived to come with pertinent responses. The research has focused on the creation of an automated audio description evaluation framework which aims to serve as a tool when required software that enables AD automation is available. In order to produce the framework, I have analysed the current audio description process workflow, with a focus on the challenges and issues which arise when creating an audio description script. By leveraging existing guidelines and regulations, I was able to first envision and then create a flexible framework that can adapt to various needs and contexts, while providing a way to evaluate and qualify the outputs of an automated audio description system.

2. Automated Audio Description System Proof of Concept

While for the best part of my research I was unable to directly automate the audio description process due the technological constraints at that time, in the last years significant progress was made which started making what was one probably into something possible. As the necessary technologies that could be used to automate the audio description process became available to the public, combined with my daily work with media workflow automation, I was able to create a proof of concept system. While far from perfect, this automated AD system represents the vision that I had when I first started this research project. Additionally, such a system will also allow, in the future, the use of the Evaluation Framework to qualify the outputs and feasibility of the technologies which have been used to automatically generate the AD track.

In conclusion, while the original goal of automating the audio description process could not be realised within the timeframe of the research project, the development of an evaluation framework has been achieved. Furthermore, the creation of a proof of concept automated system which can generate an audio description script represents an important step towards the realisation of my initial goal of a fully automated audio description generator.

Automating the creation of audio descriptions will play an important role in the life of anyone who relies on audio description tracks in order to enjoy audio-visual content. It would give people more freedom to choose what they watch with audio description, shifting the control from broadcasters to the audience. As Joel Snyder (2005) put it, “A picture is worth a

thousand words? Maybe. But the audio describer might say that a few well-chosen words conjure vivid and lasting images." The advent of automation could ensure that those "well-chosen words" are available to anyone in the world, for any audio-visual content that they would want to enjoy.

3. Further works

While this research project has created an evaluation framework and made significant steps towards the automation of the audio description process, the work is far from complete. With the sudden explosion of publicly available technologies, there are, amongst many other, several promising areas which, I believe, are worth further exploration:

1. Partial automation by integration into existing software:
   One of the first options is to integrate the automated audio description script workflow into existing tools used for creating audio description. By embedding automation into already existing software, we can focus on refining the script creation aspect and utilise existing capabilities for creating audio description instead of having to build a full software solution from scratch. The goal would be to create a workflow in which the user starts with an automatically created audio description script draft, instead of starting from an empty file. This concept will be further explored in a presentation at the Languages in the Media 2024 conference, where I will discuss practical approaches to merging automation with current industry practices.
2. Adapting and Introducing New and Improved Technologies.
   As new systems capable of describing videos and images emerge, these can be incorporated into the existing proof-of-concept system. The outputs from these advanced systems could then be assessed using the evaluation framework developed in this research. The goal would be to have an iterative approach that could allow a continuous improvement of the automated system and of the evaluation framework..
3. Automating the Evaluation Process
   An interesting idea would be to automate the evaluation of the automated audio description system. This is something that would be built on the iterative approach mentioned above. By implementing such a system, it would make identifying and comparing automated systems an easier task. Furthermore, such a system would allow the

user to observe how a system's output varies based on the algorithms that are used in the audio description automation system.

4. Introducing Customisation for Users

   Another important aspect is the user experience. A possible approach would be to allow users to tailor their audio description experience to fit their needs. It could include different types of voices, maybe varied speaking speeds or even accents and information density.

The advancements in automation technologies are opening up exciting possibilities for the future of audio description. Continuous improvements and advances could be built on my proof-of-concept system, exploring integration into existing tools, and addressing new research opportunities that can push the boundaries of what is possible. This work represents not just a technical challenge but also a chance to make audio-visual content more inclusive and accessible to everyone. I strongly believe that through collaboration, innovation, and a focus on the users of audio description, the vision of a fully automated, customizable, and accessible audio description process is not that far from us

7. Bibliography

1. (AMI), A.M.I. and (CAB), T.C.A.O.B. (2015). Live Described Video Best Practices. 0–28. Available from https://www.ami.ca/sites/default/files/2020-07/Live_Described_Video_Best Practices.pdf.
2. Aalto University, 2024. DeepCaption. GitHub. Available at: https://github.com/aalto-cbir/DeepCaption [Accessed 30 June 2024].
3. Ada Lovelace Institute, 2023. Foundation models in the public sector. Ada Lovelace Institute. Available at: https://www.adalovelaceinstitute.org/wp-content/uploads/2023/10/Foundation-models-in-the-public-sector-Oct-2023.pdf [Accessed 30 June 2024].
4. Ada Lovelace Institute, 2024. Explainer: What is a foundation model? Ada Lovelace Institute. Available at: https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/ [Accessed 30 June 2024].
5. Ada Lovelace Institute, 2024. The value chain of general-purpose AI. Ada Lovelace Institute. Available at: https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/ [Accessed 30 June 2024].
6. ADLAB, P. (2014). Pictures painted in words. ADLAB Audio Description guidelines. Available from https://accessdata.com/products-services/ad-lab [Accessed 30 June 2024].
7. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K., 2022. Flamingo: a Visual Language Model for Few-Shot Learning. arXiv. Available at: https://arxiv.org/abs/2204.14198 [Accessed 30 June 2024].
8. Alexander Thamm, 2024. An Introduction to Foundation Models. Alexander Thamm GmbH. Available at: https://www.alexanderthamm.com/en/blog/an-introduction-to-foundation-models/ [Accessed 30 June 2024].
9. Amaresh, M. and Chitrakala, S. (2019). Video captioning using deep learning: An overview of methods, datasets and metrics. Proceedings of the 2019 IEEE International Conference on Communication and Signal Processing, ICCSP 2019. 2019. IEEE, 656–661. Available from https://doi.org/10.1109/ICCSP.2019.8698097.
10. Amazon Web Services, 2024. What are Foundation Models? Amazon Web Services. Available at: https://aws.amazon.com/what-is/foundation-models/ [Accessed 30 June 2024].
11. American Council of the Blind, 2003. Guidelines for Audio Description. Available at: https://adp.acb.org/guidelines.html [Accessed 30 June 2024].
12. American Council of the Blind. (2010). Audio description guidelines and best practices: A work in progress. (September). Available from http://docenti.unimc.it/catia.giaconi/teaching/2017/17069/files/corso-sostegno/audiodescrizioni.
13. BAI. (2016). BAI Access Rules. (January), 1–15. Available from http://www.bai.ie/en/media/sites/2/2016/08/20160106_BAI_AccessRules2016_vFinal.pdf.

14. Bell, V.A. and Johnson-Laird, P.N. (1998). A model theory of modal reasoning. Cognitive Science, 22 (1), 25–51. Available from https://doi.org/10.1207/s15516709cog2201_2.
15. Biocca, F., 1997. The Cyborg's Dilemma: Progressive Embodiment in Virtual Environments. Journal of Computer-Mediated Communication, [online] 3(2). Available at: https://www.researchgate.net/publication/220438229_The_Cyborg's_Dilemma_Progressive_Embodiment_in_Virtual_Environments [Accessed 30 June 2024].
16. Bochkovskiy, A., Wang, C.-Y. and Liao, H.-Y.M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. Available from http://arxiv.org/abs/2004.10934.
17. Bourne, R.R.A., Adelson, J., Flaxman, S., Briant, P., Bottone, M., Vos, T., Naidoo, K., Braithwaite, T., Cicinelli, M., Jonas, J., Limburg, H., Resnikoff, S., Silvester, A., Nangia, V. and Taylor, H.R. (2020) 'Global Prevalence of Blindness and Distance and Near Vision Impairment in 2020: progress towards the Vision 2020 targets and what the future holds', *Investigative Ophthalmology & Visual Science*, 61(7), p. 2317. Available at: https://iovs.arvojournals.org/article.aspx?articleid=2767477 (Accessed: 1 July 2024).
18. Braun, S. (2016). The importance of being relevant? TargetTarget. International Journal of Translation Studies, 28 (2), 302–313. Available from https://doi.org/10.1075/target.28.2.10bra.
19. Braun, S. (2020). Finding the Right Words : Investigating Machine-Generated Video Description Quality Using a Corpus-Based Approach. 11–35.
20. Broadcasting Authority of Ireland, 2009. Broadcasting Act 2009. Available at: https://www.irishstatutebook.ie/eli/2009/act/18/enacted/en/html [Accessed 30 June 2024].
21. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D., 2020. Language Models are Few-Shot Learners. arXiv. Available at: https://arxiv.org/abs/2005.14165 [Accessed 30 June 2024].
22. Casile, A., Caggiano, V. and Ferrari, P.F., 2011. The mirror neuron system: a fresh view. The Neuroscientist, 17(5), pp.524-538. Available at: https://www.researchgate.net/publication/51020118_The_Mirror_Neuron_System [Accessed 30 June 2024].
23. Chen, J. et al. (2015). Déjà image-captions: A corpus of expressive descriptions in repetition. NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 504–514. Available from https://doi.org/10.3115/v1/n15-1053.
24. Chen, X. and Zitnick, C.L. (2015). Mind's eye: A recurrent visual representation for image caption generation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June, 2422–2431. Available from https://doi.org/10.1109/CVPR.2015.7298856.
25. Chen, X. et al. (2015). Microsoft COCO Captions: Data Collection and Evaluation Server. 1–7. Available from http://arxiv.org/abs/1504.00325.
26. Chen, Y. et al. (2018). Less is more: Picking informative frames for video captioning. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial

Intelligence and Lecture Notes in Bioinformatics). 2018. 367–384. Available from https://doi.org/10.1007/978-3-030-01261-8_22.
27. Cho, K. et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. 2014. 1724–1734. Available from https://doi.org/10.3115/v1/d14-1179.
28. Cromwell, H.C. et al. (2008). Sensory gating: A translational effort from basic to clinical science. Clinical EEG and Neuroscience, 39 (2), 69–72. Available from https://doi.org/10.1177/155005940803900209.
29. CRTC, 2009. Broadcasting and Telecom Regulatory Policy CRTC 2009-430. Available at: https://crtc.gc.ca/eng/archive/2009/2009-430.htm#a21 [Accessed 30 June 2024].
30. Cryer, H. and Home, S. (2009). User attitudes towards synthetic speech for talking books.
31. Dai, J. et al. (2016). R-FCN: Object detection via region-based fully convolutional networks. Advances in Neural Information Processing Systems, 379–387.
32. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. Available at: https://arxiv.org/pdf/1810.04805 [Accessed 30 June 2024].
33. Di Giovanni, E., Fryer, L., and Tor-Carroggio, I., 2023. Beyond Objectivity in Audio Description: New Practices and Perspectives. ResearchGate. Available at: https://www.researchgate.net/publication/376918219_Beyond_Objectivity_in_Audio_Description_New_Practices_and_Perspectives [Accessed 25 June 2024].
34. Donahue, J. et al. (2017). Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39 (4), 677–691. Available from https://doi.org/10.1109/TPAMI.2016.2599174.
35. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N., 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv. Available at: https://arxiv.org/abs/2010.11929 [Accessed 30 June 2024].
36. Douglas, M.R., 2023. Large Language Models. arXiv. Available at: https://arxiv.org/pdf/2307.05782 [Accessed 30 June 2024].
37. Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., and Mordatch, I., 2023. PaLM-E: An Embodied Multimodal Language Model. Available at: https://palm-e.github.io/ [Accessed 30 June 2024].
38. Elliott, D. and Keller, F. (2013). Image description using visual dependency representations. EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, (October), 1292–1302.
39. Fryer, L. (2016). An introduction to audio description a practical guide. London: Routledge. Available from https://www.dawsonera.com/readonline/9781315707228.
40. Fu, Jianlong and Zheng, Heliang and Mei, T. (2014). Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-grained Image Recognition. 4438–4446. Available from https://doi.org/https://openaccess.thecvf.com/content_cvpr_2017/papers/Fu_Look_Closer_to_CVPR_2017_paper.pdf.
41. GOV.UK, 2024. Data Ethics Framework: glossary and methodology. GOV.UK. Available at: https://www.gov.uk/government/publications/data-ethics-

framework/data-ethics-framework-glossary-and-methodology [Accessed 30 June 2024].
42. Government of Australia, 1992. Broadcasting Services Act 1992. Available at: https://www8.austlii.edu.au/cgi-bin/viewdb/au/legis/cth/consol_act/bsa1992214/ [Accessed 30 June 2024].
43. Guo, Y., Zhang, J. and Gao, L. (2019). Exploiting long-term temporal dynamics for video captioning. World Wide Web, 22 (2), 735–749. Available from https://doi.org/10.1007/s11280-018-0530-0.
44. H. P. Grice. (1969). Philosophical Review Utterer ' s Meaning and Intention Author (s): H . P . Grice Source : The Philosophical Review , Vol . 78 , No . 2 (Apr ., 1969), pp . 147-177 Published by : Duke University Press on behalf of Philosophical Review Stable URL : http. The Philosophical Review, 78 (2), 147–177.
45. Hookway, B., 2014. Interface. Cambridge, Massachusetts: The MIT Press.
46. Hori, C. et al. (2017). Attention-Based Multimodal Fusion for Video Description. Proceedings of the IEEE International Conference on Computer Vision. 2017. Available from https://doi.org/10.1109/ICCV.2017.450.
47. Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O. K., Patra, B., Liu, Q., Aggarwal, K., Chi, Z., Bjorck, J., Chaudhary, V., Som, S., Song, X., and Wei, F., 2023. Language Is Not All You Need: Aligning Perception with Language Models. arXiv. Available at: https://arxiv.org/abs/2302.14045 [Accessed 30 June 2024].
48. Huang, T.H. et al. (2016). Visual storytelling. 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference, 1233–1239. Available from https://doi.org/10.4324/9781315667829-4.
49. Ibañez, A.M. (2010). Evaluation criteria and film narrative. A frame to teaching relevance in audio description. Perspectives: Studies in Translatology, 18 (3), 143–153. Available from https://doi.org/10.1080/0907676X.2010.485682.
50. ISO, 2015. ISO 17100:2015: Translation services — Requirements for translation services. [online] Available at: https://www.iso.org/standard/59149.html [Accessed 30 June 2024].
51. ISO, 2024. Developing standards. Available at: https://www.iso.org/developing-standards.html [Accessed 30 June 2024].
52. ISO. (2015). PD ISO / IEC TS 20071-21 : 2015 BSI Standards Publication Information technology — User interface component accessibility Part 21 : Guidance on audio descriptions.
53. ITC (2000) ITC Guidance on Standards for Audio Description. Available online: http://www.ofcom.org.uk/static/archive/itc/uploads/ITC_Guidance_On_Standards_for_Audio_Description.doc [Accessed: 10 January 2022] Alternative link https://msradio.huji.ac.il/narration.doc [Accessed 30 June 2024]
54. Kiros, R. and Zemel, R. (2013). Multimodal Neural Language Models.
55. Krishna, R. et al. (2017). Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. International Journal of Computer Vision, 123 (1), 32–73. Available from https://doi.org/10.1007/s11263-016-0981-7.
56. Lakritz, J. and Salway, A. (2005). The Semi-Automatic Generation of Audio Description from Screenplays The Semi-Automatic Generation of Audio Description om Screenplays. 16. Available from http://www.bbrel.co.uk/pdfs/CS-06-05.pdf.

57. Le, J., 2023. Foundation models are going multimodal. Twelve Labs. Available at: https://www.twelvelabs.io/blog/foundation-models-are-going-multimodal [Accessed 30 June 2024].
58. Lei, J., Amrani, E., Jiang, L., He, S., Liang, C., Brown, C., Torfi, A., Gutfreund, D., Wu, W., Seidman, B., Chen, L., and Shi, H., 2022. All in One: Exploring Unified Video-Language Pre-training. arXiv. Available at: https://arxiv.org/abs/2203.07303 [Accessed 30 June 2024].
59. Lin, T.Y. et al. (2014). Microsoft COCO: Common objects in context. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8693 LNCS (PART 5), 740–755. Available from https://doi.org/10.1007/978-3-319-10602-1_48.
60. Lo, L.Y. and Lai, C.C., 2022. Visual–auditory interactions on explicit and implicit information processing. Cognitive Processing, 23(2), pp.179-189. Available at: https://doi.org/10.1007/s10339-022-01077-2 [Accessed 30 June 2024].
61. Marian, V., Hayakawa, S. and Schroeder, S.R., 2021. Cross-modal interaction between auditory and visual input impacts memory retrieval. Frontiers in Neuroscience, 15, p.661477. Available at: https://www.frontiersin.org/articles/10.3389/fnins.2021.661477/full [Accessed 30 June 2024].
62. Marian, V., Hayakawa, S. and Schroeder, S.R., 2021. Cross-modal interaction between auditory and visual input impacts memory retrieval. Frontiers in Neuroscience, 15, p.661477. Available at: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2021.661477/full [Accessed 30 June 2024].
63. Mazur, I. and Chmiel, A. (2012). Towards common European audio description guidelines: Results of the Pear Tree Project. Perspectives: Studies in Translatology, 20 (1), 5–23. Available from https://doi.org/10.1080/0907676X.2011.632687.
64. Michael, H. (1987). (Received 5 February 1987). 335–336.
65. Mitchell, M. et al. (2015). From Captions to Visual Concepts and Back - Fang_From_Captions_to_2015_CVPR_paper.pdf. Available from http://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Fang_From_Captions_to_2015_CVPR_paper.pdf.
66. Morgado, P., Li, Y. and Vasconcelos, N., 2020. Learning Representations from Audio-Visual Spatial Alignment. arXiv preprint arXiv:2011.01819. Available at: https://arxiv.org/pdf/2011.01819 [Accessed 30 June 2024].
67. Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A., 2023. A Comprehensive Overview of Large Language Models. arXiv. Available at: https://arxiv.org/pdf/2307.06435 [Accessed 30 June 2024].
68. NHS (2021) *Blindness and vision loss*. Available at: https://www.nhs.uk/conditions/vision-loss/ (Accessed: 30 June 2024).
69. Núñez, A.J.C. (2015). Multimodality and Multi-sensoriality as Basis for Access to Knowledge in Translation: The Case of Audio Description of Colour and Movement. Procedia - Social and Behavioral Sciences, 212, 210–217. Available from https://doi.org/10.1016/j.sbspro.2015.11.335.
70. OpenAI, 2023. GPT-4. OpenAI. Available at: https://openai.com/index/gpt-4-research/ [Accessed 30 June 2024].

71. OpenAI, 2024. Introducing GPT-4o and more tools to ChatGPT free users. OpenAI. Available at: https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/ [Accessed 30 June 2024].
72. Orero, P. (ed.), 2004. Audiovisual translation: A new dynamic umbrella. In: Topics in Audiovisual Translation. Amsterdam and Philadelphia: John Benjamins, pp. vii-xiii. Available at: https://benjamins.com/catalog/btl.56 [Accessed 30 June 2024].
73. Pan, Y. et al. (2016). Jointly modeling embedding and translation to bridge video and language. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-Decem, 4594–4602. Available from https://doi.org/10.1109/CVPR.2016.497.
74. Pearson, R. (2013). Described Video Best Practices. 1 (1), 42. Available from https://ecfsapi.fcc.gov/file/7520940294.pdf.
75. Piety, P.J. (2004). The language system of audio description: An investigation as a discursive process. Journal of Visual Impairment and Blindness, 98 (8), 453–469. Available from https://doi.org/10.1177/0145482x0409800802.
76. PAHO (2024) *Visual Health*. Available at: https://www.paho.org/en/topics/visual-health (Accessed: 30 June 2024).
77. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I., 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv. Available at: https://arxiv.org/abs/2103.00020 [Accessed 30 June 2024].
78. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I., 2019. Language Models are Unsupervised Multitask Learners. Semantic Scholar. Available at: https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe [Accessed 30 June 2024].
79. Rai, S., Greening, J. and Petré, L. (2010). A Comparative Study of Audio Description Guidelines Prevalent in Different Countries. 112. Available from http://audiodescription.co.uk/uploads/general/RNIB._AD_standa
80. Ramanathan, V., Liang, P. and Fei-Fei, L. (2013). Video event understanding using natural language descriptions. Proceedings of the IEEE International Conference on Computer Vision, 905–912. Available from https://doi.org/10.1109/ICCV.2013.117.
81. Redmon, J. and Farhadi, A. (2017). YOLO9000: Better, faster, stronger. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-Janua, 6517–6525. Available from https://doi.org/10.1109/CVPR.2017.690.
82. Ren, S. et al. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39 (6), 1137–1149. Available from https://doi.org/10.1109/TPAMI.2016.2577031.
83. Rensink, R.A. (2000). The dynamic representation of scenes. Visual Cognition, 7 (1–3), 17–42. Available from https://doi.org/10.1080/135062800394667.
84. RNIB, 2010. Audio Description Standards. Available at: https://web.archive.org/web/20180422091903id_/http:/audiodescription.co.uk/uploads/general/RNIB._AD_standards.pdf [Accessed 30 June 2024].
85. Roberts, A. and Raffel, C., 2020. Exploring Transfer Learning with T5: the Text-To-Text Transfer Transformer. Google Research Blog. Available at: https://research.google/blog/exploring-transfer-learning-with-t5-the-text-to-text-transfer-transformer/ [Accessed 30 June 2024].

86. Rohrbach, M. et al. (2013). Translating video content to natural language descriptions. Proceedings of the IEEE International Conference on Computer Vision, 433–440. Available from https://doi.org/10.1109/ICCV.2013.61.
87. Ruder, S., 2024. Transfer Learning - Machine Learning's Next Frontier. Ruder.io. Available at: https://www.ruder.io/transfer-learning/ [Accessed 30 June 2024].
88. Sadeghi, M.A. and Farhadi, A. (2011). Recognition using visual phrases. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1745–1752. Available from https://doi.org/10.1109/CVPR.2011.5995711.
89. Simon, C., Torcoli, M. and Paulus, J. (2019). MPEG-H Audio for Improving Accessibility in Broadcasting and Streaming. 1–11. Available from http://arxiv.org/abs/1909.11549.
90. Snyder, J. (2005). Audio description: The visual made verbal. International Congress Series, 1282, 935–939. Available from https://doi.org/10.1016/j.ics.2005.05.215.
91. Snyder, J. (2007). Audio Description. The International Journal of the Arts in Society: Annual Review, 2 (2), 99–104. Available from https://doi.org/10.18848/1833-1866/cgp/v02i02/35358.
92. Stanford HAI, 2024. Reflections on Foundation Models. Stanford HAI. Available at: https://hai.stanford.edu/news/reflections-foundation-models [Accessed 30 June 2024].
93. Sun, C., Myers, A., Vondrick, C., Murphy, K., and Schmid, C., 2019. VideoBERT: A Joint Model for Video and Language Representation Learning. arXiv. Available at: https://arxiv.org/abs/1904.01766
94. Surikuchi, A. (2019). Visual Storytelling : Captioning of Image Sequences.
95. Sutskever, I., Vinyals, O., and Le, Q. V., 2014. Sequence to Sequence Learning with Neural Networks. arXiv. Available at: https://arxiv.org/pdf/1409.3215 [Accessed 30 June 2024].
96. Szarkowska, A. (2011). Text-to-speech audio description: Towards wider availability of AD. Journal of Specialised Translation, (15), 142–162.
97. Taylor, C. (2019). Audio Description: A Multimodal Practice in Expansion. Multimodality. 195–218. Available from https://doi.org/10.1515/9783110608694-008.
98. TensorFlow, 2024. c4. TensorFlow Datasets. Available at: https://www.tensorflow.org/datasets/catalog/c4 [Accessed 30 June 2024].
99. Tong, Z., Song, Y., Wang, J., and Wang, L., 2022. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. arXiv. Available at: https://arxiv.org/abs/2203.12602 [Accessed 30 June 2024].
100. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I., 2017. Attention Is All You Need. arXiv. Available at: https://arxiv.org/abs/1706.03762 [Accessed 30 June 2024].
101. Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in mechanical translation. Ifip congress (2). 1968. 1114–1122.
102. Venugopalan, S., Rohrbach, M., et al. (2015). Sequence to sequence - Video to text. Proceedings of the IEEE International Conference on Computer Vision, 2015 Inter, 4534–4542. Available from https://doi.org/10.1109/ICCV.2015.515.
103. Venugopalan, S., Xu, H., et al. (2015). Translating videos to natural language using deep recurrent neural networks. NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human

Language Technologies, Proceedings of the Conference, 1494–1504. Available from https://doi.org/10.3115/v1/n15-1173.
104. Vinyals, O. et al. (2015). Show and tell: A neural image caption generator. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June, 3156–3164. Available from https://doi.org/10.1109/CVPR.2015.7298935.
105. Walczak, A. (2018). Audio description on smartphones: making cinema accessible for visually impaired audiences. Universal Access in the Information Society, 17 (4), 833–840. Available from https://doi.org/10.1007/s10209-017-0568-2.
106. Walczak, A. and Fryer, L. (2017). Creative description: The impact of audio description style on presence in visually impaired audiences. British Journal of Visual Impairment, 35 (1), 6–17. Available from https://doi.org/10.1177/0264619616661603.
107. Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y., 2021. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. arXiv. Available at: https://arxiv.org/abs/2108.10904 [Accessed 30 June 2024].
108. Xenonstack, 2024. Introduction to Foundation Models. Xenonstack. Available at: https://www.xenonstack.com/blog/foundation-models [Accessed 30 June 2024].
109. Xu, K. et al. (2015). Show, attend and tell: Neural image caption generation with visual attention. 32nd International Conference on Machine Learning, ICML 2015. 2015. 2048–2057.
110. Young, P. et al. (2014). From image descriptions to visual denotations. Transactions of the Association of Computational Linguistics, 2 (1), 67–78. Available from https://aclanthology.coli.uni-saarland.de/papers/Q14-1006/q14-1006%0Ahttp://aclweb.org/anthology/Q14-1006.
111. Yu, J., Wang, Z., Vasudevan, V., Yeung, S., Seyedhosseini, M., and Wu, Y., 2022. CoCa: Contrastive Captioners are Image-Text Foundation Models. arXiv. Available at: https://arxiv.org/abs/2205.01917 [Accessed 30 June 2024].
112. Zhu, L.L. and Beauchamp, M.S., 2017. Mouth and voice: A relationship between visual and auditory preference in the human superior temporal sulcus. Journal of Neuroscience, 37(10), pp.2697-2708. Available at: https://doi.org/10.1523/JNEUROSCI.2914-16.2017 [Accessed 30 June 2024].
113. Zitnick, C.L. and Parikh, D. (2013). Bringing semantics into focus using visual abstraction. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 3009–3016. Available from https://doi.org/10.1109/CVPR.2013.387.