



Degree Project in Technology

First cycle, 15 credits

Exploring AI-driven Solutions for Automated Audio Descriptions of Videos

Region Stockholm

**ZIANG WANG
HARON OSMAN**

Exploring AI-driven Solutions for Automated Audio Descriptions of Videos

Region Stockholm

ZIANG WANG

HARON OSMAN

Degree Programme in Computer Engineering

Date: June 16, 2025

Supervisors: Amna Irshad, Marcus Pettersson

Examiner: Niharika Gauraha

School of Electrical Engineering and Computer Science

Host company: Region Stockholm

Swedish title: Utforskar AI-drivna lösningar för automatiserad ljudbeskrivning av videor

Swedish subtitle: Region Stockholm

Abstract

This thesis explores the potential of using **Artificial Intelligence (AI)** to create audio descriptions for videos. AI has become more common in daily life and helps us in many areas. It supports tasks in education, public transport, healthcare and other sectors. In collaboration with Region Stockholm, the study explored the development of a system utilizing AI-driven solutions for automated audio descriptions of videos. The reason behind the project is that Region Stockholm currently struggles with producing audio described versions of videos. This has been met with complaints from **Synskadades Riksförbund (SRF)**, and it is an issue that needs to be addressed due to legal requirements for accessibility which takes effect on June 28, 2025. The goal is to successfully develop, test and evaluate a prototype of the system, with the purpose of streamlining the creation process of audio descriptions. This should enhance accessibility for visually impaired individuals since it allows more videos to be audio described. This research faced several limitations and challenges, mainly related to budget constraints. To tackle these challenges, a prototype was implemented using free and low-cost AI services. The prototype includes two main features: description generation in **SubRip Subtitle (SRT)** format (subtitles) and speech generation using **Text To Speech (TTS)**. The description generation is provided by Google's Gemini 2.0 Flash, while the service for speech generation is provided by Azure OpenAI's **TTS** model. A usability test of the prototype was conducted with participants representing the intended users at Region Stockholm. The results showed that participants were comfortable with certain features of the prototype and recognized its potential for future development. However, they also identified issues and areas that could be improved. For instance, the prototype could not sync the generated audio with the video. Furthermore, the prototype was demonstrated for the accessibility department of **Sveriges Television (SVT)**, where they shared their feedback and considerations for further development. Overall, the insights gathered show that the prototype streamlined the process of creating audio descriptions by partial automation. This could lead to more videos to become audio described, thereby enhancing accessibility for visually impaired individuals.

Keywords

Artificial Intelligent, Audio description, Large Language Model, Performance Evaluation

Sammanfattning

Denna uppsats utforskar möjligheten att använda **AI** för att skapa syntolkningar för videor. AI har blivit allt vanligare i vardagen och hjälper oss inom många områden. Den stödjer uppgifter inom utbildning, kollektivtrafik, hälso- och sjukvård samt andra sektorer. I samarbete med Region Stockholm utforskade studien utvecklingen av ett system som använder AI-drivna lösningar för automatiserade ljudbeskrivningar av videor. Anledningen till projektet är att Region Stockholm i dagsläget har svårt att producera syntolkade versioner av videor. Detta har lett till klagomål från **SRF**, och det är ett problem som måste åtgärdas på grund av lagkrav på tillgänglighet som träder i kraft den 28 juni 2025. Målet är att framgångsrikt utveckla, testa och utvärdera en prototyp av systemet i syfte att effektivisera processen för att skapa syntolkningar. Detta bör förbättra tillgängligheten för personer med synnedsättning, eftersom det möjliggör att fler videor blir syntolkade. Forskningen ställdes inför flera begränsningar och utmaningar, främst relaterade till budgetbegränsningar. För att hantera dessa utmaningar implementerades en prototyp med hjälp av gratis- och lågkostnadsbaserade AI-tjänster. Prototypen innehåller två huvudfunktioner: generering av beskrivningar i **SRT**-format (undertexter) samt talgenerering med **TTS**. Genereringen av beskrivningar tillhandahålls av Googles Gemini 2.0 Flash, medan tjänsten för talgenerering kommer från Azure OpenAI:s **TTS**-modell. Ett användartest av prototypen genomfördes med deltagare som representerade de avsedda användarna hos Region Stockholm. Resultaten visade att deltagarna var bekväma med vissa funktioner i prototypen och såg dess potential för framtida utveckling. De identifierade dock också brister och förbättringsområden. Till exempel kunde prototypen inte synkronisera det genererade ljudet med videon. Prototypen demonstrerades även för tillgänglighetsavdelningen på **SVT**, där de delade med sig av sin feedback och sina överväganden inför fortsatt utveckling. Sammanfattningsvis visar de insikter som samlades in att prototypen effektiviserar processen för att skapa syntolkningar genom delvis automatisering. Detta kan leda till att fler videor blir syntolkade och därmed förbättra tillgängligheten för personer med synnedsättning.

Nyckelord

Artificiell intelligens, Syntolkning, Large Language Model, Prestandautvärdering

Acknowledgments

We would like to express our appreciation to our supervisors from Region Stockholm, Marcus Pettersson and Rodolfo Alvarez Rosas. They supported us throughout the research process. Their guidance and encouragement helped us through stressful moments. Their dedication played a key roles in making this bachelor's thesis possible.

We would also like to thank our **Kungliga Tekniska högskolan (KTH)** supervisor and examiner, Amna Irshad and Niharika Gauraha. They showed generous support and provided valuable input during our research work.

Stockholm, June 2025

Ziang Wang

Haron Osman

Contents

1	Introduction	1
1.1	Background	1
1.2	Problem	2
1.2.1	Original problem and definition	2
1.2.2	Scientific and engineering issues	2
1.3	Purpose	2
1.4	Goals	2
1.5	Research Methodology	3
1.5.1	Philosophical assumptions	3
1.5.2	Research methods	3
1.5.3	Research approaches	4
1.6	Delimitations	5
1.7	Ethics and Sustainability	5
1.7.1	Ethical concerns	5
1.7.2	Sustainability concerns	6
1.8	Structure of the thesis	7
2	Background	8
2.1	Region Stockholm	8
2.1.1	Legal requirements for accessibility	9
2.1.2	Current State of Accessibility	9
2.2	Machine Learning	9
2.2.1	Generative AI	10
2.2.2	Computer Vision	10
2.2.3	Multimodal AI	10
2.2.4	Prompt & Prompt Engineering	11
2.3	Related work area	11
2.3.1	Toward automatic audio description generation for accessible videos	11

2.3.2	AI-Powered Scene Description Applications	12
2.3.3	Existing services	13
2.4	Summary	13
3	Method or Methods	14
3.1	Research Process	14
3.2	Pre Study & Literature Study	14
3.3	Design and Development of Prototype	18
3.4	Data Collection	21
3.4.1	Sampling	21
3.4.2	Sample Size	21
3.4.3	Target Population	21
3.5	Experimental design/Planned Measurements	21
3.5.1	Test environment/test bed/model	21
3.5.2	Hardware and Software to be used	21
3.6	Assessing reliability and validity of the data collected	22
3.6.1	Validity of method	22
3.6.2	Reliability of method	22
3.6.3	Data validity	22
3.6.4	Reliability of data	22
3.7	Planned Data Analysis	23
3.7.1	Data Analysis Technique	23
3.7.2	Software Tools	23
3.8	Evaluation framework	23
4	Automated Audio Description of Videos - Region Stockholm	24
4.1	Video Analysis	24
4.2	Text-To-Speech	27
5	Results and Analysis	29
5.1	Analysis of survey responses	29
5.1.1	Quantitative results	29
5.1.2	Qualitative results	33
5.2	Additional Results	34
5.3	Reliability Analysis	35
5.4	Validity Analysis	35
6	Discussion	36

7	Conclusions and Future work	38
7.1	Conclusions	38
7.2	Limitations	39
7.3	Future work	39
7.3.1	What has been left undone?	40
7.3.2	Next steps of the project	40
7.4	Reflections	41
	References	43

List of Figures

3.1	Research Process	15
3.2	System Architecture	19
3.3	UI Architecture	20
4.1	Usage Process	25
4.2	User prompts	26
4.3	Description Modification tool	27
4.4	Subtitle editing tool	28
4.5	Text-To-Speech	28
5.1	The participants' answers regarding the objectivity of the generated audio description.	30
5.2	The participants' answers regarding the tool's potential to enable more videos to be audio described.	30
5.3	The participants' responses on their intention to use the tool in future work.	31
5.4	The participants' responses on the ease of modifying the generated description.	31
5.5	Participants' ratings on the quality of the generated description.	32
5.6	The participants' ratings on the quality of the generated audio.	32
5.7	Participants' ratings of the perceived naturalness of the generated speech.	32

Listings

List of acronyms and abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Networks
API	Application Programming Interface
BLV	Blind and Low-Vision
DL	Deep Learning
KTH	Kungliga Tekniska högskolan
LLM	Large Language Model
ML	Machine Learning
NLG	Natural Language Generation
NLP	Natural Language Processing
OS	Operating system
SRF	Synskadades Riksförbund
SRT	SubRip Subtitle
SVT	Sveriges Television
TTS	Text To Speech

UI	User Interface
UN	United Nations
UX	User Experience

Chapter 1

Introduction

In this chapter, the project is introduced with background info on how AI can help visually impaired people by creating automated audio descriptions for videos. It lays out the main problem, purpose, and goals. The research methods, ethical and sustainability concerns, and limitations are also explained. Finally, it gives an overview of how the rest of the thesis is organized.

1.1 Background

Artificial Intelligence (AI) has displayed rapid growth over the past decade and has shown great potential in improving efficiency and accessibility in various sectors. Region Stockholm, the governing body for healthcare, public transport, and regional development in Stockholm County, is actively exploring AI-driven solutions to enhance security, safety, and accessibility [1]. Among these efforts is ensuring that visually impaired individuals have equal access to information and entertainment presented visually through videos. To address this challenge, Region Stockholm is exploring AI-driven solutions for automated audio descriptions of videos. By utilizing AI technologies such as computer vision and **Natural Language Processing (NLP)**, it may be possible to generate high-quality descriptions of visual content. This could involve extending existing **Large Language Model (LLM)** such as Google's Gemini, which have proven capabilities in understanding and generating human-like text, to handle visual data and provide accurate audio descriptions. This project aims to explore and evaluate the potential of AI-driven automated audio descriptions, assessing their effectiveness, accuracy, and future opportunities for Region Stockholm. This could lead to a more

inclusive digital space, ensuring that visually impaired individuals can access and understand video content equally.

1.2 Problem

This research study investigates how a system utilizing AI-driven solutions for automated audio descriptions of videos could improve accessibility for visually impaired individuals. How can relevant visual attributes be highlighted effectively? How can the system streamline the creation of audio descriptions for videos? And what are the challenges and future considerations for further development.

1.2.1 Original problem and definition

This research explores how AI-driven solutions for automated audio descriptions of videos provided by Region Stockholm can enhance accessibility for visually impaired users.

1.2.2 Scientific and engineering issues

Developing an AI-driven system for creating audio descriptions of videos introduces a number of scientific and engineering challenges. The system needs to accurately detect and prioritize visual elements based on user preferences. Translating the video content into clear, natural, and adaptive audio descriptions.

1.3 Purpose

The purpose of this thesis is to explore how automated audio description can streamline the creation of audio descriptions for video content by Region Stockholm and enhance accessibility for visually impaired individuals.

1.4 Goals

The goal of this project is to develop and evaluate a prototype of a system providing automated audio descriptions of videos for visually impaired individuals. The goal is broken down into the following sub goals:

1. Provide a prototype of an automated audio description system for Region Stockholm.
2. Conduct usability tests to gather data regarding ease of use and user relevance.
3. Evaluate the overall performance and identify potential improvements of the prototype.
4. Gather considerations and recommendations for future development.

1.5 Research Methodology

1.5.1 Philosophical assumptions

The thesis mainly adopts a pragmatic philosophical approach, which allows a certain freedom to choose approaches, methods and techniques that are best suited to achieve the project goal.

1.5.2 Research methods

This thesis will utilize both qualitative measures and quantitative measures, as research methods, through relevant data collection approaches.

Qualitative measures

Qualitative data allows test persons to provide descriptive feedback, with expressions in their own wordings through text boxes or interviews.

Quantitative measures

Quantitative data will be gathered by allowing test persons to provide answers that can be measured with numbers such as "yes/no" questions or rating scales etc.

Rationale

The rationale behind the chosen research methods is that combining both qualitative and quantitative measures is expected to provide valuable insights. The outcome should be practical for those creating audio descriptions for videos and individuals seeking to access visual information in alternative

ways. Qualitative feedback captures user perceptions that can not be explained with only numbers, while quantitative data gives a clear overview of general performance. Combining the two provides a more holistic evaluation in comparison to relying on only one.

1.5.3 Research approaches

The following are the research approaches that will be utilized.

Communication and communication tools

The project will mostly be conducted remotely since all development tasks are digital. Moreover, various communication tools such as teams, zoom or discord will be used throughout the project. The choice of tool will depend on the purpose such as meetings or collaborations.

Development environment

The development environment for this project will primarily use Python due to its popularity in similar projects, comprehensive library support, and available frameworks. Furthermore, Python provides great support for integration with many AI models and their corresponding **Application Programming Interface (API)**. Moreover, the system will be developed for compatibility across major **Operating system (OS)** platforms (Windows, macOS, Linux).

System architecture

The prototype will be developed using a modular system architecture, ensuring low coupling between modules, allowing easier maintenance and modifications in the code-base. Furthermore, the back-end and front-end will be properly separated, with the integration between them designed to allow updates in user experience without affecting the functionality in the back-end.

Version control

Git will be utilized as the version control system for the project due to popularity, extensive online resources available and our prior experience. Features such as branching, commit and revert will be valuable for this project. Furthermore, a clear guideline on how the version control system will be used will be established.

AI model selection

AI model selection will be based on certain criteria such as capability of video analysis, budget, API availability, as well as constraints given by Region Stockholm.

Prompt engineering

Prompt engineering techniques will be implemented to craft effective prompts for the selected AI model. This will ensure clarity, user specific adaptability (eg. relevant attributes), and context-awareness in the generated descriptions.

Usability Tests

Usability tests will be conducted to assess ease of use of the prototype, as well as the quality of descriptive responses and generated audio. Additionally, valuable data such as response time in relation to video length and size will be measured and evaluated to determine reasonability.

1.6 Delimitations

The delimitation is mainly related to the budget, since the organization struggles to fund additional AI services. As a result, the project will have to rely on free or low-cost AI tools. This delimitation could affect several things. It may reduce the ability to process longer videos and limit the amount of API requests. Most importantly, it might lower the quality of generated descriptions and audio. Moreover, since Region Stockholm is a public regional government organization, it is necessary to inform them about the AI services used and get their approval.

1.7 Ethics and Sustainability

1.7.1 Ethical concerns

AI-generated descriptions can sometimes reflect biases. These biases often come from the prompts used by developers. As a result, the audio description might be unfair or misleading. The impact of these biases can be particularly severe when dealing with diverse cultural, politics or social contexts. A description that seems neutral in one context might be offensive or misleading

in another [2].

The ideal solution to address the challenge is to create clear guidelines. Prompts should aim to be neutral and objective. Several tests need to be conducted and different prompts should be tested to see how they affect the AI's output. Since these AI-generated descriptions aims to assist visually impaired user, fairness is important. People from diverse backgrounds should review the prompts and gather feedback from people with different background can help identify potential bias and it is essential to improve both fairness and accuracy.

Additionally, several contextual tests need to be performed to analyze how the AI described scenes related to sensitive topics. This will help identify the most suitable approaches to ensure that the descriptions are as inclusive and accurate as possible.

1.7.2 Sustainability concerns

AI technology is expanding across various fields worldwide. It is becoming more accessible, even for beginners. As a result, people are using AI to simplify daily tasks to make life easier. One well known example is OpenAI's ChatGPT. It has become one of the most widely used AI tools globally. Its powerful model and User-friendly approach have attracted strong support from the user.

However, AI models require extensive cooling mechanism to prevent servers from overheating. This cooling process uses a large amount of water. Major companies such as Microsoft and Google, which run large-scale AI-systems, have reported significant increases in water usage due to AI operations. According to a research report by The Washington Post and the University of California, the GPT-4 language model uses over 520 milliliters of water just to generate a 100-word response [3].

As AI technologies become more integrated into various applications, the development of a AI-driven prototype for automated audio description prototype raises sustainability concerns. A system like automatic audio description using AI-technology require heavy computational power, especially when processing video content. This raises important questions about sustainability. It is important to consider the environmental impact of these technologies. Developers must aim for strong performance. At the same time, they should

explore energy-efficient models. Efficiency should not be an afterthought. It must be part of the design from the start.

1.8 Structure of the thesis

Chapter 1 gives an overview of the research work. It introduces the background and explains the main problem. It also outlines the purpose and potential goals of the work. In addition, the chapter highlights the possible ethical and sustainability concerns, as well as the its proposed solution.

Chapter 2 presents relevant background information related to the research area. It focuses mainly on the development of AI. This section includes a set of key findings. It also compares the current work with related studies. The goal is to support the purpose of the pre-study.

Chapter 3 presents the methodology and method used to solve the problem.

Chapter 4 presents the developed prototype and the services it offers.

Chapter 5 presents the results of the research. It includes prototype testing and feedback collected from both users and organizations.

Chapter 6 discusses the issues and suggested improvements of the prototype

Chapter 7 presents the conclusion and reflection on the research. It also outlines possible future work related to the ongoing prototype.

Chapter 2

Background

This chapter provides an overview of Region Stockholm and its accessibility goals, followed by a brief introduction to relevant machine learning concepts such as generative AI, computer vision, and multimodal AI. It also reviews related research and existing solutions in automated audio description for video accessibility, highlighting gaps that this thesis aims to address.

2.1 Region Stockholm

Region Stockholm is the governmental organization responsible for the health-care, public transport, and the regional development within Stockholm County in Sweden. The organization is one of the biggest employer in Sweden with 48000 employees with roles ranging from community planners to engineers and nurses. Their main objective is to ensure that Stockholm County is equal, open, and sustainable, offering equal chances and a high quality life for the residents [4].

Sustainability is a natural part of their work. Region Stockholm is working ambitiously with sustainability in order to ensure that the county has a sustainable future. Sustainability involves many parts, including social and environmental topics. Relevant for this thesis, topics such as disability inclusion and climate impact are highlighted [5].

Region Stockholm is actively working on ensuring that digital and physical spaces within the county are inclusive and accessible, regardless of disability. This work is based on laws and agreements such as Agenda 2030, **United Nations (UN)** convention on the rights of persons with disabilities and the

national disability policy goal [6].

Regarding the climate impact, Region Stockholm aims to shape a society with an environment that is free from pollution, promotes biological diversity and low climate impact in general. The goal is an improved environment both locally and globally [7].

2.1.1 Legal requirements for accessibility

In 2023, The Swedish Parliament (Riksdag) issued a new law – Lag (2023:254) *om vissa produkters och tjänsters tillgänglighet* – which sets accessibility requirements for certain products and services. The law, which comes into force in 2025-06-28, aims to promote equality in living conditions and participation in society, especially for individuals with disability [8].

With this, a set of accessibility requirements have been introduced. This includes requirements such as providing information about accessibility in an accessible way and accessible interfaces. Additionally, there exist various sector-specific requirements. For instance, services for access to audiovisual media services and for e-books [9].

2.1.2 Current State of Accessibility

According to **Synskadades Riksförbund (SRF)**, there are approximately 120 000 visually impaired individuals in Sweden [10]. Despite this, Region Stockholm currently lacks audio descriptions for most of their public videos. **SRF** highlight that videos, such as how to conduct a health test at home or about the expansion of the future subway station are currently inaccessible for visually impaired users [11]. Although the issue was first raised by **SRF** in 2022, it was confirmed in a recent meeting that the problem still persists. This gap in accessibility is particularly problematic given the legal requirements under SFS 2023:254, which takes effect on June 28, 2025.

2.2 Machine Learning

Machine Learning (ML) is a subfield of AI, focused on enabling systems to learn from data and mimic human-like decision-making. **ML** systems are utilized in different applications where systems have to solve complex tasks in a similar way as humans. **ML** encompasses several sub-fields, including

NLP, Artificial Neural Networks (ANN), and Deep Learning (DL). These areas power systems capable of making decisions, recognizing patterns, and generating outputs similar to human behavior. In this section, relevant ML concepts and use cases will be introduced and explained [12].

2.2.1 Generative AI

Generative AI is a field of AI that focuses on generating different types of content, including texts and images. These AI-systems rely on DL models that simulate the process of the human brain in decision-making and learning. The DL models are trained using a large number of data to identify and encode patterns and relationships. This information is then used in order to understand requests by users, and provide with relevant responses. Generative AI are applied on foundation models that serves as a basis for specific applications. The most common application are LLM [13].

Large Language Models

LLM are advanced AI-systems capable of NLP and Natural Language Generation (NLG). This means that the systems are able to understand and generate human written texts. LLMs has been groundbreaking, with many applications in various industries. For instance, numerous of organizations are currently using the GPT-series (generative pre-trained transformers). The GPT-series was first introduced by OpenAI in 2018, and has since received newer versions such as GPT-4. A few use cases of LLMs includes conversational agents, create and summarize content, language translation, and personalized recommendations [14].

2.2.2 Computer Vision

Computer vision allows systems to analyze and interpret visual data in videos and images similar to the human eye and brain. These systems are built using DL models to identify patterns, objects, and people. This means that systems can extract and interpret visual data similar to the coordination between the eyes and cognitive ability of humans [15].

2.2.3 Multimodal AI

Multimodal AIs are similar to LLMs, but capable of understanding and generating content of more than one means of communication. For instance,

these AI-systems can generate images based on given descriptions, or provide with a written summary of an image input. Furthermore, multiple means of inputs allows these AI-systems to utilize information from varying data sources, such as an image and text [16].

2.2.4 Prompt & Prompt Engineering

LLMs and Multimodal AIs require instructions provided by users in order to generate the desired outputs. These inputs are referred to as prompts. To yield consistency in the generated outputs, the AI-systems require enough context and detailed information about what the desired output should include. To achieve such consistency, users can systematically design the prompt, and evaluate the outputs through trial and error. If the outputs are inconsistent, refine the prompt and try again. This process is called prompt engineering. The prompt engineers choose specific formats, words and phrases to guide the AI, generating more meaningful outputs. Prompt engineering makes the overall utilization of generative AIs in applications more efficient and effective. Additionally, the developers often encapsulate the user input in a prompt, allowing the users to focus less on how they structure their requests in the applications [17].

2.3 Related work area

This section highlights research and solutions in the field of automated video accessibility. This includes a similar research work of automated audio descriptions, research about AI-powered scene descriptions and existing commercial tools such as verbit and viddyscribe.

2.3.1 Toward automatic audio description generation for accessible videos

Automatic audio description generation is an area that has been previously investigated by Yujia Wang et al in 2021. In their study, the researchers developed a system that generates an audio description of visual information using three different modules.

The researchers evaluated the system through conducted tests with 5 different video types and a total of 32 individuals, 12 of those were visually impaired. Their analysis revealed that audio description preferences varied among users

and the type of videos. These preferences include level of detail and amount of content generated by the system. Furthermore, the researchers mentioned that existing solutions usually do not include features for user preferences that allows for tailoring the audio description, emphasizing the demand of it. Additionally, the researchers observed that the output usually missed certain events, but that most visually impaired users were content with some information rather than none. Moreover, they emphasize that their system is a prototype and believe their study's analysis offers useful recommendations for further development in the area [18].

2.3.2 AI-Powered Scene Description Applications

In 2024, a team of researcher - Ricardo E Gonzalez Penuela, Jazmin Collins, Cynthia Bennett, Shiri Azenkot conducted a research study on investigating Use Cases of AI Powered Scene Description Applications for Blind and Low Vision People [19]. The tool generates scene descriptions from photos. The main purpose for this research work is to help **Blind and Low-Vision (BLV)** individuals understand their surroundings.

In a period of 12 days, sixteen **BLV** participants used the AI-tool in their daily lives, documenting their experiences. They shared when and how they used it, which features helped them most, and the challenges they faced during usage.

The studies reveals a variety of use cases. Participants used the tool to detect and identify texts, objects and actions within a scene. These examples highlight how flexibility and helpful the tool can be in supporting daily activities for **BLV** individuals. However, the researchers also mentioned some limitations. A majority of the participants did question the accuracy of the tool after receiving a vague or incomplete descriptions. This raises the concerns about trust in the system.

Despite these concerns, the study offered valuable insights into the users' perspectives. Privacy was a significant factor. Some participants preferred using the AI tool over asking people for help. This was particularly true when dealing with sensitive or personal information. The AI provided a greater sense of independence.

This thesis work offers meaningful connections to that study. It supports ongoing efforts to develop and evaluate a prototype of a system providing automated audio description of video material for visually impaired individuals.

2.3.3 Existing services

There exists several commercial solutions that offer services for automated audio description such as ViddyScribe and AudibleSight. Both platforms include features that allows for video editing to integrate the audio description. Also, AudibleSight allows for a more customized service, such as selecting specific scenes to include in the descriptions. These platforms operate on annual subscription models and require users to upload their videos to external servers for processing. Furthermore, large-scale or governmental use involve customized enterprise agreements without determined pricing [20][21].

2.4 Summary

Automated audio description of videos and scenes is a technique that has been previously studied by researchers and is offered through several commercial services today. However, in their study, Yujia Wang et al. observed that their prototype lacked features for user preferences, such as relevance and level of detail in descriptions. Our prototype attempts to address this limitation by allowing users to customize the prompt through parameters, making it possible to explicitly define what is and what is not relevant. This approach could also allow for adjustments to the level of detail of descriptions, depending on the configuration of the prompt. Furthermore, Yujia Wang et al. describes that the prototype lacked quality and missed certain key events in the generated descriptions. Their study was conducted in 2021, and since that advancements in high-quality AI services have made seamless video analysis more accessible. Our prototype aims to integrate these improved services for better accuracy and coverage.

The commercial solutions is a practical alternative for smaller companies and organizations. The services discussed earlier in this chapter allow for seamless process of video analysis, and video editing. However, the services operates on annual subscriptions with pricing varying depending on usage and features required. Most importantly, the price is undetermined for large-scale and governmental use which could introduce complexities in agreements and potentially high costs, particularly for organizations like Region Stockholm. Additionally, relying on third-party services for automated audio description could limit the videos eligible for the service, depending on data privacy concerns. Furthermore, Region Stockholm would have less control over how the data is processed when relying on third-party services.

Chapter 3

Method or Methods

This chapter outlines the research process and the methods employed, including data collection and the design of the prototype used in the study.

3.1 Research Process

The research focuses on developing AI-driven solution for automated audio description. These tools are designed for Region Stockholm, and the main purpose is to analyze and provide descriptions of visual elements in video. The research process can be divided into several key stages:

Step 1 Pre study & Literature study

Step 2 Design and Development of Prototype

Step 3 Data Collection

Step 4 Discussion and Implications

3.2 Pre Study & Literature Study

The pre-study phase mainly covered research and other work in the area of audio and scene description for visually impaired individuals, prior to this thesis. This phase put together the findings of the referenced studies, their purpose, and recommendations for further work. The findings of the pre study are presented in sections [2.3](#) and [2.4](#).

The literature study included a broad evaluation of various AI-service providers, and their specification. The goal was to find the most suitable

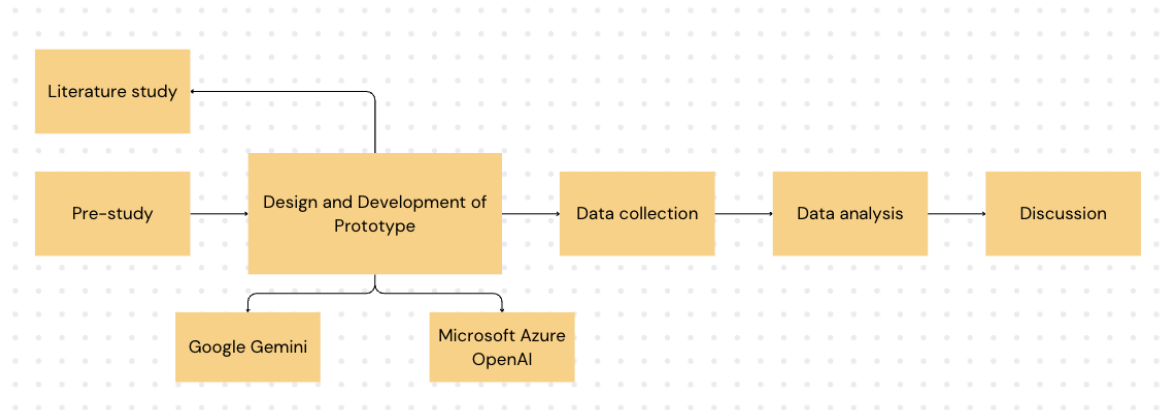


Figure 3.1: Research Process

AI-service provider for the development of the prototype. To reach a final decision, the research involved a detailed comparison of various aspects such as functionality, flexibility, accessibility and pricing models. After the initial review, three AI-service provider were identified as the main contenders for temporary use. Since the prototype focuses on video analysis, the nature of video data became a key consideration. Unlike a single image, which captures a single frame, a video consists of a sequence of images displayed in rapid succession. This sequence creates the illusion of motion, which adds complexity to the analysis process.

Some AI service providers are well-suited for analyzing individual frames. However, the primary goal of this research is to analyze continuous video content. As a result, providers that specialize in video analysis were preferred. Below is a breakdown of the three selected candidates.

Google Gemini Flash 2.0

Gemini Flash 2.0 is a AI service developed by Google. It is designed for fast and responsive tasks. The model can handle different types of data, including images and short video. The model can detect objects, actions, scenes, and contextual elements when using for analyzing videos [22]. It is also capable of generating textual descriptions in a key-frame mode or a full video mode. This allows for quick and efficient description of video content. Its low latency makes it ideal for interactive use.

Key attributes

- Fast and low latency model.
- Easy to use and user-friendly.
- Provides decent video analytics for detecting objects, scenes, and actions.
- Frame extraction or chunking is required for generating descriptions of longer videos.

OpenAI GPT-4

OpenAI's GPT-4 is a language model developed by OpenAI. It is the fourth iteration in the GPT series. This model is designed to help users generate human-like responses. GPT-4 is a large language model, enabling it to perform various **NLP** tasks. However, GPT-4 can not process video files directly, since it lacks build-in video or audio processing capabilities [23]. This makes GPT-4 an inappropriate choice for this research, as a transcript of the video content is required to generate a description of the video.

Key Features

- User-friendly.
- Generates fluent and human-like textual outputs.
- Does not support direct video or audio input.
- Requires pre-extracted transcripts for video content.

Microsoft Azure AI

Microsoft Azure is a cloud computing platform. It provides tools and AI services to help build, run, and manage AI solutions. One of its strengths lies in generating video descriptions using complex video analytics. It covers more than just basic parts like scenes, objects, people, and speech. It also adds some additional emotion and sentiment analysis. Unlike OpenAI, Azure accepts complete video files and processes both audio and visual content [24]. However, due to its broad capabilities, Azure Video Indexer can be challenging for beginners, especially for developers who are not familiar with Azure's cloud ecosystem. The pricing model is complex and may not be ideal for small to medium-sized projects [25].

Key Features

- Offers detailed analysis of both video and audio content.
- Accepts full video files and processes visual and audio elements.
- May be challenging for beginners due to the complexity of the Azure ecosystem.
- More expensive than Gemini or GPT-4 for small to medium-sized projects.
- Does not require pre-extracted transcripts. Users can upload video files directly through the web portal or API.

Tillgänglig Video

Tillgänglig Video is a Swedish organization focuses on making video content accessible to all individuals, regardless of their abilities. Since this current research is facing visual impaired user, the information provided by Tillgänglig Video can be highly valuable. Their insights can help improve accessibility to a new level.

Sight interpretation can be categorized into different types. The most basic type involves reading visual elements and briefly explains the action from one scene to another. A more advanced type of sight interpretation conveys emotions by describing facial expressions, character movements, and ongoing events. It is difficult to define a universal standard, as each individual is unique, and their needs can be very different. Adding audio descriptions to a video while maintaining consistency requires careful planning in advance. A well-planned film can sometimes reduce the need for audio descriptions, as key information may already be included in the original soundtrack [26].

A meeting was held with Malmö Stad to discuss regarding the use of audio descriptions in videos. They emphasized that creating audio descriptions is a team effort between video directors and accessibility experts. For instance, a well-structured film with intentional pauses creates better opportunities for professionals to add audio descriptions [27]. This improves the user experience and helps avoid unnecessary interruptions of the original audio.

3.3 Design and Development of Prototype

Features

The design of the prototype was entirely guided by the specific needs and requirements of Region Stockholm. To improve the user experience for both the video director and users with visual impairments, a series of meetings were conducted. These sessions focused on discussing and refining the prototype's features. Reaching an agreement on which features to include required several iterations and careful consideration.

- User Interface (UI)
 - Region Stockholm logotype display.
 - Simple video uploader.
 - AI-powered video analysis section
 - Basic video editing tools
- Response
 - Timestamp tagging
 - Video description in Swedish
 - Bias-free content
 - Gender-neutral language
 - SRT subtitle file output
 - Audio file export

System Architecture

Figure 3.2 shows the system architecture of the prototype. It provides an overall flow of how each component interacts, from user input to back-end processing and AI services.

When a user accesses the prototype, their device connects to the front-end via the internet. As a result, internet access and file upload permissions are required to allow users to upload video files and receive responses from the system.

The front-end handles all the user interactions. It sends forward the required

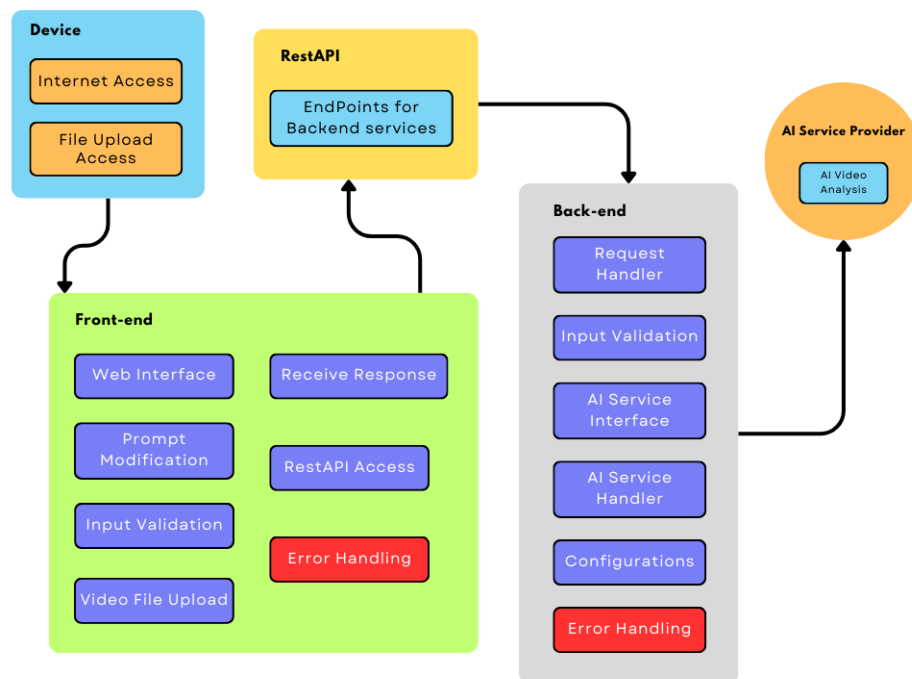


Figure 3.2: System Architecture

data to the back-end for processing. The **User Interface (UI)** features a clean, functional design that allows user to easily upload videos with just making a few click. After uploading, a clear and well-organized AI analysis shows up on the screen. Also, a set of input validations should be implemented to keep the system stable and prevent future errors in the backend. For example, the system needs to check whether the uploaded video file format or the file size is supported by the AI service provider. This helps avoid unnecessary errors in the back-end and additional costs when using the AI.

If the video passes the validation phase, it will send video files to the backend through a backend service via Rest API. The Rest API is a endpoints for backend service and it provides APIs that the front-end can call to trigger backend operations.



Figure 3.3: UI Architecture

The backend handles the direct communication with the AI service provider. It collects data from the front-end, processes it and reorganize it. The very first thing the request reaches is the Request Handler. Its purpose is to accept and interpret incoming API requests. Before sending any data forward to the AI service, the system performs a new round of validation check. This makes sure the input is valid and also helps prevent potential errors from happening.

Given that this prototype is designed for the public sector, the choice of the AI-service provider needs to be flexible. This ensures the solution can adapt to future requirements and possible limitations. For this reason, the AI service handler was divided into two distinct components to ensure better modularity and flexibility in the system design. The first component is the AI Service Interface. This component is responsible for preparing and formatting requests sent to external AI services. Due to the current economic conditions, a modular approach enables Region Stockholm to easily switch between different AI service providers with minimal impact on the overall system. The design maintains high cohesion and low coupling. This improves system flexibility and simplifies the debugging process. The AI Service Handler is responsible for managing all interaction with the selected AI service provider. It handles both the submission of requests and the processing of responses. Most importantly, this component also performs prompt engineering to optimize the input sent to the AI model.

The configuration component makes sure the system runs reliably. It manages all settings specific to the specific environment. This includes protecting sensitive information. API keys are kept secure and are excluded from version control by using the .git-ignore files.

3.4 Data Collection

3.4.1 Sampling

Purposeful sampling: test participants are selected from Region Stockholm by the supervisor.

3.4.2 Sample Size

An estimated 3-5 test participants are expected to take part in the user experience test.

3.4.3 Target Population

Content creators from Region Stockholm involved in accessibility or video/audio work.

3.5 Experimental design and Planned Measurements

3.5.1 Test environment/test bed/model

The test overseer will receive the project files along with a setup guide and carry out the test with selected participants. The participants will try all services the tool offers, simulating a simplified workflow of creating an audio description. Thereafter, participants will be able to complete the survey, allowing us to evaluate the tool based on the criteria presented in section [3.8](#).

3.5.2 Hardware and Software to be used

Google Forms will be used for conducting surveys after testing. The backend is developed in python, while the frontend is built with React. Any computer

with internet connection and basic development environment support is sufficient to carry out the test.

3.6 Assessing reliability and validity of the data collected

3.6.1 Validity of method

The validity of the method will be assured by letting participants utilize the tool as intended. That is, allowing the users to upload a video file for video analysis and generation of description, modification of the description, and generation of the audio.

3.6.2 Reliability of method

The reliability of method is ensured by standardizing the test procedure and survey questions. That is, the participants will test the prototype in the same manner, and answer the same questions in the survey. Furthermore, the test will be overseen by the supervisor at Region Stockholm to ensure that the participants have the necessary context.

3.6.3 Data validity

The data validity is ensured if users clearly understands and interact with the prototype as intended. Thereafter, the participants will answer a survey to see whether they perceive the output as correct, useful, and aligned with the purpose of streamlining the process of creating audio descriptions. Moreover, the survey is designed to collect both qualitative and quantitative feedback. This will allow participants to share their opinions in their own words while also leaving measurable responses that can be analyzed statistically.

3.6.4 Reliability of data

The data is considered reliable if more than one user give similar feedback or identify similar issues. Additionally, allowing users to freely comment gives them a chance to explain anything unclear or why they could not provide with accurate quantitative responses.

3.7 Planned Data Analysis

3.7.1 Data Analysis Technique

Quantitative analysis of closed-form responses (eg. yes/no, linear scales) in the survey will be utilized to get measurable insights on tool effectiveness and user satisfaction.

Qualitative thematic analysis of free text responses in survey will be used in order to identify common themes, suggestions, and user pain points.

3.7.2 Software Tools

Google Forms will be used for conducting surveys after testing.

3.8 Evaluation framework

The prototype will be evaluated based on how test participants experience the quality of descriptions, how easily descriptions can be created and modified, and quality of the generated audio. Since the target group are the ones intended to utilize a finalized system of the prototype, their assessment will provide valuable insights of the prototype's performance.

Key evaluation answers the following questions:

- **Functionality** – Does the prototype perform the intended tasks reliably?
- **Description quality** – How clear, accurate, and useful are the generated descriptions?
- **Objectivity** – Are the descriptions neutral and free from description?
- **Speech quality** – Is the generated audio clear and natural-sounding?

Chapter 4

Automated Audio Description of Videos - Region Stockholm

A working prototype based on an AI-driven solution was successfully developed. The prototype contains two main features. The first is a video analysis tool that extracts useful information from video content based on user preferences. The second feature is a **Text To Speech (TTS)** conversion tool that converts an **SubRip Subtitle (SRT)** file into AI-generated spoken audio. The video creator for Region Stockholm can easily upload videos using the designed **UI**. The interface includes a few buttons and text fields. The overall layout is clean and organized, and the design can be updated later to meet the company's specific requirements.

4.1 Video Analysis

Usage Process and User Interface

The prototype has a relatively basic **UI**, as its main purpose is to analyze video content while improving user experience. Users can easily upload a video using the video uploader and then click "Analyze" to generate the description. Thereafter, the user may modify the description as desired and choose to download the SRT file if they wish. Finally, the user can generate and download the completed audio description using the selected AI voice. Additionally, users may upload any SRT file to utilize the text-to-speech functionality.

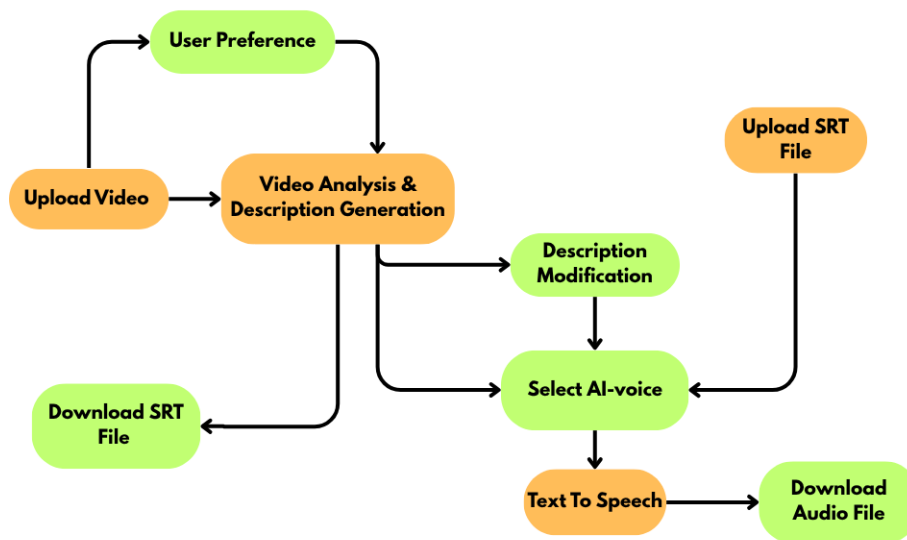


Figure 4.1: Usage Process

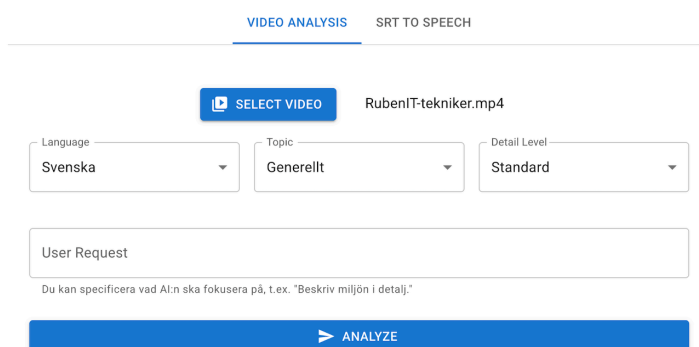
Token limitation

A potential challenge is analyzing extremely long or large video files. Gemini 2.0 Flash is used for analyzing the video content. It processes videos by sampling frames and tokenizing both the frames and audio. It has a token limit of 1 048 576 tokens. This limit is roughly equal to one hour of video. Videos longer than one hour may exceed this limit. This could affect the model's ability to process longer videos effectively.

User Prompt Modifications

Since every video is unique, different aspects may need to be highlighted in the audio descriptions depending on the specific context. To increase the flexibility of the audio description, a user prompt modification tool was introduced. This feature allows users to partially modify the prompt based on

their specific needs. Currently, users can specify the output language, define the video topic, and select the desired level of detail for the audio description. At last, users can extend the base prompt by adding their own input as a user request. This feature allows them to tailor the audio description to better suit their specific needs. However, the introduction of user preferences brings both advantages and limitations. On one hand, user input allows for more tailored and relevant results. On the other hands, an inappropriate or poorly written prompts can interfere with the original prompt. This may lead to incorrect or misleading audio descriptions.



The screenshot shows a web interface for video analysis. At the top, there are two tabs: "VIDEO ANALYSIS" (active) and "SRT TO SPEECH". Below the tabs, there is a blue button labeled "SELECT VIDEO" with a video icon, followed by the text "RubenIT-tekniker.mp4". Underneath, there are three dropdown menus: "Language" set to "Svenska", "Topic" set to "Generellt", and "Detail Level" set to "Standard". Below these is a text input field labeled "User Request" with a placeholder text: "Du kan specificera vad AI:n ska fokusera på, t.ex. 'Beskriv miljön i detalj.'". At the bottom, there is a large blue button labeled "ANALYZE" with a right-pointing arrow.

Figure 4.2: User prompts

Description Modification

After the analysis phase is completed, the system receives an **SRT** file generated by Google Gemini. The file contains the video description, with specific timestamps and corresponding descriptions for each time segment. To improve the user experience, the generated audio description is displayed alongside the video. The subtitles are presented with a black background for a better visibility. Users can easily edit subtitles by clicking on the subtitle, edit the description, and click on the save button. All changes are saved immediately and the displayed subtitle should be visible. Since the prototype is powered by generative AI, the audio description may not always be fully accurate or perfect. The subtitle editing tool allows video creators to manually adjust the text according to their own specific preferences.

Edit SRT Content

```
1
00:00:00,100 --> 00:00:04,180
En blå skärm visas med vit text som lyder: Ruben, IT-support på Service
Desk.
Region Stockholms logotyp visas i det övre högra hörnet.

2
00:00:04,840 --> 00:00:14,230
En person öppnar ett fack med nummer 682 och tar ut en laptop.

3
00:00:14,230 --> 00:00:25,360
Personen går längs en korridor till ett skrivbord med två skärmar.

4
```

Figure 4.3: Description Modification tool

4.2 Text-To-Speech

Improving accessibility involves more than simply providing descriptions in **SRT** format. It is also necessary to include an audio file which an AI voice reads the interpreted text with appropriate pauses. Currently, the system can generate an AI voice using Microsoft Azure **TTS** model to read each description by converting the **SRT** files, but it does so without incorporating any pauses. The system allows users to download the audio file. However, the users may need to use other video editing applications to further work with the file.

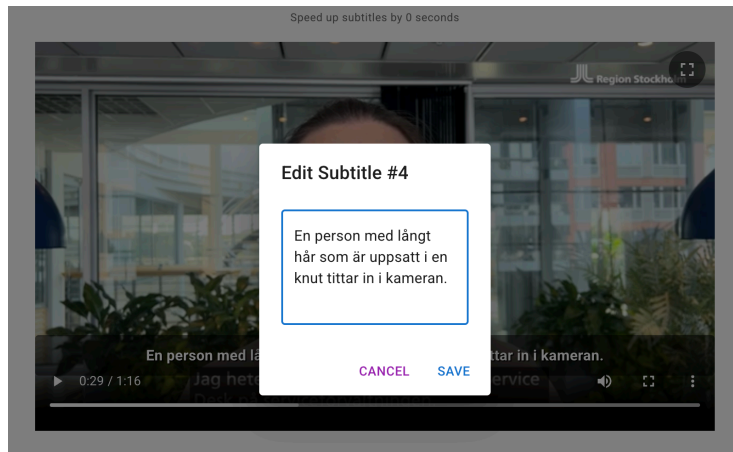


Figure 4.4: Subtitle editing tool

Selection of AI Voices

Regarding the audio component, six different AI voices are available for selection. However, the quality of the audio may be inconsistent, and a slightly noticeable American accent could impact the user experience. At this stage, the prototype relies on free-tier services from various AI-service providers, which limit the overall quality and consistency. With some further investment and access to advanced resources, the limitation can be solved to improve the overall audio experience.

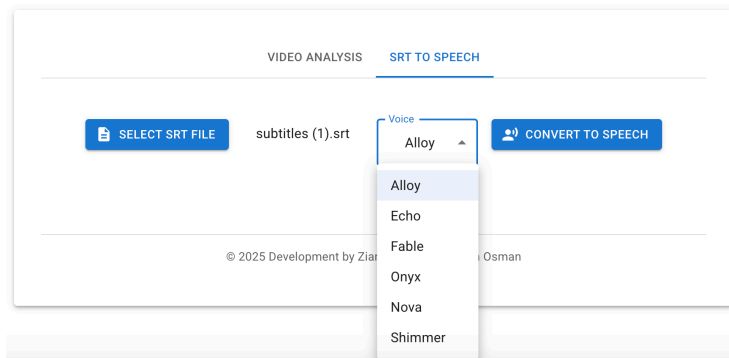


Figure 4.5: Text-To-Speech

Chapter 5

Results and Analysis

After successful implementation, a test was conducted covering the usage process. In this chapter, the results of the survey regarding the conducted test will be presented and analyzed. Due to time constraints, the test was conducted with 5 participants and 4 of them answered the survey. The survey was provided in Swedish, and the response figures shown in this chapter will also be in Swedish. However, the figure captions will be in English to explain the content. Furthermore, additional feedback from the accessibility section of Swedish television will be presented.

5.1 Analysis of survey responses

5.1.1 Quantitative results

Yes/no questions

Figure 5.1 shows that all participants found the generated audio description to be objective.

Var den genererade ljudbeskrivningen **objektiv** (utan personliga åsikter eller tolkningar)?

4 svar

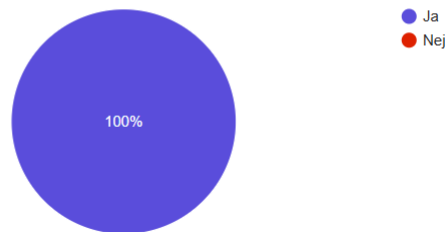


Figure 5.1: The participants' answers regarding the objectivity of the generated audio description.

The results displayed on figure 5.2 shows that all participants agree that the tool could allow more videos to be audio described.

Tror du att verktyget kan bidra till att fler videor kan **syntolkas**?

4 svar

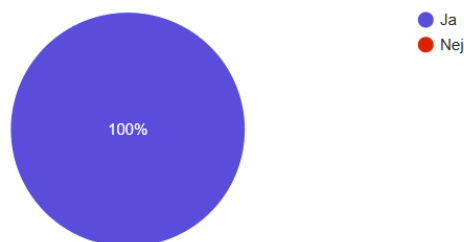


Figure 5.2: The participants' answers regarding the tool's potential to enable more videos to be audio described.

Figure 5.3 shows that all participants would consider using this tool to generate audio descriptions in the future.

Skulle du kunna tänka dig att använda detta verktyg i framtiden?

4 svar

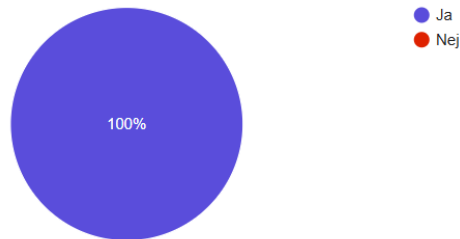


Figure 5.3: The participants' responses on their intention to use the tool in future work.

Figure 5.4 shows that all participants agreed it was easy to modify the generated description.

Det var enkelt att redigera den syntolkade texten.

4 svar

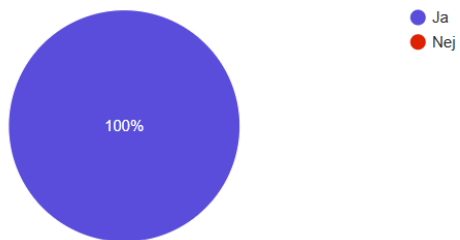


Figure 5.4: The participants' responses on the ease of modifying the generated description.

Linear scale questions

The linear scale ranges from 1-6, where 1 corresponds to "very bad" and 6 is "very good".

Figure 5.5 shows that most participants considered the quality of the generated description (textual content) to be good, with a median rating of 5.



Figure 5.5: Participants' ratings on the quality of the generated description.

Figure 5.6 shows that the participants had mixed opinions about the quality of the generated audio, with responses distributed evenly on the "very bad" and "very good" ends of the scale. The median rating is 3.5.



Figure 5.6: The participants' ratings on the quality of the generated audio.

Figure 5.7 shows an even distribution of responses, with a median rating of 3.5. This may reflect individual preferences for the AI voice and the absence of Swedish AI voices in the selected TTS model.

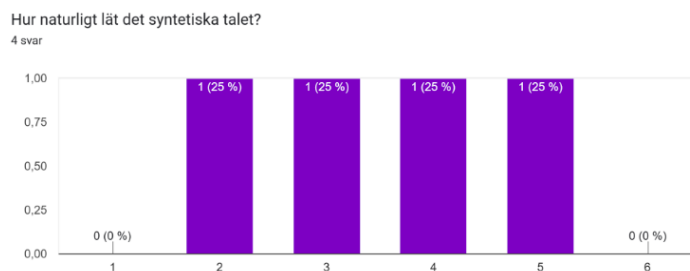


Figure 5.7: Participants' ratings of the perceived naturalness of the generated speech.

5.1.2 Qualitative results

Participants were asked to answer several open-ended questions. Below are selected quotes along with a summary of the main points raised.

What worked well with the tool?

"Very easy to upload film, analyze and edit in the text. Everything was smooth and fast."

"That it did part of the job so that we could build on it."

Most of the participants found the tool easy to use, and the experience was smooth and intuitive. According to them, it helps video creators by simplifying their workflow and makes the video creation process more efficient.

What did not work as well?

"The English accent in generated speech is distracting. I lack an educational tutorial on how to use the tool. Sound cacophony when both audio interpretation and film are played."

"You couldn't hear what the midwife said at all. And neither could you hear what the generated voice said. They were talking over each other. It will be like that even if you shorten it because you need to think about the audio interpretation of a video as soon as you record it. You need to leave a few seconds without speech so that the audio description has time to say something before the person in question starts speaking."

Some valuable feedback was received on the prototype regarding areas that did not work so well, and some points were common. They mentioned that the English accent in the generated speech, which they found somewhat distracting, which can cause a concern in the future use. Furthermore, several users indicated that the audio output was a bit messy, as the soundtrack and the original video sound was playing at the same time, making it difficult to hear either clearly. This is a feature that due to time constraints, wasn't fully completed, but has the potential for future development. At last, two users suggested that the prototype was missing clear instructions showing all the steps involved in editing description subtitles within the box.

What suggestions do you have for future features or improvements?

”Let the text remain a little longer in certain places. As for the sound, you could perhaps let the soundtrack play when there is no sound from the film, or lower the volume of the film so that the soundtrack drowns out the sound of the film.”

”Develop an editor where it is possible to freeze a frame for as long as the audio interpretation requires. And then export a new version.”

According to the feedback from the previous question, the potential improvements should mainly focus on finding better synchronization between the generated speech and the audio of the video. Apart from this, some users wanted a more advanced editor which would provide an option to freeze a video frame for the duration required by the audio description.

Any other comments or opinions you would like to share?

”Good job! I think it will help more videos be subtitled.”

”Good thing we’re working to reach more people! It was difficult to rate the quality of the generated speech and whether it sounded natural because it was hard to hear. It was above the midwife’s speech.”

Overall, the majority of users provided positive feedback on the prototype. They appreciated its potential to increase accessibility for visually impaired users and simplify the video creator’s workflow. However, since the combination of each soundtrack is not yet perfectly set, some users found it difficult to rate the quality of the generated speech, as the soundtracks were distracting each other. This is a point that should be considered for addressing later.

5.2 Additional Results

Sveriges Television (SVT)

A meeting was held at **SVT**’s head office to discuss potential improvements to the prototype. Their collaboration is key for the research study and it also

ensures the feedback received is of high quality.

SVT was impressed by the overall performance of the prototype. The prototype was specifically adapted for visually impaired users. It also gives the video creator a platform to adjust the audio description based on their own needs. Beside this, **SVT** also praised the strong modularity of the prototype and the quality of the audio descriptions. The flexibility is important as the AI market changes quickly due to political and economical factors. The prototype maintains low coupling between modules, allowing companies to easily switch AI service providers if necessary. The audio description remains neutral and bias-free, it captures small details, from paper notes and various items on a box to changes in a person's emotions over time. **SVT** was very satisfied with the overall results at the end, and they are interested in following the future progress of this prototype.

5.3 Reliability Analysis

The method used and data collected are reliable. All participants used the tool in the same way and followed the same process. They were provided the same questions and context about the tool, ensuring that they did not experience the test differently.

5.4 Validity Analysis

The method and data should be considered valid. The users could not see what other users had answered previously, ensuring that their answers and opinions were not affected by others. Moreover, the combining quantitative and qualitative questions led to better insights from intended users, in turn increasing the validity of the data.

Chapter 6

Discussion

The research was successful, a working prototype was developed and tests were conducted, providing insights from intended users at Region Stockholm. Furthermore, the prototype was presented at the **SVT** headquarters, allowing us to gather their opinions and suggestions, gaining additional valuable insights.

The feedback that was gained from the tests were mainly positive. Everyone who attended the demonstration and tested the prototype agreed on its potential positive impact on the creation of audio descriptions. However, there were some negative experiences in the tests and a lot of considerations that were brought up.

Firstly, the audio description could not be played smoothly over the original video, making the audio messy. The reason behind this is mainly that the audio synchronization functionality was not completed in this project. For now, the users would have to download the audio and integrate it separately in their go-to video editor software.

Secondly, although it is possible to edit the timings of the subtitles and modify the description, participants of the tests highlighted that it could be slightly improved to further improve the user-experience. One pointed out that the description modification should be in one box rather than two. Meaning that it would be smoother to avoid using a pop up box for modification of generated description.

Thirdly, two participants mentioned that instructions and a guidance should be included. Even though the overall experience was that the tool was intuitive,

certain features and functionalities could be considered advanced and hard to understand. A tutorial or additional guidance of specific functionalities would help ensure that any user can understand the tool, regardless of their technical background.

Lastly, participants in the demonstration for **SVT** suggested more functionalities of the tool. For instance, the ability to download the description as pure text to use as a script. This would allow some users to automate parts of the process in creating audio descriptions, and then continue with recording the speech manually. This could be useful in scenarios where manual narration is preferred.

Chapter 7

Conclusions and Future work

This chapter wraps up the project by summarizing the main outcomes and reflecting on what was achieved. It also talks about the challenges faced and points out areas where there's room for improvement. Finally, it highlights possible directions for future work and shares some overall thoughts on the project's impact.

7.1 Conclusions

To conclude this research, a working prototype that automates large parts of creating audio descriptions was successfully developed, tests were conducted with the intended users, and a demonstration was held for a major actor within the Swedish media industry. This allowed us to gain insights from multiple people who actively work with public video material, giving valuable considerations, suggestions, and feedback. This means that all the sub goals presented in section 1.4 were met.

The goals were general, due to the lack of previous experience of working with audio descriptions. Also, the requirements for the project from Region Stockholm were not specific, the goal was simply to build a foundation and be able to show a proof of concept, as well as evaluate future opportunities.

The results presented in chapter 5 displayed a mostly positive response. All participants of the test believe that continued development of the tool could allow them to produce more audio described videos, which could in turn enhance accessibility for visually impaired individuals. The participants agreed that the services the tool offers were quick and easy to utilize, meaning

that the process of creating audio descriptions was successfully streamlined. Furthermore, the feedback gathered from **SVT** shows that even major actors are interested and optimistic about how AI could enhance accessibility.

Our suggestions for others working in this area is to thoroughly study how audio descriptions are currently created. At first, it was believed that the whole process should be automated, including syncing the audio with the video. However, while it is true that some test participants mentioned that they would want this functionality improved, other insights shows that many users want partial automation, and continue from there on their own. A fully automated solution will not guarantee the result they desire. With that said, study how the professionals are currently working to build tools that actually support their workflow.

7.2 Limitations

The project was large, but the time was limited. As a result, our efforts were limited, and some originally planned components had to be omitted. For instance, correct timings between the descriptions in the audio were not fully implemented, which meant that the syncing functionality was not done. Also, **SRF** could not participate in this research and assist with tests with visually impaired users. This limited our insights in how visually impaired users would experience the generated audio descriptions. Because of this, it was not possible to capture their opinions about the prototype, and their considerations for future work. However, these limitations were anticipated. As a result, none of this was formally included in our goals, as they were kept more general which was explained in section **7.1**.

7.3 Future work

Due to the large scale of the project, not everything that was initially intended could be done in this research. In this section, the parts that was left undone as well as the next steps in the project will be explained.

7.3.1 What has been left undone?

Test with Visually Impaired Users

While a **User Experience (UX)** test was conducted with content creators, the initial intention was also to test the system with visually impaired users. However, this could not be carried out within the scope of this thesis due to practical constraints. Some insights were still gathered through the pre-study, but a full user test with the target audience remains an important step for future work.

Complete Automation

At first, the intention was to automate the whole process of creating audio descriptions. However, this approach proved to be both complicated and inefficient in practice. Automatically integrating the audio description to the original video would likely lead to undesired results, such as overlapping or interrupting important parts of the original audio. Instead, the generated audio track was provided separately, giving creators control over the final integration. This strikes a better balance between automation and quality.

7.3.2 Next steps of the project

Cost analysis

The prototype does work as intended, but the performance from a cost perspective is necessary in order to determine its practicality. Examples of cost factors could include, deployment, data storage, and choice of AI models.

Due to the research being carried out on a free budget, free plans and student subscriptions were utilized. A detailed cost analysis will be performed if the project continues with funding.

Audio Synchronization

While a fully automated audio description system is believed to produce bad results, implementing audio synchronization is still vital. The functionality was not completed, and feedback from test participants emphasized that it should be included. Furthermore, audio synchronization would simplify the work process of integrating the speech with original audio, and also provide a clearer understanding of the audio timing.

Potential for an Integrated Video Editor

Exploring the potential of integrating a video editor tool into the prototype could prove rewarding. It could allow users to integrate the audio with the original video, directly after generating the audio description. This would further streamline the workflow and eliminate the need for external tools to integrate the audio.

Prompt Engineering and User Preferences

Crafting and testing new prompts allows comparison of outputs in terms of consistency. Additionally, different prompts could lead to generated results being more influenced by user preferences. While the current prompt has produced descriptions of sufficient quality, other prompts could potentially improve relevance and better match the user preferences, enhancing the overall user experience.

7.4 Reflections

The project displays a meaningful utilization of AI services, streamlining the process of creating audio descriptions for videos. From a social perspective, it could enhance accessibility of videos and thereby contribute to inclusion for visually impaired individuals.

Environmentally, it is important to acknowledge that AI services have a carbon and water footprint. However, given the clear social benefit, this use case could justify the trade off. In our opinion, there has to be a balanced view between environmental impact and the value of social sustainability.

Economically, while AI services such as google's Gemini and OpenAI's GPT have operational costs, they offer a scalable alternative to the otherwise expensive and labor-intensive manual production of audio descriptions.

Ethically, the system would not replace sight interpreters role in creating audio descriptions, it simply streamlines their work. By being able to automate the video analysis, generating descriptions, generating the speech, more videos could be provided with audio descriptions. This ultimately makes their work more efficient. Furthermore, the utilization of the system in this project concerns public video material. Private data introduces more ethical concerns, as AI services are carried out by third parties.

The modularity of our system also allows for flexibility, allowing developers to implement compatibility for other AI service providers without affecting other parts of the system. This means that switching AI service providers for economical or ethical reasons is doable as long as the new AI service provider can perform the required services.

References

- [1] R. Stockholm, "Ai-projekt och ai-initiativ i region stockholm," vol. RFC 791 (Standard), n.d. [Online]. Available: <https://www.regionstockholm.se/om-region-stockholm/sa-jobbar-vi-med-ai/ai-projekt-och-ai-initiativ-i-region-stockholm/> [Page 1.]
- [2] National Center for State Courts, "Interpreter code of ethics," 2025, accessed: 2025-03-27. [Online]. Available: https://www.ncsc.org/__data/assets/pdf_file/0022/19606/interpreter-code-of-ethics.pdf [Page 6.]
- [3] J. Thier and F. Editors. (2025, feb) California wildfires raise alarm on water-guzzling ai like chatgpt. Accessed: 2025-03-27. [Online]. Available: <https://fortune.com/article/how-much-water-does-ai-use/> [Page 6.]
- [4] Region Stockholm, "Vårt uppdrag," 2025, accessed: 2025-05-13. [Online]. Available: <https://www.regionstockholm.se/om-region-stockholm/det-har-gor-region-stockholm/vart-uppdrag/> [Page 8.]
- [5] —, "Hållbarhet," 2025, accessed: 2025-05-13. [Online]. Available: <https://www.regionstockholm.se/hallbarhet/> [Page 8.]
- [6] —, "Funktionshinderfrågor," 2025, accessed: 2025-05-13. [Online]. Available: <https://www.regionstockholm.se/om-region-stockholm/hallbarhet/social-hallbarhet/funktionshinderfragor/> [Page 9.]
- [7] —, "Miljöarbete i region stockholm," 2025, accessed: 2025-05-13. [Online]. Available: <https://www.regionstockholm.se/om-region-stockholm/hallbarhet/miljo/> [Page 9.]
- [8] Sveriges Riksdag, "Lag (2023:254) om vissa produkters och tjänsters tillgänglighet," 2023, accessed: 2025-05-13. [Online]. Available: <https://www.riksdagen.se/sv/dokument-och-lagar/dokument/svensk-for>

- [fattningssamling/lag-2023254-om-vissa-produkters-och-tjansters_sfs-2023-254/](#) [Page 9.]
- [9] P. och telestyrelsen, “Lagen om vissa produkters och tjänsters tillgänglighet,” 2024, accessed: 2025-05-13. [Online]. Available: <https://pts.se/digital-inkludering/lagen-om-vissa-produkters-och-tjansters-tillganglighet/> [Page 9.]
- [10] Synskadades Riksförbund (SRF), “Vem är synskadad?” 2024, accessed: 2025-05-13. [Online]. Available: <https://srf.nu/synskador/om-synskador/vem-ar-synskadad/> [Page 9.]
- [11] S. S. Gotland, “Region stockholm måste följa lagen om tillgänglighet till digital offentlig service (dos),” 2025, accessed: 2025-05-13. [Online]. Available: <http://www.srfstockholmgotland.se/Pressrum/Uttalanden/Region-Stockholm-maste-folja-lagen-om-tillganglighet-till-digital-offentlig-service-DOS-/> [Page 9.]
- [12] S. Brown, “Machine learning, explained,” 2021, accessed: 2025-05-13. [Online]. Available: <https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained> [Page 10.]
- [13] C. Stryker and M. Scapicchio, “What is generative ai?” 2025, accessed: 2025-05-13. [Online]. Available: <https://www.ibm.com/think/topics/generative-ai> [Page 10.]
- [14] Microsoft, “What are large language models (llms)?” 2025, accessed: 2025-05-13. [Online]. Available: <https://azure.microsoft.com/sv-se/resources/cloud-computing-dictionary/what-are-large-language-models-llms> [Page 10.]
- [15] M. Azure, “What is computer vision?” 2025, accessed: 2025-05-13. [Online]. Available: <https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-is-computer-vision> [Page 10.]
- [16] C. Stryker, “What is multimodal ai?” 2024, accessed: 2025-05-13. [Online]. Available: <https://www.ibm.com/think/topics/multimodal-ai> [Page 11.]
- [17] A. W. Services, “What is prompt engineering?” 2025, accessed: 2025-05-13. [Online]. Available: <https://aws.amazon.com/what-is/prompt-engineering/> [Page 11.]

- [18] Y. Wang, W. Liang, H. Huang, Y. Zhang, D. Li, and L.-F. Yu, “Toward automatic audio description generation for accessible videos,” in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. New York, NY, USA: Association for Computing Machinery, 2021. doi: 10.1145/3411764.3445347 pp. 1–12. [Online]. Available: <https://doi.org/10.1145/3411764.3445347> [Page 12.]
- [19] R. E. G. Penuela, J. Collins, C. Bennett, and S. Azenkot, “Exploring ai-based scene description tools with blind and low vision users,” vol. 8, 2024. [Online]. Available: <https://dl.acm.org/doi/fullHtml/10.1145/3613904.3642211> [Page 12.]
- [20] ViddyScribe, “Viddyscribe: Video script and captioning tool,” 2025, accessed: 2025-05-13. [Online]. Available: <https://www.viddyscribe.com/> [Page 13.]
- [21] Audible Sight, “Audible sight: Ai-powered audio description for videos,” 2025, accessed: 2025-05-13. [Online]. Available: <https://www.audiblesight.ai/> [Page 13.]
- [22] Google AI, “Gemini api models documentation,” 2025. [Online]. Available: <https://ai.google.dev/gemini-api/docs/models> [Page 15.]
- [23] OpenAI, “Gpt-4,” 2025. [Online]. Available: <https://openai.com/index/gpt-4/> [Page 16.]
- [24] Microsoft, “Azure video indexer overview,” 2025. [Online]. Available: <https://learn.microsoft.com/en-us/azure/azure-video-indexer/video-indexer-overview> [Page 16.]
- [25] —, “Azure account purchase options,” 2025. [Online]. Available: <https://azure.microsoft.com/en-us/pricing/purchase-options/azure-account?icid=pricing-calculator> [Page 16.]
- [26] P. AB, “Tillgänglig video,” jun 2020. [Online]. Available: www.popularte.se [Page 17.]
- [27] M. stad, “Rundtur på tygelsjöbiblioteket,” 2025. [Online]. Available: <https://malmo.se/Uppleva-och-gora/Biblioteken/Vara-bibliotek/Tygelsjobiblioteket/For-barn-och-unga/Rundtur-pa-Tygelsjobiblioteket.html> [Page 17.]

TRITA – EECS-EX-2025:512
Stockholm, Sverige 2025

www.kth.se