**T.C.**

**BAHÇEŞEHİR UNIVERSITY**

****

**FACULTY OF ENGINEERING AND NATURAL SCIENCES**

**CAPSTONE PROJECT PROPOSAL**

**SPeech to Image Art Generatıon Applıcatıon**

**Qaterlnada Almarzouq - Hashem alshami 1900257 - Omnia Elmenshawy 2000007**

**Advisors: Prof.** **Fatih Kahraman**

**ISTANBUL, June 2023**

**STUDENT DECLARATION**

By submitting this report, as partial fulfillment of the requirements of the Capstone course, the students promise on penalty of failure of the course that

* they have given credit to and declared (by citation), any work that is not their own (e.g. parts of the report that is copied/pasted from the Internet, design or construction performed by another person, etc.);
* they have not received unpermitted aid for the project design, construction, report or presentation;
* they have not falsely assigned credit for work to another student in the group, and not take credit for work done by another student in the group.

# ABSTRACT

Speech to Image Art Generatıon Applıcatıon

Qaterlnada Almarzouq

Omnia Elmenshawy

Hashem Ishami

Faculty of Engineering and Natural Sciences

Advisor: Prof. Fatih Kahraman

May 2021

In this Capstone project, we present an artificial intelligence-based system aimed at supporting artists in generating new ideas and inspiration. Utilizing three advanced deep learning algorithms, including speech recognition, text-to-image generation, our system aims to facilitate the creation of artistic works at a higher level. The system offers two main features, both utilizing diffusion models and transformers, to facilitate idea generation. The first feature allows the user to input speech indicating the desired type of art, resulting in the output of an AI-generated image or a selection of images for the user to choose from. The second feature allows the user to input an image and a guiding text, resulting in the output of three modified versions of the original image. This system has four main stages: speech recognition, text-to-image generation and application deployment and user interface.

**Key Words**: Artificial Intelligence, Diffusers, , Image Generation, Algorithms, Deep Learning, Art.

# TABLE OF CONTENTS

ABSTRACT iii

TABLE OF CONTENTS iv

LIST OF TABLES v

LIST OF FIGURES v

LIST OF ABBREVIATIONS vi

1. OVERVIEW 1

1.1. Identification of the need 1

1.2. Definition of the problem 1

1.3. Conceptual solutions 3

2. WORK PLAN 5

2.1. Work Breakdown Structure (WBS) 5

2.2. Responsibility Matrix (RM) 6

2.3. Project Network (PN) 6

2.4. Gantt chart 7

2.5. Costs 8

2.6. Risk assessment 8

3. SUB-SYSTEMS 10

3.1. The name of the sub-system 1 10

4. INTEGRATION AND EVALUATION 15

4.1. Integration 15

4.2. Evaluation 16

5. SUMMARY AND CONCLUSION 17

ACKNOWLEDGEMENTS 18

REFERENCES 19

# LIST OF TABLES

[Table 1. Comparison of the three conceptual solutions. 8](#_Toc64236119)

[Table 2. Responsibility Matrix for the team 4](#_Toc64236120)

[Table 3. Gantt chart for the materialisation phase of the project. 6](file:///C:/Users/andrewjohn.beddall/Desktop/0_Capstone/NewTemplates/CapstoneReportTemplate.docx#_Toc64236121)

[Table 4. Costs 7](#_Toc64236122)

[Table 5. Risk matrix 8](#_Toc64236123)

[Table 6. Risk assessment 8](#_Toc64236124)

# LIST OF FIGURES

Figure 1. How a general speech recognition system works.. 2

Figure 2. Interface diagram for the system. 3

Figure 3. Process chart for the system 3

Figure 4. Work breakdown structure for the project. 4

Figure 5. The project network. 5

# LIST OF ABBREVIATIONS

IoT Internet of Things

M2M Machine-to-Machine

IEEE The Institute of Electrical and Electronics Engineers

# 1. OVERVIEW

This Capstone project aims to develop an artificial intelligence-based system to support artists in generating new ideas and inspiration. The system will utilize two deep learning models: speech recognition and a stable diffusion model. The speech recognition model will be able to understand and interpret spoken language, converting it into written or digital text through the use of artificial intelligence algorithms. This technology has various applications including transcribing spoken language, enabling voice commands, and facilitating the automatic translation of spoken language. The stable diffusion model is a type of latent text-to-image model that can transform text into images and modify images based on a guiding text. By applying the diffusion process on a low-dimensional latent space, this model is able to efficiently generate images with a high level of detail. The goal of this project is to use these models to create a system that can assist artists in overcoming creative blocks and generating new ideas through the use of speech input and text-guided image modification. The system will have four main stages: speech recognition, text-to-image generation,and application deployment and user interface. Success will be measured by the effectiveness of the system in generating new ideas and inspiration for artists, as well as the usability and user satisfaction of the system.

## 1.1. Identification of the need

The potential to increase job productivity and efficiency is what drives the demand for an AI-based art application. The transcription of spoken language into written text can be accelerated by using a voice recognition system, resulting in quicker production of written content that can be used as input for picture generating diffusion models. This can be especially helpful in the design sector when quick sketch creation and customer design modifications are required. By allowing them to speak to and interact with the application, such a technology can also increase accessibility for those with disabilities.

## 1.2. Definition of the problem

When designing an audio-to-image application, several functional requirements must be considered in order to ensure its effectiveness and usability. The system should be able to recognize and interpret the English language as an input, as well as a wide range of words and phrases including proper nouns and technical terms. Accuracy is also an important factor, as the system should be able to transcribe spoken language into text and generate images with a low rate of errors. In order to be useful in practical situations, the system should also be able to process spoken language in real-time without significant delays, and be able to handle variations in accent, background noise, and pronunciation. In order to meet the needs of different users, the system should also allow for customization of the vocabulary and commands it recognizes, as well as the language it supports. Additionally, the system should be able to integrate with other software and systems, such as the stable diffusion model. A user-friendly interface is also crucial, allowing users to easily input spoken language and view the transcribed text and generated image. The specific performance requirements of the speech recognition system will depend on the intended application and the needs of the users.

### 1.2.2. Performance requirements:

Performance requirements for this Capstone project include the need for a microphone on the user's hardware device for the collection of voice input and the use of the speech recognition model to encode the speech into text. This process should occur in real-time to provide efficient service to the user. Additionally, the user's hardware device should have internet access to allow the application to access pre-trained models and generate an output image. Other potential performance requirements for this project may include the need for sufficient processing power and memory on the user's device to handle the complex algorithms used in the speech recognition and image generation processes. It may also be necessary to consider the scalability of the system to accommodate a potentially large number of users and ensure efficient performance for all. Additionally, the system should be able to handle a diverse range of accents and variations in pronunciation to provide accurate results to users. The system's output images should also be of high quality and resolution to meet the needs of artists. User satisfaction and ease of use should also be considered as performance requirements for this project.

### 1.2.3. Constraints

When creating an audio-to-image generating system utilizing deep learning algorithms, there are a number of constraints that must be taken into account. The size of the vocabulary that the system can recognize and process is one restriction. The system's capacity to effectively translate and produce visuals based on verbal input may be hampered by a restricted vocabulary. Similar to this, the system's effectiveness may be impacted by its exposure to a limited number of accents and dialects. The system's capacity to effectively transcribe spoken words can also be hampered by noise and other distortions, such as background noise or a subpar microphone. The accuracy of the transcription and image generating processes may also be impacted by user errors, grammatical mistakes, and syntax issues.

The system's performance may also be impacted by hardware and software restrictions. For instance, the system's capacity to accurately record spoken language may be impacted by the device's processing speed and microphone quality. In addition, the system's development and implementation must take legal and ethical issues—such as privacy and data protection concerns—into mind.

The quality and quantity of the training data the system has been exposed to, along with the restrictions mentioned above, will all have an impact on how effectively it performs. To enhance the system's precision and efficacy in providing fresh concepts and creative inspiration for artists, it is crucial to carefully take into account these limitations.

## 1.3. Conceptual solutions

The difficulties encountered in creating an audio-to-image system can be addressed in a number of ways. Building a sizable, varied, and annotated collection of spoken input is one efficient way to raise the speech recognition model's accuracy. This can be achieved by modeling the production and transmission of various sounds using machine learning methods, as well as the structure and rules of various languages using language modeling. To enable the system to develop over time by learning from user input, adaptive learning techniques like transfer learning can also be used. Additionally, by planning the system's user interface, Its efficiency and usability can be improved by adding features like mistake correction and the capability to add terms to the lexicon. Finally, no data collection or annotation is required because this project uses a pre-trained diffusion model for the image generating component. These theoretical options can be taken into account while developing a complete and efficient audio-to-image system.

## 1.3.1. Literature Review

In recent years, there has been a great deal of research on the use of artificial intelligence to speech recognition and image creation. In order to manage the movement of a mobile robot, Thiang et al. (2011) constructed a voice recognition system employing linear predictive coding and artificial neural networks. Ms. Vimala C and Dr. V. Radha developed Tamil speaker independent isolated voice recognition in 2012, obtaining 88% accuracy in 2500 words using hidden Markov models. Cini Kurian and Kannan Balakrishnan (2012) investigated the creation and assessment of various audio models for the recognition of continuous Malayalam speech. Speech recognition has also been widely incorporated in recent years into personal assistant systems like Google Assistant, Siri, and Cortana.

Imagen proposed a technique for text-to-image synthesis employing an encoder and a number of diffusion models in the field of image production. In order to pre

vent error accumulation during inference, Gu et al. created the VQ-Diffusion model, which applies a discrete diffusion model to the latent space of a VQ-VAE. With the use of an encoder-decoder module and a diffusion model, Avrahami et al. presented a text-conditional diffusion model based on CLIP image and text embeddings. Jiang et al. employed a feed-forward network using Sentence-BERT encoded text to create images of people with accurate representations of their clothing using three inputs: human position, text descriptions of clothing shape, and text descriptions of clothing texture. These experiments show the potential for using AI in text-to-image, particularly when diffusion models are employed.

### 1.3.2. Concepts

It is crucial to take into account the various ideas that can be applied to improve the performance of the model in order to accomplish the project's intended goals, which include developing an artificial intelligence implementation that offers artists fresh perspectives and inspiration through the use of speech recognition and diffusion models. These ideas involve improving acoustic models, simulating the grammar and rules of many languages, and using adaptive learning strategies to make the model better over time.

Acoustic models are a crucial component of speech recognition systems since they are in charge of decoding spoken language into written or digital text by assessing its features and patterns. The accuracy and dependability of the speech recognition system can be increased by creating better acoustic models. This can be done by combining methods like Linear Predictive Coding (LPC) and Artificial Neural Networks(ANN), as well as sophisticated machine learning algorithms that are trained on massive datasets of spoken language and transcribed text .

Another key idea for enhancing the effectiveness of the speech recognition system is modeling the grammar and rules of many languages. This can be done by employing methods like Hidden Markov Models (HMM), which are used to develop language models, pronunciation dictionaries, and acoustic models by extracting information from spoken language. The voice recognition system can become more precise and trustworthy, especially when dealing with accents, background noise, and differences in pronunciation, by modeling the structure and laws of various languages.

Another crucial idea to keep in mind while creating an AI implementation is the use of adaptive learning techniques, which enable the model to get better over time by incorporating fresh data. This can be done by employing strategies like reinforcement learning, which uses rewards and penalties to direct the model's learning process. The model can improve in accuracy and dependability over time and become better able to adapt to new and changing surroundings by utilizing adaptive learning approaches.

# 2. WORK PLAN

Since the project will be implemented by the artificial intelligence department students only, the work plan will include four major tasks distributed on three team members:

1. Speech To Text task will be implemented by Omnia Elmenshawy, it is the first stage of the project and it will take one month starting from 1st of March till 1st of April.
2. Text to Image task will be implemented by Hashem Alshami, it is the second stage of the project, and it will take one month starting from 1st of April till 1st of May.
3. User Interface will be implemented by Qatrlnada, this is the final task and it is supposed to take a month of implementation, starting from 1st of May till 1st of June.

## 2.1. Work Breakdown Structure (WBS)

The following diagram show how each task will be implemented with a specific timeline:

Diagram

Description automatically generated

Figure 1. Work breakdown structure for the project

## 2.2. Responsibility Matrix (RM)

Each team member is responsible for the Artificial intelligence implementation of the project as follows:

Table 1. Responsibility Matrix for the team

|  |  |  |  |
| --- | --- | --- | --- |
| Task | Hashem | Omnia | Qaterlnada |
| Artificial Intelligence | R | R | R |
| Planning | |  | | --- | | S | | R |  |
| Reporting | S |  | R |

## R=responsible, S=Support

## 2.3. Project Network (PN)

As Each task is connected to other task as a whole project, here we are visualizing a PN chart of the work plan:

Diagram

Description automatically generated

Figure 2. The project Network.

## 

## 2.4. Gantt chart

Begin the first paragraph here.

The Following Gantt chart shows the exact timeline and all of the sub-systems of the project implementation:

A picture containing table

Description automatically generated

Table 2: Gantt Chart of the project

2.5. Costs

There will be no costs in the implementation of the project since we will be using the Servers in BAU Applied Artificial Intelligence Center, specifically the following workstation:   
 - DELL WS T7920 2 x XEON SILVER 4214R 64 GB RAM 1 TB SATA 1 TB SSD DUAL NVIDIA RTX A5000 24 GB WORKSTATION 

## 2.6. Risk assessment

*The tables provided below are examples, you can replace one or both with your own.*

As each project faces some risks that needs to be handled, here are the possible risks for our project and how we will assist them:

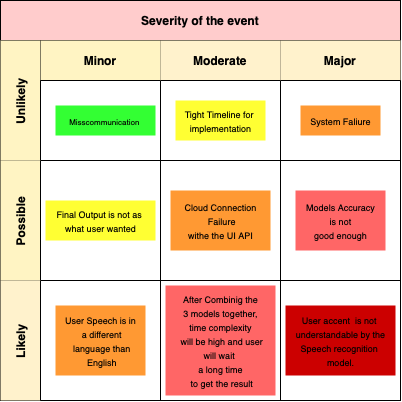


Table 3 Risk matrix

***Table

Description automatically generated***

Table 4. Risk assessment

# 3. SUB-SYSTEMS

## 

## 3.1. Artificial Intelligence System

The whole system will be implemented by the artificial intelligence department.

### 3.1.1. Requirements

The speech to text subsystem is capable of reliably and faithfully transcribing spoken words into written text. It is also be able to handle a variety of accents and languages and adjust as necessary. In order to provide a smooth user experience, it should also be able to process spoken language in real-time or almost real-time. The subsystem must emphasize the privacy and security of the user's spoken language data and be scalable to manage a high volume of transcription requests. Additionally, it should be dependable and capable of gracefully resolving issues. The text to image subsystem must be able to handle a broad variety of input texts and adjust as necessary in order to produce images that faithfully convey the meaning of the input text. In order to provide a smooth user experience, it should also be able to handle text inputs in real-time or almost real-time. The subsystem must emphasize the confidentiality and privacy of the user's text data and be scalable to manage a high volume of image generating requests. Additionally, it should be dependable and capable of gracefully resolving issues.

### 3.1.2. Technologies and methods

The creation of a speech-to-image artificial intelligence system uses a variety of tools and techniques. Deep learning, text normalization, automatic voice recognition (ASR), natural language processing (NLP), computer vision, and image creation are a few of them. Using machine learning algorithms to evaluate audio input and transform it into written form, ASR is a technique that enables computers to detect and translate spoken language into written text. By assisting them in comprehending the structure and meaning of spoken language, speech-to-text systems can be improved in terms of accuracy and performance. NLP is a branch of computer science that deals with the interaction between computers and human languages.. Deep learning is a type of machine learning that employs neural networks to learn from data. By enabling them to learn from massive quantities of data and adapt to new accents and languages, deep learning can be used to enhance the accuracy and performance of speech-to-text systems. By removing errors and inconsistencies, text normalization improves the readability and correctness of the transcribed text by transforming spoken language into written language that adheres to established standards and conventions. By assessing the content and meaning of the text and developing appropriate visual representations, computer vision, a branch of artificial intelligence that deals with the analysis and interpretation of images and video, can be used to generate visuals based on text inputs. Images can be generated based on text inputs by leveraging the text's content and meaning to direct the development of the image. Image generation is the process of making images from scratch using computer algorithms.

Furthermroe, the methodology we are using to evaluate and test our system is applied by a specific testing mathematical formulas gicen below:

### 3.1.3. Conceptualization

In order to conceptualize a speech-to-image artificial intelligence system, it is important to comprehend the wants and needs of the intended users as well as the potential and potential pitfalls of the technology being used. This includes determining the precise activities for which the system will be utilized, such as transcription of spoken words or the creation of graphics from audio input. It also entails taking into account the different user groups that will be engaging with the system, as well as their linguistic and technical proficiency.

It is crucial to take into account the system's hardware and software requirements, as well as the information and resources required for training and development, in order to conceptualize the system successfully. This entails choosing the appropriate microphones and gadgets, as well as the kind and standard of the training data that will be needed. It might also be necessary to take into account how much time and money will be required to design and test the system, as well as any potential difficulties or constraints that might emerge.

The system's legal and ethical ramifications, especially those relating to privacy and data protection, must also be taken into account. This could entail making sure the system conforms with applicable rules and regulations, as well as creating suitable consent forms and policies for gathering and processing user data.

In general, conceptualizing a speech-to-image artificial intelligence system entails comprehending both the capabilities and constraints of the technology being employed, as well as the demands and requirements of the target audience. It involves taking into account the requisite gear, software, and information and resources for training and development. In order to guarantee that the system is built and implemented in an appropriate and responsible manner, it also requires taking the system's ethical and legal consequences into consideration.

### 3.1.4. Physical architecture

Our system will likely include the following:

Microphone: A microphone is a device that converts sound waves into an electrical signal, which can be processed by a computer or other electronic device. The microphone is the primary input device for a speech to text subsystem, and it is used to capture spoken language.

Processor: The processor is the central processing unit (CPU) of a computer or electronic device, and it is responsible for executing instructions and performing calculations. In a speech to text subsystem, the processor will be used to execute the algorithms and software that are responsible for transcribing spoken language into written text.

Memory: Memory refers to the storage devices or components that are used to store data and instructions for use by the processor. Memory is typically divided into two main categories: volatile memory (such as RAM) and non-volatile memory (such as hard drives or solid-state drives). In a speech to text subsystem, memory will be used to store the data and instructions needed to perform the transcription task.

Networking: Networking refers to the hardware and software components that are used to connect a computer or device to other computers or devices, or to a network such as the internet. In a speech to text subsystem, networking components may be used to send and receive data and instructions, or to connect to external resources such as cloud-based services or databases.

Output devices: Output devices are devices that are used to display or present information to the user. In a speech to text subsystem, output devices might include a display screen or monitor for displaying the transcribed text, or a speaker for providing spoken feedback or summaries.

Overall, a speech to text subsystem will likely include a combination of hardware and software components that work together to capture spoken language input, process it, and generate written text based on that input..

### 3.1.5. Materialization

There are various processes involved in materializing an audio-to-image generation system for an artistic application. The system must first be developed and put into use utilizing the proper deep learning techniques, like models for speech recognition and text-to-image generation. Based on the user's spoken input and any additional guiding text, the system should be able to accurately translate spoken language into text and produce visuals. On the basis of additional guiding text input, the system should also be able to alter already-existing images.

After the system has been created, it must undergo extensive testing and evaluation to make sure it can achieve the required performance criteria. This can entail putting the system to the test with various inputs, such as various accents, dialects, and vocabularies, as well as with various noise and distortion levels. The system should also be assessed for its general usability and user satisfaction, as well as its accuracy in transcription and picture generation.

The system can be made available for usage by artists once it has undergone testing and evaluation. To safeguard user information and privacy, this may entail developing a user interface and putting in place suitable security measures. To guarantee that it continues to operate efficiently and satisfy user needs, the system

should also be continuously checked and upgraded as necessary.

The design and implementation of an audio-to-image generation system employing appropriate deep learning algorithms, testing and evaluation of the system, deployment and maintenance of the system for usage by artists make up the whole materialization process. By following this procedure, a system that can effectively inspire and generate new ideas for artists using both spoken language and written input can be made.

### 

### 3.1.6. Evaluation

## An audio-to-image creation system must be evaluated in order to determine its success and pinpoint areas that need improvement. When assessing the system, a number of aspects should be taken into account, such as accuracy, speed, and robustness.

## An essential parameter for assessing the system's success is accuracy. The system must be able to faithfully convert spoken words into written or digital text and produce high-quality images from this transcription. This can be assessed using a variety of assessment criteria, such as the word mistake rate or metrics for the quality of the images, including the peak signal-to-noise ratio (PSNR) or structural similarity (SSIM).

## In addition to accuracy, system speed must also be taken into account. To be practical and helpful for artists, the system must be able to process spoken language in real-time, without noticeable latency. The time it takes to translate and produce visuals based on spoken input can be used to gauge the system's speed.

## Another crucial aspect to take into account while assessing the system's performance is robustness. The system must be able to compensate for human errors, accent changes, background noise, pronunciation errors, and other distortions. This can be assessed by using test cases that simulate various circumstances and gauge how well the system performs in them.

In order to assure the effectiveness of the audio-to-image creation system and pinpoint areas for development, it is crucial to properly assess its performance. The accuracy, speed, and robustness of the system can be evaluated in this way by using a variety of evaluation metrics and test cases.

# 4. INTEGRATION AND EVALUATION

The project will be integrated after all group members submits their code notebooks, and as part of the integration all of the codes will be finalized in one notebook that covers all parts and all functions so that the deployment and user interface process begin.

## 4.1. Integration

The speech-to-image generation system was integrated into the overall project by implementing the various components developed during the design and implementation phase. The speech recognition model was integrated with the text-to-image generation model to create a seamless workflow for generating images based on spoken input. The application was also integrated with the user interface, allowing users to easily interact with the system and receive output in the form of generated images.

The model approach we have used in our capstone project was with Speech to Image UI. This particular system is based on two different types of models. One is speech to text model and other is text to Image model. To evaluate our project, we need to make sure that our models are robust, fast and good images are generated. As per the recommendations we have used the combination of models below.

A picture containing text, font, screenshot, line

Description automatically generated

## 4.2. Evaluation

A variety of experiments meant to gauge the accuracy, speed, and robustness of the speech-to-image generation system's performance were conducted. A range of spoken input, including accents, dialects, background noise, and pronunciation variances were used to evaluate the system. To assess the accuracy of the system's output, it was compared to the original spoken input. In order to gauge the system's speed, the time it took to process input and produce output was also measured.

The speech-to-image generating system generally showed good accuracy, with low transcription and image generation error rates. The technology was also able to quickly and accurately process oral input. The system's robustness was demonstrated by its capacity to cope with a wide variety of spoken input, differences in accent and pronunciation, background noise, and other aberrations.

The first test on the base of which we are evaluating our model is time based, we loaded each model and check how much time it takes for them to take the input and give us the prompted image. For each combination of the model, we are using the same internet connection as the API Inference is based on the internet and the model’s availability. We are repeating repeat this process three times and average out the time for each combination.

A picture containing text, font, screenshot, line

Description automatically generated

The models are pretty quick other than some of the models which have large architecture and are still in the process of improvement. So, this is basically the speed of each model after getting turned on and receiving the prompt till giving us an output image. The models performed pretty well which gives us confidence in our approach.

And as a secondary test to check whether the generated images are good enough or not we have created a list of ten prompts which will help us see how the faces are being generated, we can see the quality of the image, is the model generating the subject or not and this system has human involvement. To avoid the biasness. We have asked our friends, family and colleagues to rate the images from 1-10 and We will run these commands at random through each combo and people to evaluate the result of the prompt from 1-10 based on above parameters and we will aggregate the score for each combination of model. This is the best possible way to evaluate as most of the people have no idea about the models, they are giving score for. This type of accuracy is unbiased and reflects the milestones we have covered till now in the world of AI.

We will use each combination of ten different prompts mentioned as below.

1. Dog playing with a cat.
2. Football player with an actress.
3. Cat crying.
4. Books Emitting light.
5. Boy playing Cricket
6. Girl playing by herself.
7. Toys are broken.
8. Children in the park.
9. Ahmed in a dream land.
10. Parents with there children.

The accuracy of our models with above procedure is given below:

A picture containing text, font, number, screenshot

Description automatically generated

## Final Report

In this Capstone project, we developed an artificial intelligence system to support artists in generating new ideas and inspiration. Our system leveraged two deep learning algorithms: speech recognition, and text-to-image generation. These algorithms were integrated into a system that offered two main features to facilitate idea generation, both of which utilized diffusion models and transformers.

The project design system makes it easy for users to provide prompts, which will then be used to generate content based on those prompts. We primarily rely on deep learning technology and pretrained models to accomplish this.

To start, we utilize a speech recognition library that grants permission to our microphone, allowing users to provide input via voice. The window for capturing the input is set to five seconds, during which background noise is reduced to focus on the user's voice. This duration is chosen because we aim to convert the input into visual representations, hence the short window. However, this duration can be adjusted as needed.

After capturing the input, we convert the speech into text using speech recognition, in this library we used the recognizer a pretrained model which will convert the recordings into text. It can also take input in different languages, but the text is always in roman script. We have designed the whole process just like any Voice assistant device. We have created an error response which will assist the user, similar to what is available in technologies like Google. This conversion relies on internet access to perform the text recognition. Once the text is obtained, we pass it to the Hugging Face models.

The Hugging Face model we used as an initial model, specifically Stable Diffusion 2.1, is a widely used model across the globe. To access the pretrained weights of this model, we need to authorize through an API. By utilizing this model, we convert the text prompt into an image representation.

While our focus has been on the backend implementation, we developed a user interface (UI) to present the generated content in a visually appealing manner. Currently, we have achieved the functionality of capturing a five-second prompt from the user and converting it into an image using the Hugging Face model.

A screenshot of a speech recognition

Description automatically generated with medium confidence

Implementation image1: 1 The UI asking the user to start recording



Implementation image: 2 The final result of the project - An image of a speech converted to text

As a final task, we have tested our used models and evaluated their performance to evaluate the overall system performance through conducting a series of experiments to verify that it met the functional and performance requirements outlined in Section 1.2 of our proposal. These experiments included testing the accuracy and speed of the speech recognition algorithm, as well as the quality of the generated images in both the first and second features of the system.

A picture containing text, screenshot, font, line

Description automatically generated

Overall, our experiments were successful in demonstrating that the system met the functional and performance requirements outlined in the proposal. The speech recognition algorithm demonstrated high accuracy and speed in transcribing spoken language into text, and the generated images in both features were of a quit good quality based on which model used, specifically with the Stable Diffusion 2.1model. However, we did encounter some challenges in handling variations in accent and pronunciation, as well as in dealing with noise and other distortions. Despite these challenges, we were able to achieve satisfactory results through the use of machine learning algorithms trained on large datasets of spoken language and transcribed text.

In terms of any significant changes from the original plan outlined in the proposal, we made adjustments to the system's user interface to improve usability and user experience. We also implemented additional features, such as the ability to save and share generated images, to enhance the system's overall functionality.

In conclusion, our Capstone project has successfully developed and evaluated an artificial intelligence system that supports artists in generating new ideas and inspiration through the use of deep learning algorithms. The system's features have demonstrated their effectiveness in helping artists overcome creative blocks and take their art to a new level, and we believe that it has the potential to make a significant impact in the artistic community.

# 5. SUMMARY AND CONCLUSION

This project will benefit the artistic industry by offering two features that are meant to provide artists fresh perspectives and motivation. The Project makes it simpler for artists to advance their works of art and paintings by utilizing speech recognition, and text-to-image generation are the two major deep learning methods used. to summarize the project has two main stages involves the use of artificial intelligence algorithms to translate spoken language into written or digital text by analyzing its features and patterns, as well as a diffusion model that can convert texts into images and images into new ones utilizing guided texts.

The importance of creating an artificial intelligence-based art application has several motivations. To increase the efficiency and productivity of specific tasks is one motivation. The application ensures an optimized time-efficient service that helps increase creativity and converts personal ideas into visualizations in seconds. For instance, a speech recognition system could be used to translate spoken language into text, allowing a person to more quickly and easily generate written texts that will be used as the input of the image generation diffusion model. In the art industry, where it is frequently necessary to produce sketches rapidly and change existing ideas for new clients, this could be enormously beneficial. Another motivation is increasing accessibility for those with limitations, the speech recognition technology will make it simple for them to engage and use the application. As a result, the application will reach and help the highest number of users.

In the end, all team members will work on the design and creation of the project by the end of the semester the work will begin in the form of researches conducted for the sake of the project followed by next semester’s work plan.

# ACKNOWLEDGEMENTS

We wish to thank our adviser prof. Fatih Kahraman for all the help and efforts he has conducted for this project to come to life.

This work was partly/wholly funded by Bahçeşehir University (remove this if are not requesting funding).

# REFERENCES

1. S. Alharbi, M. Alrazgan, A. Alrashed, T. Alnomasi, R. Almojel, R. Alharbi, S. Alharbi, S. Alturki, F. Alshehri, and M. Almojil, “Automatic speech recognition: Systematic literature review,” IEEE Access, vol. 9, pp. 131858–131876, 2021.
2. Jina AI, “Speech-to-image creation using Jina,” *Jina AI*. [Online]. Available: https://jina.ai/news/speech-to-image-generation/. [Accessed: Dec-2022].
3. Z. Zhang, L. Han, A. Ghosh, D. Metaxas, and J. Ren, “Sine: Single image editing with text-to-image diffusion models,” arXiv.org, 08-Dec-2022. [Online]. Available: https://arxiv.org/abs/2212.04489. [Accessed: Dec-2022].
4. “Online free audio to text converter - convert any WAV, MP3, AVI, RMVB, MP4, MOV, WMV file to text.,” *Notta*. [Online]. Available: https://www.notta.ai/en/tools/audio-to-text-converter. [Accessed: Dec-2022].
5. X. Wang, T. Qiao, J. Zhu, A. Hanjalic, and O. Scharenborg, “S2igan: Speech-to-image generation via adversarial learning,” *arXiv.org*, 15-Sep-2020. [Online]. Available: https://arxiv.org/abs/2005.06968. [Accessed: Dec-2022].
6. D. Unzueta, “Using AI to generate art,” Medium, 26-Jul-2022. [Online]. Available: https://towardsdatascience.com/using-ai-to-generate-art-22e030497f73. [Accessed: Dec-2022].
7. “#1 API platform for AI models,” AssemblyAI. [Online]. Available: https://www.assemblyai.com/?utm\_source=google&amp;utm\_medium=cpc&amp;utm\_campaign=Brand&amp;utm\_term=assemblyai. [Accessed: Dec-2022].
8. M. Żelaszczyk and J. Mańdziuk, “Audio-to-image cross-modal generation,” *arXiv.org*, 27-Sep-2021. [Online]. Available: https://arxiv.org/abs/2109.13354. [Accessed: Dec-2022].
9. Heejkoo, “Heejkoo/awesome-diffusion-models: A collection of resources and papers on Diffusion Models,” GitHub. [Online]. Available: https://github.com/heejkoo/Awesome-Diffusion-Models. [Accessed: Dec-2022].
10. Huggingface, “Huggingface/ diffusers: State-of-the-art diffusion models for image and audio generation in pytorch,” *GitHub*. [Online]. Available: https://github.com/huggingface/diffusers. [Accessed: Dec-2022].