**MIT School of Engineering**

**Department of Computer Science and Engineering**

**Project Synopsis**

**Project Title:** Text to speech & speech to text

**Problem Statement:**

The current technological advancements have made speech to text and text to speech conversion essential for many applications such as language translation, assistive technology, and automated customer service. This poses a significant challenge to users who require accurate and user-friendly speech to text and text to speech transformation.

To address this challenge, the aim of this project is to design and implement a Java-based speech to text and text to speech transformer that provides accurate and user-friendly conversion for various applications.

The main objective of this project is to provide a robust, scalable, and user-friendly Java-based speech to text and text to speech transformer that can be integrated into various applications. This project will help users with hearing or speech impairments to communicate better and improve the efficiency of language translation and customer service.

**Abstract:**

Text-to-speech (TTS) and speech-to-text (STT) are two important technologies that have revolutionized the way we interact with machines. TTS involves the synthesis of natural-sounding speech from written text, while STT involves converting spoken words into text. Both technologies have numerous applications in various industries, including accessibility, language translation, virtual assistants, and automated customer service.

TTS systems typically involve several components, including text input, text analysis, acoustic modeling, signal processing, and output. The text input is the written text to be converted to speech, and the text analysis involves applying natural language processing techniques to extract linguistic features such as parts of speech, syntax, and semantics. The acoustic modeling stage involves converting the linguistic features into acoustic parameters, which represent the speech waveform. In signal processing, the acoustic parameters are used to generate a synthetic speech waveform. Finally, the synthesized speech is played through a speaker or output device.

STT systems, on the other hand, involve audio input, pre-processing, feature extraction, acoustic modeling, language modeling, and output. The audio input is the spoken language to be converted to text, and pre-processing involves removing background noise and enhancing speech quality. Feature extraction involves extracting spectral and temporal characteristics of the speech signal from the pre-processed audio. In acoustic modeling, a statistical model is trained on a large corpus of speech data to map acoustic features to linguistic units, such as phonemes or words. A language model is then used to predict the most likely sequence of words given the acoustic features. Finally, the text output is generated and can be used for further processing or analysis.

.

**Literature Survey: Detail survey done**

Speech-to-Text Models:

"Listen, Attend and Spell" by William Chan et al. (2016): This paper introduced an attention-based sequence-to-sequence model for speech recognition, which achieved state-of-the-art performance on several benchmark datasets.

"Deep Speech 2: End-to-End Speech Recognition in English and Mandarin" by Dario Amodei et al. (2016): This paper proposed a deep neural network architecture for speech recognition, which improved on the state-of-the-art performance on several datasets.

"Streaming End-to-End Speech Recognition for Mobile Devices" by Tara Sainath et al. (2015): This paper introduced a streaming end-to-end model for speech recognition, which can operate on mobile devices with limited computational resources.

"Recent Advances in Deep Learning for Speech Research at Microsoft" by Frank Seide et al. (2011): This paper presented an overview of recent advances in deep learning for speech recognition and introduced a deep neural network architecture that achieved state-of-the-art performance on several benchmark datasets.

Text-to-Speech Models:

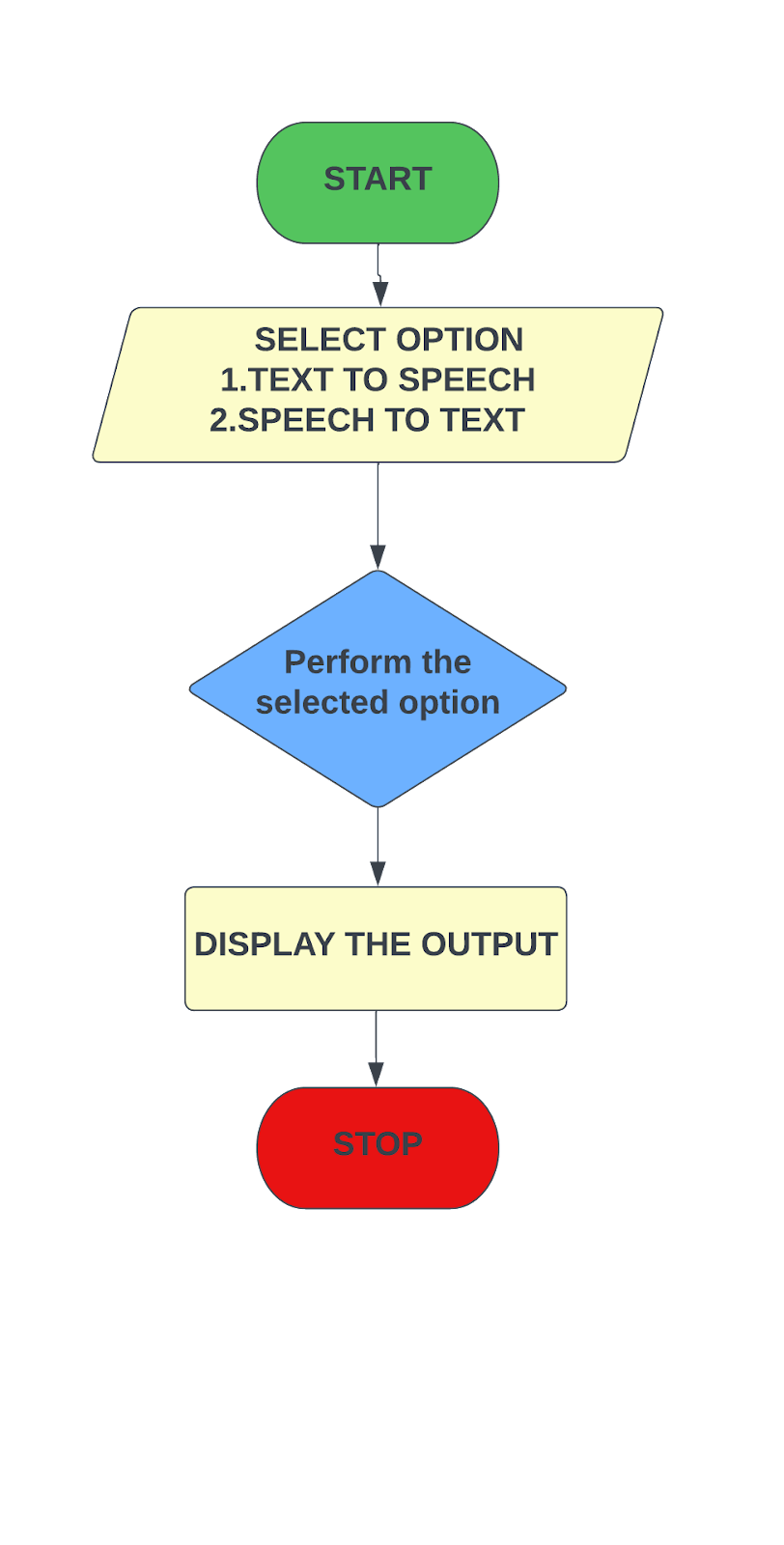
"Tacotron: Towards End-to-End Speech Synthesis" by Jonathan Shen et al. (2017): This paper introduced a sequence-to-sequence model for speech synthesis, which can generate speech with natural-sounding intonation and expressiveness.

"WaveNet: A Generative Model for Raw Audio" by Aaron van den Oord et al. (2016): This paper proposed a deep generative model for speech synthesis, which can generate high-quality speech with a high level of naturalness and expressiveness.

"Deep Voice: Real-time Neural Text-to-Speech" by Sercan Ö. Arık et al. (2017): This paper introduced a deep neural network architecture for text-to-speech, which can generate high-quality speech in real-time.

"Neural Speech Synthesis with Transformer Network" by Wei Ping et al. (2018): This paper proposed a transformer-based architecture for text-to-speech, which achieved state-of-the-art performance on several benchmark datasets.

**Proposed System (Block Diagram):**

****

**Conclusion:**

Text-to-speech (TTS) and Speech-to-text (STT) technologies are two powerful tools that enable computers to process and convert human language into machine-readable format and vice versa. TTS technology can convert written text into spoken words, while STT technology can convert spoken words into written text.

**References:**

1] [www.stackoverflow.com](http://www.stackoverflow.com/)

[2] [www.pythonprogramming.net](http://www.pythonprogramming.net/)

[3] [www.codecademy.com](http://www.codecademy.com/)

[4]<https://en.wikipedia.org/wiki/Face_detection>

[5] <https://en.wikipedia.org/wiki/Facial_recognition_system>

**Annexure:**

**Annexure I: Form A-Title Approval (for offline mode)**

**Project Title Evaluation Parameters:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sr.**  **No.** | **Parameters** | **Topic 1** | **Topic 2** | **Topic 3** |
| 1. | Title | **Speech/Text Transformer** | **Library Management System** | **Music Player** |
| 2. | Domain  Expertise | Java Programming language | Java Programming language | Java Programming language |
| 3. | Technical  Feasibility | In this project we will use Java speech recognition libraries such as Sphinx4 and CMUCLMTK to identify speech patterns. These libraries are designed to provide an easy-to-use, flexible and efficient speech recognition engine.  We will also be using FreeTTS which is a free, open-source Text-to-Speech (TTS) synthesis system written in the Java programming language. It can be used to convert written text into spoken words | We will be using Java object-relational mapping libraries such as Hibernate that simplifies the process of mapping Java objects to database tables.  We will also be using Apache POI which is a Java library that can be used to read and write Microsoft Office documents, such as Excel spreadsheets, which can be useful for storing and retrieving data related to books, authors, and borrowers**.** | We will be using Java libraries used for audio capturing, processing, and playing such as Java Sound API, JLayer, JAudiotagger .  We will also be using Log4j and JUnit which are Java libraries that can be used to log system messages and errors. They can be used for monitoring and debugging the music player. |
| 4. | Future Scope | Improved Accuracy: One of the main challenges in speech-to-text and text-to-speech transformation is achieving high accuracy in the recognition and synthesis of speech. | Mobile Accessibility  Data Analytics  Integration with E-Resources  Sustainability | Integration with AI  Improved User Interface  Integration with Streaming Services  Advanced Audio Processing |
| 5. | Applicability | Convert doctors' dictations into written notes. create audio versions of textbooks or to provide voice feedback for students' written assignments. It can be used to create automated voice responses for phone or chat-based customer service interactions. | library management systems can be used to manage collections of books, journals, and other resources. They can also be used to track borrowing history, manage fines and fees, and provide access to electronic resources.. | Music player apps can be used by individuals for personal use to play music stored on their devices, as well as to stream music from various online sources.Music player software can also be used by professionals in the music industry such as music producers, audio engineers, and DJs to play and manage music tracks during production, recording, or live performances. |
|  | Approved  (✓) | ✓ |  |  |
| **Remark:** | | | | |

**Annexure II: Form B-Market and financial feasibility (verify from guide)**

**Project Title Evaluation Parameters:**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Parameters** | **Description About Project** |
| 1. | Business Ideas and Implementation from project Marks(10) | Product can be used by anyone to convert **Text to Speech or Speech to Text.** |
| 2. | Market Survey (competitors, substitute products, potential market, etc.) Marks(10) | **Text-to-Speech** market was valued at USD 2.06 billion in 2021 and is expected to reach the value of USD 17.01 billion by 2029, at a CAGR of 30.20% during the forecast period of 2022-2029.  **Speech-to-Text** API market size was valued at USD 2.32 billion in 2021 and is expected to expand at a compound annual growth rate (CAGR) exceeding 15.2% from 2022 to 2030 |
| 3. | Market Acceptability of Product Marks(5) | Rising preference for handheld devices and the increasing number of people suffering from visual impairments and learning disabilities propels growth of the global text-to-speech and Speech-to-Text market. In addition, rising adoption of voice assistants & smart speakers to aid growth fuel growth of the text-to-speech and Speech-to-Text market. |
| 4. | Emerging Trends about Project and Product Marks(10) | Emerging trends for Text-to-Speech are Neural Text-to-Speech,Emotional TTS and Multilingual TTS. Emerging trends for Speech-to-Text Voice-based chatbots,voice assistants and Voice Recognition for Translation Applications. |
| 5. | Income Generation ideas through Project Marks(5) | 1. **Transcription services:** STT technology can be used to create transcripts of audio and video recordings. You can offer transcription services to individuals and businesses who need written transcripts of their recordings.  2.  **Captioning and subtitling:** TTS technology can be used to generate captions and subtitles for videos.  3. **Voiceover services:** TTS technology can also be used to create professional voiceovers for videos, commercials, and other content.  4.  **Language translation:** You can offer language translation services to businesses and individuals who want to reach a global audience.  5.  **Personal assistant services:** You can use STT and TTS technology to create a virtual personal assistant service. |
| 6. | Project Profitability Marks(5) | 1.  **Cost of technology:** The cost of acquiring and maintaining STT and TTS technologies can be a significant expense.  2.  **Revenue streams:** There are many revenue streams possible for projects involving STT and TTS technologies, such as transcription services, captioning and subtitling services etc.  3.**Use cases:** There are many use cases for STT and TTS technologies, ranging from business applications to consumer products. It's important to consider the specific use case and target audience when developing and marketing these technologies. |
| 7. | Cost Benefit Analysis Marks(5) | 1. **Costs of Development:** The cost of developing STT and TTS products can vary significantly depending on the complexity of the technology and the level of expertise required. This can include costs associated with research and development, software development, and hardware development.    2.  **Licensing Costs:** If using licensed STT and TTS products, there may be licensing fees and other costs associated with the use of the technology.    3. **Infrastructure Costs**: Depending on the size of the project and the scale of the technology, there may be infrastructure costs associated with the development and deployment of STT and TTS products. This can include costs associated with server infrastructure, networking, and data storage.    4. **Maintenance Costs:** Once the product is deployed, there may be ongoing maintenance costs associated with maintaining and updating the technology. This can include costs associated with bug fixes, upgrades, and customer support. |
|  |  |  |
| **Remark:** | | |

**Annexure III: Literature survey paper or links**

Speech-to-Text Models:

"Listen, Attend and Spell" by William Chan et al. (2016): This paper introduced an attention-based sequence-to-sequence model for speech recognition, which achieved state-of-the-art performance on several benchmark datasets.

"Deep Speech 2: End-to-End Speech Recognition in English and Mandarin" by Dario Amodei et al. (2016): This paper proposed a deep neural network architecture for speech recognition, which improved on the state-of-the-art performance on several datasets.

"Streaming End-to-End Speech Recognition for Mobile Devices" by Tara Sainath et al. (2015): This paper introduced a streaming end-to-end model for speech recognition, which can operate on mobile devices with limited computational resources.

"Recent Advances in Deep Learning for Speech Research at Microsoft" by Frank Seide et al. (2011): This paper presented an overview of recent advances in deep learning for speech recognition and introduced a deep neural network architecture that achieved state-of-the-art performance on several benchmark datasets.

Text-to-Speech Models:

"Tacotron: Towards End-to-End Speech Synthesis" by Jonathan Shen et al. (2017): This paper introduced a sequence-to-sequence model for speech synthesis, which can generate speech with natural-sounding intonation and expressiveness.

"WaveNet: A Generative Model for Raw Audio" by Aaron van den Oord et al. (2016): This paper proposed a deep generative model for speech synthesis, which can generate high-quality speech with a high level of naturalness and expressiveness.

"Deep Voice: Real-time Neural Text-to-Speech" by Sercan Ö. Arık et al. (2017): This paper introduced a deep neural network architecture for text-to-speech, which can generate high-quality speech in real-time.

"Neural Speech Synthesis with Transformer Network" by Wei Ping et al. (2018): This paper proposed a transformer-based architecture for text-to-speech, which achieved state-of-the-art performance on several benchmark datasets.