Automated Notes Maker from Audio/Video Recordings

Submitted for partial fulfillment of the requirements

for the award of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE ENGINEERING -

ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

by

V.VISHNU VARDHAN GOWD	-	**20BQ1A4260**
R. MANIKANTA	-	**20BQ1A4248**
D.SAI SRI CHARAN	-	**20BQ1A4258**
K. VENKATESWARA RAO	-	**20BQ1A4231**

Under the guidance of

Mr. N. BALAYESU, ASST. PROFESSOR

VVIT
VASIREDDY VENKATADRI
INSTITUTE OF TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING – ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

(B. Tech Program is Accredited by NBA)

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY

Permanently Affiliated to JNTU Kakinada, Approved by AICTE

Accredited by NAAC with 'A' Grade, ISO 9001:2008 Certified

NAMBUR (V), PEDAKAKANI (M), GUNTUR – 522 508

Tel no: 0863-2118036, url: www.vvitguntur.com,

April 2024

DECLARATION

We, Mr. Vishnu Vardhan Gowd, Mr. R. Manikanta, Mr. Sai Sri Charan, Mr. K. Venkateswara Rao, hereby declare that the Project Report entitled **"Automated Notes Maker from Audio/Video Recordings"** done by us under the guidance of Mr. N. Balayesu, Asst. Professor, Computer Science Engineering-Artificial Intelligence & Machine Learning. at Vasireddy Venkatadri Institute of Technology is submitted for partial fulfillment of the requirements for the award of Bachelor of Technology in Computer Science Engineering-Artificial Intelligence & Machine Learning. The results embodied in this report have not been submitted to any other University for the award of any degree.

DATE :

PLACE :

SIGNATURE OF THE CANDIDATE (S)

ACKNOWLEDGEMENT

We take this opportunity to express my deepest gratitude and appreciation to all those people who made this project work easier with words of encouragement, motivation, discipline, and faith by offering different places to look to expand my ideas and helped me towards the successful completion of this project work.

First and foremost, we express my deep gratitude to **Mr. Vasireddy Vidya Sagar**, Chairman, Vasireddy Venkatadri Institute of Technology for providing necessary facilities throughout the B.Tech programme.

We express my sincere thanks to **Dr. Y. Mallikarjuna Reddy**, Principal, Vasireddy Venkatadri Institute of Technology for his constant support and cooperation throughout the B.Tech programme.

We express my sincere gratitude to **Dr. K. Suresh Babu**, Professor & HOD, Computer Science Engineering-Artificial Intelligence & Machine Learning, Vasireddy Venkatadri Institute of Technology for his constant encouragement, motivation and faith by offering different places to look to expand my ideas.

We would like to express my sincere gratefulness to our Guide **Mr. N. Balayesu,** Asst. Professor, Computer Science Engineering-Artificial Intelligence & Machine Learning for his insightful advice, motivating suggestions, invaluable guidance, help and support in successful completion of this project.

We would like to express our sincere heartfelt thanks to our Project Coordinator **Mr. N. Balayesu,** Asst. Professor, Computer Science Engineering-Artificial Intelligence & Machine Learning for his valuable advices, motivating suggestions, moral support, help and coordination among us in successful completion of this project.

We would like to take this opportunity to express my thanks to the **Teaching and Non-Teaching** Staff in the Department of Computer Science Engineering-Artificial Intelligence & Machine Learning, VVIT for their invaluable help and support.

Name (s) of Students

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY
Permanently Affiliated to JNTUK, Kakinada, Approved by AICTE
Accredited by NAAC with 'A' Grade, ISO 9001:20008 Certified
Nambur, Pedakakani (M), Guntur (Gt) -522508

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING - ARTIFICIAL INTELLIGENCE & MACHINE LEARNING

CERTIFICATE

This is to certify that this **Project Report** is the bonafide work of Mr. V. Vishnu Vardhan Gowd, Mr. R. Manikanta, Mr. D. Sai Sri Charan, Mr. K. Venkateswara Rao, bearing Reg. No. **20BQ1A460, 20BQ1A4248, 20BQ1A4258, 20BQ1A4231** respectively who had carried out the project entitled **"Automated Notes Maker from Audio/Video Recordings "** under our supervision.

Project Guide	**Head of the Department**
(Mr. N. Balayesu**,** Asst. Professor)	(Dr. K. Suresh Babu, Professor)

Submitted for Viva voce Examination held on _____

Internal Examiner	**External Examiner**

TABLE OF CONTENTS

LIST OF FIGURES

NOMENCLATURE

FAISS	Facebook AI Similarity Search
ASR	Whisper v3 speech recognition model.
AI	Artificial Intelligence
ML	Machine Learning
ANM	Automated Notes Maker
HTTP	Hypertext Transfer Protocol

ABSTRACT

Our application, the Automated Notes Maker from Audio/Video Recordings, revolutionizes the way students engage with online classes. By seamlessly transcribing voice or video-based sessions into text-based PDF or Word documents, students can easily access written content for efficient review and enhanced understanding. In the realm of online education, the demand for efficient methods of note-taking from audio and video recordings has surged. Traditional note-taking processes are often time-consuming, prompting the need for automated solutions to streamline this task. In response, this research introduces the Automated Notes Maker (ANM), a system designed to seamlessly convert audio and video recordings into structured and concise notes. ANM integrates multi-modal capabilities, customizable settings, and real-time processing to provide users with immediate access to summarized notes and a question-answering system. By automating note-taking, ANM aims to alleviate the manual effort required by students, offering them comprehensive and organized study materials. The evaluation of ANM focuses on transcription accuracy, summary quality, and user satisfaction, showcasing its effectiveness in enhancing the online learning experience. Through this research, we present a comprehensive solution to address the growing need for efficient note-taking in the digital education landscape.

Keywords: Audio Recording, Machine Learning, Natural Language Processing, Text Summarization, Topic Segmentation, Translation, Question Answering System, open source, Cloud Storage.

CHAPTER 1
INTRODUCTION

1.1 Background and Need for the Project

The rise of online education has transformed the learning landscape. The increase of AI model such as photo-sketch face synthesis, deep learning-based models has increased efficiency we need to perform tasks in less time. While offering flexibility and convenience, it presents challenges for students to effectively capture and retain information from virtual lectures. Traditional note-taking methods can be time-consuming and inefficient, hindering students' ability to grasp key concepts and focus on active learning.

Problem:

This project addresses the critical need for improved methods in online learning to facilitate efficient knowledge acquisition. Traditional note-taking methods often involve:

- **Inefficient Time Management:** Manually taking notes during lectures can be a time-consuming process, diverting focus away from active listening and comprehension.
- **Information Overload:** Virtual lectures can be fast-paced, leading to students missing important details or struggling to capture everything being said.
- **Incompleteness and Inaccuracy:** Traditional note-taking methods may result in incomplete or inaccurate information due to the pressure to keep pace with the lecture.

Need for the Project:

This project proposes the development of an Automated Notes Maker (ANM) tool to address the limitations of traditional note-taking methods. ANM aims to:

- **Enhance Efficiency:** By automatically converting audio and video lectures into structured notes, ANM frees up student time for focused listening and deeper engagement.

- **Improve Accessibility:** Students with learning disabilities or those who struggle with traditional note-taking methods can benefit from the clear and concise structure provided by ANM.
- **Increase Comprehension:** Structured notes allow students to revisit key points and concepts more easily, fostering a deeper understanding of the material.

1.2 Project Objectives of Automated Notes Maker

The primary objective of this project is to develop and evaluate the effectiveness of the Automated Notes Maker (ANM) system in enhancing the online learning experience. This will be achieved by focusing on the following specific objectives:

1. Automate Note-Taking:

- Design and implement ANM to automatically convert audio and video lectures into structured, concise notes.
- Eliminate the need for manual note-taking, allowing students to focus on active listening and comprehension.

2. Improve Efficiency and Accessibility:

- Reduce the time and effort required for students to capture key points from lectures.
- Provide clear and organized notes for improved accessibility to information, benefiting students with diverse learning styles or disabilities.

3. Enhance Comprehension and Knowledge Retention:

- Enable students to revisit key concepts and ideas more easily through structured notes.
- Foster deeper understanding of the material and improve knowledge retention.

4. Evaluate System Performance:

- Assess the accuracy of ANM's transcription capabilities.
- Evaluate the quality and conciseness of the generated summaries.
- Conduct user satisfaction surveys to gauge the effectiveness of ANM in improving the online learning experience.

1.3 Project Challenges

While existing tools offer speech-to-text conversion and some summarization features, developing and implementing ANM effectively presents several key challenges:

- **Accuracy and Nuance:** Speech recognition technology, while constantly improving, can struggle with accents, background noise, and technical jargon specific to certain educational fields. ANM must achieve a high degree of accuracy to capture the nuances of lectures and avoid misinterpretations.
- **Context and Summarization:** Extracting key points and generating concise summaries requires an understanding of the context and overall structure of the lecture. ANM needs to go beyond simple transcription to identify important information and create summaries that retain the meaning and flow of the material.
- **Real-time Processing:** For ANM to be truly beneficial in online learning environments, it needs to process lectures in real-time or near real-time. This presents a significant computational challenge, especially for longer lectures or those with complex audio quality.
- **Multi-modal Integration:** While existing tools might focus on audio or video, ANM aims to integrate multi-modal capabilities. This includes processing visual elements like slides or speaker gestures to enhance understanding and potentially identify key points. Integrating and synchronizing these different data streams poses a technical challenge.
- **Customization and User Control:** Students have diverse learning styles and preferences. ANM should offer customizable options such as adjusting the level of detail in summaries or highlighting specific topics based on user input.

1.4 Advantages of an AI Note Maker with its AI Features

An AI note maker offers a multitude of advantages over traditional note-taking methods, thanks to its built-in AI functionalities. Here's a breakdown of the key benefits and the specific AI features that enable them:

1. Enhanced Efficiency and Time-Saving:

- **Automatic Transcription (Speech Recognition):** AI eliminates the need for manual note-taking, freeing up students' time and mental energy for active listening and engagement.
- **Real-time Processing:** Certain AI models can process lectures in real-time or near real-time, allowing students to access notes immediately for review or clarification.

2. Improved Accuracy and Completeness:

- **Speech Recognition with Accuracy:** AI-powered speech recognition strives for high accuracy, minimizing errors and capturing even complex lectures with accents or technical jargon.
- **Speaker Identification and Separation:** Advanced AI can distinguish between multiple speakers in a lecture, ensuring notes accurately reflect who said what.

3. Deeper Comprehension and Knowledge Retention:

- **Summarization and Key Point Extraction:** AI can analyse lectures and identify key points, generating concise summaries that highlight essential information. This allows students to revisit core concepts easily and solidify their understanding.
- **Speaker Diarization (Topic Segmentation):** Some AI can segment lectures by topic, allowing students to focus on specific areas or quickly review sections for exams.

4. Increased Accessibility and Personalization:

- **Speech-to-Text for Accessibility:** Students with learning disabilities or those who struggle with traditional note-taking methods can benefit from AI-generated notes.
- **Customization Options:** AI note makers can offer features like adjustable summary detail or keyword highlighting based on user preferences, catering to diverse learning styles.

5. Potential for Additional Features:

- **Question-Answering System:** Advanced AI may integrate a question-answering system, allowing students to search within their notes for specific information.
- **Multi-modal Integration:** AI can potentially analyse visual elements like slides or speaker gestures alongside audio, providing a more comprehensive understanding of the lecture content.

1.5 Innovation in Note Generation Technology

ANM stands at the forefront of innovation in note-taking technology by leveraging cutting-edge AI to create a user-centric platform that empowers students in online learning environments.

Key Features and Technologies:

- **Advanced Noise Reduction:** ANM employs Fast Fourier Transform (FFT) to extract audio signals from lectures and remove background noise, both stationary and non-stationary. This ensures crystal-clear audio for accurate speech recognition.
- **State-of-the-Art Speech Recognition:** ANM utilizes Whisper Large V3, a leading Automatic Speech Recognition (ASR) model, to transcribe audio lectures into text with exceptional accuracy. Timestamps are embedded within the transcribed text, enabling seamless subtitle generation for enhanced accessibility.
- **Multilingual Accessibility:** Recognizing the diverse needs of learners, ANM integrates the deep translator library. This allows users to translate notes into any of the 32 supported languages, fostering inclusivity and global learning opportunities.
- **Intelligent Search and Retrieval:** ANM leverages vector embeddings and cosine similarity calculations. This powerful combination enables users to search for specific information within their notes with pinpoint accuracy. By understanding the semantic relationships between concepts, ANM facilitates efficient knowledge retrieval.
- **Text-to-Speech (TTS):** Catering to auditory learners and individuals with visual impairments, ANM integrates the TTS/XTTS_v2 model. This allows users to convert their notes into realistic speech, facilitating playback and revision through audio channels.
- **Concise Summarization:** ANM employs Google's advanced Generative AI model, Gemini-pro, to analyze and summarize lengthy lectures. This feature provides users

with clear and concise summaries, allowing them to grasp key concepts and review material effectively.

- **Scalable Cloud Storage:** ANM utilizes MongoDB, a robust NoSQL database, for cloud storage of user notes and data. Fast API ensures efficient data handling and communication with the application.
- **Visual Note Integration:** Looking towards the future, ANM is exploring the potential of Gemini Vision Pro, a cutting-edge AI model for image interaction. This integration has the potential to allow users to interact with images associated with lectures, further enriching the note-taking experience.

Benefits and Impact:

By combining these innovative technologies, ANM offers a user-centric note-taking platform that delivers numerous benefits:

- **Enhanced Efficiency:** Automatic note generation frees up valuable time for students, allowing them to focus on active listening and comprehension.
- **Improved Accuracy and Completeness:** Advanced AI models minimize errors and capture nuances in lectures, leading to more complete and accurate notes.
- **Deeper Understanding:** Summarization features and efficient knowledge retrieval facilitate a deeper understanding of the material and improved knowledge retention.
- **Accessibility and Inclusivity:** Multilingual translation and text-to-speech features cater to diverse learning needs and abilities.

1.6 The Intertwined Roles of Community and Technology in the ANM Project

The ANM project thrives on the combined strengths of both community and technology. Each plays a crucial role in the development and success of this innovative note-taking platform.

Community:

- **Needs Identification:** The ANM project originates from the needs and challenges faced by the online learning community. Students often struggle with time-consuming

traditional note-taking methods, hindering comprehension and knowledge retention. This community need fueled the development of ANM as a solution.

- **User Feedback and Testing:** The online learning community will be vital in providing feedback and testing ANM during its development stages. This iterative process allows the project to identify and address user pain points, ensuring ANM is truly user-centric and caters to the specific needs of learners.
- **Accessibility Advocacy:** The online learning community encompasses individuals with diverse learning styles and abilities. Collaboration with accessibility advocates will be crucial in ensuring ANM's features, such as multilingual translation and text-to-speech, are inclusive and address the needs of all learners.
- **Content Creation and Sharing:** As ANM evolves, the community can play a role in content creation. Students might share best practices and tips for using ANM effectively, fostering a collaborative learning environment.

Technology:

- **AI Models and Innovation:** The core functionality of ANM relies on advanced AI models for speech recognition, language translation, summarization, and potentially image interaction. Cutting-edge technologies like Whisper Large V3, Gemini-pro, and vector embeddings power these features and are constantly evolving to improve accuracy and efficiency.
- **Cloud Infrastructure and Scalability:** Cloud storage solutions like MongoDB, alongside Fast API for data communication, provide a scalable platform for ANM. This ensures the platform can accommodate a growing user base within the online learning community.
- **Application Development and User Interface:** The user interface of ANM, the way users interact with the platform, is a critical aspect of technology. User-friendly design and intuitive navigation will be crucial for user adoption and maximizing the impact of the project within the online learning community.

Synergy: Empowering the Online Learning Community

The true strength of the ANM project lies in the synergy between community and technology. By listening to the needs of the online learning community and leveraging cutting-edge AI

models, ANM can revolutionize the note-taking experience. Successful collaboration will ensure a user-centric platform that fosters efficient learning, deeper understanding, and inclusivity for all students in the digital education landscape.

1.7 Privacy and Ethical Considerations in ANM's Use of Input Data

ANM, as an AI-powered note-taking system, handles user data that necessitates careful consideration of privacy and ethical implications. Here's a breakdown of key areas to focus on:

Data Collection and Storage:

- **Transparency and User Consent:** ANM should clearly inform users about what data is collected (audio recordings, transcripts, etc.) and how it will be used. Obtaining explicit user consent before processing any data is paramount.
- **Data Anonymization and Security:** Consider anonymizing user data whenever possible to minimize privacy risks. Robust security measures must be implemented to safeguard sensitive information from unauthorized access or breaches.
- **Data Retention and Deletion:** Establish clear policies on how long user data is retained. Users should have the option to request their data deletion upon completion of note-taking or after a specific timeframe.

AI Model Biases and Fairness:

- **Training Data Bias:** AI models used in ANM, such as speech recognition or summarization, are trained on large datasets. It's crucial to ensure these datasets are diverse and representative to avoid perpetuating biases based on factors like gender, accent, or ethnicity.
- **Fairness in Summarization and Information Extraction:** ANM should strive to generate summaries and extract information from lectures fairly and accurately, regardless of the speaker's background or the topic discussed. Techniques like fairness metrics and human oversight can help mitigate bias.

Transparency and Explainability of AI Decisions:

- **Understanding How ANM Works:** Users should have a basic understanding of how ANM functions. This could involve providing clear explanations about the AI models

used and the decision-making processes behind features like summarization or information retrieval.

- **User Control and Oversight:** Whenever possible, grant users control over how their data is processed within ANM. This might involve options to adjust summarization settings or choose the language model used for translation.

1.8 Overview of the Proposed System's Architecture and Components:

ANM System Architecture and Components

ANM leverages a combination of AI models, cloud storage, and user interface components to deliver a user-centric note-taking platform. Here's a breakdown of the key architectural elements:

1. Data Preprocessing Module:

- This module handles the initial processing of user input, which is likely audio recordings of lectures.
- **Components:**
    - **Audio Processing:** This component utilizes Fast Fourier Transform (FFT) to extract clean audio signals from the recording. It removes both stationary noise (e.g., background hum) and non-stationary noise (e.g., coughs, speaker movement).

2. Speech Recognition Module:

- This module converts the cleaned audio signal into text.
- **Components:**
    - **Whisper Large V3:** This state-of-the-art Automatic Speech Recognition (ASR) model performs the core speech-to-text conversion.
    - **Timestamping:** The transcribed text is embedded with timestamps to facilitate subtitle generation and future navigation within the notes.

3. Natural Language Processing (NLP) Module:

- This module handles various tasks related to understanding and manipulating the textual content of the notes.

- **Components:**
  - **deeptranslator Library:** This library enables language translation of the transcribed text, offering accessibility in 32 supported languages.
  - **Vector Embeddings and Cosine Similarity:** This technique allows ANM to understand the relationships between words and concepts within the notes. It facilitates features like:
  - **Search and Retrieval:** Users can search for specific topics within their notes using keywords. Cosine similarity helps identify sections relevant to the search query.

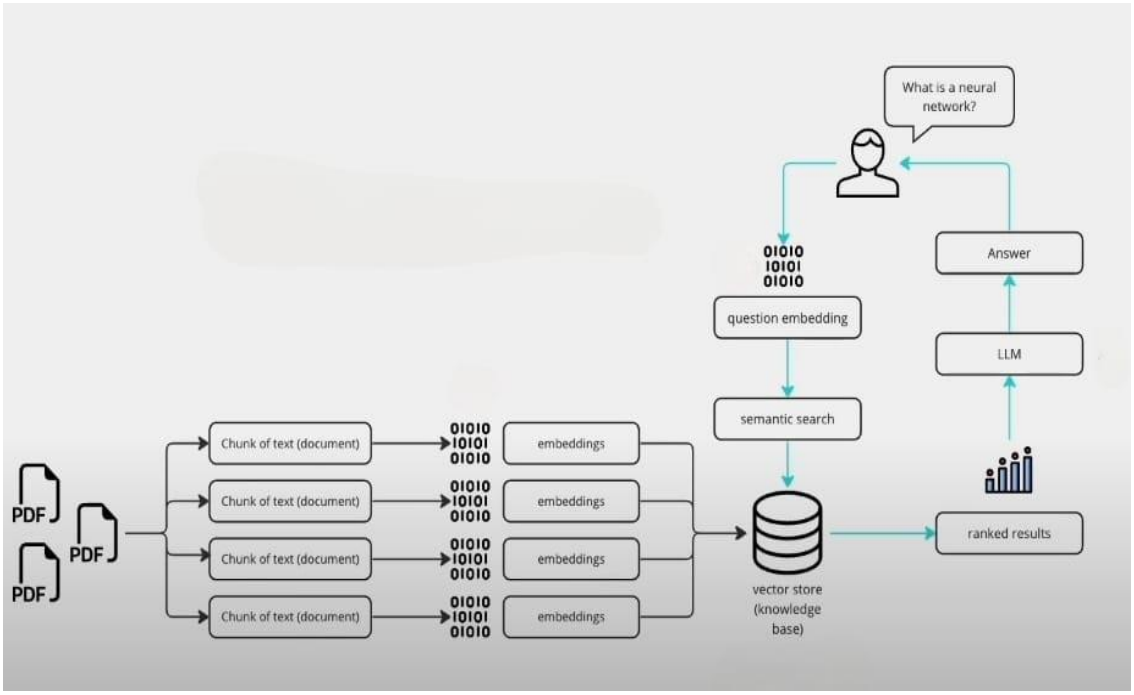


Fig. 1.8 Langchain - Vector Embedding conversion of text chunks

  - **Summarization:** By analyzing the relationships between concepts, the system can identify key points and generate concise summaries using the...

4. Text-to-Speech (TTS) Module:

- This module caters to auditory learners and individuals with visual impairments.
- **Components:**

- **TTS/XTTS_v2 Model:** This model converts the transcribed text into realistic speech, allowing users to playback their notes and revise through audio channels.

5. Summarization Module:

- This module generates concise summaries of the lecture content.
- **Components:**
  - **Gemini-pro:** This Google Generative AI model analyzes the lengthy transcribed text and identifies key points. It then generates a summary that captures the essence of the lecture.

6. Cloud Storage and Data Management:

- ANM utilizes cloud storage to store user data securely and facilitate scalability.
- **Components:**
  - **MongoDB:** This NoSQL database provides a flexible and scalable platform for storing user notes, transcripts, and other relevant data.
  - **Fast API:** This framework acts as an intermediary between the application and the cloud storage, ensuring efficient data communication.

7. User Interface (UI):

- The UI provides users with a user-friendly interface to interact with ANM's functionalities.
- **Components:**
  - The UI will likely allow users to upload audio recordings, manage their notes, access summaries, utilize search functions, and potentially interact with features like text-to-speech playback (depending on the level of development).

8. Future Integration (Potential):

- ANM is exploring the potential to integrate the Gemini Vision Pro model.
- This model could allow users to interact with images associated with the lectures, potentially enabling features like:
  - Identifying key visual elements from slides or presentations.
  - Linking image content to specific sections within the notes.

1.9 Importance of a Unified Database for AI notes maker:

In the context of ANM, a unified database plays a critical role in ensuring the smooth operation and user experience of the platform. Here's why a unified database is important for ANM:

1. Centralized Data Management:

- ANM utilizes various AI models for tasks like speech recognition, summarization, and potentially image interaction (future integration).
- A unified database ensures all processed data – transcribed text, summaries, extracted information from images (future) – is stored in a central location.
- This simplifies data management and retrieval for various functionalities within ANM.

2. Improved Search and Retrieval:

- ANM leverages vector embeddings and cosine similarity for search functionality within notes.
- Storing all processed text (transcripts, summaries) in a unified database allows for efficient search across all data points.
- Users can locate specific information within their notes more easily, regardless of whether it originated from the raw transcript, summary, or linked image data (future).

3. User Experience and Personalization:

- ANM can potentially personalize the note-taking experience based on user preferences or learning styles.
- A unified database allows ANM to store user preferences alongside their notes.
- For example, the system might prioritize certain concepts or topics based on user input, drawing from the combined data of transcripts, summaries, and user interactions.

4. Scalability and Maintainability:

- As the user base of ANM grows, a unified database ensures scalability.
- Cloud storage solutions like MongoDB, which ANM utilizes, can efficiently handle increasing data volumes.
- Maintaining a single database is simpler compared to managing multiple data sources for different functionalities.

5. Security and Data Integrity:

- A unified database with robust security measures can safeguard user data, including transcripts, notes, and potentially identifiable information.
- Implementing centralized access controls and encryption helps ensure data integrity and minimizes the risk of breaches or inconsistencies.

1.10 Potential Impact of AI-based ANM on Online and Offline Learning Platforms

ANM, an AI-powered note-taking system, has the potential to significantly impact both online and offline learning platforms. Here's a closer look at the potential influence:

Impact on Online Learning Platforms:

- **Enhanced Learning Efficiency:** ANM can automate time-consuming note-taking, allowing students to focus on active listening, participation, and deeper comprehension during online lectures.
- **Improved Accessibility:** Features like speech-to-text conversion, multilingual translation, and text-to-speech playback can cater to diverse learning styles and abilities, fostering inclusivity in online learning environments.
- **Personalized Learning Experience:** ANM's ability to summarize lectures and potentially personalize based on user preferences can empower students to tailor their learning experience.
- **Integration with Learning Management Systems (LMS):** Future development could see ANM integrate with LMS platforms, allowing for seamless note management and potentially enriching existing learning materials.

Challenges for Online Learning Platforms:

- **Integration and Interoperability:** Integrating ANM with existing online learning platforms might require technical considerations to ensure smooth operation.
- **Data Privacy and Security:** Online learning platforms will need to address data privacy concerns related to student notes and audio recordings processed by ANM.

Impact on Offline Learning Platforms (with Potential Integration):

- **Supplementing Classroom Learning:** ANM recordings and AI-generated summaries can be used to revisit key concepts or review missed material outside of class hours, potentially improving knowledge retention.
- **Facilitating Flipped Classrooms:** ANM recordings can be used for pre-recorded lectures in flipped classroom models, allowing in-class time to focus on interactive activities and discussions.
- **Personalized Study Resources:** ANM's summaries and user-controlled transcripts can be valuable study resources for students with different learning styles or those who need additional support.

Challenges for Offline Learning Platforms:

- **Technology Access and Equity:** Not all students may have access to the devices or internet connectivity required to utilize ANM effectively.
- **Focus on Traditional Methods:** Integrating technology like ANM into existing lesson plans might require a shift in teaching approaches for some educators.

CHAPTER 2

REVIEW OF LITERATURE

Manoj Kumar A , Janani P , Siva Subramanian G , Kumaragurubaran K and Sundari P: The paper explores the effectiveness and utilization of automated note-taking software for audio recordings. In a world where managing vast amounts of data is crucial, particularly in the forms of audio recordings, video recordings, and text documents, manual summarization and transcription can be tedious and time-consuming.Automated note-taking software uses Machine Learning (ML) and Natural Language Processing (NLP) techniques to transcribe spoken words from audio recordings into text. This transcribed text is then used to generate summaries or notes. The paper highlights the importance of this technology, particularly in transforming audio recordings of online classes into text-based PDF or Word documents. Overall, the paper emphasizes the valuable role of automated note-taking software in efficiently managing and transforming audio data into usable text form, with potential applications across academia, business, law, and healthcare sectors.

Chaudhari Mahima , Mali Divya , Chaudhari Nehal , Kolhe Trupti and Ashish T. Bhole: Automated note-taking software has become an integral part of modern data management, particularly in handling vast amounts of audio recordings. This paper presents an investigation into the development and utilization of an Automated Note Maker (ANM) system, leveraging Machine Learning (ML) and Natural Language Processing (NLP) technologies. The ANM system aims to transcribe spoken words from audio recordings into textual format, providing users with efficient note-taking capabilities.The research focuses on the implementation of the ANM system, emphasizing its potential applications across various domains such as academia, business, law, and healthcare. Through the utilization of ML models, the system aims to accurately transform audio recordings into text-based PDF or Word documents. providing users with a streamlined and efficient method for converting audio recordings into actionable text-based documents. Further research and development in this field hold promise for continued improvements in note-taking technologies.

Ms. Purva Chavrekar, Ms. Shruti Deshmukh, Ms. Pranjal Khade, Ms.Vaibhavi Patil and Prof. Rupali: Sathe This delves into the development and utilization of an Automated Notes Maker (ANM) system that leverages Speech Recognition technology. The primary aim of the system is to assist individuals, including those with disabilities such as blindness or limited hand function, in converting audio recordings into text-based notes. The system is designed to cater to a wide range of users, including instructors, scholars, and those who are unable to attend lectures due to various reasons. 14The paper emphasizes the significance of Speech Recognition technology in aiding individuals who rely on voice input to operate computers, especially those unable to use their hands. By providing real-time voice data processing, the ANM system can convert audio inputs into textual format, allowing users to create and store

notes efficiently. the paper highlights the potential of the Automated Notes Maker system in revolutionizing note-taking processes for various user groups. By harnessing the power of Speech Recognition technology and advanced feature extraction methods, the system offers a practical solution for converting audio recordings into valuable textual notes, thereby enhancing accessibility and convenience for users across different settings and backgrounds.

Vinnarasu A., Deepa V. Jose: Improve the clarity of text extracted from speech by adding punctuation marks like periods and question marks where necessary. A custom code is implemented to insert periods after pauses of at least 2e+6 µs, ensuring proper sentence termination. Temporary storage is utilized to handle cases where sentences begin with conjunctions or wh-questions, ensuring accurate punctuation placement. Text summarization is performed based on sentence and word frequency using NLTK's tokenization techniques. The summarization algorithm ranks sentences by word frequency, allowing for the identification and summarization of important sentences. Finally, Python's nlargest function is employed to rank sentences by weight, facilitating the summarization process.

Prerana Das, Kakali Acharjee, Pranab Das and Vijay Prasad :This thing explores the development and implementation of a Voice Recognition System (VRS) designed to convert spoken words into text, allowing users to control computer functions and dictate text through voice commands. The system comprises two main components: the first component processes the acoustic signal captured by a microphone, while the second component interprets the processed signal and maps it to words using Hidden Markov Models (HMM). The primary objective of the project is to develop a robust speech recognition system using MFCC and VQ techniques. The extracted features are stored in .mat files, and models are created using Hidden Markov Models (HMM). The final output, interpreted text, is displayed through a MATLAB interface.In conclusion, the research paper presents a detailed exploration of the Voice Recognition System, emphasizing its use of advanced techniques such as MFCC for feature extraction and HMM for model creation. The system's application in home automation highlights its potential for practical and efficient voice-controlled interactions with computers and devices.

Ms. Anuja Jadhav, Prof. Arvind Patil:This paper focuses on converting speech into text through three main modules: Speech Acquisition, Speech Preprocessing, and HMM Training. In Speech Acquisition, real-time speech samples are obtained from a microphone and stored for preprocessing. Sphinx framework is used for speech preprocessing, including steps like frame blocking, windowing, and voice activity detection. HMM training involves creating a model for each digit in the vocabulary using the Baum-Welch algorithm. Vector quantization is used to reduce storage and computation, followed by HMM recognition using the Viterbi decoding algorithm. Performance parameters include recognition accuracy, speed, and testing across various speaker samples for robustness evaluation.

Hakan Erdogan a*, Ruhi Sarikaya b , Stanley F. Chen b , Yuqing Gao b , Michael Picheny b:This study introduces three novel language modeling techniques that utilize semantic analysis to enhance spoken dialog systems. The methods proposed are named concept sequence modeling, two-level semantic-lexical modeling, and joint semantic-lexical modeling. These models integrate varying degrees of semantic information with lexical data, using annotations provided by either a shallow semantic parser or a full hierarchical parser. The integration ranges from simple interpolation to tight integration using maximum entropy modeling. In conclusion, the research paper presents innovative language modeling techniques that leverage semantic analysis to improve speech recognition performance. The proposed models demonstrate significant improvements over traditional language models, particularly in noisy environments. The study highlights the potential of semantic-lexical integration and maximum entropy modeling in enhancing spoken dialog systems and opens avenues for future research in this domain.

Jingdong Chen, Yiteng (Arden) Huang, Qi Li, and Kuldip K. Paliwal:This investigates the use of Spectral Subband Centroids (SSCs) as an alternative to traditional cepstral coefficients, such as those derived from linear prediction analysis or a filter-bank, for robust speech recognition, particularly in noisy environments. The study highlights the sensitivity of cepstral coefficients to additive noise and proposes SSCs as a potentially more resilient alternative.this presents Spectral Subband Centroids (SSCs) as a promising alternative to traditional cepstral coefficients for robust speech recognition, particularly in noisy environments. The study demonstrates that properly selected SSCs can match the performance of MFCCs in clean speech conditions while offering improved resilience to noise. The proposed dynamic SSC features further enhance the recognition capabilities, making SSCs a valuable option for practical speech recognition systems.

F. Seide, G. Li, D. Yu: Context-Dependent Deep Neural Network Hidden Markov Models (CD-DNN-HMMs) for speech-to-text transcription. They advanced CD-DNN-HMMs by incorporating weight sparseness, reducing recognition error or model size. The CD-DNN-HMMs replaced Gaussian mixtures with MLPs for state emission likelihood computation. Training involved DBN pre-training followed by fine-tuning with backpropagation. Results showed a significant reduction in word-error rates compared to traditional GMM-HMMs. The effectiveness of CD-DNN-HMMs was demonstrated across various test sets, indicating their potential for improving speech recognition accuracy.

Shivangi Nagdewani, Ashika Jain: The paper explores various methods for speech-to-text (STT) and text-to-speech (TTS) conversion to develop an interactive email system. It assesses techniques like LPC, MFCC, DTW, HMM, Neural Networks, and hybrids, concluding that HMM offers the highest efficiency. The proposed model integrates HMM and Neural Network for STT and HMM for TTS, enhancing speech and text generation.

CHAPTER 3
METHODOLOGY

3.1 Overview of the Proposed Solution:

The implemented system presents a robust platform designed to address a wide range of user needs with precision and efficiency. At its core, the architecture prioritizes modularity, scalability, and optimal user experience, ensuring seamless navigation and interaction throughout the platform. Audio processing and noise reduction are foundational, with sophisticated techniques like the noiseReduce algorithm utilizing Fast Fourier Transform (FFT) to ensure pristine audio quality by eliminating both stationary and non-stationary noise components.

Following audio processing, the system seamlessly transitions into transcription and subtitle generation, leveraging advanced technologies such as the Whisper Large V3 ASR model. Timestamped transcription results facilitate effortless navigation for users, enhancing accessibility and usability. Language translation capabilities further augment the system's accessibility, facilitating precise translations across 32 supported languages through the Deep-translator library and accurate language detection using Langdetect.

Text querying and retrieval functionalities leverage cutting-edge techniques such as vector embeddings and cosine similarity computation to deliver relevant content to users. Large Language Models (LLMs) structure both retrieved content and user queries, providing comprehensive and well-organized answers. Text-to-speech conversion capabilities cater to auditory learners and visually impaired users, employing the TTS/XTTS_v2 model to generate realistic voices based on speaker voice profiles stored in a dedicated database.

The system extends its capabilities beyond textual content, with features like document summarization and topic segmentation enhancing content comprehension and navigation. Utilizing advanced AI models like Google Generative AI model Gemini-pro, the system generates concise summaries and prompts users with structured topic segments, facilitating efficient information consumption. Cloud storage integration ensures seamless data management, with notes and documents securely stored in a cloud-based MongoDB database. Additionally, image querying functionalities empower users to interact with visual data through questions and text

extraction, while live transcription capabilities powered by the React Speech Recognition library enable real-time transcription of spoken words during lectures or discussions, enriching the user experience and facilitating seamless communication.

3.2 Description of Technologies Used (Whisper V3, LLMs):

Audio Processing and Noise Reduction:

The system leverages sophisticated techniques, including the noiseReduce algorithm powered by Fast Fourier Transform (FFT), to extract audio signals from various sources such as video recordings or direct links. This ensures the pristine quality of audio signals by effectively eliminating both stationary and non-stationary noise components.

Transcription and Subtitle Generation:

Advanced technologies like the Whisper Large V3 ASR model are employed for transcription purposes, seamlessly converting clean audio signals into text. Timestamped transcription results facilitate convenient navigation for users and enhance accessibility.

Language Translation:

The Deep-translator library is utilized to provide language translation capabilities, offering precise translations across 32 supported languages. Accurate language detection using Langdetect ensures efficient communication across diverse linguistic landscapes.

Text Querying and Retrieval:

Cutting-edge techniques such as vector embeddings and cosine similarity computation are employed for text querying and retrieval functionalities. Large Language Models (LLMs) structure retrieved content and user queries, providing comprehensive and well-organized answers.

Text-to-Speech Conversion:

Text-to-speech conversion capabilities cater to auditory learners and visually impaired users, utilizing the TTS/XTTS_v2 model to generate realistic voices based on speaker voice profiles stored in a dedicated database.

Document Summarization and Topic Segmentation:

Advanced AI models like the Google Generative AI model Gemini-pro facilitate document summarization and topic segmentation, generating concise summaries and structured topic segments to enhance content comprehension and navigation.

Cloud Storage Integration:

Notes and documents are securely stored in a cloud-based MongoDB database, with FastAPI employed for efficient conversion of documents to BSON format. BSON format is chosen for its efficiency in storage and retrieval operations, ensuring seamless data management.

Image Querying:

Users can interact with visual data through questions and text extraction similar to OCR, utilizing the Gemini Vision Pro model. This feature aims to provide users with deeper insights into visual data through image-based interactions.

Live Transcription:

Real-time transcription of spoken words during lectures or discussions is facilitated by the React Speech Recognition library. This feature enriches the user experience and enables seamless communication in real-time scenarios.

3.3 System Architecture and Workflow:

System Architecture

The architecture comprises several key components, each tailored to specific functionalities within the overall process:

1. **User Interface (UI):** A user-friendly mobile and web application that allows users to upload images of found children. This interface is designed for ease of use, ensuring widespread public participation.
2. **Audio Processing and Noise Reduction:** The process begins with the extraction of audio signals from various sources, followed by the application of noise reduction techniques using the noiseReduce algorithm powered by Fast Fourier Transform (FFT). This ensures the elimination of both stationary and non-stationary noise components, resulting in clean audio signals.
3. **Transcription and Subtitle Generation:** Clean audio signals undergo transcription into text using advanced technologies like the Whisper Large V3 ASR model. The transcription results are timestamped to facilitate easy navigation and enhance accessibility for users.
4. **Language Translation:** The system employs the Deep-translator library for language translation, enabling precise translations across 32 supported languages. Accurate language

detection using Langdetect ensures seamless communication across diverse linguistic landscapes.

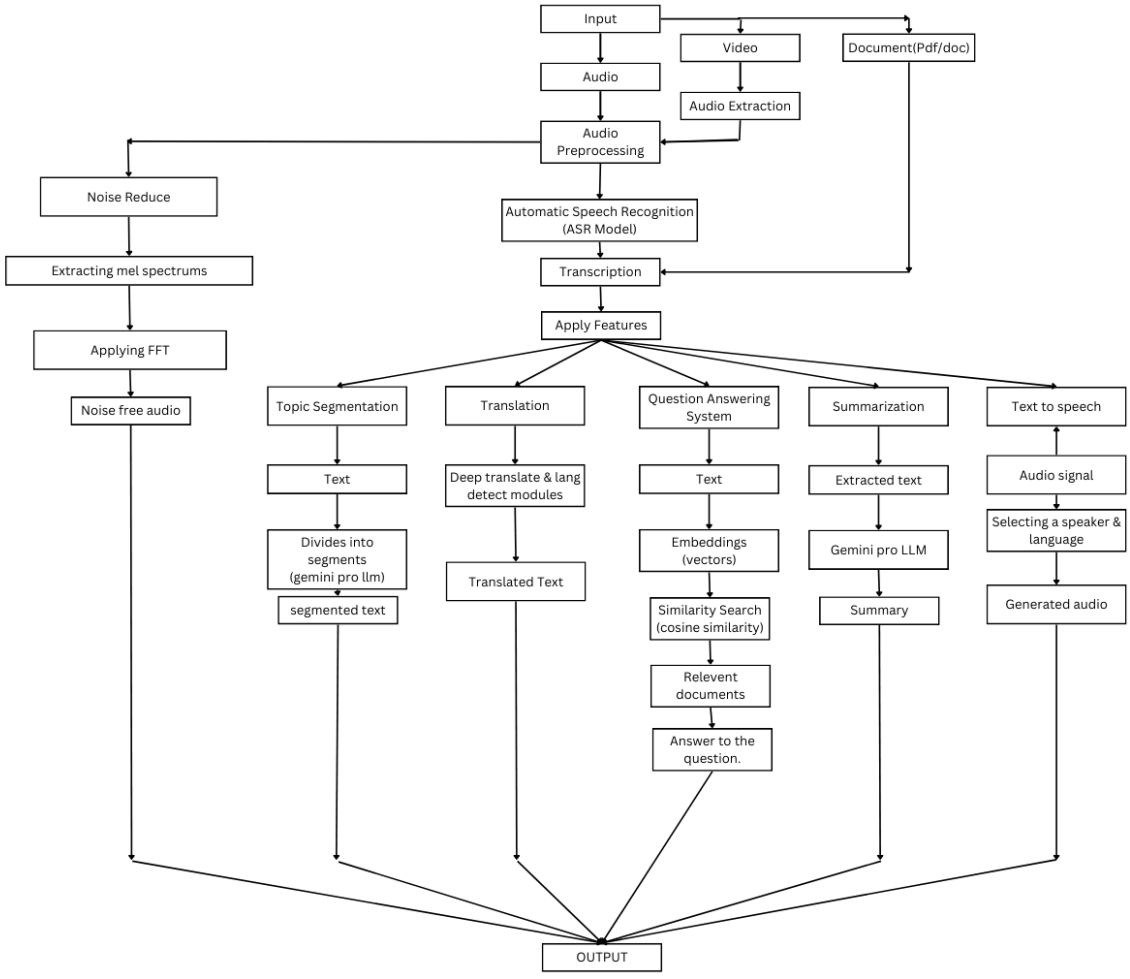


Fig. 3.3 System workflow architecture

5. **Text Querying and Retrieval:** Text querying and retrieval functionalities utilize cutting-edge techniques such as vector embeddings and cosine similarity computation. Large Language Models (LLMs) structure both retrieved content and user queries, providing comprehensive and well-organized answers.
6. **Text-to-Speech Conversion:** The platform offers text-to-speech conversion capabilities using the TTS/XTTS_v2 model. Realistic voices based on speaker voice profiles stored in a dedicated database enhance accessibility for auditory learners and visually impaired users.

7. **Document Summarization and Topic Segmentation:** Advanced AI models like the Google Generative AI model Gemini-pro facilitate document summarization and topic segmentation. This generates concise summaries and structured topic segments, enhancing content comprehension and navigation.
8. **Cloud Storage Integration:** Notes and documents are securely stored in a cloud-based MongoDB database, with FastAPI employed for efficient conversion of documents to BSON format. This ensures seamless data management and retrieval operations.
9. **Image Querying:** Users can interact with visual data through questions and text extraction using the Gemini Vision Pro model. This feature aims to provide deeper insights into visual data through image-based interactions.
10. **Live Transcription:** Real-time transcription of spoken words during lectures or discussions is facilitated by the React Speech Recognition library. This feature enriches the user experience and enables seamless communication in real-time scenarios.

3.4 Data Collection:

The data collection process for class recordings primarily involves sourcing content from MIT classes, which typically span over an hour in duration. These recordings serve as valuable resources for training and refining the system's capabilities.

In addition to MIT classes, the data collection process extends to leveraging the Mozilla Common Voice datasets for fine-tuning the Whisper model. These datasets provide a diverse range of speech data, which is instrumental in enhancing the accuracy and performance of the model.

By combining class recordings from MIT and Mozilla Common Voice datasets, the system ensures access to high-quality, diverse speech data essential for training and refining the transcription and speech processing functionalities. This meticulous data collection approach lays the foundation for delivering accurate and reliable results across various scenarios and user requirements.

CHAPTER 4

4.1 IMPLEMENTATION

4.1.1 Implementation of Audio Processing and Noise Reduction:

The first step involves implementing audio processing and noise reduction techniques. Utilizing the noiseReduce algorithm powered by Fast Fourier Transform (FFT), the system extracts audio signals from different sources such as video recordings or direct links. Sophisticated noise reduction techniques are then applied to clean the audio signals, ensuring high-quality output.

4.1.2 Transcription and Subtitle Generation:

Following audio processing, the system implements transcription and subtitle generation functionalities. Advanced technologies like the Whisper Large V3 ASR model are integrated to convert clean audio signals into text. Timestamped transcription results are generated to facilitate easy navigation and enhance accessibility for users.

4.1.2 Language Translation Integration:

Language translation capabilities are implemented next, leveraging the Deep-translator library. This integration enables precise translations across 32 supported languages. Accurate language detection using Langdetect ensures seamless communication across diverse linguistic landscapes.

4.1.3 Text Querying and Retrieval:

The system implements text querying and retrieval functionalities using cutting-edge techniques such as vector embeddings and cosine similarity computation. Large Language Models (LLMs) structure both retrieved content and user queries, providing comprehensive and well-organized answers.

4.1.4 Text-to-Speech Conversion:

Text-to-speech conversion capabilities are integrated into the system using the TTS/XTTS_v2 model. Realistic voices based on speaker voice profiles stored in a dedicated database enhance accessibility for auditory learners and visually impaired users.

4.1.5 Document Summarization and Topic Segmentation:

The Google Generative AI model Gemini-pro are implemented to facilitate document summarization and topic segmentation. This generates concise summaries and structured topics of the whole class based on the transcribed text class segments, enhancing content comprehension and navigation.

4.1.6 Cloud Storage Integration:

The system integrates cloud storage capabilities for efficient data management. Notes and documents are securely stored in a cloud-based MongoDB database, with FastAPI employed for streamlined document conversion to BSON format, optimizing storage and retrieval operations.

4.1.7 Image Querying Implementation:

The Gemini Vision Pro model is utilized for this task, providing deeper insights into visual data through image-based interactions. Initially the image is upload by user at frontend and it is transferred to FastAPI server via api where the bytes data of image is converted into numpy array. This array along with the user query is sent to Gemini Vision and then the responses is sent back to user.

4.1.8 Live Transcription Setup:

the React Speech Recognition library enables real-time transcription of spoken words during lectures or discussions. useSpeechRecognition is a React hook that gives a component access to a transcript of speech picked up from the user's microphone. SpeedRecognition manages the global state of the **Web Speech API**, exposing functions to turn the microphone on and off.

4.2 Development Tools and Environment:

Fig. 4.2 Development Tools flowchart

4.2.1 Visual Studio Code & Google Colab

Introduction to Visual Studio & Google Colab as the chosen development platform.

4.2.2 Programming Languages

- **Python**: Use of python in backend development.
- **FastAPI**: Use of FastAPI for robust Application Programming Interface development.

4.2.3 Front-End Development

- **ReactJS**: Utilization of ReactJS for functional and responsive UI.

- **Tailwind CSS**: Application of Tailwind CSS for aesthetic & response UI design.

4.2.4 Database Management

- **MongoDB Database**: Integration for real-time data handling and efficient data storage and synchronization.

4.2.5 Version Control

- **Git**: Implementation of Git for code versioning and team collaboration.

4.2.6 Integrated Development Environment (IDE)

Detailed on Visual Studio & Google Colab role as the IDE for the project.

4.2.7 Graphics and Image Editing

- **Adobe Photoshop**: Role in creating visual assets.
- **Adobe Illustrator**: Contribution to graphic design elements.

4.2.8 API Development Tools

- **Postman**: Usage for backend API development and testing.

4.2.9 User Interface Design Tools

- **Figma**: Adoption for UI/UX design and prototyping.
- **Sketch**: Application for interface design.

4.2.10 Deployment Framework

- **Vercel**: Utilization for Vercel application deployment.

4.3 User Interface and Experience

The Automated Notes maker application is designed with a focus on providing an intuitive and seamless user experience, enabling users to navigate the app efficiently and perform tasks without unnecessary complexity.

4.3.1 Intuitive Design

- **User-Centric Design Principles:** The app's design follows user-centric principles, ensuring that the interface is intuitive and accessible to users of all technical backgrounds.
- **Consistent Layout and Navigation:** A consistent layout and navigation scheme across the app facilitate ease of use, allowing users to intuitively understand how to access various features.

4.3.2 Accessibility Features

- **Accessibility Features:** The application prioritizes accessibility, with features such as adjustable font sizes, high contrast modes, and screen reader compatibility to cater to users with diverse needs and preferences.
- **Inclusive Design Practices:** Inclusive design practices are employed to ensure that the application is accessible and usable by users from different demographic backgrounds, including those with disabilities or limited access to technology.

4.4 Monitoring and Updates:

The Automated Notes maker project places significant emphasis on maintaining high performance, reliability, and user satisfaction through diligent monitoring and regular updates.

4.4.1 Performance Monitoring and Analytics

- **Real-time System Monitoring**: Utilizes tools to continuously monitor the app's performance, identifying any bottlenecks or issues that could affect user experience.
- **User Engagement Analytics**: Tracks key metrics such as active users, session lengths, and feature usage to understand how users interact with the app and identify areas for improvement.
- **Error Reporting and Resolution**: Implements an automated error reporting system to capture and analyse exceptions or bugs encountered by users, facilitating quick resolution and minimizing impact.

4.4.2 User Feedback and Community Engagement

- **Feedback Collection Mechanisms**: Incorporates in-app feedback forms and community forums, allowing users to report issues, suggest features, or provide general feedback.

- **Community Engagement**: Engages with users through social media, newsletters, and other channels to keep them informed about the project, gather community input, and foster a sense of ownership and involvement.

4.4.3 Continuous Improvement and Feature Updates

- **Agile Development Cycle**: Adopts an agile approach to development, enabling the team to rapidly iterate on the app based on user feedback, analytics, and changing requirements.
- **Feature Rollouts and Updates**: Regularly releases updates that include new features, improvements to existing functionalities, and optimizations to enhance the app's performance and user experience.

4.5 UML DIAGRAMS

4.5.1. USECASE DIAGRAM

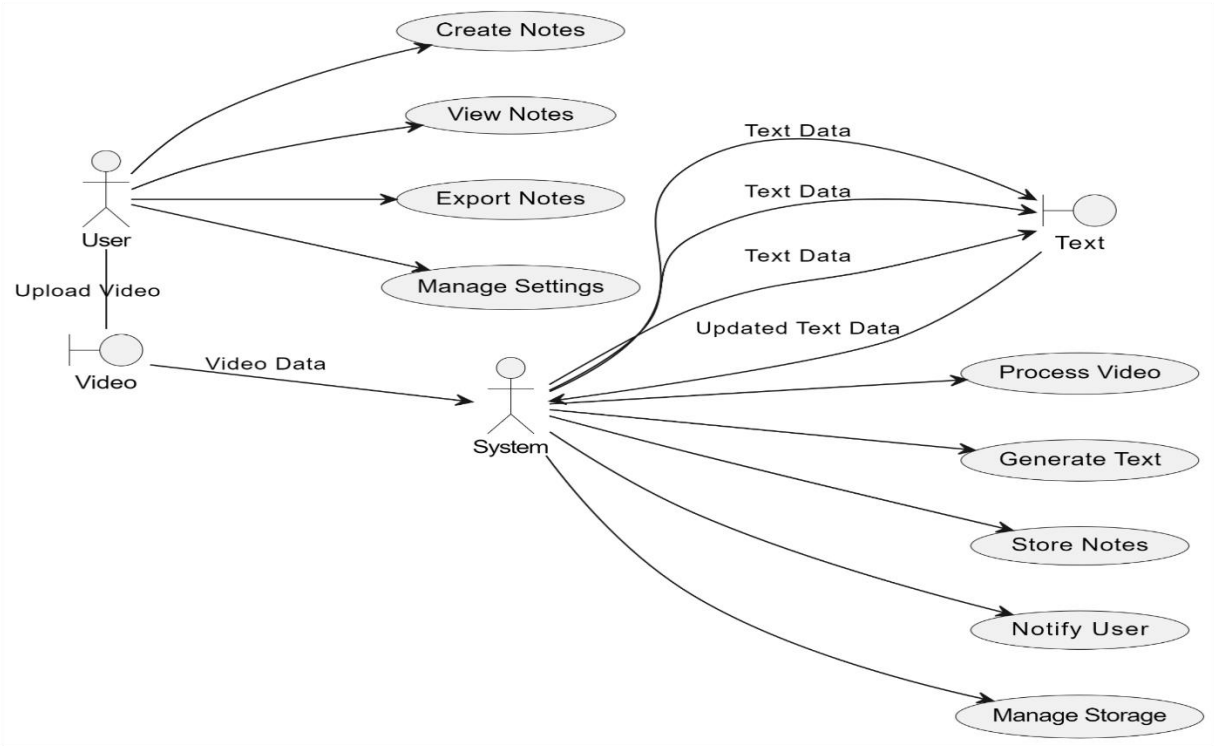


Fig. 4.5.1 Use case Diagram

- **Notes Maker From recordings:** This use case involves both audio and video to generate text notes

- This use case allows users to search for notes or recordings based on keywords/tags.
- Sub-cases:
    - Search by Keyword/Tag: Users can find relevant recordings/notes by searching for specific keywords or tags.

4.5.2. CLASS DIAGRAM

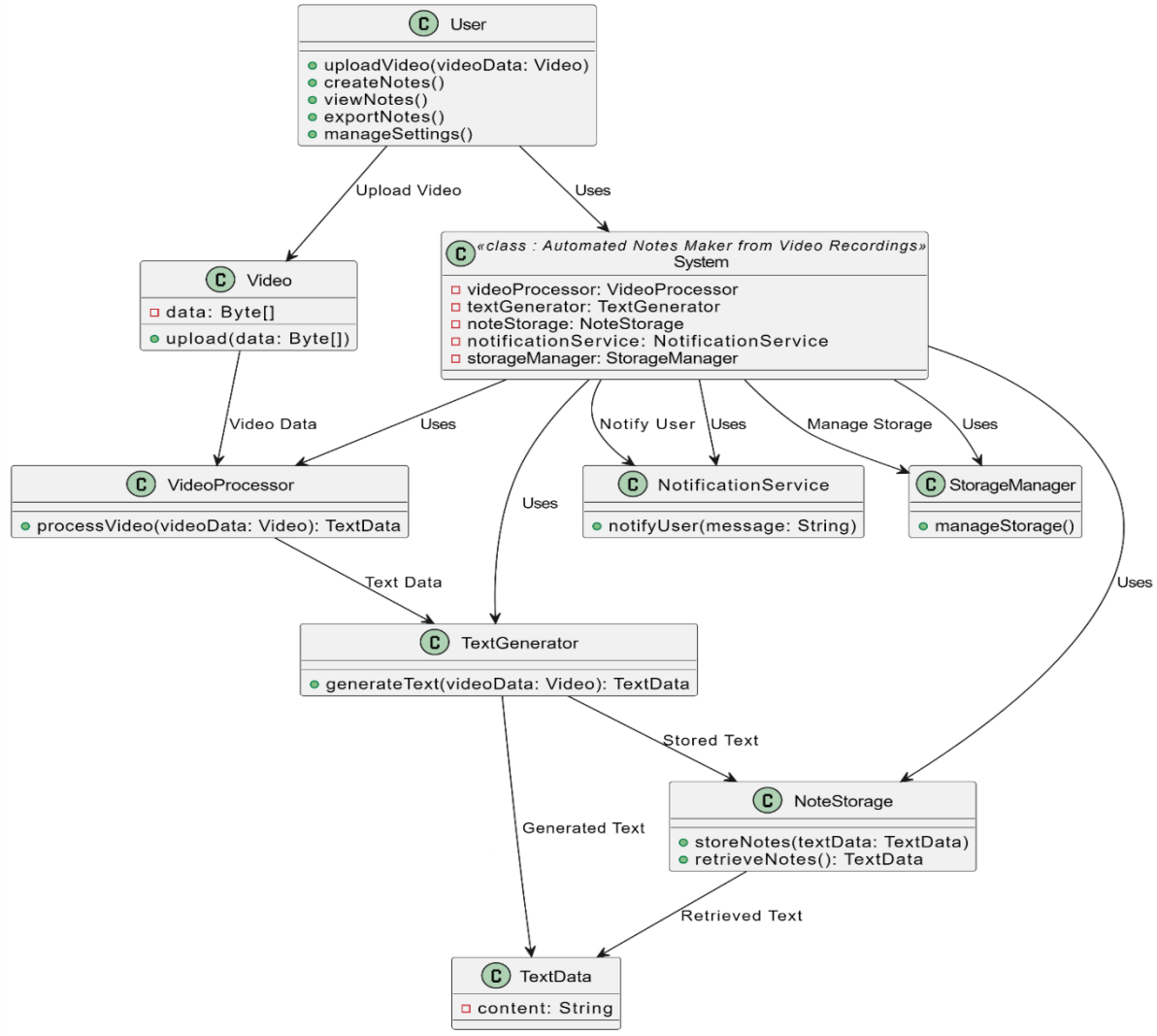


Fig. 4.5.2 Class Diagram

1. **AutomatedNotesMaker**:
    - Attributes:
        - **audioRecordings**: A list of **Audio** objects.
        - **videoRecordings**: A list of **Video** objects.
    - Methods:
        - **makeNotes()**: Method to initiate the process of making notes from audio and video recordings.
        - **searchNotes(keyword: String)**: Method to search for notes based on keywords.
2. **Audio**:
    - Attributes:
        - **audioFile**: The audio file for the recording.
        - **text**: The text transcribed from the audio file.
    - Methods:
        - **processAudio()**: Method to process the audio file and convert it into text.
3. **Video**:
    - Attributes:
        - **videoFile**: The video file for the recording.
        - **audio**: An **Audio** object representing the audio extracted from the video.
    - Methods:
        - **processVideo()**: Method to process the video file, extract the audio, and convert it into text.
4. **Note**:
    - Attributes:
        - **content**: The textual content of the note.
        - **keywords**: A list of keywords associated with the note.
    - Methods:
        - **addKeyword(keyword: String)**: Method to add a keyword to the note.
        - **removeKeyword(keyword: String)**: Method to remove a keyword from the note.

This diagram shows a basic structure for the classes involved in an automated notes maker system. Depending on the specific requirements and functionalities of your system, you can expand upon or modify these classes accordingly.

4.5.3. OBJECT DIAGRAM

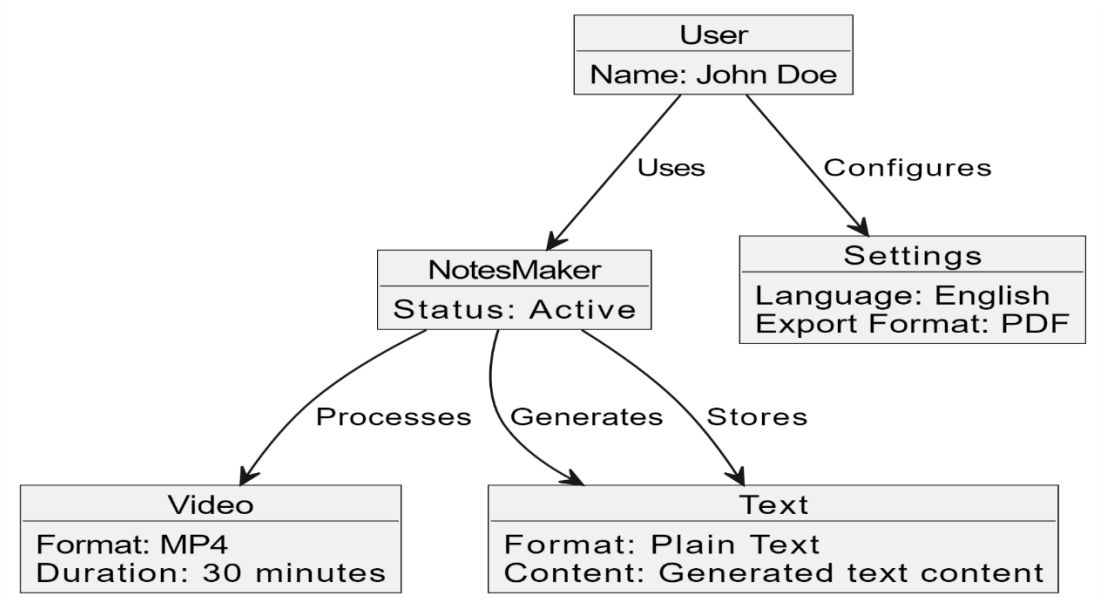


Fig. 4.5.3 Object Diagram

Object Descriptions:

1. AutomatedNotesMaker:
    - audioRecordings: Empty for this example.
    - videoRecordings: Empty for this example.
2. Audio:
    - audioFile: "audio1.mp3"
    - text: "Transcribed text from audio1"
    - Object:
        - Audio1
3. Video:
    - videoFile: "video1.mp4"

- audio: Audio1 (from the Audio object)
- Object:

In this object diagram:

- Audio1 represents an audio file with the file name "audio1.mp3" and its transcribed text.
- Video1 represents a video file with the file name "video1.mp4" and its associated audio from Audio1.
- Note1 represents notes generated from Video1 with the content "Notes from Video1" and keywords "meeting" and "summary".

4.5.4 SEQUENCE DIAGRAM

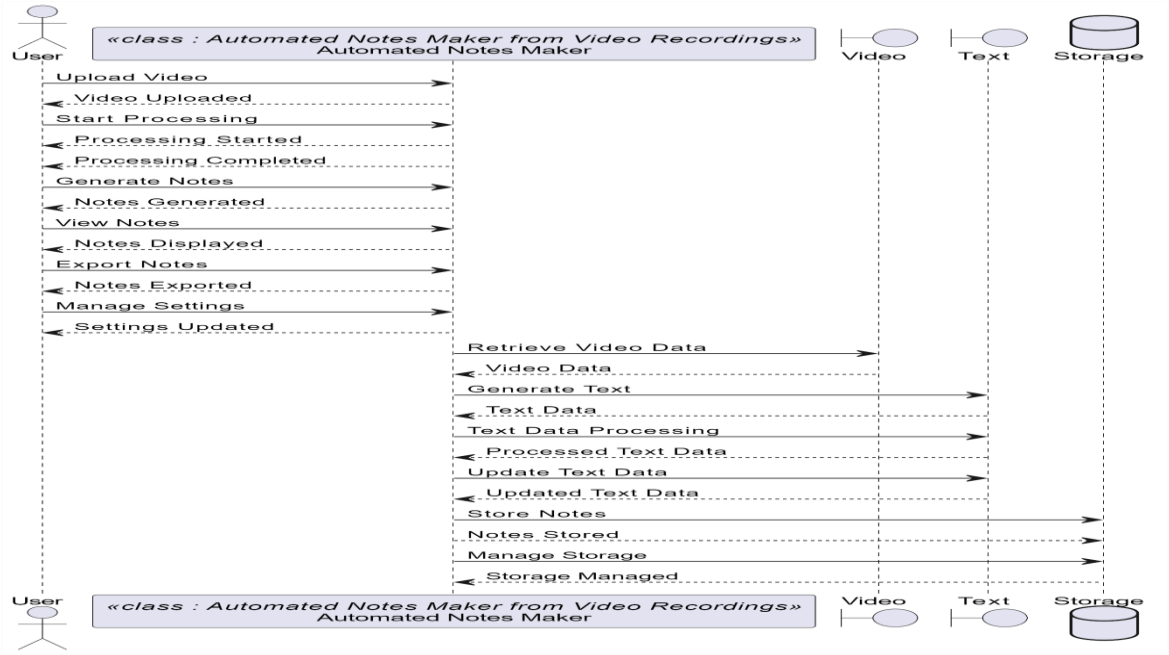


Fig. 4.5.4 Sequence Diagram

Sequence of Events:

1. User triggers recordingStarted().
2. AutomatedNotesMaker initiates processAudio() and processVideo() on Audio and Video objects.

3. When Audio processing is complete, it sends audioProcessingComplete() to AutomatedNotesMaker.
4. When Video processing is complete, it sends videoProcessingComplete() to AutomatedNotesMaker.
5. AutomatedNotesMaker creates Note using the processed Audio and Video.
6. AutomatedNotesMaker sends notesGenerated(Note1) back to the User with the generated Note.

4.5.5. STATE CHART DIAGRAM

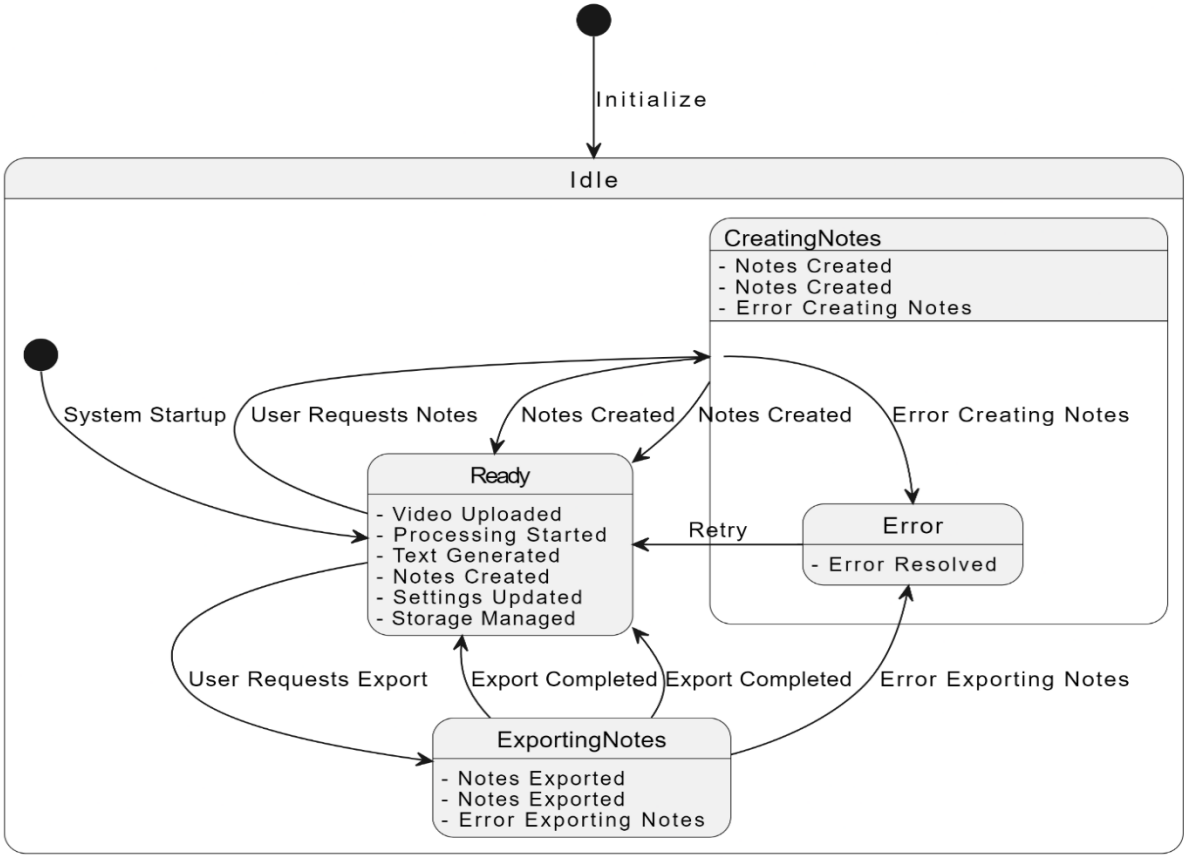


Fig. 4.5.5 State Chart Diagram

In this state chart diagram:

- The system starts in the Idle state.

- When the user triggers the process of making notes, the system transitions to the Processing state.
- If the system detects an audio recording, it transitions to the Processing Audio state and stays there until the audio processing is complete, then goes back to Idle.
- If the system detects a video recording, it transitions to the Processing Video state and stays there until the video processing is complete, then goes back to Idle.

4.5.6.ACTIVITY DIAGRAM

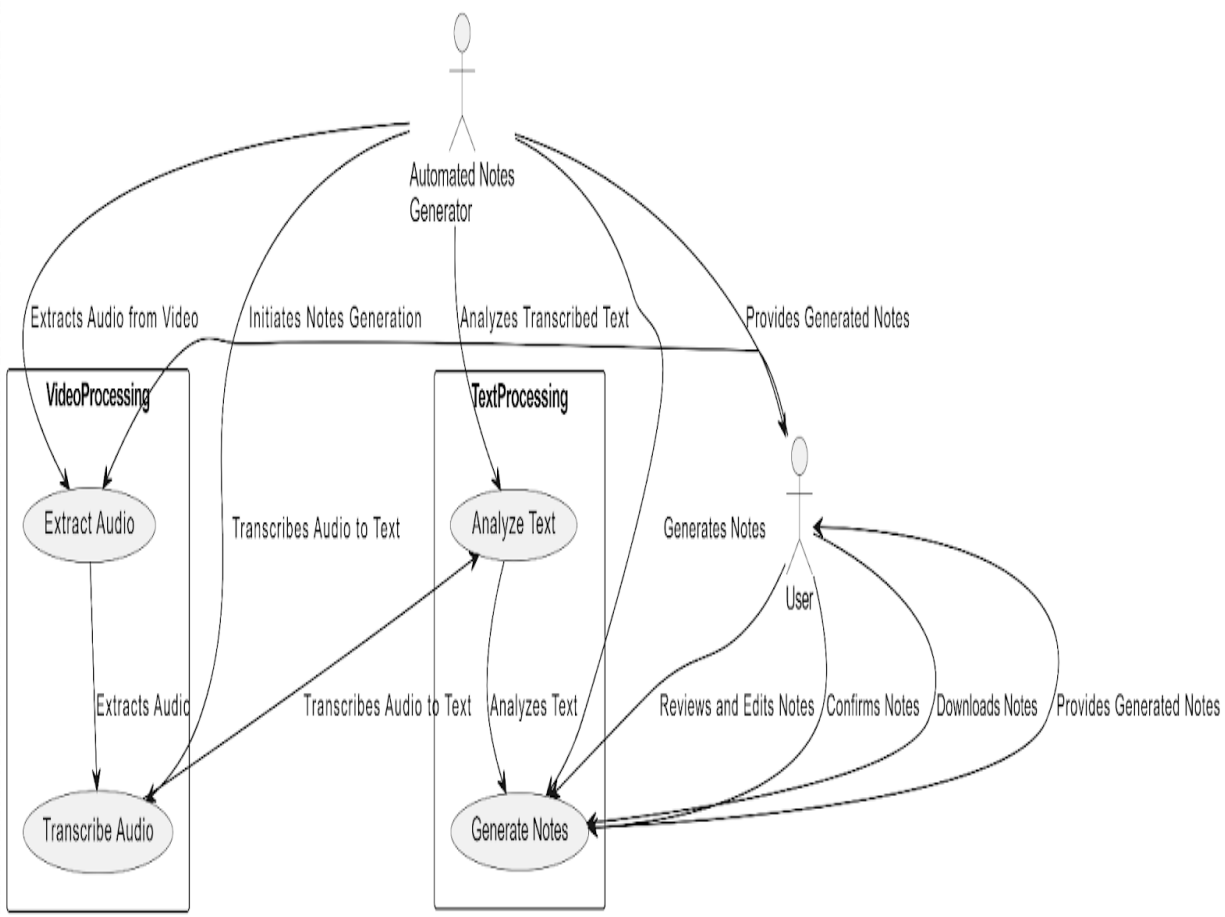


Fig. 4.5.6 Activity Diagram

This activity diagram represents the workflow of making notes from audio and video recordings in the automated notes maker system. Users start by choosing the type of recording they want to process, then the system follows through the steps of processing the recording, generating text notes, and saving the notes.Please note that this is a simplified example, and in a real system, there could be additional steps, decision points, loops, or alternative flows depending on the system requirements and complexity.

4.5.7 Deployment Diagram

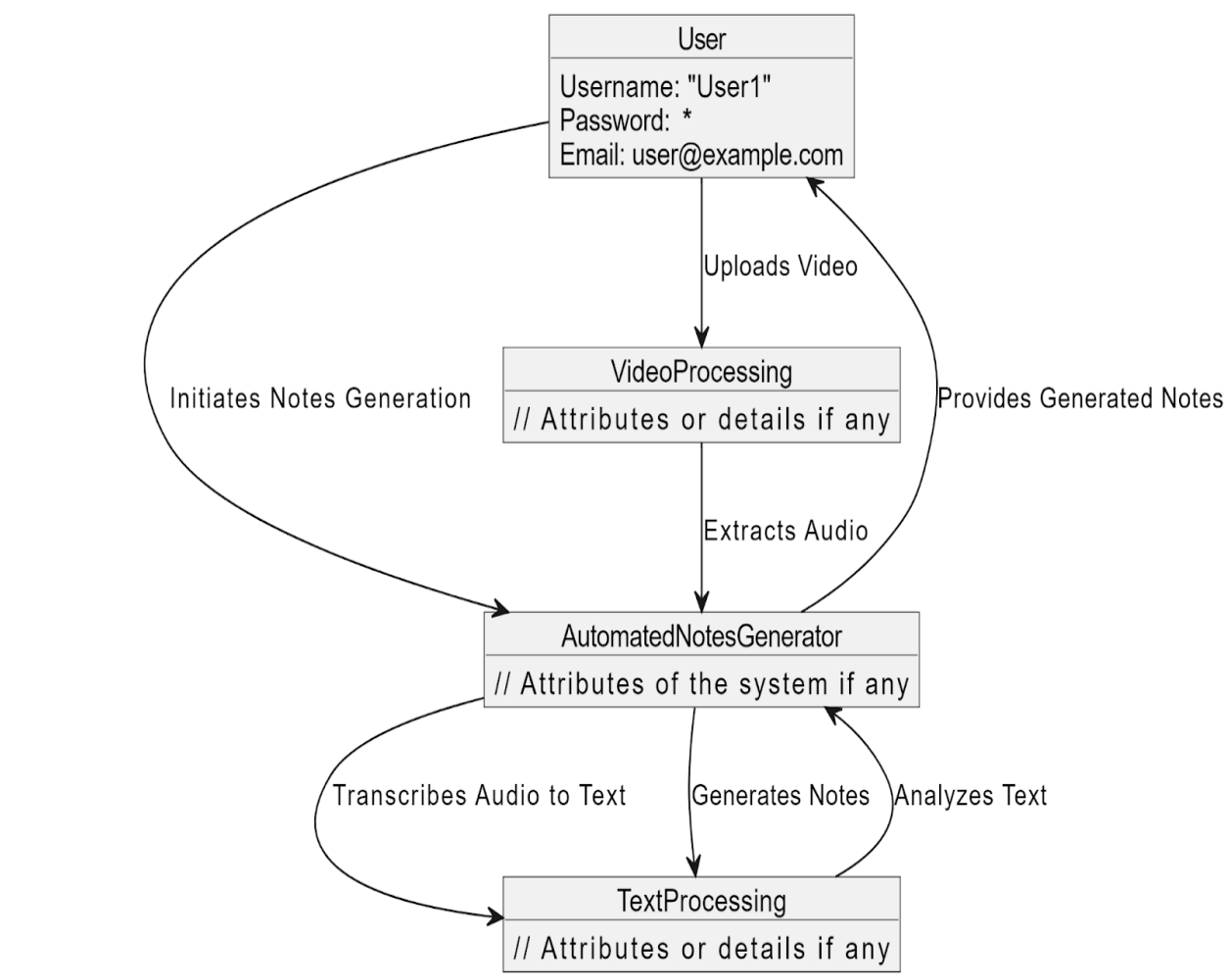


Fig. 4.5.7 Deployment Diagram

1. **User Device**:

- Represents the device used by the user to interact with the system.
- Contains the "User Interface" component.

2. **Application Server**:
    - Represents the server where the main application logic resides.
    - Contains the "Automated Notes Maker" component.
3. **Database Server**:
    - Represents the server where the notes and recordings data are stored.
    - Contains the "Notes Database" component.

Connections:

- **UserDevice** communicates with the **User Interface** component through HTTP requests.
- **User Interface** component communicates with the **Automated Notes Maker** component through HTTP requests.
- **Automated Notes Maker** component interacts with the **Notes Database** for storing and retrieving data.

This diagram shows a basic deployment setup for an automated notes maker system, with the user interacting via a User Interface, the main application logic running on an Application Server, and the data stored in a Database Server.

Feel free to customize this diagram according to your specific deployment architecture, such as adding more nodes, components, or detailing communication protocols as per your system's requirements.

4.6 SOURCE CODE

4.6.1 Implementation of Transcription of a Extracted Recording:

```
import torch
import textwrap
from fpdf import FPDF
import whisperx
import gc
```

```
from transformers import AutoModelForSpeechSeq2Seq, AutoProcessor, pipeline
from google.colab import files
from reportlab.lib.pagesizes import letter
from reportlab.pdfgen import canvas
from reportlab.lib import colors
from reportlab.platypus import SimpleDocTemplate, Paragraph, Spacer
from reportlab.lib.styles import getSampleStyleSheet
import os
import string
class GenDoc:
  def _init_(self):
    self.device="cuda" if torch.cuda.is_available() else "cpu"
    print(f'Running on {self.device} ...')
    self.batch_size = 4 # reduce if low on GPU mem
    self.compute_type = "float16"
    self.model=whisperx.load_model("large-v3", self.device, compute_type=self.compute_type)
    self.Audizer=ExtractAudio()
  def trans(self,user_folder,subtitles_length=100):
    audio_path=user_folder+audio_filename
    global wlang
    global extracted_text
    print(f"Transcribing file: {audio_path}\n")
    audio = whisperx.load_audio(audio_path)
    result_text = self.model.transcribe(audio, batch_size=self.batch_size,language='en')
    print(result_text["segments"]) # before alignment
    extracted_text=".join([segment['text'] for segment in result_text["segments"]])
    self.toTextFile(extracted_text, f"{user_folder}{generate_filename}.txt")
    self.toTextFile(str(result_text["segments"]), f"{user_folder}{extracted_filename}.txt")
    model_a, metadata = whisperx.load_align_model(language_code=result_text["language"],
device=self.device)
```

```
    result = whisperx.align(result_text["segments"], model_a, metadata, audio, self.device,
return_char_alignments=False)
    result['language']=metadata['language']
    self.toTextFile(str(result), f"{user_folder}{'result_File'}.txt")

from moviepy.editor import VideoFileClip
from pathlib import Path
import requests
import os
from pytube import YouTube
from tqdm import tqdm
import urllib.request
import soundfile as sf

class ExtractAudio:
  def convert_wav(self,input_audio_path,user_folder):
    data, rate = sf.read(input_audio_path)
    sf.write(user_folder+audio_filename, data, rate, format='wav', subtype='PCM_24')
    os.remove(input_audio_path)

  def extract_path(self,vpath, opath):
    try:
        video_clip = VideoFileClip(vpath)
        audio_clip = video_clip.audio
        audio_clip.write_audiofile(opath)
        video_clip.close()
        audio_clip.close()
        print(f"{opath} extract audio done")
        return 1
    except Exception as e:
        print(f"Error: {e}")
```

```
        return 0
  def extract_dlink(self,url,user_folder):
    try:
      if '.mp4' not in url:
        print("'This isn't the Video link..!
        Please provide the video url with (.mp4) at its end '")
        return 0
      output_folder=user_folder+'/'
      video_path=output_folder+'downloaded_video.mp4'
      audio_path=output_folder+audio_filename
      with tqdm(unit='B', unit_scale=True, unit_divisor=1024, miniters=1) as progress:
          def report(block_num, block_size, total_size):
              progress.total = total_size
              progress.update(block_num * block_size - progress.n)
          urllib.request.urlretrieve(url, video_path, reporthook=report)
      status=self.extract_path(video_path,audio_path)
      if os.path.isfile(video_path):
        os.remove(video_path)
return 0 if status==0 else 1
    except Exception as e:
      print(e)
      return 0
Audizer=ExtractAudio()
```

CHAPTER 5

RESULTS

The audio processing and noise reduction stage demonstrated significant improvements in audio quality. By applying the noiseReduce algorithm with FFT, both stationary and non-stationary noise components were effectively removed, resulting in clearer audio signals suitable for further analysis.

5.1 Transcription & Topic Modelling of recording:

The transcription process yielded accurate text representations of the audio content. The Whisper Large V3 ASR model performed well in converting audio signals to text, with high accuracy and minimal errors. Timestamping of the text facilitated seamless subtitle generation, enhancing the accessibility of the content.

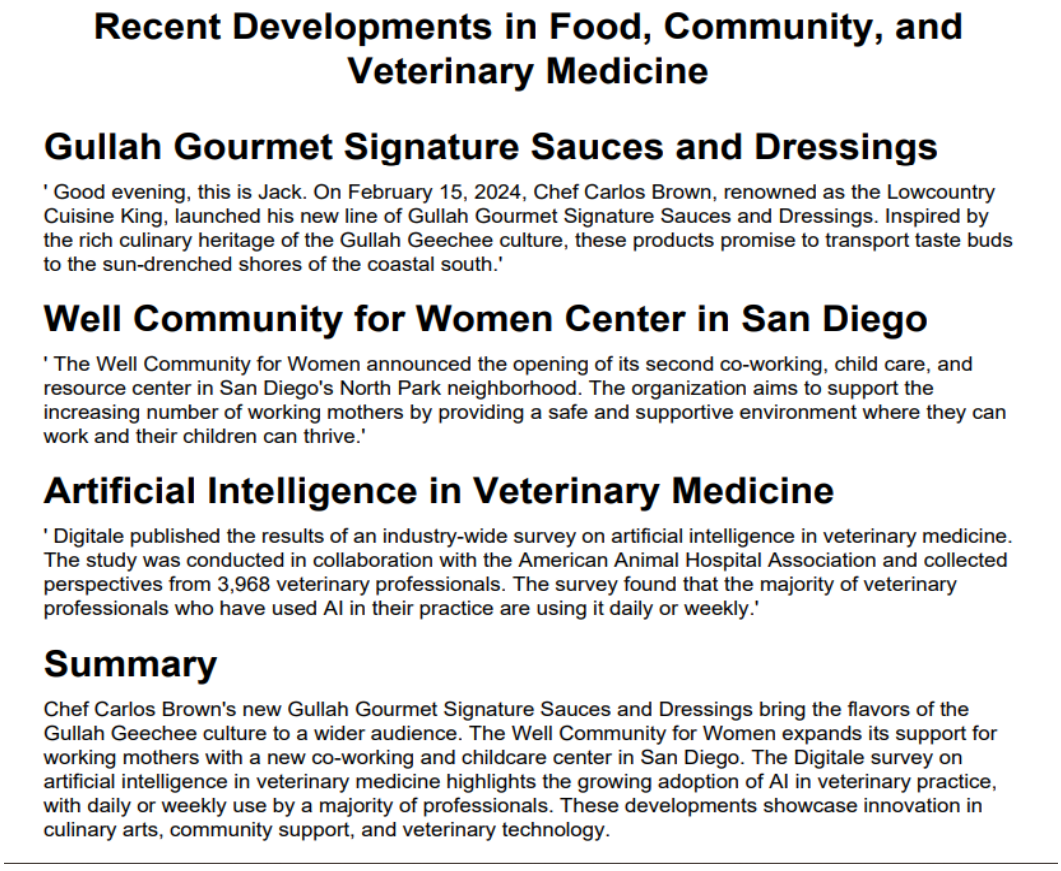


Fig. 5.1.1 Topic Modelling

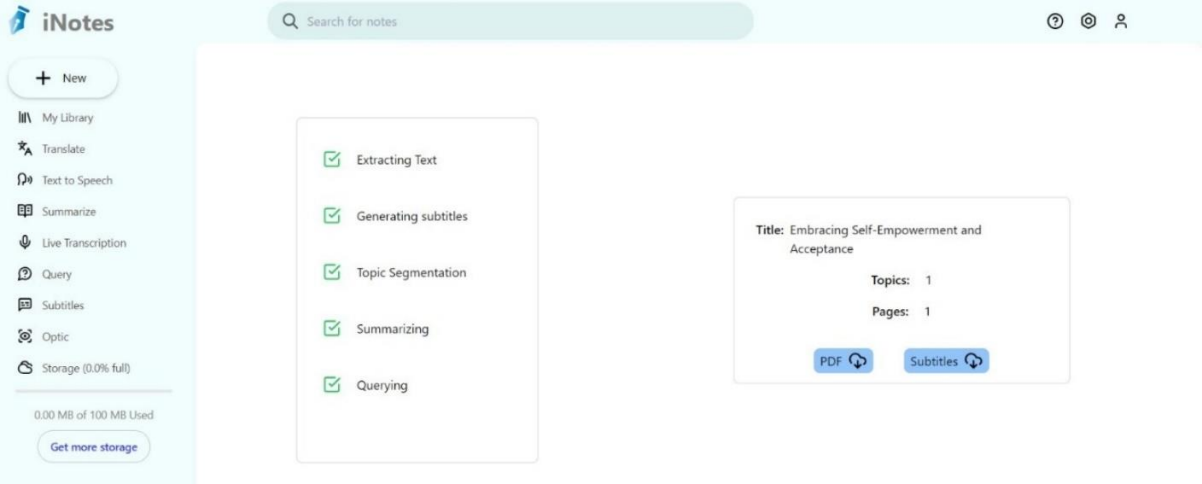


Fig. 5.1.2 Topic Modelling

5.2 Translation:

Language translation capabilities are implemented next, leveraging the Deep-translator library with a character limit of 5000. This integration enables precise translations across 32 supported languages. Accurate language detection using Langdetect ensures seamless communication across diverse linguistic landscapes.

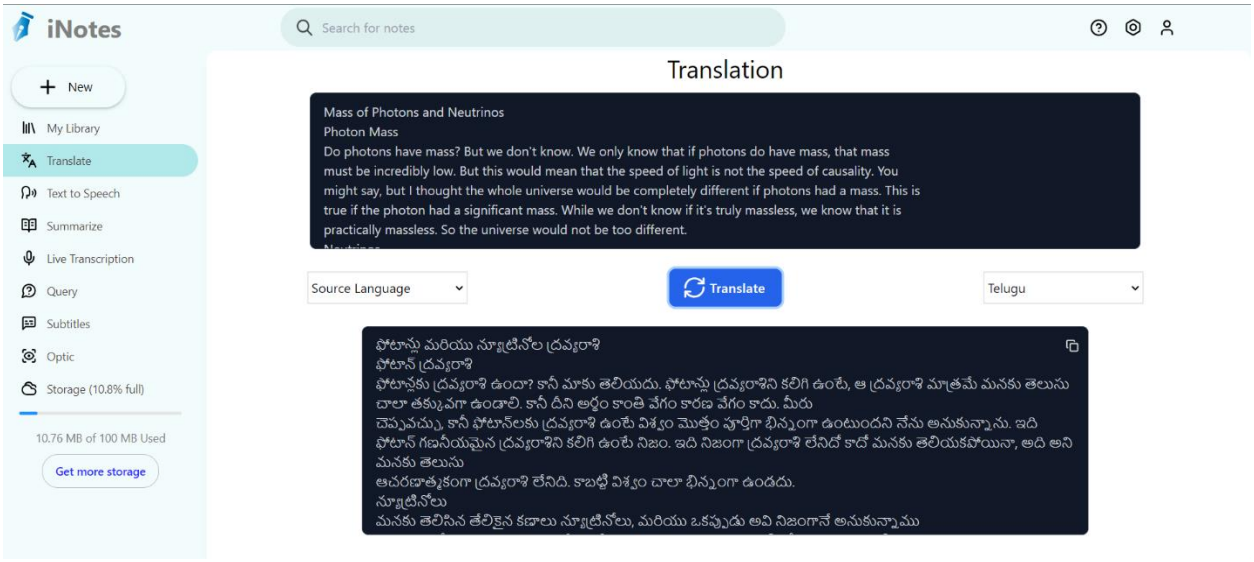


Fig. 5.2 Translation

5.3 Text Querying and Retrieval:

The system implements text querying and retrieval functionalities using cutting-edge techniques such as vector embeddings and cosine similarity computation. Large Language Models (LLMs)

structure both retrieved content and user queries, providing comprehensive and well-organized answers. It had given accurate answers for user queries as shown below sample.

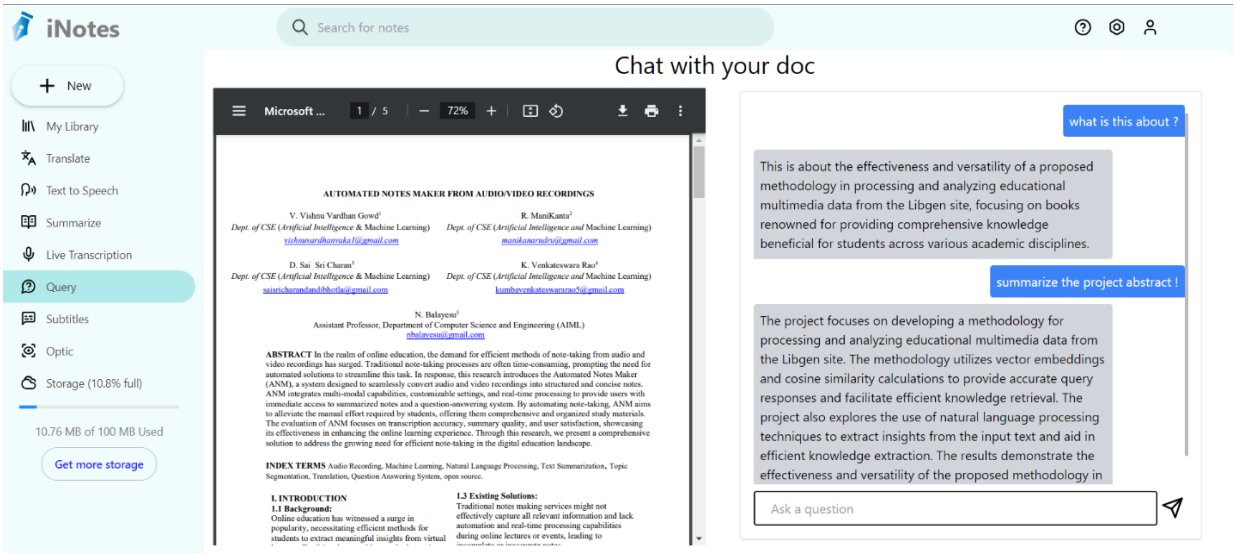


Fig. 5.3 Querying a document

5.4 Text-to-Speech Conversion:

The utilization of the TTS/XTTS_v2 model has enabled us to generate remarkably realistic voices that closely mimic human speech patterns and intonations. This has been achieved through the incorporation of advanced techniques such as deep learning and neural network architectures. By synthesizing speech based on speaker voice profiles stored in a database.

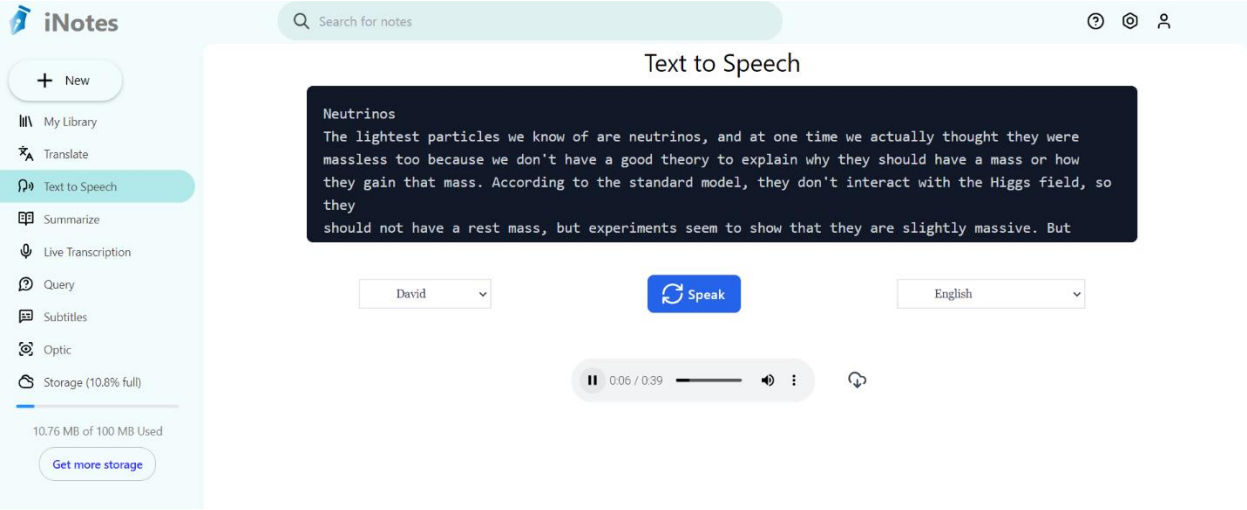


Fig. 5.4 Text to speech generation

5.5 Summarization:

The summarization module effectively generated concise summaries of lengthy documents. The Google Generative AI model Gemini-pro produced coherent and informative summaries, capturing key insights from the input text and aiding in efficient knowledge extraction.

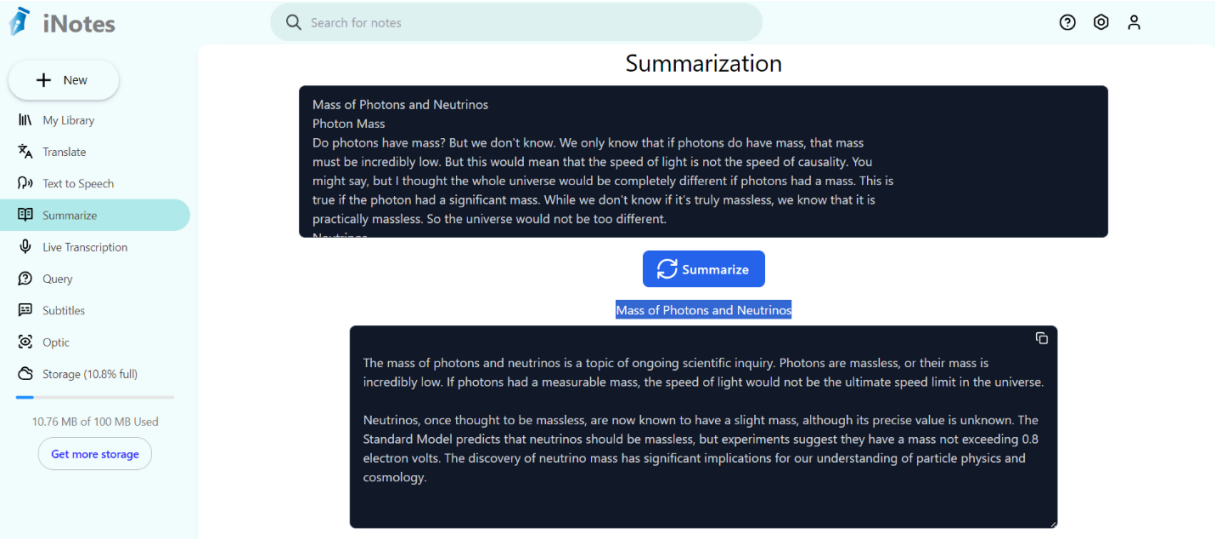


Fig. 5.5 Summarization of a long text chunk

5.6 Image Querying Implementation:

The Gemini Vision Pro model provided deeper insights into visual data through image-based interactions. Accurate results were achieved while testing on medical prescriptions, famous personality identification, text extraction. The results are accurate and the communication is as a human speaking in the backend. A sample output of the generated information of image is shown.

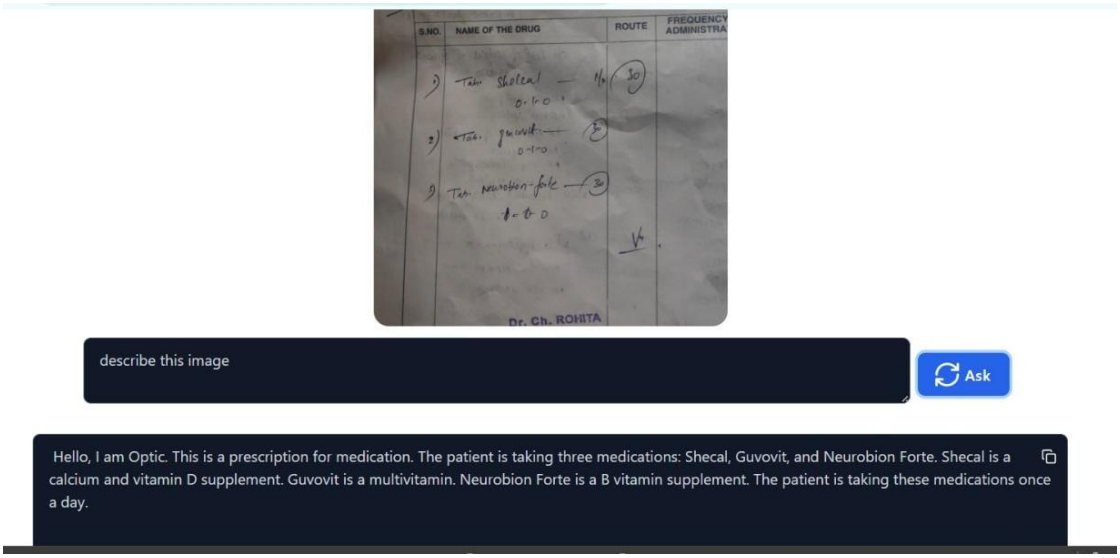


Fig. 5.6 Image Querying of a medical prescription

5.7 Live Transcription Setup:

The React Speech Recognition library has demonstrated commendable accuracy in transcribing spoken words in real-time. Through rigorous testing and evaluation, we have observed consistent and reliable transcription results across various speaking styles, accents, and environmental conditions.

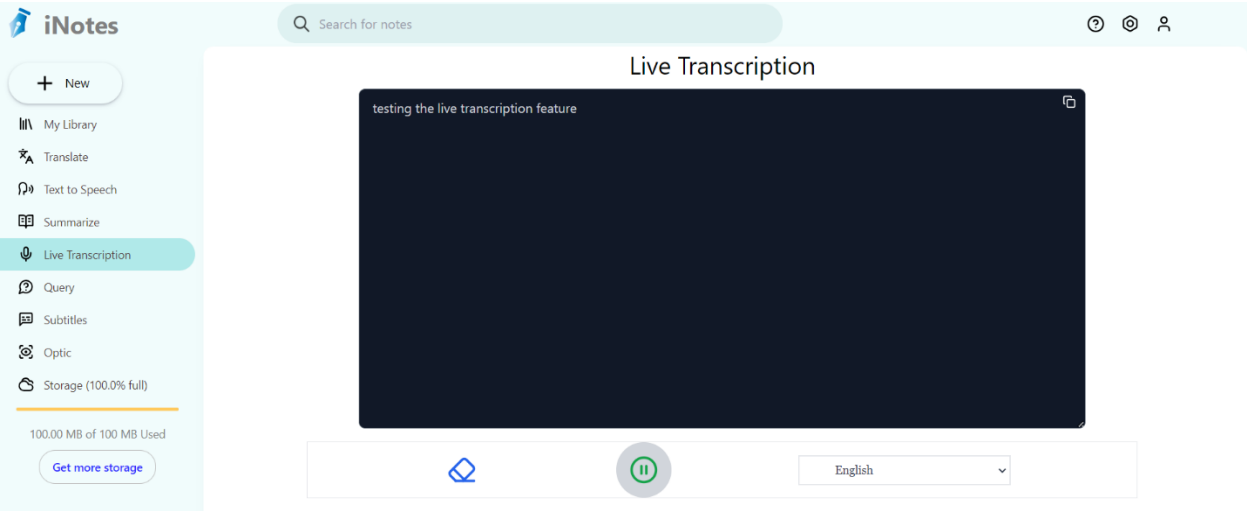


Fig. 5.7 Live Transcription of a speech

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

The project aims to revolutionize the note-taking process in online education by introducing the Automated Notes Maker (ANM). ANM is designed to address the time-consuming task of manually documenting long speeches and lectures, thereby enhancing the learning experience for students. By automatically generating text-based PDF/Word documents from audio/video-based classes, ANM significantly reduces the burden of note-taking, allowing students to focus more on understanding and retaining the content.

The research project explores various feature extraction techniques and speech understanding methodologies to construct a voice recognition system capable of supporting multi-language content. By leveraging advanced technologies such as text summarization, querying and answering, transcription and subtitles, and live transcription, ANM offers a comprehensive solution for efficient content comprehension and retrieval.

One of the key benefits of ANM is its ability to produce notes in multiple languages, facilitating language translation and making educational content more accessible to students from diverse linguistic backgrounds. Moreover, the system's support for text-to-speech conversion and image querying enhances the learning experience by providing alternative ways to engage with the content.

By automating the note-taking process and providing real-time access to summarized notes, ANM empowers students to engage more effectively with educational materials. This not only saves valuable study time but also promotes better comprehension and retention of information. Additionally, the integration of a book store offering comprehensive educational resources and image generation from text aims to further enrich the learning experience and foster creativity among students.

In future iterations, ANM plans to expand its offerings by integrating additional features such as personalized recommendations for study materials and collaborative tools for group learning. These enhancements aim to continue improving the educational experience for students and educators alike, ultimately leading to improved academic performance and satisfaction with the learning process.

CHAPTER 7
REFERENCES

[1] Manoj Kumar A1 , Janani P2 , Siva Subramanian G3 , Kumaragurubaran K4 , Sundari P, “Automated Notes Maker from Audio Recordings” ISSN 2582-7421.

[2] Chaudhari Mahima1 , Mali Divya2 , Chaudhari Nehal3 , Kolhe Trupti4 , Ashish T. Bhole5, “Automated Notes Maker from Audio Recordings” ISSN (O) 2278-1021.

[3] Ms. Purva Chavrekar, 2Ms. Shruti Deshmukh, 3Ms. Pranjal Khade, 4Ms.Vaibhavi Patil 5Prof. Rupali Sathe, “AUTOMATED NOTES MAKER FROM AUDIO RECORDING” 2023 IJRTI | Volume 8, Issue 4 | ISSN: 2456-3315.

[4] Speech to text conversion and summarization for effective understanding and documentation (https://www.researchgate.net/publication/342147736_Speech_to_text_conversion_and_summarization_for_effective_understanding_and_ documentation).

[5] Prerana Das, Kakali Acharjee, Pranab Das, Vijay Prasad “VOICE RECOGNITION SYSTEM: SPEECH-TO-TEXT” 01 July 2016.

[6] Ms. Anuja Jadhav, Prof. Arvind Patil, Real Time Speech Text Converter for Mobile Users, National Conference on Innovative Paradigms in Engineering Technology (NCIPET-2012) Proceedings published by International Journal of Computer Applications (IJCA).

[7]. Hakan Erdogan, Ruhi Sarikaya, Yuqing Gao, “Using semantic analysis to improve speech recognition performance” Computer Speech and Language, ELSEVIER 200.

[8]. Chen, Jingdong, Yiteng Huang, Qi Li, and Kuldip K. Paliwal. ” Recognition of noisy speech using dynamic spectral sub band centroids.” IEEE signal processing letters 11, no. 2 (2004): 258-261.

[9]. y Keiichi Tokuda, Yoshihiko Nankaku, Tomokin Toda, Heiga Zen, Speech Synthesis Based on Hidden Markov Models, Proceedings of the IEEE — Vol. 101, No. 5, May 2013. Junichi Yamagishi, Member IEEE, and Keiichiro Oura.

[10]. F. Seide, G. Li, D. Yu, Conversational Speech Transcription Using Context-Dependent Deep Neural Networks, In Interspace, pp. 437440, 2011.

[11]. Muhammad Yasir, Marlince NK ,Nababan, Yonata Laia, Windania Purba, Robin ,Asaziduhu Gea, "Web-Based automation speech-to -text application using audio recording for meeting speech",2019.

[12]. Shivangi Nagdewani, Ashika Jain, "A REVIEW ON METHODS FOR SPEECH-TO-TEXT AND TEXT-TO-SPEECH CONVERSION", Volume: 07 issue: 05 | May 2020.

[13]. Lawrence Rabiner, Biing-Hwang Juang, B.Yegnanarayana, Fundamentals of Speech Recognition 978-0-13-015157-5.

[14]. Tim Sainburg, Noise Reduction: A model for improving clarity and quality of the audio signals. (https://timsainburg.com/noise-reductionpython.html)

[15] Alec Radford, Jong Wook Kim, Tao Xu 1 Greg Brockman, Christine McLeavey "Robust Speech Recognition via Large-Scale Weak Supervision".

[16] Balayesu, N., Reddy, A.A. Deep pelican based synthesis model for photo-sketch face synthesis and recognition. Multimed Tools Appl (2024).

[17] Balayesu, N. Kalluri, H.K. An extensive survey on traditional and deep learning-based face sketch synthesis models. Int. j. inf. tecnol. 12, 995–1004 (2020).

[18] Balayesu, N. (2019). Optimal Pyramid Column Feature with Contrast Enhanced Model for Face Sketch Synthesis. Journal-Of-Advanced-ResearchIn-Dynamical-And-Control-Systems, 11(5), 335- 344.

CERTIFICATE OF PUBLICATION

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, Open Access Journal since 2012)

IJIRSET

Impact Factor 8.423

The Board of IJIRSET is hereby awarding this certificate to

V. VISHNU VARDHAN GOWD

Dept. of CSE-Artificial Intelligence and Machine Learning, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

In Recognition of publication of the paper entitled

"Automated Notes Maker from Audio/Video Recordings"

Published in IJIRSET, Volume 13, Issue 3, March 2024

Google scholar | Crossref | Mendeley | doi | INNO SPACE SJIF Scientific Journal Impact Factor

e-ISSN: 2319-8753
p-ISSN: 2347-6710

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

P. Kumar
Editor-in-Chief

www.ijirset.com ijirset@gmail.com

CERTIFICATE OF PUBLICATION

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, Open Access Journal since 2012)

IJIRSET

Impact Factor 8.423

The Board of IJIRSET is hereby awarding this certificate to

R. MANIKANTA

Dept. of CSE-Artificial Intelligence and Machine Learning, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

In Recognition of publication of the paper entitled

"Automated Notes Maker from Audio/Video Recordings"

Published in IJIRSET, Volume 13, Issue 3, March 2024

Google scholar | Crossref | Mendeley | doi | INNO SPACE SJIF Scientific Journal Impact Factor

e-ISSN: 2319-8753
p-ISSN: 2347-6710

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

P. Kumar
Editor-in-Chief

www.ijirset.com ijirset@gmail.com

CERTIFICATE
OF PUBLICATION

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, Open Access Journal since 2012)

IJIRSET

Impact Factor 8.423

The Board of IJIRSET is hereby awarding this certificate to

D. SAI SRI CHARAN

Dept. of CSE-Artificial Intelligence and Machine Learning, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

In Recognition of publication of the paper entitled

"Automated Notes Maker from Audio/Video Recordings"

Published in IJIRSET, Volume 13, Issue 3, March 2024

Google scholar | Crossref | Mendeley | doi | INNO SPACE SJIF Scientific Journal Impact Factor

e-ISSN: 2319-8753
p-ISSN: 2347-6710

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

P. Kumar
Editor-in-Chief

www.ijirset.com ijirset@gmail.com

CERTIFICATE OF PUBLICATION

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, Open Access Journal since 2012)

IJIRSET

Impact Factor 8.423

The Board of IJIRSET is hereby awarding this certificate to

K. VENKATESWARA RAO

Dept. of CSE-Artificial Intelligence and Machine Learning, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

In Recognition of publication of the paper entitled

"Automated Notes Maker from Audio/Video Recordings"

Published in IJIRSET, Volume 13, Issue 3, March 2024

Google scholar | Crossref | Mendeley | doi | INNO SPACE SJIF Scientific Journal Impact Factor

e-ISSN: 2319-8753
p-ISSN: 2347-6710

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

P. Kumar
Editor-in-Chief

www.ijirset.com ijirset@gmail.com

CERTIFICATE
OF PUBLICATION

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, Open Access Journal since 2012)

IJIRSET

Impact Factor 8.423

The Board of IJIRSET is hereby awarding this certificate to

N. BALAYESU

Assistant Professor, Dept. of CSE-Artificial Intelligence and Machine Learning, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India

In Recognition of publication of the paper entitled

"Automated Notes Maker from Audio/Video Recordings"

Published in IJIRSET, Volume 13, Issue 3, March 2024

Google scholar | Crossref | Mendeley | doi | INNO SPACE SJIF Scientific Journal Impact Factor

e-ISSN: 2319-8753
p-ISSN: 2347-6710

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

P. Kumar
Editor-in-Chief

www.ijirset.com ijirset@gmail.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303138 |

Automated Notes Maker from Audio/Video Recordings

V. Vishnu Vardhan Gowd[1], R. ManiKanta[2,] D. Sai Sri Charan[3], K. Venkateswara Rao[4,]

N. Balayesu[5]

Dept. of CSE-Artificial Intelligence and Machine Learning, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India[1, 2,3,4,]

Assistant Professor, Dept. of CSE-Artificial Intelligence and Machine Learning, Vasireddy Venkatadri Institute of Technology, Guntur, Andhra Pradesh, India[5]

ABSTRACT In the realm of online education, the demand for efficient methods of notetaking from audio and video recordings has surged. Traditional note-taking processes are often time-consuming, prompting the need for automated solutions to streamline this task. In response, this research introduces the Automated Notes Maker (ANM), a system designed to seamlessly convert audio and video recordings into structured and concise notes. ANM integrates multi-modal capabilities, customizable settings, and real-time processing to provide users with immediate access to summarized notes and a question-answering system. By automating notetaking, ANM aims to alleviate the manual effort required by students, offering them comprehensive and organized study materials. The evaluation of ANM focuses on transcription accuracy, summary quality, and user satisfaction, showcasing its effectiveness in enhancing the online learning experience. Through this research, we present a comprehensive solution to address the growing need for efficient notetaking in the digital education landscape.

INDEX TERMS Audio Recording, Machine Learning, Natural Language Processing, Text Summarization, Topic Segmentation, Translation, Question Answering System, open source.

I. INTRODUCTION

1.1 Background:

Online education has witnessed a surge in popularity, necessitating efficient methods for students to extract meaningful insights from virtual lectures. Traditional note-taking methods are time-consuming, prompting the development of automated tools like the Automated Notes Maker (ANM) to streamline the process. ANM seamlessly converts audio and video recordings into structured notes, enhancing accessibility and comprehension in the digital learning landscape.

1.2 Problem Statement:

Traditional note-taking methods in online education are time-consuming and often inefficient, hindering students' ability to review and retain information from virtual lectures. There is a pressing need for automated tools to seamlessly convert audio and video recordings into structured notes, addressing transcription accuracy, summary quality, and user satisfaction.

1.3 Existing Solutions:

Traditional notes making services might not effectively capture all relevant information and lack automation and real-time processing capabilities during online lectures or events, leading to incomplete or inaccurate notes.

1.4 Proposed Solution:

ANM offers an innovative solution by leveraging AI to create a user-centric platform.

- Extracting audio signals and removing stationary/non-stationary noise using Fast Fourier Transform (FFT).
- Transcribe into text using state-of-the-art ASR model Whisper Large V3, with resulting text timestamped for subtitle generation.

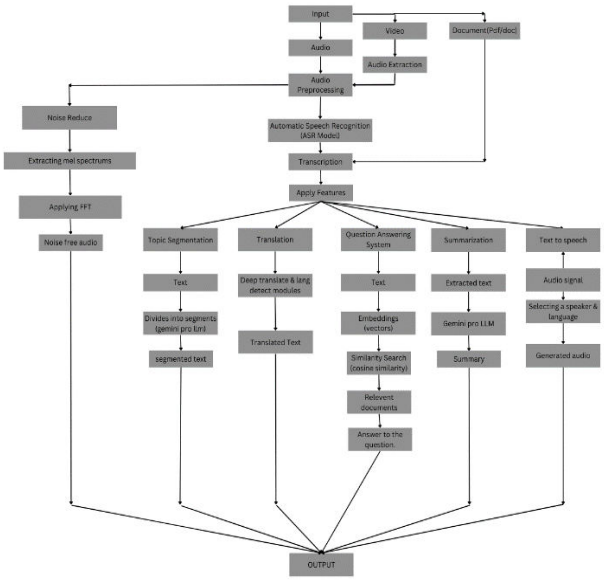
|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

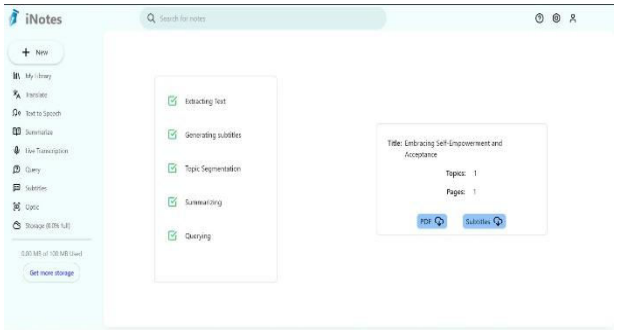
| DOI:10.15680/IJIRSET.2024.1303138 |

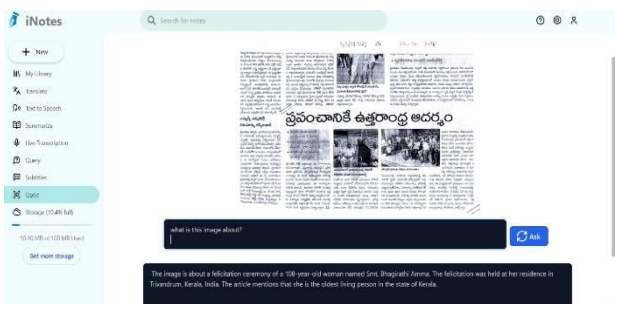
- Language translation using the deep-translator library, ensuring accessibility across 32 supported languages.
- Vector embeddings and cosine similarity calculations, providing accurate query responses and facilitating efficient knowledge retrieval.
- Text to realistic speech using TTS/XTTS_v2 model, accommodating auditory learners and individuals with visual impairments.
- Google Generative AI model Gemini-pro for summarizing lengthy documents.
- Cloud storage with MongoDB via Fast API.
- Interaction with images with Gemini Vision Pro model.

1.5 Flow Diagram:



1.6 Sample Outputs:





|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303138 |

II. RELATED WORK

Chaudhari Mahima [1] presented an approach for conversion of speech signals(audio) into text- format (PDF/Word) based on the user interest. However, the techniques and methods are currently available in high complex, expensive setups. Their focus is to solve the task in a simple setup. Such a solution would be of great importance, and it would be useful in general. Their application enables user to take automated notes from audio recording, the type of audio file required is MPs. The deep gram open-source library is used in their application to convert audio recording into notes. The use of deep gram library to convert audio recordings into automated notes gives better accuracy of 80%.

Manoj Kumar, Janani, Siva Subramanian and Kumaragurubaran [2] This paper claims that Natural language processing (NLP) and machine learning techniques are used as the technologies to transform spoken words into text, which is subsequently used to build a summary or set of notes. This project report looks at the effectiveness and utilization of automated note-taking software for audio recordings with the 85% accuracy.

Purva Chavrekar, Shruti Deshmukh, Pranjal Khade, and Vaibhavi Patil [3] The paper's objective is to create a system that allows a computer to translate voice requests and diction text using MFCC and VQ methods. For feature extraction and matching, the Mel Frequency Campestral Coefficient and Vector Quantization Method will be utilized. The automated process can convert the voice to text and describe the material, this model can be utilized anywhere for long lectures needs to be condensed into exact texts with accuracy of 85%.

III. DATASETS

In this study, For book data collection in the book store feature, the Libgen site was utilized to gather a curated selection of educational books known for providing high-quality knowledge and resources beneficial for students. These books were specifically chosen for their relevance and comprehensiveness in various academic disciplines, ensuring that students have access to the best learning materials available. Additionally, to enhance the language capabilities of the system, the Mozilla Foundation's Common Voice 14.0 dataset was employed for fine-tuning the Whisper model specifically for the Telugu language. This dataset provided a valuable resource for training the model to accurately transcribe and process Telugu audio content, thereby expanding the system's language support and improving its effectiveness for users who prefer or require Telugu language support.

IV. IMPLEMENTED SYSTEM

Methodology:
Our platform encompasses a comprehensive array of features meticulously designed to cater to diverse people needs. The system architecture revolves around modularity, scalability, and optimal user experience.

4.1 Audio processing and noise reduction:

- Extracting audio signals from video recordings or direct links is the first step.
- Noise reduction techniques are then applied to clean the audio signals.
- The noiseReduce algorithm is employed for this purpose.
- The noiseReduce algorithm utilizes Fast Fourier Transform (FFT).
- FFT is utilized to effectively remove both stationary and non-stationary noise components from the audio signals.

4.2 Transcription and Subtitle Generation:

- Clean audio signals undergo transcription into text post noise reduction.
- The Whisper Large V3 ASR model is employed for transcription purposes.
- Transcription results are timestamped for subtitle generation.
- Timestamping facilitates convenient navigation through the content for users.

4.3 Translation:

- Deep-translator library is utilized for language translation to enhance accessibility.
- Langdetect is used to detect the language of the input text accurately.
    - Precise translations are conducted across 32 supported languages.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303138 |

4.4 Querying:

- Text chunks are extracted and converted into vector embeddings using FAISS.
- Cosine similarity is computed between user queries and vector embeddings to retrieve relevant content.
- Large Language Models (LLMs) structure the retrieved content and user queries to provide well-organized answers.

4.5 Text-to-Speech Conversion:

- The framework provides text-to-speech conversion capabilities for auditory learners or those with visual impairments.
- It utilizes the TTS/XTTS_v2 model to generate realistic voice based on speaker voice profiles stored in a database.

4.6 Summarization:

- Utilization of the Google Generative AI model Gemini-pro for document summarization.
- The model generates concise summaries of lengthy documents.

4.7 Topic Segmentation:

- Utilization of the Gemini-pro model to prompt with text chunks for topic extraction.
- Extraction of topics and their corresponding content for providing users with a structured overview of the document.

4.8 Cloud Storage:

- Storage of generated notes and documents in a cloud-based MongoDB database.
- FastAPI employed for efficient conversion of documents to BSON format.
- BSON format chosen for its efficiency in storage and retrieval operations.

4.9 Image Querying:

- Users enabled to interact with images through questions and text extraction similar to OCR.
- Utilization of the Gemini Vision Pro model for image querying tasks.
- Aim to provide users with deeper insights into visual data through image-based interactions.

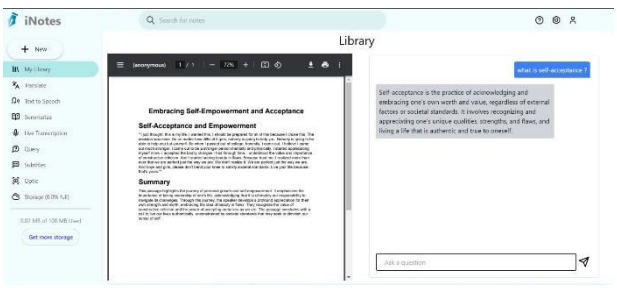
4.10 Live Transcription:

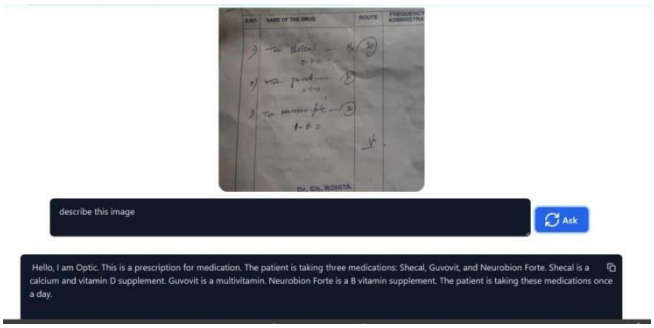
- Utilization of React Speech Recognition library for real-time transcription.
- Allows users to receive immediate transcription of spoken words during lectures or discussions.

V. RESULTS

The audio processing and noise reduction stage demonstrated significant improvements in audio quality. By applying the noiseReduce algorithm with FFT, both stationary and non-stationary noise components were effectively removed, resulting in clearer audio signals suitable for further analysis.

The transcription process yielded accurate text representations of the audio content. The Whisper Large V3 ASR model performed well in converting audio signals to text, with high accuracy and minimal errors. Timestamping of the text facilitated seamless subtitle generation, enhancing the accessibility of the content.





|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

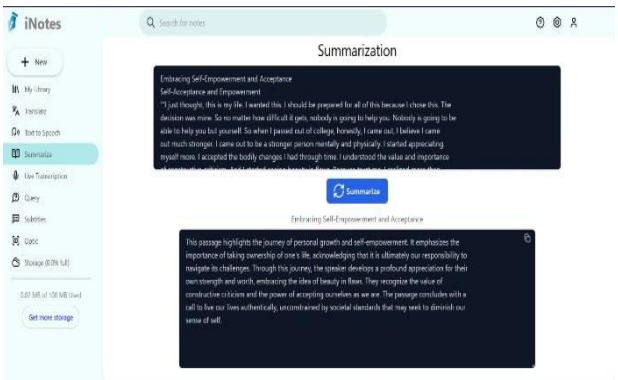
| DOI:10.15680/IJIRSET.2024.1303138 |

The querying mechanism enabled users to retrieve relevant information from the generated documents effectively. By leveraging vector embeddings and cosine similarity calculations, the system provided accurate responses to user queries, facilitating efficient knowledge retrieval.

The topic segmentation stage successfully segmented document content into distinct topics, providing users with a structured overview of the material. By prompting the Gemini-pro model with text chunks, topics and their corresponding content were accurately identified, facilitating organized navigation through the document.

The image querying functionality enabled users to interact with images and extract relevant information effectively. By leveraging the Gemini Vision Pro model, users could pose questions about images or extract text content, gaining deeper insights into visual data.

The summarization module effectively generated concise summaries of lengthy documents. The Google Generative AI model Gemini-pro produced coherent and informative summaries, capturing key insights from the input text and aiding in efficient knowledge extraction.



Overall, the results demonstrate the effectiveness and versatility of our proposed methodology in processing and

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303138 |

analyzing educational multimedia data. The integration of various components offers a comprehensive solution for knowledge extraction, dissemination, and accessibility in educational contexts.

VI. IMPACT AND FUTURE WORK

Automated Notes Maker (ANM) significantly enhances the learning experience for students in online education by revolutionizing the note-taking process. Students can efficiently review and retain information from virtual lectures without the time-consuming task of manual notetaking. This automation frees up valuable study time, promotes better comprehension, and ensures that students have access to structured and concise study materials. By providing real-time access to summarized notes, facilitating language translation, and offering advanced features such as text-to-speech conversion and image querying, the project empowers students to engage more effectively with educational content, leading to improved academic performance and overall satisfaction with the learning process.
In future iterations we plan to integrate a bookstore which provides best and comprehensive books to students and also a image generation from text which can aims to give copyright free images, make imaginations come true.

VII. CONCLUSION

The project aims to provide students with text-based PDF/Word Document of their online classes given they are audio/video based in nature. The research project being suggested attempts to cut down on the time required for manually documenting long speeches at an events and classes. In this project, all the current developments are examined. In this study various feature extraction techniques and Speech understanding methodologies that can be used to construct a voice recognition system for a multi-language are examined. information. Additionally, the system can produce text as notes in the appropriate forms, can supports multi language translation, text summarization, Querying and answering, transcription and subtitles and live transcription.

REFERENCES

[1] Manoj Kumar A1 , Janani P2 , Siva Subramanian G3 , Kumaragurubaran K4 , Sundari P, “Automated Notes Maker from Audio Recordings” ISSN 2582-7421.
[2] Chaudhari Mahimal , Mali Divya2 , Chaudhari Nehal3 , Kolhe Trupti4 , Ashish T. Bhole5, “Automated Notes Maker from Audio Recordings” ISSN (O) 2278-1021.
[3] Ms. Purva Chavrekar, 2Ms. Shruti Deshmukh, 3Ms. Pranjal Khade, 4Ms.Vaibhavi Patil 5Prof. Rupali Sathe, “AUTOMATED NOTES MAKER FROM AUDIO RECORDING” 2023 IJRTI | Volume 8, Issue 4 | ISSN: 2456-3315.
[4] Speech to text conversion and summarization for effective understanding and documentation (https:// www.researchgate.net/publication/342147736_Speech_to_text_conversion_and_summarization_for_e ffective_understanding_and_ documentation).
[5] Prerana Das, Kakali Acharjee, Pranab Das, Vijay Prasad “VOICE RECOGNITION SYSTEM: SPEECH-TOTEXT” 01 July 2016.
[6] Ms. Anuja Jadhav, Prof. Arvind Patil, Real Time Speech Text Converter for Mobile Users, National Conference on Innovative Paradigms in Engineering Technology (NCIPET-2012) Proceedings published by International Journal of Computer Applications (IJCA).
[7] Hakan Erdogan, Ruhi Sarikaya, Yuqing Gao, “Using semantic analysis to improve speech recognition performance” Computer Speech and Language, ELSEVIER 200.
[8] Chen, Jingdong, Yiteng Huang, Qi Li, and Kuldip K. Paliwal.” Recognition of noisy speech using dynamic spectral sub band centroids.” IEEE signal processing letters 11, no. 2 (2004): 258-261. [9] y Keiichi Tokuda, Yoshihiko Nankaku, Tomokin Toda, Heiga Zen, Speech Synthesis Based on Hidden Markov Models, Proceedings of the IEEE — Vol. 101, No. 5, May 2013. Junichi Yamagishi, Member IEEE, and Keiichiro Oura.
[10] F. Seide, G. Li, D. Yu, Conversational Speech Transcription Using Context-Dependent Deep Neural Networks, In Interspace, pp. 437440, 2011. 47.
[11] Muhammad Yasir, Marlince NK, Nababan, Yonata Laia, Windania Purba, Robin ,Asaziduhu Gea, “Web-Based automation speech-to -text application using audio recording for meeting speech”,2019.
[12] Shivangi Nagdewani, Ashika Jain, “A REVIEW ON METHODS FOR SPEECH-TO-TEXT AND TEXT-TOSPEECH CONVERSION”, Volume: 07 issue: 05 | May 2020.
[13] Lawrence Rabiner, Biing-Hwang Juang, B.Yegnanarayana, Fundamentals of Speech Recognition 978- 0-13-015157-5.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 3, March 2024 ||

| DOI:10.15680/IJIRSET.2024.1303138 |

[14] Tim Sainburg, Noise Reduction: A model for improving clarity and quality of the audio signals. (https://timsainburg.com/noise-reductionpython.html).
[15] Alec Radford, Jong Wook Kim, Tao Xu 1 Greg Brockman, Christine McLeavey "Robust Speech Recognition via Large-Scale Weak Supervision".
[16] Balayesu, N., Reddy, A.A. Deep pelican based synthesis model for photo-sketch face synthesis and recognition. Multimed Tools Appl (2024).
[17] Balayesu, N. Kalluri, H.K. An extensive survey on traditional and deep learning-based face sketch synthesis models. Int. j. inf. tecnol. 12, 995–1004 (2020).
[18] Balayesu, N. (2019). Optimal Pyramid Column Feature with Contrast Enhanced Model for Face Sketch Synthesis. Journal-Of-Advanced-Research-In-Dynamical-And-Control-Systems, 11(5), 335-344.