IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

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

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 3, March 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.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.

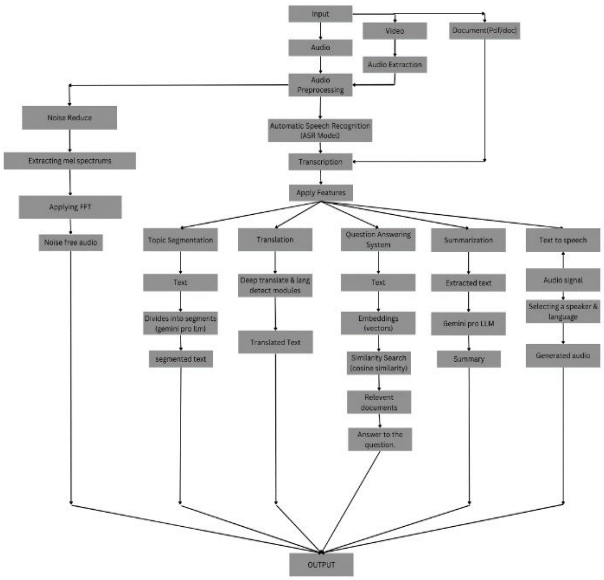
|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

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

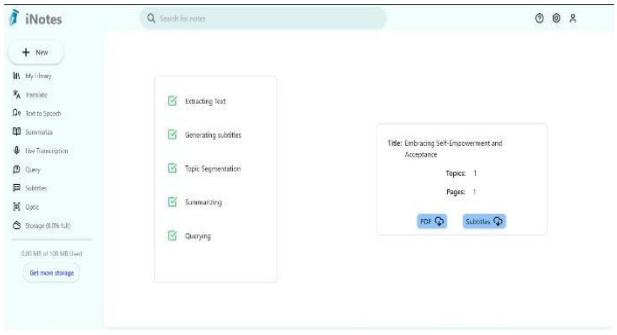
| DOI:10.15680/IJIRSET.2024.1303138 |

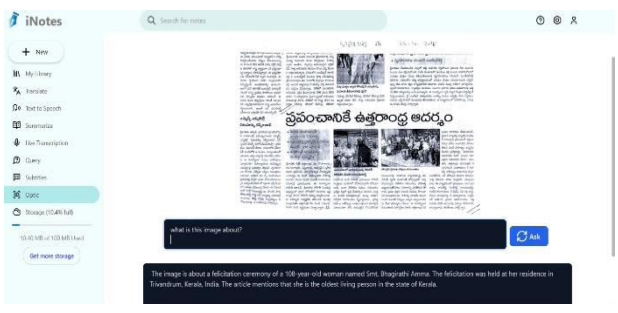
- Language translation using the deep-translator library, ensuring accessibility across 32 supported languages.
- Vector embeddings and cosine similarity calculations, providing accurate query responses and facilitating efficient knowledge retrieval.
- Text to realistic speech using TTS/XTTS_v2 model, 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.

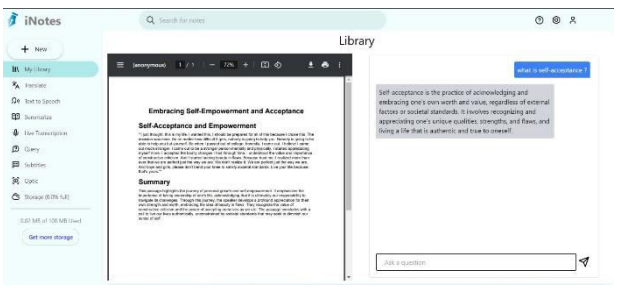
4.10 Live Transcription:

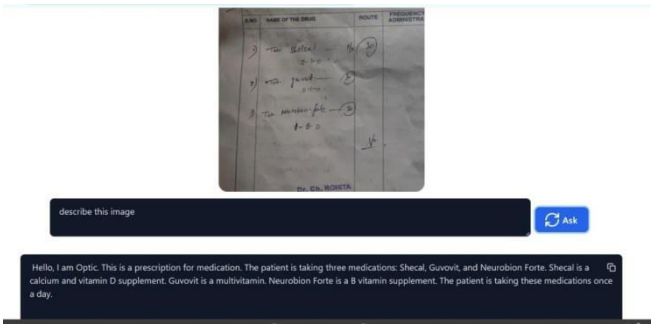
- Utilization of React Speech Recognition library for real-time transcription.
- Allows users to receive immediate transcription of spoken words during lectures or discussions.

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 ||

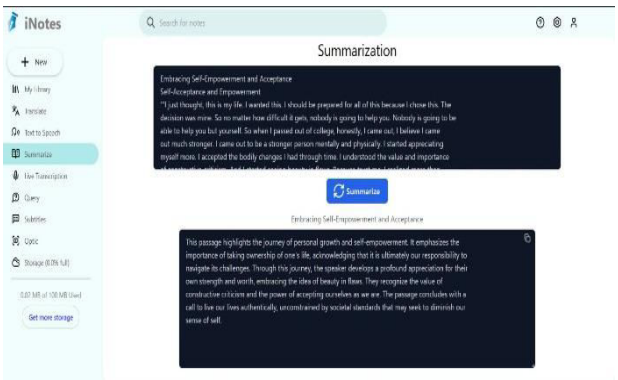
| DOI:10.15680/IJIRSET.2024.1303138 |

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.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details