VOICE-TO-TEXT SYNTHESIS: NLP-DRIVEN NOTE GENERATION FROM AUDIO RECORDINGS

Devashree Pawar
Dept. Artificial Intelligence and Data Science
Vidyavardhini's College of Engineering and Technology
Mumbai, India
devashree.203889201@vcet.edu.in

Chetan Jawale
Dept. Artificial Intelligence and Data Science
Vidyavardhini's College of Engineering and Technology
Mumbai, India
chetan.203719107@vcet.edu.in

Prasad Shah
Dept. Artificial Intelligence and Data Science
Vidyavardhini's College of Engineering and Technology
Mumbai, India
prasad.203292101@vcet.edu.in

Prof. Sejal D'mello
Dept. Artificial Intelligence and Data Science
Vidyayardhini's College of Engineering and Technology
Mumbai, India
sejal.dmello@vcet.edu.in

Abstract*—The paper introduces a novel online application called a video summarizer created with Python Flask and leverages cutting-edge AI technologies like the Gemini AI Transcription API and the YouTube API to satisfy the increasing demand for effective solutions for summarising and transcribing videos. This tool simplifies the transcribing and summarizing procedures by allowing users to save summaries in text format for later reference. Its core feature comprises getting video data from YouTube and accurately transcribing it using the Gemini AI Transcription API, followed by a summary that generates notes, providing a complete assessment in significantly less time than watching the entire video. Notably, the application has comprehensive multilingual support, allowing for translation of both text and interface into languages such as Marathi and Hindi via the Google Translation Module. This feature improves accessibility for users with different linguistic backgrounds, promoting inclusive content consumption and distribution. The study's significance stems from its offering of accessible and adaptable tools for content consumption and information distribution across linguistic boundaries. The web application enables users to efficiently transcribe and summarize video information by leveraging advanced AI technology inside an intuitive interface, resulting in easy access to useful insights. This improvement represents a significant step forward in bridging the gap between audiovisual content and textual information, meeting the changing needs of a digital society. The research proposes a ground-breaking approach to the issues of video transcription and summarization, utilizing AI technology to improve productivity and accessibility. With its user-friendly interface and broad multilingual features, the web tool is a great asset for researchers, educators, content creators, and professionals looking to extract relevant insights from video content across linguistic boundaries.

Keywords*— Python Flask, Gemini AI Transcription API, YouTube API, Video Transcription, Summarization, Multilingual Support, and Google Translation Module.

I. INTRODUCTION

Over the exceeding years of technology, we have come across various methods of interacting and manipulating videos more efficiently [11]. The internet is a vast multimedia content library that includes a wide range of videos covering different genres, subjects, and languages in today's digital world. The development of video material has profoundly changed how people consume information, providing a dynamic and engaging medium for communication, education, and entertainment. However, alongside the plethora of video content, there is a pressing need for effective tools to transcribe and synthesize this wealth of information. It is an impressive yet alarming fact that there is far more video being captured—by consumers, scientists, defense analysts, and others—than can ever be watched or browsed efficiently. For example, 144,000 hours of video are uploaded to YouTube daily; life loggers with wearable cameras amass Gigabytes of video daily; 422,000 CCTV cameras perched around London survey happenings in the city 24/7[12].

Recognizing the importance of this topic, this research paper presents a trailblazing answer in the shape of a transcription web application, painstakingly built using the Python Flask framework and enhanced with cutting-edge AI technology. At its core, this online application uses complex algorithms and APIs, such as the Gemini AI Transcription API [7] and the YouTube API[8], to automate video transcription and summarization, thus generating notes.

Online learning and E-Learning platforms have been in the headlines in recent years and have altered the way we perceive education to some extent [2]. In today's hectic digital environment, there is a greater demand than ever for prompt and accurate transcribing services. Audio-to-text

software has developed into an innovative tool that fundamentally alters how we organize and extract pertinent information from spoken speech. Decades worth of hand-engineered domain knowledge has gone into current state-of-the-art automatic speech recognition (ASR) pipelines [1]. Automatic speech recognition is also called speech recognition; it can be defined as graphical representations of frequencies emitted as a function of time[5]. By converting audio/video recordings into text using Automatic Speech Recognition (ASR) algorithms built into Gemini AI transcription API, this method provides a practical and efficient replacement for manual transcription. The main objective is to shorten a video while preserving the important and relevant information it contains [3].

These programs can do more than just transcription, though. One such advanced technique uses the NLTK (Natural Language Toolkit) toolkit along with the powers of Natural Language Processing (NLP) to advance the retrieved text. This dynamic application uses natural language processing (NLP) techniques to convert audio from video recordings to text and analyze, interpret, and summarize the content. This technology transforms nonsensical transcriptions into meaningful understandings by applying sophisticated language and semantic analysis.

This software's fusion of ASR and NLP makes it simple for professionals and organizations to convert spoken input into organized, useful information. This technology is revolutionary in the field of information management and analysis, whether it is utilized to extract crucial insights from consumer feedback or to streamline the transcription of meetings and interviews. By enabling businesses to utilize their spoken content fully, it becomes an invaluable tool for improved communication, knowledge extraction, and decision-making. The combination of ASR with NLP is an effective strategy for staying ahead in the cutthroat competitive landscape of the digital age when data-driven insights are critical.

Voice-to-text synthesis, a dynamic and expanding field at the intersection of Natural Language Processing (NLP) and audio/video processing, addresses the growing need to turn spoken words quickly and consistently into written text. In a time when textual data collection and management are crucial in many disciplines, the development of advanced systems that can transform audio recordings into coherent and contextually suitable notes offers enormous promise. Many voice recognizer components have been improved because of Deep Neural Networks (DNNs). They are commonly used in hybrid DNN-HMM speech recognition systems for acoustic modeling [4]. The limitations of traditional transcribing techniques, such as their time-consuming nature and human error risk, have spurred interest in innovative approaches that make use of NLP and ASR technologies. Voice-to-text synthesis essentially aims to close the gap between spoken language and written records. Its goal is to create systems that can precisely record spoken words while also capturing the subtleties of speech, like the tone of the speaker, pauses, and the conversation's overall context. This thorough approach is essential for creating notes that are coherent and contextually relevant in addition to being verbatim.

Additionally, the online application has a wide range of cutting-edge features created to meet the various demands and preferences of users. One such feature is the ability to save summaries in text format, which gives users a handy way to store and reference important information. The world appears to be heading toward greater reliance on technology to meet the increasing demand for translation services[13]. Due to the increased need for global communication, multilingual machine translation is the propel for researchers [6]. Therefore, to reflect the global nature of digital communication, the application has multilingual support, allowing users to translate both the transcribed text and the application interface into languages such as Marathi and Hindi. This is done by using a Google Translate Module[14] This multilingual approach not only improves accessibility for users with different linguistic backgrounds but also demonstrates the application's commitment to inclusivity and cultural diversity. By breaking down language barriers and promoting cross-border contact, the web application emerges as an effective instrument for encouraging cross-cultural understanding and collaboration in an increasingly interconnected world.

II. METHODOLOGY

Video watching is one of the most popular digital activities worldwide. A large amount of video content is captured, produced, and distributed over the Internet every minute and this number has been growing rapidly over the past few years [16]. To create a transcription web application that may leverage films viewed online to obtain the video's summarised information, this research study presents a revolutionary methodology. This methodology provides a comprehensive strategy to automate the transcription process, simplify multilingual communication, and improve accessibility in response to the growing need for effective transcription solutions.

A. Problem Statement

In an age dominated by digital multimedia material, obtaining and analyzing massive amounts of video data creates considerable hurdles. Existing transcription methods are typically time-consuming and lack multilingual support, making it difficult for users with varied linguistic backgrounds to access information. This paper tries to overcome these issues by creating a transcription web application using Python Flask and powerful AI technologies such as the Gemini AI Transcription API[7] and YouTube API[8]. The application will rapidly transcribe and summarize videos, and you may save the summaries as text. It will also provide multilingual capability, allowing translation into Marathi and Hindi using the Google Translation Module[14]. The major goal is to improve accessibility and usability, allowing users to easily navigate and extract meaningful insights from multimedia information, independent of language obstacles.

B. Implementation

Our research aims to use Natural Language Processing (NLP) to automatically translate spoken words from audio files into notes that are both contextually appropriate and fully formed. This section provides a detailed overview of our methods for achieving this goal. It meticulously details our strategy, which includes the careful selection of NLP models and cutting-edge technology. These methods and approaches are carefully selected to enable the system to transcribe a wide range of spoken languages and understand the nuances inherent in accents, colloquialisms, and specialist vocabulary important to specific disciplines. This paper presents an application that simplifies the process of transcribing and summarizing audio from video information by utilizing the capabilities of cutting-edge AI technologies, such as the YouTube API and the Gemini AI Transcription API. The process consists of several interrelated processes, starting with setting up the development environment and integrating necessary dependencies. The program then uses the YouTube API to retrieve video content, the Gemini AI Transcription API to start transcription and options for text summarizing and storage. The program also has multilingual capabilities, allowing for translation using the Google Translation Module into languages like Marathi and Hindi. Every part of the process is carefully thought out to guarantee effectiveness, scalability, and dependability. To confirm functionality and improve user experience, rigorous testing and continuous integration techniques are used. The transcribing web application has the potential to completely transform digital information accessibility and comprehension through this all-encompassing approach.

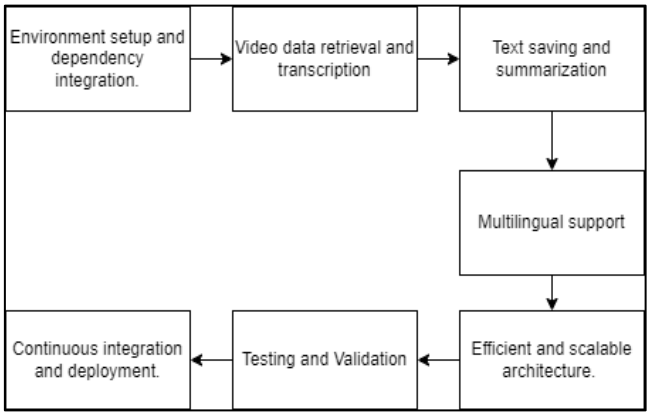

Fig 1. Proposed System

As per the figure shown above let us go over the steps one by one.

1. Environment Setup and Dependency Integration: Setting up the development environment for the transcribing web application begins with establishing the Python Flask framework, which is lightweight and versatile for web programming. Flask provides the flexibility required to incorporate external APIs effortlessly. The essential dependencies are included alongside Flask, including the requests library, which allows for easier communication with external APIs. This library enables the application to send queries to the Gemini AI Transcription API and the YouTube API, retrieve data, and handle responses efficiently. The Gemini AI Transcription API and YouTube API have been registered and incorporated into the program, allowing for core transcription and video data retrieval functions.
2. Video Data Retrieval and Transcription: When a user requests a video transcription, the program uses the YouTube API to collect the video's information. This metadata contains information about the video's title, description, and other important facts. The obtained data is then used to start the transcription process using the Gemini AI Transcription API. Gemini AI is multimodal, which means it takes audio, video, or text as input. The API translates the video's audio content into text format, using modern algorithms and speech recognition technology, such as ASR to assure accuracy and efficiency. The application then retrieves and processes the transcribed text in preparation for additional features such as note generation and translation. Gemini AI transcription API uses certain inbuilt algorithms such as Latent Semantic Analysis[10] or Transformer models such as Bert Model[9]
3. Text Saving and Summarization: The application provides the ability to store the transcription as text to improve user experience and facilitate easy access to transcribed content. This tool allows users to save transcribed text for future reference, making it easier to retrieve and utilize vital information. Additionally, the application has features for summarizing transcribed text. The summarization method reduces the lengthy transcription to a succinct summary, emphasizing key themes and important information for easy comprehension and reference which are provided as notes. The application meets the different needs and interests of users while dealing with video content by offering transcription and summary features.
4. Multilingual Support: The transcribing web application has multilingual support since it understands how important it is to serve consumers with a variety of linguistic backgrounds. With the use of the Google Translation Module, users can translate the summaries and the transcribed material into languages like Hindi and Marathi. Text can be translated between languages with ease and preservation of context and meaning thanks to the integration with the Google Translation Module. Users may efficiently access and grasp content in their choice language by interacting with the Google

Translate API, which guarantees correctness and dependability in the translation process. This Google Translation Module API works on a powerful model known as Seq2Seq (Sequence-to-Sequence) which translates one sequence to another sequence. In seq2seq models, the input is a sequence of certain data units and the output is also a sequence of data units[15]. Consequently, translating the generated notes into languages like Marathi and Hindi is made easier by using the Google Translate Module API.

5. Efficient and Scalable Architecture: The transcribing web application's architecture places a high priority on efficiency and scalability, guaranteeing reliable operation even with a large volume of users. Sturdy error-handling techniques are used to handle exceptions and unexpected behaviors with grace, improving the stability and dependability of the application. The program meets user needs for video transcription and summary services by using effective coding techniques and maximizing resource utilization to give responsive and consistent performance.
6. Testing and Validation: Throughout the development process, rigorous testing procedures are used to find and fix potential problems or flaws. Extensive testing scenarios are carried out to verify the application's operation and performance on various platforms and environments. This includes end-to-end testing to confirm the application's functionality, integration testing to analyze interactions between modules, and unit testing to examine individual components. The application is iteratively improved and refined based on user feedback and testing results, guaranteeing its dependability and efficacy in real-world usage circumstances.
7. Continuous Integration and Deployment: By streamlining the development lifecycle, continuous integration and deployment techniques allow for the smooth rollout of regular updates and improvements to users. Changes to the application code are methodically evaluated and distributed through automated testing and deployment pipelines, reducing downtime, and guaranteeing a seamless transition for customers. Because of its iterative development process, the program can adapt quickly to user feedback and new needs, staying competitive and relevant in the quickly changing fields of accessibility and digital content management.

III. RESULTS AND DISCUSSIONS

The voice-to-text synthesis application, built with Python Flask, the YouTube API, the Gemini AI Transcription API, and the Google Translation Module, has enhanced accessibility and usability for users who want to translate, summarise, and transcribe video footage. The results are encouraging. Users may now obtain concise summaries of films for improved comprehension and reference owing to the integration of the YouTube API and the Gemini AI Transcription API, which allows for accurate and efficient transcription of video material. Furthermore, features such as text preservation and multilingual support have increased user comfort by making it easy for people from diverse linguistic backgrounds to access and consume transcribed content. Positive comments have been received regarding the user-friendly interface design. Users find it straightforward and intuitive, and excellent error-handling procedures provide dependable performance even under high user traffic.

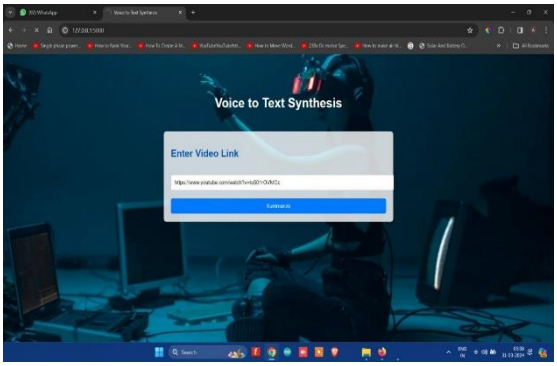

Fig 2. Front web page

Figure 2 depicts the initial page of our video summarizing the website, which acts as a gateway to the platform's functions. Users are presented with a streamlined interface that allows for seamless interaction. At the top of this page, viewers are requested to enter a YouTube video link, which initiates a procedure that makes use of the YouTube API. Following the submission of the video link, the YouTube API is smoothly invoked, orchestrating the collection of critical metadata connected with the video content. The recovered information includes key aspects such as the video title and description. These vital components serve as basic information, allowing for a thorough understanding of the video's content. Our platform provides Gemini AI with a comprehensive understanding of the video at hand by utilizing the YouTube API to obtain necessary metadata. This basic data is the cornerstone of our video summarizing technique, allowing Gemini AI to successfully traverse and distill video information. Users can start on a journey of effective video summarizing using this straightforward interface and integrated API capabilities, which are supported by a thorough grasp of the content's context and relevance.

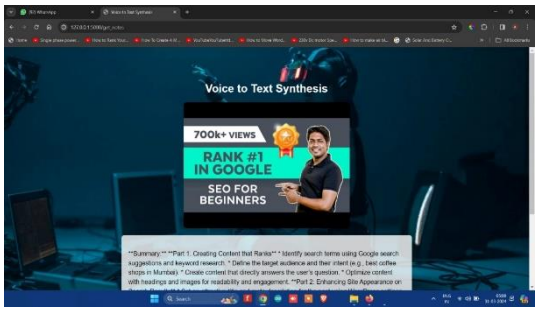

Fig 3. Notes generated through YouTube Video.

Figure 3 displays a user interface designed for efficient dissemination and storage of summary notes. The web page offers consumers a streamlined experience with simple navigation options and intuitive design features. Users may easily access and analyze summarised content, with the opportunity to save notes directly to their system. This feature promotes accessibility and convenience by allowing users to save vital information for later use. The design is user-centered, emphasizing clarity and ease of use to improve the note-taking experience. By providing a comprehensive solution for note transmission and storage, the website contributes to increased productivity and efficiency in information management tasks. The image shows a user-friendly interface that supports effective information acquisition and retention via streamlined note-taking features.

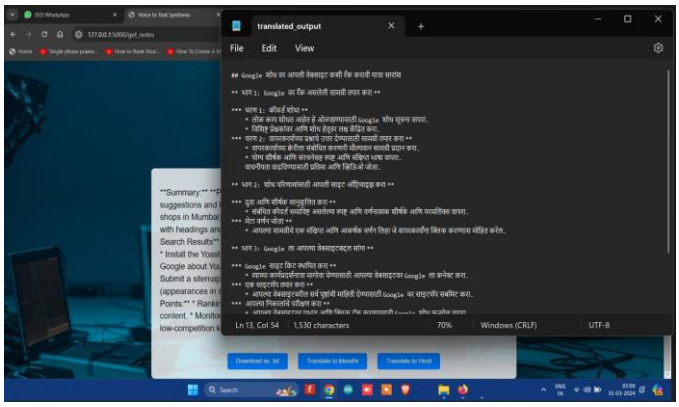

Fig 4. Translation of the generated notes.

Figure 4 depicts a critical feature of our research endeavor, the supply of created notes in both Hindi and Marathi, with seamless integration for effective saving within the system. This multilingual capacity is provided by incorporating the Google Translation Module API into our program, which allows for smooth and accurate translation procedures. By leveraging the Google Translation Module API, our system ensures that users may access the summarized notes in their choice language, increasing accessibility and inclusion. The API enables real-time translation, allowing for the speedy and accurate conversion of text from one language to another. This integration not only broadens the scope of our system but also demonstrates our commitment to tackling linguistic diversity and encouraging equitable access to information. Furthermore, by allowing users to save translated notes immediately within the system, we simplify the process of knowledge retention and allow for seamless integration into users' workflows.

IV. CONCLUSION

In conclusion, voice-to-text synthesis is a web application built with Python Flask, the Gemini AI Transcription API, and the YouTube API that has shown great promise in improving access to video material. It is a web-based transcription and summary program. By seamlessly merging transcription and summarizing functions, as well as the ability to save summaries as text, the program satisfies the growing demand for efficient content consumption and information retrieval. The key findings of this research indicate the effectiveness of NLP in overcoming difficulties like as background noise and accent changes, resulting in improved transcription accuracy. Furthermore, text summary techniques are used to reduce transcriptions into succinct summaries while keeping vital information, hence improving understanding and information retrieval.

Furthermore, the Google Translation Module allows users to translate material into Marathi and Hindi, promoting greater accessibility and inclusivity across linguistic boundaries. This multilingual strategy not only serves a varied user base, but it also helps to democratize knowledge distribution.

There are several future research scopes of this web application, which include firstly, there is a need for improvement in the accuracy of translations supplied by the Google Translation Module for Marathi and Hindi. This could include looking into advanced language models or fine-tuning parameters to improve the quality of multilingual summaries. Also incorporating additional APIs such as Google Cloud speech-to-text and IBM Watson speech-to-text APIs, as well as Gemini AI for speech-to-text transcription and text summarizing, might considerably expand the web app's functionality. Researchers could investigate new methods of properly processing and presenting information by incorporating more modern tools.

REFERENCES

[1]. Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin", in *ICML'16: Proceedings of the 33rd International Conference on International Conference on Machine Learning*, arXiv:1512.02595v1 [cs.CL] 8 Dec 2015 (2015).

[2]. Chengpei Xu, Ruomei Wang, Shujin Lin, Xiaonan Luo, Baoquan Zhao, Lijie Shao, Mengqiu Hu, "Lecture2Note: Automatic Generation of Lecture Notes from Slide-Based Education Videos", in *IEEE International Conference on Multimedia and Expo*, ICME.2019,(2019).

[3]. Ke Zhang, Kristen Grauman, and Fei Sha, "Retrospective encoders for video summarization," in *ECCV - 15th European Conference on Computer Vision. 2018*, vol. 11212, pp. 391–408, Springer (2018).

[4]. Navdeep Jaitly, Quoc V. Le, Oriol Vinyals, "Listen, Attend and Spell", in *Interspeech 2015 conference*, arXiv:1508.01211v2 [cs.CL] 20 Aug 2015 (2015).

[5]. Saliha Benkerzaz, Youssef Elmir, Abdeslam Dennai, "A Study on Automatic Speech Recognition", in

Journal of Information Technology Review, Volume 10, 3 August 2019 (2019).

[6]. Madhura Mandar Phadke, Dr. Satish R. Devane, " Multilingual Machine Translation : An Analytical Study", in *International Conference on Intelligent Computing and Control Systems*, ICICCS 2017, pp. 881 – 884, IEEE 2017.

[7]. Gemini Team Google: Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai et al., "Gemini: A Family of Highly Capable Multimodal Models", in *Google DeepMind*, arXiv:2312.11805v1 [cs.CL] 19 Dec 2023 (2023).

[8]. Joseph Kready, Shishila Awung Shimray, Muhammad Nihal Hussain, Nitin Agarwal, "YouTube Data Collection Using Parallel Processing", in *IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pp. 1119-1122, IEEE 2020 (2020).

[9]. Koroteev M.V., "BERT: A Review of Applications in Natural Language Processing and Understanding" Retrieved 22 March 2021 DOI: 10.48550/arXiv.2103.11943.

[10]. Peter W. Foltz, "Latent Semantic Analysis for text-based research", in *Behavior Research Methods, Instruments and Computers,* pp. 197 – 202, February 1996.

[11]. Smita Jawale, Ankit D. Singh, Pritesh C. Mane, Abhishek V. Thakur, "Automatic Subtitle Generation in Real-Time", in *International Journal of Computer science engineering Techniques*, pp. 14 – 18, Volume 1 Issue 2, ISSN: 2455-135X, 2016.

[12]. Boqing Gong, Wei-Lun Chao, Kristen Grauman, Fei Sha, "Diverse sequential subset selection for supervised video summarization", in *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, Volume 2, pp. 2069–2077, December 2014 (2014).

[13]. Reem Alsalem, "The Effects of the Use of Google Translate on Translation Students' Learning Outcomes", in *AWEJ for Translation & Literary Studies*, Volume3, Number4, pp. 46-60, DOI: http://dx.doi.org/10.24093/awejtls/vol3no4.5.

[14]. Saliha Benkerzaz, Youssef Elmir, Abdeslem Dennai, "A Study on Automatic Speech Recognition", in *Journal of Information Technology Review*, Volume 10, Number 3, pp. 77-85, August 2019.

[15]. Yaser Keneshloo, Tian Shi, Naren Ramakrishnan, Chandan K. Reddy, "Deep Reinforcement Learning for Sequence-to-Sequence Models", in *IEEE Transactions on Neural Networks and Learning Systems*, Volume: 31, Issue: 7, pp. 2469 - 2489, July 2020.

[16]. Hongxiang Gu, Stefano Petrangeli, Viswanathan Swaminathan, "SumBot: Summarize Videos Like a Human", in *IEEE International Symposium on Multimedia (ISM)*, pp. 210-217, 22 January 2021 (2021).