iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** III **Month of publication:** March 2025

DOI: https://doi.org/10.22214/ijraset.2025.68048

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

Smart Speech to Text Accurate Transcription with ASR and NLP

Mr.K.Pranay Kumar Reddy[1], Mr.R.Likith Kumar[2], Mr.P.Bharath Kumar Reddy[3], Mr.R.Mohith Santhosh[4],
Dr.R.Karunia Krishnapriya[5], Mr. Pandreti Praveen[6], Mr.V Shaik Mohammad Shahil[7], Mr.N.Vijaya Kumar[8]
[1, 2, 3, 4]UGScholar,[5]Associate Professor, [6, 7, 8]Assistant Professor, Sreenivasa Institute of Technology and Management Studies, Chittoor,India.

***Abstract:** This paper investigates the incorporation of sophisticated natural language processing (NLP) methods for automatic summarization and speech-to-text conversion. It investigates the google's speech-to-text API for high- accuracy transcription with Facebook's BART model for succinct text summarizing. The study covers a number of use cases, including as automated meeting documentation, content structure, search ability, and podcast and video transcription. It also tackes important issues like minimizing prejudice and preserving contextual correctness. In order to improve text processing, the research also examines deep learning techniques including generative adversarial networks (GAN) and LSTM networks. This work attempts to improve audio data processing and enable efficient information retrieval from spoken content by integrating state-of-the-technology.*

***Keywords:** Natural language processing (NLP), speaker diarization, speech recognition, LSTM networks,BART Model Transformer-based model.*

I. INTRODUCTION

When there is a large amount of audio content accessible, it could be challenging to efficiently useful information. This project deftly blends Google's speech-text-Text API with Facebook's BART (Bidirectional and Auto Regressive Transformer) model, two powerful natural language processing tools, get around this challenge. The objective is to enable precise audio-to-text conversion and then automatically provide concise lucid written summaries.the growing volume of audio material,such as podcasts customer service calls,corporatemeetings,and instructional lectures,necessitates the use of advanced technology for effective content processing. The initiative acknowledges that accurate transcription and and effective summarization are necessary to extract meaningful information from spoken speech.

The growing volume of audio material, such as podcasts, customer service calls,corporatemeetings,and instructional lectures, necessitates the use of advanced technology for effective content processing. The poll acknowledges that accurate transcription and effective summarization are necessary to extract meaningful information from spoken.

The integration of state-of-art technologies from Google and Facebook demonstrates the commitment to applying the latest advancements in natural language processing. This program has a wide range of users, from improving business analytic and educational resource to increasing the accessibility and search ability of content.the next sections will go into greater detail about the approach, technical details,and potential application of this integrated system, with focus on how revolutionary it may be in terms of how we evaluate and infer information from audio.

II. LITERRATURE REVIEW

The need to extract valuable insights from spoken discourse has increased demand for effective audio processing solutions in the age of ever expanding digital material. In order to achieve smooth transcription and subsequent summarization of audio information, this study proposes an integrated approach that sits at the nexus of sophisticated natural language processing (NLP) techniques and audio-to-text conversion. This technique is a new method for turning spoken speech into succinct and logical textual summaries by utilizing Facebook's Bidirectional and Auto-Regressive Transformers(BART) model and Google's speech-to-Text API.

Recurrent neural networks (RNN's) and transformer-based architectures are among the models included in the survey,which also discusses their uses in summarization, machine translation,conversation systems, and creative writing. In addition to addressing issues like context preservation and bias prevention, it investigates assessment approaches like perplexity and BLEU score.

The survey provides insightful viewpoints on the present and potential future paths of text production in deep learning through case studies and experimental findings. The survey tackles issues tackles issues like context preservation and bias avoidance while assessing these models using measures like perplexity and BLEU score.

The survey offers useful viewpoints on the present situation and potential future paths of text production in the field of deep learning through perceptive case studies and experimental findings.

A. Working Principle

Multiple deep learning text generation models, including generative adversarial networks (GANS), recurrent neural networks (RNNs), and transformer-based models like BERT and GPT, are incorporated into the working concept. These models use deep learning techniques to generate text sequences from input data. The performance of these models may be assessed using evaluation metrics such as perplexity and BLEU score. Additionally, the practical applications of these technologies, such as machine translation, conversation systems, and summarization, may be explored .

The survey [2] thoroughly examines the use of GANs in textual content production, emphasizing its potential in a number of fields, including conversation systems, creative writing, and natural language generation. The survey clarifies the benefits and difficulties of this method by examining the fundamental ideas and architectures of GANs in text production. In order to evaluate the quality and coherence of generated text, evaluation techniques unique to GAN-based text creation are covered, along with relevant performance indicators. Additionally, the study highlights the importance of GANs in developing the area of natural language processing by discussing noteworthy developments, new trends, and potential avenues for future research in using them for text production.The survey carefully investigates the use of GANs to a variety of text creation problems, such as conversation production, language modeling, and tale development. The survey clarifies the potential and difficulties associated with this method by breaking down the design and training processes of GANS in text production

B. Working Principle

As part of the working idea, text production is done using generative adversarial networks (GANS). In order to generate realistic text sequences, deep learning models called GANs are composed of a discriminator and a generator that are trained against one another. A variety of GAN-based text production structures and training techniques may be included in the survey, along with evaluation tools to determine the quality of the generated text. It is likely that the practical applications of GANS in text creation tasks, such as conversation systems, language modeling, and creative writing, will be examined.

The capacity of LSTM networks to capture sequential relationships in text input is highlighted in the study [3], which explores the construction and operation of these networks. The effectiveness of LSTM networks in producing logical and contextually appropriate textual material is demonstrated by the author through an analysis of context-based text creation approaches. The importance of contextual information in tasks like machine translation, conversation systems, and summarization is highlighted in the research, which explores many methods for integrating it into LSTM-based text production models. Additionally, the author provides performance evaluations and experimental data to show how well LSTM networks operate in context-based text production, opening the door for further developments in artificial intelligence and natural language processing research.The efficiency of LSTM networks in producing logical and contextually appropriate textual output is demonstrated by the author through a thorough analysis of context-based text creation approaches. The paper explores methods for incorporating contextual signals into LSTM-based models, demonstrating how they may be used in a variety of fields, including summarization, conversation production, and machine translation.

C. Working Princple

The journal explores context-based text creation using Long Short-Term Memory (LSTM) networks. For applications like language modeling and text synthesis, recurrent neural networks (RNNs) with long-range dependency capture capabilities—such LSTM networks—are ideal. The article may offer techniques for incorporating contextual information into LSTM-based text generation models in addition to training protocols and assessment metrics to determine the quality of generated text. The potential applications of LSTM networks in the actual world for machine translation, conversation systems, and summarization may also be studied.

An overview of the advancements in automatic text summarizing is given in the survey [4], which also describes the fundamental procedures and techniques used to produce succinct summaries from massive amounts of text. Mridha and Lima discuss the advantages, disadvantages, and uses of many extraction-based, abstraction-based, and hybrid techniques to automatic text summarization. The poll also addresses the difficulties that come with automated text summarizing, including conserving important information, managing domain-specific material, and ensuring coherence. The authors give scholars and practitioners in the field of natural language processing useful insights into the present and future directions of automatic text summarization by emphasizing new developments and summarizing research findings.Additionally, the study explores the inherent difficulties of artificial text summarization, including problems with scalability, maintaining coherence, and content selection.

As a thorough resource for natural language processing scholars and practitioners, the authors provide insightful information on the developments and difficulties in automatic text summarization by combining previous studies and suggesting future study avenues.

D. Working Principle

The article provides a broad overview of artificial text summarizing techniques, including advancements, fundamental principles, and challenges faced. It could discuss several methods of automatic text summarization, such as hybrid, abstraction-based, and extraction-based approaches, and how to implement them using technologies like natural language processing (NLP), machine learning algorithms, and deep learning models. The research may also address the criteria and evaluation techniques used to assess the quality of summaries generated by different methodologies. It could also discuss real-world applications and advancements in automated text summarization, including details on possible directions for further research in the area.

The TAVT framework is suggested as a way to support transfer learning in the research [5], which focuses on bridging the gap between audio and visual modalities in text creation tasks. The authors show how well the TAVT framework works to produce comprehensible written output from audio-visual inputs through careful testing and analysis. The framework shows encouraging results in tasks including multi-modal translation, picture captioning, and speech recognition by utilizing transfer learning approaches. The study also highlights the importance of transferable audio-visual text production in promoting multi-modal comprehension and communication, as well as possible uses and future research avenues in this area.

E. Working Principle

This paper most likely examines the development of the TAVT (Transferable Audio-Visual Text Generation) framework, which attempts to generate textual output from audio-visual inputs. Through the application of transfer learning strategies, it may address the integration of visual and auditory modalities in text production tasks to facilitate knowledge transfer between domains. The article may discuss the architecture, training methodologies, and applications of the TAVT framework in speech recognition, picture captioning, and multi-modal translation. The significance of transferable audio-visual text production in improving multimodal understanding and communication may also be discussed, along with performance metrics and potential applications**.**

In an effort to improve accessibility and user experience, the study [6] presents a novel approach to text summarization using voice input. The authors show how well their method works to produce succinct summaries from spoken input through careful testing and analysis. The suggested approach has potential for uses like automatic audio content transcription and summarization, which would enable effective information extraction from spoken dialog. The study advances natural language processing methods by bridging the gap between text summarizing and voice-based input, providing insightful information on the possibilities of voice-based text summarization systems." The efficiency of the suggested approach in producing succinct and logical summaries from spoken input is shown through thorough testing and analysis. The study outlines the possible uses of voice-based text summary, such as automated transcription, audio recording content summarization, and accessible services for people with visual impairments. Limitations. The study advances natural language processing technology by combining text summarizing methods with voice-based interaction, opening up new possibilities for effective information retrieval from spoken content.

F. Working Principle

The journal most likely presents a unique approach to voice-based text summarization, with an emphasis on extracting concise summaries from spoken input. It may cover techniques for converting spoken input into text, such speech recognition algorithms or voice-to-text APIs. The report may also cover the use of NLP approaches, such as extractive or abstractive summarization, to summarize the transcribed text. With potential applications in automated transcription and audio material summarizing, the method's capacity to generate coherent and significant summaries from spoken input may be evaluated.

III. COMPARATIVE STUDY

Table1ComparisonofvariousAlgorithms

Text creation developed from rule-based systems (2016), which were interpretable but lacked scalability, to Hidden Markov Models (2017) for speech recognition. While multi-model integration (2019) integrated audio-visual data for richer text production, RNNs (2018) enhanced context modeling but had training issues. While LSTMs (2021) enhanced sequential modeling, GANs (2020) produced realistic text but struggled with instability. While Google's Speech-to-Text API (2023) achieved great transcription accuracy but necessitated internet connectivity, Facebook's BART (2022) improved summarization using

Year	Algorithm	Key Developments	Pros	Cons
2016	Rule-based Systems [4]	Integration with linguistic rules and heuristics	- Transparent and interpretable text generation	-Limited Scalability, requires manual rule crafting
2017	Hidden Markov Models	Application in speech recognition and text generation	- Widely used for text modelling and generation	-Limited capability to captu re complex dependencies
2018	Recurrent Neural Networks [1]	Mechanisms for improved context modelling	-Handles sequential data effectively	- Gradient vanishing, Slow training convergence
2019	Multi-Model Integration [5]	Integration of audio-visual data for text generation	-Enhances text generation with audio- visual cues	-Increased complexity, potential data synchronization issues
2020	Generative Adversarial Networks [2]	Enhanced training techniques for stability	-Generates diverse and realistic text samples	-Training instability, Mode collapse
2021	LSTM Networks [3]	Improved architectures for sequential modelling	-captures sequential dependencies in text data	- Limited holding of long term dependencies
2022	Facebook's BART Model [6]	Transformers-based architecture for summarization	- Coherent and concise text summarization	-Training complexity,Resource- intensive computation
2023	Google's Speech-to-Text API	Advanced Deep learning models for transcription	- High accuracy in audio transcription	-Limited customization options,requires internet connection
2024	Bart	Enhanced fine-tuning for Diverse NLP tasks,improved Text generation efficiency.	Strong performance in summarization on translation and text competion.	Computationally expensive,require Large datasets for fine-tuning

IV. PROPOSED SYSTEM

This project proposes a smart, AI-powered transcription system that combines Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) for accurate, real-time speech-to-text conversion, overcoming the limitations of existing solutions. The system is designed to provide high accuracy, multilingual support, real-time processing, and enhanced privacy while ensuring cost-effectiveness and user accessibility. One of the system's key features is its integrated functionality, which combines speech transcription, summarization, punctuation correction, and speaker diarization into a single platform. This enables users to create concise summaries in addition to transcribing speech, making it especially useful for meetings, lectures, interviews, and voice-driven applications. In contrast to traditional cloud-based solutions, this system is made to operate both online and offline.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

V. SYSTEM ARCHITECTURE

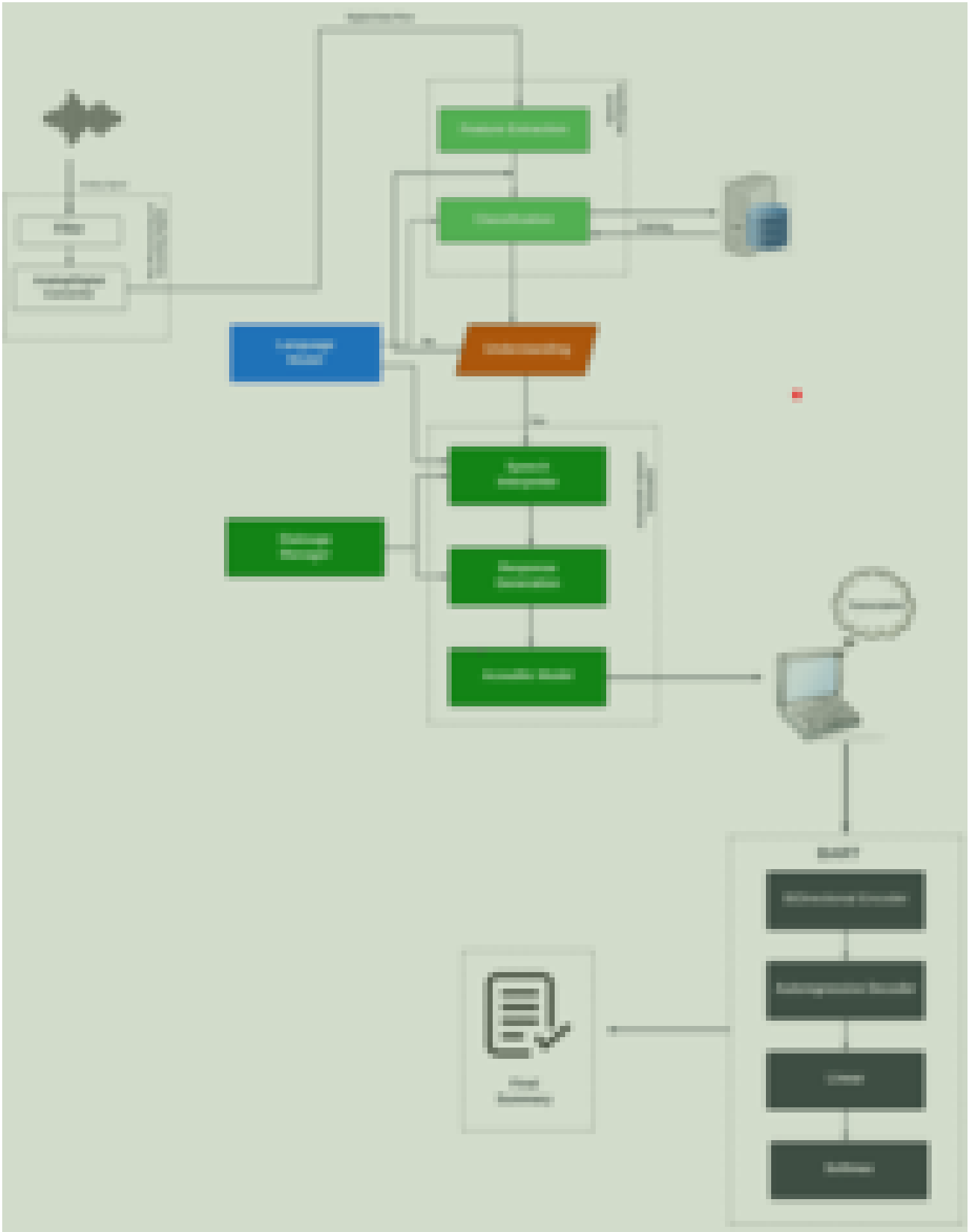


Figure1Extractivealgorithmonabstractiveinputand Abstractivealgorithmonextractiveinput

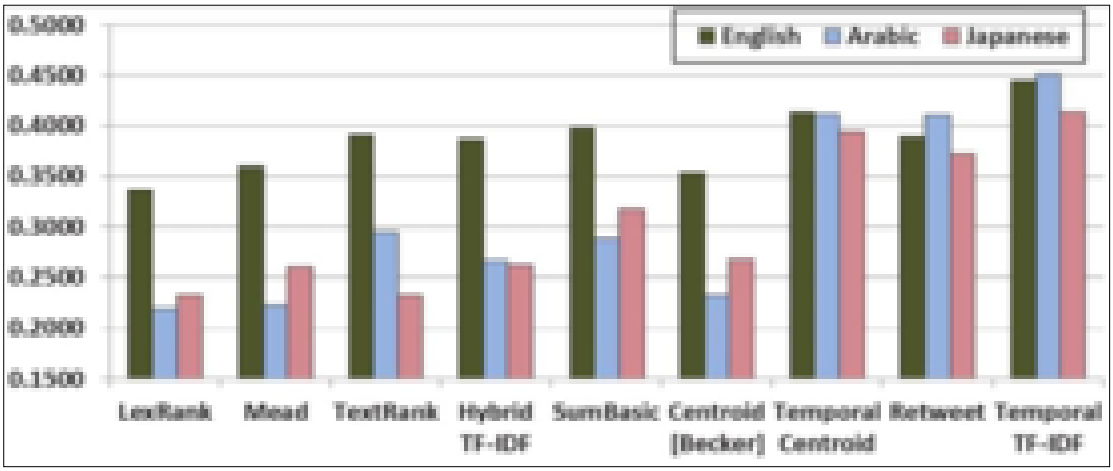


Figure2Variousmethodsondifferentlanguages

VI. CONCLUSION

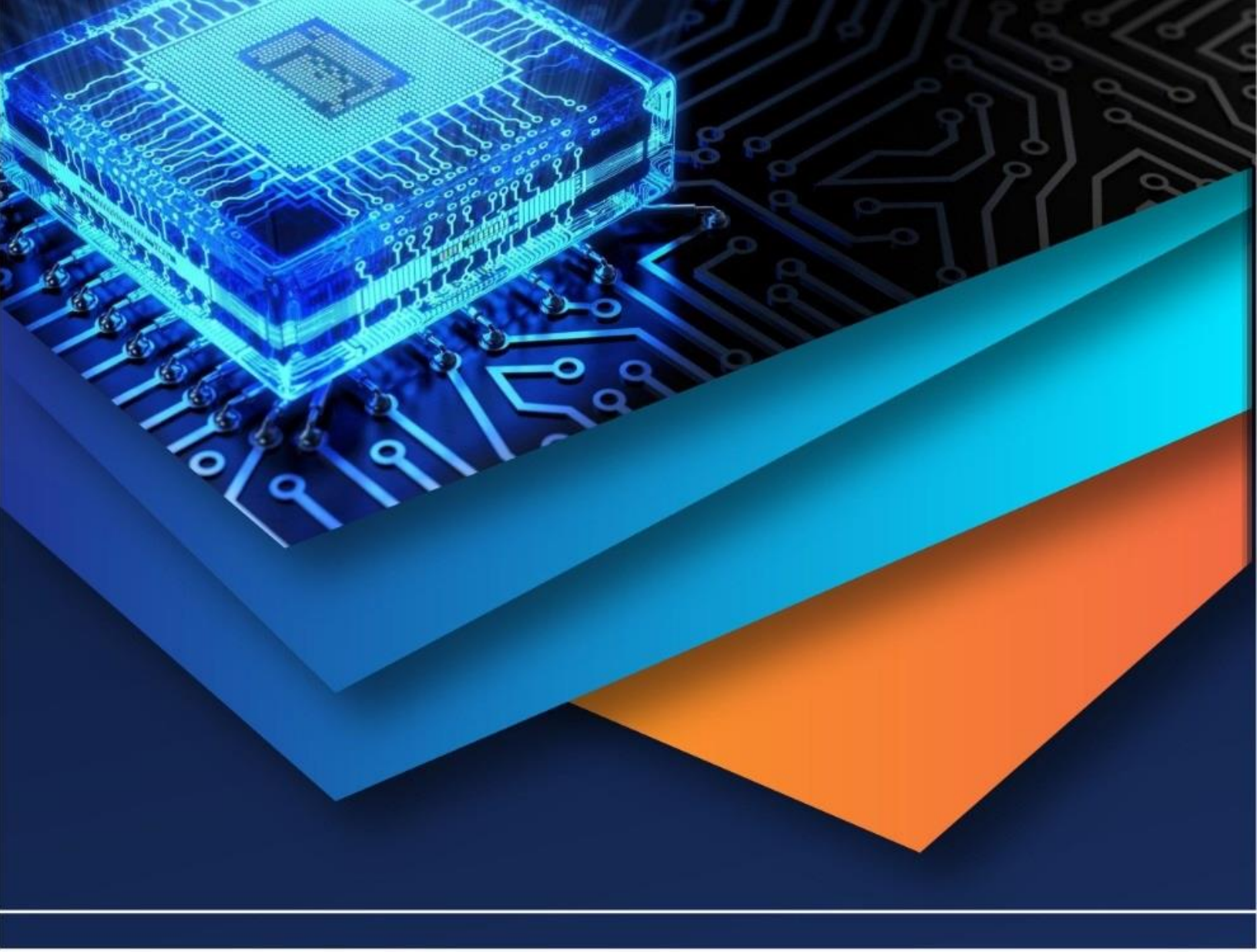
We examined many journal publications concerning audio to text creation and text summarization. We understood the advantages and disadvantages of using various strategies. Given the limitations, we will try to overcome them by utilizing transformers in our proposed methods. Our approach combines state-of-the-art technologies like Google's Speech-to-Text API and Facebook's BART Model to create a seamless solution to manage large volumes of audio data processing. This journal paper provides a comprehensive analysis of methods, highlighting both their benefits and drawbacks, making it a priceless resource for researchers and practitioners in natural language processing. This project will be used in journalism, education, podcasting, and business meetings.

VII.ACKNOWLEDGWMENTS

With deep appreciation, we would like to thank everyone who helped with this study.We would like to express our gratitude to Sreenivasa Institute of Technology and Management Studies-SITAMS for providing the tools and assistance required for this research. We would especially like to thank Dr. R. Karunia Krishnapriya, Mr. Pandreti Praveen, Mr.V Shaik Mohammad Shahil, Mr. Vijay Kumar for their significant advice and knowledge in the areas of machine learning and hepatocellular carcinoma. Their observations greatly improved the caliber of our work.

REFERENCES

[1] TouseefIqbal,ShaimaQureshi.TheSurvey:Textgenerationmodelsindeeplearning.JournalofKingSaud University – Computer and Information Sciences 34 (2022) 2515–2528.
[2] Gustavo H. de Rosa , Joao P. Papa: A survey on text generation using generative adversarial networks. Pattern Recognition 119 (2021) 108098.
[3] SivasuryaSanthanam:ContextbasedText-GenerationUsingLSTMNetworks,InstituteforSoftwareTechnology ,GermanAerospaceCenter(DLR),2020.
[4] M.F.Mridha:ASurveyofAutomaticTextSummarization:Progress,ProcessandChallenges,2021.
[5] Wang Lin,TaoJin,YeWang:TAVT:TowardsTransferableAudio-VisualText Generation,Zhejiang University,2022.
[6] PratimaMohanThorat,Prof.Dr.M.S.Bewoor:ANovelApproachforVoiceBasedTextSummarizer,2022
[7] M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/ *FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.
[8] PDCA12-70 data sheet," Opto Speed SA, Mezzovico, Switzerland.
[9] A. Karnik, "Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP," M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)