AUDIOINSIGHT PERFORMANCE ANALYSIS: EXPLORING SPEECH AND TEXT TECHNOLOGIES

A.N.Adapanawar, Tanvi Gaikwad, Sharva Khandagale, Subrat Dhapola, Mayank Wakdikar

[1] *Professor, Department of Computer Engineering, Sinhgad Academy of Engineering, Maharashtra, Pune*
[2] *Student, Department of Computer Engineering, Sinhgad Academy of Engineering, Maharashtra, Pune*
[3] *Student, Department of Computer Engineering, Sinhgad Academy of Engineering, Maharashtra, Pune*
[4] *Student, Department of Computer Engineering, Sinhgad Academy of Engineering, Maharashtra, Pune*
[5] *Student, Department of Computer Engineering, Sinhgad Academy of Engineering, Maharashtra, Pune*

ABSTRACT

In a time when digital content is growing at an exponential rate, organizing and understanding large amounts of audio data effectively presents daunting obstacles. This study recognizes the need for novel approaches in this field and presents "AudioInsight," a sophisticated summarization system that aims to transform the way audio content is handled by strategically integrating cutting-edge machine learning and natural language processing (NLP) techniques. At its center, AudioInsight speaks to a worldview move in sound information handling, offering users a comprehensive toolkit to explore the complexities of advanced substance. Its essential work revolves around the consistent transformation of talked substance into brief literary outlines and vice versa, leveraging state-of-the-art NLP methods to distill key experiences from audio sources. Additionally, the framework brags vigorous linguistic use adjustment capabilities, guaranteeing that the passed-on data is not only concise but also syntactically accurate—a significant perspective in encouraging successful communication and comprehension. Past its summarization and linguistic use rectification functionalities, AudioInsight addresses a bunch of subordinate challenges predominant in audio information handling. These incorporate relieving issues such as inaccurate word tally and compatibility disparities, subsequently enhancing the overall accuracy and reliability of the summarized content. Moreover, a notable enhancement lies within the integration of Optical Character Recognition (OCR) innovation, enabling clients to consistently handle and translate different sorts of information past conventional audio formats. This expansion essentially extends the system's utility and pertinence, situating AudioInsight as a flexible arrangement able of dealing with a diverse range of content with ease and accuracy. In pith, AudioInsight stands as a confirmation to the meeting of cutting-edge innovations and user-centric plan standards, culminating in a comprehensive arrangement custom fitted to streamline the preparing and comprehension of audio data within the computerized scene. Its natural interface and strong highlight set offer clients unparalleled adaptability and proficiency in analyzing and controlling audio content. Whether utilized by people looking for to improve individual efficiency or organizations endeavoring to open bits of knowledge from tremendous stores of sound information, AudioInsight emerges as a trusted partner, enabling clients to distill complex data into significant insights. By leveraging progressed NLP and machine learning strategies, coupled with natural client interfacing and consistent integration of subordinate technologies such as OCR, AudioInsight sets a new standard in audio content handling, clearing the way for improved efficiency, educated decision-making, and unparalleled experiences.

Keyword : - *BERT, NLP, ASR, T5, OCR, Tokenization, Summarization, Whisper, Transcribing.*

1. INTRODUCTION

One of the key applications of automatic speech recognition is to transcribe speech documents such as talks, presentations, lectures, and broadcast news [1]. Although speech is the most natural and effective method of communication between human beings, it is not easy to quickly review, retrieve, and reuse speech documents if they

are simply recorded as audio signal. Therefore, transcribing speech is expected to become a crucial capability for the coming IT era. Although high recognition accuracy can be easily obtained for speech read from a text, such as anchor speakers' broadcast news utterances, technological ability for recognizing spontaneous speech is still limited [2]. The issue of effectively processing and comprehending large volumes of audio data has grown significantly with the exponential expansion of digital material. "AudioInsight" is presented in this work. With the added feature of OCR for users, AudioInsight is a comprehensive summarization system that uses cutting-edge natural language processing (NLP) and machine learning techniques to transform spoken content into succinct textual summaries and vice versa. It also provides grammar error correction, handles word count inconsistencies, and visualizes important insights.

1.1 Speech to Text

Speech, is the foremost capable way of communication with which human beings express their considerations and sentiments through distinctive dialects. The highlights of discourse vary with each dialect. However, while communicating within the same dialect, the pace and the lingo changes with each individual. This makes trouble in understanding the passed-on message for a few individuals. Sometimes lengthy speeches are quite troublesome to follow due to reasons such as distinctive articulation, pace and so on. Discourse acknowledgment which is an associate disciplinary field of computational linguistics helps in creating advances that enables the acknowledgment and interpretation of discourse into content. Content summarization extricates the utmost important data from a source which may be a content and provides the satisfactory rundown of the same. The inquire about work displayed in this paper depicts a straightforward and successful strategy for discourse acknowledgment. The discourse is converted to the comparing content and produces summarized content. This has different applications like lecture notes creation, summarizing catalogues for long archives and so on. React Speech Recognition gives a command alternative to perform a certain task based on a specific speech phrase.

1.2 Text to Speech

Speech-to-text technology, commonly known as ASR, turns talked dialect into composed content. It permits for the translation of talked words into computerized format, making it a valuable device for assortment of applications. Recent progresses in deep learning and natural language processing have driven to outstanding advance in speech-to-text frameworks. These frameworks are valuable in numerous distinctive spaces, including as voice assistants, accessibility solutions for individuals with disabilities, and translating services. They have streamlined the process of turning talked words into content, empowering more compelling and helpful communication, documentation, and data retrieval.

1.3 OCR (Optical Character Recognition)

OCR processes for uploaded files include a few complex steps that collectively change image-based or scanned substance into text. At first, the file is pre-processed to improve its quality, counting assignments like picture improvement, noise-reduction, and deskewing to guarantee ideal OCR results. At that point, content detection distinguishes ranges within the document containing textual substance, recognizing it from other graphical components. Once the content locales are recognized, the OCR engine employs pattern recognition calculations to distinguish and decipher individual characters and words. Language modelling and context analysis play a crucial part in rectifying mistakes and improving recognition precision by considering the setting of words inside the record. Post-processing steps, such as spell-checking and formatting, refine the extricated content to make it more coherent and discernable. At last, the OCR framework yields the digitized content, which can be put away, altered, or searched, offering users the comfort of working with substance from their uploaded files in a advanced and editable format. These processes have advanced essentially over the years, driven by progressions in artificial intelligence and profound learning, coming about in progressively precise and proficient OCR arrangements.

1.4 Summarization

Extractive text summarization using BERT represents a powerful approach to distilling key information from textual documents. BERT (Bidirectional Encoder Representations from Transformers), a state-of-the-art language model, is fine-tuned on a specific summarization task, enabling it to identify and extract salient sentences directly from the input text. By leveraging BERT's contextual understanding of language, this method selects sentences that capture the essence of the document, preserving the original context and meaning. This extractive approach offers a

streamlined solution for generating concise summaries while maintaining the coherence and relevance of the source material.

2. SYSTEM ARCHITECTURE

The system architecture is designed to provide a robust and scalable solution for text and audio summarization, as well as optical character recognition (OCR). It consists of a backend built using Flask, a Python-based web framework, and a frontend developed with ReactJS, a JavaScript library for building user interfaces. Additionally, Tailwind CSS is used for styling and UI design.

2.1 Backend Architecture (Flask):

The Flask backend serves as the core of the system, handling client requests, data processing, and interaction with external services. It is structured using the Model-View-Controller (MVC) pattern for modularity and maintainability.

1. Controllers: Flask routes are defined to handle HTTP requests from the frontend. Each route corresponds to a specific functionality such as user authentication (login/register), text/audio summarization, and OCR.
2. Models: The backend includes models to represent data entities such as users, text/audio documents, and OCR results. These models are defined using an ORM (Object-Relational Mapping) library such as SQLAlchemy for interacting with the database.
3. Services: Various services are implemented to encapsulate business logic and interact with external APIs or libraries. For grammar correction, the system utilizes the Vennify-T5 model, while the OpenAI_Whisper model is employed for audio processing. Additionally, BERT (Bidirectional Encoder Representations from Transformers) is used for text summarization. These services handle their respective tasks efficiently and provide accurate results.
4. Database: Flask interacts with a relational database (e.g., PostgreSQL, SQLite) to store user information, documents, and summarization results. SQLAlchemy is used to abstract database operations and facilitate data manipulation.

2.2 Frontend Architecture (ReactJS with Tailwind CSS):

The frontend is developed using ReactJS to create a dynamic and responsive user interface. Tailwind CSS is utilized for styling, providing a utility-first approach for rapid UI development.

1. Components: React components are organized hierarchically to represent different UI elements and views. Components are reusable and encapsulate specific functionality such as user authentication forms, document upload interfaces, and summarization results display.
2. State Management: React's state management capabilities (e.g., useState, useContext) are employed to manage application state and facilitate data flow between components. State is updated based on user interactions and backend responses.
3. API Integration: The frontend communicates with the Flask backend via RESTful APIs. Axios or Fetch API is used to make HTTP requests to the backend endpoints, enabling data retrieval and submission for tasks such as user authentication, document upload, and summarization.
4. Routing: React Router is utilized for client-side routing, enabling navigation between different views and components within the application. Route configuration ensures that users can access relevant pages for functionalities like login, registration, and document processing.

3. PERFORMANCE TESTING

For summarization through text we selected the following paragraph “There is a great deal of talk and endeavor to protect nature, the animals, the birds, the whales and dolphins, to clean the polluted rivers, lakes, fields and so on. Nature is not put together by thought, as religion and belief are. Nature is the tiger, that extraordinary animal with its energy, its great sense of power. Nature is the solitary tree in the field, the meadows and the grove; it is that squirrel shyly hiding behind a bough. Nature is the ant, the bee and all the living things of the earth. Nature is the river, not a particular river, whether the Ganga, the Thames or the Mississippi. Nature is those mountains, snow-clad, with dark blue valleys and range of hills meeting the sea. The universe is part of nature. One must have a feeling for all this, not destroy it, not kill for one’s pleasure or one’s table. We do kill cabbages, the vegetables we eat, but one must

draw the line somewhere. If you do not eat vegetables, how will you live? So one must intelligently discern." The performance with respect to some of the popular websites is plotted: -

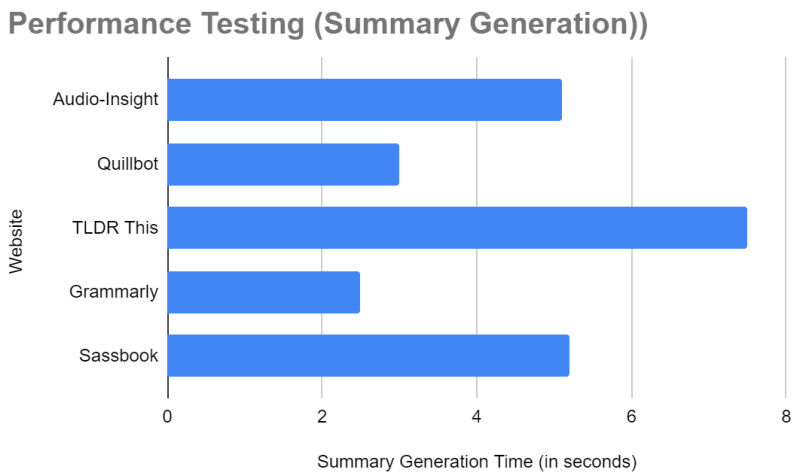

Fig-1 Performance testing of text summary

The model used in audio transcribing is openai whisper with the help of speech recognition module and ffmpeg. To test the performance of audio file processing we used a '.mp3' format of 5.36 mb and recorded the amount of time required for different websites along with our web application. The results are plotted: -

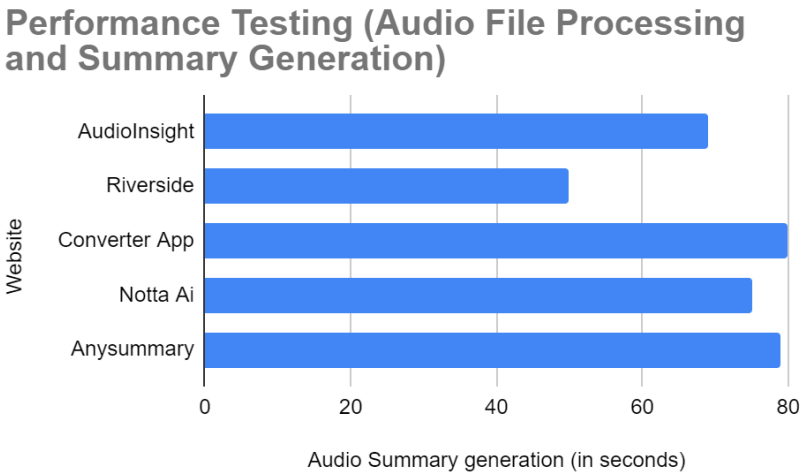

Fig-2 Performance testing of audio summary

4. FUTURE WORK

In our never-ending quest to develop artificial intelligence and its useful applications, we have discovered Three exciting prospects with great potential are image-based text generation, multilingual support, and speech emotion recognition. By allowing computers to recognize human emotions from speech patterns, voice emotion recognition has the potential to completely transform human-computer interaction. From improving customer service encounters to promoting tailored mental health and well-being applications, this technology has a wide range of uses. In addition to these initiatives, we are fully dedicated to expanding Multi-Language Support in order to develop Natural Language Processing (NLP) systems that can produce and process information in several languages

with ease. This project will enable worldwide enhance dialogue and broaden the educational experience for language learners. Our vision of a future where technology links emotional, visual, and linguistic applications, eventually creating a more connected, accessible, and efficient global society, drives us as we navigate this revolutionary terrain.

5. CONCLUSIONS

In conclusion, our study has demonstrated the efficacy of AudioInsight in comparison to similar websites, showcasing its robust performance across various metrics. Through benchmarking, we have validated the effectiveness of AudioInsight's processing technologies, affirming its position as a reliable platform for extracting insights from text as well as audio data. The favorable outcomes of our evaluation underscore the potential of AudioInsight in real-world applications, such as transcription, translation, and sentiment analysis. As technology continues to advance, AudioInsight stands as a testament to the progress in leveraging cutting-edge technologies to unlock valuable insights from audio sources. Moving forward, further research and development efforts can continue to refine and enhance it, ensuring its continued excellence in facilitating text and audio data analysis and interpretation.

6. ACKNOWLEDGEMENT

This paper and the research behind it would not have been possible without the exceptional support of our guide, Adapanawar sir. His enthusiasm, knowledge and attention to detail have been an inspiration and has given valuable inputs in our work on track from our first discussion with the concepts to the final draft of this paper. The group members have contributed greatly with their ideas, finding previous research along with designing the system. We are also grateful for the insightful comments offered by the anonymous peer reviewers at books & texts. The generosity and expertise of one and all have improved this study in innumerable ways and saved us from many errors.

7. REFERENCES

[1]. K. Padmanandam, S. P. V. D. S. Bheri, L. Vegesna and K. Sruthi, ”A Speech Recognized Dynamic Word Cloud Visualization for Text Summarization,” 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2021, pp. 609-613, doi: 10.1109/ICICT50816.2021.9358693.

[2]. K. A. Bharadwaj, M. M. Joshi, N. S. Kumbale, N. S. Shastry, K. Panimozhi and A. Roy Choudhury, ”Speech Automated Examination for Visually Impaired Students,” 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bangalore, India, 2020, pp. 378-381, doi: 10.1109/ICIMIA48430.2020.9074847.

[3]. X. Chang, W. Zhang, Y. Qian, J. L. Roux and S. Watanabe, ”MIMO-Speech: End-to-End Multi-Channel Multi-Speaker Speech Recognition,” 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 2019, pp. 237-244, doi: 10.1109/ASRU46091.2019.9003986.

[4]. W. Minhua, K. Kumatani, S. Sundaram, N. Strom and B. Hoffmeis- ¨ ter, ”Frequency Domain Multi channel Acoustic Modeling for Distant Speech Recognition,” ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 6640-6644, doi: 10.1109/ICASSP.2019.8682977.

[5]. K. S, S. R, S. R and T. S V, ”Survey on Automatic Text Summarization using NLP and Deep Learning,” 2023 International Conference on Advances in Electronics, Communication, Computing and Intelligent Information Systems (ICAECIS), Bangalore, India, 2023, pp. 523-527, doi: 10.1109/ICAECIS58353.2023.10170660.

[6]. S. Furui, T. Kikuchi, Y. Shinnaka and C. Hori, ”Speech-to-text and speech-to-speech summarization of spontaneous speech,” in IEEE Transactions on Speech and Audio Processing, vol. 12, no. 4, pp. 401-408, July 2004, doi: 10.1109/TSA.2004.828699

[7]. Manoj Kumar, and O. Kumar. ”Speech recognition: A review.” International Journal of Advanced Networking and Applications (IJANA) (2014): 62-71.

[8]. Juang, B.H. and Rabiner, L.R., 2005. Automatic speech recognition–a brief history of the technology development. Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara, 1, p.67

[9]. Meng, J., Zhang, J. and Zhao, H., 2012, August. Overview of the speech recognition technology. In 2012 fourth international conference on computational and information sciences (pp. 199-202). IEEE.

[10]. P. Jayasuriya, M. Wijesundara, S. Thelijjagoda and N. Kodagoda, ”Grammar Error Correction for Less Resourceful Languages: A Case Study of Sinhala,” 2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS), Peradeniya, Sri Lanka, 2023, pp. 169-174, doi: 10.1109/ICIIS58898.2023.10253578.
[11]. R. S., V. S., S. T., R. K. and L. Gadhikar, ”Vyakranly : Hindi Grammar & Spelling Errors Detection and Correction System,” 2023 5th Biennial International Conference on Nascent Technologies in Engineering (ICNTE), Navi Mumbai, India, 2023, pp. 1-6, doi: 10.1109/ICNTE56631.2023.10146610.
[12]. X. Xu, ”Design and Implementation of English Grammar Error Correction System Based on Deep Learning,” 2022 3rd International Conference on Information Science and Education (ICISE-IE), Guangzhou, China, 2022, pp. 78-81, doi: 10.1109/ICISE-IE58127.2022.00023.
[13]. Kulkarni, N. D, A. Joshi, M. H. M and N. S. Kumar, ”Automatic Syntax Error Correction,” 2021 Asian Conference on Innovation in Technology (ASIANCON), PUNE, India, 2021, pp. 1-7, doi: 10.1109/ASIANCON51346.2021.954476
[14]. M. Kim, S. -K. Choi and H. -C. Kwon, ”Context-Sensitive Spelling Error Correction Using Inter-Word Semantic Relation Analysis,” 2014 International Conference on Information Science & Applications (ICISA), Seoul, Korea (South), 2014, pp. 1-4, doi: 10.1109/ICISA.2014.6847379.
[15]. T. -T. -H. Nguyen, A. Jatowt, M. Coustaty, N. -V. Nguyen and A. Doucet, ”Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing,” 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), Champaign, IL, USA, 2019, pp. 29-38, doi: 10.1109/JCDL.2019.00015.
[16]. A. B. Salah, J. p. Moreux, N. Ragot and T. Paquet, ”OCR performance prediction using cross-OCR alignment,” 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 2015, pp. 556-560, doi: 10.1109/ICDAR.2015.7333823.
[17]. K. Woo, ”Improving OCR Accuracy on Images with Motion Blur via GAN Derivatives,” 2020 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, USA, 2020, pp. 1-4, doi: 10.1109/URTC51696.2020.9668859.
[18]. L. R. Blando, J. Kanai and T. A. Nartker, ”Prediction of OCR accuracy using simple image features,” Proceedings of 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 1995, pp. 319-322 vol.1, doi: 10.1109/ICDAR.1995.599003.