IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AUDIOINSIGHT: AN APPLICATION OF SPEECH-TO-TEXT, TEXT-TO-SPEECH, AND LANGUAGE PROCESSING TECHNOLOGIES

Mr.A.N.Adapanawar,
Assistant Professor,
Dept. of Computer Engineering,
STES SAE Pune, India

Sharva Khandagale,
Student,
Dept. of Computer Engineering,
STES SAE Pune, India

Tanvi Gaikwad,
Student,
Dept. of Computer Engineering,
STES SAE Pune, India

Mayank Wakdikar,
Student,
Dept. of Computer Engineering,
STES SAE Pune, India

Subrat Dhapola,
Student,
Dept. of Computer Engineering,
STES SAE Pune, India

Abstract: With the exponential growth of digital content, efficiently processing and understanding vast amounts of audio data has become a significant challenge. This paper presents "AudioInsight". AudioInsight is a comprehensive Summarization System that leverages advanced natural language processing (NLP) and machine learning techniques to convert spoken content into concise textual summaries and vice versa, visualize key insights, and provide grammar error correction as well as handle Wrong word count incompatibilities with the added feature of OCR for users.

I. INTRODUCTION

As we reflect on our journey through the captivating realm of Natural Language Processing (NLP) during AudioInsight, our exploration of Text-to-Speech using BERT, Speech-to-Text, Text Summarization with BERT, Grammar Error Correction, and Wrong Word Count has been nothing short of transformative. These five pillars of NLP have illuminated the vast potential of technology in enhancing our interaction with language, paving the way for more efficient, accessible, and error-free communication. Text-to-Speech using BERT has demonstrated the power of technology to give voice to written words, enriching the user experience and making content more accessible to a diverse audience. The ability to hear words come to life holds promise for those with visual impairments, and it's poised to revolutionize audiobooks, navigation systems, and beyond. Conversely, Speech-to-Text technology has broken down barriers in transcription services and voice assistants. This application empowers efficient content creation, particularly in fields where spoken language is integral. It's not just a technological marvel but a real-world solution that streamlines data entry and enhances communication across diverse sectors. Text Summarization with BERT stands out as an invaluable tool for condensing large volumes of text into concise, coherent summaries. It promises to save time and facilitate better decision-making, catering to professionals, researchers, and students alike. Grammar Error Correction and Wrong Word Count tools, on the other hand, are set to refine the quality of written content by identifying and rectifying linguistic and grammatical errors, ensuring our messages, documents, and publications are clear, precise, and linguistically sound. As we conclude our journey, it's clear that these five pillars of NLP are not merely technological marvels but practical solutions with the potential to enhance how we communicate, learn, and share information.

Abbreviations and Acronyms

BERT	Bidirectional Encoder Representations from Transformers
NLP	Natural Language Processing
ASR	Automated Speech Recognition
CLS	Classification Task
SEP	Separate Token
MASKED	Masked Token
T5	Text-to-Text Transfer Transformer
OCR	Optical Character Recognition

II. METHODOLOGY

2.1) Speech-to-text conversion

Speech, is the most powerful way of communication with which human beings express their thoughts and feelings through different languages. The features of speech differ with each language. However, even while communicating in the same language, the pace and the dialect varies with each person. This creates difficulty in understanding the conveyed message for some people. Sometimes lengthy speeches are also quite difficult to follow due to reasons such as different pronunciation, pace and so on. Speech recognition which is an inter disciplinary field of computational linguistics aids in developing technologies that empowers the recognition and translation of speech into text. Text summarization extracts the utmost important information from a source which is a text and provides the adequate summary of the same. The research work presented in this paper describes an easy and effective method for speech recognition. The speech is converted to the corresponding text and produces summarized text. This has various applications like lecture notes creation, summarizing catalogues for lengthy documents and so on. React Speech Recognition provides a command option to perform a certain task based on a specific speech phrase. For example, when a user asks for weather information, you can perform a weather API call. This is just a basic example, but when it comes to voice assistance and control, the possibilities are endless.

2.2) Text to Speech

Speech-to-text technology, commonly known as ASR, turns spoken language into written text. It allows for the transcription of spoken words into digital format, making it a useful tool for a variety of applications. Recent advances in deep learning and natural language processing have led to notable progress in speech-to-text systems. These systems are useful in many different domains, including as voice assistants, accessibility solutions for people with disabilities, and transcribing services. They have simplified the process of turning spoken words into text, enabling more effective and convenient communication, documentation, and information retrieval.

2.3) Summarization

1) BERT Language Model

BERT is pre-trained language model which comprises of a set of transformer encoders which represents the text at word and

sentence level with the help of unsupervised training techniques like masked language modeling and next sentence prediction.

BERT being a pre-trained model, it is trained on 3300M words. To learn contextual relations between words in a text BERT uses

a transformer encoder with an attention mechanism. Transformer [9] in their native form consists of the encoder as well as decoder

where encoder learns the text input and decoder is tuned to conduct a specific task. As BERT is a language model, understanding

the input text is the only important factor. Due to this reason only transformer encoders are used in BERT. Rather than reading

the text input sequentially like various directional models [3][5][6][7], the transformer encoder reads the entire sequence of words

at once. This functionality of the transformer helps generate context by calculating the relevance of each word concerning the

presence of other words in the sentence. The level of contextual understanding is directly proportional to several transformer

encoder layers. Fig. 2 shows BERT's architecture. BERT-Base, Uncased is used for the purpose which has 12 transformer layers,

768 hidden state, 12 attention heads, 110M parameters.

2) Information Flow in BERT

In this language model, language modeling objective is achieved by learning contextual representation from large scale corpora which extends the idea of word embeddings. In BERT two special tokens are inserted into a sentence. [CLS] token is added at the beginning for the sentences, it is accumulated with the understanding developed by the language model for those sentences at the output. [SEP] is appended at the end to represent the end of a sentence. These tokens are added into the text where the text is represented as a series of tokens, X = [word1, word2, ..., wordn]. In this modified text representation, for every token 3 different embeddings are allocated. First is Token embedding, it helps to convert word into a vector with a fixed dimension which contains the meaning of the word. In BERT this is done by WordPiece tokenization [11]. Second is Sentence embeddings, they help with differentiation between two sentences. Third, Position embedding represents the position of every token in the text. These embeddings are then added up to one single input vector and are passed on to a multi-layered transformer. A multi-layered transformer consists of transformer encoders which are stacked on each other. Only the first encoder receives the input vector while the other encoder's input is output from the previous encoder. Each transformer encoder consists of a multi-head attention layer and a feedforward layer with layer normalization at the output of each layer. In an encoder, every word (token) flow in their path. In the attention layer, their paths interact whereas in feedforward there is no interaction between them.

3) Training in BERT

BERT natively uses masked language models and "next sentence prediction" to train on tokens. In the masked language model, some percentage of the input token is replaced by [MASKED] token. In the case of BERT, it is 15 percentage. Then the model is trained by predicting the masked words by only taking help from contextual understanding provided by the rest of the non-masked words. In the next sentence prediction, a pair of sentences are passed to the model and it is trained by classifying whether the subsequent sentence in the original document is the second sentence in the pair. While in this training, half of the two-sentence pair input, where the second sentence is random, but for the remaining half, the second sentence is the subsequent sentence from the text corpus. [CLS] token, [SEP] token with the sentence, and position embedding helps BERT to distinguish between two sentences while training. generated for every attention head. These attention score matrices are combined and then reduced to match the input dimension of the feed-forward layer.

4) Summarization Encoder

In "next sentence prediction", BERT deals only with sentence pair input. In this summarization model, as the model needs to develop a contextual understanding between multiple sentences it should be able to deal with multiple sentence input. To enable multi-sentence input, this model changes how [CLS] and [SEP] were used natively in BERT. In this summarizer encoder [CLS] tokens are added at the start of every sentence instead of at the start of the first sentence which helps to aggregate the information of the sentence preceding it. Fig. 2 shows the architecture of BERT for summarization. The document is represented in such a way that lower transformation layers represent adjacent sentences and higher layers of the transformer represent the multi-sentence in combination with self-attention. As BERT is pre-trained, encoder requires less training.

5) Dataset and Pre-Processing

The model is trained on CNN/DailyMail News highlights dataset. This dataset contains news articles and associated highlights which serve as a reference summary. Standard splitting of the dataset is used for training, validation, and testing of the model. The data is split and pre-processed with the help of the Stanford CoreNLP toolkit [12] and the input documents are limited to 512 tokens.

2.4) Grammar Error Correction

To establish a grammar error correction system using a T5 model currently fine-tuned on the C4-200M dataset, we are following

this process. Firstly, we are gathering a dataset of sentences containing grammar errors. Next, we are pre-processing the data and

fine-tuning a pre-trained T5 model on this dataset, employing the C4-200M dataset as the basis for pre-training. As of now, we

are defining the training configurations and training the model while actively monitoring its performance. We are continuously

evaluating its grammar correction capabilities and making adjustments when necessary. Following the fine-tuning process, we

are utilizing the model to correct sentences in real time. We have implemented post-processing steps as needed, deployed the

model, and are considering a feedback loop for ongoing enhancements. Utilizing a T5 model that has been trained on the C4-

200M dataset proves to be a powerful approach for grammar error correction, with our ongoing attention to training and evaluation

ensuring its current effectiveness.

2.5)Wrong Word Count

Wrong word count, often referred to as inaccurate or incorrect word count, is a prevalent issue in various forms of written communication, ranging from academic essays and professional reports to creative writing and online content. This problem arises when the actual number of words in a document does not align with the stated word count, whether it is too high or too low. Inaccurate word counts can be a significant concern, particularly in contexts where specific word limits are essential for compliance, such as in academic assignments or submissions to publishers. Students may find themselves penalized for exceeding or falling short of specified word counts in their essays, potentially affecting their grades. Likewise, authors submitting manuscripts to journals or publishers might face rejection or revision requests due to word count discrepancies. The T5 model, officially known as the Text-to-Text Transfer Transformer, represents a pinnacle of achievement in the domain of natural language processing, setting new standards in the correction of grammatical errors. T5's exceptional proficiency extends well beyond its singular expertise in grammar correction, encompassing a wide array of text-based tasks that have redefined the landscape of language processing.

At its core lies a formidable transformer architecture, a testament to the transformative potential of deep learning models in understanding and generating human language. T5's modus operandi, which approaches grammar correction as a text-to-text task, distinguishes it from traditional rule-based grammar checkers. This ingenious approach allows T5 to not only detect grammatical errors but to also generate coherent and contextually sensitive corrections. By analyzing the context in which these errors manifest, T5 ensures that its corrections seamlessly integrate with the surrounding text, enhancing both the correctness and overall quality of the written content. Its versatility is astonishing, enabling T5 to address a vast spectrum of grammatical issues, from the intricacies of subject-verb agreement and punctuation to the complexities of sentence structure. This makes T5 an invaluable tool across diverse applications, from refining academic prose to facilitating content generation for businesses and individuals.

T5's grammar correction capabilities exemplify the immense potential of transformer-based models in natural language understanding and generation, transcending the boundaries of mere error correction to become a pivotal contributor to the evolution of language technologies. Its far-reaching applications span a multitude of fields, asserting itself as an indispensable asset in the ever-advancing domain of language processing and communication. The determination of word count using the T5 model, or any analogous language model, follows a structured procedure to ensure precision. T5, a robust model developed by Google, is capable of handling various natural language processing tasks, including word count estimation. In the context of research papers, the process can be elucidated as follows:

1) **Tokenization**: The initial step involves tokenizing the input text, which breaks it down into smaller units, typically sub word tokens. Each word or sub word is represented by a unique token. Tokenization is a pre-requisite to enable the model to effectively process and understand the text.

2) Model Input: The tokenized text is then supplied as input to the T5 model. T5, being a text-to-text model, frames tasks as text prompts. In the context of word count, the input serves as a textual prompt instructing the model to ascertain the word count within the given text.

3) Inference: The T5 model generates an output based on the provided text prompt. In the case of word count, the model's response is a numerical value signifying the word count in the input text.

4) Post-processing: Following the model's output generation, post-processing is required to extract the word count value and make it accessible. This involves extracting the numerical figure generated by the model and utilizing it as the final word count.

It is important to emphasize that T5, akin to other language models, calculates word count in terms of tokens rather than actual words. Tokens may vary in length, with some tokens representing complete words, while others represent portions of a word. Hence, to obtain an accurate word count, a conversion of token count to actual word count may be necessary. This conversion should consider the average token length for words in the specific language under analysis. Utilizing the T5 model for word count determination can be an invaluable tool in the realm of research, offering applications such as content analysis, document summarization, and validation of text length. Nevertheless, researchers should remain cognizant of tokenization intricacies and the fact that the model quantifies tokens, not words. This awareness is vital when interpreting and applying results to specific research tasks.

2.6)Optical Character Recognizer

OCR processes for uploaded files involve several intricate steps that collectively transform image-based or scanned content into machine-readable text. Initially, the uploaded file is pre-processed to enhance its quality, including tasks like image enhancement, noise reduction, and deskewing to ensure optimal OCR results. Then, text detection identifies areas in the document containing textual content, distinguishing it from other graphical elements. Once the text regions are recognized, the OCR engine employs pattern recognition algorithms to identify and interpret individual characters and words. Language modelling and context analysis play a vital role in correcting errors and improving recognition accuracy by considering the context of words within the document. Post-processing steps, such as spell-checking and formatting, refine the extracted text to make it more coherent and readable. Finally, the OCR system outputs the digitized text, which can be stored, edited, or searched, offering users the convenience of working with content from their uploaded files in a digital and editable format. These processes have evolved significantly over the years, driven by advancements in artificial intelligence and deep learning, resulting in increasingly accurate and efficient OCR solutions for uploaded files.

III. System Architecture

In our work, we've witnessed the power of Text-to-Speech to give voice to written content, making audiobooks, navigation systems, and accessibility tools more engaging and inclusive. This technology has also found its place in the entertainment industry, where it breathes life into virtual assistants, enhancing our interactions with smart devices. Speech-to-Text has emerged as a game-changer in transcription services, voice assistants, and communication tools, allowing for the accurate conversion of spoken words into text and facilitating effortless content creation and data entry. This application has revolutionized industries like healthcare, where doctors can quickly dictate notes, and in the education sector, where it enables real-time transcription for students with hearing impairments. Text Summarization, another remarkable application, streamlines information consumption and knowledge management. It simplifies the process of sifting through vast amounts of news articles or research papers, providing concise summaries for efficient information gathering and decision making in fields like journalism, finance, and academia. Moreover, BERT's capabilities in Grammar Error Correction and Wrong Word Count have a pivotal role in enhancing the quality of written content across industries. It ensures that marketing materials, legal documents, academic papers, and medical reports are free from errors, with far-reaching implications in real-world scenarios. Additionally, Optical Character Recognition (OCR) technology has revolutionized the way we convert printed or handwritten text into machine-readable data. Its applications are diverse, from digitizing paper documents for archiving to assisting visually impaired individuals in accessing printed content. As we conclude our exploration, it's evident that these practical applications are reshaping the landscape of language interaction, communication, and information processing.

IV. Future Work

In our relentless pursuit of advancing artificial intelligence and its practical applications, we have identified three dynamic frontiers that hold immense promise: Speech Emotion Recognition, Image based Text Generation, and Multi-Language Support. Speech Emotion Recognition is poised to revolutionize human-computer interaction by enabling machines to discern human emotions from speech patterns. This technology has wide-ranging applications, from enhancing customer service interactions to fostering personalized mental health and well-being applications. Simultaneously, our venture into Image Recognition harnesses deep learning models to empower machines to identify objects and patterns within images, offering potential benefits in security and convenience. Alongside these endeavors, we are deeply committed to advancing Multi-Language Support, with the aim of creating Natural Language Processing (NLP) systems capable of seamlessly processing and generating content in multiple languages. This initiative will facilitate global communication and make education more inclusive for language learners. Our journey through these innovative landscapes is motivated by the vision of a future where technology bridges emotional, visual, and linguistic applications, ultimately fostering a more connected, accessible, and efficient global society.

V. Acknowledgment

This paper and the research behind it would not have been possible without the exceptional support of our guide, Adanapanawar sir. His enthusiasm, knowledge and exacting attention to detail have been an inspiration and kept 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.

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] harma, 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.