e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

International Journal Of Multidisciplinary Research In Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 4, April 2024 |

| DOI:10.15680/IJMRSET.2024.0704213 |

Minutes of Meeting using Deep Learning

G Kushal, N Sukesh, T Jayanth, S Praveena

IV B. Tech, Department of ECE, MGIT, Gandipet, Hyderabad, India

IV B. Tech, Department of ECE, MGIT, Gandipet, Hyderabad, India

IV B. Tech, Department of ECE, MGIT, Gandipet, Hyderabad, India

Associate Professor, Department of ECE, MGIT, Gandipet, Hyderabad, India

ABSTRACT: This paper leverages state-of-the-art technologies in speech recognition, natural language processing (NLP), and summarization to automate insights extraction from audio recordings, such as meetings. It employs the Whisper model for transcription tasks, integrated with PyAudio for cross-platform audio handling. The system processes audio data through various steps, including noise reduction, speaker diarization, and transcription. Summarization is achieved using transformer-based models like BERT and LED (The Longformer Encoder-Decoder), enabling concise representations of transcripts. Furthermore, multilingual support is ensured via Google Translate for translations and language identification models. The approach involves integrating these components into a scalable pipeline with user-friendly web interfaces for accessibility and robust security measures. Comprehensive testing metrics and user feedback are utilized for iterative improvements. The results include multilingual transcripts, summaries, and actionable insights, demonstrating a holistic solution for managing audio-based information efficiently.

KEYWORDS: Deep leaning, BERT and LED

I. INTRODUCTION

The rapid advancements in speech recognition and natural language processing (NLP) technologies have opened avenues for automating the extraction of actionable insights from audio data. Meetings, interviews, and lectures, which often contain valuable information, are commonly stored as audio recordings, making efficient transcription and summarization essential for accessibility and analysis. This paper integrates cutting-edge methodologies in audio processing, transcription, and summarization to address the challenges of extracting and presenting information from audio recordings. Leveraging the Whisper model, a state-of-the-art encoder-decoder Transformer trained on extensive labeled datasets, this system ensures robust transcription performance across languages and domains. Complementing this is the use of advanced NLP models like BERT and LED for summarization, allowing for condensed yet comprehensive representations of audio content. Additionally, the system is designed with multilingual support and incorporates Google Translate API for translation tasks, enhancing its applicability across global use cases. Key features include audio preprocessing for quality enhancement, speaker diarization for segmentation, and an intuitive web interface for seamless user interaction. Scalability and security considerations ensure the system's suitability for real-world deployment.This work aims to provide a comprehensive, automated solution for transcription, summarization, and translation of audio recordings, offering significant value in domains such as education, business, and research. The methodology, results, and potential applications are discussed in detail in the subsequent sections.

II. RELATED WORK

To gain a comprehensive understanding of the domain and the prior research conducted in this field, several studies and research papers were reviewed. Below is an overview of some of the referenced papers, organized chronologically from the most recent to the oldest:

In 2021, Megha Manuel et al. [1] introduced the Automated Minute Book Creation (AMBOC) model, utilizing machine learning to extract key information from meeting discussions. While the model achieved high accuracy and could differentiate between male and female speakers, its performance was suboptimal for languages other than English. Also in 2021, Jia Jin Koay et al. [2] implemented a sliding window approach combined with the BART summarizer to generate meeting summaries. The method's scalability and lack of reliance on annotated data were significant advantages, but its accuracy faltered in detecting highly relevant utterances in certain scenarios. In 2020, Chenguang Zhu et al. [3] proposed the Hierarchical Meeting Summarization Network Model (HMNet), based on an encoder-decoder transformer architecture. The model demonstrated strong performance across automated and human evaluation metrics. However, it struggled to provide adequate coverage of detailed items in lengthy meeting transcripts. In 2019,

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 4, April 2024 |

| DOI:10.15680/IJMRSET.2024.0704213 |

Anna Nedoluzhko et al. [4] conducted an analysis of the tools and datasets required for automatic meeting minutes generation. They outlined a topology of methods and meeting types but failed to develop a fully functional model for practical use. Guokan Shang et al. [5], in 2018, presented a framework leveraging Multi Sentence Compression Graph (MSCG) to create summaries. While the method produced well-structured summaries comparable to human-written ones, the lack of coherence between certain entities rendered some outputs unusable. Siddhartha Banerjee et al. [6], in 2016, proposed a supervised learning method for segmenting topics and identifying key utterances. These utterances were combined, and the best sub-graph was selected through integer linear programming (ILP) as the final output. In 2014, Tatsuro Oya et al. [7] introduced a novel multi-sentence fusion algorithm for generating summary templates using lexico-semantic information. Templates were selected based on their alignment with human-evaluated summaries and source transcripts. Yashar Mehdad et al. [8], in 2013, utilized an Entailment Graph to identify semantic relationships between sentences, which were then used to construct a word graph model for summarizing meeting transcripts. While the model outperformed baselines in ROUGE-1 scoring, it ranked third in ROUGE-2. Finally, in 2012, Lu Wang et al. [9] proposed an unsupervised framework that employed an in-domain relation learner to process clusters of Decision-Related Dialogue Acts (DRDAs) for abstractive summarization. The model outperformed several baseline systems and produced competitive results when compared to supervised frameworks.

III. METHODOLOGY

Figure 1 shows the methodology followed in this work.

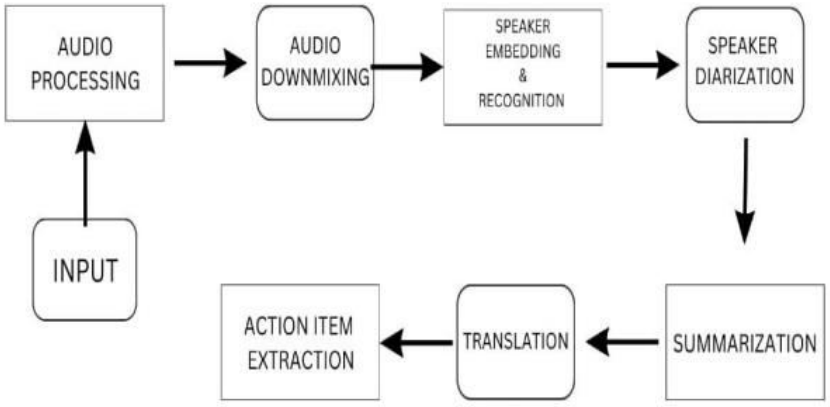

Figure 1

This Figure 1 represents the workflow of an advanced audio processing system designed for tasks such as speaker recognition, diarization, translation, summarization, and action item extraction. The process begins with raw audio input, which undergoes preprocessing in the Audio Processing stage to enhance quality and prepare it for analysis. The audio is then subjected to Audio Downmixing, where multi-channel audio is combined into a single channel or a standardized format. Next, the system performs Speaker Embedding and Recognition, generating unique embeddings to identify speakers based on prior data. This is followed by Speaker Diarization, which segments the audio to distinguish and track speakers over time. After diarization, the content is processed for Summarization, where the system condenses the information into a concise format. If needed, the summarized content is further translated into a target language in the Translation stage. Finally, the system identifies and extracts actionable tasks or decisions during the Action Item Extraction phase. This workflow is particularly useful in applications like meeting transcription, call center analytics, and intelligent virtual assistants.

1. Input Data to the Whisper Model: Whisper Model[11] is a model designed for audio processing, which is likely used for tasks such as speech recognition or audio transcription. The Whisper model takes audio data as input and processes it to generate transcriptions or other relevant information.

2. Training the Whisper Model: After providing the input data, the Whisper model undergoes a training phase. During training, the model learns patterns and features from the input audio data. The training process involves adjusting the model's parameters to minimize the difference between its predicted outputs and the actual outputs (ground truth).

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 4, April 2024 |

| DOI:10.15680/IJMRSET.2024.0704213 |

3. Input Audio File for Transcription: Once the Whisper model is trained, it can be used to process new audio files. In this case, an audio file is given as input to the trained Whisper model.

4. Transcription with Whisper Model: The Whisper model transcribes the audio file, converting spoken words into text. The output is a transcript of the audio content.

5. Text Summarization with BERT Model: BERT[10] (Bidirectional Encoder Representations from Transformers) is a pre-trained natural language processing model. In this step, the transcript obtained from the Whisper model is processed using a BERT model that has been fine-tuned for text summarization. Fine-tuning involves training the BERT model on a specific task, in this case, summarizing text.

6. Summarizing the Transcript: The BERT model generates a summary of the transcript. Text summarization aims to condense the content of the transcript while 23 retaining key information. This summary provides a concise representation of the audio content.

7. Translation with Google Translate API: Google Translate API allows for the translation of text from one language to another. In this step, the summarized transcript is translated using the Google Translate API.
8. Final Output: The final output is likely the translated and summarized version of the original audio content. This could be useful for scenarios where there is a need to understand or communicate the content of the audio in a different language and in a condensed form.

IV. EXPERIMENTAL RESULTS

Figure 2 represents the conversation between the two speakers. It is generated by Figure3 is the transcript summary obtained by LED model. Figure 4 shows telugu conversation by GOOGLE translate model . Figure 5 shows the Figure 5 shows the telugu summary output obtained by LED model.

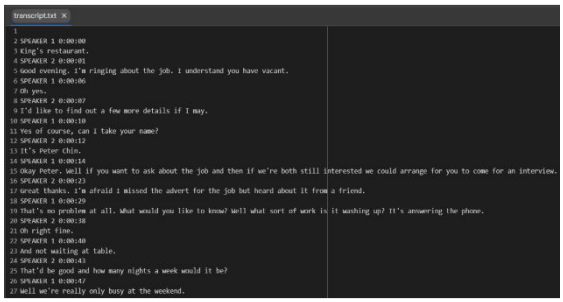

Figure 2

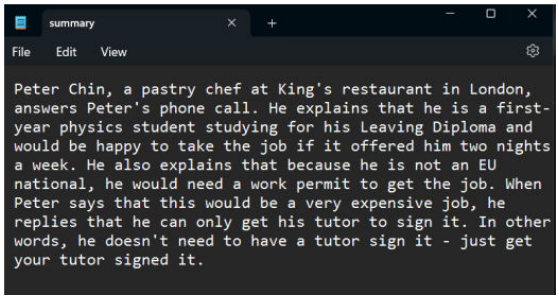

Figure 3

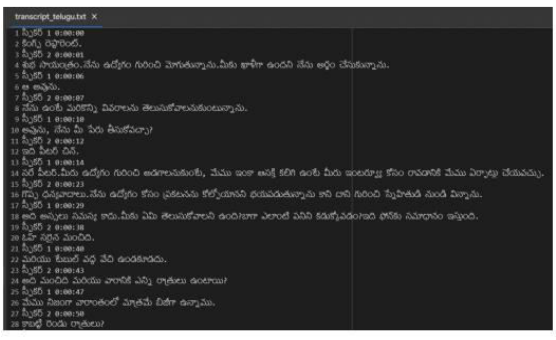

Figure 4

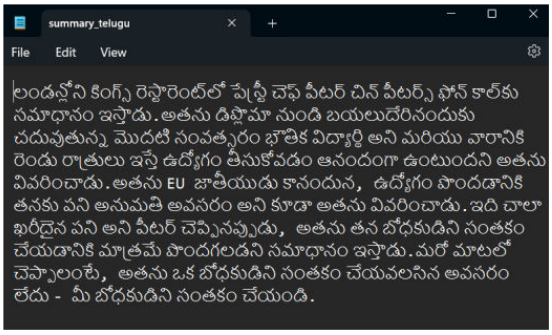

Figure 5

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 4, April 2024 |

| DOI:10.15680/IJMRSET.2024.0704213 |

REFERENCES

[1] Manuel, Megha; Menon, Amritha S; Kallivayalil, Anna; Isaac, Suzana; K.S, Dr. Lakshmi" Automated Generation of Meeting Minutes Using Deep Learning Techniques" IJCDS July 2021

[2] Koay, Jia Roustai, Alex Dai, Xiaojin Liu, Fei. (2021). A Sliding-Window Approach to Automatic Creation of Meeting Minutes.2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop. 68-75. 10.18653/v1/2021.naacl-srw.10.

[3] Zhu, Chenguang Xu, Ruochen Zeng, Michael Huang, Xuedong. (2020). A Hierarchical Network for Abstractive Meeting Summarization with Cross-Domain Pretraining. Findings of the Association for Computational Linguistics: EMNLP 2020. 194-203. 10.18653/v1/2020.findings-emnlp.19.

[4] Nedoluzhko, Anna and Ondrej Bojar. "Towards Automatic Minuting of the Meetings." ITAT (2019). 19th Conference Information Technologies - Applications and Theory (ITAT 2019).

[5] Shang, Guokan Ding, Wensi Zhang, Zekun Tixier, Antoine Meladianos, Polykarpos Vazirgiannis,Michalis LORRE, Jean-Pierre. (2018). Unsupervised Abstractive Meeting Summarization With Multi-Sentence Compression and Budgeted Submodular Maximization. 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)664- 674.10.18653/v1/P18-1062.

[6] Banerjee, Siddhartha & Mitra, Prasenjit & Sugiyama, Kazunari. (2016). Abstractive Meeting Summarization Using Dependency Graph Fusion.The 24th International World Wide Web Conference (WWW 2015 Companion).

[7] Oya, T., Mehdad, Y., Carenini, G., Ng, R. (2014). A Template-based Abstractive Meeting Summarization: Leveraging Summary and Source Text Relationships. INLG. 8th International Natural Language Generation Conference (INLG).

[8] Mehdad, Yashar Carenini, Giuseppe Tompa, Frank Ng, Raymond. (2013). Abstractive Meeting Summarization with Entailment and Fusion. 136-146. 14th European Workshop on Natural Language Generation.

[9] Wang, Lu Cardie, Claire. (2012). Focused Meeting Summarization via Unsupervised Relation Extraction. SIGDIAL Conference 5 July 2012.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805v2 [cs.CL] 24 May 2019

[11] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever, Robust Speech Recognition via Large-Scale Weak Supervision, Xiv:2212.04356v1 [eess.AS] 6 Dec 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com