iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.61710

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 12 Issue V May 2024- Available at www.ijraset.com

Summarization for Enhanced Video Conference Collaboration

Aswin Thampi[1], Shahinsha P S[2], Vishnu V V[3], Pristy Paul T[4]
Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala

Abstract: ***Efficient collaboration during video conferences is crucial for team productivity. However, the absence of effective mechanisms to capture and preserve key insights often results in misunderstandings and reduced efficiency. The proposed initiative focuses on developing a web application to improve collaboration in video conferences by converting speech conversations into summarized text. Leveraging machine learning models, the system transcribes spoken content using whisper model and employs abstractive summarization using Pegasus X-Sum model. The Whisper architecture is implemented as an encoder-decoder Transformer. Input audio is split into 30-second chunks, converted into a log-Mel spectrogram, and then passed into an encoder. A decoder is trained to predict the corresponding text caption, intermixed with special tokens that direct the model to perform transcription. The goal is to empower users to effortlessly capture and retain crucial discussions, promoting better understanding and decision-making.***

Index Terms: PDF- Portable Document Format, MLM - Mask Language Modelling, GSM – Gap Sentence Generation, RNN-Recurrent Neural Network, HTML - Hyper Text Markup, Language, CSS - Cascading Style Sheets

I. INTRODUCTION

I. Introduction In today's digital era, video conferencing has become an integral part of modern workplace communication. To address the challenges associated with information retention and comprehension in video conferencing, our project introduces a web application leveraging machine learning technology to incorporate transcription and summarization models. II. Literature Review A web application is introduced to address the current problems faced in video conferencing platforms. Key features include transcription of speech content, Summarization of transcribed content, and a video conferencing platform. However, reliance on consistent internet connectivity may pose a limitation. Nonetheless, the application enhances the virtual conferences. III. Methodology The methodology involves a model utilizing the Transformer Encoder Decoder architecture for transcribing the speech content. The audio is processed by encoder decoder to convert it into a text representation , further another Transformer model processes the text to convert it into a summary. By utilising the attention mechanism ,the transformer architecture enhances its ability to transcribe and summarize the inputs. IV. Implementation Details The implementation involves using python for capturing audio , further the captured audio is given as input to the transcription model and summarization model. Both models employees a Transformer encoder decoder architecture.

Transcription model capture the audio signals , converts it into text representation. Summarization model uses MLM and GSG techniques to generate summary from the output text of transcription model. These models are integrated into a web application, offering a seamless user experience. V. Experimental Evaluation The experimental evaluation involves using a WER metric for evaluating the transcription model and ROUGE metric analysis and human evaluation is done on Summarization model. VI. Conclusion and Future Work Application of transcription and summarization in video conferencing platforms enhances the efficiency of virtual meetings. Expanding transcription and Summarisation capabilities to accommodate multiple speakers and incorporating translation services are the future considerations. VII. References A comprehensive list of references used throughout the project will be provided. Note: This condensed version of the project outline has been formatted to fit a single page, providing an overview of the project's key aspects while maintaining its main structure.

II. RELATED WORKS

A. Data Collection

A diverse dataset of audio recordings for training the Whisper model. This dataset should cover a range of speakers, accents, and speech patterns to ensure robustness. Collect text data for training and evaluating the Pegasus model. This text data can be sourced from various sources such as news articles, blogs, or books.

B. Data Preprocessing

Preprocess the audio data by converting it into a suitable format for input to the Whisper model. This involve converting audio files into spectrograms suitable for neural network processing. Clean and preprocess the text data for input to the Pegasus model. This include tokenization, lowercasing, removing special characters, and handling any language-specific nuances.

C. Feature Extraction

Train the model to extract features from the audio data, such as phonetic representations or spectral features, using an encoder-decoder architecture. For summarization, using transformer-based architecture extract features from the transcribed text data.

D. Training and Evaluation

The performance of both the transcription and summarization models is evaluated based on various metrics and qualitative assessments. For the transcription model, metrics such as word error rate (WER) are computed to quantify the model's ability to accurately transcribe audio input into text. Furthermore, the summarization model's performance is assessed using metrics such as ROUGE scores (Recall Oriented Understudy for Gisting Evaluation), which measure the overlap between generated summaries and reference summaries. Manual inspection of summarised data were also done.

E. Model Integration and Deployment

Implement the Whisper and Pegasus models within a Flask application, exposing two separate endpoints for transcription and summarization. Incorporate the Flask application on high-end servers with GPU /Google collab support for enhanced performance. Develop a React frontend to interact with the Flask backend, allowing users to send audio files for transcription and view the generated summaries. Ensure seamless communication between the frontend and backend, handling file uploads, API requests, and responses efficiently. Test the integration thoroughly to ensure reliability, scalability, and usability of the end-to-end system. Use WebRTC and Socket for Video conferencing and real time communication.

III. PROPOSED MODEL

First the audio is recorded using python , then it is converted to an audio file , which is processed by the Transcription model, further the output from the transcription model is given as input to the summarization model.

A. Transformer for Transcription

The transformer architecture, pivotal in transcription tasks, comprises an encoder and decoder. The encoder transforms input audio features, like spectrograms, into meaningful representations. A key innovation, the self-attention mechanism, enables the model to weigh the importance of each input token, capturing contextual dependencies effectively. Employing multiple attention heads enhances its ability to capture diverse relationships concurrently. As the transformer architecture lacks inherent sequential order capture, positional encoding is incorporated to inform the model about each token's position in the sequence. Multiple layers of transformer blocks in the encoder enable the model to learn hierarchical representations, capturing both low and high-level information. The final layer of the

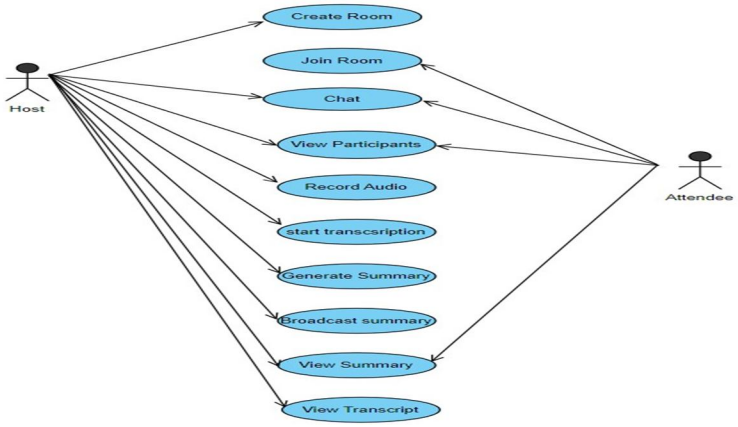


Fig. 1. Use Case Diagram of Our Proposed Model

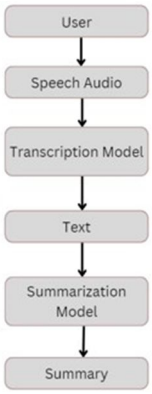


Fig. 2. Data Flow Diagram of Our Proposed Model

Encoder produces contextualized representations of the input audio features. The decoder, receiving the contextualized representation from the encoder as its initial state, initiates the generation process with a special token (e.g., ¡start¿). At each step, considering the previously generated token, it predicts the next word in the transcription sequence. This iterative process continues until a special end-of-sequence token (e.g., ¡end¿) is generated or a predefined maximum sequence length is reached.

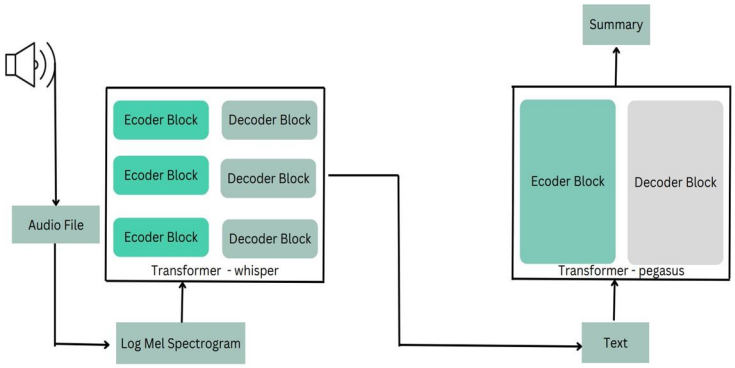


Fig. 3. proposed model

B. Transformer for Summarization

In the summarization process, the encoder plays a pivotal role in processing the input text sequence, typically the transcribed document. Comprising multiple layers of Transformer blocks, each equipped with self-attention mechanisms, the encoder meticulously analyzes the relationships between words within the text, capturing nuanced context and meaning. Through iterative computation, the encoder distills the essential information contained in the input text into a compressed representation, effectively summarizing its content.

Subsequently, the decoder takes on the responsibility of generating the summary text based on the compressed representation provided by the encoder. Employing another set of Transformer blocks with attention mechanisms, the decoder meticulously attends to the encoder output, retrieving pertinent information for summarizing the text. In a step-by-step fashion, each decoder layer generates one word or phrase at a time, progressively building the summary.

Furthermore, advanced techniques such as gap sentence generation and mask language modeling may be incorporated into the decoder's architecture to further enhance the quality and coherence of the generated summary. These techniques enable the decoder to fill in missing information, predict the next word or phrase based on context, and ensure that the generated summary maintains syntactic and semantic coherence.

Overall, the synergy between the encoder and decoder components in the summarization process enables the model to effectively distill complex textual information into concise and coherent summaries, facilitating efficient information retrieval and comprehension

IV. RESULT

The performance of both the transcription and summarization models is evaluated based on various metrics and qualitative assessments. For the transcription model, metrics such as word error rate (WER) are computed to quantify the model's ability to accurately transcribe audio input into text. Furthermore, the summarization model's performance is assessed using metrics such as ROUGE scores (RecallOriented Understudy for Gisting Evaluation), which measure the overlap between generated summaries and reference summaries. Manual inspection of summarized data was also done.

Fig. 4. Landing Page

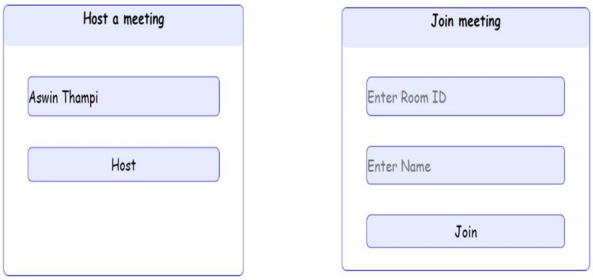


Fig. 5. Creating and Joining Room

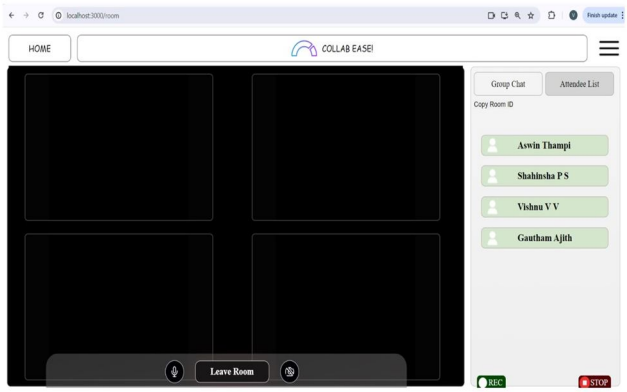


Fig. 6. Home Page

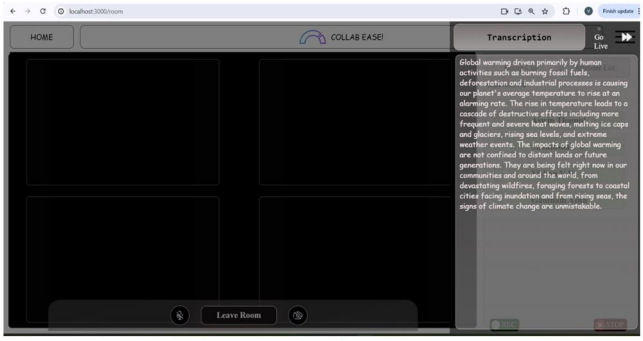


Fig. 7. Prepared Transcript

Fig. 8. Prepared Summary

V. FUTURE SCOPE

Our video conferencing platform has seen significant success, offering essential features like chatting, transcription, and summarization. However, to stay ahead in the rapidly evolving landscape of digital communication, we recognize the need to expand and enhance our capabilities. In response to this, we have identified several areas for future development.

Enhanced Transcription and Summarization Focusing Multiple Speakers: One key area of focus is to expand our transcription and summarization capabilities to accommodate multiple speakers. By implementing advanced algorithms and machine learning models, we aim to accurately transcribe and summarize conversations involving multiple participants. This enhancement will improve the overall efficiency and effectiveness of communication during video conferences.

Global Accessibility with Translation Services: To facilitate seamless communication across languages, we plan to incorporate real-time translation services into our platform. By leveraging state-of-the-art natural language processing techniques, we aim to provide instant translation of spoken and written content, enabling users to interact effortlessly regardless of their language proficiency. This enhancement will promote inclusivity and expand the reach of our platform to a global audience.

Real-Time Transcription: Incorporating real-time transcription capabilities is another crucial aspect of our future development strategy. Instead of relying solely on post-recorded audio transcription, we intend to implement real-time transcription algorithms that can accurately transcribe spoken content as it occurs during video conferences. This enhancement will enable users to access transcribed content in real-time, enhancing accessibility and improving the overall user experience.

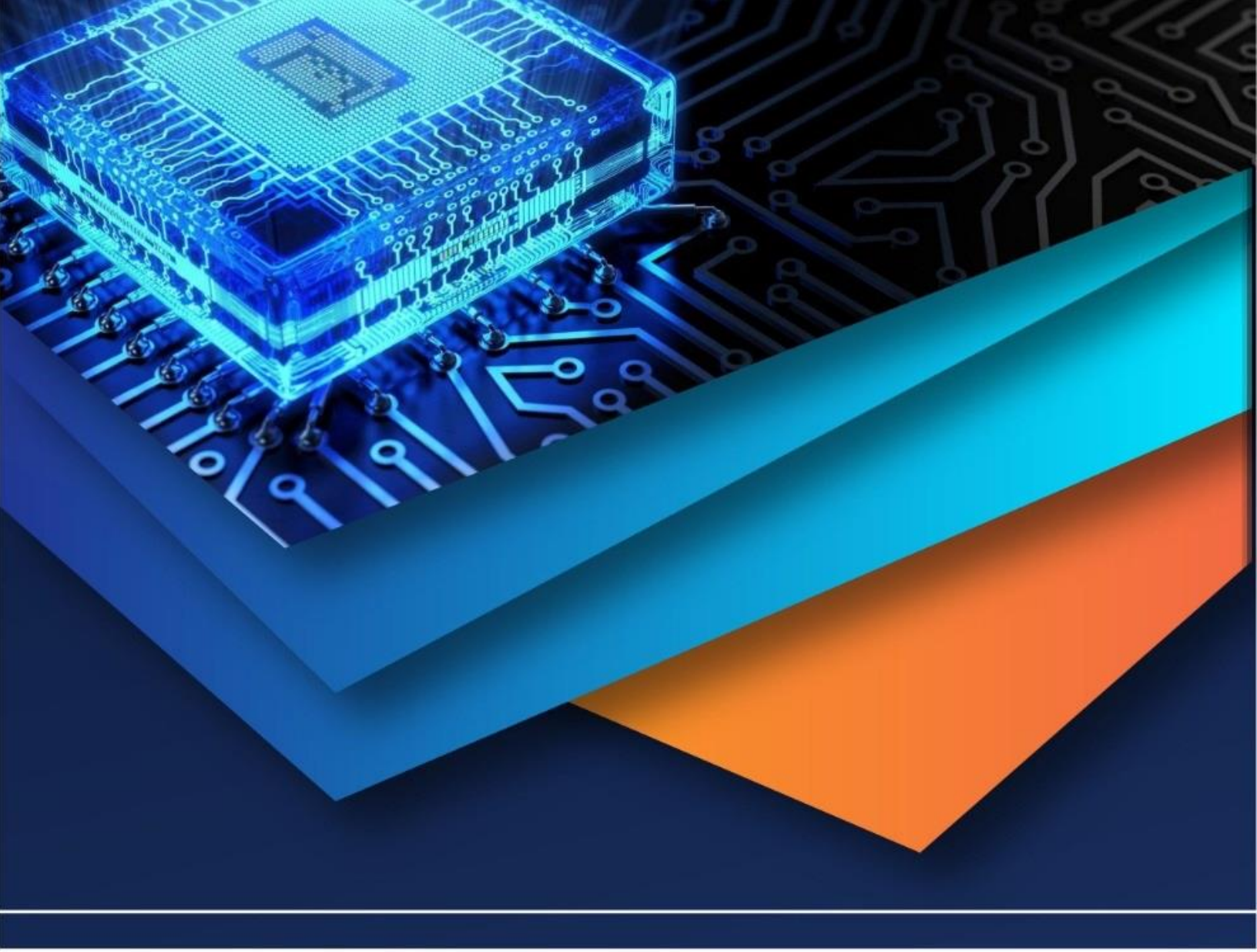
VI. CONCLUSION

Our project embarked on a journey to revolutionize video conferencing by integrating transcription and summarization functionalities using advanced transformer-based models. The primary objective was to develop a platform that not only facilitates real-time communication but also enhances productivity through transcription and summarization of conversations. Central to our approach was the utilization of encoder-decoder transformer architecture, a powerful framework known for its effectiveness in natural language processing tasks. We implemented two separate models: a transcription model responsible for converting speech inputs into text, and a summarization model tasked with condensing the transcribed text into concise summaries. The workflow of our system begins with the recording of user speech during video conferences.

This speech data is then fed into the transcription model, which accurately transcribes the spoken content into text. Subsequently, the transcribed text is passed to the summarization model, which generates concise summaries of the conversation's key points.The impact of our project extends beyond mere convenience; it fosters inclusivity by providing accessible communication tools for individuals with hearing impairments or language barriers. Moreover, the automated summarization feature enhances productivity by enabling users to quickly grasp the essence of discussions without the need to sift through lengthy transcripts.Looking ahead, we recognize the potential for further innovation and improvement. Future iterations of the platform may explore additional features such as real-time translation services, advanced speaker recognition, and customizable summarization options.

REFERENCES

[1] Dilawari, A., Khan, M.U.G., Saleem, S. and Shaikh, F.S. (2023). Neural attention model for abstractive text summarization using linguistic feature space. IEEE Access, 11, pp. 23557-23564.
[2] Zhang, J., Zhao, Y., Saleh, M. and Liu, P. (2020, November). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In International Conference on Machine Learning (pp. 11328-11339). PMLR.
[3] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning (pp. 28492-28518). PMLR.
[4] Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W. S., and Nguyen, A. (2019). Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4845–4854.
[5] Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., et al. (2015). Deep speech 2: end-to-end speech recognition in English and Mandarin. arXiv preprint arXiv:1512.02595.
[6] Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., et al. (2021). XLS-R: Self-supervised cross-lingual speech representation learning at scale. arXiv preprint arXiv:2111.09296.
[7] Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). Wav2vec 2.0: A framework for selfsupervised learning of speech representations. arXiv preprint arXiv:2006.11477.
[8] Baevski, A., Hsu, W.-N., Conneau, A., and Auli, M. (2021). Unsupervised speech recognition. Advances in Neural Information Processing Systems, 34, 27826–27839.
[9] Chen, G., Chai, S., Wang, G., Du, J., Zhang, W.-Q., Weng, C., Su, D., Povey, D., Trmal, J., Zhang, J., et al. (2021). Gigaspeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio. arXiv preprint arXiv:2106.06909.
[10] Zeyer, Albert, et al. (2019). A comparison of transformer and LSTM encoder-decoder models for ASR. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY