VIDEO CONFERENCING WITH CC FOR IMPAIRED HEARING USING NLP

Bhuvanesh V T[1], Mukesh N[2], Praveen S[3], Vaanathi S[4]

[1] *Student, Information Technology, Bannari Amman Institute of Technology, Tamil Nadu, India*
[2] *Student, Information Technology, Bannari Amman Institute of Technology, Tamil Nadu, India*
[3] *Student, Information Technology, Bannari Amman Institute of Technology, Tamil Nadu, India*
[4] *Assistant Professor, Artificial Intelligence and Data Science, Bannari Amman Institute of Technology, Tamil Nadu, India*

ABSTRACT

This paper proposes an innovative approach to address the communication challenges faced by individuals with impaired hearing during video conferencing. The purpose is to improve the accessibility and inclusivity of virtual communication platforms by integrating Natural Language Processing (NLP) techniques and Closed Captioning (CC) functionalities. The proposed system employs advanced NLP algorithms to enhance real-time speech recognition and language understanding, catering specifically to the nuances of diverse accents, speech patterns, and contextual cues. By leveraging these capabilities, the system aims to provide accurate and contextually relevant transcriptions of spoken content within the video conferencing environment. In addition to the NLP-driven transcription, the research focuses on the integration of dynamic Closed Captioning to generate synchronized, visually accessible text alongside the video feed. This feature ensures that participants with impaired hearing can follow the conversation more effectively, capturing not only spoken words but also non-verbal cues, enhancing the overall communication experience.

Keyword: *Closed Caption, real-time speech recognition, Natural Language Processing, Transcription.*

1. INTRODUCTION

In recent years, the technological landscape has witnessed an unprecedented surge in the development and utilization of communication platforms, transforming the way individuals connect and collaborate globally. Video conferencing, in particular, has become an integral part of everyday life, facilitating seamless communication across geographical boundaries. While these advancements have undeniably bridged gaps and improved accessibility for many, there exists a significant challenge in ensuring inclusivity for individuals with impaired hearing.

Hearing impairment affects millions of people worldwide, making it imperative for technology to evolve in a way that addresses their unique communication needs. In this context, this paper explores a ground breaking approach to video conferencing, leveraging Natural Language Processing (NLP), Web Real-Time Communication (WebRTC), Flask for backend development, and React for frontend design. The central goal is to enhance the communication experience for individuals with impaired hearing by integrating real-time closed captioning capabilities into video conferencing platforms.

1.1 Natural Language Processing

Natural Language Processing (NLP) is a cornerstone in our pursuit of inclusivity in video conferencing for individuals with impaired hearing. NLP algorithms, grounded in machine learning and artificial intelligence, are instrumental in enabling our system to dynamically recognize and transcribe spoken language during live video conferences. These algorithms go beyond simple voice recognition; they comprehend linguistic nuances, diverse

accents, and contextual intricacies to deliver accurate real-time closed captions. By harnessing the power of machine learning, our system adapts to a myriad of communication scenarios, ensuring an inclusive experience for users across various linguistic backgrounds.

1.2 Web real-time communication

WebRTC serves as the linchpin in our architecture, facilitating the seamless transmission of audio and video data in real-time. This section delves into the technical intricacies of WebRTC, emphasizing its role in establishing low-latency communication channels between users during video conferences. This technology ensures that the spoken words are transmitted swiftly and efficiently to the NLP engine for transcription. By providing a robust infrastructure for real-time audio and video streaming, WebRTC is pivotal in enhancing the responsiveness and overall effectiveness of our closed captioning system.

2. LITERATURE SURVEY

”Peer to Peer Multimedia Real-Time Communication System based on WebRTC Technology”. This paper proposes a practical implementation of a web peer-to-peer real-time communication system utilizing WebRTC technology, HTML5, and a Node.js server address. The key aspects highlighted include Real-Time Communication (RTC), WebRTC Technology, JavaScript APIs, Node.js Server, Multimedia Transmission, Security, and Compatibility. The focus is on enabling high-speed data transmission without the need for plugins.

“WebRTC Peer to Peer Learning Technology”, to explore the application of WebRTC in the context of a peer-to-peer learning system. It emphasizes the foundation role of WebRTC in real-time communication capabilities within web browsers and its application in facilitating direct connections between learners and experts. The advantages of the proposed system include real-time communication, dynamic interactions, public availability, decentralization, and a plugin-free mechanism. Technical aspects, such as using STUN and SDP for NAT traversal and session negotiation, are discussed.

“WebRTC Enabled Video Communication”, to focuses on the use of WebRTC for establishing and conducting communications in a web application. It praises WebRTC for simplifying communication by eliminating external dependencies and enabling one-click browser embedded calls. The advantages include secure voice and video calls using the Secure RTP protocol (SRTP), preventing eavesdropping, and ensuring advanced voice and video quality. The paper acknowledges the continuous development of WebRTC.

2.1 Web Socket

“Research and Implementation of WebRTC Signaling via Web Socket-based for Real-time Multimedia Communications”, Delves into the core functions of WebRTC, describing audio input/output, network connections, and data encoding/decoding for multimedia communication. It illustrates the WebRTC architecture, emphasizing libjingle for session management and the role of Web Socket-based signaling. The proposed signaling mechanism involves the exchange of Offer and Answer signals and the use of Session Description Protocol (SDP) for peer-to-peer connection parameters.

“Exploring WebRTC Technology for Enhanced Real-Time Services”, Provides insights into the two levels of WebRTC architecture: WebRTC C++ API and Web API. It explores the functions of Peer Connection, Media Streams, and Data Channels in the WebRTC API. The role of Peer Connection in direct peer-to-peer communication through signaling negotiation, and the importance of ICE, STUN, and TURN protocols in traversing NAT and firewalls are discussed. The paper introduces Media Streams as abstract representations of audio and video data streams, including Local and Remote Media Streams.

2.2 Natural Language Processing

Natural Language Processing: State of The Art, Current Trends and Challenges”, Offers a comprehensive overview of Natural Language Processing (NLP) applications, including machine translation, named entity recognition, optical character recognition, and part-of-speech tagging. Noteworthy is the discussion on cross-lingual event detection frameworks and modular systems for multilingual event extraction. The interdisciplinary nature of NLP is

emphasized, involving computer scientists, linguists, psychologists, and philosophers. The paper traces the historical evolution of NLP from its origins in machine translation research to modern developments, including computational grammar theory and the transition to statistical language processing in the 90s.

Natural Language Processing (almost) from Scratch". Introduces a unified neural network architecture and learning algorithm for various NLP tasks, such as part-of-speech tagging, chunking, named entity recognition, and semantic role labeling. The system achieves or surpasses state-of-the-art performance in these tasks without extensive task-specific feature engineering. Instead of relying on handcrafted features, the system learns internal representations from large amounts of mostly unlabeled training data. The authors demonstrate the creation of a high-performance tagging system with minimal computational resources.

3. METHODOLOGY

Real-time captioning converts spoken audio from the video call into text captions displayed on the screen simultaneously. This allows hearing impaired users as well as other audiences to follow the conversation by reading the captions instead of relying solely on audio. Real-time captioning offers an alternative access point to the spoken content, ensuring they can fully participate in the video call and grasp the information being shared.

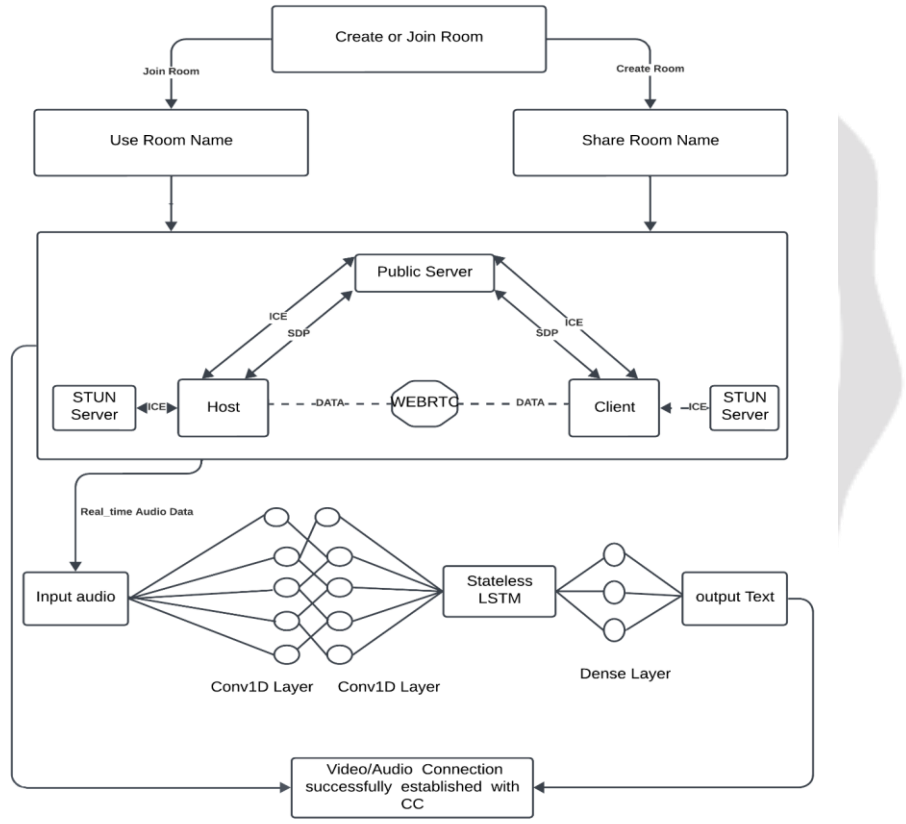

Fig - 1: Flow diagram of video conferencing

3.1 Enhanced Communication through NLP

NLP plays a crucial role in real-time captioning by enabling the system to understand and convert spoken language into text. Once words are recognized, NLP techniques are used for further processing. This involves Grammar Correction, Correcting any grammatical errors introduced during the speech recognition process and Punctuation Insertion, Adding appropriate punctuation marks to the generated captions for improved readability and comprehension.

NLP algorithms are used to operate efficiently, enabling real-time processing of the audio stream. This ensures minimal delay between spoken words and the corresponding captions appearing on the screen. Furthermore, NLP allows for customization of the captions (font size, color) for optimal readability based on individual needs.

WebRTC is an open-source technology that allows web browsers to directly establish real-time audio and video connections without requiring additional plugins or software installations. This makes it a powerful tool for building browser-based video conferencing applications. While WebRTC is a powerful tool for building a video conferencing application with its browser-based functionality, scalability, security, and network performance require careful consideration. To handle a large number of concurrent calls, the architecture needs to be designed for scalability.

4. LSTM NETWORK

Converting video to text involves extracting information from the visual content of the video and converting it into a textual representation. Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN), can be used for this task. LSTMs are particularly useful for sequential data like video frames. video-to-text conversion, also known as video captioning, is a technique that aims to bridge this gap by automatically generating textual descriptions or captions for video content that plays a crucial role in making video content more accessible, searchable, and understandable.

4.1 Sequential Data Handling

LSTMs are specifically designed to handle sequential data, making them well-suited for processing sequences of video frames. Each video frame can be treated as a time step in the sequence, allowing LSTMs to effectively analyse the temporal dynamics present in the video data.

4.2 Temporal Relationships

Video data inherently contains temporal relationships between consecutive frames. LSTMs are capable of capturing these temporal dependencies, allowing them to understand the progression of actions and events over time in a video. This is crucial for accurately captioning videos, as it enables the model to contextualize each frame within the broader sequence.

4.3 Memory Cells

LSTMs utilize memory cells to store and retrieve information over sequences. This ability to retain relevant information over long time intervals is essential for video captioning tasks, where contextual understanding often requires information from multiple past frames to generate accurate captions for the current frame.

4.4 Vanishing Gradient Problem

The vanishing gradient problem, common in traditional recurrent neural networks (RNNs), can hinder the training of models on long sequences of data. LSTMs address this issue by using specialized gating mechanisms that allow for better gradient flow through the network, mitigating the vanishing gradient problem and facilitating more effective training on video data.

4.5 Gating Mechanisms

LSTMs incorporate input, forget, and output gates, which control the flow of information within the network. These gating mechanisms enable LSTMs to selectively update and forget information based on the current context, enhancing their ability to model complex temporal relationships in video data.

4.6 Variable Sequence Lengths

Videos can vary in length, with different numbers of frames representing different durations of content. LSTMs are capable of processing sequences with variable lengths, providing flexibility in handling videos of different durations without the need for fixed-size inputs.

4.7 Feature Extraction

In video captioning tasks, LSTMs are often used in conjunction with convolutional neural networks (CNNs) for feature extraction. CNNs are adept at capturing spatial features from individual video frames, while LSTMs excel at

modelling temporal dynamics. By combining the strengths of both architectures, LSTM-CNN models can effectively extract spatial and temporal features from video data, leading to improved captioning performance..

5. CONCLUSIONS

This project successfully developed an online video conferencing application specifically designed to cater to users with hearing impairments. The core functionalities include: Real-time video conferencing, Speech-to-text conversion, Firebase storage, Agora SDK integration and Web Sockets.

The combined use of these technologies addresses a critical need within the video conferencing landscape, promoting inclusivity and accessibility for users with hearing impairments. Through the implementation of LSTM models, the project demonstrates the potential for improved speech recognition accuracy, especially in challenging audio environments. Firebase storage offers a secure and convenient solution for managing conference transcripts.

6. REFERENCES

[1]. Zinah Nayyef, Sarah Fairs Amer, Zena Hussain ,"Peer to Peer Multimedia Real-Time Communication System based on WebRTC Technology" in 2019

[2]. H. Fateh Ali Khan, A. Akash, R. Avinash, C. Lokesh in "WebRTC Peer to Peer Learning Technology" in 2018.

[3]. Naman Avasthi, Palakh Mignonne Jude, Rhea Thomas in "WebRTC Enabled Video Communication "in 2018.

[4]. Cui Jian, Zhuying Lin in "Research and Implementation of WebRTC Signaling via WebSocket-based for Real-time Multimedia Communications"in 2022.

[5]. Vasco Amaral, Solange Rito Lima, Telma Mota, Paulo Chainho in "Exploring WebRTC Technology for Enhanced Real-Time Services" 2022.

[6]. Ronan Collobert, Jason Weston, L´eon Bottou, Michael Karlen, Koray Kavukcuoglu, Pavel Kuksa ,"Natural Language Processing (almost) from Scratch" in 2011.

[7]. Basu Dev Shivahare, Arun Kumar Singh, Nilesh Uppal, Ashar Rizwan, Vangala Sri Vaathsav, Shashikant Suman , "Study of Natural Language Processing and its Recent Applications" in 2023.