Automated Caption Generation for Video Call with Language Translation

Sanjeeva Polepaka1*, Varikuppala Prashanth Kumar1, S. UmeshChandra1, Hema Nagendra Sri Krishna1, Gaurav Thakur2

1Department of CSE (AIML), GRIET, Hyderabad, Telangana State, India
2Uttaranchal Institute of Technology, Uttaranchal University, Dehradun, 248007, India

Abstract. In the modern era, virtual communication between individuals is common. Many people's lives have been made simpler in a number of circumstances by providing subtitles, generating automated captions for social media videos, and language translation from a source language to a targeted language. Both are included, which offers face-to-face translated captions during video conversations. React is used for application development. To send the data, socket programming is utilized. Context is understood and translated using Google translate API and speech recognition modules. With OpenAI and Whisper, captions are generated. This paper will directly create the translated user voice rather than translating the text and creating subtitles.

1 Introduction

In the modern world, communication is essential in day-to-day life, especially in an interconnected and globally networked society. We now communicate almost exclusively through video conversations, which provide a virtually in-person experience that enables us to convey not just our voices but also our expressions. Effective communication is essential, especially when people from various cultures and backgrounds must collaborate. Multilingual caption generating technology in video conversations is developing as a feasible solution to this problem. With the help of this technology, captions are created automatically in several languages during video conversations, bridging the language barrier between speakers of other languages. Throughout the call, the system will interpret and translate spoken words in real-time using machine learning techniques, presenting the captions on the screen.

It will be practical for individuals from all around the world since this real-time, multilingual caption creation system for video chats will be made to accommodate several languages and dialects. With the use of this technology, video chats might become more linguistically diverse and inclusive. This method can improve greater intercultural communication and understanding, improves better collaboration and cooperation in a

* Corresponding author: sanjeeva1690@grietcollege.com
variety of industries including business, education, healthcare, and more. Technology for generating automated captions in many languages during video conversations is an innovative breakthrough that has many advantages. Real-time captioning enables improved communication and comprehension amongst speakers of various languages. In circumstances like international business meetings or remote team cooperation, where language limitations may be problematic, this technology is very helpful.

Machine learning techniques will be used to develop a system that can translate and transcribe spoken language in real-time. Artificial intelligence known as machine learning enables computers to learn from experience and advance without explicit programming. As a result, the system can adjust to diverse languages and dialects, which over time will improve its accuracy and effectiveness. Those with hearing difficulties or those who prefer to read captions instead of listening to the audio can also benefit from the real-time multilingual caption creation system for video conversations. It can also be helpful for language learners who want to use video calls to develop their language abilities by reading the subtitles. The use of this technology can benefit a number of industries, including healthcare, where it can improve communication between patients and healthcare professionals who speak different languages. It may also be helpful in the classroom, where it can make lectures and conversations easier for students to grasp, particularly in schools with a varied or multinational student body. Overall, the method for creating multilingual captions for video conferences in real-time is a promising innovation that has the potential to enhance interaction and cooperation between individuals from various backgrounds and cultures.

2 Literature survey

Xavier suggested Audio-a-test alignment for speech recognition with very limited resources [1]. The objective was to build speech recognition systems for new language with very limited resources. The speech recognizer was well trained from scratch using recorded audio aligned with text transcripts. The parameters adopted to evaluate the system proposed by looking both at the accuracy of the alignment at word level and its usability to train our ASR system.

Kasiprasad worked on accent detection of telugu speech using prosodic and dormant features [2]. The objective was to recognize speech other than the system was protracted by using Nearest Neighborhood Classifier. Accuracy of the model obtained was 72% for the best recognition. The accent corresponding to the least Euclidian distance was identified as the absolute accent. The study concluded that speech recognition for a language in different dialect can be achieved.

Aditya focus on Automatic Subtitle Generation for videos [3] and [13]. The aim is to overcome issues in incorporating subtitles into videos, generally downloaded to third party source and adding manually to media player which requires quality internet and readily available subtitles. The comparative study among speech recognition engines, Bathaloori performed Classification of Analysed Text in Speech Recognition using KNN-LSTM is comparison with convolutional Neural Network to 8 improve Precision for Identification of key words [4]. The aim is to classify the features from language translation in speech recognition from English to Telugu using RNTV - LSTM comparison with CNN. The model precision & Accuracy were performed with dataset alexa additional with English-Telugu of size 8166 sentences. They concluded that RNN- LSTM from the dataset speech recognition, feature Telugu-id produce accuracy 93% precision 68.40% which was comparatively higher than CNN accuracy 66.11% and precision is 61.90%. Suryakanthi and Sherma done translate discourse, keeping the context in mind such that actual meaning doesn't change [5]. To evaluate system performance in translating discourse formed using
connectives a test suite of around 400 sentences was developed. It is concluded that the results were more accurate than compared with results obtained by Google translator. Vaishnavi developed an Android-based application programming interface (API) using Firebase for language translation [6].

Junkun provided an alternative solution for improving Speech-to-Text translation via Spectrogram Reconstruction [7]. In this model the proposed spectrogram reconstruction performs self-supervised training without using transcription whereas the other models like ASR-ST involves multitasking which is transcription and translation. Sachi presented a paper on Machine Translation and various methods used for translating one language into another language [8], [12]. They proposed that, to develop a Machine Translation process it requires five major features to work proper, they are: 1. Morphological Analyser 2. Parts of speech tagging 3. Chunk 4. Parsing 5. Word Sense Disambiguation They aim to increase the performance of Chunk for Hindi language with the existing approaches by using maximum entropy models.

Vinnarasu developed a model for the text summarization in which input is given through speech [9] and [11]. The objective is to convert the speech to text and deliver the summarized text as a output to the users. This model involves two modules which is speech recognition and text summarization. Natural Language Processing (NLP) is used to extract the features from the speech which has a value to convert it to the text. The speech from the source is recorded using a microphone and the feature is extracted in text format using Google Application Programming Interface (API). Tie-Yan developed a Simultaneous Speech to text translation to reduce the translation delay [10] and [14]. The objective of their model is to translate the speech to text simultaneously, their simultaneous 10 translation works on wait-k strategy, and Connectionist Temporal Classification (CTC) loss for segmentation.

3 Problem statement and objectives

3.1 Problem statement

The problem statement is to develop a video call application that enables real-time translation of spoken English to Telugu and vice versa and displays the translated text as captions. This will facilitate communication between English and Telugu speakers who are not fluent in each other's languages and improve the overall user experience of the video call.

3.2 Objectives

- To create a system that can provide automated multilingual captions for video calls in real-time.
- To make sure the system can support a variety of languages and dialects, making it accessible to users worldwide.
- To design a user-friendly interface that can show many languages of captions during video calls.

4 Proposed method

The architecture diagram shows the main components of the system and their interactions with each other. In this case, the system consists of four main components: the user interface module, the Video SDK cloud service, the speech recognition module, and the
translation module. The user interface module is developed using React, a popular JavaScript library for building user interfaces. The user interface module provides the user interface for the video call application and handles user interactions such as starting and ending video calls. The Video SDK cloud service is used to provide video call functionality. The translation module is responsible for translating the text data from one language to another.

4.1 Modules in the proposed work

4.1.1 Module 1: Video call application

In this module, we built an application to provide virtual conversations among users. We adopt technologies like React in order to develop the application. To construct this, we must have a computer system provided with one IDE. Then we need to set up some configuration in order to develop applications. This can be done in two ways, the first one saves development time, removes difficulties dealing with many dependencies and system configurations.

4.1.2 Module 2: Audio extraction

Audio to text transformation is a well-liked application of machine learning and Natural Language Processing (NLP). This are the steps involved in speech to text:

- Pre-processing: Before being transformed into a digital signal that a computer can process, the audio signal has to be pre-processed to remove any noise or unwanted sounds. Techniques like filtration, normalization, and feature extraction may be used in this.

- Acoustic Modelling: An acoustic model, a deep learning model trained to recognize the sounds and phonemes of a specific language, is then fed the pre-processed audio signal. The fundamental units of spoken language, known as phonemes or sub-words, are used by the acoustic model to map the audio signal.

- Language Modelling: The next step is to produce the corresponding words and sentences after the acoustic model generated an order of phonemes or sub-word units. This is accomplished by using a language model, a deep learning model trained on sizable textual datasets to forecast the most probable word order given the input phonemes or sub-word units. To produce the most likely transcription of the spoken audio, the language model considers elements like grammar, vocabulary, and context.

- Post-processing: The language model's output is finally post-processed to eliminate any errors or discrepancies and format the text so that it is comprehensible and readable. Techniques like spell-checking, grammar-checking, and punctuation-checking can be used for this.

4.1.3 Module 3: Translation

Module 3 builds upon the text generated in Module 2 as input.

- Caption Generation: Once the speech has been transcribed into text, the next step is to generate captions that can be displayed on the screen in real-time. This can be done using NLP techniques such as named entity recognition, sentiment analysis, and part-of-speech tagging to identify the key elements of the text and generate automated captions that accurately reflect the content of the conversation.
• **Language Translation:** Once the captions have been generated in the original language, the next step is to translate them into other languages for multilingual communication. This can be done using machine learning models that are trained on large datasets of text in different languages. The models use statistical analysis and NLP techniques to identify the meaning and context of the text and translate it accurately into the target language.

• **Displaying Captions:** Finally, the captions and translations can be displayed on the screen in real-time for all participants in the video call to see. This can help to improve communication and understanding between people who speak different languages and make video calls more accessible to people with hearing impairments.

• **Text Gets Translated into Targeted Language:** The next step is to convert the generated captions into the intended language for the video call’s spoken language. To do this, machine learning models that have been trained on large text datasets in both the original language and the intended language are used.

• **Translated Text is provided as Captions for the Users:** After being produced and transcribed into the desired language, the captions can then be made available to users as subtitles. The subtitles can be added as closed captions to the video recording or shown on the screen in real-time. Users who don't speak the actual language can still read and comprehend the dialogue thanks to this.

5 Results and discussions

In this paper, our main aim is translation of languages from one source to targeted languages. The following datasets are utilized to train the neural network models. There is a dataset called ‘Hindi_English_Truncated_Corpus.csv’ available in Kaggle. It consists of fifty thousand of English sentences and their corresponding Hindi transcripts. Shape of the dataset is like this (50000,2), first row of datasets generally describes source from where this sentence is took and English sentence from the source mentioned formerly, Hindi sentences that matches the exact meaning of English sentence in the same row. They are separated by delimiter ‘,’. Example: ted, politicians do not have permission to do what needs to be done,"राजनीतिज्ञोंके पासजोकायकराचाहए, वहकरनोके अनुमानित नहीं है।" This is just one row in the above dataset so ted the source from where this this is collected and had its corresponding Hindi sentence which matches meaning to the English sentence. This bulk dataset is split into train and test datasets to train the model and test the performance of our model. Training dataset consists of 40000 English sentences with their corresponding Hindi translated sentence and test dataset of 10000 English sentences with their corresponding Hindi transcripts.

![Fig. 1. English to hindi dataset.](image-url)
5.1 English to telugu translation

There is no dataset available in Kaggle for this purpose but with the help of Google translation API we generated english_to_telugu.csv dataset. We took all the English sentences from above dataset and feed these sentences to google API, which in return gives translated Telugu sentence. Map the English sentences with its corresponding translated Telugu sentences. Same as English to Hindi translation dataset maintain three columns one is source; second one is English Sentence and third one is Telugu Sentence. Separate these three columns by comma. Now our dataset is ready, which consists of fifty thousands of English sentences and their corresponding Telugu sentences. Shape of dataset is (50000,2), first column source tells from where this sentence is collected, second column English sentence have a English sentence and third column has corresponding Telugu sentence. Example:“Metapoliticians do not have permission to do what needs to be done.,మెటాపలిటికియన్‌లు ఎందుకు కావచ్చు చేసాలంటి పరిచితి విషయం ఉంచడానికి సంవేదన అవలంబించలేయాలి.” This is the single instance of the dataset. So there are 49999 more such instances in our dataset. Here also we use pandas python library to perform text pre-processing and see how data is.

Fig. 2.English to telugu dataset.

Above entire dataset is split into training and testing datasets of sizes respectively 40000 English sentences with their corresponding translated Telugu sentence and 10000 English sentences with their corresponding translated Telugu sentence.

5.2 Video call

The video call feature can improve communication and collaboration among people who are in different locations. It can also provide a more convenient and cost-effective way for people to receive medical consultations, counselling sessions, or educational classes without having to travel long distances. This feature can also have significant benefits for businesses that have remote teams or serve customers from different regions. One of the main advantages of video calls is that they enable face-to-face communication, which can help to build stronger relationships and enhance understanding. For example, in educational settings, teachers can use video calls to interact with their students and deliver lectures remotely. This can help to provide equal educational opportunities for students who live in remote areas or cannot attend school in person. In addition, video calls can be used for telemedicine, where doctors can diagnose and treat patients remotely. This can be particularly helpful for patients who live in rural or underserved areas, where healthcare access may be limited.
5.3 Home screen

The home screen of the video call application provides users with options to permit audio and video access permissions. This feature is important because it enables the participants to join the meeting with the correct settings that ensure they can hear and see each other. The home screen also offers two options for users: create a meeting or Interactive Live Streaming. You can observe figure 5 to understand the above description better and how our home screen appears.

![Home screen of video call application.](image)

5.4 Meeting creation process

When the user clicks on the create button, a unique meeting ID is generated. The meeting ID is important because it serves as a link for other users to join the meeting. The host can share the meeting ID with other participants via email or other messaging platforms. The meeting creation process is straightforward and easy to follow. Once the user clicks on the create button, they are prompted to enter their name and select whether they want to enable audio or video. The meeting ID is then generated and the user can copy and share the ID with others who wish to join the meeting.

5.5 Join screen

The join screen is an essential part of the video call application as it enables participants to join a meeting that has already been created by the host. When participants receive the meeting ID from the host, they can navigate to the join screen and enter their name and the
meeting ID to join the meeting. The join screen is designed to be simple and user-friendly. Participants are prompted to enter their name and the meeting ID, which they received from the host. Once the user enters the required information, they can click on the "Join Meeting" button to enter the meeting room. Before joining the meeting, participants are given the option to check their audio and video quality. This feature is essential because it ensures that participants have the correct settings before joining the meeting, which helps to avoid technical difficulties during the meeting. This feature provides participants with more control over their audio and video settings, which can help to create a more comfortable and secure meeting environment.

In summary, the join screen is a crucial part of the video call application, which enables participants to join a meeting that has already been created by the host. The join screen is designed to be simple and user-friendly, and it offers participants the option to check their audio and video quality before joining the meeting. The join screen also allows participants to select their audio and video preferences, which can help to create a more comfortable and secure meeting environment.

5.6 Meeting room

The meeting room provides participants with a simple and user-friendly interface. Participants can check their audio and video quality before joining the meeting, which is an essential feature that ensures a smooth and uninterrupted meeting experience. The interface also allows users to turn off and turn on the microphone and webcam. By default, the webcam is turned on, but participants can turn it off if they do not wish to use it.

In conclusion, the use of React frameworks has enabled the development of a user-friendly front-end UI for a video call application. The home screen provides users with the option to permit audio and video access permissions and offers two options for users: create a meeting or Interactive Live Streaming. The meeting creation process is straightforward and easy to follow, and the meeting room provides participants with a simple and user-friendly interface. Overall, this UI is a significant step towards improving the user experience of video call applications.
5.7 Audio to text

The audio to text feature can enhance accessibility for people who are deaf or hard of hearing, as well as for those who prefer to read rather than listen to audio content. This feature can also help researchers, journalists, or content creators to transcribe interviews, speeches, or podcasts more efficiently and accurately. One of the main advantages of audio to text is that it can make audio content more accessible to people with hearing impairments. For example, in educational settings, teachers can use this feature to provide transcripts of their lectures for students who are deaf or hard of hearing. This can help to provide equal educational opportunities for all students. In addition, audio to text can be used for transcription purposes in various industries, such as journalism and content creation. This can help to save time and improve accuracy, as human transcription can be time-consuming and error prone.
5.8 Text translation

The text translation feature can facilitate language learning and cultural exchange by enabling users to communicate with people who speak different languages. This feature can also support businesses that operate in multiple countries or serve customers from diverse backgrounds. One of the main advantages of text translation is that it can help to break down language barriers and promote cross-cultural communication. For example, students can use this feature to translate their assignments into different languages or communicate with their peers from different countries. In addition, businesses can use text translation to expand their reach and serve customers from diverse backgrounds. This can help to increase revenue and improve customer satisfaction.

![Fig. 8. English caption in video conference.](image)

Although the caption feature was not completed, it remains a valuable potential feature that could improve accessibility for people who are deaf or hard of hearing. It could also enhance the user experience for people who prefer to read captions while watching videos. One of the main advantages of captions is that they can make videos more accessible to people with hearing impairments. For example, in educational settings, teachers can use this feature to provide captions for their videos for students who are deaf or hard of hearing. This can help to provide equal educational opportunities for all students. In addition, captions can be used to improve the user experience for people who prefer to read captions while watching videos. This can help to provide a more immersive and engaging experience for viewers.

![Fig. 9. English to telugu translation output from trained model](image)
6 Conclusion and future enhancements

The automated Multilingual Caption Generation in Video Calls technology is important because it encourages diversity as well as effective communication between individuals from different cultures and backgrounds. Language limitations can make it difficult for people to express their ideas and thoughts clearly, which can lead to misunderstandings. This solution eliminates language barriers and facilitates efficient communication, enhancing cooperation and production. It is simple to use and open to anyone with different levels of technical proficiency. This is an important aspect since it guarantees that even those without technical knowledge may use the system and learn from it. Additionally, the device improves 48 communications for those with hearing problems by delivering multilingual captions in real-time during video chats.

The approach that was used in this paper includes the gathering and preparation of data, the creation of a machine learning model, and the implementation of a system. In order to convert spoken language into text and convert it into the required language, the system makes use of speech recognition and machine translation models. After the development and improvement of the machine learning models, the automated multilingual caption generating system was implemented in reality.

References