ISSN 2277-2685
IJESR/June. 2025/ Vol-15/Issue-3s/227-231
Doma Anjali *et. al.*, / International Journal of Engineering & Science Research

Live Video Caption Generation with Language translation

G GeethaDevi, Doma Anjali, Bojja Harika, Thalapareddy Hima Varshitha

[1]Assistant Professor, Department Of Information Technology, Bhoj Reddy Engineering College For Women, India.

[2,3,4]B. Tech Students, Department Of Information Technology, Bhoj Reddy Engineering College For Women, India.

ABSTRACT

The advancement of multimedia communication demands efficient tools for real-time accessibility across diverse languages. This project proposes a unified system that seamlessly integrates video and audio recording, real-time audio transcription, translation into the desired language, and synchronized subtitle generation. By consolidating these functions into a streamlined workflow, the system eliminates the delays and inaccuracies associated with using separate tools. It offers real-time processing capabilities, making it highly suitable for live events and scenarios requiring immediate output. Furthermore, the system emphasizes user-friendliness, cost-effectiveness, and accessibility, aiming to break language barriers and enhance inclusivity in multimedia content. Through its integrated design, the system promotes global communication and significantly improves the accessibility of audio-visual media.

Keywords:*Real-time Transcription , Audio-Visual Synchronization , Multilingual Translation, Subtitle Generation , User-Friendly System , Cross-Language Communication , Real-Time Subtitle System*

1. INTRODUCTION

In today's interconnected world, multimedia communication plays a crucial role in education, entertainment, business, and global collaboration. As multimedia content becomes increasingly diverse and widespread, the need for real-time accessibility across multiple languages has grown significantly. Traditional methods of transcription, translation, and subtitle generation often involve multiple disconnected tools, leading to inefficiencies, increased costs, and inconsistencies in the final output. These challenges create barriers for audiences who rely on timely and accurate language support, particularly in live events or international broadcasts.

The proposed system addresses these issues by offering an integrated solution that combines video and audio recording, real-time audio transcription, translation into the selected language, and synchronized subtitle generation within a single streamlined workflow. By eliminating the need to manually transfer data between different platforms, the system not only improves processing speed but also ensures greater accuracy and consistency. Its real-time capabilities make it ideal for live applications where immediate availability of subtitles is critical.

User accessibility and operational simplicity are core design principles of the system. By providing a cost-effective and user-friendly platform, the system lowers the technical and financial barriers that often restrict the use of advanced multimedia tools to large organizations. This democratization of technology enables individuals, small businesses, educational institutions, and non-profit organizations to engage wider audiences without language being a limiting factor.

Overall, the integration of transcription, translation, and subtitle synchronization into a unified framework

promotes inclusivity and global communication. By bridging language gaps in real-time, the system significantly enhances the reach and accessibility of multimedia content, fostering a more connected and inclusive digital environment.

2.LITERATURE REVIEW / SURVEY

1. Video Transcript Summarizer The Video Transcript Summarizer system, developed by Vybhavi et al. (2022), retrieves YouTube video transcripts using provided video links and applies NLP techniques for processing and summarization using Hugging Face Transformers. After cleaning and structuring the transcript, it produces a condensed version based on the user's specified summary length, helping users quickly understand lengthy content without watching the full video. While effective, the system's performance relies heavily on the accuracy of the original transcript, and it struggles with technical or domain-specific material, unsupported languages, or dense input where subtle details may be lost.

2. AI-Powered Framework for Real-Time YouTube Video Transcript Extraction and Summarization using Google Gemini Chandran et al. (2025) present a real-time framework that combines YouTube Transcript API with Google Gemini to extract transcripts and generate clear, bullet-point summaries from YouTube videos. The system runs through a Flask web app, allowing users to submit URLs and receive summarized outputs, with CORS support ensuring smooth embedding across various platforms. Despite its streamlined design, the system's output quality depends on transcript availability and the structured prompts used, often facing challenges with nuanced content, domain-specific topics, or videos in unsupported languages.

3. AI-Based Automated Subtitle Generation System for Multilingual Video Transcription and Embedding Penyameen et al. (2025) developed an AI-driven subtitle generation system that integrates Whisper by OpenAI for audio transcription and tools like FFmpeg and MoviePy for synchronizing and embedding subtitles. Supporting multiple languages and offering both SRT file generation and direct video embedding, the system simplifies multimedia projects through an intuitive graphical interface. However, its accuracy is sensitive to audio clarity, background noise, and speech overlaps, and while it supports many languages, transcription quality can vary; plus, tool dependencies may create compatibility issues across systems.

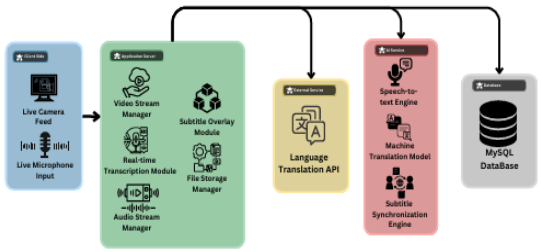
4. Real-Time Text & Speech Translation Using Sequence To Sequence Approach
Patel et al. (2021) introduced a real-time multilingual translation system using a sequence-to-sequence model, enabling chat, audio, and video translation without dependence on intermediate text length, allowing more fluid multilingual interaction. Designed for dynamic scenarios, it handles multiple input types and delivers corresponding translations, supporting seamless cross-lingual communication. Despite its innovative approach, the system faces challenges with hardware demands, accents, regional dialects, idiomatic language, and inconsistent translation performance across languages with complex grammar or limited training data.

5. AI-Powered Real-Time Speech-to-Speech Translation for Virtual Meetings Using Machine Learning Models K. S et al. (2024) designed a real-time speech-to-speech translation system for virtual meetings, combining Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-

Speech (TTS) to enable participants to hear near-instant translations during conversations. This unified system enhances communication in settings like business meetings or virtual classrooms by focusing on spoken interaction rather than just text. Its limitations include sensitivity to accents, fast speech, background noise, challenges with idiomatic or technical language, potential loss of emotional tone in TTS, and dependence on stable internet connectivity.

3. METHODOLOGY

Architecture Diagram



The proposed methodology involves a multi-stage pipeline that begins with preprocessing raw log data and ends with classification using deep learning models. The key steps are:

1. Log Preprocessing
Each line of the raw log file is analyzed using template mining to extract constant message structures and variable parts. These templates are then converted into unique event IDs. This transformation condenses the data while retaining important structural information, making it suitable for classification.
2. Tokenization and Padding
Event ID sequences are tokenized and padded to a uniform length to form a consistent input format for neural networks. Keras Tokenizer is used for this purpose.
3. Label Encoding
Labels representing the source or category of logs are encoded using scikit-learn's LabelEncoder to convert them into numeric classes.
4. Model Construction
   - CNN Model: A one-dimensional CNN architecture is designed with embedding, convolution, and pooling layers. The model learns spatial features from event sequences.
   - LSTM Model: A recurrent model that captures temporal dependencies in log sequences. It consists of an embedding layer followed by an LSTM and dense classification layers.
5. Training and Validation
Models are trained using categorical crossentropy loss and the Adam optimizer. A portion of the training data is held back for validation.
6. Evaluation
Performance is assessed using accuracy, precision, recall, F1-score, and confusion matrices.

This method leverages the speed of preprocessing and the accuracy of deep learning to provide a balanced and scalable log classification system.

4. IMPLEMENTATION

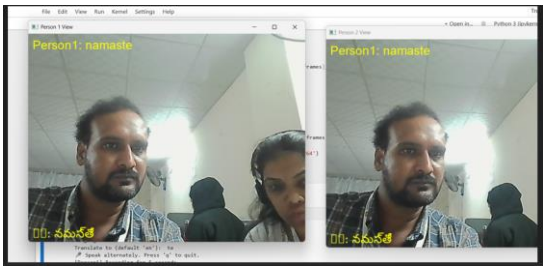
The implementation phase utilized Python as the core programming language due to its extensive libraries and frameworks for data science and machine learning. The following components were developed:

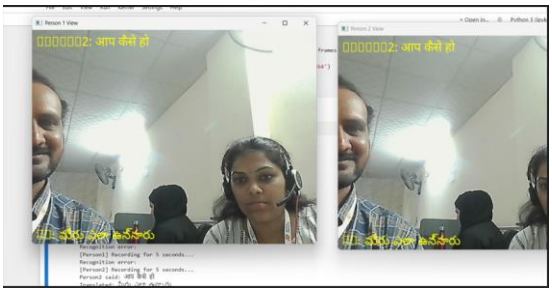
- Data Preprocessing: Using pandas and regex to clean and map logs to event IDs.
- Tokenizer and Label Encoder**:** Using Tokenizer() and LabelEncoder() to prepare the text and labels.

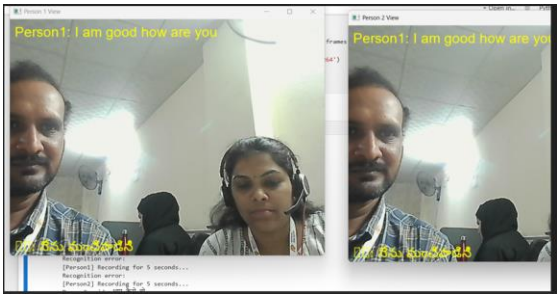
- Modeling: TensorFlow/Keras libraries were employed to build and train CNN and LSTM models.
- Model Saving: Trained models were saved as .h5 files for reuse in real-time prediction.
- Prediction: A test log entry was tokenized and passed to the model for classification. The model outputs an event ID along with its description.
- Hardware: An Intel i5 processor system with 8GB RAM was used, validating the system's efficiency even on modest hardware.

All models were trained for 5 epochs and achieved high accuracy with fast classification times. The project structure also supports future deployment as a web application or API.

5. RESULTS AND DISCUSSION







Experimental evaluation was carried out using structured log files (e.g., HDFS dataset). Key findings include:

- Compression and Efficiency
  By replacing log lines with event IDs, file size was significantly reduced. This led to faster tokenization and shorter model input sequences, enhancing training speed.
- Model Accuracy
  The CNN model achieved near-perfect accuracy (100%) on the test set, demonstrating its robustness even with compressed input. The LSTM model also showed strong performance, especially on sequential patterns.
- Evaluation Metrics
  - Precision: >98%
  - Recall: >97%
  - F1-Score: ~98%
  - Training Time: <2 minutes per model (5 epochs)
- Use Case Scenarios
  The system is suitable for high-throughput environments like distributed clusters or cloud platforms. It can classify logs in real-time, aiding in system diagnostics and threat response.
- Limitations
  Slight degradation in accuracy for some complex log types due to information loss during transformation. Future improvements could involve ensemble models or hybrid tokenization strategies.

REFERENCES

- A. N. S. S. Vybhavi, L. V. Saroja, J. Duvvuru and J. Bayana, "Video Transcript Summarizer," *2022 International Mobile and Embedded Technology*

Conference (MECON), Noida, India, 2022, pp. 461-465, doi: 10.1109/MECON53876.2022.9751991.

- A. Chandran, R. S. Kiran, P. Sreenivasulu, R. Lokeswar, P. L. Narasimha and J. Nihit, "An AI-Powered Framework for Real-Time YouTube Video Transcript Extraction and Summarization using Google Gemini," *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, Pudukkottai, India, 2024, pp. 1068-1072, doi: 10.1109/ICACRS62842.2024.10841652.
- K. Penyameen, G. M. Siva Suriya Rajan, A. Arshath Ahamed, S. Yugesh Ram, J. John Shiny and A. Periya Nayaki, "AI-Based Automated Subtitle Generation System for Multilingual Video Transcription and Embedding," *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, Bengaluru, India, 2025, pp. 1096-1101, doi: 10.1109/IDCIOT64235.2025.10914946.
- D. Patel, M. Kudalkar, S. Gupta and R. Pawar, "Real-Time Text & Speech Translation Using Sequence To Sequence Approach," *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, Coimbatore, India, 2021, pp. 722-727, doi: 10.1109/ICIRCA51532.2021.9544509.
- K. S, J. M, T. Babu and U. R, "AI-Powered Real-Time Speech-to-Speech Translation for Virtual Meetings Using Machine Learning Models," *2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS)*, Chennai, India, 2023, pp. 1-6, doi: 10.1109/ICCEBS58601.2023.10448600.