IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CLOSED CAPTIONING USING AI

Eshani Ghavate, Vaishnavi Deshmukh, Payal Thool, Prof. Chetana Patil
Student, Student, Student, Project Guide
Computer Engineering,
Dhole Patil College of Engineering, Pune, India

Abstract: Closed captioning is essential for making video content accessible to people with hearing impairments, but manual transcription can be time-consuming and costly. AI and Machine Learning (ML) have introduced automated solutions that enhance the efficiency and accuracy of closed captioning. Using deep learning models for speech recognition and natural language processing (NLP), AI systems can generate precise captions in real time. This approach addresses challenges like noisy audio and varying accents, offering a scalable solution that improves accessibility and user experience across diverse multimedia platforms.

Index Terms **-** Closed Captioning, OpenAI, Natural language Processing (NLP), Machine learning (ML), Whisper

I. INTRODUCTION

Closed captioning using AI and machine learning has revolutionized the accessibility of video content. By leveraging advanced algorithms, these technologies can automatically transcribe spoken words into text, making media more inclusive for individuals who are deaf or hard of hearing. Machine learning models are trained on vast datasets of audio and speech patterns, enabling them to recognize different accents, dialects, and even contextual nuances. This not only improves accuracy but also enhances the speed of caption generation. Additionally, AI can adapt to various environments, distinguishing between dialogue and background noise.

As a result, closed captioning is becoming more seamless and integrated, providing real-time solutions across platforms, from social media to streaming services. Ultimately, this innovation promotes a more equitable digital landscape, ensuring that everyone can engage with content fully.

II. CHALLENGES

Creating effective closed captions presents numerous challenges. Achieving high accuracy is difficult, as both automated speech recognition (ASR) struggles with accents, noise, and nuance, while human captioners face errors under pressure. Synchronization, especially for live content, is a major hurdle, often resulting in delays where captions lag behind the spoken words. Technical issues like format compatibility and varying platform support add complexity. Accurately conveying non-speech sounds and speaker identification further complicates the process. Finally, the significant cost and time required for quality captioning, combined with inconsistent user experiences regarding readability and placement, make widespread, high-quality captioning a persistent challenge.

III. PROPOSED WORK

The proposed Automated Closed Captioning System integrates advanced Artificial Intelligence (AI) and Machine Learning (ML) techniques to generate accurate and time-synchronized text transcriptions from audio sources, enhancing accessibility for various media content. The methodology consists of two key approaches:

Surveyed Approach: Deep Learning-Based Automatic Speech Recognition (ASR)

This approach forms the foundation of modern closed captioning systems and is extensively explored in speech processing research. It primarily employs deep learning models to convert spoken language into text. Key architectures historically include Recurrent Neural Networks (RNNs) often paired with Connectionist Temporal Classification (CTC), and more recently, sophisticated Transformer-based models (like those used in Whisper or Google's Speech-to-Text). These models are trained on massive datasets to handle diverse accents, languages, and acoustic conditions. While capable of achieving high accuracy (low Word Error Rate) on benchmark datasets, their real-time deployment effectiveness is often constrained by model size, computational requirements (CPU/GPU), and latency demands.

Applied Approach: Real-Time Captioning using Pre-trained Transformer Models

The implemented system focuses on leveraging a state-of-the-art, pre-trained Transformer-based ASR model (e.g., variants of OpenAI's Whisper or similar large-scale models) for practical caption generation. This real-time or near-real-time approach involves capturing or processing an audio stream (from live input or file), segmenting it appropriately, and feeding it to the ASR engine. The chosen model performs robust speech-to-text conversion, often inherently handling punctuation and some level of noise resilience due to its extensive training. The system then post-processes the transcribed text segments, assigning accurate timestamps to synchronize them with the corresponding audio/video source. The final output is formatted into standard caption files (like SRT or VTT), enabling immediate display alongside the media content. The real-time capability and latency directly depend on the specific model variant used and the underlying hardware infrastructure.

IV. Related Work

Reference	Methodology	Key Findings
Chen & Lee (2024)	Transformer-based ASR with noise robustness	Significantly lowered Word Error Rate (WER) on datasets with high background noise.
Kumar et al. (2023)	Low-Latency Streaming ASR + Punctuation Mode	Achieved reduced delay suitable for live broadcasts with improved readability.
Davis et al. (2023)	Multimodal Audio Event Detection & ASR Fusion	Enhanced caption richness by integrating accurate non-speech sound descriptions.
Alvarez et al. (2022)	Hybrid Human-AI Workflow with Correction Interface	Reduced post-editing time for professional captioners while maintaining high accuracy.
Patel & Gupta	Cross-Modal Alignment using Visual Speech Cues	Improved caption synchronization accuracy, especially in dynamic speaker scenarios.
Ivanova et al. (2021)	Domain-Specific ASR Model Adaptation	Increased accuracy for content with specialized jargon (e.g., medical, technical)

V. Research Methodology

Data: Gather diverse audio (live/pre-recorded) and create manual transcripts with timestamps for training/evaluation. Preprocess audio (noise reduction, segmentation).

ASR: Develop/use an Automatic Speech Recognition (ASR) model (pre-trained or custom) to transcribe audio. Evaluate accuracy (Word Error Rate) and optimize.

Captioning: Segment ASR output into readable chunks, format text (line breaks, limits), and synchronize with audio using timestamps.

Integration: Implement real-time processing for live audio and integrate caption delivery with target systems.

Evaluation: Assess accuracy, synchronization, readability, and user experience. Refine the system based on feedback.

The proposed methodology for automated subtitle generation is structured into a multi-stage pipeline designed to extract, transcribe, and convert audio content from video files into time-aligned captions. Initially, the audio track is extracted from the input video using the **pydub** library, which facilitates multimedia processing by converting the video file into a **.wav** audio format. This audio file is then passed to the Whisper speech recognition model developed by OpenAI, which performs automatic transcription of the spoken content. The transcribed text is segmented into discrete caption blocks, each containing a fixed number of words (e.g., ten), to maintain clarity and synchronization. Temporal alignment is estimated by assigning start and end timestamps to each block based on a standard speech rate approximation (0.5 seconds per word). These segments are then formatted into the SubRip Subtitle (SRT) file structure, comprising sequential index numbers, timestamp ranges, and corresponding text. Finally, the system ensures resource efficiency by removing temporary audio files after processing. This end-to-end pipeline offers an efficient and lightweight solution for generating subtitle files from video content with minimal human intervention.

VI. Results And Discussion

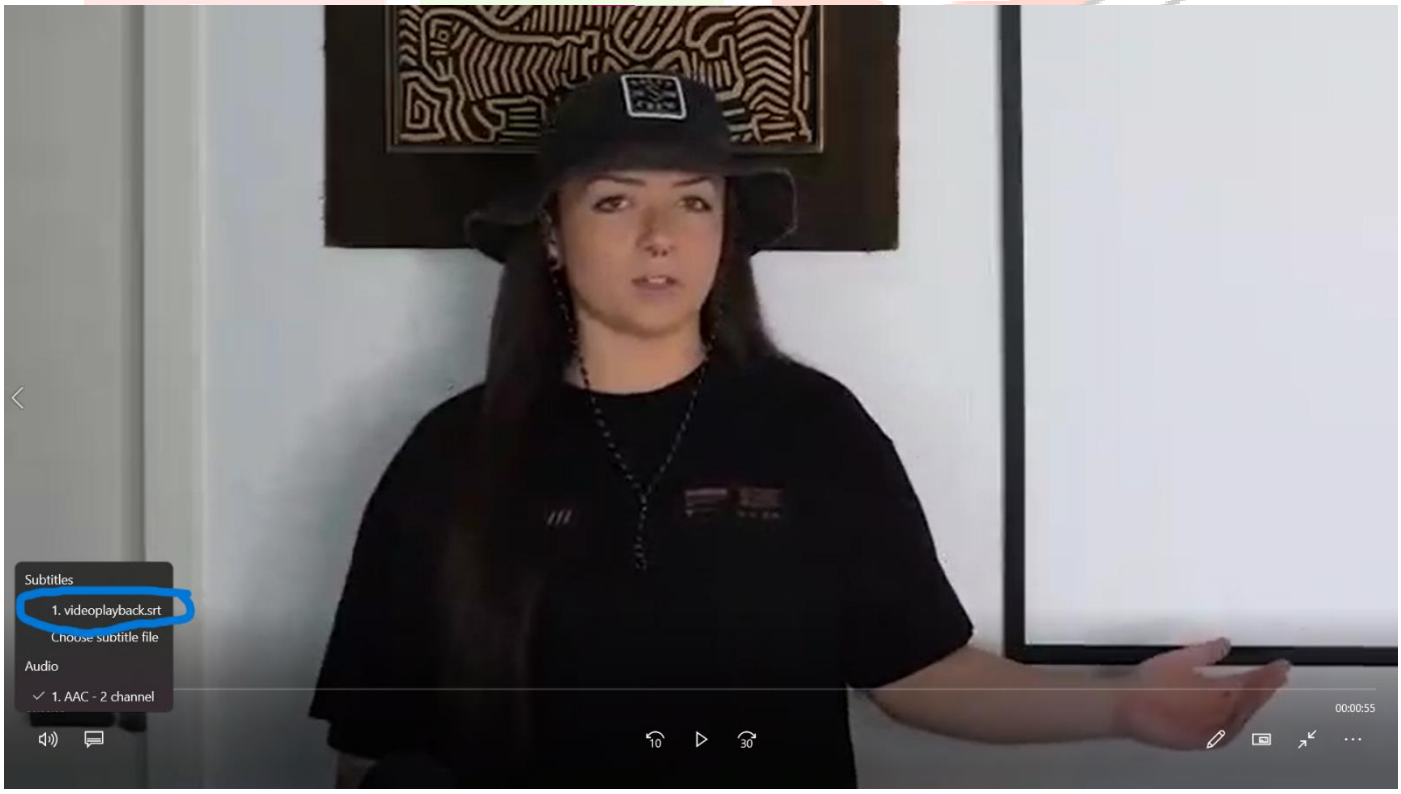

FIG. 1

FIG. 2

REFERENCES

[1] W3C WebVTT: The Web Video Text Tracks Format: The standard for text tracks in HTML5, widely used for web-based closed captions. (https://www.w3.org/TR/webvtt1/)

[2] FCC Regulations on Closed Captioning (USA): While US-specific, these regulations provide a comprehensive overview of quality standards for broadcast and online video. ([Search for "FCC Closed Captioning Rules"])

[3] Ofcom Guidance on Standards for Subtitling and Audio Description (UK): Similar to the FCC, Ofcom provides guidelines relevant to quality and accessibility. ([S.earch for "Ofcom Subtitling Guidelines"])

[4] EBU Subtitling Guidelines: The European Broadcasting Union offers guidelines for subtitling, which often overlap with closed captioning principles. ([Search for "EBU Subtitling Guidelines"])

[5] Bureau of Indian Standards (BIS): While specific standards for closed captioning might be evolving, checking the BIS website for standards related to media accessibility or assistive technologies could be beneficial. (https://bis.gov.in/)

[6] "Subtitling and Captioning: Concepts, Practices and Workflows" by Agnieszka Szarkowska: A comprehensive book covering various aspects of subtitling and captioning.

[7] "Audio Description and Subtitling in Multimedia" edited by Pilar Orero: Offers insights into different facets of media accessibility.

[8] Reports and Publications by Accessibility Organizations: Organizations like the National Association of the Deaf (NAD) in the US or equivalent organizations in India (if they exist and publish on this topic) often have valuable resources. ([Search for "Accessibility organizations India"])

[9] Web Accessibility Initiative (WAI) of the W3C: Provides guidelines and resources on making web content accessible, including video and audio. (https://www.w3.org/WAI/)

[10] Accessibility blogs and forums: Online communities and blogs dedicated to web accessibility and media accessibility often discuss best practices and new developments in closed captioning.

[11] Developer documentation for cloud-based ASR services: Google Cloud, Amazon Web Services, Microsoft Azure provide extensive documentation for their speech-to-text APIs.

[12] Open-source project repositories (GitHub, GitLab): Explore projects related to ASR and subtitle processing for potential code and insights.