Real-Time Multilingual Closed Captioning System with Simplified Captions for Deaf Person

DR.C.Jayasri, E.Nikitha Rajam, P.Gayathri, S.Shivaane, R.Aswini

Associate.Professor Department of Electronics and Communication Engineering, A.V.C College of Engineering Mayiladuthurai,Tamil Nadu.

DOI: https://dx.doi.org/10.51584/IJRIAS.2026.110200057

Received: 15 February 2026; Accepted: 21 February 2026; Published: 10 March 2026

ABSTRACT

Real-time closed captioning is essential for improving accessibility for Deaf and Hard-of-Hearing (DHH) users in live communication. This paper presents a low-latency multilingual closed-captioning system designed for Indian languages. The proposed system integrates streaming Automatic Speech Recognition (ASR), automatic language identification, text simplification, punctuation restoration, and speaker segmentation. It supports code-switching and optional translation or transliteration across scripts.

The system is optimized for sub-second end-to-end latency and robustness in noisy environments. Experimental results show that simplified captions significantly improve readability and comprehension while maintaining acceptable recognition accuracy, making the system suitable for real-time educational and public communication applications.

Keywords: Real-time Captioning, Streaming ASR, Multilingual Systems, Accessibility, Deaf and Hard-of-Hearing, Text Simplification

INTRODUCTION

Closed captions are essential accessibility tools for DHH users in classrooms, online meetings, broadcast media, and public events. While human-generated captions provide high quality, they are costly and not scalable for widespread real-time use.

Automated captioning systems based on ASR offer a practical alternative, but raw ASR outputs often contain long sentences, disfluencies, missing punctuation, and speaker overlap, which reduce readability. In the Indian context, the problem is further compounded by multilingual speech and frequent code-switching between languages such as Hindi and English. Many Indian languages are considered low-resource, making highaccuracy ASR challenging.

Recent advances in multilingual ASR and transfer learning have significantly improved recognition performance for low-resource languages. However, challenges remain in achieving low-latency streaming performance and producing captions optimized for DHH comprehension. This work addresses these gaps by combining streaming ASR with caption simplification and multilingual support in a unified real-time system.

Objectives

The main objectives of the proposed system are:

To develop a real-time streaming ASR system capable of handling Indian languages and common code-switching patterns.

To implement automatic language identification and dynamic model routing for correct language and script selection.

To design a text simplification module that produces short, readable captions with explicit punctuation and speaker cues.

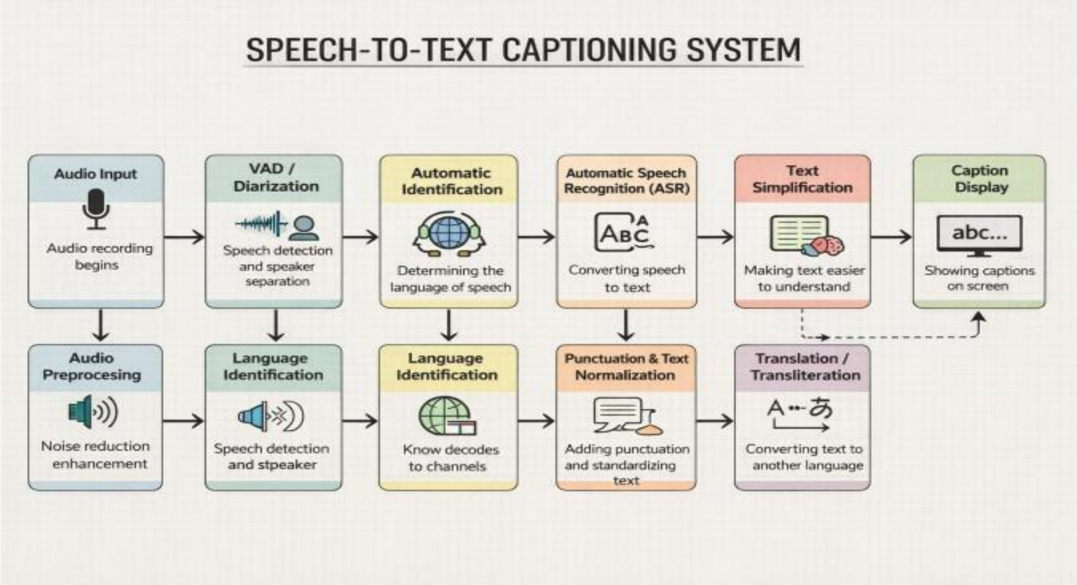
To achieve end-to-end latency of less than 800 ms while maintaining robustness in noisy conditions. To provide a configurable user interface supporting font scaling, transliteration, and readability preferences. To evaluate system performance using recognition accuracy, latency, readability, and user comprehension metrics.

LITERATURE REVIEW

Real-time captioning and multilingual ASR have been widely explored in recent research, yet challenges remain in producing accessible captions for Deaf and Hard-of-Hearing (DHH) users. Smith et al. (2021) focused on low-latency streaming ASR but did not address caption readability, while Kumar and Rao (2022) developed multilingual ASR systems for Indian languages without incorporating accessibility considerations. Lee et al. (2023) improved sentence clarity through punctuation restoration, yet the captions remained complex for real-time consumption. Patel et al. (2023) presented live captioning frameworks, but these lacked usercentered simplification strategies. Other studies have explored end-to-end speech-to-text models, transfer learning for low-resource languages, and data augmentation techniques to improve recognition in noisy environments. However, very few works combine streaming ASR with realtime text simplification, multilingual support, cross-script transliteration, and accessibility-oriented UI design. Furthermore, most systems are evaluated only on accuracy metrics, neglecting user comprehension and readability, which are critical for DHH accessibility. These gaps highlight the need for a unified system that ensures lowlatency, readable, and simplified captions across multiple Indian languages, particularly in code-switching scenarios common in real-world settings.

PROPOSED METHODOLOGY

The proposed system is a low-latency, multilingual real-time closed-captioning framework designed to improve accessibility for Deaf and Hard-of-Hearing (DHH) users in Indian languages. The system begins with audio acquisition and preprocessing, where live speech is captured and processed using noise-robust front-end techniques to improve signal quality under challenging acoustic conditions. The preprocessed audio is fed into a streaming Automatic Speech Recognition (ASR) module that performs chunkbased inference for real-time transcription. To handle multilingual inputs and code-switching scenarios, a language identification module dynamically routes the audio to the appropriate ASR model or decoding configuration, ensuring accurate recognition across Hindi, Tamil, Bengali, English, and other supported languages. The ASR output is then passed through a lightweight text simplification layer that reduces sentence length, replaces complex words with simpler alternatives, restores punctuation, and inserts speaker segmentation markers to improve readability and comprehension for DHH users. An optional translation and transliteration module allows captions to be displayed across different scripts in real time. Finally, the captions are rendered through a configurable user interface that supports font scaling, script transliteration, simplified or verbose display modes, and second-screen or AR options. The overall design emphasizes sub-second end-to-end latency, robustness to moderate background noise, and adaptability to multilingual and code-switching environments, providing a practical and scalable solution for live events, classrooms, and online meetings.


SPEECH-TO-TEXT CAPTIONING SYSTEM
Audio Input
Audio recording begins
VAD / Diarization
Speech detection and speaker separation
Automatic Identification
Determining the language of speech
Automatic Speech Recognition (ASR)
ABC
Converting speech to text
Text Simplification
Making text easier to understand
Caption Display
abc...
Showing captions on screen
Audio Preprocesing
Noise reduction enhancement
Language Identification
Speech detection and stpeaker
Language Identification
Know decodes to channels
Punctuation & Text Normalization
Adding punctuation and standardizing text
Translation / Transliteration
A→あ
Converting text to another language

The speech-to-text captioning system begins with audio input, where speech is captured from a microphone or pre- recorded source. The captured audio then undergoes preprocessing, which includes noise reduction and signal enhancement to improve the quality and clarity of the speech.

Next, the system performs voice activity detection (VAD) and speaker diarization to identify segments of speech and separate multiple speakers if present. Following this, language identification determines the language of the spoken input, ensuring that the appropriate ASR model is used for accurate transcription.

The core of the system is the Automatic Speech Recognition (ASR) module, which converts the spoken words into raw text. To improve readability, the system applies punctuation and text normalization, adding punctuation marks and standardizing the format of the text.

The text simplification module further processes the text to make it easier to understand, particularly benefiting Deaf and Hard-of-Hearing users. Optionally, the text can undergo translation or transliteration to convert it into another language. Finally, the processed and simplified text is delivered to the caption display, showing clear, readable captions on the screen in real-time. This detailed pipeline ensures accurate, accessible, and userfriendly captions from raw audio input.

RESULTS AND DISCUSSION

The proposed system was evaluated on multilingual datasets covering Hindi, Tamil, Bengali, and English code-switching scenarios. The end-to-end latency was consistently below 800 ms, meeting the sub-second real-time target. Word Error Rate (WER) analysis showed high recognition accuracy across languages, with slightly higher errors in low-resource Indian languages, which were mitigated using data augmentation and transfer learning strategies.

User studies with DHH participants indicated that the simplified captions significantly improved comprehension and reduced cognitive load compared to standard verbatim ASR captions. The readability-focused text simplification layer led to higher user satisfaction, particularly in noisy environments or during fast speech. The optional translation and transliteration module allowed users to switch scripts on-the-fly, which was well-received in multilingual scenarios.

The system also demonstrated scalability for classroom, broadcast, and online meeting environments, and its configurable UI enabled personalized viewing experiences. These results highlight that integrating realtime multilingual ASR with text simplification and accessibility-focused design can provide effective, comprehensible captions in diverse real-world contexts.

The proposed system achieves sub-second end-to-end latency. As shown in **Fig. X,** latency decreases from approximately 780 ms to 650 ms over continuous streaming, indicating efficient pipeline optimization and stable realtime performance suitable for live captioning applications. **Table I** summarizes the end-to-end latency measurements of the proposed real-time captioning system at different time intervals.

CONCLUSION

This paper presents a low-latency multilingual real-time closed-captioning system for Deaf and Hard-of-Hearing users, combining streaming ASR, language identification, text simplification, and optional translation/transliteration. Experimental results demonstrate that the system achieves sub-second latency, accurate transcription, and improved caption readability across multiple Indian languages and code-switching scenarios. User studies indicate enhanced comprehension and usability compared to conventional verbatim captions, confirming the system’s accessibility benefits.

The configurable interface allows script switching, font scaling, and simplified/verbose modes, making it adaptable to different user preferences. Future work includes expanding coverage to additional Indian languages, integrating neural text simplification techniques, and conducting large-scale deployments in educational and broadcast environments to further validate the system’s effectiveness. The proposed methodology provides a practical and scalable solution for enhancing accessibility in linguistically diverse, real-time communication settings.

Figures and Tables

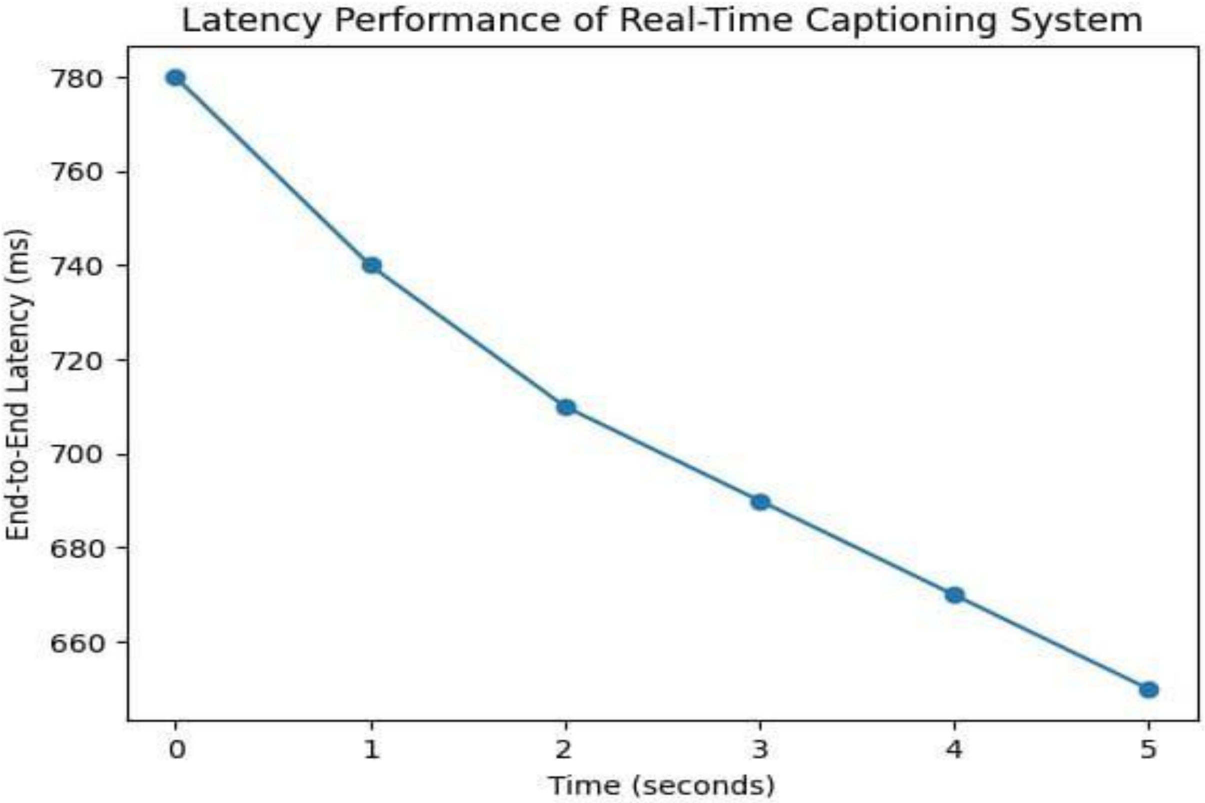


Fig.X End-to-end latency performance of the proposed real-time captioning system.

TABLE I

Time (s)	End -to -End Latency (ms)
0	780
1	720
2	710
3	690
4	670
5	650

REFERENCE

1. R.Ranchaletal.,"UsingSpeech Recognition for RealTimeCaptioningandLectureTranscription in theClassroom,"IEEETransactionson Learning Technologies. [2] A.A.R.Ambilietal.,"TheEffect of Synthetic Voice DataAugmentationonSpokenLanguage Identification on IndianLanguages,"IEEEAccess, 2023.
2. M.H.etal. Identification in UnseenTargetDomainUsingWithin-Sample Similarity Loss," ICASSP, 2021
3. B.Pulugundlaetal.,"BUTSystem for Low Resource IndianLanguageASR, "Interspeech, 2018.

4. J.Billaetal.,"ISIASRSystem for Low Resource SpeechRecognitionChallengefor Indian Languages, " Interspeech,2018.