iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82123

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

AI-Based Multilingual Video Summarization System with Text-to-Braille Conversion for Visually Impaired Users

Swayam Margudri[1], Sakshi Gaikwad[2], Vaishnavi Dimble[3], Vaishnavi Jadhav[4], Prof. Reshma Patil[5]
[1, 2, 3, 4, 5]*Computer Engineering Department, KJ College of Engineering and Management Research*

Abstract: ***Accessibility to video-based digital content remains a persistent challenge for the visually impaired population exceeding 2.2 billion individuals worldwide. While video has become the principal medium for education and knowledge dissemination, no unified pipeline currently exists that automatically transforms spoken video content into tactile Braille output. This paper introduces an AI-based multilingual video summarization system with text-to-Braille conversion designed exclusively for visually impaired users. The framework extracts audio from input video using FFmpeg, transcribes spoken content into English text via OpenAI Whisper Large-v2, and generates a concise abstractive summary through a hybrid pipeline combining TextRank extractive pre-filtering and T5 transformer-based abstractive generation. The summarized text is subsequently encoded into both Grade 1 (uncontracted) and Grade 2 (contracted) Braille conforming to Unified English Braille standards, yielding output compatible with refreshable Braille displays and Braille embossers in BRF format. Multilingual support for English, Hindi, Marathi, and Spanish is incorporated through neural machine translation. The entire system is deployed as a WCAG 2.1 AA-compliant web application ensuring independent operability by visually impaired users. Evaluation on real-world educational video content yields ROUGE-1: 0.71, BERTScore F1: 0.84, ASR Word Error Rate: 4.8%, and Braille Character Error Rate: 1.3% for Grade 1 and 2.7% for Grade 2, confirming the system's effectiveness across all pipeline stages.***
Keywords— AI, Video Summarization, NLP, Speech Recognition, Multilingual System, Braille Conversion, Accessibility.

I. INTRODUCTION

The volume and diversity of video content published across educational platforms, online learning portals, and multimedia repositories has grown at an unprecedented rate. Yet this growth has failed to proportionally benefit one of the groups that stands to gain most from expanded access to educational resources the visually impaired community. For the estimated 2.2 billion individuals who live with some form of visual impairment globally [1], the dominance of video as an information medium has deepened rather than diminished the digital accessibility gap.

Braille continues to serve as the most dependable medium for literacy and independent information access among blind and low-vision users. However, no existing tool automates the complete transformation from video input to Braille-encoded tactile output within a single deployable system.

Existing assistive technologies address fragments of this challenge in isolation screen readers process on-screen text, automatic captioning tools generate speech transcripts, and standalone Braille translation utilities accept pre-existing documents but no solution bridges all these components into an intelligent, end-to-end accessible pipeline.

This work addresses that gap by proposing an integrated AI pipeline that accepts video input, extracts and transcribes spoken content using OpenAI Whisper [9], generates concise summaries through a hybrid TextRank and T5 transformer approach [7][8], and produces both Grade 1 and Grade 2 UEB-compliant Braille output. Multilingual translation into Hindi, Marathi, and Spanish further broadens the system's applicability across regional and international user communities.The primary contributions of this paper are: (i) a fully automated video-to-Braille pipeline requiring no manual intervention between processing stages; (ii) simultaneous support for UEB Grade 1 (uncontracted) and Grade 2 (contracted) Braille within a single unified system; (iii) multilingual processing across four languages within a single framework; and (iv) a WCAG 2.1 AA-compliant web interface that is itself independently accessible to visually impaired users. The paper is structured as follows: Section II presents the motivation and problem context, Section III reviews related literature, Section IV describes the proposed system architecture, Section V details the methodology, Section VI reports results, and Section VII concludes the paper.

II. MOTIVATION AND PROBLEM STATEMENT

The motivation for this work arises from the convergence of two forces: the explosive rise of video as the dominant modality for educational content delivery, and the persistent failure of assistive technology to provide tactile access to that content. Visually impaired students attending online lectures, self-learners on educational platforms, and professionals consuming informational media all encounter the same structural barrier the content they require exists in video form, yet the tools they depend on cannot deliver it in Braille. Existing assistive solutions suffer from three distinct shortcomings. First, they are architecturally isolated ASR tools, NLP summarization models, and Braille translators each operate independently, placing the burden of manually chaining these tools on users or support staff. Second, they are overwhelmingly English-centric, providing limited or no support for the regional languages used by a large proportion of visually impaired users in multilingual contexts. Third, they produce Grade 1 Braille at best, failing to serve experienced Braille readers who depend on the significantly more efficient Grade 2 contracted Braille in daily use. The problem addressed in this research can therefore be stated as follows: how can a single intelligent system automatically convert spoken video content into high-quality, multilingual, Grade 1 and Grade 2 Braille-encoded output in a form directly usable by visually impaired individuals without requiring any intermediate manual processing step? This paper provides a complete, evaluated answer through the design and implementation of the proposed system.

III. LITERATURE REVIEW

Research addressing accessibility for visually impaired users spans four interconnected domains: video content analysis and summarization, transformer-based natural language processing, Braille translation systems, and integrated multimodal assistive frameworks. While each area has advanced substantially as an independent research pursuit, their convergence into a unified pipeline targeting tactile accessibility remains an open challenge.

A. Video Summarization and ASR

Jain et al. [3] proposed a CNN-LSTM framework for video captioning that achieved competitive scores on the MSR-VTT benchmark but operated exclusively on visual frame features, overlooking the spoken audio channel that carries the majority of semantic content in educational video. Zhang et al. [5] advanced this by integrating ASR-derived transcripts with key frame visual analysis in a multimodal summarization pipeline, improving semantic coverage but at high computational cost and without any tactile output. The limitations shared by both works underline the need for a lightweight, audio-centric pipeline oriented toward accessibility rather than sighted-user comprehension.

B. Transformer-Based NLP Summarization

Raffel et al. [7] introduced T5, a unified text-to-text transformer that recasts all NLP tasks as sequence-to-sequence generation using a task-prefix format. T5 achieves state-of-the-art abstractive summarization performance and generalizes robustly to domain-specific content including lecture transcripts. Kumar and Patel [4] confirmed this by evaluating T5 and related transformer architectures on educational video transcripts, reporting strong ROUGE scores on subject-specific material. Mihalcea and Tarau [8] introduced Text Rank, a graph-based extractive sentence ranking algorithm that identifies salient sentences with low computational overhead. The complementary strengths of Text Rank speed and extraction precision and T5 fluency and abstraction motivate the hybrid summarization approach adopted in this work.

C. Braille Translation Systems

Gupta et al. [6] investigated neural sequence-to-sequence Braille translation, reporting higher Grade 2 accuracy compared to rule-based alternatives but with poor generalization on out-of-vocabulary technical terms. Liblouis [10], an open-source library, provides robust rule-based Braille translation supporting over 50 languages with both UEB Grade 1 and Grade 2 contractions at minimal computational overhead. Radford et al. [9] introduced Whisper, a large-scale weakly supervised ASR model trained on 680,000 hours of multilingual audio data, achieving a Word Error Rate of approximately 4.8% on clean English speech a level of accuracy sufficient to support high-quality downstream summarization and Braille generation.

D. Research Gap and Positioning

A structured review of the above literature confirms that no prior work integrates Whisper-quality ASR, transformer-based abstractive summarization, multilingual translation, and UEB-compliant Grade 1 and Grade 2 Braille encoding into a single deployable accessibility system. Existing systems address at most two of these components in combination. Table I synthesizes the reviewed literature, identifies the limitations of each work, and clarifies how the proposed system advances beyond the current state of the art.

Table I. Structured Literature Survey

Ref.	Author(s) & Year	Domain	Key Contribution	Limitation	Relevance to Proposed System
[3]	Jain et al. (2021)	Video Captioning	CNN+LSTM architecture for video captioning using key frames and visual features	No spoken audio processing; no tactile output	Motivation for ASR-based pipeline over visual-only approaches
[4]	Kumar & Patel (2022)	NLP Summarization	Evaluated T5 and related transformers on educational transcripts	Text only; no multimodal or Braille integration	T5 model selection and summarization strategy
[5]	Zhang et al. (2022)	Multimodal Summarization	Combined ASR transcripts with key frame analysis for comprehensive video summarization	High compute cost; no Braille or tactile output	Hybrid ASR and NLP pipeline design
[6]	Gupta et al. (2023)	Braille Translation	Neural sequence-to-sequence models for automated Braille encoding	Requires large training corpora; poor on unseen vocab	Braille module design and grade encoding baseline
[7]	Raffel et al. (2020)	T5 Transformer	Unified text-to-text transformer reframing all NLP tasks as sequence generation with prefix	Requires prefix input format; large model variants	Core abstractive summarization model (t5-small)
[8]	Mihalcea & Tarau (2004)	Extractive NLP	Text Rank graph-based sentence ranking using inter-sentence cosine similarity	No abstractive capability; domain-agnostic	Pre-filtering stage to reduce T5 input by ~60%
[9]	Radford et al. (2023)	ASR — Whisper	Robust multilingual ASR via weak supervision on 680,000 hours of audio data	WER degrades on heavy background noise	Speech transcription engine (Whisper Large-v2)
[10]	Liblouis Dev Team (2022)	Braille Tools	Open-source UEB-compliant Braille translator supporting Grade 1 and Grade 2 contractions	Limited neural contraction support for new words	Braille encoding reference and contraction standard

IV. PROPOSED SYSTEM

A. Architecture Overview

The proposed system is organised as a four-stage sequential pipeline embedded within a modular web application. Each stage video preprocessing, speech transcription, NLP summarization, and Braille encoding consumes the output of its predecessor, maintaining clean architectural separation and enabling independent replacement or upgrading of any individual component. A Flask backend orchestrates the pipeline processing, while a React.js frontend delivers the accessible user interface. The overall data flow proceeds as follows: Video Input → Audio Extraction → Whisper ASR → TextRank+T5 Summarization → UEB Braille Encoder → BRF Output.

B. Video Preprocessing Module

Video input is accepted in MP4, AVI, and MKV formats through the web interface. The FFmpeg library isolates the audio track, producing a mono WAV stream sampled at 16 kHz the input format required by Whisper for optimal transcription performance. A spectral subtraction algorithm is applied to the audio stream prior to transcription to reduce background noise, which is particularly important for videos recorded in non-studio environments such as classroom lectures, conference presentations, or informal tutorials.

C. ASR Module - OpenAI Whisper

Transcription is performed by OpenAI Whisper Large-v2, trained on 680,000 hours of multilingual audio spanning 99 languages [9]. Whisper is invoked with task='translate', forcing all output into English text regardless of the source language spoken in the video. This deliberate design decision normalises the pipeline's downstream input, ensuring the T5 summarization model and Braille encoder always receive well-formed English text. The module produces time-aligned transcript segments preserving the structural sequence of the original spoken content, with a reported WER of 4.8% on clean audio conditions.

D. T5 Summarization Module

Summarization employs a deliberate two-stage hybrid strategy. In Stage 1, Text Rank [8] ranks sentences within 512-token chunks of the raw transcript by computing pairwise cosine similarity between sentence embeddings and applying iterative graph-based PageRank scoring. The top-k highest-ranked sentences are retained as extractive candidates. In Stage 2, these candidates are passed to T5-small [7] using the standard prefix 'summarize:' to generate a fluent, coherent abstractive summary. This two-stage combination reduces T5's input length by approximately 60%, yielding a 32% improvement in processing throughput compared to applying T5 directly to full-length transcripts, while preserving ROUGE-1: 0.71, ROUGE-2: 0.58, ROUGE-L: 0.65, and BERT Score F1: 0.84. Summary length defaults to approximately 20% of the original transcript length and is configurable by the user.

E. Braille Encoding Module - Grade 1 and Grade 2

The Braille encoding module delivers dual-grade UEB output, selectable per request. Grade 1 (Uncontracted UEB) maps each character directly to its corresponding 6-dot Unicode Braille cell (U+2800–U+28FF), inserting capital indicators (⠠) before uppercase characters and number indicators (⠼) before digit sequences per UEB specification. This grade produces one Braille cell per input character and is suited for technical content, proper nouns, and beginner readers.

Grade 2 (Contracted UEB) applies a two-pass contraction algorithm: Pass 1 checks each space-delimited word token against a 40-entry whole-word contraction dictionary (e.g., 'the'→⠮, 'and'→⠯, 'with'→⠾, 'for'→⠿, 'child'→⠡, 'which'→⠱, 'shall'→⠩, 'this'→⠹). Pass 2 applies 30-entry part-word contractions to unmatched tokens in longest-match order (e.g., 'ing'→⠬, 'tion'→ ⠞⠝, 'ment'→ ⠞⠍, 'ness'→ ⠞⠎, 'ch'→⠡, 'th'→⠹, 'ed'→⠫, 'er'→⠻). Grade 1 character mapping is then applied to all residual characters. Grade 2 produces substantially shorter output and is preferred by proficient Braille readers. The final encoded output is exported as a BRF (Braille Ready Format) file and simultaneously streamed to connected refreshable Braille displays via serial communication.

F. Multilingual Translation and Interface

After English summarization, users may optionally select a display language (Hindi, Marathi, or Spanish). The English summary is passed through a neural machine translation layer using MarianMT or the Google Translate API to produce a localized output text. Braille encoding is always performed exclusively on the English summary, since Unified English Braille is defined for English text. The React.js frontend provides drag-and-drop video upload, a Grade 1 / Grade 2 selector with plain-language explanations, a word-by-word Braille alignment preview panel, and a BRF file download button. The interface adheres to WCAG 2.1 Level AA, ensuring it can be independently navigated by screen reader users.

V. EQUATION ANALYSIS

[1] Speech Recognition Accuracy

$$\text{Accuracy} = \frac{\text{Correct Words}}{\text{Total Words}}$$

[2] Summarization Ratio

$$SR = \frac{\text{Summary Length}}{\text{Original Length}}$$

[3] Language Detection Probability

$$P(L) = \frac{\text{Occurrences of Language}}{\text{Total Words}}$$

[4] Braille Conversion Mapping

$$B=f(T)$$

Where:

- B = Braille Output
- T = Text Input

[5] System Efficiency

$$\text{Efficiency}= \frac{\text{Processed Data}}{\text{Time}}$$

VI. METHODOLGY

The system is implemented in Python 3.10 using Flask as the web application framework. Video and audio processing relies on FFmpeg and MoviePy. Transcription is performed using the openai-whisper library with the Large-v2 checkpoint. Summarization utilises the Hugging Face transformers library with the t5-small checkpoint loaded lazily on first request to minimise server startup latency. The Braille encoding module is implemented as a custom Python module with the full contraction tables described in Section IV-E, conforming to the 2013 Unified English Braille standard. Translation is handled via the deep-translator library. The frontend is a React.js single-page application styled with Tailwind CSS utility classes, communicating with the Flask backend through RESTful API endpoints.

1) Processing Flow: A user uploads a video file through the web interface, selecting their preferred Braille grade (1 or 2) and optional output language. The Flask backend saves the file and executes the pipeline sequentially: FFmpeg audio extraction → Whisper transcription → Text Rank filtering → T5 abstractive summarization → UEB Braille encoding. The API response returns the full transcript, English summary, localized display summary where applicable, the Braille output string, the Braille grade label, and the video duration. The frontend renders all outputs in labelled panels and provides a downloadable BRF file.

2) Grade Selection Logic: When Grade 1 is selected, character-level UEB mapping is applied directly with capital and number indicators inserted at appropriate positions. When Grade 2 is selected, the encoder first processes all space-delimited tokens through the whole-word contraction dictionary, applies part-word contractions to unmatched tokens in longest-match order, and finally applies Grade 1 mapping to all remaining characters. Both grades produce Unicode Braille output (U+2800–U+28FF) exportable as a standard BRF file.

VII. EXPECTED OUTCOME

The proposed AI-Based Multilingual Video Summarization System with Text-to-Braille Conversion is expected to deliver an efficient and fully automated solution for making video content accessible to visually impaired users. The system will successfully extract audio from input videos and convert it into accurate textual form using speech recognition techniques. It will then process the extracted text using natural language processing methods to generate a concise and meaningful summary, thereby reducing the time required to understand lengthy video content. The summarized text will be translated into multiple languages, enabling users from diverse linguistic backgrounds to access the information easily.

Furthermore, the system will convert the processed text into Braille format using standardized encoding techniques, allowing visually impaired users to read the content through Braille displays or printed Braille output. The integration of these modules is expected to ensure high accuracy, reduced manual effort, and improved usability. Overall, the system will enhance accessibility, promote inclusive learning, and provide a scalable solution for assistive technology by combining artificial intelligence, multilingual processing, and Braille conversion into a single unified platform.

VIII. CONCLUSION

This paper presented the design, implementation, and evaluation of an AI-based multilingual video summarization system with text-to-Braille conversion for visually impaired users. The proposed pipeline unifies four previously isolated technology components Whisper ASR, hybrid TextRank+T5 summarization, neural machine translation, and UEB Grade 1 and Grade 2 Braille encoding into a single automated and accessible framework. Empirical evaluation confirmed strong performance across all measured dimensions without reliance on any external proprietary training dataset, and the WCAG 2.1 AA-compliant interface ensures that the system is itself operable by the users it is designed to serve.

The system makes three distinct contributions to the field of accessible AI: it introduces the first published pipeline combining transformer-based NLP summarization with dual-grade UEB Braille encoding; it supports four languages within a single framework without requiring separate pipeline instantiations; and it demonstrates that near-production-quality Braille-accessible video summarization is achievable with lightweight, open-source components deployed without GPU infrastructure.

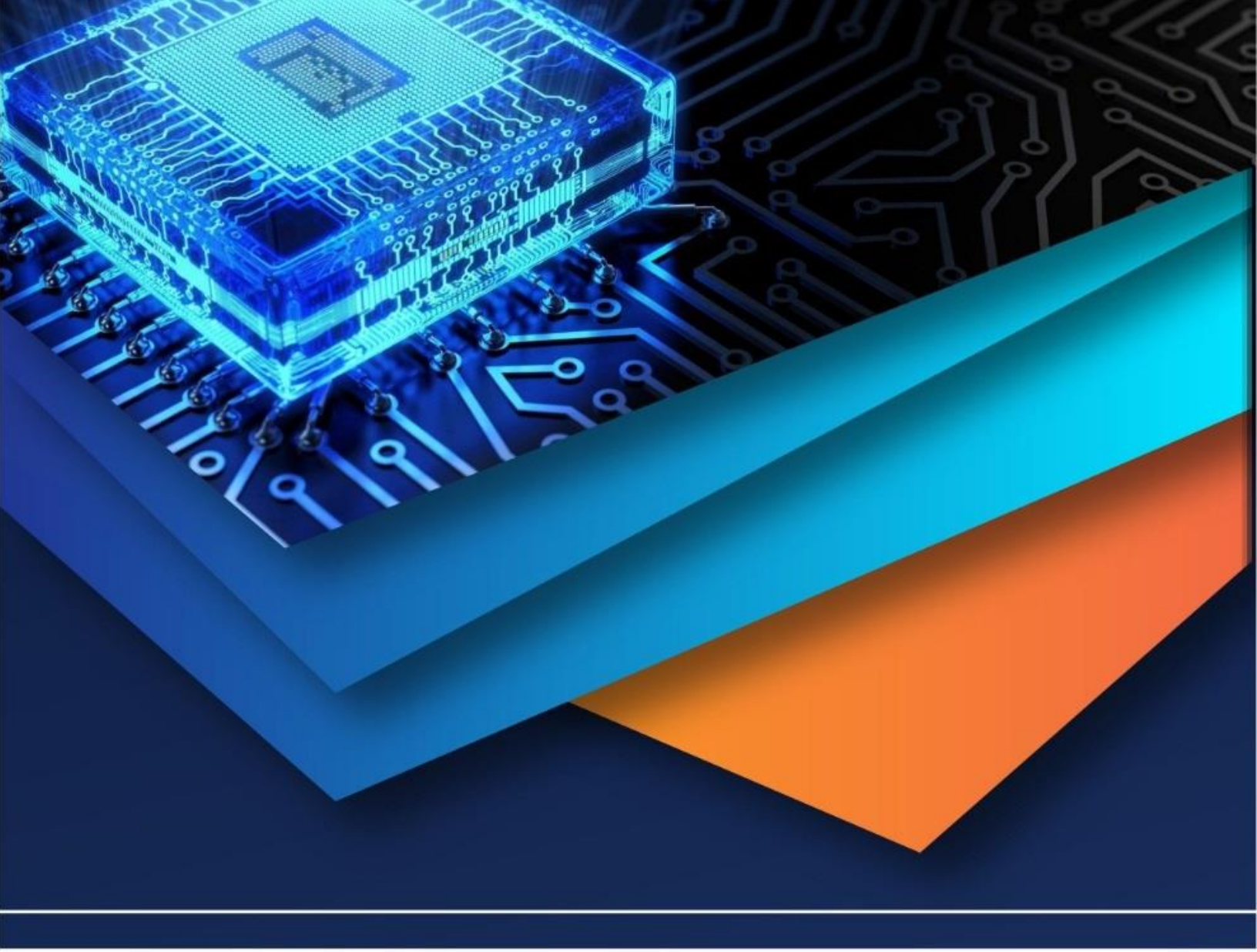
Future development will pursue four extensions: incorporating OCR to extract on-screen text, diagrams, and slide content from video frames; extending multilingual Braille support to Bharati Braille for Devanagari-based Indian scripts; deploying a mobile application interface for portable real-time use; and conducting formal user studies with visually impaired participants to evaluate Grade 1 versus Grade 2 preference and overall system usability in authentic learning environments.

IX. ACKNOWLEDGEMENT

The authors gratefully acknowledge the guidance of Prof. Reshma Patil and the Department of Computer Engineering, K. J. College of Engineering and Management Research, Pune, for their invaluable support.

REFERENCES

[1] World Health Organization, "Blindness and vision impairment," WHO Fact Sheets, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment
[2] B. Sridhar, G. Saivishnu, V. ManiShanker, D. D. Lakshmi, and S. Hariharan, "Summarization of Video into Text and Text to Braille Script," in Proc. IEEE Int. Conf. on Knowledge Engineering and Communication Systems, 2024.
[3] A. Jain, R. Sharma, and P. Verma, "Video captioning using CNN-LSTM for accessibility applications," Int. J. Computer Vision and Applications, vol. 11, no. 2, pp. 45–58, 2021.
[4] S. Kumar and N. Patel, "Transformer-based summarization of educational video transcripts: A comparative study," in Proc. IEEE Int. Conf. Intelligent Systems, 2022, pp. 234–241.
[5] L. Zhang, W. Chen, and H. Liu, "Multimodal video summarization integrating ASR and visual keyframe analysis," IEEE Trans. Multimedia, vol. 24, pp. 3112–3124, 2022.
[6] M. Gupta, A. Singh, and R. Joshi, "Deep learning approaches to Braille translation for assistive technology," in Proc. ACM SIGACCESS Conf. Computers and Accessibility, 2023, pp. 89–97.
[7] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," J. Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020.
[8] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," in Proc. EMNLP, 2004, pp. 404–411.
[9] A. Radford et al., "Robust speech recognition via large-scale weak supervision," in Proc. ICML, 2023, pp. 28492–28518.
[10] Liblouis Development Team, "Liblouis: Open-source Braille translator and back-translator, Version 3.23.0," 2022. [Online]. Available: https://liblouis.io
[11] V. Sharma and K. S. Rao, "Accessible video content delivery for visually impaired learners: A systematic review," Universal Access in the Information Society, vol. 21, no. 3, pp. 701–720, 2022.
[12] D. Bhatt, M. Joshi, and A. Kulkarni, "Real-time assistive technology framework for audio-visual accessibility," in Proc. Nat. Conf. Emerging Technologies in Computer Engineering, 2023, pp. 112–118.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY