ISSN No. 0976-5697

VOLUME 17 SPECIAL ISSUE 1, JULY 2026

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

AI-POWERED MULTILINGUAL CONTENT LOCALIZATION ENGINE FOR SKILL COURSES

Pankaj Vaishnav, Dushyant Salvi, Kushal Cheeta, Mahesh Vyas
Department of Computer Science and Engineering
Geetanjali Institute of Technical Studies (GITS)
Dabok, Udaipur (Raj.), India
pankaj.vaishnav@gits.ac.in, salvidushyant17@gmail.com, Kushal.cheeta5@gmail.com, vy.maheshv@gmail.com

Abstract: This paper presents an AI-powered multilingual content localization engine designed to bridge the language accessibility gap in India's vocational and skilling ecosystem. The proposed system integrates multiple AI technologies including Automatic Speech Recognition (ASR) using Whisper, Neural Machine Translation (NMT) using MarianMT, domain- specific glossary alignment via Retrieval-Augmented Generation (RAG), and Text-to-Speech (TTS) synthesis using Coqui TTS into a unified pipeline for both text-based course content localization and video speech dubbing. The engine targets all 22 scheduled Indian languages and incorporates a modular microservice architecture with LMS integration through FastAPI REST endpoints. Key components include a vocal isolation layer using Spleeter, structured PDF and text parsing, structured assessment localization, and a feedback-driven quality improvement loop. The framework addresses challenges of domain terminology misalignment, regional voice naturalness, audio-video synchronization, and scalability for national-level deployment. Evaluation is designed around BLEU scores for translation, Word Error Rate for transcription, and Mean Opinion Score for TTS naturalness. This work contributes a feasible, cost-efficient, and scalable architecture for automated multilingual localization of vocational training material with significant implications for learner inclusion and certification success across India's diverse linguistic communities.

Keywords: multilingual localization; automatic speech recognition; neural machine translation; text-to-speech; skill courses; AI localization; Whisper ASR; MarianMT; Coqui TTS; vocational training; LMS integration; Indian languages; RAG glossary; content accessibility

1 INTRODUCTION

India's vocational education and skilling landscape faces a fundamental challenge: the overwhelming majority of training content is created and delivered in English, a language that a significant portion of the target learner population is not proficient in. According to national census data, fewer than 12% of Indians use English as a primary or secondary language, yet skilling platforms, safety training modules, and certification assessments continue to be developed in English-first formats. This creates a structural exclusion of rural, regional, and non-English-speaking communities from meaningful participation in national skilling initiatives.

The Government of India's flagship Skill India program aims to train hundreds of millions of youths in vocational skills to enhance employability. However, the linguistic gap between content creators and learners significantly reduces course completion rates, comprehension quality, and certification success. Traditional solutions involving manual translation and video dubbing are slow, expensive, inconsistent in quality, and impossible to scale to national levels across dozens of regional languages.

The emergence of deep learning-based natural language processing and speech technologies has opened new possibilities for automated, high-quality multilingual content localization. Models such as OpenAI's Whisper for speech recognition, MarianMT for neural machine translation, and Coqui TTS for text-to-speech synthesis provide powerful building blocks for an end-to-end localization pipeline. However, applying these models directly to vocational training content presents unique challenges: domain-specific terminology must be preserved accurately, voice naturalness must be region-appropriate, and the pipeline must integrate seamlessly with existing Learning Management Systems (LMS). This paper introduces a conceptual AI-powered multilingual content localization engine specifically designed for skill courses. The proposed system provides a unified pipeline for localizing both textual study material and video speech through transcription, translation, and re-dubbing. The system is designed with a modular microservice architecture enabling incremental deployment and updates without full system retraining.

- A unified AI localization pipeline combining ASR, NMT, domain glossary normalization, and TTS into a single coherent system for skill course content.
- A domain-specific glossary alignment mechanism using Retrieval-Augmented Generation (RAG) to ensure terminology accuracy across 22 Indian languages.
- An evaluation framework covering BLEU score for translation, Word Error Rate (WER) for transcription, and Mean Opinion Score (MOS) for audio naturalness.

The rest of this paper is organized as follows: Section II

reviews related work. Section III describes the system architecture and methodology. Section IV presents the system design and requirements. Section V defines the evaluation framework. Section VI discusses implications and limitations. Section VII concludes the paper.

2. RELATED WORK

A. Neural machine Translation.

Neural Machine Translation has advanced significantly since the transformer architecture introduced by Vaswani et al. [1]. MarianMT, developed by the Helsinki-NLP team [2], provides pre-trained translation models for numerous language pairs including several major Indian languages such as Hindi, Bengali, Tamil, and Telugu. These encoder-decoder transformer models are well-suited for CPU-constrained deployment due to their relatively compact size. Efforts specifically targeting Indian languages include the IndicNLP initiative [3] and AI4Bharat [4], which have built large-scale parallel corpora and pre- trained models for all 22 scheduled Indian languages. These works highlight that generic translation models often fail to accurately translate domain-specific content, making glossary augmentation a necessary component.

B. Automatic Speech Recognition for Multilingual Content

OpenAI's Whisper [5] is a general-purpose speech recognition model trained on 680,000 hours of multilingual data, achieving competitive Word Error Rates across multiple languages including several Indian languages. Whisper's encoder-decoder architecture supports both transcription and cross-lingual translation in a single pass, making it well-suited for the initial speech processing stage of a localization pipeline. Earlier ASR research for Indian languages was fragmented across individual language-specific models [6], but recent multilingual models have significantly improved coverage, though domain- specific vocabulary continues to challenge error rates.

C. Text-to-Speech Synthesis and Voice Localization

Modern end-to-end neural TTS systems have replaced older concatenative and parametric approaches, producing highly natural speech. Coqui TTS [7] is an open-source neural TTS framework supporting multiple voice styles and languages with a training pipeline for custom voice models. Research on Indian language TTS, including the IndicTTS project by Baby et al. [8], has demonstrated that regional prosodic characteristics significantly improve listener acceptance and MOS scores compared to generic synthesis, validating the importance of region-specific voice profiles in a localization context.

D. Retrieval-Augmented Generation for Domain Alignment

Retrieval-Augmented Generation (RAG), introduced by Lewis et al. [9], combines generative models with a retrieval mechanism that fetches relevant content from an external knowledge base. In the localization context, RAG-based glossary lookup provides a principled mechanism for ensuring domain-specific terms are correctly translated and preserved across language pairs. This approach is particularly relevant for vocational training content where terms such as safety protocols, technical specifications, and certification requirements must be translated with precision rather than semantic approximation.

E. Existing Localization Systems for E-Learning

Prior automated e-learning localization systems have focused primarily on subtitle generation and basic text translation through APIs such as those integrated into Coursera and edX. These systems lack domain-specific alignment and do not address video audio dubbing or regional voice generation. The absence of a unified pipeline handling both text and speech localization with domain glossary grounding represents the primary gap that the proposed system addresses.

3. METHODOLOGY

The proposed localization engine is designed as a multi-layered architecture that ensures flexibility, scalability, and efficient processing of multimedia content. It is structured into five logical design layers: Input Acquisition, Speech Isolation and Detection, Localization and Normalization, Audio Synthesis and Merge, and Integration and Feedback. Each of these layers operates as an independent modular microservice, enabling seamless updates, maintenance, and scaling without disrupting the overall system. This modular approach also supports parallel development and deployment, making the system adaptable to evolving technologies and requirements.

The first layer, Input Acquisition, is responsible for collecting and preparing the source media. It accepts various input formats, including video and audio files, and ensures compatibility by converting them into standardized formats if necessary. This layer also extracts the audio track from video content and prepares metadata required for downstream processing. By handling diverse input types, it establishes a robust entry point for the localization pipeline.

The second layer, Speech Isolation and Detection, focuses on identifying and separating spoken content from background noise, music, or other non-speech elements. Advanced signal processing and machine learning techniques are employed to detect speech segments accurately. This step is critical for ensuring that only relevant linguistic content is passed forward for localization, thereby improving efficiency and reducing processing errors.

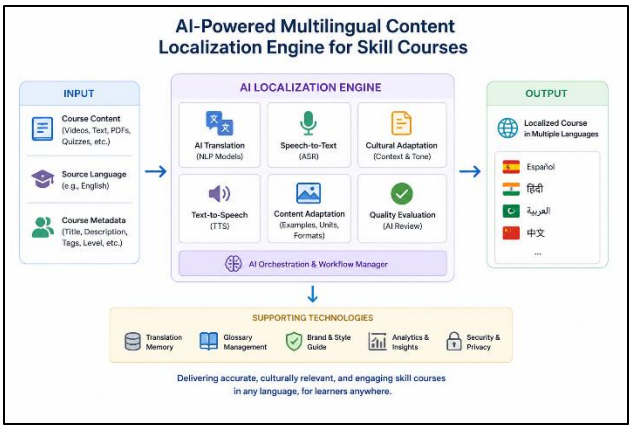

Figure 1: System Architecture of the AI-Powered Multilingual Content Localization Engine

Input Acquisition Layer

The Input Acquisition Layer serves as the entry point for all content entering the localization pipeline. Users or connected LMS platforms supply skill course content in three primary formats: raw text documents, structured PDFs including course modules, assessment banks, and safety instruction sheets, and multimedia files such as video lectures and audio lessons. An intelligent orchestrator component analyzes the incoming content type and routes it to the appropriate processing sub-pipeline. For structured document inputs, a document parser extracts content while preserving document hierarchy including headings, tables, bullet lists, assessment questions, and embedded skill keywords. Assessment questions are extracted into a structured JSON format to enable accurate re-injection of translated content while preserving option mapping integrity. Audio lesson files are routed directly to the Localization and Normalization Layer.

A. Speech Isolation and Detection Layer

For video-format skill training content, vocal separation is a necessary preprocessing step before transcription can proceed. The Speech Isolation and Detection Layer employs Spleeter [10], a source separation library developed by Deezer, to split mixed audio tracks into separate vocal and background stems. This separation ensures that the ASR model receives clean speech signals, reducing transcription errors from background interference. Following vocal isolation, a speech activity detection module scans the extracted vocal track to identify active speech regions and their precise timestamps using energy thresholding, spectral flux analysis, and silence window detection. These timestamps are critical for accurate audio-video.

B. Localization and Normalization Layer

The Localization and Normalization Layer is the core intelligence component of the pipeline. Speech segments are first transcribed into text using Whisper [5], with outputs timestamped to preserve synchronization with the original video timeline. Text content from both transcription and document parsing is passed to the NMT translation module using MarianMT [2] as the primary translation engine, selected for its balance of translation quality and CPU inference speed.

Before final output, all translated content is passed through the Domain Skill Vocabulary (DSV) normalization step. The DSV is a custom glossary database covering skill-domain terminology across fields such as electrical work, construction, healthcare, agriculture, and information technology. The glossary is indexed and queried using a RAG-based lookup mechanism where translated terms are compared against the glossary and corrected when domain-specific alternatives are available. This step ensures technical terms retain their precise meaning across language pairs.

C. Audio Synthesis and Merge Layer

Normalized translated transcripts are passed to the Audio Synthesis and Merge Layer for voice generation and video reconstruction. Neural TTS synthesis is performed using Coqui TTS [7] configured with four broad regional voice profile groups: North Indian (Hindi belt prosody), South Indian (Tamil-Telugu cadence), East Indian (Bengali-Odia prosody), and West Indian (Gujarati-Marathi tone banks). The synthesized audio segments are aligned with the original video timeline using timestamps from the speech detection stage. The multimedia merge operation is handled through a workspace block backed by FFmpeg [11], which manages audio-video synchronization, timeline buffering for duration differences between original and synthesized speech, and final output encoding to produce a fully dubbed video file in the target regional language.

D. Integration and Feedback Layer

The Integration and Feedback Layer exposes the localization pipeline to external LMS platforms through a REST API service built using FastAPI [12]. Standard endpoints allow LMS platforms to submit content for localization, retrieve localized outputs, and manage glossary updates. Offline glossary packs are provided for low-connectivity environments. A quality feedback loop allows localized outputs to be rated by human trainers and domain experts on a 1-5 quality scale across translation accuracy, voice naturalness, and terminology precision. These ratings are stored in a feedback log database and serve as signals for future model fine-tuning and glossary enrichment.

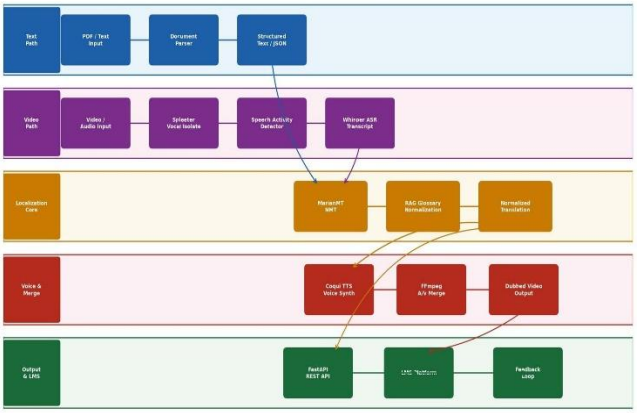

Figure 2: Data Flow Across All Pipeline Components of the Localization Engine

4. SYSTEM DESIGN AND REQUIREMENTS

A. Hardware Requirements

The system is designed to operate without GPU dependency, making it deployable on standard institutional hardware. Minimum requirements include a multi-core CPU (Intel i5/i7 or equivalent), 8GB RAM with 12-16GB recommended for concurrent inference, and 256GB storage for audio-video processing. Stable internet is required only for initial model downloads; subsequent inference operates fully offline.

B. Software Stack

The software stack is built entirely on open-source components to ensure cost efficiency and accessibility. Table 1 summarizes the technology choices across pipeline layers.

Table 1: Software Components of the Localization Engine

Layer	Technology	Function
Speech Isolation	Spleeter	Vocal track extraction from video
ASR Transcription	OpenAI Whisper	Speech-to-text conversion
Translation	MarianMT	Neural machine translation
Glossary Alignment	RAG Lookup (Custom)	Domain terminology normalization
Voice Synthesis	Coqui TTS	Regional TTS voice dubbing
Multimedia Merge	FFmpeg	Audio-video synchronization
API Integration	FastAPI	LMS REST endpoint exposure
Document Parsing	Python Ecosystem	PDF and text structure extraction

C. Supported Languages and Formats

The engine targets all 22 officially scheduled Indian languages as defined by the Eighth Schedule of the Indian Constitution, including Hindi, Bengali, Tamil, Telugu, Marathi, Gujarati, Kannada, Malayalam, Odia, Punjabi, and others, with Hinglish normalization also supported. Supported content formats include course PDFs, plain text modules, structured assessment banks in JSON, audio lesson files, and video lecture files in standard multimedia formats.

4. EVALUATION FRAMEWORK

Since the system is proposed as a conceptual architecture grounded in proven AI components, evaluation is defined using industry-standard metrics for each subsystem. The following metrics are proposed for validation upon system deployment.

A. Translation Quality

Translation accuracy is measured using the BLEU score [13], which computes n-gram overlap between machine-translated output and human reference translations. A BLEU score above 30 is generally considered acceptable for domain-specific content, with scores above 40 indicating high quality. Glossary-hit accuracy measures the percentage of domain-specific skill terms correctly matched and preserved from the Domain Skill Vocabulary. Semantic similarity scoring using cosine distance between sentence embeddings provides a complementary measure of meaning preservation beyond surface-level n-gram matching.

B. Speech Recognition Quality

ASR quality is evaluated using Word Error Rate (WER), which measures the proportion of words incorrectly transcribed relative to a reference transcript. Whisper's reported WER for Hindi and major Indian languages ranges from 8 to 15 percent on clean speech, with higher values expected for domain-specific technical vocabulary. Reducing WER for skill-domain content is a priority that the glossary correction step partially addresses by post-processing transcription outputs.

C. TTS Naturalness and Overall Benchmarks

Voice synthesis quality is assessed using Mean Opinion Score (MOS) ratings collected from domain trainers and end learners on a 1-5 scale across clarity, naturalness, regional appropriateness, and pace. Table 2 presents the target quality benchmarks for each evaluation metric.

Table 2: Target Evaluation Benchmarks

Metric	Component	Target
BLEU Score	Translation (NMT)	> 35
Glossary-Hit Accuracy	Domain Alignment (RAG)	> 90%
Word Error Rate (WER)	ASR (Whisper)	< 15%
MOS Rating	TTS (Coqui TTS)	> 3.5 / 5.0
JSON Schema Validity	Assessment Localization	100%
Option Mapping Integrity	Assessment Localization	100%

Accessibility is evaluated by verifying screen-reader compatibility through plain text structural output, caption completeness for all STT-segmented speech regions, and audio lesson clarity ratings for TTS outputs. These validations ensure that localized content remains accessible to differently-abled learners alongside the primary target audience.

5. DISCUSSION

A. System Strengths

The proposed localization engine addresses several critical gaps in India's skilling content ecosystem. The unified pipeline approach eliminates fragmentation from using separate tools for text translation and video dubbing, reducing

content processing overhead. The Domain Skill Vocabulary with RAG-based lookup provides a principled mechanism for preserving technical terminology that generic NMT models tend to mistranslate. The modular microservice architecture enables incremental deployment, where individual components can be updated as better models become available without disrupting the entire pipeline. The feedback loop creates a self-improving system that becomes more accurate over time through trainer input, aligning improvement with actual usage patterns of target learner communities.

B. Limitations and Challenges

Several practical limitations must be acknowledged. Translation quality for low-resource Indian languages such as Bodo, Dogri, Santhali, and Kashmiri is likely to be lower than for high-resource languages like Hindi and Bengali, due to limited parallel training data availability for MarianMT in these language pairs. Voice naturalness for TTS synthesis in regional languages remains a significant challenge, as producing highly natural voices for South Indian and Northeast Indian language varieties requires dedicated voice recording and model training efforts.

C. Deployment Considerations

For national-scale deployment under programs like Skill India Digital, the system must integrate with existing LMS infrastructure through FastAPI REST endpoints. Offline glossary packs are provided for low-bandwidth rural environments. The design without GPU dependency enables deployment on standard government and institutional hardware without specialized infrastructure investment. Privacy is addressed by ensuring the pipeline does not retain user-specific learning data beyond feedback ratings, with all processing performed on course content rather than learner behavioral data.

6. CONCLUSION AND FUTURE WORK

This paper has presented an AI-powered multilingual content localization engine for skill courses designed to bridge the language accessibility gap in India's vocational training ecosystem. The proposed system integrates Whisper ASR, MarianMT NMT, RAG-based domain glossary alignment, and Coqui TTS into a unified five-layer pipeline capable of localizing both text- based study content and video speech dubbing across 22 Indian languages. The architecture addresses key challenges including domain terminology preservation, regional voice naturalness, audio-video synchronization, and LMS platform integration. The modular design and feedback-driven improvement loop ensure progressive refinement to meet the evolving needs of national skilling programs.

Several directions for future research are identified:

1. Fine-tuning of MarianMT and Whisper models on skill-domain corpora to improve accuracy for technical vocabulary and domain-specific speech patterns.
2. Development of dedicated regional voice models for underrepresented language groups to improve TTS naturalness MOS scores.
3. Investigation of speech rate normalization algorithms to improve audio-video synchronization for language pairs with significant duration differences.
4. Integration with actual LMS platforms such as Skill India Digital and SWAYAM to validate the system in real-world educational deployment scenarios.
5. Exploration of large multilingual language models such as IndicBERT and mBART50 for improved cross-lingual transfer, particularly for low-resource Indian languages.

7. ACKNOWLEDGMENT

The authors would like to thank the Department of Computer Science and Engineering at Geetanjali Institute of Technical Studies, Dabok, Udaipur, Rajasthan, for their support and resources during the development of this project. The authors also express gratitude to Project Guide Mr. Pankaj Vaishnav and Head of Department Dr. Mayank Patel for their invaluable guidance and constructive feedback throughout the project work.

8. REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[2] J. Tiedemann and S. Thottingal, "OPUS-MT - Building open translation services for the World," in Proc. 22nd Annual Conference of the European Association for Machine Translation, 2020, pp. 479-480

[3] A. Kunchukuttan, P. Mehta, and P. Bhattacharyya, "The IIT Bombay English-Hindi Parallel Corpus," in Proc. 11th International Conference on Language Resources and Evaluation, 2018, pp. 2828-2835.

[4] Tiwari, K., Patel, M. (2020). Facial Expression Recognition Using Random Forest Classifier. In: Mathur, G., Sharma, H., Bundele, M., Dey, N., Paprzycki, M. (eds) International Conference on Artificial Intelligence: Advances and Applications 2019. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-15-1059-5_15.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in Proc. 40th International Conference on Machine Learning, 2023.

[6] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, "JHU ASpIRE system: Robust LVCSR with time delay neural networks and IVECTOR adaptation," in Proc. IEEE ASRU Workshop, 2015, pp. 539-546.

[7] Shekhawat, V.S., Tiwari, M., Patel, M. (2021). A Secured Steganography Algorithm for Hiding an Image and Data in an Image Using LSB Technique. In: Singh, V., Asari, V.K., Kumar, S., Patel, R.B. (eds) Computational Methods and Data Engineering. Advances in Intelligent Systems and Computing, vol 1257. Springer, Singapore.

https://doi.org/10.1007/978-981-15-7907-3_35

[8] Patel, Mayank, Neelam Badi, and Amit Sinhal. "The role of fuzzy logic in improving accuracy of phishing detection system." International Journal of Innovative Technology and Exploring Engineering (2019): 3162-3164.

[9] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Golge, and M. A. Ponti, "YourTTS: Towards zero-shot multi-speaker TTS and zero- shot voice conversion for everyone," in Proc. 39th International Conference on Machine Learning, 2022.

[10] A. Baby, A. P. Thomas, N. Nishanthi, and H. A. Murthy, "Resources for Indian languages," in Proc. 1st Workshop on Technologies for MT of Low Resource Languages, 2016, pp. 46-54.

[11] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. Yih, T. Rocktaschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 9459-9474.