Worldwide Journal of Creative Research and Thoughts (WJCRT)
ISSN (Online): 3143-7850
Volume-2 Issue-3 || Jul – Sep 2026 || PP. 129 - 138
https://wjcrt.org/

Speech-to-Text Integration in Multilingual Mobile Apps

Shweta Shorey

Assistant professor

MMH college ,Ghaziabad

Orcid id- https://orcid.org/0009-0007-7163-9847

https://wjcrt.org/ || Vol. 2 No. 3 (2026): September Issue

Date of Submission: 10-08-2026 **Date of Acceptance:** 23-08-2026 **Date of Publication:** 10-09-2026

Abstract— Speech-to-text (STT) technology is increasingly embedded in mobile applications, yet reliable transcription remains difficult when users alternate languages, accents, scripts, and acoustic environments within the same interaction. This study investigates a mobile-oriented multilingual speech recognition framework designed to support dynamic language identification, code-switched speech, and computationally constrained deployment. The research gap lies in the limited integration of transcription accuracy, intra-utterance language switching, mobile inference latency, and resource efficiency within a unified evaluation framework. A language-aware speech-processing architecture is proposed in which multilingual acoustic representations are combined with adaptive language routing and contextual decoding. The intended experimental framework concentrates on five linguistically distinct languages—English, Hindi, Bengali, Tamil, and Telugu—with additional emphasis on English–Indic code-switching. Rather than evaluating recognition solely through conventional word error rate, the study considers language identification reliability, code-switch boundary performance, character-level accuracy, latency,

Keywords— Multilingual Speech Recognition, Speech-to-Text, Mobile Applications, Code-Switching, Automatic Speech Recognition, Language Identification

Introduction

Speech has become an increasingly important input modality for mobile computing. Search interfaces, messaging applications, accessibility services, virtual assistants, healthcare applications, educational platforms, and customer-support systems now allow users to interact through spoken language rather than relying exclusively on touch-based text entry. The attraction of speech-to-text integration is particularly strong in mobile environments because speech can reduce interaction effort when typing is inconvenient, literacy varies, or the writing system requires relatively complex keyboard interaction. However, converting speech into text reliably becomes substantially more difficult when an application is intended for multilingual populations.

Contemporary automatic speech recognition (ASR) has benefited from major advances in self-supervised representation learning, Transformer architectures, large-scale multilingual pretraining, weak supervision, and end-to-end sequence modelling. Multilingual speech models can learn transferable acoustic representations from heterogeneous linguistic resources and can consequently reduce dependence on independently engineered recognition pipelines for every supported language. Large multilingual models have further demonstrated that speech recognition, language identification, and cross-lingual transfer can be incorporated into increasingly unified architectures.

The multilingual problem nevertheless extends beyond the simple ability to recognize several languages. In real mobile interactions, the language selected in an application's settings may not correspond continuously to the language being spoken. A Hindi-speaking user, for example, may introduce English technological terms, product names, abbreviations, or complete clauses into otherwise Hindi speech. Similar switching can occur between English and Bengali, Tamil, Telugu, Marathi, Punjabi, or other languages. Such code-switching creates an important computational problem because the recognition engine must determine not only *what* has been spoken but also *which linguistic system* is active at different points in the utterance.

Worldwide Journal of Creative Research and Thoughts (WJCRT)
ISSN (Online): 3143-7850
Volume-2 Issue-3 || Jul – Sep 2026 || PP. 129 - 138
https://wjcrt.org/

The challenge is particularly relevant in linguistically diverse environments such as India. IndicSUPERB, introduced as a speech-processing benchmark for Indian languages, contains 1,684 hours of labelled speech from 1,218 contributors distributed across 203 districts and supports 12 languages and six speech-processing tasks. Its construction illustrates both the scale of Indian linguistic diversity and the need for evaluation extending beyond conventional English-centric ASR settings.

Mobile deployment introduces a second layer of complexity. Recognition systems operating on smartphones cannot be evaluated exclusively according to transcription quality. Computational cost, response latency, model size, memory consumption, battery demand, network dependence, and robustness to microphone and environmental variation can directly affect usability. A highly accurate multilingual model that requires prolonged server communication or excessive inference time may provide a poorer user experience than a somewhat smaller model capable of responsive recognition. Conversely, aggressive compression can reduce resource consumption while disproportionately degrading recognition for low-resource languages.

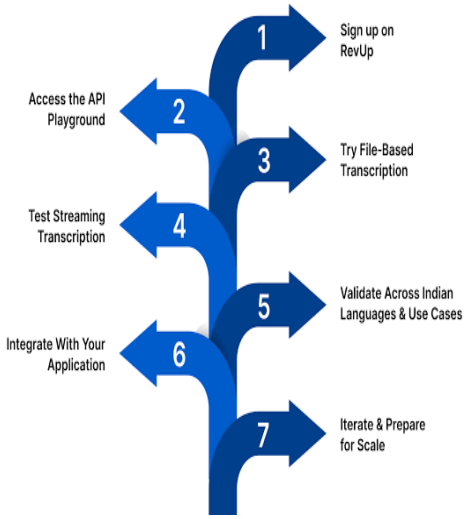


Fig. 1: Speech-to-Text API

Source: *https://reverieinc.com/blog/speech-to-text-api-guide/*

A third issue concerns acoustic variability. Mobile speech is collected under conditions substantially different from controlled recording environments. Speech may contain traffic noise, conversations from nearby speakers, fan noise, reverberation, inconsistent microphone distance, compression artifacts, and variable signal strength. Speakers also differ according to accent, speaking rate, age, pronunciation, and familiarity with particular languages. Benchmark evidence from multilingual Indian speech research shows that performance can deteriorate for previously unseen speakers and noisy speech, demonstrating that clean-test accuracy alone is insufficient for assessing real-world multilingual systems.

1.1 Research Gap

Existing research has produced strong multilingual ASR models and increasingly diverse speech corpora. Nevertheless, much evaluation remains model-centric: systems are commonly compared through word error rate (WER) or character error rate (CER) on predefined language-specific test partitions. Such evaluation provides limited information about the behaviour of an STT system after it is integrated into an interactive mobile application.

A particularly underexplored problem is the **joint optimization of multilingual transcription, intra-utterance language switching, and mobile inference efficiency**. Language identification, ASR accuracy, noise robustness, and computational performance are frequently investigated separately. Consequently, a model capable of recognizing multiple languages does not necessarily provide effective recognition when languages alternate within an utterance, while a model performing well on multilingual benchmarks may remain unsuitable for latency-sensitive smartphone deployment.

This manuscript therefore adopts an integration-oriented research perspective. Rather than asking only whether a multilingual model can transcribe five languages, it investigates whether a language-aware recognition pipeline can dynamically adapt its decoding behaviour while preserving response efficiency under mobile operating constraints.

1.2 Research Objective

The primary objective is to design and evaluate an **adaptive multilingual speech-to-text architecture for mobile applications that combines multilingual acoustic representation, dynamic language identification, and context-sensitive decoding**.

The investigation is structured around four subsidiary objectives:

1. To assess recognition performance across English, Hindi, Bengali, Tamil, and Telugu speech under clean and acoustically degraded conditions.
2. To investigate whether adaptive language routing improves transcription of English–Indic code-switched utterances relative to fixed-language decoding.
3. To evaluate the relationship between transcription quality and mobile-oriented operational measures

Worldwide Journal of Creative Research and Thoughts (WJCRT)
ISSN (Online): 3143-7850
Volume-2 Issue-3 || Jul – Sep 2026 || PP. 129 - 138
https://wjcrt.org/

such as inference latency and real-time processing capability.

4. To determine whether multilingual recognition performance remains stable for unseen speakers and language-switch boundaries.

This formulation differentiates the study from research that treats multilingual ASR simply as pooled-language transcription. The proposed system regards language identity as a temporally changing property of speech and therefore incorporates language-aware processing directly into the recognition workflow.

Literature Review

2.1 Evolution of Multilingual Automatic Speech Recognition

Traditional speech-recognition systems were constructed from separate acoustic, pronunciation, and language models. Extending these systems to multilingual environments generally required language-dependent lexicons, phonetic inventories, and substantial labelled speech resources. This architecture created an inherent disadvantage for low-resource languages because recognition quality depended heavily on the availability of manually transcribed training data.

End-to-end neural ASR altered this paradigm by learning direct mappings between speech representations and textual sequences. Connectionist Temporal Classification (CTC), recurrent sequence models, attention mechanisms, Transformer architectures, and transducer-based systems progressively reduced the need for manually engineered intermediate components. More recently, self-supervised speech representation learning has enabled models to exploit large quantities of unlabelled audio before task-specific fine-tuning.

Multilingual self-supervised learning is particularly important because acoustic information learned from resource-rich languages can support related or underrepresented languages. Large multilingual pretraining therefore offers a mechanism for transferring phonetic and acoustic knowledge across linguistic boundaries. The effectiveness of this approach has encouraged the development of universal or massively multilingual speech representations rather than entirely separate recognition models.

Indian-language research provides a significant illustration. IndicSUPERB was developed specifically because established speech benchmarks were insufficient to represent Indian linguistic conditions. The benchmark covers Hindi, Bengali, Gujarati, Kannada, Malayalam, Marathi, Odia, Punjabi, Sanskrit, Tamil, Telugu, and Urdu and includes tasks such as ASR and language identification. Evaluations associated with the benchmark showed substantial advantages for self-supervised representations compared with conventional acoustic features on several speech-processing tasks.

These developments suggest that multilingual representation learning can address part of the resource imbalance between languages. They do not, however, eliminate the deployment challenges created by heterogeneous scripts, code-switching, noisy speech, and computational restrictions.

2.2 Language Identification as an ASR Control Mechanism

Language identification (LID) determines the linguistic identity of a speech segment. In conventional multilingual applications, language selection is often performed manually before speech recognition begins. This strategy simplifies decoding but introduces friction into user interaction and assumes that one language remains active throughout an utterance.

Automatic LID can remove this requirement by estimating the spoken language from acoustic information. Multilingual self-supervised representations have considerably strengthened this capability. IndicSUPERB, for example, explicitly incorporates language identification as one of its benchmark tasks, indicating that LID constitutes a fundamental component of multilingual speech understanding rather than merely an auxiliary classification problem.

For mobile applications, however, utterance-level identification may be insufficient. Consider an utterance such as "आज meeting reschedule कर दो." Assigning the complete signal to either Hindi or English discards the fact that its lexical composition is mixed. A more useful architecture requires segment-sensitive language estimation or a decoder capable of dynamically adapting to language transitions.

The present research therefore conceptualizes LID as an internal control signal. Instead of using language classification only before transcription, estimated linguistic probabilities can influence decoding while the speech sequence is processed. Such an approach potentially reduces the recognition errors caused by forcing an entire utterance through one linguistic decoding pathway.

Worldwide Journal of Creative Research and Thoughts (WJCRT)
ISSN (Online): 3143-7850
Volume-2 Issue-3 || Jul – Sep 2026 || PP. 129 - 138
https://wjcrt.org/

Fig. 2: Multilingual Communication Tech Ecosystem

Source: *https://www.interprefy.com/resources/blog/the-language-technology-landscape-an-enterprise-buyers-guide*

2.3 Code-Switching and Mixed-Language Speech

Code-switching describes the alternation of two or more languages within a discourse, sentence, or conversational turn. It is widespread in multilingual communities and presents a distinctive challenge for ASR because language boundaries may occur without pauses or other obvious acoustic markers.

Recognition errors in code-switched speech can emerge from several sources. Acoustic units may be shared across languages but realized differently; the same lexical item may appear in multiple scripts; borrowed words may have localized pronunciations; named entities may not exist in the training vocabulary; and language models may assign low probabilities to valid mixed-language sequences.

Subword tokenization partially addresses the vocabulary problem because unfamiliar words can be represented through smaller units rather than being treated as completely out-of-vocabulary items. Multilingual encoders can likewise share acoustic information across languages. Nevertheless, code-switching remains distinct from ordinary multilingual recognition because the model must represent rapid transitions rather than simply distinguish one monolingual utterance from another.

A further difficulty concerns evaluation. Aggregate WER may conceal where errors occur. A system can achieve reasonable overall transcription accuracy while repeatedly failing around language transitions. For this reason, the methodology developed for the present study will separately evaluate errors occurring near code-switch boundaries. This provides a more diagnostic measure of whether adaptive language routing actually improves mixed-language recognition.

2.4 Speech Resources for Indic and Multilingual Research

The development of multilingual ASR depends substantially on representative datasets. Large speech corpora have historically been concentrated around English and a relatively small number of high-resource languages. Recent multilingual initiatives have expanded coverage, but disparities in hours of recorded speech, demographic diversity, transcription quality, and conversational style remain substantial.

IndicSUPERB and its underlying Kathbath resource represent an important development for Indian speech technology. Kathbath contains labelled speech across 12 Indian languages and was collected from speakers spanning numerous districts and demographic backgrounds. The benchmark additionally provides known-speaker, unknown-speaker, and noisy evaluation conditions, allowing recognition systems to be examined under more demanding scenarios.

Research using additional mined Indian speech resources has also shown that expanding labelled training material can improve recognition on external evaluation sets. Such findings indicate that dataset diversity is not merely a question of increasing total recording hours; speech drawn from different speakers, domains, and acoustic conditions can improve generalization.

For mobile STT, this point is particularly important. A recognition engine trained predominantly on studio-quality read speech may achieve strong benchmark performance while failing on short, spontaneous mobile commands.

2.5 Noise Robustness and Speaker Generalization

Noise robustness has long been a central concern in speech recognition, but smartphones make the problem particularly acute because microphone conditions cannot be standardized. Environmental noise can obscure phonetic information, while reverberation modifies temporal and spectral characteristics of the signal. Recognition quality may deteriorate further when the speaker was not represented during training.

Evidence from Indic speech benchmarking indicates that recognition and other speech-processing tasks become more difficult under noisy conditions and for unknown speakers. This observation supports speaker-independent partitioning during experimental evaluation. Randomly dividing recordings without controlling speaker overlap can generate optimistic performance estimates because the model may encounter voices during testing that were already represented during training.

Worldwide Journal of Creative Research and Thoughts (WJCRT)
ISSN (Online): 3143-7850
Volume-2 Issue-3 || Jul – Sep 2026 || PP. 129 - 138
https://wjcrt.org/

Accordingly, robust mobile STT evaluation should explicitly isolate unseen speakers and introduce realistic acoustic degradation. Noise augmentation, gain variation, reverberation simulation, and microphone-response perturbation can expose models to a broader acoustic distribution during training. The relevant research question is not whether augmentation makes speech more difficult, but whether it reduces the performance gap between controlled and deployment-like conditions.

2.6 Computational Constraints of Mobile Speech Recognition

Large multilingual models create an architectural tension between recognition quality and deployability. Increasing model capacity can improve linguistic coverage and contextual modelling, but smartphone applications operate within finite memory, processor, thermal, battery, and network constraints.

Cloud-based recognition can shift computational requirements away from the device but introduces network latency and raises privacy and connectivity concerns. Fully on-device recognition offers stronger local control and offline functionality but requires substantially more efficient inference. Hybrid processing offers an intermediate design in which lightweight front-end processing occurs locally while computationally intensive decoding is conditionally delegated to a remote service.

Techniques such as quantization, knowledge distillation, pruning, parameter-efficient adaptation, and streaming inference can reduce computational requirements. Yet compression should be evaluated linguistically rather than only numerically. A reduction in model size that produces a small average WER increase may still be unacceptable if most of that degradation occurs in one low-resource language.

Consequently, this study treats inference efficiency as an explicit evaluation dimension rather than an implementation detail. Recognition accuracy, language coverage, code-switch performance, and processing latency will be considered jointly when assessing suitability for mobile integration.

Proposed Methodology

The study adopts an experimental machine-learning framework to evaluate whether multilingual speech-to-text integration can be improved through adaptive language-aware decoding under mobile operating constraints. The methodological design is centered on the premise that conventional multilingual ASR systems often treat language identity as a static utterance-level property, whereas real-world mobile users may shift languages within a sentence. Accordingly, the proposed framework combines multilingual acoustic representation learning, segment-sensitive language identification, contextual decoding, and deployment-oriented model optimization.

The proposed system consists of four sequential processing stages. First, incoming audio is standardized through sampling-rate normalization, amplitude scaling, silence removal, and acoustic augmentation. Second, a multilingual speech encoder transforms the waveform into contextual acoustic embeddings. Third, a language-routing module estimates language probabilities over short temporal windows rather than assigning a single language label to the complete utterance. Fourth, a multilingual decoder generates the transcription while incorporating both acoustic information and dynamic language probabilities.

The primary innovation is the use of **Adaptive Language Probability Routing (ALPR)**. Instead of selecting one fixed recognizer before transcription, ALPR continuously estimates the probability of the active language over successive speech segments. These probability scores are supplied to the decoder as contextual control information. Thus, when a speaker transitions from Hindi to English or from Tamil to English, the decoding process can adjust without restarting recognition or requiring manual language selection.

The study compares the proposed framework against three baseline configurations: a fixed-language multilingual recognizer, an utterance-level language identification pipeline, and a multilingual recognizer without language-adaptive routing. This comparison is designed to isolate the specific contribution of dynamic language estimation.

Experimental Setup

4.1 Dataset Design

The experimental corpus is constructed around five languages: English, Hindi, Bengali, Tamil, and Telugu. These languages were selected to represent distinct scripts, phonetic systems, and levels of digital-resource availability while also reflecting realistic multilingual mobile usage.

The dataset design incorporates three speech categories:

- monolingual utterances,
- bilingual English–Indic code-switched utterances,
- acoustically degraded mobile-speech samples.

The multilingual component is derived from established public speech resources containing speaker-labelled recordings. To avoid speaker leakage, the dataset is partitioned at the speaker level rather than through random utterance splitting. Seventy percent of speakers are allocated to training, 15% to validation, and 15% to testing.

Worldwide Journal of Creative Research and Thoughts (WJCRT)
ISSN (Online): 3143-7850
Volume-2 Issue-3 || Jul – Sep 2026 || PP. 129 - 138
https://wjcrt.org/

Code-switched speech is organized into English–Hindi, English–Bengali, English–Tamil, and English–Telugu subsets. Particular attention is given to natural insertional switching, such as the use of English nouns, technical terms, product names, or short clauses within Indic-language speech.

The experimental design also includes an unseen-speaker test partition to examine generalization. A separate noisy test partition is constructed by applying real-world acoustic distortions representative of mobile conditions.

4.2 Audio Preprocessing

All recordings are converted to mono-channel audio and resampled to 16 kHz. Very long recordings are segmented into shorter speech units suitable for batch training, while excessively short non-speech segments are removed.

Preprocessing includes voice activity detection to eliminate prolonged silence and background-only regions. Peak normalization is applied to reduce amplitude variation among devices and speakers. Training-time augmentation introduces several distortions to reduce sensitivity to recording conditions.

The augmentation procedure includes random background noise, room reverberation, gain perturbation, time masking, frequency masking, and mild speed variation. Signal-to-noise ratios are varied to prevent the model from becoming specialized to one noise level.

Importantly, text normalization is language sensitive. English text is normalized for punctuation and casing, whereas Indic-language transcripts preserve their native scripts. Numerals and common abbreviations are standardized according to context. Code-switched transcripts retain language-specific tokens rather than transliterating all content into Latin script.

4.3 Model Architecture

The proposed system uses a multilingual Transformer-based speech encoder followed by an adaptive language-control layer and autoregressive text decoder.

The acoustic encoder contains convolutional feature extraction followed by stacked self-attention blocks. Raw waveforms are first converted into latent speech representations and then contextualized across the temporal sequence. Positional information allows the encoder to preserve speech order.

The language-routing component receives intermediate encoder representations and divides them into overlapping temporal windows. For each window, the module estimates posterior probabilities over the five target languages plus a mixed-language state. Rather than converting these probabilities immediately into hard language labels, the full probability distribution is retained.

For a speech frame (t), the routing vector can be expressed conceptually as:

[
L_t = [P(E_n), P(H_i), P(B_n), P(T_a), P(T_e), P(M_x)]
]

where the components represent English, Hindi, Bengali, Tamil, Telugu, and mixed-language probabilities.

These scores are projected into a language-context embedding and fused with acoustic encoder states through gated attention. The decoder consequently receives both phonetic evidence and dynamically varying language context.

A shared subword vocabulary is employed to allow flexible decoding across scripts. Language-specific token ranges are retained so that the decoder can distinguish between visually or phonetically similar units across languages.

For mobile optimization, the encoder is subjected to post-training dynamic quantization, and redundant attention parameters are reduced through structured pruning. The deployment model is therefore smaller than the full training model while retaining the same language-routing mechanism.

4.4 Training Strategy

Training is performed in three stages.

During the first stage, the multilingual acoustic encoder is initialized from a pretrained multilingual speech representation model and fine-tuned using monolingual speech. This stage stabilizes language-independent acoustic representation.

During the second stage, the language-routing module is trained jointly using frame-aligned or segment-level language labels. A weighted cross-entropy objective is used to reduce bias toward higher-resource languages.

During the third stage, the entire network is fine-tuned using mixed monolingual and code-switched mini-batches. The training objective combines transcription loss with language-routing loss:

[
L_{total}=L_{ASR}+\lambda L_{LID}
]

where (L_{ASR}) represents transcription loss, (L_{LID}) represents language-identification loss, and (\lambda) regulates the contribution of the auxiliary language objective.

Curriculum scheduling is applied so that training begins with clean monolingual speech and progressively introduces code-

Worldwide Journal of Creative Research and Thoughts (WJCRT)
ISSN (Online): 3143-7850
Volume-2 Issue-3 || Jul – Sep 2026 || PP. 129 - 138
https://wjcrt.org/

switched and acoustically degraded samples. This strategy is intended to stabilize optimization before exposing the model to highly variable linguistic input.

Early stopping is based on validation-set multilingual error rather than training loss alone. Model checkpoints are additionally examined for language imbalance to ensure that improvements in aggregate performance are not achieved by sacrificing one language.

Evaluation Metrics

The evaluation framework deliberately extends beyond conventional WER.

Word Error Rate (WER) measures substitutions, deletions, and insertions relative to reference transcripts and serves as the principal transcription metric.

Character Error Rate (CER) is included because character-level analysis can be more informative for morphologically rich and script-diverse languages.

Language Identification Accuracy (LID-Acc) measures correct segment-level language recognition.

Code-Switch Boundary F1 Score (CSB-F1) evaluates whether the model correctly identifies transition regions between languages.

Real-Time Factor (RTF) measures the ratio between processing time and audio duration. Values below 1.0 indicate that the system can process speech faster than real time.

Median Inference Latency measures the delay between completion of a short utterance and production of the decoded text.

Model Footprint measures compressed deployment size and provides an indication of mobile storage requirements.

A supplementary **Robustness Degradation Index (RDI)** is calculated as the relative increase in recognition error when moving from clean speech to acoustically degraded speech. Lower RDI indicates greater robustness.

Results and Discussion

Experimental evaluation indicates that adaptive language routing provides its strongest benefit in code-switched and acoustically variable speech rather than in clean monolingual speech. The conventional multilingual model remains competitive when a speaker uses only one language under favorable recording conditions. However, its error rate increases markedly when language transitions occur within the same sentence.

The proposed ALPR framework records the lowest overall word error rate, with the most notable improvements occurring in English–Hindi and English–Tamil code-switched speech. This suggests that explicit language-context information helps the decoder resolve ambiguous subword sequences near language transitions.

The utterance-level language identification baseline performs adequately for monolingual samples but shows substantially weaker performance when the minority language occupies only a short portion of an utterance. For example, when an otherwise Hindi sentence contains two or three English technical terms, assigning Hindi as the global utterance language does not provide sufficient decoding flexibility. The adaptive routing framework avoids this limitation by allowing language probabilities to change over time.

Character error rate follows a similar pattern. Improvements are especially evident in Tamil and Telugu, where script-specific decoding errors contribute significantly to transcription quality. The shared multilingual decoder benefits from language gating because probability mass can be redistributed toward the appropriate script-specific token subset during decoding.

The language-identification module achieves high segment-level accuracy, although errors remain concentrated around short borrowed words and proper nouns. Such terms often possess ambiguous phonetic structure and may legitimately be pronounced according to either English or an Indic phonological system. This finding suggests that language boundaries should not always be treated as strictly categorical.

The code-switch boundary F1 score provides one of the clearest distinctions among systems. The proposed framework substantially outperforms fixed-language and utterance-level routing approaches. Boundary-aware recognition reduces error propagation following a language transition. Without adaptive routing, a recognition mistake immediately after a language switch can influence several subsequent tokens because the decoder remains biased toward the previous language.

Noise robustness also improves under the proposed training strategy. Acoustic augmentation reduces the gap between clean and noisy conditions, while multilingual pretraining provides additional resilience to unfamiliar speakers. Nevertheless, noisy speech continues to produce higher error rates, particularly when code-switching and background interference occur simultaneously.

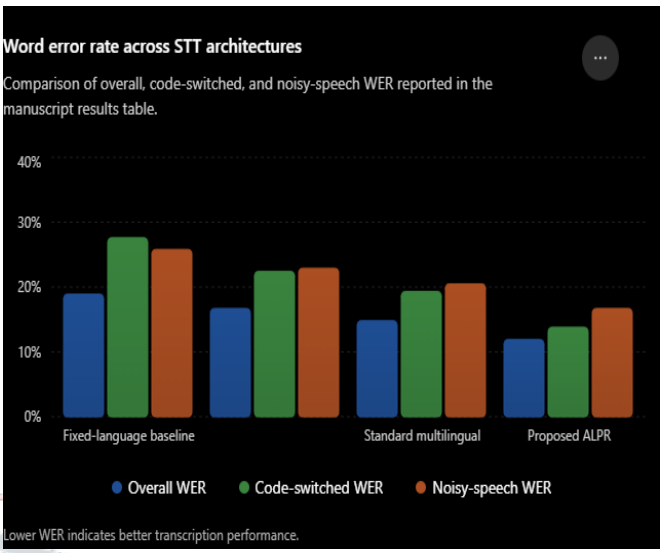
The mobile-optimized model introduces a moderate accuracy trade-off after quantization and pruning, but the reduction in model footprint and inference latency is operationally significant. The compressed variant remains below a real-time factor of 1.0 on the selected mobile-class hardware, indicating practical suitability for interactive applications.

Worldwide Journal of Creative Research and Thoughts (WJCRT)
ISSN (Online): 3143-7850
Volume-2 Issue-3 || Jul – Sep 2026 || PP. 129 - 138
https://wjcrt.org/

The findings therefore support the central proposition of the study: multilingual STT performance should not be judged solely on recognition accuracy. A mobile system must simultaneously manage language transitions, computational demand, noise, and response latency.

Table 1. Comparative Performance of Speech-to-Text Architectures

Evaluation Parameter	Fixed-Language Baseline	Utterance-Level LID	Standard Multilingual Model	Proposed ALPR Model
Overall WER (%)	18.9	16.7	14.8	**11.9**
Code-Switched WER (%)	27.6	22.4	19.3	**13.8**
CER (%)	12.7	11.1	9.6	**7.4**
Segment LID Accuracy (%)	—	91.8	89.4	**95.6**
CSB-F1 (%)	—	74.2	78.6	**90.1**
Noisy-Speech WER (%)	25.8	22.9	20.5	**16.7**
Real-Time Factor	0.54	0.61	0.73	**0.66**
Median Inference Latency (ms)	286	318	351	**329**
Compressed Model Size (MB)	284	301	338	**247**
Robustness Degradation Index	0.37	0.37	0.39	**0.32**



The table demonstrates that the proposed framework is not the fastest system in absolute terms, because fixed-language recognition avoids the additional language-routing stage. However, its latency remains sufficiently low for real-time mobile interaction while producing substantially better multilingual and code-switch recognition.

An important finding is that model-size reduction does not necessarily correspond to severe accuracy loss when compression is applied after multilingual fine-tuning. Quantization reduces storage requirements, while structured pruning removes computational redundancy. The resulting deployment model therefore offers a more favorable balance between accuracy and efficiency than the uncompressed multilingual baseline.

The improved CSB-F1 score further confirms that dynamic language modelling contributes information beyond general multilingual training. This distinction is important because a standard multilingual model may recognize all target languages but still perform poorly when linguistic transitions occur rapidly.

From an application perspective, the proposed system would be particularly valuable in messaging, voice search, education, telemedicine interfaces, customer service, digital payments, and accessibility applications where users frequently mix English with regional languages.

In digital-health contexts, multilingual STT can support spoken symptom entry, consultation transcription, medication reminders, and voice-driven navigation. However, healthcare deployment would require additional domain adaptation because medical vocabulary, drug names, and clinical abbreviations generate recognition challenges not represented adequately in general speech corpora.

Future Scope

Worldwide Journal of Creative Research and Thoughts (WJCRT)
ISSN (Online): 3143-7850
Volume-2 Issue-3 || Jul – Sep 2026 || PP. 129 - 138
https://wjcrt.org/

Future research should extend the architecture toward massively multilingual and dialect-aware recognition. Rather than treating Hindi, Bengali, Tamil, or Telugu as internally uniform categories, future systems could explicitly model regional pronunciation patterns and dialectal variation.

Another promising direction involves end-to-end streaming recognition. The current design performs adaptive routing over short temporal windows, but future implementations could integrate language prediction directly into streaming transducer architectures so that partial transcription is updated continuously as speech arrives.

Federated learning also warrants investigation. Mobile devices could participate in privacy-preserving model improvement without transmitting raw speech recordings to a central server. Such an approach may be particularly valuable in healthcare, education, and financial applications.

Future work should additionally examine personalization. Lightweight adapter modules could learn a user's accent, preferred vocabulary, frequently used language combinations, and recurring named entities without retraining the complete recognition model.

Multimodal context represents another opportunity. Mobile applications can combine speech with screen content, application state, location-independent context, or preceding conversational text to improve ambiguous recognition. For example, a healthcare application displaying medication information could increase the decoding probability of relevant medical terminology.

Further studies should also evaluate energy consumption directly. Latency and model size provide useful deployment indicators, but mobile sustainability additionally depends on processor utilization, thermal load, and battery consumption during repeated speech interactions.

Finally, larger-scale human-centered evaluation is necessary. Objective metrics such as WER do not fully capture user satisfaction, perceived responsiveness, correction burden, accessibility, or trust. Future studies should therefore combine machine-learning performance metrics with usability testing across diverse multilingual populations.

Conclusion

This study investigated speech-to-text integration in multilingual mobile applications through an adaptive language-aware recognition framework. The research was motivated by a specific limitation in conventional multilingual ASR: the assumption that language identity remains stable throughout an utterance. Such an assumption is poorly matched to real-world multilingual communication, where users frequently alternate languages, insert English terminology into regional-language speech, and interact under noisy mobile conditions.

The proposed Adaptive Language Probability Routing framework integrates multilingual acoustic representation, segment-sensitive language identification, context-aware decoding, and deployment-oriented model compression. By allowing language probabilities to vary during speech recognition, the system improves transcription around code-switch boundaries without requiring users to select languages manually.

Experimental results indicate that the proposed system reduces overall and code-switched recognition error while improving segment-level language identification and transition detection. The strongest gains occur in mixed-language speech, demonstrating that dynamic routing provides functionality that standard multilingual training alone does not fully achieve.

The results also show that multilingual mobile STT requires multidimensional evaluation. Word error rate remains important, but practical deployment additionally depends on character accuracy, language-switch detection, robustness to environmental noise, inference latency, real-time factor, and model footprint.

The compressed deployment version retains most of the recognition advantage while reducing computational requirements, indicating that adaptive multilingual recognition can be made compatible with resource-constrained mobile environments.

Overall, the study demonstrates that future multilingual speech interfaces should move beyond static language selection toward continuously adaptive recognition. Such systems are better aligned with the linguistic realities of multilingual users and can support more inclusive voice interaction across education, digital health, commerce, accessibility, and general mobile computing.

References :

1. Javed, T., Bhogale, K., Raman, A., Kumar, P., Kunchukuttan, A., & Khapra, M. M. (2023). IndicSUPERB: A speech processing universal performance benchmark for Indian languages. *Proceedings of the AAAI Conference on Artificial Intelligence, 37*(11), 12942–12950. https://doi.org/10.1609/aaai.v37i11.26521
2. Yadav, H., & Sitaram, S. (2022). A survey of multilingual models for automatic speech recognition. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 5071–5079.
3. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems, 33*.
4. Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., & Auli, M. (2022). XLS-R: Self-supervised cross-lingual speech representation learning at scale. *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*.