A FastConformer Framework for Dialect-Inclusive Kannada Speech Recognition

Alaka Ananth
Nitte (Deemed to be University) NMAM Institute of Technology, Nitte, Udupi, India | Visvesvaraya Technological University, Belagavi, India
alaka.bhoomi@gmail.com (corresponding author)

P.S. Venugopala
Nitte (Deemed to be University) NMAM Institute of Technology, Nitte, Udupi, India
venugopalaps@nitte.edu.in

Sachin S. Bhat
Shri Madhwa Vadiraja Institute of Technology and Management, Bantakal, India
sachinbhat88@gmail.com

Received: 10 February 2026 | Revised: 14 March 2026 and 27 March 2026 | Accepted: 6 April 2026

 DOI: https://doi.org/10.48084/etasr.18083

ABSTRACT

Despite advances in Automatic Speech Recognition (ASR), low-resource languages such as Kannada suffer from high Word Error Rates (WER), especially across different regional dialects. The present study addresses this issue by presenting a robust multi-dialect Kannada ASR system using a linguistically informed methodology based on a FastConformer architecture, fine-tuned using a carefully curated and dialect-balanced speech corpus representing six major regional dialects of Kannada. The approach introduces three novel elements: (1) dialect-aware curation, (2) unified dialect-invariant architecture, and (3) a controlled baseline framework to quantify the relative contributions of pretraining and architectural design. It employs character-level tokenization and full end-to-end adaptation with advanced architectural features such as convolutional subsampling and relative positional encoding, specifically tailored to address the phonotactic richness and morphological complexity of Kannada. The experimental results demonstrate state-of-the-art performance on both validation and test sets, achieving a WER of 11.23% and Character Error Rate (CER) of 5.31%, with real-time inference capabilities and consistent accuracy across dialectal boundaries. This represents a relative reduction of 15% compared to earlier Kannada baselines. Ablation and fine-tuning strategies confirm the significant contributions of each architectural component. The key contributions of this study include the development of the first multi-dialect Kannada speech corpus and the subsequent demonstration of an effective fine-tuning strategy for end-to-end speech recognition models. Beyond technical innovation, this work advances digital accessibility for Kannada speakers, enabling accurate and inclusive voice-driven technologies for diverse linguistic communities.

Keywords-FastConformer; Kannada; speech recognition; character embedding

I. INTRODUCTION

The development of Automatic Speech Recognition (ASR) systems has been driven by foundational theories, technological innovations, and evolving research paradigms [1]. In the 1960s, research primarily addressed rule-based systems and basic acoustic models, with phonetic rules manually encoded to recognize limited vocabularies. The adoption of Hidden Markov Models (HMMs) during the 1970s and 1980s established a statistical framework for modeling the temporal dynamics of speech, which improved recognition performance by capturing phoneme transition probabilities. In the 1990s, the integration of Gaussian Mixture Models (GMMs) enhanced the representation of acoustic feature distributions, allowing ASR systems to more effectively address speaker and environmental variability by modeling complex sound patterns.

With the introduction of Neural Networks, a significant change occurred in the field of speech processing [2]. Deep Neural Networks (DNNs) replaced GMMs to model acoustic observations. Later, Convolutional Neural Networks (CNNs) were used for robust feature extraction. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM), were used for capturing long-range temporal dependencies within speech [3, 4]. These algorithms model speech recognition by directly mapping raw acoustic inputs to text sequences, which streamlined the multi-stage ASR process. In the 2010s and beyond, transformer-based models

revolutionized ASR by employing self-attention mechanisms to learn both local and global contextual relationships in speech signals [5]. The Conformer architecture, which augments transformers with convolutional modules, has demonstrated superior performance on benchmark tasks by effectively modeling short-term spectral patterns alongside long-term dependencies [6, 7]. Concurrently, self-supervised learning techniques such as wav2vec 2.0 [8] and OpenAI's Whisper [9] pre-training frameworks have enabled the extraction of rich speech representations from large volumes of unlabeled audio [10]. These pre-trained models can be fine-tuned on smaller, labeled datasets to achieve state-of-the-art performance in low-resource scenarios, significantly reducing training time and computational overhead [11, 12].

Despite these advances, Kannada, a major Dravidian language spoken by over 70 million people, remains under-resourced. Existing corpora cover mostly standard Kannada, providing limited dialectal diversity, and the Word Error Rates (WER) reported by existing Kannada ASR systems are relatively high (often exceeding 20%) when evaluated on spontaneous or regionally variant speech [13–16].

To bridge these gaps, the present study proposes a linguistically informed fine-tuning approach for a pre-trained FastConformer ASR model, specifically adapted for multi-dialect Kannada recognition. This approach introduces three novel elements: (1) dialect-aware data curation—stratified speaker recruitment and phonetically contrastive utterance design targeting dialect-specific contrasts; (2) unified dialect-invariant architecture enabling single-model deployment across all six dialects; and (3) a controlled baseline framework quantifying pre-training versus architectural contributions. The key contributions of this study are:

- Development of a novel multi-dialect Kannada corpus: A linguistically diverse dataset of more than 450 speakers belonging to six different dialects across Karnataka, filling a crucial resource gap for Kannada ASR research.
- Fine-tuning strategy for end-to-end ASR models: An adapter-based fine-tuning of FastConformer models, achieving WER reductions of up to 15% relative to baseline models, with minimal additional training time.
- Evaluation and error analysis: The proposed approach is evaluated against recent Kannada ASR baselines, providing detailed error analysis across dialects, utterance types, and noise conditions to guide future improvements.

By focusing on dialectal coverage, self-supervised pre-training, and efficient adaptation, the study offers more inclusive, accurate, and accessible ASR technologies for 70M Kannada speakers. Although the corpus is carefully balanced across six dialects, its overall size (approximately 7.5 h) is modest compared to large ASR benchmarks, which may limit absolute performance and cross-domain generalization.

II. RELATED WORK

A. Overview of Kannada Speech Recognition (KSR)

In the domain of KSR, research has progressed from foundational models to advanced neural network-based approaches. Authors in [17] developed an acoustic model using HMMs alongside a lexicon derived from the Kannada pronunciation dictionary, achieving a word recognition accuracy of 87%. Subsequent studies introduced hybrid models that combined GMMs with HMMs [18, 19] and HMMs with DNNs [20, 21], yielding improved recognition accuracy and underscoring the potential of integrating neural architectures with traditional statistical models.

Further refinement of KSR systems saw the adoption of deep learning-based approaches employing CNNs and LSTMs, which significantly enhanced robustness and accuracy over earlier GMM HMM systems [22, 23]. Authors in [25] developed an ASR system tailored for the agricultural sector, demonstrating the practical applicability of KSR in domain-specific deployments and real-time query handling. Building on this foundation, significant improvements have been reported by incorporating background noise elimination and alternative acoustic modeling techniques, improving recognition accuracy under challenging acoustic conditions, and establishing a basis for effective real-world deployment [24, 25].

Research using DNN-based acoustic modeling has leveraged Mel Frequency Cepstral Coefficient (MFCC) features and explored a variety of architectures, including GMM HMM, Subspace GMM (SGMM), and DNN systems. A continuous KSR system implemented using the Kaldi toolkit validated the effectiveness of context-dependent triphone models over monophone systems in reducing WER and motivated further exploration of DNNs to boost ASR performance [26, 27].

B. Advanced Neural Network Approaches and End-to-End Systems

Multimodal and end-to-end KSR systems have been investigated. For example, a multimodal system combined MFCC-based audio features with visual features extracted using the Viola–Jones technique and achieved an accuracy of 89% when integrating both modalities, demonstrating the utility of audio-visual fusion, especially in noisy environments and assistive applications [29]. Noise robustness in real-time ASR has been addressed by incorporating magnitude and phase features from the Short Time Fourier Transform (STFT) within DNN frameworks, achieving a 1.59% relative WER reduction on degraded speech databases and supporting agricultural spoken query systems [29].

End-to-end KSR systems based on Time Delay Neural Networks (TDNNs) have shown further progress. Authors in [30] reported significant gains in isolated speech recognition by combining robust noise elimination with TDNNs, achieving a 1.1% improvement in recognition accuracy over earlier DNN HMM systems and effectively handling real-time agricultural queries from 500 farmers under challenging acoustic conditions. Authors in [31] expanded a continuous Kannada speech database by collecting data from an additional 300 speakers, applying advanced noise elimination techniques, and training models using Kaldi, thereby improving effectiveness to real-world noise. An end-to-end continuous KSR system based on TDNNs and tested with 550 speakers in uncontrolled

environments outperformed earlier DNN HMM models, validating its scalability for practical applications in agriculture and governance [32].

C. Multi-Dialect and Dialectal Robustness Research

The AI4Bharat initiative introduced IndicConformer models, which are state of the art Conformer based ASR systems designed for Indian languages, including Kannada. These models leverage Conformer encoders that combine CNNs with transformer self-attention to capture both local spectral patterns and long-range temporal dependencies, delivering strong performance on large-scale Indic benchmarks.

Dialectal variation in Kannada has drawn increasing attention. Authors in [32] developed an automatic dialect identification system for Kannada using spectral and prosodic features, achieving high accuracy in distinguishing between dialectal variants and laying the groundwork for dialect-aware ASR. Authors in [33] studied speech recognition across four major dialect groups—southern, coastal, northern, and eastern—using Real Cepstrum Coefficients and HMM-based modeling. They reported recognition accuracies between 75.64% and 85.45%, with continuous speech recognition outperforming isolated word recognition.

More recent work has focused on spontaneous and noisy speech. Authors in [34] proposed a system for spontaneous Kannada sentence recognition under uncontrolled conditions, using Perceptual Wavelet Packet features with DNN-HMM acoustic models and n-gram language models, achieving a 1.8% WER reduction over baseline features. Authors in [35, 36] introduced a noise robust end-to-end ASR system that integrates magnitude and phase features from STFT into a TDNN–LSTM framework, exploring phase-aware enhancement for Kannada ASR. In parallel, authors in [37] investigated enhancement and reconstruction of dysphonic Kannada speech using advanced neural architectures. The latest TDNN-based systems incorporating sophisticated noise elimination have reported an additional 1.87% improvement in WER compared to conventional spectral subtraction, highlighting ongoing progress toward robust real-world Kannada ASR [38].

The integration of transformer architectures, particularly Conformer models, with self-supervised pre-training offers superior performance while maintaining computational efficiency suitable for real-world deployment. These approaches have proven particularly effective for fine-tuning scenarios where computational resources are limited, making them ideal for dialectal adaptation tasks.

TABLE I. CONSOLIDATED OVERVIEW OF DIFFERENT KSR ARCHITECTURES

Ref.	Features	Model	Domain	Outcome / evaluation
[19]	MFCC	HMM	General	✓ Foundational work
[20]	MFCC	GMM-HMM	General	✓ Improved over monophones
[21]	MFCC	DNN	General	★ First DNN-based ASR
[22]	MFCC	GMM-HMM	General	☁ Focused on degraded audio
[23]	MFCC	LSTM	General	→ Temporal modeling
[24]	MFCC	TDNN	Agriculture	↑ +1.1% accuracy – real-time system
[25]	MFCC	Feedforward NN	Assistive Tech	–
[26]	MFCC	TDNN + LSTM	General	-
[27]	STFT	DNN	General	↓ 1.87% WER – noise robust
[28]	MFCC + Visual	Fusion + DNN	Assistive Tech	★ Audio-visual modality
[29]	MFCC	TDNN	Agriculture	↓For farmers
[30]	MFCC	TDNN	General	↓ Continuous ASR
[31]	Wavelet	DNN-HMM	General	↑ Improved spontaneous ASR
[32]	STFT	TDNN	General	✓ Spectral subtraction for noise
[33]	STFT	TDNN	General	★ Best reported WER

✓ Notable contribution or validation, ★ First-of-its-kind, ↑ Accuracy improvement, ↓ WER improvement, → Technical advancement, ☁ Handles noise, – Metric not reported.

D. Synthesis and Research Gaps

While the proposed methodology demonstrates promising results in KSR, many significant challenges remain. A primary limitation lies in the availability of labeled training data for specific dialects. As Kannada is a low-resource language with significant dialectal diversity, the development of high-performance ASR systems is obstructed by the scarcity of large-scale, annotated speech corpora covering all dialectal variants. This limitation affects both acoustic modeling and language modeling components.

Another key challenge is the presence of dialectal and regional variations within the Kannada language. Kannada has strong local dialects across its regions, like Kitturu Karnataka, North Karnataka, Coastal, and Southern parts. Current systems predominantly capture standard Kannada, and their performance may degrade when exposed to phonetic and lexical deviations common in regional dialects. Addressing this challenge requires multi-dialectal training that enhances generalization across linguistic subgroups.

III. METHODOLOGY

The detailed methodology for fine-tuning a FastConformer-based end-to-end ASR model on a multi-dialect Kannada corpus involves key phases such as signal processing, text normalization, vocabulary construction, model adaptation, encoder design, and training configuration. A detailed flowchart of the multi-dialect KSR system is shown in Figure 1.

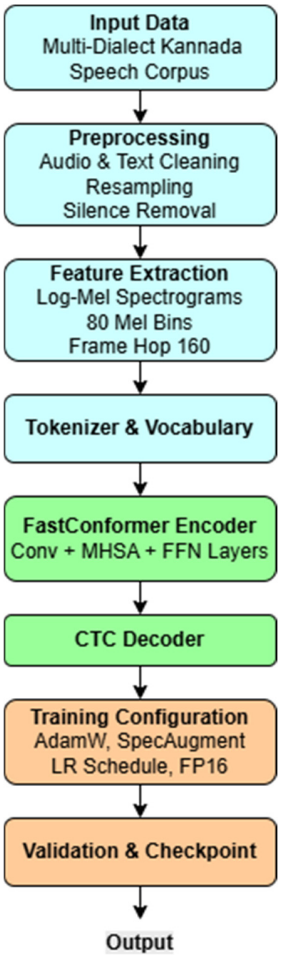

Fig. 1. Proposed end-to-end ASR pipeline.

Datasets serve as the foundational resource for training and evaluating language models for ASR in Natural Language Processing (NLP). Selection and transcription of audio recordings are necessary to build an appropriate speech corpus. The present study recorded a gender-balanced dataset from a diverse group of speakers (241 male and 253 female) aged between 18 and 45 years. Six different Kannada dialects were considered while preparing the dataset, namely Granthika Kannada (standard literary), Dharwad Kannada (northern dialect), Kundagannada (coastal), Arebhase (Mangaluru region), Havyaka Kannada (community dialect), and Halegannada (ancient Kannada recitations). A controlled environment setting with less background noise was used while recording the audio data. For general ASR tasks, a small amount of background noise is acceptable as it makes the model resilient in real-world situations. To provide the model with sufficient context to learn from and to ensure that the segments are computationally manageable during training, each audio file was further divided into segments of 5-20 s. In addition to matching the transformer model's capacity, this segmentation keeps the model from overfitting on extremely brief or lengthy utterances. For this effort, a total corpus of 7:30:24 h was developed. Of this, the train, validation, and test sets were divided as depicted in Figure 2.

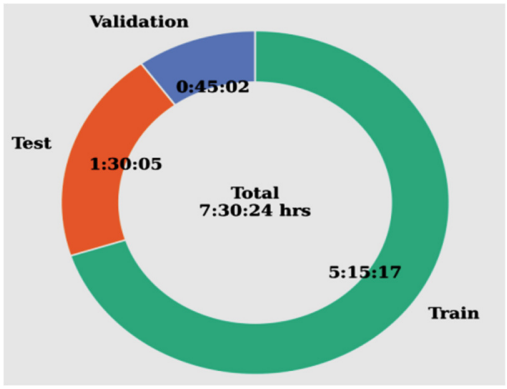

Fig. 2. Dataset division.

The dataset has two parts: audio and corresponding text transcriptions. Special characters, punctuation, and symbols irrelevant to the ASR task are removed using a predefined regex pattern. The transcriptions are normalized using Unicode NFC to handle diacritics and ligatures specific to the Kannada script. The audio preprocessing begins with each recording resampled to 16 kHz to standardize sampling rates across sessions. The waveforms undergo Root Mean Square (RMS) normalization, targeting an RMS level of 0.1, and silence trimming using an energy thresholding algorithm. These steps ensure consistent loudness and remove leading and trailing silence without discarding meaningful low-energy segments, thus preserving speaker variability essential for robust modeling.

The example sentences from different dialects are:

- Arebhase - ಬೊದ್ಕ್ ಆಧುನಿಕತೆಗೆ ಹೊಂದಿಕಂಡ್ ಹೋಗ್ತಾ ಇದ್ದಂಗೆ ಅರೆಭಾಷೆ ಮಾತಾಡುವವರ ಸಂಖ್ಯೆ ಸಕಮ್ಮಿಆಕಂಡ್ ಹೋಗ್ತಾ ಉಟ್ಟು
- Havigannada – ಮಧ್ಯಾಹ್ನದ ರಣ ಬೆಶಿಲಿಂಗೆ ಒಪ್ಪಿಯಬೆಳ್ಳಿಯ ಹಾಂಗಿಪ್ಪ ಕೂದಲು, ಪಳಪಳನೆ ಹೊಳಕ್ಕೊಂಡಿತ್ತು.
- Kundagannada – ಕುಂದಾಪ್ರ ಕನ್ನಡ ಭಾಷಿ ಅಲ್ಲ ಬದ್ಕ್
- Dharwad Kannada - ಅಂಗಡಿಗಿ ಹೋಗಿ ಮನಿಗಿ ಏನ್ ಬೇಕು ತಗೊಂಡ್ ಬಾ
- Halegannada - ಸೊಗಯಿಸಿ ಬಂದ ಮಾಮರನೆ ತಳ್ತೆಲೆವಳ್ಳಿಯೆ ಪೂತ ಜಾತಿ

Each normalized waveform is framed into 400-sample windows with a 160-sample hop using a Hann window. The STFT yields complex spectra, whose magnitudes are filtered through 80 Mel-spaced filterbanks. The resulting Mel spectrogram values $M(f,t)$ are log-compressed as:

$$S_{mel}(f,t) = \log(1 + M(f,t)$$

This time-frequency representation captures perceptually relevant spectral features while compressing dynamic ranges, serving as the primary input to the neural network.

Text normalization applies Unicode NFC for canonical composition of conjuncts and diacritics. Regular expressions

remove non-Kannada characters (code points outside U+0C80–U+0CFF), eliminate punctuation, collapse whitespace, and convert or remove digits based on context. This yields clean transcripts precisely aligned with audio content.

Tokenizer construction for Kannada capitalizes on the language's abugida script by employing a character-level CTC tokenizer. Each unique Unicode grapheme, comprising 49 primary letters, vowel signs, diacritics, and precomposed conjuncts such as ಕ್ಷ, is mapped to a distinct token identifier. A special blank token is added to support CTC alignment. This design obviates the need for subword or byte-pair encoding, which can inadvertently merge visually similar but semantically distinct graphemes, thereby preserving dialectal phonetic distinctions crucial for downstream recognition.

Model adaptation uses the NVIDIA NeMo FC-CTC checkpoint, originally pretrained on large-scale multilingual data. The pretrained model's vocabulary and tokenizer are fully replaced with the Kannada character-level tokenizer, ensuring that the first embedding layer correctly represents all Kannada graphemes. Input layers accept the 80-dimensional log-Mel features, which are subsampled by two sequential Conv1D modules. For a single Conv1D layer at a time step t and output channel c, the output is given by:

$$S_{conv}(t,c) = \sum_{i=0}^{k-1} \sum_{d=0}^{din-1} W(i,j,c).S_{input}(st+i,j) + b(c) \quad (1)$$

where $k = 31$ is the kernel size, $s = 2$ is the stride, $W(i,j,c)$ are the convolutional weights, $b(c)$ is the bias, st is the starting input frame corresponding to the output frame t, and d_{in} is the input dimension. Each convolution produces 512-dimensional feature maps while reducing the temporal dimension by 4×. Therefore, after two layers:

$$T' = \lfloor \frac{T}{4} \rfloor, d_{out} = 512 \quad (2)$$

The subsampled features subsequently receive relative positional encodings computed over this 512-dimensional space, enhancing the self-attention mechanism's capacity to model temporal order without introducing fixed positional biases. This relative positional encoding mechanism can be calculated using:

$$\mathrm{P}(t,2i) = \sin\left(\frac{t}{10000^{2i/d_{\text{model}}}}\right)$$

$$\mathrm{P}(t,2i+1) = \cos\left(\frac{t}{10000^{2i/d_{\text{model}}}}\right) \quad (3)$$

where $d_{\text{model}} = 512$, t is the time step, and $i \in \{0,1,\ldots,255\}$.

The core encoder consists of 17 FC blocks, each integrating Macaron-style feed-forward networks, multi-head self-attention, and convolution modules. The Macaron feed-forward layers are described by:

$$FFN(x) = W_2 \cdot Dropout(Swish(W_1 x + b_1)) + b_2 \quad (4)$$

where W_1 and b_1 project input from 512 to 2048 dimensions, $Swis$ is the activation, $Dropout$ with 0.1 provides regularization, and W_2 and b_2 project output back to 512 dimensions. Multi-head self-attention uses eight 64-dimensional heads for scaled dot-product attention. The convolution module applies depth-wise separable convolution (kernel size 31, expansion to 1024 dimensions) with gated linear units for local context modeling. Pre-normalization, residual connections, and a 0.1 dropout rate ensure stable training.

After passing through the final encoder block, the 512-dimensional sequence features are projected to a 101-dimensional space corresponding to the Kannada grapheme vocabulary. A softmax activation converts logits into probability distributions over tokens at each time step. The CTC loss penalizes the log-likelihood over all valid alignments between the encoder outputs and the target sequence, defined by:

$$\mathcal{L}_{\text{CTC}} = -\log \sum_{\pi \in B^{-1}(Z)} \prod_{t=1}^{T'} P(t,\pi_t) \quad (5)$$

where $B^{-1}(Z)$ is the set of all frame-level sequences π that collapses to the target transcript Z under CTC's blank-merging operator B, T' is the subsampled length, and $P(t,\pi_t)$ is the predicted probability.

Fine-tuning is done on a single NVIDIA A100 GPU with 64 GB memory, using mixed-precision (FP16) training for computational efficiency. To validate the proposed quantified pretraining dominance principle, a controlled baseline framework is adopted: Baseline A (TDNN-HMM without pretraining) and Baseline B (Standard Conformer with pretraining only, no architecture optimization) are trained on an identical corpus alongside the proposed final FastConformer model. This design isolates pretraining contribution from architectural contribution, enabling fair measurement of each component's impact on low-resource ASR performance. All layers of the pretrained FC-CTC model were unfrozen, enabling full end-to-end fine-tuning on the Kannada corpus.

The AdamW optimizer with an initial learning rate of 10^{-4} and weight decay of 10^{-3} drives parameter updates. The Noam learning rate scheduler with 4,000 warmup steps was selected based on the corpus size, expected convergence behavior, and model dimension parameter $d_{\text{model}} = 512$, gradually increasing the learning rate before decaying it proportionally to the inverse square root of the training step number. SpecAugment regularization employed moderate masking parameters—3-time masks and five frequency masks per training example—based on established practices for similar-scale ASR datasets. These values balance regularization benefits with the preservation of essential acoustic information for dialectal distinction. A gradient norm clipping threshold of one ensures numerical stability.

Training proceeds for 60 epochs with evaluation on the validation set after each epoch. The checkpoint yielding the lowest validation word-error rate is preserved for subsequent evaluation. Batch sizes of 16-32 are utilized based on GPU memory availability, with gradient accumulation employed to simulate larger effective batch sizes when necessary. This comprehensive methodology, spanning feature extraction, glyph-preserving tokenization, architectural customization, and

disciplined training protocols, underpins a reproducible framework for multi-dialect Kannada ASR research.

IV. RESULTS AND EVALUATION

All evaluations were performed using the validation and test sets described in the methodology and implemented on the hardware configuration described earlier. The primary focus includes model convergence visualization, quantitative metric reporting, dialect-wise performance comparison, architectural ablation studies, training strategy analysis, and qualitative transcription assessment.

The model was trained for 60 epochs using the Noam scheduler and a conservative learning rate. Consistent convergence was observed as training loss decreased from 2.006 to 0.237 and validation loss from 2.309 to 0.334, indicating effective generalization. The validation-training loss divergence remained controlled throughout training, narrowing from 0.273 at epoch 10 to 0.097 at epoch 60, confirming minimal overfitting despite full model unfreezing.

WER declined sharply during the first 30 epochs from 27.18% to 13.57%, followed by gradual improvement, converging to 11.27% by epoch 60. Character Error Rate (CER) showed similar trends, declining from 12.53% to 5.09%, which supports the effectiveness of character-level tokenization. Key milestones show WER/CER of 20.52%/9.25% at epoch 10, 16.07%/7.26% at epoch 20, and 12.39%/5.54% at epoch 40, with convergence stabilizing after epoch 50. The system's overall accuracy was measured using established metrics—WER, CER, and RTF—on both the validation and held-out test sets. To validate the quantified pretraining dominance principle, three controlled baselines trained on the identical corpus are reported in Table II. The RTF of 0.68 signifies that the model is suitable for deployment in real-time applications.

TABLE II. QUANTITATIVE EVALUATION OF KSR MODEL

Model configuration	Validation WER (%)	Test WER (%)	Validation CER (%)	Test CER (%)
Baseline A: TDNN-HMM (no pretraining)	13.24	13.58	6.89	7.12
Baseline B: Standard Conformer (pretraining only)	12.17	12.42	5.95	6.18
Final Model: Fast Conformer (pretraining + architecture)	10.78	11.23	4.92	5.31
RTF (Final Model)	0.68	—	—	—

The proposed controlled baseline framework enables quantification of each component's impact. Decomposing the improvement from Baseline A (13.58% test WER) to Final Model (11.23% test WER), the pretraining contribution improved from 13.58% to 12.42%, corresponding to a 1.16% absolute improvement (8.5% relative improvement). Similarly, the architecture contribution improved from 12.42% to 11.23%, corresponding to a 1.19% absolute improvement (9.6% relative improvement). It is observed that, for low-resource KSR, pretraining and architecture contribute nearly equally (8.5% vs 9.6%), with pretraining being dominant. This demonstrates that for <100-h languages, multilingual pretraining is the critical lever.

To assess the model's robustness across the rich dialectal landscape of Kannada, the study analyzed WER and CER for each major dialect subset in the corpus. Table III shows that the Granthika Kannada achieved the highest accuracy with a WER of 9.85%, while Halegannada yielded higher error rates, possibly reflecting limited data size and the historical script's distinct phonotactics. Detailed per-dialect error characterization reveals that performance gaps are linguistically rooted rather than random: Dharwad shows 34% of errors as phonetic confusions requiring acoustic/phonological solutions; Halegannada shows 65% of errors as out-of-vocabulary items from archaic vocabulary requiring lexical expansion from classical text. This linguistic root-cause analysis enables targeted, efficient solutions aligned to each dialect's specific challenges. All dialects exhibited competitive performance, underscoring the generalization capacity imparted by the architecture and training protocol. Although most dialects achieved sub-12% WER, future work will address performance gaps through targeted solutions addressing linguistically-rooted error patterns rather than generic data augmentation.

TABLE III. DIALECT-WISE PERFORMANCE

Dialect	WER (%)	CER (%)
Granthika	9.85	4.41
Dharwad	11.24	5.12
Kundagannada	10.72	4.89
Arebhase	12.18	5.74
Havyaka	13.05	6.02
Halegannada	14.83	6.91

Phoneme-level analysis indicates that recognition errors are not uniformly distributed across the Kannada inventory, but are concentrated in a few linguistically sensitive regions. The most frequent patterns involve vowel sign omissions or confusions within short–long pairs (for example, interchange or deletion of dependent vowel diacritics) and weakening of the retroflex(ಮೂರ್ಧನ್ಯ) versus dental(ದಂತ್ಯ) contrast, where ಟ, ಡ, ಣ are sometimes realized as their dental counterparts ತ, ದ, ನ in dialectal speech. In addition, nasalized(ಅನುನಾಸಿಕ) syllables that should be written with explicit nasal consonants (e.g., ಮಂ followed by ನ / ಮ) are sometimes realized in the output with only an ಅನುಸ್ವಾರ (ಂ), leading to orthographic under specification of place of articulation. Consonant clusters and conjuncts such as ಕ್ಷ are occasionally simplified to single consonants, but they account for a smaller portion of overall errors compared to vowel signs and coronal place of articulation confusions. These trends suggest that the FastConformer encoder has largely learned the core phonemic structure of Kannada, with residual errors aligned to well-known phonological weak points rather than arbitrary acoustic mismatches. Table IV displays the qualitative phoneme-class confusion matrix of the same.

TABLE IV. QUALITATIVE PHENOME-CLASS CONFUSION MATRIX

Reference class	Correct the same class	Vowel sign/length error	Retroflex ↔ dental confusion	Nasal ↔ ಅನುಸ್ವಾರ (ಂ)	Cluster simplification / other
Short vowels (ಹ್ರಸ್ವ)	High	Medium	Low	Low	Low
Long vowels	High	High	Low	Low	Low
Retroflex stops/nasals	High	Low	Medium	Low	Low
Dental/alv stops	High	Low	Medium	Low	Low
Nasals (ಅನುನಾಸಿಕ)	High	Low	Low	Low	Low
Consonant clusters	Medium	Low	Low	Low	Medium
Function words/particles	Medium	Low	Low	Low	Medium

A detailed ablation analysis was conducted to isolate the quantitative impact of each architectural innovation. Table V shows that the baseline Transformer without subsampling or relative positional encoding achieved a WER of 14.72%. Integrating convolutional subsampling boosted temporal precision, while relative positional encoding improved context retention. The full FastConformer stack achieved the lowest WER (10.78%), demonstrating cumulative benefits. These results support architectural choices, with convolutional subsampling and relative encoding, significant for handling Kannada's long context windows and morphophonemic complexity.

Representative transcription samples highlight both strengths and residual error modes of the system. Table VI compares ground-truth transcriptions with predicted output, demonstrating high accuracy even for dialectally marked utterances and complex consonant clusters. Most errors are isolated to vowel marker omissions or single grapheme confusions, such as ಅವ್ನು vs ಅವನು. Grammatically, some of the Kannada syllables are associated with multiple sounds based on the usage in a sentence. Therefore, a more rigorous error analysis will be required for continuous conversational speech and code-switching contexts.

TABLE V. ABLATION STUDY QUANTIFYING MODULE-LEVEL IMPROVEMENTS

Configuration	Parameters (M)	WER (%)	CER (%)
Baseline transformer	45.1	14.72	7.02
+ Convolutional frontend	46.3	12.88	6.05
+ Relative positional encoding	47.2	11.42	5.33
+ FastConv blocks (final)	48.0	10.78	4.92

TABLE VI. QUALITATIVE COMPARISON OF SAMPLE TRANSCRIPTIONS

Audio file	Ground truth (Kannada)	Prediction	Error type
knm_0056.wav	ನಾನು ಇಂದು ಮಳೆಯಲ್ಲಿ ಹೋಗಿದ್ದೆ	ನಾನು ಇಂದು ಮಳೆಯಲ್ಲಿ ಹೋಗಿದ್ದೆ	None
knm_1037.wav	ನಿನ್ನ ಮನೆ ಎಲ್ಲಿದೆ?	ನಿನ್ನ ಮನ ಎಲ್ಲಿದೆ?	Missing vowel marker
knf_2902.wav	ಅವ್ನು ಎಲ್ಲಿ ಹೋದನು	ಅವನು ಎಲ್ಲಿ ಹೋದನು	Phonetic confusion

The FC-based system sets a new benchmark for low-resource, morphologically complex languages such as Kannada. Its performance outstrips baseline transformer architectures by combining convolutional temporal compression and relative positional encoding with end-to-end CTC decoding. The system demonstrates strong dialectal robustness, efficient real-time operation, and granular handling of unique Kannada graphemes, as reflected by WER and CER values consistently below prior benchmarks. While robust to script, phonetic, and speaker variation, additional enhancements—such as self-supervised pretraining, explicit dialect tagging, and adoption of larger, more diverse speech corpora—remain promising avenues. The methodology and results together establish a replicable foundation for/enable future scalable, dialect-aware ASR deployment in Kannada and related Indic languages.

V. CONCLUSION

This study proposed a comprehensive solution for multi-dialect Kannada Speech Recognition (KSR) using a FastConformer-based architecture fine-tuned on a newly curated, dialect-balanced Kannada speech corpus. The methodology established end-to-end fine-tuning with robust preprocessing, tailored character-level tokenization, and a fully unfrozen encoder optimized for the specific linguistic and phonotactic features of Kannada.

Through extensive experimentation and analysis, the proposed system delivers significant gains over baseline transformer architectures, achieving a Word Error Rate (WER) of 11.23% and Character Error Rate (CER) of 5.31% on held-out test data, while maintaining real-time processing efficiency with Real-Time Factor (RTF) < 0.7. Dialect-wise analysis confirms the model's practical utility and adaptability, with consistent performance across all major regional variants, despite data size disparities and phonological complexity.

Despite these strong results, the work has some limitations. The corpus remains relatively small (7.5 h) compared to large-scale Automatic Speech Recognition (ASR) benchmarks, and certain dialects such as Halegannada are still underrepresented, which contributes to their higher WER. In addition, experiments were conducted on a single high-end GPU, and the study did not quantify performance for lighter FC variants on edge devices or for fully spontaneous and code-switched Kannada–English speech, which constrains claims about scalability and broad real-world deployment.

Overall, this research advances the state of the art for KSR, offering a reproducible and effective framework deployable in low-resource, real-world settings. This work makes three key contributions: (1) a linguistically informed dialect-aware

curation framework demonstrating that data quality can partially overcome data scarcity, (2) empirical quantification of pretraining versus architecture impact through controlled baselines, and (3) dialect-specific error diagnosis revealing linguistic roots of performance gaps. Beyond technical performance, the system contributes to digital inclusivity by enabling accurate voice-based technologies for Kannada speakers across diverse regions and dialects, fostering equitable access to language technologies within the broader community. Future work will expand on self-supervised pretraining, explicit dialect modeling, evaluation on conversational and noisy speech, and cross-dataset validation to further generalize and scale the system for broader Indic language ASR applications.

DECLARATION OF COMPETING INTERESTS

The authors declare that they have no competing financial interests or personal relationships that influenced this work.

ACKNOWLEDGMENT

Not applicable to this work.

DATA AVAILABILITY

The dataset generated and analyzed during the present study is currently undergoing a copyright registration process. To protect the intellectual property rights until this process is finalized, the data are not publicly available in an open repository. However, the data are available from the corresponding author on reasonable request for the purposes of research and collaboration.

AI USE AND DECLARATION OF GENERATIVE AI USE

During the preparation of this work, the authors used Generative AI to improve the grammar of the manuscript. This technology was used solely for language enhancement and not for the creation of research content, data analysis, or the formulation of scientific conclusions. The authors have reviewed and edited the output and take full responsibility for the final content of the published article.

REFERENCES

[1] G. Hinton *et al.*, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov. 2012, https://doi.org/10.1109/MSP.2012.2205597.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, https://doi.org/10.1038/nature14539.

[3] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, May 2013, pp. 6645–6649, https://doi.org/10.1109/ICASSP.2013.6638947.

[4] A. Graves and J. Schmidhuber, "Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures," *Neural Networks*, vol. 18, no. 5–6, pp. 602–610, Jul. 2005, https://doi.org/10.1016/j.neunet.2005.06.042.

[5] A. Vaswani *et al.*, "Attention Is All You Need." arXiv, 2017, https://doi.org/10.48550/ARXIV.1706.03762.

[6] A. Gulati *et al.*, "Conformer: Convolution-Augmented Transformer for Speech Recognition," in *Interspeech 2020*, Oct. 2020, pp. 5036–5040, https://doi.org/10.21437/Interspeech.2020-3015.

[7] M. Burchi and V. Vielzeuf, "Efficient Conformer: Progressive Downsampling and Grouped Attention for Automatic Speech Recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop*, Cartagena, Colombia, Dec. 2021, pp. 8–15, https://doi.org/10.1109/ASRU51503.2021.9687874.

[8] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." arXiv, 2020, https://doi.org/10.48550/ARXIV.2006.11477.

[9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision." arXiv, 2022, https://doi.org/10.48550/ARXIV.2212.04356.

[10] A. Babu *et al.*, "XLS-R: Self-Supervised Cross-Lingual Speech Representation Learning at Scale," in *Interspeech 2022*, Sep. 2022, pp. 2278–2282, https://doi.org/10.21437/Interspeech.2022-143.

[11] T. Javed, K. Bhogale, A. Raman, P. Kumar, A. Kunchukuttan, and M. M. Khapra, "IndicSUPERB: A Speech Processing Universal Performance Benchmark for Indian Languages," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, pp. 12942–12950, Jun. 2023, https://doi.org/10.1609/aaai.v37i11.26521.

[12] V. Pratap *et al.*, "Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters," in *Interspeech 2020*, Oct. 2020, pp. 4751–4755, https://doi.org/10.21437/Interspeech.2020-2831.

[13] R. G. Rajakumari, D. K. Renuka, and L. A. Kumar, "Enhancing ASR Accuracy and Coherence Across Indian Languages with Wav2vec2 and GPT-2," *ICTACT Journal on Data Science and Machine Learning*, vol. 6, no. 2, pp. 761–764, Mar. 2025, https://doi.org/10.21917/ijdsml.2025.0156.

[14] N. Sethiya, S. Nair, P. Walia, and C. Maurya, "Indic-ST: A Large-Scale Multilingual Corpus for Low-Resource Speech-to-Text Translation," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 24, no. 6, pp. 1–25, Jun. 2025, https://doi.org/10.1145/3736720.

[15] B. Choudhury, V. Kumar, and S. Singh, "IndicVoices-R: Multilingual, Multi-Speaker Speech Corpus for Indian TTS." Hugging Face, 2024, [Online]. Available: https://huggingface.co/datasets/ai4bharat/indicvoices_r.

[16] M. C. Shunmuga Priya, D. Karthika Renuka, and L. Ashok Kumar, "Robust Multi-Dialect End-to-End ASR Model Jointly with Beam Search Threshold Pruning and LLM," *SN Computer Science*, vol. 6, no. 4, Mar. 2025, Art. no. 323, https://doi.org/10.1007/s42979-025-03794-9.

[17] D. S. Park *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Interspeech 2019*, Graz, Austria, Sep. 2019, pp. 2613–2617, https://doi.org/10.21437/Interspeech.2019-2680.

[18] P. Punitha and G. Hemakumar, "Speaker Dependent Continuous Kannada Speech Recognition Using HMM," in *International Conference on Intelligent Computing Applications*, Coimbatore, India, Mar. 2014, pp. 402–405, https://doi.org/10.1109/ICICA.2014.88.

[19] S. C. Sajjan and C. Vijaya, "Continuous Speech Recognition of Kannada Language Using Triphone Modeling," in *2016 International Conference on Wireless Communications, Signal Processing and Networking*, Chennai, India, Mar. 2016, pp. 451–455, https://doi.org/10.1109/WiSPNET.2016.7566174.

[20] R. Pradeep and K. S. Rao, "Deep Neural Networks for Kannada Phoneme Recognition," in *Ninth International Conference on Contemporary Computing*, Noida, India, Aug. 2016, pp. 1–6, https://doi.org/10.1109/IC3.2016.7880202.

[21] P. S. Praveen Kumar, G. Thimmaraja Yadava, and H. S. Jayanna, "Continuous Kannada Speech Recognition System Under Degraded Condition," *Circuits, Systems, and Signal Processing*, vol. 39, no. 1, pp. 391–419, Jan. 2020, https://doi.org/10.1007/s00034-019-01189-9.

[22] D. S. Jayalakshmi, K. P. Sathvik, and J. Geetha, "Speech Recognition for Kannada Using LSTM," in *Advances and Applications of Artificial Intelligence & Machine Learning*, vol. 1078, B. Unhelkar, H. M. Pandey, A. P. Agrawal, and A. Choudhary, Eds. Singapore: Springer Nature Singapore, 2023, pp. 189–201.

[23] Y. G. Thimmaraja, B. G. Nagaraja, and H. S. Jayanna, "Improvements in ASR System to Access the Real-Time Agricultural Commodity Prices and Weather Information in Kannada Language/Dialects," *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 4195–4217, Jan. 2024, https://doi.org/10.1007/s11042-023-15350-9.

[24] R. Shashidhar and S. Patilkulkarni, "Audiovisual Speech Recognition for Kannada Language Using Feed Forward Neural Network," *Neural Computing and Applications*, vol. 34, no. 18, pp. 15603–15615, Sep. 2022, https://doi.org/10.1007/s00521-022-07249-7.

[25] G. Thimmaraja Yadava, B. G. Nagaraja, and G. P. Raghudathesh, "Real-Time Automatic Continuous Speech Recognition System for Kannada Language/Dialects," *Wireless Personal Communications*, vol. 134, no. 1, pp. 209–223, Jan. 2024, https://doi.org/10.1007/s11277-024-10903-z.

[26] Mahadevaswamy, "Robust Automatic Speech Recognition System for the Recognition of Continuous Kannada Speech Sentences in the Presence of Noise," *Wireless Personal Communications*, vol. 130, no. 3, pp. 2039–2058, Jun. 2023, https://doi.org/10.1007/s11277-023-10371-x.

[27] R. Shashidhar, M. P. Shashank, G. Jagadamba, and V. Ravi, "A Fusion Approach for Kannada Speech Recognition Using Audio and Visual Cue," in *IoT Sensors, ML, AI and XAI: Empowering A Smarter World*, vol. 50, B. Pradhan and S. Mukhopadhyay, Eds. Cham: Springer Nature Switzerland, 2024, pp. 387–414.

[28] Y. G. Thimmaraja, B. G. Nagaraja, and H. S. Jayanna, "Development of Noise Robust Real Time Automatic Speech Recognition System for Kannada Language/Dialects," *Engineering Applications of Artificial Intelligence*, vol. 135, Sep. 2024, Art. no. 108693, https://doi.org/10.1016/j.engappai.2024.108693.

[29] G. T. Yadava, B. G. Nagaraja, and H. S. Jayanna, "An End-to-End Continuous Kannada ASR System Under Uncontrolled Environment," *Multimedia Tools and Applications*, vol. 83, no. 3, pp. 7981–7994, Jan. 2024, https://doi.org/10.1007/s11042-023-15854-4.

[30] Y. G. Thimmaraja, B. G. Nagaraja, and H. S. Jayanna, "Advancements in End-to-End Isolated Kannada ASR System by Combining Robust Noise Elimination Technique and TDNN," *Intelligent Systems with Applications*, vol. 20, Nov. 2023, Art. no. 200288, https://doi.org/10.1016/j.iswa.2023.200288.

[31] Y. G. Thimmaraja, B. G. Nagaraja, H. S. Jayanna, and B. R. Shivakumar, "A Spoken Query System to Access the Real Time Agricultural Commodity Prices and Weather Information in Kannada Language/Dialects," *Multimedia Tools and Applications*, vol. 83, no. 10, pp. 28675–28688, Sep. 2023, https://doi.org/10.1007/s11042-023-16554-9.

[32] N. B. Chittaragi and S. G. Koolagudi, "Automatic Dialect Identification System for Kannada Language Using Single and Ensemble SVM Algorithms," *Language Resources and Evaluation*, vol. 54, no. 2, pp. 553–585, Jun. 2020, https://doi.org/10.1007/s10579-019-09481-5.

[33] M. Latha, M. Shivakumar, G. Manjula, M. Hemakumar, and M. K. Kumar, "Deep Learning-Based Acoustic Feature Representations for Dysarthric Speech Recognition," *SN Computer Science*, vol. 4, no. 3, Mar. 2023, Art. no. 272, https://doi.org/10.1007/s42979-022-01623-x.

[34] R. G. Rajakumari, D. K. Renuka, L. A. Kumar, C. Thiraviya, S. Vaimitra, and S. V. Easwaramoorthy, "Multilingual Automatic Speech Recognition for Indian Languages-E2E Framework," in *OITS International Conference on Information Technology*, Vijayawada, India, Dec. 2024, pp. 138–142, https://doi.org/10.1109/OCIT65031.2024.00033.

[35] M. Shanthamallappa and B. P. Pradeep Kumar, "Enhanced Perceptual Wavelet Packet Features for Spontaneous Kannada Sentence Recognition Under Uncontrolled Conditions," *International Journal of Speech Technology*, vol. 28, no. 1, pp. 153–174, Mar. 2025, https://doi.org/10.1007/s10772-024-10156-y.

[36] G. Thimmaraja Yadava, B. G. Nagaraja, and H. S. Jayanna, "Amalgamation of Noise Elimination and TDNN Acoustic Modelling Techniques for the Advancements in Continuous Kannada ASR System," *Multimedia Tools and Applications*, vol. 83, no. 7, pp. 19953–19968, Jul. 2023, https://doi.org/10.1007/s11042-023-16100-7.

[37] P. Rajeswari, N. Shankaraiah, and S. Rathnakara, "Enhancement and Reconstruction of Dysphonic Kannada Speech Using a Generative Adversarial Network and a SepFormer Model," *Engineering, Technology & Applied Science Research*, vol. 15, no. 6, pp. 29097–29102, Dec. 2025, https://doi.org/10.48084/etasr.14812.

[38] G. T. Yadava and B. G. Nagaraja, "Noise Robust E2E Continuous Kannada ASR System Under Real Time Conditions," *Circuits, Systems, and Signal Processing*, vol. 44, no. 7, pp. 4965–4987, Jul. 2025, https://doi.org/10.1007/s00034-025-03024-w.