



# Multilingual low resource Indian language speech recognition and spell correction using Indic BERT

M C SHUNMUGA PRIYA<sup>1,\*</sup> , D KARTHIKA RENUKA<sup>1</sup>, L ASHOK KUMAR<sup>2</sup> and S LOVELYN ROSE<sup>3</sup>

<sup>1</sup>Department of IT, PSG College of Technology, Coimbatore 641 004, India

<sup>2</sup>Department of EEE, PSG College of Technology, Coimbatore 641 004, India

<sup>3</sup>Department of CSE, PSG College of Technology, Coimbatore 641 004, India

e-mail: shunmugapriya.mc@gmail.com; dkr.it@psgtech.ac.in; lak.eee@psgtech.ac.in; slr.cse@psgtech.ac.in

MS received 18 May 2021; revised 23 May 2022; accepted 16 August 2022

**Abstract.** India is a land of unity; it is home to 122 major languages and 1599 other languages. Around 70% of people in India speak Indo-Aryan languages whereas 19% speak Dravidian languages which are agglutinative morphologically rich. Speech is a lucid, time-saving, and effortless means of communication. Automatic speech recognition (ASR) is a process that accurately transcribes spoken utterances into text. Speech recognition in Indian languages will empower people to easily access their regional language to any content they desire. The ultimate goal of this proposed work is to develop a novel deep sequence modeling-based ASR system with improved spell corrector for seven low-resource languages. The efficacy of our proposed model is evaluated using word error rate (WER) and sequence match ratio. The end-to-end ASR system based on a recurrent neural network-gated recurrent unit (RNN-GRU) achieves plausible results with average WER of 0.62. Indeed, one of the key concerns in the ASR system is spelling errors in transcribed text. Despite the intricacy involved in spell correction of Natural Language Processing, the transformer-based INDIC Bidirectional Encoder Representations from Transformers language model yields a significant improvement in performance by 10% and reduces the average WER to 0.52.

**Keywords.** Automatic speech recognition; deep learning; gated recurrent unit; indic BERT; indic spell corrector.

## 1. Introduction

A multilingual speech recognition system transcribes input audio waveforms in any language to their corresponding readable text. Indo-Aryan and Dravidian are the major two groups of a language spoken in India. Some of the Indo-Aryan languages are Sanskrit, Hindi, Bengali, Gujarati, and Punjabi which are widely spoken in the North West region. Dravidian or agglutinative languages are rich in the morphological structure such as Tamil, Telugu, Malayalam, and Kannada which are uttered in south India. As per the 2011 census, Hindi is spoken by 43.63% of its population whereas, the other scheduled languages such as Bengali, Marathi, Telugu, Tamil, Gujarati, Kannada, and Malayalam are spoken by 18%, 6.8%, 6.7%, 5.7%, 4.58%, 3.61%, 2.88% of its population respectively. To preserve the low resources of Indian languages from dwindling, there exist a definite need for

techniques like multilingual speech recognition, which inspire people to use their native Indian languages for human-computer interaction.

Owing to the high necessity of multilingual speech recognition for low-resource Indian languages, an illustration of conventional ASR workflow is required. Phonemes are the smallest units of sound whereas graphemes are the smallest units of text. Automatic Speech Recognition (ASR) aims to convert raw audio into a sequence of graphemes [1]. An ASR system identifies the most probable text given an input speech signal [2]. Conventional speech recognition models [3, 4] using the Hidden Markov model-Gaussian Markov model (HMM-GMM) are folded by three different models namely acoustic, pronunciation and language models. Acoustic models convert raw audio features to phonemes. Pronunciation model or lexicon model map phonemes to understandable words. The language model outputs the most probable or likelihood word sequence.

Given the acoustic feature vectors of a speech signal  $\hat{x} = \{x_1, x_2, \dots, x_t\}$ , the goal of ASR system is to find a word

\*For correspondence

Published online: 05 November 2022

sequence  $\hat{w} = \{w_1, w_2, \dots, w_n\}$ .  $\hat{w}$  is derived based on Bayes rule as in equation (1).

$$\hat{w} = \operatorname{argmax} p(x) = \operatorname{argmax} \frac{p(x|w)p(w)}{p(x)} \quad (1)$$

where,  $p(x|w)$  represents the acoustic model,  $p(w)$  is the prior probability of the word sequence and it represents the language model,  $p(x)$  is a constant over  $w$ .

Conventional ASR models have several demerits such as being speaker-dependent, recognizing isolated words, and based on small vocabulary, which makes them unviable for a wide range of users. The conventional ASR models are superseded by the recent end-to-end deep learning-based speech recognition models, which greatly simplify the complexity of traditional speech recognition [5]. The proposed work ideally put forward a novel approach using a deep sequential neural network-based end-to-end multilingual low resource Indian language speech recognition with an improved spell corrector module.

The development of a multilingual low resource Indian language speech recognition model involves several challenges, such as the availability of quality speech data. Yet, another crucial challenge is to collect the lexicon of scripts for the Indian languages. Complex patterns such as different dialects and complex scripts make a speech-to-text model for Indian languages a more challenging task. To surmount the challenges of speech-to-text models and to perform more accurate transcription, RNN-GRU and Indic BERT models are fused in our proposed work.

The contribution of our work is as follows

- Monolingual speech data annotated with text labels spoken by native speakers is gathered and collected in scripts for seven Indian languages (Tamil, Telugu, Malayalam, Kannada, Bengali, Gujarati, and Marathi).
- The Mel frequency cepstral coefficients (MFCC) features are extracted and developed an end-to-end ASR model (phonemes to graphemes).
- Furthermore, to accelerate the performance and rectify the spelling errors, Indic BERT a deeply bidirectional language model for spelling correction is exploited.

The content of this paper is organized as follows: a summary of recent research in the field of speech recognition is presented in section 2, followed by the dataset statistic description in section 3. The system architecture is explained in section 4 and detailed discussions in section 5.

## 2. Related work

Sir George Abraham Grierson [6] conducted a survey of language diversity in India from 1898 to 1928 and he reported that around 544 dialects are present. A wide variety of dialects in Indian languages, transform the ASR

system into a more complex one. Complex multi-dialect issues in ASR are analyzed by B. Li *et al* [7] using a single sequence-to-sequence model. However, extremely complex patterns of speech such as speaker age, and background noise degrade the ASR system performance. Most recently, H. Miao *et al* [8] proposed a hybrid CTC-Attention-based model which improves the real-time human-computer interaction performance. Recently, Multilingual ASR (MLASR) systems is explored by Oliver Adams *et al* [9] using a common attention-based language independent encoder structure. Moreover, a Multilingual ASR model for Indonesian and Tibetan languages using deep learning models was developed by [10, 11]. In another study, multilingual ASR for four Ethiopian languages using a deep neural network based multitask learning approach was developed [12].

In 2019, Google Alexa recognized the widely spoken Indian language Hindi for the first time. Moreover, Google research reported that in a country like India, 72% of users prefer voice search in regional languages. Singh *et al* [13] performed a detailed survey on speech recognition systems in Indian languages and explored the recent trends. Voice search in regional language allows users to save time, makes them more convenient, and is widely utilized for many applications such as subtitle creation, speech interfaces, tele medicine, and dictation tools. Despite, their extended benefits there are also notable challenges such as how accurately it recognizes the Indian languages. But, better tuned neural network algorithms can exponentially increase the performance of such a speech recognition system. Several studies proved that recurrent neural network is specialized for sequential data. In particular, B.Li *et al* [14], Z.Wu *et al* [15] illustrated the improved performance of two gating structures of RNN such as GRU and LSTM for speech recognition applications. Cho, K., *et al* [16] proposed GRU by dropping gates from LSTM and GRU models overcome vanishing and exploding gradient problems [17]. In addition, Listen attend and spell (LAS) modeled by W.chan *et al* [18] outperforms the RNN-Transducer technique for streaming speech transcribing applications. Although much research has been carried out, Deep Speech2 is one of the simple yet powerful GRU-CTC based recognition models [19].

Shaohua Zhang *et al* [20] studied a soft masked BERT model for Chinese spell error correction. Influenced by the success of BERT, Hao *et al* [21] analyzed how to fine-tune the BERT and also inspected the dynamics of BERT. Similarly, to perform spell correction in Indian languages, Indic BERT is engaged which is a pre-trained language model with two variants such as base (12 layers) and large (24 layers). Indeed, Indic BERT is modeled on Indic Corp monolingual corpora of Indian languages which are evaluated using Indic GLUE with various downstream natural language tasks Kakwani *et al* [22]. Jain *et al* [23] published the performance of the BERT language model on various NLP tasks for 3 Indian languages. The five methods for

spell error correction proposed by Neto *et al* [24] are as follows, rule-based technique, similarity key-based technique, edit distance, n-gram based method, and neural network based methods. Didenko *et al* [25] studied that BERT can also be applied to grammatical error correction scenarios.

The research gaps identified from this review are as follows:

- The lack of sufficient data in Indian languages is an overarching issue in the field of Indian computational linguistics.
- Limited studies are evidenced on Indian languages except for Hindi in the field of computational linguistics.
- Lack of effective speech recognition and spell corrector in Indian languages. Understanding the research gaps, lead us to develop a novel speech recognition model with improved spell correctors in low-resource Indian languages.

### 3. Dataset statistic

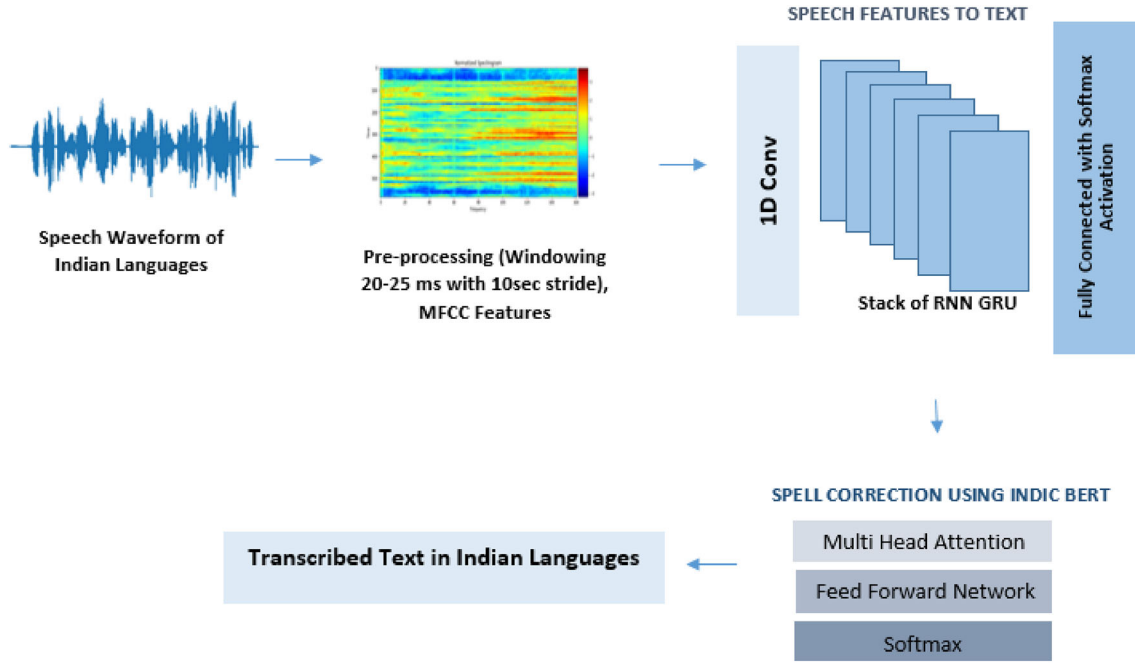
The gathered monolingual corpora of seven low-resource Indian languages are spoken by native speakers with varied dialects and have finally undergone a quality check [26]. An average of 3500 sentences were spoken in Indo-Aryan and Dravidian languages except for Marathi. Speech corpora were recorded based on sentences from Wikipedia, real-world sentences (eg: weather forecast), sentences created from templates, and original handwritten sentences. The speech corpora were noise free including ambient noise, and breathing sounds, as it is recorded in an acoustic environment at 48kHz(16 bits per sample). Indian Script Code for Information Interchange (ISCII) is utilized as a benchmark reference for collecting the scripts. The statistical distribution of Indian Speech data and script for the seven Indian languages are gathered and collective inference is given with a higher level of granularity in table 1. The dataset can be downloaded from the link (<https://www.openslr.org/resources.php>).

### 4. System architecture

Speech recognition has created a technological impact on society and is expected to flourish further in the area of human-machine interaction. In recent years, neural network architecture has ushered advancements in computational linguistics [27]. The proposed neural network based architecture is categorized into two subsections. Subsection 4.1 discusses the architecture of the automatic speech recognition module. In subsection 4.2, the post-processing spell correction module is described in detail.

**Table 1.** Indian Language speech data distribution.

Languages	Number of sentences (F-Female, M-Male)	Number of speakers	Total		Scripts
			duration in hours	Number of distinct utterances	
Tamil	4291 (2335 F, 1956 M)	50 (25F, 25M)	4.01	12779 (6620F, 6159M)	247 (12 vowels, 18 consonants, 241 combinant letter, 1 special letters)
Telugu	4448 (2294 F, 2154 M)	47 (24F, 23M)	2.73	8554 (4218F, 4336M)	60 (16 vowels, 3 modifiers, 41 consonants)
Malayalam	4126 (2103 F, 2023 M)	42 (24F, 18M)	3.02	12120 (5713F, 6407M)	57 (15 vowel letters, 42 consonants)
Kannada	4400 (2186 F, 2214 M)	59 (23F, 36M)	4.31	16003 (8622F, 7381M)	49 (13 vowels, 25 structured consonants, 9 unstructured consonant and 2 partly vowels and 2 partly consonants)
Bengali	3200M	15M	3.30	13100	41 (28 consonants and 13 vowels)
Marathi	1569F	9F	3.02	3072	52 (14 vowels, 36 consonants and 2 sound modifiers)
Gujarati	4272 (2219F, 2053M)	36 (18M, 18F)	4.30	16021 (8203F, 7818M)	47 (13 Vowels and 34 Consonants)



**Figure 1.** System architecture.

#### 4.1 Automatic speech recognition (Phonemes to Graphemes)

Speech is time-varying, quasi-stationary natured signals with 16000 samples/second. To inspect every instance of speech over time and to achieve good spectral representation, initially windowing and framing are performed. The input speech waveforms are framed by applying a window of size 20 to 25 milliseconds with 10 milliseconds stride. Following, discrete cosine transform (DCT) is applied over the sequence of input speech frames, for attaining MFCC. Given the input sequence of cepstral features with  $t$  number of frames  $x = \{x_1, \dots, x_t\}$  is passed into a deep sequential block as depicted in figure 1. The deep sequential recurrent block consists of 1D convolution layer followed by a stack of seven GRU layers with 250 cells each. The one dimensional convolution layer with a stride size of 2, shift along the dimension of time, frequency is convolved at each step. GRU is folded by two gated units such as update and reset gate and it handles long term dependencies more efficiently. The hidden units  $h = \{h_1, \dots, h_t\}$  of GRU is updated at each time step based on the update rule of GRU as given in equation (2–4).

$$h_t = z_t \otimes h_{t-1} + (1 - z_t) \otimes \text{ReLU}(\text{BN}((W_x x_t + r_t \otimes W_h h_{t-1} + b_h))) \quad (2)$$

$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_z) \quad (3)$$

$$r_t = \sigma(W_r[h_{t-1}, x_t] + b_r) \quad (4)$$

where,  $W_z, W_r, W_h$  are the weight parameters of update  $z$ , reset  $r$  and hidden  $h$  states respectively,  $b_{[z,r,h]}$  is the bias,  $\sigma$  is the sigmoid activation function. Batch normalization with ReLU activation is applied for normalizing the mean and variance of GRU layers as shown in equation (5).

$$\text{BN}(G) = \frac{\gamma(G - \text{mean})}{\sqrt{(\text{variance}^2 + \epsilon) + \beta}} \quad (5)$$

where,  $\gamma$  and  $\beta$  are scaling and shifting parameters,  $\epsilon$  is small constant.

To optimize the models, an adaptive moment estimation (ADAM) optimizer is used. To generalize the model that is to predict new unseen data more accurately, optimized hyperparameter values are incorporated such as L2 regularization of 1e-2, weight decay 1e-6, momentum values as 0.9, 20% of dropout, learning rate value as 1e-3. The model was trained with 50 epochs and a batch size of 20, early stopping hyperparameter is incorporated as a measure of generalization. The learned features are then fed into a fully connected layer followed by a multi-class softmax classifier.

#### 4.2 Spell correction using Indic BERT

The output transcripts produced by the acoustic model may have spelling errors. To explicitly correct the spelling errors and to post-process the output text transcribed from the RNN-GRU model, Indic BERT a pre-trained language model on 12 Indian languages is applied. BERT is an

Input Sentence  $X$  “இசை நிகழ்ச்சியை யாவினத தொழில்நுட்ப வசதியுடன் பிரம்மாண்டமான முறையில் நடத்தை திட்டமிற்றிடிருக்கிறோம்”

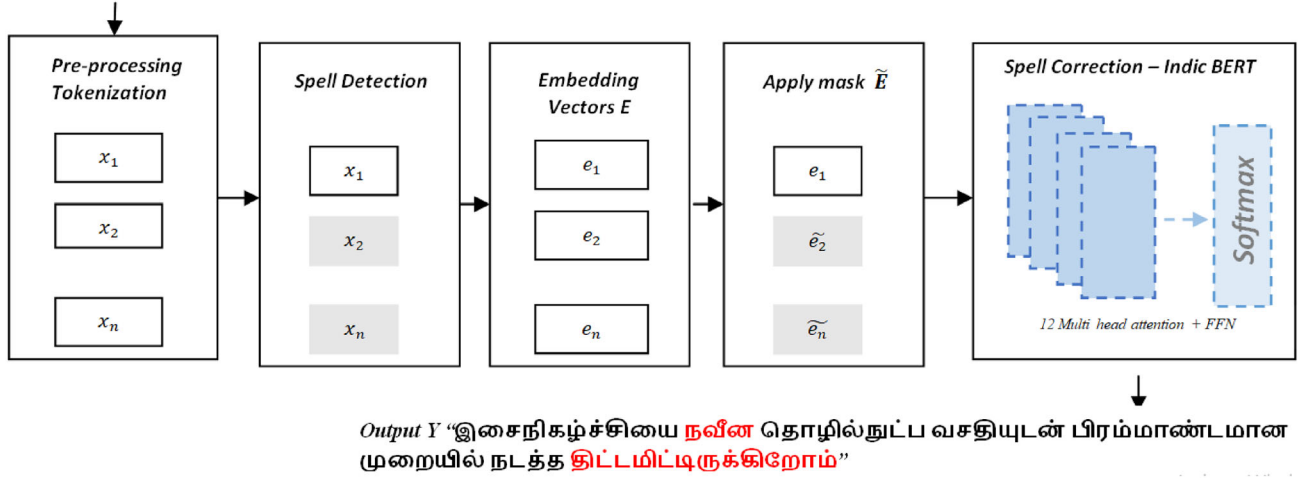


Figure 2. Spell correction.

unobtrusive technique of Google, a transformer-based pre-trained model for many of the natural language processing tasks such as text classification, named entity recognition, paraphrase detection and spell correction to extract context-sensitive features from the input text.

Figure 2 describes the state-of-the-art for spell correction framework. The natural language toolkit iNLTK is intended

for pre-processing the Indic text. Initially, the input sentence  $X$  is split into sequence of tokens  $X = \{x_1, x_2, \dots, x_n\}$ . The spell error detection is performed to identify the misspelled tokens. Spell error detection is a binary classifier problem based on Levenshtein distance, which gives 0 for words without spell error whereas 1 for misspelled words as given in equation (6).

இசை நிகழ்ச்சியை நவீன தொழில்நுட்ப வசதியுடன் நடத்த திட்டமிட்டிருக்கிறோம்

(We have planned to conduct the music concert with modern technology)

Ranking

$(c_1, p_1) \rightarrow$  நவீன (modern), 0.59

$(c_2, p_2) \rightarrow$  புதிய (latest), 0.21

$(c_3, p_3) \rightarrow$  சிறந்த (best), 0.20

Candidate List

இசை நிகழ்ச்சியை யாவினத தொழில்நுட்ப வசதியுடன் நடத்த திட்டமிட்டிருக்கிறோம்

(We have planned to conduct the music concert with moden technology)

Figure 3. Sample Candidate Word Selection (Red Coloured - Tokens with spelling errors, Blue Coloured - Candidate word list, Green Coloured - Spell corrected words).



$$\text{Spell Error Detection} = \begin{cases} 0 & \text{if } Lev_{(a,p)} = 0 \\ 1 & \text{else if } Lev_{(a,p)} > 0 \end{cases} \quad (6)$$

$$Lev_{(a,p)}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \left( \begin{aligned} &lev_{(a,p)}(i-1,j) + 1, \\ &lev_{(a,p)}(i,j-1) + 1, \\ &lev_{(a,p)}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{aligned} \right) & \text{otherwise} \end{cases} \quad (7)$$

The levenshtein distance  $Lev_{(a,p)}$  measures the difference between the actual  $a$  and predicted  $p$  tokens by calculating the total number of operations such as insertion, exclusion and substitution required to match it as given in the equation (7).

The word embedding sequence  $E = \{e_1, e_2, \dots, e_n\}$  is generated using Indic fast text (IndicFT) toolkit. The mask is applied for the identified misspelled tokens as  $\tilde{E} = \{e_1, \tilde{e}_2, \dots, e_n\}$ . Spell error correction is solved using Indic BERT base model, which has 12 layers of multi-head self-attention followed by a feed forward network as shown in equation (8–9).

$$\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V \quad (8)$$

$$\text{Multihead}(Q, K, V) = \text{Head}_1 \text{Attn}(QW_1^Q, KW_1^K, VW_1^V) || \dots || \text{Head}_n \text{Attn}(QW_n^Q, KW_n^K, VW_n^V) \quad (9)$$

where  $Q, K, V$  are the Query, Key Value and are their corresponding weight matrices.

It contextually analyses the masked sequence  $\tilde{E}$  bidirectionally and produces a candidate word list. The candidate word list consists a list of words that are similar to the masked tokens as shown in figure 3. To choose a single candidate word from the list of words created, it is considered a multi class classification problem. Finally, Indic BERT has a softmax layer for performing a multi class classification, to identify the most likely word that occurs in the place of masked tokens. Thereby improving the performance of the RNN-GRU based acoustic modeling, Indic BERT yields the candidate word with higher probability and thus producing the spell corrected output sequence  $Y = \{y_1, y_2, \dots, y_n\}$ .

## 5. Results and analysis

The proposed model has yielded good performance and the performance improvements with the state of art methods are compared in this section. Furthermore, the results and

analysis section is categorized into two subsections. Subsection 5.1 discusses the experimental results of the proposed system. In subsection 5.2, we perform an ablation study that explores the effects of layers and analysis of different input features.

### 5.1 Experimental results

The Indian speech corpora consist of recorded speech of seven low resource Indian languages. We have considered 1000 audio samples for each language and split them into training and validation set by 80:20 ratio. All models are trained using Tesla A100 GPU Server using deep learning libraries such as tensorflow and keras. An average computational time of 2.3 hours has been recorded for training and validating the model.

RNN-GRU model captures the frequency pattern of MFCC at each time step and produce the transcribed text which is further processed by Indic BERT model. The advantage of using Indic BERT model is that it is able to mitigate spell error by analysing the sentence more contextually. The sample transcribed output on each Indian language of the proposed system is shown in figure 4 with International Phonetic Alphabet (IPA), where the unigrams with spell errors are highlighted in red colour. IPA provides an internationally recognized set of phonetic symbols to represent the pronunciation of languages. Hence, it is used for the readability of the Indian languages text.

The effectiveness of speech to text model is evaluated using two metrics such as word error rate (WER) and sequence match ratio. WER is a common evaluation metric of ASR system, which can be calculated using the Damerau-Levenshtein distance as given in equation (10).

$$WER = \frac{(I + D + S)}{N} \quad (10)$$

Where,  $S$  is the number of substitutions,  $D$  is the number of deletions,  $I$  is the number of insertions and  $N$  is the number of words in the sentence.

From the results it is inferred that, when Indic BERT model is combined with GRU, it achieves an average WER of 0.52 which is a reliable improvement over GRU model. Table 2 compares the performance of GRU and GRU-Indic BERT. The results show that Indic BERT spell corrector model improved the overall accuracy and reduced the average WER by 10%.

Sequence match ratio (SMR) is another metric for evaluating an actual sentence to a predicted sentence by counting matching unigram tokens as in equation (11). Table 3 shows the sequence match ratio of first hundred samples, with an average match ratio of 85%.

Languages	Recurrent Neural Network (Gated Recurrent Unit)	
	True Transcription	Predicted Transcription
Tamil	இசை நிகழ்ச்சியை நவீன தொழில்நுட்ப வசதியுடன் பிரம்மாண்டமான முறையில் நடத்த திட்டமிட்டிருக்கிறோம்  (ʔɪsəj nɪgəʔtʃɪʔjəj nəvi:nə tɔɻɪlnɪpə vəsətɪjɐdɐn pɪrəmmɑ:ndəma:nə mɔɻəjɪl nədətɪ ə tɪttəmiɪtɪrɪkkɪrɔ:m)	இசை நிகழ்ச்சியை யாவினத தொழில்நுட்ப வசதியுடன் பிரம்மாண்டமான முறையில் நடத்தை திட்டம்மிறுதிருக்கிறோம்  (ʔɪsəj nɪgəʔtʃɪʔjəj jɑ:vɪnədə tɔɻɪlnɪpə vəsətɪjɐdɐn pɪrəmmɑ:ndəma:nə mɔɻəjɪl nədətɪ tɪttəmmɪrɪrɪkkɪrɔ:m)
Telugu	అనగ మన దేశానికి హిందీ అధికర భష  (əɳəɡɑ: məɳə dɛ:ʃɑ:nɪki hɪndi: ədʰɪkɑ:rə bʰɑ:ʃə)	అనగ మన దేశానికి హిందీ అధికర భష  (əɳəɡɑ: məɳə dɛ:ʃɑ:nɪki hɪndi: ədʰɪkɑ:rə bʰɑ:ʃə)
Malayalam	അവസാനം ചെ സാന്താ ക്ലാർ ആക്രമിക്കാനായി തന്നെ ആത്മഹത്യാ സംഘത്ത് തയ്യാറാക്കി  (əvəʃɑ:nɔ̃ tʃe ʃɑ:ntɑ: kla:ra ɑ:kɾəmi:kka:nɑ:jɪ tənne ɑ:tmaɦətjɑ: sə̃ɡʱətte tɔjja:ra:kkɪ)	അവസാനം ചെ സാന്താക്ലാർ ആക്രമിക്കാനായി തന്നെ ആത്മഹത്യാ സംഘത്ത് തയ്യാറാക്കി  (əvəʃɑ:nɔ̃ tʃe ʃɑ:ntɑ:kla:ra ɑ:kɾəmi:kka:nɑ:jɪ tənne ɑ:tmaɦətjɑ: sə̃ɡʱətte tɔjja:ra:k)

Figure 4. Transcriptions of Indian Languages (with IPA).

Table 2. WER of RNN-GRU and RNN-GRU + Spell Correction (INDIC BERT).

Language	GRU (train)	GRU (validation)	GRU+Indic BERT (train)	GRU+Indic BERT (validation)
Tamil	0.59	0.84	0.48	0.73
Telugu	0.57	0.83	0.46	0.71
Malayalam	0.48	0.87	0.39	0.78
Kannada	0.51	0.83	0.42	0.75
Bengali	0.47	0.80	0.39	0.72
Marathi	0.30	0.65	0.21	0.53
Gujarati	0.36	0.67	0.28	0.56
Average	0.46	0.78	0.37	0.68

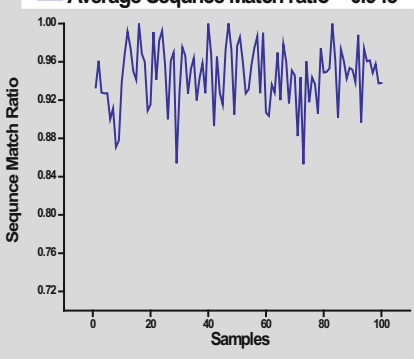
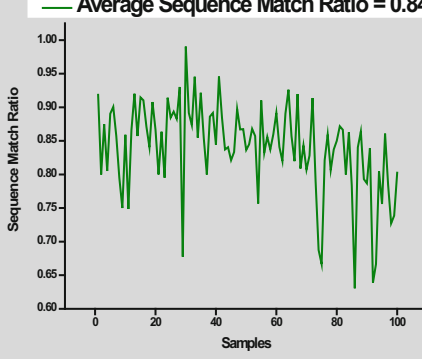
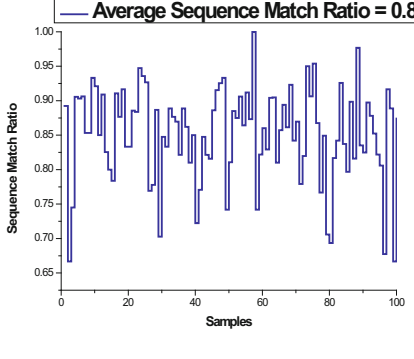
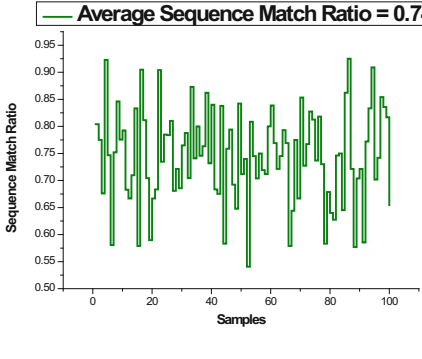
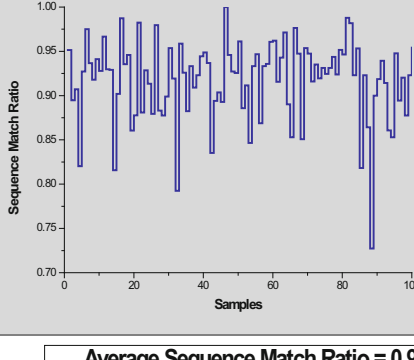
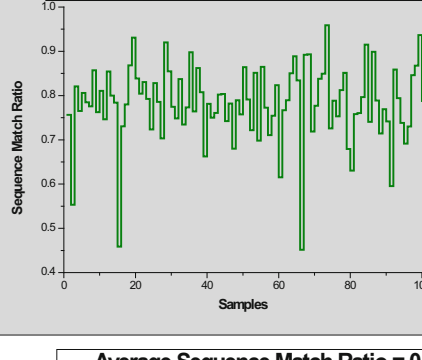
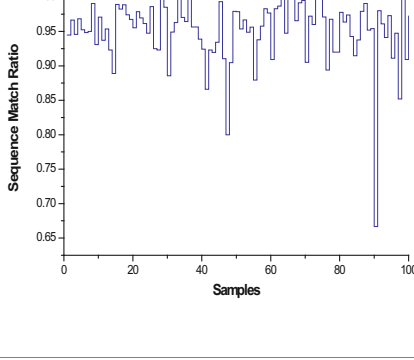
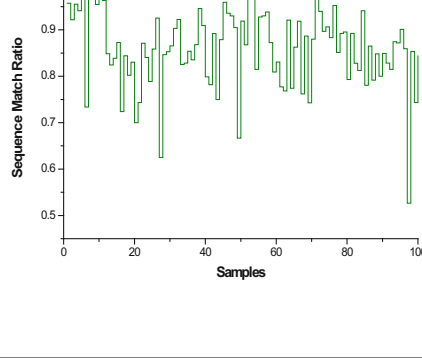
Sequence Match Ratio (actual, predicted)

$$= \frac{\text{No of Matching Unigrams}}{\text{Max}(\text{len}(\text{actual}), \text{len}(\text{predicted}))} * 100 \quad (11)$$

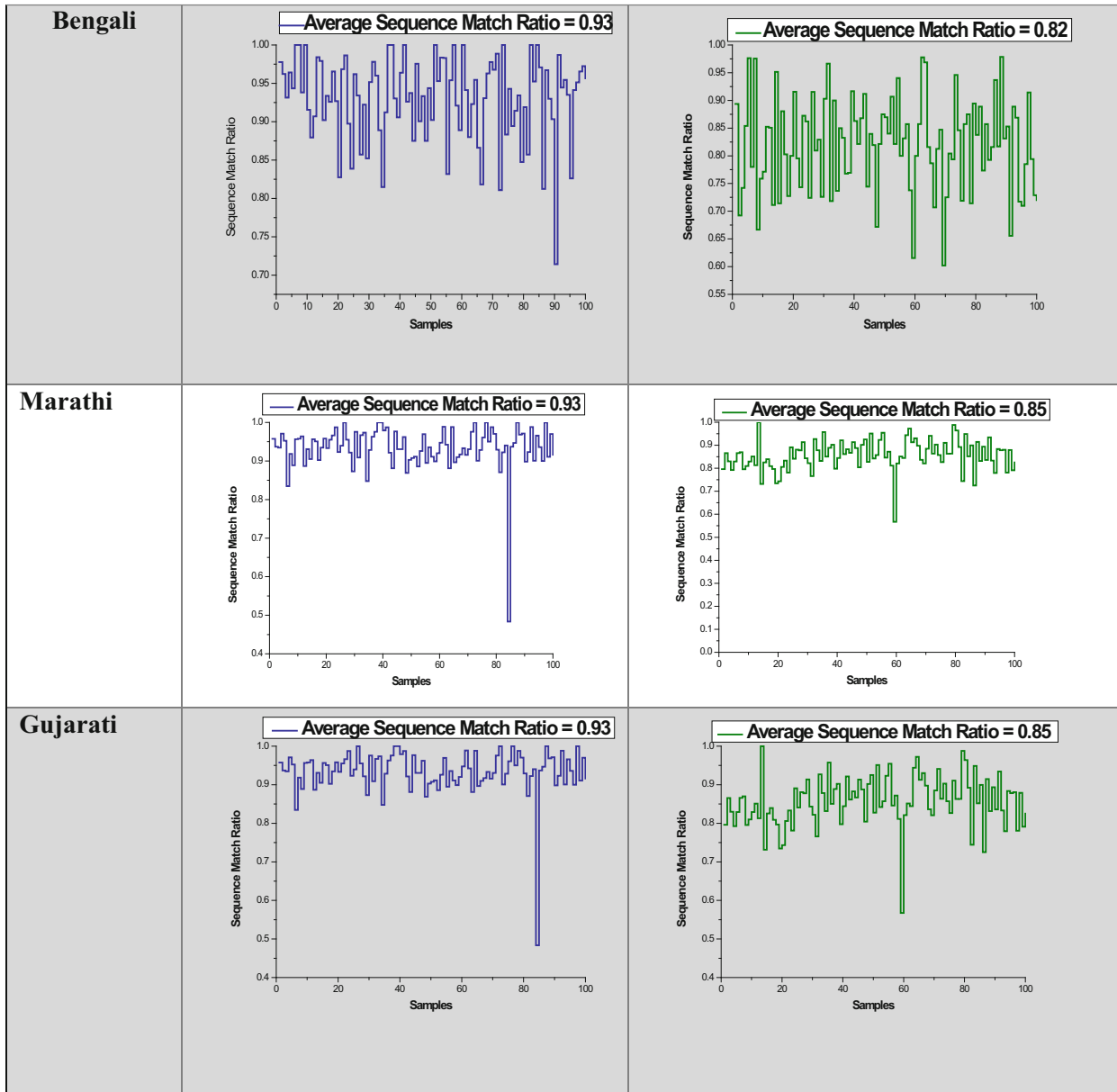
Where, len(actual), len(predicted) is the number of tokens in actual and predicted sentence.

Intuitively, from the results, it is inferred that the large number of morphological variations in Indian languages causes the validation error to be higher which has a significant impact on the increase in generalization error. One

**Table 3.** Summary of our results: sequence match ratio of seven Indian Languages.

Languages	Train Sequence Match Ratio	Validation Sequence Match Ratio
Tamil	<p>— Average Sequence Match ratio = 0.945</p> 	<p>— Average Sequence Match Ratio = 0.84</p> 
Telugu	<p>— Average Sequence Match Ratio = 0.85</p> 	<p>— Average Sequence Match Ratio = 0.74</p> 
Malayalam	<p>— Average Sequence Match Ratio = 0.91</p> 	<p>— Average Sequence Match Ratio = 0.78</p> 
Kannada	<p>— Average Sequence Match Ratio = 0.95</p> 	<p>— Average Sequence Match Ratio = 0.86</p> 



**Table 3.** continued

way to achieve better generalization is to leverage the size of high quality Indian speech data.

### 5.2 Ablation study

To analyze the effectiveness of BERT based spell corrector module, we experimented with an ablation study on the English language using GRU models with and without BERT. We present the ablation study results by comparing the WER on RNN-GRU and RNN-GRU + BERT base model using Librispeech corpus as shown in table 4 with the same training settings. The results are compared against

the state-of-the-art systems and it is inferred that the WER of GRU-BERT is comparatively better as shown in table 5. Removing the spell corrector module raises the WER of the model. This implies that employing the BERT spell corrector module along with end-to-end speech recognition can collectively improve the performance of the overall system.

To understand the importance of feature representation, we performed an ablation study using three different features such as spectrogram, MFCC and log-mel filterbank energies. The spectrogram is a visual time-frequency representation of spectral energies at different frequency

**Table 4.** Librispeech data.

Subset	Hours	Female speakers	Male speakers
Dev-clean	5.4	20	20
Test-clean	5.4	20	20

**Table 5.** Comparative analysis on Librispeech Corpus (LM - Language Model, SC-Spell Correction).

Model	WER (%) on dev clean	WER (%) on dev clean
ASR with Bi-Directional Transformer [28]	7.65	18.97
Word Level CTC + 4 gram [29]	6.3	6.8
Listen Attend Spell + LSTM (SC) [30]	5.04	5.08
Proposed System (RNN-GRU + BERT Base)	4.95	5.01

**Table 6.** Ablation studies of different features.

Input features	WER on dev-clean (Clean speech) (%)	WER on dev-other (noisy speech) (%)
MFCC	6.1	19.8
Log mel filter bank energies	4.9	24.7
Spectrogram	11.3	35.1

values. MFCC is a representation of speech features, which can be obtained by applying the discrete cosine transform (DCT) to a Mel-frequency spectrogram. Spectrogram uses 161 input dimensions whereas, MFCC utilizes only 13 cepstral dimensions. Log mel filterbank energies (MFB) is widely used features for clean speech ASR system as their delta parameters degrade the performance of noisy speech [31]. We present the results by comparing the WER on the RNN-GRU model using dev-clean(clean speech) and dev-other(noisy speech) of LibriSpeech with the same training settings. From the results, it is clear that MFCC and log mel-filterbank energies encourage the learning better for RNN-GRU based ASR system in dev-clean(clean speech) whereas log mel-filterbank energies show lower WER for noisy speech as shown in table 6.

The following are the main takeaways from the ablation study

- The BERT language model corrects the spelling errors of the ASR system and enhances the overall speech recognition performance.
- MFCC features yield better performance for noisy input speech when compared to log mel-filter bank energies and spectrogram.

## 6. Conclusion

In this work, we proposed a fused framework of RNN-GRU with Indic BERT, to enrich the user experience of uttering native Indian languages to interact with devices around them ubiquitously. Results witnesses that RNN GRU blended with Indic BERT achieved consistent performance across the Indian languages with an average word error rate of 0.52. Our exploration will open up an opportunity to develop ASR for other low-resource Indian languages. In the future, our proposed approach can be extended to other Indian languages as well and explore the possibilities of cross-lingual Indian speech recognition.

## Acknowledgements

Our sincere thanks to Department of Science and Technology, Government of India for funding this project under Department of Science and Technology Interdisciplinary Cyber Physical Systems (DST - ICPS) scheme.

## References

- [1] Miao H, Cheng G, Zhang P and Yan Y 2020 Online hybrid CTC/attention end-to-end automatic speech recognition architecture. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 28: 1452–1465
- [2] Graves A and Jaitly N 2014 Towards end-to-end speech recognition with recurrent neural networks. In: *International conference on machine learning*. 1764–1772
- [3] Zhang Y, Alder M and Togneri R 1994 Using Gaussian mixture modeling in speech recognition. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. 1: 1–613
- [4] Baum LE, Petrie T, Soules G and Weiss N 1970 A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*. 41: 164–171
- [5] Aldarmaki H, Ullah A, Ram S and Zaki N 2022 Unsupervised automatic speech recognition: A review. *Speech Communication*. 139: 76–91
- [6] Sir George Grierson 1928 Sir George Grierson and the Linguistic Survey of India. *Journal of the Royal Asiatic Society of Great Britain and Ireland*. 3: 711–718
- [7] Li B, Sainath T N, Sim K C, Bacchiani M, Weinstein E *et al* 2018 Multi-Dialect Speech Recognition with a Single Sequence-to-Sequence Model. *IEEE International Conference on Acoustics, Speech and Signal Processing*. 4749–4753
- [8] H Miao, G Cheng P Zhang and Y Yan 2020 Online Hybrid CTC/Attention End-to-End Automatic Speech Recognition Architecture. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 28: 1452–1465
- [9] Adams O, Wiesner M, Watanabe S and Yarowsky D 2019 Massively multilingual adversarial speech recognition. In: *Proceedings of the 2019 Conference of the North American*

- Chapter of the Association for Computational Linguistics: *Human Language Technologies; Minneapolis, Minnesota, USA*. pp 96–108
- [10] Wu B, Sakti S, Zhang J and Nakamura S 2022 Modeling unsupervised empirical adaptation by DPGMM and DPGMM-RNN hybrid model to extract perceptual features for low-resource ASR. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 30: 901–916
- [11] Qin S, Wang L, Li S, Danj J and Pan L 2022 Improving low-resource tibetan end-to-end ASR by multilingual and multilevel unit modeling. *EURASIP Journal on Audio, Speech, and Music Processing*. 2: 1–10
- [12] Tachbelie M Y, Abate S T and Schultz T 2022 Multilingual speech recognition for global phone languages. *Speech Communication*. 140: 71–86
- [13] Singh A, Kadyan V, Kumar M and Bassan N 2020 ASRoIL: a comprehensive survey for automatic speech recognition of Indian languages. *Artificial Intelligence Review*. 53: 3673–3704
- [14] Li B, Chang S Y, Sainath T N, Pang R, He Y, Strohmaier T and Wu Y 2020 Towards fast and accurate streaming end-to-end ASR. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 6069–6073
- [15] Wu Z, Li B, Zhang Y, Aleksic P S and Sainath TN 2020 Multistate encoding with end-to-end speech RNN transducer network. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 7819–7823
- [16] Cho K, Van Merriënboer B, Bahdanau D and Bengio Y 2014 On the properties of neural machine translation: Encoder-decoder approaches. In: *Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. 103–111
- [17] Bengio Y, Simard P and Frasconi P 1994 Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*. 5: 157–166
- [18] Chan W, Jaitly N, Le Q and Vinyals O 2016 Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: *IEEE international conference on acoustics, speech and signal processing*. 4960–4964
- [19] Amodei D, Ananthanarayanan S, Anubhai R, Bai J, Battenberg E *et al* 2016 Deep speech 2: end-to-end speech recognition in English and mandarin. In: *International Conference on Machine Learning*. 48: 173–182
- [20] Zhang S, Huang H, Liu J, Li H 2020 Spelling Error Correction with Soft-Masked BERT. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 882–890
- [21] Hao Y, Dong L, Wei F and Xu K 2020 Investigating Learning Dynamics of BERT Fine-Tuning. In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. 87–92
- [22] Kakwani D, Kunchukuttan A, Golla S, Gokul NC *et al* 2020 IndicNLP Suite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. *Findings of the Association for Computational Linguistics: EMNLP*. 4948–4961
- [23] Jain, K, Deshpande A, Shridhar k, Laumann, F and Dash A 2020 Indic-Transformers: An Analysis of Transformer Language Models for Indian Languages. ArXiv, abs/2011.02323
- [24] Neto A F, Bezerra B and Toselli A 2020 Towards the Natural Language Processing as Spelling Correction for Offline Handwritten Text Recognition Systems. *Applied Sciences*. 10: 7721
- [25] Didenko B and Shaptala Julia 2019 Multi-headed Architecture Based on BERT for Grammatical Errors Correction. In: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. 246–251
- [26] He F, Chu S C, Kjartansson O, Rivera C, Katanova A *et al* 2020 Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. 6494–6503
- [27] Lovelyn Rose S, Ashok Kumar L and Karthika Renuka D 2019 *Deep Learning using Python*, Wiley
- [28] Liao L, Afedzie Kwofie F, Chen Z, Han G, Wang Y *et al* 2022 A bidirectional context embedding transformer for automatic speech recognition. *Information*. 13
- [29] Collobert R, Hannun A, Synnaeve G 2020 Word-level speech recognition with a letter to word encoder. In: *International Conference on Machine Learning*. 2100–2110
- [30] Guo J, Sainath TN, Weiss RJ 2019 A spelling correction model for end-to-end speech recognition. In: *IEEE International Conference on Acoustics, Speech and Signal Processing; Brighton, United Kingdom*. 5651–5655
- [31] Yun-Peng Wu, Jia-Min Mao, Wei-Feng Li 2016 Robust speech recognition by selecting mel-filter banks. In: *2nd Annual International Conference on Electronics, Electrical Engineering and Information Science; Xian, China*. 407–416