EXPLORING DEEP SPECTRAL AND TEMPORAL FEATURE REPRESENTATIONS WITH ATTENTION-BASED NEURAL NETWORK ARCHITECTURES FOR ACCENTED MALAYALAM SPEECH- A LOW-RESOURCED LANGUAGE

Rizwana Kallooravi Thandil[1*][0000-0001-9305-681X] , Mohamed Basheer K.P[2][0000-0003-3847-4124]

Article History: **Received: 28.03.2023** **Revised: 20.04.2023** **Accepted: 06.05.2023**

Abstract.
Constructing an Accented Automatic Speech Recognition (AASR) system for a language is a challenging endeavor due to variations in pronunciation, intonation, and rhythm. This study proposes a novel approach to AASR for Malayalam speech, employing Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and BiDirectional Long Short-Term Memory (BiLSTM) architectures, each incorporating an attention block.To conduct the study, the authors assembled an accented speech corpus containing samples from five different accents in Malayalam. The research was carried out in six distinct phases, involving different combinations of features and model architectures. In the initial phase, Mel Frequency Cepstral Coefficients (MFCC) were used for feature vectorization, while RNN was employed for modeling the accented speech data. This phase resulted in a Word Error Rate (WER) of 11.98% and a Match Error Rate (MER) of 76.03%.The second phase employed MFCC and tempogram methods for feature vectorization, combining them with RNN and an attention mechanism for constructing a unified model for the accented data. This phase achieved a WER of 7.98% and an MER of 82.31%. In the third phase, MFCC and Tempogram feature vectors were utilized with LSTM for modeling the accented data, resulting in a WER of 8.95% and an MER of 83.64%.The fourth phase used the same feature set as phases two and three, incorporating LSTM with attention mechanisms for constructing the accented model. This phase yielded a WER of 3.8% and an MER of 87.11%. The fifth and sixth phases utilized BiLSTM and BiLSTM with attention mechanisms, respectively, while maintaining the same feature set. These phases achieved WERs of 3.5% and 3.25% and MERs of 90.12% and 92.25%, respectively.The experiments demonstrate that the BiLSTM with attention mechanism architecture, incorporating appropriate accent attributes, performed well even for unknown accents. Performance evaluation using WER and MER indicated a reduction of 50% to 65% when employing attention mechanisms with RNN, LSTM, and BiLSTM approaches.

Keywords: Accented Speech Recognition, Human Machine Interaction, Spectral Feature Vectorization, RNN, LSTM, BiLSTM, Attention Mechanisms, Malayalam Speech Recognition, ASR for Low Resourced Language

[1*,2]Sullamussalam Science College, Areekode, Kerala, India, [1]Email: ktrizwana@gmail.com,
[2]Email: mbasheerkp@gmail.com

***Corresponding Author:** Rizwana Kallooravi Thandil
*Sullamussalam Science College, Areekode, Kerala, India, E-mail: ktrizwana@gmail.com

DOI: 10.53555/ecb/2023.12.si5a.0388

1 Introduction

Malayalam is predominantly spoken in the Indian state of Kerala and the Lakshadweep Islands, with a wide range of accents used by its speakers. Consequently, the significance of Accented Automatic Speech Recognition (AASR) systems capable of handling accented Malayalam speech has been increasing. This study focuses on the development of a unified speech recognition model specifically designed for accented speech in the Malayalam language. The primary objective is to assess the performance of the proposed techniques in recognizing speech signals from multi-accented speakers. The dataset employed in this study comprises accented Malayalam speech samples collected from native Malayalam speakers residing in five different districts of Kerala. The study investigates the effectiveness of Mel Frequency Cepstral Coefficients (MFCC) and Tempogram features, in conjunction with Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), BiDirectional Long Short-Term Memory (BiLSTM), and attention mechanisms, to enhance the accuracy of accented speech recognition. MFCC and Tempogram features are widely used in speech recognition tasks and have demonstrated promising outcomes in capturing the acoustic characteristics of speech.

The results of this study illustrate the practicality of employing deep learning techniques and engineering the accent vectors to recognize accented speech in Malayalam. To improve the precision of our models, we incorporate attention mechanisms, enabling the network to concentrate on the most significant features for recognition. We evaluate the performance of our proposed models against baseline models trained on conventional feature extraction methods and establish the superiority of our models in accurately recognizing accented speech.

The study reveals that the utilization of MFCC and tempogram features in conjunction with attention mechanisms in the proposed models surpass the performance of traditional models when it comes to recognizing accented speech. In addition, we present a comprehensive analysis of the outcomes, emphasizing the obstacles encountered and the prospects for further investigation in this field. The findings indicate that employing BiLSTM and attention mechanisms holds promise for enhancing the accuracy and reducing word error rates in ac cented Malayalam speech recognition, thus carrying significant implications for applications such as language learning, speech-to-text transcription, and voice-enabled interfaces. The paper presents several key contributions, as outlined below:

1. Construction of Accented Speech Dataset: The authors have developed a dataset specifically for accented speech, which serves as the foundation for conducting the study. This dataset facilitates the evaluation and analysis of various approaches to recognizing accented speech.
2. Experimental Phases and Approaches: The experiment is conducted in six distinct phases, employing six different approaches. This comprehensive evaluation enables the researchers to compare and identify the most effective approach for accented speech recognition.
   3. Analysis of Spectral Features in Accent Identification: The study delves into the analysis of spectral features and their implications in accent identification. This analysis sheds light on the relevance and effectiveness of these features in accurately recognizing accents in speech.
4. Novel Approach for Gradient Optimization: The paper introduces a novel approach aimed at mitigating the issues of vanishing and exploding gradients while maximizing the expectation. This technique enhances the training process and improves the overall performance of the accented speech recognition models.

These contributions collectively contribute to advancing the understanding and capabilities of accented speech recognition, offering insights into dataset construction, experimental methodologies, spectral feature analysis, and gradient optimization techniques.

2 Related Work

Accented Automatic Speech Recognition (AASR) has presented challenges due to the inherent variability in speech patterns stemming from factors such as pronunciation, intonation, and rhythm. However, recent advancements in deep spectral feature representations and attention-based neural network architectures have shown promise in enhancing accented speech recognition. This is particularly crucial for low-resourced languages that encompass a wide range of accents, as accurately recognizing speech in such languages poses a significant challenge. Recent studies have thus focused on leveraging deep spectral feature representations and attention-based neural network architectures to address this challenge and improve the accuracy of accented speech recognition. These approaches have the potential to overcome the difficulties associated with diverse accents and contribute to more robust and effective AASR systems.

The study conducted by H. Bao et. al [1] demonstrated the effectiveness of the multitask learning

approach with the attention mechanism in handling the challenges of accented speech recognition. The authors' findings suggest that leveraging shared knowledge and incorporating attention mechanisms can lead to significant improvements in the accuracy of recognizing accented speech. Wu et al. [2] present a survey on accent-robust acoustic modeling for automatic speech recognition (ASR). The paper provides an overview of recent advancements in this field, including deep learning approaches, and discusses techniques for improving ASR performance in the presence of accents. Fernández-Gavilanes et al. [3] propose an unsupervised approach for accent classification in accented speech recognition. They leverage clustering algorithms and deep neural networks to automatically classify accents without relying on labeled accent data, offering a promising method for handling accents in ASR systems. Dai et al. [4] explore the application of transfer learning techniques for accented speech recognition. They investigate pretraining models on large-scale multilingual datasets and fine-tune them for accented speech recognition tasks, demonstrating improved performance by leveraging transfer learning. Ma et al. [5] propose a multilingual and multi-accent end-to-end speech recognition system based on the Transformer architecture. Their system is designed to recognize accented speech from various languages and accents, and it achieves competitive performance, highlighting the effectiveness of the Transformer model in handling accented speech.

Baevski et al. [6]proposes techniques and architectures that can be adapted for self-supervised learning in accent identification and accented speech recognition tasks.Jansen et al.[7] discusses the use of self-supervised pretraining for speech recognition on low-resource languages. It explores techniques to leverage unlabeled data to improve the performance of automatic speech recognition (ASR) systems. The study demonstrates the effectiveness of self-supervised learning in addressing the challenges of limited training data in low-resource language scenarios.Leng et al.[8]investigates the use of contextualized representations to capture the distinctive features of different accents, leading to improved recognition accuracy. Choudhry et al.[9] explores the discriminative representations for accent identification tasks by investigating the effectiveness of transfer learning tasks to improve accent classification performance.Bansal et al.[10][11]in their work explores methods to learn informative representations from unlabeled accented speech data, which can then be used to enhance the performance of ASR systems.

All the above studies have demonstrated the effectiveness of deep spectral feature representations and attention mechanisms in improving accented speech recognition. Attention-based neural network architectures have shown promising results in improving accented Malayalam speech recognition. These studies have demonstrated the effectiveness of deep spectral feature representations in capturing acoustic characteristics of accented speech and the importance of attention mechanisms in improving recognition accuracy. These findings have significant implications for developing accurate speech recognition systems for low-resourced languages like Malayalam and could be useful in various dimensions.

3 Methodology

The task of Accented Automatic Speech Recognition (AASR) poses significant challenges, especially for low-resource languages such as Malayalam. ASR and AASR for the Malayalam language are still in their early stages of development. The availability of publicly accessible data for conducting studies is extremely limited. Therefore, to address this limitation, the authors created an accented dataset consisting of multisyllabic words specifically for this study. The experiment was conducted in six distinct phases, and the detailed steps involved in the experiment are illustrated in Figure 1.

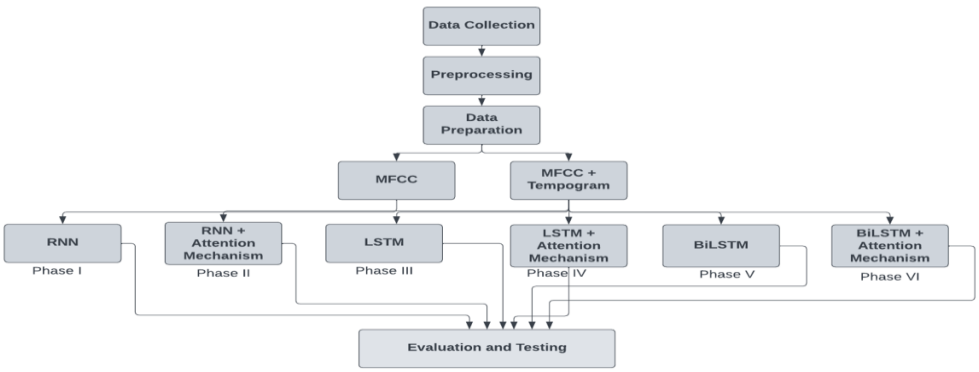

Figure 1 Steps Involved in the Proposed Methodology

3.1Data collection

The primary challenge encountered during this study was the lack of a benchmark dataset containing accented speech. The unavailability of such data significantly hindered the progress of the experiment. To address this issue, the authors took the initiative to construct a speech corpus specifically for this study. The corpus was created under natural recording conditions and consisted of approximately 1.17 hours of accented speech. The construction of the corpus involved capturing individual utterances of multisyllabic words lasting between two to five seconds. Data collection was conducted in five different districts of Kerala, where speakers of Malayalam exhibit diverse accents. Forty participants, comprising twenty males and twenty females, were involved in the data collection process. The selected districts were Kasaragod, Kannur, Kozhikode, Wayanad, and Malappuram in Kerala. The accents included in the corpus were heavily influenced by the languages spoken in the neighboring states that share borders with these districts.

3.2Data Preprocessing

In this study, the data preprocessing phase involved the utilization of two methods: the MFCC (Mel-frequency cepstral coefficients) and the Tempogram approach. MFCC algorithm is used in this study owing to its capability in feature vectorization, which aids in capturing important speech characteristics. Additionally, Tempogram features were extracted from the speech signals during audio preprocessing. These Tempogram features capture the tempo and rhythmic information that closely align with the accent details present in the speech. For this experiment, the MFCC vectors and a combination of the vectors derived from both MFCC and Tempograms were concatenated together at various stages of the experiment. This approach aimed to incorporate both the spectral information from MFCC and the rhythmic information from Tempograms, enhancing the overall representation of the speech data for improved analysis and recognition.

Table 1 The statistics of the accented samples

District	No. of Audio Samples
Kasaragod	1360
Kannur	1360
Kozhikode	1690
Malappuram	1360
Wayanad	1300
Total	7070

In the domain of accented speech recognition, the MFCC approach assumes a pivotal role in capturing and representing the distinctive spectral attributes of accented speech. Accents introduce variations in pronunciation, phonetic patterns, and prosody, which consequently impact the acoustic properties of speech signals. By extracting MFCC features from accented speech, it becomes possible to capture significant spectral information related to the specific accent being considered. The MFCC algorithm effectively calculates the mel-frequency filterbank energies and employs a discrete cosine transform (DCT) to derive the cepstral coefficients. These coefficients form a representation of the speech signal's spectral envelope, encapsulating crucial details about its phonetic content. In the context of accented speech recognition, the utilization of MFCC features enables discrimination between different accents and enhances the accuracy of the recognition process. By employing MFCCs, the recognition system can effectively capture accent-specific characteristics, including spectral variations, differences in vowel quality, and prosodic patterns. The steps involved in extracting the first 13 MFCC coefficients in this study are:

1. Preprocessing:
Apply a windowing function, such as the Hamming window, to the speech signal to segment it into frames of equal duration: x_w[n] = x[n] * w[n], where w[n] is the window function;x[n] is the speech signal, where n is the sample index.

2. Fast Fourier Transform (FFT):
Compute the N-point discrete Fourier transform (DFT) of each frame to obtain the magnitude spectrum: X[k] = |FFT(x_w[n])|, where X[k] represents the magnitude spectrum, and k is the frequency bin index.

3. Mel-filterbank:
Apply a mel-filterbank to the magnitude spectrum to emphasize relevant frequencies: H[m, k] = |H_m[k]|, where H[m, k] represents the magnitude response of the m^{th} mel-filter at the k^{th} frequency bin.

4. Logarithm:
Take the logarithm of the filterbank outputs to convert the magnitudes to a logarithmic scale: M[m] = log10(sum(H[m, k])), where M[m] represents the logarithmic filterbank output for the m^{th} filter.

5. Discrete Cosine Transform (DCT):
Apply a discrete cosine transform (DCT) to the logarithmic filterbank outputs to decorrelate the

coefficients: C[k] = sum(M[m] * cos((m * pi * (2k + 1)) / (2N))), where C[k] represents the DCT coefficient for the k^{th} index, and N is the total number of mel-filterbank outputs.

6. Selecting the 13 coefficients:

Retain the first 13 DCT coefficients to capture the most relevant information about the spectral envelope of the speech signal: MFCC = [C[0], C[1], ..., C[12]].

The C[0] coefficient represents the overall energy in the signal, while the C[1] corresponds to the spectral flatness, indicating the uniformity of the signal's spectrum. The C[2] represents the spectral centroid, which is the frequency dividing the power spectrum into two halves. The C[3] denotes the spectral roll-off, indicating the frequency below which a certain percentage of the total power in the spectrum is contained. Furthermore, the C[4] to C[6] represent the first three formants of the signal, representing the resonant frequencies of the vocal tract. These formants play a crucial role in determining the phonetic content of the speech. Finally, the C[7] to C[12] coefficients capture the higher-order cepstral coefficients, which provide detailed information about the spectral envelope of the speech signal. By extracting these 13 coefficients, we obtain a comprehensive representation of the accented speech signals, encompassing important aspects such as energy, spectral characteristics, vocal tract resonances, and fine spectral details.

After the initial extraction of the 13 coefficients, we proceeded to compute their first and second derivatives, resulting in the formation of 39 vector representations. The first derivative coefficients, obtained by differentiating the coefficients with respect to time, capture the rate of change of spectral features over time. This information includes variations in pitch, loudness, and spectral content. On the other hand, the second derivative of the MFCC coefficients yields coefficients that portray the acceleration of spectral features over time. This reveals rapid changes in the signal, such as the onset and offset of speech sounds. By incorporating the first and second derivatives, we effectively obtained a set of 39 feature vectors. To finalize the feature representation, we computed the mean value of all 39 coefficients and appended it to the vector list, resulting in the formation of 40 MFCC coefficients for this study. This comprehensive feature set captures relevant information regarding both the dynamic changes and overall spectral characteristics of the accented speech signals.

Tempogram features were employed in this study to focus on the accent and rhythm-specific features of the signals. We have extracted 384 speech vectors using tempogram speech extraction techniques to conduct this study. Tempogram is a speech analysis tool used to represent the rhythmic structure of speech and music. It is similar to a spectrogram, which represents the frequency content of a sound signal over time, but instead of frequency, the y-axis of a tempogram represents the tempo of the speech, while the x-axis represents time. This is an effective technique in capturing the tempo and accent-specific features of the speech signal and improving the performance of speech recognition systems, particularly in noisy environments or when the speech signal is distorted. It involves segmenting the speech signal into short frames, computing the spectrogram, calculating the autocorrelation to obtain the tempogram, and normalizing the tempogram to a fixed tempo.

3.3Accented Model Construction

Accented speech recognition for the Malayalam language has emerged as a prominent area of research, with numerous studies conducted in recent years. Malayalam, known for its distinctive phonological features, presents a significant challenge for speech recognition systems in accurately transcribing accented speech. The variations in pronunciation, intonation, and rhythm between accented speech and standard speech further complicate the task for these systems. To address this challenge, the authors have developed six models specifically tailored for recognizing accented speech in Malayalam. The experiment encompasses six distinct phases, each meticulously examined to explore the most effective approach for tackling this issue. The subsequent sections will provide a comprehensive discussion of each phase of the experiment.

3.3.1 Phase 1: The RNN Approach

The utilization of recurrent neural networks (RNNs) offers a valuable solution for accented speech recognition by effectively processing audio signals in a sequential manner. RNNs excel in capturing the temporal dependencies that exist among individual audio frames, making them particularly well-suited for this task. A notable advantage of RNNs lies in their ability to handle audio data of varying lengths and learn intricate long-term relationships within the sequences. By sequentially analyzing the audio signal, RNNs enable the modeling of contextual and temporal information, thereby facilitating accurate recognition of accented speech. The inherent capability of RNNs to capture dynamic patterns within sequential data

positions them as a highly suitable approach for addressing the specific challenges presented by accented speech in the Malayalam language.

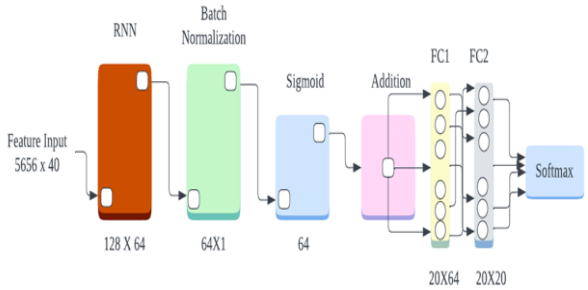

Figure 2 Proposed RNN

The RNN architecture, depicted in Figure 2, receives a feature input of 40 Mel-frequency cepstral coefficient (MFCC) features. These features serve as the input to the RNN network layers. The RNN network operates by sequentially processing the input sequences, one at a time, while retaining a contextual understanding within the network. The resulting output from the RNN layer is then forwarded to the batch normalization layer, where the data is normalized. Once the normalization is completed, the data is further passed to the Sigmoid layer.

Subsequently, the output vectors from the various layers are concatenated and summed together before being transmitted to the softmax layer. The softmax layer employs a probabilistic function to predict the target class, determining the class with the highest probability. This final output represents the classification result for the given input. Overall, this architecture employs a systematic flow, incorporating RNN processing, normalization, activation functions, and concatenation, to enable accurate prediction of the target class using the provided MFCC features..

3.3.2 Phase 2: RNN with Attention Mechanism

In this phase of the experiment, the focus was on utilizing 424 Mel-frequency cepstral coefficient (MFCC) and tempogram feature vectors. To construct the accented model, a recurrent neural network (RNN) with an attention block architecture was employed. This architecture allows the model to selectively emphasize relevant information within the accented speech.The feature input is initially fed into the RNN architecture, followed by a dense layer. To mitigate overfitting, a dropout layer was integrated into the architecture, reducing the likelihood of the model excessively relying on specific features. Subsequently, another dense layer is incorporated. The activation functions used in the network are Sigmoid and Relu, enabling non-linear transformations to be applied during the processing.

The predictions generated by the softmax layer are then fed back into the RNN network, specifically through the attention layer. This iterative process facilitates an improved output compared to previous approaches. The integration of the attention mechanism allows the model to effectively focus on relevant aspects of the accented speech, leading to enhanced performance in recognizing and transcribing accented speech patterns.

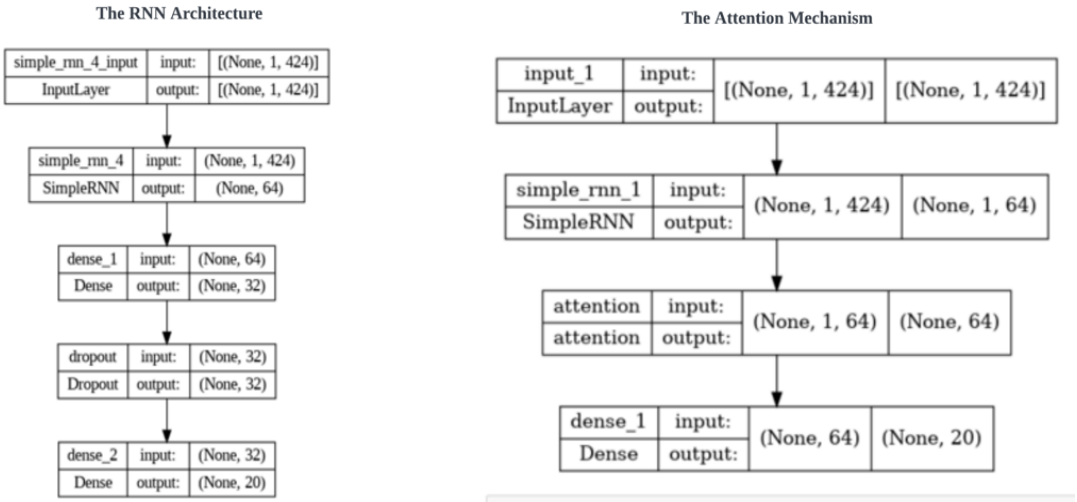

Figure 3 Proposed RNN with Attention Mechanism

3.3.3 Phase 3: The LSTM Approach

In this phase of the experiment, the feature input layer consists of concatenated vectors derived from applying the MFCC and Tempogram methods to the speech signals. Specifically, a total of 424 feature vectors were extracted from each speech signal for the purposes of this study. These feature vectors serve as the input to the LSTM layers,

which effectively mitigate the vanishing gradient issues commonly encountered in traditional RNN architectures.The output from the LSTM layers is subsequently normalized by passing it through the batch normalization layer. This normalization process ensures that the activations within each layer of the network possess zero mean and unit variance, which aids in preventing gradient explosion. The normalized output is then fed into the Relu layer, which applies a rectified linear unit activation function to introduce non-linearity.The concatenated output from these layers is further passed through dense layers before reaching the final softmax layer, which generates predictions. The softmax layer utilizes a probabilistic function to assign probabilities to different classes, ultimately determining the most likely target class. This comprehensive architecture allows for efficient processing and analysis of the concatenated feature vectors derived from the MFCC and Tempogram methods, facilitating improved accuracy in prediction tasks..

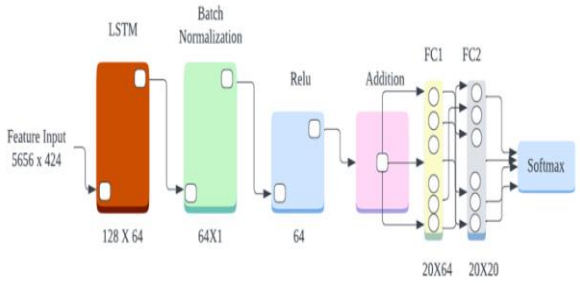

Figure 4 Proposed LSTM

3.3.4 Phase 4: LSTM with Attention Mechanism

In the fourth phase of the experiment, an LSTM model with an attention block architecture was employed. The feature vectors utilized in this phase consisted of 424 MFCC and Tempogram vectors extracted from the accented audio data.

The proposed LSTM architecture consists of two primary branches: an operational block and a skip connection block. The skip connection block branch plays a crucial role in highlighting only the relevant activations during the training process. By incorporating this skip connection, the model can focus on the most informative and discriminative features, thereby enhancing its performance.The attention block within the proposed LSTM architecture plays a vital role in reducing computational resources wasted on irrelevant activations. By selectively attending to relevant information, the attention block directs the model's focus to the most significant parts of the input sequence, optimizing computational efficiency.The inclusion of the attention block and skip connection within the LSTM model architecture allows for improved performance in accented speech recognition. These components contribute to reducing unnecessary computations and emphasizing relevant activations, ultimately enhancing the accuracy and efficiency of the system..

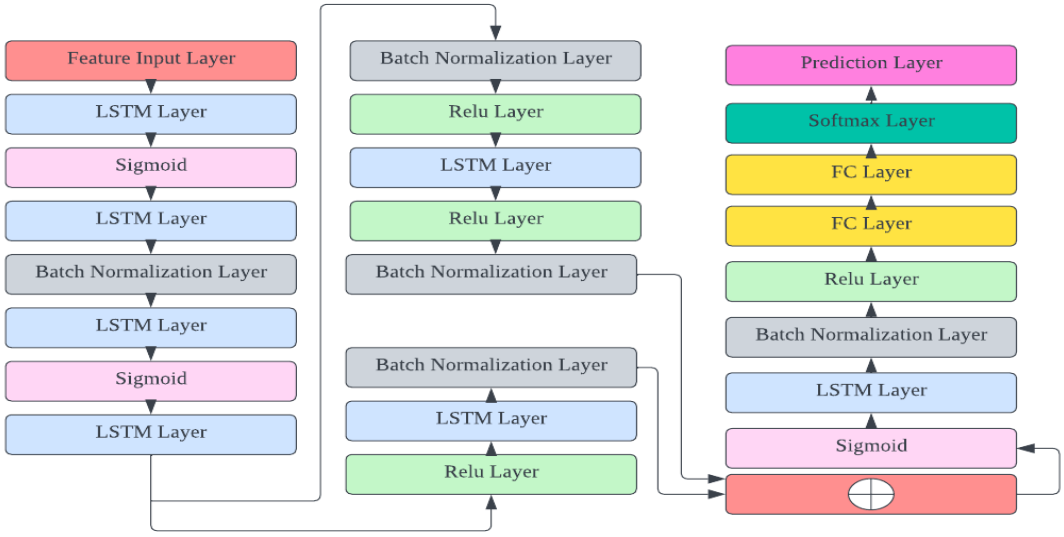

Figure 5 Proposed LSTM with Attention Block

3.3.5 Phase 5: BiLSTM Approach

In the fifth phase of the experiment, the focus was on utilizing accented speech vectors obtained by combining MFCC and Tempogram vectors. To address this, the BiLSTM methodology was employed as the model architecture. BiLSTMs leverage the strength of both forward and backward

LSTM layers to capture contextual information from both past and future contexts, enabling a comprehensive understanding of the input sequence.The architecture of a BiLSTM model typically comprises two sets of LSTM layers, with one processing the input sequence in the forward direction and the other in the backward direction. This bidirectional processing allows the model to effectively capture relevant temporal dependencies and better comprehend the intricate patterns and dynamics present in accented speech.

By combining the information from both directions, the BiLSTM model gains a holistic view of the input sequence, enabling it to accurately recognize and interpret the unique characteristics of accented speech patterns. This bidirectional approach enhances the model's ability to capture contextual information and improve its performance in accented speech recognition tasks.The utilization of BiLSTM models in this phase allowed for the effective integration of both past and future context, leading to a comprehensive understanding of the input sequence and improved recognition and interpretation of accented speech patterns.

3.3.6 BiLSTM with Attention Mechanism

This phase of experiment combines the strengths of BiLSTM and attention mechanisms to improve the accuracy of recognizing and transcribing accented speech.In this approach, the BiLSTM model is designed to capture contextual information by processing the input sequence in both forward and backward directions. The forward LSTM layer analyzes the sequence from the beginning, while the backward LSTM layer processes it from the end. This bidirectional processing allows the model to consider past and future context, enabling a more comprehensive understanding of the accented speech.

The attention mechanism assigns different weights to specific parts of the input sequence, highlighting the most relevant information for accurate recognition. By dynamically focusing on important segments, the attention mechanism helps the model adaptively allocate its resources and give more weight to crucial features.

The combination of BiLSTM and attention mechanisms enables the model to effectively capture long-term dependencies and concentrate on the salient aspects of the accented speech. This approach significantly enhances the recognition and transcription accuracy, addressing the challenges posed by pronunciation variations, intonation, and rhythm in accented speech.

3.3.7 Result and Evaluation

The study encompassed six distinct phases, each employing different sets of feature vectors, methodologies, and experimental parameters, while exploring the area of Accented Speech Recognition (AASR) for the Malayalam language. Throughout the experiments, consistent training parameters and environmental setups were maintained to ensure fair comparisons.In phase 1, the Adam optimizer was utilized, while the subsequent phases (2-6) employed the rmsprop optimizer. The initial learning rate was set at 0.001 for phases 1 and 2, and 0.01 for phases 3 to 6. The duration of the experiments varied, with phases 1 and 2 running for 3000 epochs, phases 3 and 4 for 2000 epochs, and phases 5 and 6 for 68 epochs, as the models demonstrated faster learning rates in these later phases.Throughout all phases, the categorical cross-entropy loss function was employed to evaluate and optimize the models. By consistently using this loss function, the study aimed to assess and compare the performance of different architectures and methodologies.

By conducting the experiments with varying feature vectors, methodologies, and experimental parameters, the study aimed to derive comprehensive insights into AASR for the Malayalam language. The diverse experimental setups allowed for a thorough exploration of the factors influencing the performance of the models, ultimately leading to nuanced conclusions regarding the most effective approaches for accented speech recognition in Malayalam.

The architecture underwent extensive fine-tuning through multiple experimental trials in each phase. Various combinations of optimizers, loss functions, and neural network architectures were explored during this process. However, only the refined and optimized version of the architecture is discussed and presented in this paper.The performance of the architecture in each phase was evaluated based on test and validation accuracies. These accuracies for the four different phases of the experiment are summarized in Table 2, providing a comprehensive overview of the model's performance at different stages of the study. The accuracies serve as a quantitative measure of the model's ability to accurately recognize and transcribe accented speech in the Malayalam language.The fine-tuning process involved iteratively adjusting and optimizing the architecture's parameters, configuration, and training techniques to enhance its performance. By carefully selecting the most effective combinations and configurations, the researchers aimed to achieve the best possible results in accented speech recognition for Malayalam.

The presented results in Table 2 reflect the outcomes of the fine-tuned architecture, showcasing the achieved accuracies and highlighting the progress made in each phase of the experiment.2.

Table 2 Evaluation Metrics in Terms of Accuracy and Loss

Phase	Train Accuracy	Validation Accuracy	Train Loss	Validation Loss	No.of Epochs
Phase I	87.18%	65.15%	0.0096%	0.0277%	3000
Phase II	92.02%	72.61%	0.0074%	0.0317%	3000
Phase III	94.10%	64.87%	0.0050%	0.0309%	2000
Phase IV	96.27%	73.03%	0.0031%	0.0291%	2000
Phase V	96.25%	72.45%	0.0026%	0.1093%	68
Phase VI	97.37 %	74.27%	0.0036%	0.0025%	68

Upon evaluating the performance of the experiments conducted in different phases, it is evident that the utilization of RNN with attention mechanism in Phase II significantly outperformed Phase I, where only RNN was employed. Phase II exhibited higher accuracy rates and lower loss rates compared to Phase I. Further evaluation of Phase III, where LSTM was employed, and Phase IV, where LSTM with attention mechanism was used, revealed a notable improvement in Phase IV. Phase IV demonstrated enhanced performance compared to the preceding phases of experiments, exhibiting reduced error rates.

Moreover, the phase of the experiment that employed BiLSTM in Phase V demonstrated superior performance in comparison to the phase that employed LSTM with attention mechanism. The highest performance was achieved when the model incorporated BiLSTM with attention mechanisms. The evaluation of accuracies and loss rates across the entire experiment indicates that this phase exhibited greater accuracy and lower error rates.

In addition to accuracy and loss rates, other metrics employed to evaluate the performance of this study include Word Error Rate (WER) and Match Error Rate (MER). WER is calculated as the ratio of the total number of incorrectly recognized or missing utterances to the total number of utterances in the recognized speech dataset. To compute the WER, several error types are considered:

1. Substitution errors: These errors occur when the system incorrectly recognizes a speech and replaces it with a different utterance. The count of substitution errors is denoted by S.
2. Insertion errors: These errors occur when the system adds extra utterances to the output that are not present in the original speech. The count of insertion errors is denoted by I.
3. Deletion errors: These errors occur when the system fails to recognize a speech that exists in the original speech dataset. The count of deletion errors is denoted by D.

To calculate the WER, the total number of utterances in the speech dataset (N) is determined, followed by the count of errors in the system's output that do not match the original speech data. By considering these error types, the WER provides a comprehensive evaluation of the system's performance in accurately recognizing and reproducing speech data.

Once the counts of substitution, insertion, and deletion errors are obtained, the Word Error Rate (WER) can be calculated using the formula: WER = (S + D + I) / N. The WER represents the percentage of incorrectly recognized utterances in relation to th total number of utterances, denoted by N.On the other hand, the Match Error Rate (MER) measures the error rate when the system is optimized to minimize errors. It represents the percentage of correctly recognized utterances in the recognized speech, considering the reference speech. MER can be calculated as MER = (N - S - D - I) / N, where N, S, D, and I correspond to the total number of utterances, substitutions, deletions, and insertions, respectively.The WER and MER values generated in this experiment are presented in Figure 10. Notably, Phase VI, which involved the use of BiLSTM with attention mechanism, yielded the lowest error rates for both WER and MER metrics.

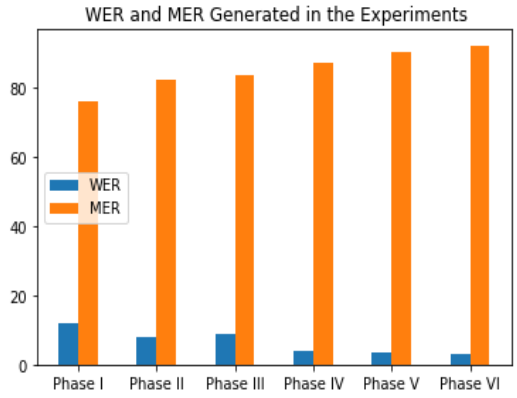

Figure 6 Performance Evaluation Using WER and MER

3.4Conclusion

The authors propose a novel methodology for enhancing Automatic Accent Speech Recognition

(AASR) models specifically designed for the Malayalam language. This involves exploring various combinations of spectral features and architectural frameworks. The experimental findings demonstrate that the utilization of a BiLSTM architecture with an attention block yields lower Word Error Rate (WER) and higher Match Error Rate (MER) compared to other approaches. Consequently, the study concludes that employing an attention block with BiLSTM architecture, along with appropriate feature vectors, is optimal for modeling accented speech in low-resourced languages.The novelty of this research also lies in the extraction of accented speech features, which contributes to the improved construction of the accented model. Furthermore, the constructed model exhibits satisfactory performance when tested with unknown accents. The dataset used in this study encompasses representations from both male and female voices across different age groups. Consequently, the feature vectors employed in this research adequately capture the variations in prosodic values associated with gender and age, further enhancing the model's effectiveness.

Constructing Automatic Accent Speech Recognition (AASR) models for Malayalam presents a challenge due to the diverse range of accents that exist within the language. However, the lack of a benchmark dataset for conducting research in this area further compounds the difficulty and creates a significant research gap. In light of this, the authors aim to address this issue by initiating the construction of an accented dataset, which will be made publicly available. This dataset will serve as a valuable resource for conducting various studies related to AASR.

Looking ahead, the authors plan to propose improved approaches for developing unified accented models capable of recognizing all accents within the Malayalam language. These approaches can subsequently be adapted for modeling accent recognition in other low-resourced languages as well. By advancing the understanding and development of AASR techniques, this research has the potential to contribute significantly to the field and address the challenges posed by diverse accents in various languages.

References

1. H. Bao, T. Wu, and H. Meng, "Improving accented speech recognition using multitask learning with attention mechanism," IEEE Signal Processing Letters, vol. 27, 2020, pp. 1084-1088.
2. Wu, Z., Li, Z., Meng, H., Xie, L., & Yu, K. (2021). Accent Robust Acoustic Modeling for Automatic Speech Recognition: A Survey. arXiv preprint arXiv:2105.03247.
3. Fernández-Gavilanes, L., Patel, R., Ghoraani, B., & Zhi, S. (2021). Unsupervised Accent Classification for Accented Speech Recognition. arXiv preprint arXiv:2104.14439.
4. Dai, W., Qian, Y., Zhang, D., Xu, X., & Yu, K. (2020). Exploring Transfer Learning for Accented Speech Recognition. arXiv preprint arXiv:2012.06662.
5. Ma, X., Qian, Y., Xiao, K., Dai, W., & Yu, K. (2020). Multilingual and Multi-Accent End-to-End Speech Recognition with Transformer. arXiv preprint arXiv:2012.15456.
6. Baevski, A., & Auli, M. (2019). VirelBERT: A Self-Supervised Visual Reasoning Language Model. arXiv preprint arXiv:1908.02265.
7. Jansen, A., & Scharenborg, O. (2020). Self-Supervised Pretraining for Speech Recognition on Low-Resource Languages. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7474-7478). IEEE.
8. Leng, Y., Zetterholm, E., & Sahidullah, M. (2021). Self-supervised learning for accented speech recognition using contextualized representations. In Proceedings of the 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (pp. 748-755). IEEE.
9. Chaudhry, A., & Rezaei, M. A. (2020). Improving Accent Classification with Self-Supervised Learning and Transfer Learning. In Proceedings of the 2020 IEEE Spoken Language Technology Workshop (SLT) (pp. 1066-1072). IEEE.
10. Bansal, A., Huang, G. B., & Ramakrishnan, S. (2020). Self-Supervised Representation Learning for Accented Speech Recognition. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7824-7828). IEEE.
11. Muneer, V.K., Mohamed Basheer, K.P. (2023). A Collaborative Destination Recommender Model in Dravidian Language by Social Media Analysis. In: Khanna, A., Polkowski, Z., Castillo, O. (eds) Proceedings of Data Analytics and Management . Lecture Notes in Networks and Systems, vol 572. Springer, Singapore. https://doi.org/10.1007/978-981-19-7615-5_45