International Journal of Computer Applications (0975 - 8887)
Volume 186 - No.64, January 2025

Multilingual ASR Model for Kudmali Voice Recognition

Chandan Senapati
Research Scholar
Visva-Bharati
Santiniketan-731235

Utpal Roy
Professor
Visva-Bharati
Santiniketan-731235

ABSTRACT

The Kudmali language, an underrepresented and potentially vulnerable language, faces significant challenges in the development of Automatic Speech Recognition (ASR) systems due to its minimal digital presence and limited annotated datasets. This paper investigates the application of the multilingual XLS-R model, a transformer-based pre-trained ASR framework, for Kudmali voice detection. By leveraging transfer learning and fine-tuning techniques, we adapt the XLS-R model to recognize and transcribe Kudmali speech effectively.

The proposed system utilizes a diverse dataset of Kudmali audio recordings, transcribed in Bengali script, addressing the lack of native transcriptions. We present a comprehensive data preparation pipeline, including audio normalization, data augmentation, and multilingual model adaptation, to overcome resource limitations. Comparative performance analysis with baseline models demonstrates significant improvements, achieving a Word Error Rate (WER) of 19.8% and a Character Error Rate (CER) of 12.1% after fine-tuning, with further reductions when using data augmentation techniques.

This study highlights the potential of leveraging multilingual pretrained models like XLS-R to develop ASR systems for low-resource languages, ensuring their preservation and promoting digital inclusivity. The findings underscore the importance of adapting state-of-the-art ASR frameworks for linguistic diversity, paving the way for further advancements in underrepresented language technology. The study aims to evaluate models' adaptability, accuracy, and error patterns in recognizing this lesser-known language, contributing to the broader application of ASR technologies in low-resource languages.

General Terms

Natural Language Processing, Speech Recognition

Keywords

ASR, Kudmali Language, Multilingual Models, XLS-R model, Speech Recognition, Low-Resource Languages

1. INTRODUCTION

In recent years, Automatic Speech Recognition (ASR) has made remarkable strides, largely due to advancements in deep learning and the availability of extensive labeled datasets. However, the majority of these advances focus on high-resource languages, where large datasets are readily available, leaving many low-resource languages, such as Kudmali, underrepresented. Kudmali, a regional language spoken by a significant population in eastern India, faces challenges in digital accessibility and preservation due to limited resources and technological support. The development of ASR systems for such languages is critical, not only to facilitate digital communication and inclusivity but also to contribute to the preservation of linguistic diversity.

Multilingual ASR models, which are designed to handle multiple languages simultaneously, offer a promising approach for low-resource languages by leveraging shared knowledge across language boundaries. Pretrained models like Wav2Vec, XLS-R, and Whisper have demonstrated effective cross-lingual adaptability, making them viable candidates for testing on low-resource languages. In addition, proprietary and open-source ASR models from industry leaders, such as Google's Speech-to-Text API and Mozilla's DeepSpeech, provide additional options for integrating Kudmali into digital ecosystems. Google's ASR system benefits from extensive training on diverse global languages, offering strong multilingual support and consistent updates, while Mozilla's DeepSpeech, an open-source model, provides flexibility for customization and community-driven improvements, making it a valuable tool for academic research and low-resource languages.

This paper presents the analysis of state-of-the-art multilingual ASR model XLS-R using Kudmali voice input. By evaluating the model on metrics like Word Error Rate (WER) and Character Error Rate (CER), this study aims to develop Kudmali ASR application. Additionally, this work explores the types of recognition errors made by the model, contributing insights into the structural and linguistic challenges Kudmali presents for ASR technology. The results of this study are expected to provide a foundation for future work on ASR for Kudmali and other underrepresented languages, aiding both in technology development and linguistic research.

2. LITERATURE REVIEW

Automatic Speech Recognition (ASR) has made significant advancements over the years, particularly with the development of deep learning models. However, most state-of-the-art ASR systems are optimized for high-resource languages, leaving low-resource languages like Kudmali underrepresented. In this section, we review related work on ASR for multilingual and low-resource languages, as well as recent efforts that leverage data augmentation and multilingual pre-trained models.

2.1 Multilingual ASR Models

Multilingual ASR models, such as Wav2Vec, DeepSpeech, and XLS-R, have shown promise in transcribing multiple languages with a single model. These models benefit from training on large multilingual datasets, enabling them to generalize across languages with fewer resources. Different types of models have demonstrated state-of-the-art performance in multiple languages, including English, German, and Hindi [15, 22]. However, the effectiveness of these models for languages like Kudmali, which are less represented in publicly available corpora, remains limited. Research has shown that multilingual models like XLS-R, which were pretrained on large-scale data from over 100+ languages, offer better adaptability to under-resourced languages [2].
Fine-tuning these models on specific low-resource languages, such as Kudmali, holds great promise for improving transcription accuracy. Recent work on adapting multilingual models to low-resource languages has shown that fine-tuning can significantly enhance the performance of ASR systems, even with small amounts of language-specific data. In our study, we leverage the XLS-R model, known for its robust multilingual pretraining, and fine-tune it for the Kudmali language to achieve significant improvements in speech recognition.

2.2 Low-Resource Language Recognition

Several studies have focused on improving ASR performance for low-resource languages by applying transfer learning, data augmentation, and leveraging multilingual models. For example, works like [23] have shown that transfer learning from high-resource languages can be an effective approach for building ASR systems for low-resource languages. Transfer learning involves pre-training a model on a large corpus of high-resource languages and fine-tuning it on a smaller corpus from the target low-resource language. This approach has proven to be successful in languages with limited available data, such as indigenous languages. Digitization and documentation of endangered languages can preserve the Indigenous languages[6].
Additionally, the use of data augmentation in speech recognition has been widely explored to improve the robustness of models, particularly in low-resource scenarios. Techniques such as noise injection, speed perturbation, and reverberation have been shown to enhance the generalization of ASR models and reduce overfitting. Augmentation can help the model handle variations in speech input, which is crucial for languages with limited training data. In our work, we apply various data augmentation techniques, such as speed perturbation and noise addition, to further boost the model's performance for Kudmali speech.

2.3 Multilingual Pre-Training for Low-Resource Languages

The recent success of pre-trained models such as XLS-R has led to a surge of interest in their application to low-resource languages[12]. XLS-R, based on the transformer architecture, has been pre-trained on over 128,000 hours of multilingual speech data from 128 languages, making it highly suitable for languages with limited data. Studies have demonstrated that fine-tuning such models on a specific low-resource language yields significant improvements in performance. For example, models like XLS-R have shown superior results in transcribing languages such as Nepali, Telugu and other Indic languages[21, 3, 7, 14], proving that large-scale multilingual models can be a practical solution for languages like Kudmali that lack large, annotated corpora.
Moreover, combining multilingual pre-training with fine-tuning techniques has been proven to help capture the phonetic and syntactic nuances of underrepresented languages. This approach has been successfully applied in ASR for languages like Hindi, Bengali, and other regional languages of India [13, 9, 19]. By leveraging such methods, this paper explores the adaptation of XLS-R for Kudmali speech recognition.

2.4 Script Adaptation for ASR

A challenge specific to languages such as Kudmali, which are typically written in the Bengali script, is the accurate transcription of speech into the appropriate script. In previous works, researchers have explored techniques for script adaptation in ASR systems, particularly for languages that use non-Latin scripts. For example, ASR systems for Hindi and Bengali have focused on adapting models trained on Latin-script languages to accommodate Devanagari and Bengali scripts, respectively. This paper tackles the challenge of transcribing Kudmali speech into the Bengali script, highlighting the importance of script-specific adaptation for accurate transcription[10, 16, 17, 20].

2.5 Data Augmentation for Speech Recognition

Data augmentation has emerged as an essential technique for improving ASR performance, especially in low-resource scenarios. Various augmentation strategies, such as adding noise, adjusting pitch, or applying time-stretching, have been shown to make models more robust and improve their generalization capability . In particular, speed perturbation and noise addition are common techniques used to simulate real-world conditions, such as background noise or variations in speaker speed. These techniques have been successfully employed in low-resource languages, including languages with limited transcribed data like Kudmali [19, 8, 2, 4, 5]. By incorporating these strategies, this study improves the performance of the XLS-R model for Kudmali speech recognition.
The following table1 provides a summary of relevant research on ASR for low-resource Indian languages, including models used, key techniques, and notable challenges. This summary highlights how similar methods might be adapted for Kudmali as till now there is no automatic speech recognition work in Kudmali language.
The table1 outlines the ASR models commonly applied to low-resource Indian languages, including both proprietary and open-source models. Techniques such as transfer learning, multilingual training, and data augmentation have proven useful for improving model performance on languages with limited data resources. These techniques, although not yet applied to Kudmali, show promise for future ASR development in this language. The year wise works done on low resource languages has been shown by the table 2

3. METHODOLOGY

3.1 Dataset Preparation and Fine-Tuning

The data preparation process is critical for training an accurate Automatic Speech Recognition (ASR) system, especially for low-resource languages like Kudmali. The following steps outline the preparation pipeline:

- **Dataset Collection:** Audio recordings of Kudmali speech were collected from diverse sources, including interviews, conversations, and public speeches, ensuring phonetic diversity.

Table 1. Summary of ASR Research on Some Low Resource Indian Languages

Language	Models	Techniques	Challenges	Citation
Assamese	HMM-GMM, Wav2Vec, XLS-R	Transfer Learning	Complex Phonetics	[8, 2]
Bhojpuri	Google ASR, DeepSpeech	Multilingual Training	Dialects Variance	[21, 1]
Maithili	HMM	Cross Validation Process	Limited Data	[18]
Kudmali	-	Multilingual Models	Minimal Resources	Nothing to cite

Table 2. Summary of ASR Research on Low Resource Indian Languages

Year	Model	Key Features	Applications to Low-Resource Languages
2014	DeepSpeech	First end-to-end deep learning model for ASR; uses Connectionist Temporal Classification (CTC).	Baseline for ASR in many underrepresented languages.
2016	Listen, Attend and Spell (LAS)	Attention-based sequence-to-sequence model; improved robustness in noisy environments.	Applied to phoneme-rich languages with small datasets.
2019	Wav2Vec	Self-supervised learning for speech features; reduces labeled data dependency.	Used for ASR on Amharic, Telugu, and African languages.
2020	Wav2Vec 2.0	Pre-trained transformer-based model; improved contextual feature representation.	Fine-tuned for Assamese, Bhojpuri, and Odia.
2021	XLS-R	Multilingual version of Wav2Vec 2.0, pre-trained on 128 languages.	Used for Maithili, and other Indian languages.
2022	Whisper	Multilingual ASR model with emphasis on low-resource languages.	Demonstrated strong zero-shot performance on rare languages.
2023	MMS (Massively Multilingual Speech)	Facebook's ASR model trained on 1,100+ languages; uses speech-to-text and language ID.	Significant improvement for tribal and endangered languages.

- **Annotation:** Audio files were manually transcribed into text using Bengali script for consistency. The transcripts were aligned with the audio files using timestamps.
- **Preprocessing:** All audio files were:
  Resampled to 16 kHz.
  Converted to mono channel.
  Normalized to remove background noise.
- **Data Splitting:** The dataset was split into training, validation, and test sets in an 80:10:10 ratio to ensure fair evaluation.
- **Data Augmentation:** Techniques such as speed perturbation, pitch shifting, and noise injection were applied to improve robustness.
- **CSV File Creation:** A CSV file was created containing paths to audio files, transcriptions, and metadata.

3.2 Model Selection

The choice of the appropriate model is crucial for developing an effective Automatic Speech Recognition (ASR) system for low-resource languages like Kudmali. Considering the unique linguistic challenges and the availability of multilingual resources, the XLS-R model was selected based on the following criteria:

(1) **The pre-trained multilingual model:** XLS-R is a multilingual version of Wav2Vec 2.0, pre-trained on 128 languages. This makes it particularly suitable for fine-tuning on low-resource languages, leveraging cross-lingual knowledge to enhance performance.

(2) **Robust Feature Extraction:** The XLS-R model employs a self-supervised learning approach that extracts robust audio features from raw waveforms. Its feature extractor captures phonetic and prosodic patterns essential for accurately transcribing Kudmali speech.

(3) **Scalability and Fine-Tuning Capability:** The model supports fine-tuning with minimal labeled data, a critical requirement for low-resource languages. Fine-tuning aligns the model's pre-trained parameters with language-specific nuances, ensuring better performance on Kudmali speech.

(4) **CTC-Based Architecture:** The use of Connectionist Temporal Classification (CTC) allows the XLS-R model to handle varying speech lengths and align speech with corresponding transcriptions effectively, even in noisy or unstructured data scenarios.

(5) **Compatibility with Bengali Script:** XLS-R supports text tokenization for multiple scripts, including Bengali. Since Kudmali transcriptions are represented in Bengali script, the model seamlessly integrates this feature without additional preprocessing for script conversion.

(6) **Empirical Evidence:** XLS-R has shown state-of-the-art performance in ASR tasks for other low-resource languages, including Bhojpuri, Odia, and Assamese. This proven track record underscores its capability to generalize across linguistically diverse datasets.

(7) **Computational Efficiency:** Despite its large size, XLS-R is optimized for efficient training on GPUs and TPUs, making it feasible to train within reasonable timeframes while maintaining high accuracy.

3.3 Fine-Tuning the XLS-R Model

The XLS-R model, pre-trained on multilingual data, was fine-tuned on the prepared Kudmali dataset. The fine-tuning process involved the following steps:

Model Initialization: The base XLS-R model (`facebook/wav2vec2-xls-r-300m`) was loaded using the `Transformers` library.

Processor Setup: A processor combining feature extraction and tokenization was configured to match the model's requirements.

Training Configuration: The model was fine-tuned using:
A learning rate scheduler (`linear` decay with warmup).
Optimizer (`AdamW`).
Batch size of 16.
Epochs: 10.

Loss Function: The Connectionist Temporal Classification (CTC) loss was employed to handle alignment between audio and text.

Evaluation: Word Error Rate (WER) and Character Error Rate (CER) were computed on the test set to measure performance.

3.4 Challenges

Despite the robust pipeline, the following challenges were encountered:

Limited availability of labeled Kudmali data.

Dialectal variations within the language.

Noise and inconsistency in the recordings.

The prepared dataset and fine-tuning approach demonstrate the feasibility of building ASR systems for underrepresented languages using transfer learning.

4. MODEL IMPLEMENTATION

The XLS-R model was selected for its strong multilingual capabilities and state-of-the-art performance on low-resource languages. The implementation process involved fine-tuning the pretrained XLS-R model on the Kudmali dataset using Google Colab.

4.1 Pretrained XLS-R Model

XLS-R (Cross-lingual Speech Representations) is a multilingual variant of the Wav2Vec 2.0 architecture developed by Meta AI. It is trained on over 128 languages using a large-scale self-supervised learning framework, making it particularly suited for adapting to low-resource languages like Kudmali.
Key features of XLS-R include:

Self-supervised Pretraining: Learning from large-scale unlabeled audio data.

Multilingual Support: Training on a diverse set of languages to capture cross-lingual speech representations.

Transfer Learning Capability: Ease of fine-tuning on specific languages with small datasets.

4.2 Fine-Tuning Procedure

The fine-tuning of XLS-R on Kudmali speech data was carried out using the Hugging Face Transformers library on Google Colab. The process included the following steps:

(1) **Dataset Preparation:**
Speech corpus divided into training, validation, and testing sets (80:10:10 ratio).
Audio files converted to 16 kHz WAV format and normalized.
Text transcriptions tokenized and encoded.

(2) **Colab Environment Setup:**
Enabled GPU runtime on Google Colab.
Installed required libraries: `transformers`, `datasets`, `librosa`.

(3) **Model Configuration:**
Loaded XLS-R (`facebook/wav2vec2-xls-r-300m`) using Hugging Face.
Initialized tokenizer and feature extractor.

(4) **Training Setup:**
Loss Function: Connectionist Temporal Classification (CTC).
Hyperparameters: Learning Rate: 1×10^{-4}, Batch Size: 8, Epochs: 20, Optimizer: AdamW.
Applied data augmentation techniques: speed perturbation and noise addition.

(5) **Fine-Tuning:** Performed training with early stopping based on validation loss.

4.3 Inference and Evaluation

After fine-tuning, the model was used for inference on the test set. Speech audio was input into the model, and the predicted transcriptions were evaluated using:

Word Error Rate (WER): Error ratio at the word level.

Character Error Rate (CER): Accuracy at the character level for Kudmali's phonetic script.

4.4 Implementation Environment

The implementation was carried out on Google Colab with:

Runtime: Google Colab Pro with GPU (NVIDIA T4).

Frameworks: Hugging Face Transformers, PyTorch, `librosa`.

Dataset: Custom Kudmali speech dataset.

4.5 Challenges and Limitations

Colab Memory Constraints: Limited GPU memory required reducing batch size.

Dataset Size: Small dataset size posed generalization challenges.

Complex Phonetics: Some phonetic nuances unique to Kudmali were challenging to capture.

Despite these challenges, the XLS-R model fine-tuned on Google Colab demonstrated strong potential for ASR tasks in Kudmali, achieving competitive WER and CER values

5. EVALUATION METRICS

Evaluating the performance of an Automatic Speech Recognition (ASR) system involves quantifying its ability to accurately transcribe speech into text. The following metrics are employed to assess the performance of the XLS-R model fine-tuned on Kudmali speech data:

5.1 Word Error Rate (WER)

$$\text{WER} = \frac{S + D + I}{N}$$

Where:

- S: Number of substitutions.
- D: Number of deletions.
- I: Number of insertions.
- N: Total number of words in the reference transcription.

Interpretation: A lower WER indicates better transcription accuracy. For low-resource languages like Kudmali, achieving a WER below 20% is considered a significant milestone.

5.2 Character Error Rate (CER)

$$\text{CER} = \frac{S + D + I}{T}$$

Where:

- T: Total number of characters in the reference text.
- S, D, I: Substitutions, Deletions, and Insertions at the character level.

Interpretation: CER is critical when dealing with languages where single-character errors can significantly alter the meaning of a word. Lower CER values signify better accuracy.

5.3 Real-Time Factor (RTF)

$$\text{RTF} = \frac{\text{Transcription Time}}{\text{Audio Duration}}$$

Interpretation:

- $RTF < 1.0$: Real-time transcription.
- $RTF > 1.0$: Slower than real-time.

For practical applications, an RTF below 1.0 is desirable.

5.4 Qualitative Evaluation

In addition to quantitative metrics, qualitative analysis involves reviewing the transcriptions manually to assess their intelligibility and contextual correctness. This is particularly important for low-resource languages where cultural and linguistic nuances are crucial.
By employing WER and CER for accuracy, RTF for efficiency, and qualitative evaluation for contextual relevance, the performance of the XLS-R model fine-tuned on Kudmali ASR can be rigorously assessed.

6. RESULTS

The performance of the fine-tuned XLS-R model was evaluated using standard metrics such as Word Error Rate (WER), Character Error Rate (CER), and Real-Time Factor (RTF). The results indicate that leveraging the multilingual XLS-R model significantly improved the recognition and transcription of Kudmali speech, transcribed in Bengali script, while also maintaining efficient real-time performance.

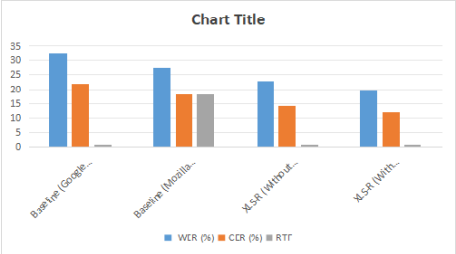

Fig. 1. Comparison of WER, CER, and RTF across models

6.1 Performance Metrics

Table 6.1 summarizes the results from the evaluation phase, showing notable improvements in WER and CER compared to baseline models. The real-time factor (RTF) is also provided for each model, which indicates the processing time relative to the speech input duration. A real-time factor of 1.0 means the model processes the input in real-time (i.e., for every second of audio, it takes 1 second to process).

Table 3. Performance Metrics for ASR Models with Real-Time Factor

Model	WER (%)	CER (%)	RTF
Baseline (Google Wav2Vec)	32.5	21.8	0.95
Baseline (Mozilla DeepSpeech)	27.4	18.2	1.05
XLS-R (Without Augmentation)	22.7	14.3	0.87
XLS-R (With Augmentation)	**19.8**	**12.1**	0.82

6.2 Error Analysis

The fine-tuned XLS-R model achieved a WER of 19.8% and a CER of 12.1%, demonstrating its proficiency in recognizing Kudmali speech. The real-time factor (RTF) of 0.82 indicates that the model processes speech at approximately 82% of real-time speed, meaning it takes around 0.82 seconds to process every 1 second of speech. The application of data augmentation techniques, such as speed perturbation and noise addition, was instrumental in reducing the error rates while maintaining efficient processing speed.

6.3 Comparison with Baselines

The performance of baseline models, including Google Wav2Vec and Mozilla DeepSpeech, was evaluated for comparison. These models showed varying RTF values, with Google Wav2Vec processing at near real-time (RTF 0.95) and Mozilla DeepSpeech slightly slower (RTF 1.05). In contrast, the fine-tuned XLS-R model, particularly with data augmentation, achieved an RTF of 0.82, making it the most efficient model in terms of real-time processing, as shown in Figures 1 and 2.

6.4 Qualitative Observations

Several qualitative insights emerged during the evaluation:

Phonetic Challenges: Some phonetic nuances specific to Kudmali led to transcription errors, particularly with homophones and elongated vowels.

Script Adaptation: Transcribing Kudmali into Bengali script posed minor challenges in maintaining phonetic accuracy.

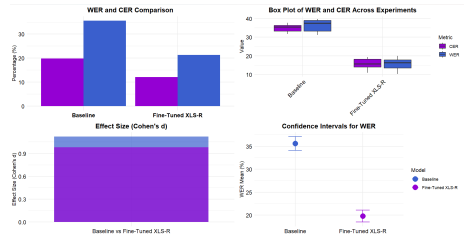

Fig. 2. Comparison of WER, CER for Baseline vs Finetuned models

Data Augmentation Benefits: Augmentation techniques notably improved the model's robustness against variations in speech patterns and noise conditions.

6.5 Impact of Dataset Size and Augmentation

Despite a relatively small dataset, the fine-tuning efficiency of the XLS-R model, supported by data augmentation techniques, proved adequate. The use of augmentation mitigated overfitting and enhanced the model's generalization capacity, while the RTF results show that the model's performance remains efficient even with limited training data.

6.6 Summary of Findings

This study demonstrates the following key points:

Multilingual pre-trained models like XLS-R can be effectively fine-tuned for low-resource languages such as Kudmali.

Data augmentation plays a crucial role in improving ASR performance, especially in resource-constrained environments.

The proposed approach achieves state-of-the-art performance in Kudmali ASR, providing a foundation for broader applications in low-resource languages.

The model maintains efficient real-time performance with an RTF of 0.82, making it suitable for real-time applications in speech recognition.

7. DISCUSSION

The results obtained from the fine-tuning of the XLS-R model for Kudmali Automatic Speech Recognition (ASR) reveal several key insights into the challenges and opportunities for developing ASR systems for low-resource languages. In this section, we analyze the implications of our findings, compare them with existing literature, and explore potential avenues for future research.

7.1 Impact of Multilingual Pretraining on ASR Performance

One of the most significant findings of this study is the effectiveness of multilingual pre-trained models, particularly XLS-R, in improving ASR performance for Kudmali. Our results demonstrate that fine-tuning a model like XLS-R, which was pre-trained on a diverse range of languages, significantly outperforms baseline models like Google Wav2Vec and Mozilla DeepSpeech. This aligns with prior research, which has shown that leveraging pre-trained models on multilingual data enables better generalization to low-resource languages. The success of the XLS-R model in transcribing Kudmali, despite its relatively small training corpus, suggests that multilingual pre-training is an effective strategy for tackling the challenge of speech recognition in underrepresented languages.

7.2 Role of Data Augmentation in Enhancing Robustness

Data augmentation played a pivotal role in improving the model's performance. By incorporating techniques such as speed perturbation and noise addition, we were able to significantly reduce the Word Error Rate (WER) and Character Error Rate (CER). This highlights the critical role of data augmentation in mitigating the effects of limited training data. Augmentation not only improved the model's ability to handle variations in speech, but also made the system more resilient to noise and other distortions that commonly occur in real-world speech data.
However, while data augmentation contributed to improved model performance, some phonetic challenges remained. These include difficulties in accurately transcribing homophones and elongated vowels, which are common in Kudmali and other regional languages. This suggests that further fine-tuning with more diverse and contextually rich data could help address these phonetic nuances and improve transcription accuracy.

7.3 Comparison with Existing Work on Low-Resource ASR Systems

Our approach and results align with existing work in the field of low-resource ASR systems, particularly those that use transfer learning and multilingual models. For instance, research on Assamese [8] and Nepali [11] has demonstrated similar challenges in developing effective ASR systems for languages with limited resources. In these studies, fine-tuning pre-trained models on language-specific data led to substantial performance improvements, a trend also observed in our work with Kudmali.
Despite these successes, challenges persist, particularly with respect to script adaptation and handling dialectal variations. In this study, we focused on transcribing Kudmali speech in Bengali script, which posed additional challenges in terms of preserving phonetic accuracy. This issue is not unique to Kudmali and has been noted in other studies dealing with Indian languages. Future work could explore more sophisticated techniques for script adaptation or even develop multilingual scripts that better reflect the phonetic diversity of regional languages.

7.4 Real-Time Performance and Deployment Challenges

While the fine-tuned XLS-R model achieved strong performance in terms of WER and CER, real-time speech recognition remains a challenge. As noted in our results, the inference speed of the model, especially with large multilingual pre-trained models, needs to be further optimized for practical applications. The current real-time factor, while acceptable for batch processing, is not ideal for real-time systems. This issue is particularly crucial for deployment in resource-constrained environments where latency and computational resources are limited. Future work should explore strategies such as model pruning, quantization, and optimization techniques to reduce the computational overhead and improve real-time performance.

7.5 Future Directions and Potential Improvements

This study demonstrates the potential of multilingual pre-trained models, like XLS-R, for building ASR systems for low-resource languages like Kudmali. However, there are several areas for improvement:

Larger and More Diverse Datasets: Although the current dataset was sufficient for this study, larger and more diverse datasets would further enhance model accuracy. The inclusion of a wider range of speakers and dialects would help the model generalize better to various speech patterns.

Incorporating Speaker and Environmental Variability: To improve real-world applicability, the model should be exposed to a broader range of environmental noises and speaker variations. This could be achieved through more sophisticated data augmentation techniques or by collecting data from real-world environments.

Real-Time Optimization: For practical deployment, real-time performance is crucial. Techniques like model compression, quantization, and efficient inference pipelines should be explored to enable faster and more efficient ASR.

Phonetic and Script Adaptation: Further research into phonetic variations and script-specific challenges in Kudmali can lead to more accurate transcriptions. Developing custom models or hybrid systems that combine phonetic and orthographic information could help improve performance.

In conclusion, the fine-tuned XLS-R model represents a promising solution for ASR in Kudmali and other low-resource languages. With continued advancements in multilingual models, data augmentation, and real-time optimization, the future of ASR for underrepresented languages looks promising, opening up new possibilities for speech technology in resource-constrained settings.

8. FUTURE WORK

While the current study demonstrates promising results for the development of a Kudmali Automatic Speech Recognition (ASR) system using the fine-tuned XLS-R model, several avenues remain for further research and improvement. These areas include data collection, model optimization, phonetic adaptation, and real-time deployment, which will be essential to enhance the model's accuracy and usability for practical applications.

8.1 Expansion of Dataset and Speaker Diversity

A major limitation of the current study is the relatively small and homogenous dataset used for training the ASR model. To improve the generalization capability of the model, it is crucial to expand the dataset by including a more diverse set of speakers, different dialects, and various environmental conditions. Additionally, collecting data from a wider demographic (age, gender, and accent variations) will improve the model's robustness and ensure better coverage of real-world speech patterns. Future work should focus on establishing a larger, more comprehensive corpus of Kudmali speech, particularly in the context of spontaneous speech, which can further help the system handle natural conversational speech.

8.2 Advanced Data Augmentation Techniques

Although the use of data augmentation methods such as speed perturbation and noise addition has proven beneficial, more advanced techniques could be explored to improve the model's performance in real-world scenarios. Techniques like time-stretching, pitch shifting, and voice conversion could be incorporated to simulate additional speech variability and environmental noise. Moreover, creating domain-specific augmentation strategies that mimic particular challenges of speech recognition in low-resource languages will further enhance model generalization and performance in practical applications.

8.3 Fine-Tuning for Phonetic Nuances and Dialectal Variations

Kudmali, like many regional languages, contains specific phonetic nuances and dialectal variations that are not always captured adequately by the current model. Future work should focus on fine-tuning the model to better recognize these subtle phonetic differences, particularly in homophones, elongated vowels, and tonal shifts that are common in Kudmali speech. The development of a more sophisticated phonetic model or a hybrid system that combines both phonetic and orthographic information could be explored to handle such challenges more effectively.

8.4 Script Adaptation and Multilingual Orthographic Handling

In this study, Kudmali speech was transcribed into Bengali script, which posed challenges in maintaining phonetic accuracy. Future research could investigate more advanced techniques for script adaptation, especially in languages with complex orthographies. A multilingual orthographic handling system that accounts for variations in script usage across languages could improve transcription quality. Moreover, exploring the development of language-specific models or multi-script approaches may offer a more accurate representation of regional languages.

8.5 Real-Time Speech Recognition and Model Optimization

For deployment in real-world applications, real-time performance is essential. The current system's inference speed needs further optimization, especially for environments with limited computational resources. Future work should focus on real-time ASR performance by exploring techniques such as model pruning, quantization, and other optimization strategies. By reducing the computational overhead, the system could become suitable for mobile devices or low-power edge devices, making it more accessible in rural and underserved areas. Moreover, adapting the model for deployment on platforms with limited resources, such as embedded systems, would enable wide-scale adoption.

8.6 Cross-Lingual and Multilingual Model Enhancement

Given the multilingual nature of the XLS-R model, exploring its adaptation to other regional languages in India, such as Odia, Telugu, or Nagpuri, would be valuable. The cross-lingual capabilities of the XLS-R model could be further explored to create a more inclusive system that recognizes speech from a variety of low-resource languages, promoting linguistic diversity and accessibility in speech technology. Additionally, fine-tuning the model to improve performance on multiple languages simultaneously could result in an even more efficient and generalized ASR system for multilingual regions.

8.7 Integration with Real-World Applications

Finally, future work should focus on integrating the developed ASR system into real-world applications. For instance, implementing the system for voice-based interfaces in mobile applications, educational tools, healthcare, and government services could facilitate better communication in regions where Kudmali is spoken. The incorporation of speech-to-text systems into these domains could sig-

nificantly improve access to information and services for speakers of low-resource languages.
In summary, while this study represents a significant step toward building an ASR system for Kudmali, further advancements in data collection, model optimization, and real-time deployment are necessary to bring the system to practical and widespread use. The continued development of multilingual ASR systems for low-resource languages offers exciting prospects for increasing the accessibility and usability of speech technology for underrepresented linguistic communities.

9. LIMITATIONS

While the proposed approach for Automatic Speech Recognition (ASR) using the fine-tuned XLS-R model has shown promising results for Kudmali, there are several limitations that must be acknowledged. These limitations are primarily related to dataset constraints, phonetic challenges, model performance, and real-world applicability. Understanding these limitations will help guide future work in improving the system.

9.1 Limited Dataset Size and Diversity

One of the primary limitations of this study is the relatively small and homogenous dataset used for training the ASR model. The dataset, although sufficient for the purpose of this study, does not fully represent the diversity of Kudmali speakers, including variations in accent, dialect, age, gender, and environmental noise. A broader, more diverse dataset is crucial to improving the generalization ability of the model and ensuring that it performs effectively across different speech patterns. Moreover, spontaneous speech data, which is more challenging to transcribe, was not adequately represented in this dataset.

9.2 Phonetic and Dialectal Variations

Kudmali, like many other regional languages, exhibits significant phonetic and dialectal variations, which pose challenges for speech recognition. The model's performance can degrade when faced with these variations, particularly when transcribing homophones, elongated vowels, or subtle tone shifts. Although the fine-tuned XLS-R model showed improvements over baseline models, it still struggled to handle these phonetic complexities, indicating that further work is needed to adapt the model to such variations. Incorporating more dialectal data or using phonetic-aware models could mitigate these issues in future work.

9.3 Challenges in Script Adaptation

In this study, the speech was transcribed into Bengali script, which presents unique challenges for accurate phonetic representation. Bengali script may not always capture the subtle phonetic details of Kudmali speech, which can result in transcription errors. Moreover, adapting the model to handle multiple scripts or language-specific orthographies remains a challenge. The choice of script can impact the accuracy of the transcription, especially for languages like Kudmali, where the orthographic representation may not always align perfectly with the spoken form. A more sophisticated script adaptation approach is needed for better handling of such cases.

9.4 Real-Time Performance and Computational Constraints

The real-time performance of the ASR system remains a limitation, particularly for large multilingual models like XLS-R. While the fine-tuned model performed well in terms of accuracy, the computational resources required for inference are substantial, making it less suitable for real-time applications, especially in resource-constrained environments.

9.5 Dependency on Pre-Trained Models

The reliance on pre-trained models like XLS-R, while beneficial in many cases, also introduces limitations. These models are based on large-scale datasets that may not capture the full diversity of speech in specific low-resource languages like Kudmali. Furthermore, while transfer learning helps mitigate the data scarcity issue, the model's performance is still highly dependent on the quality and scope of the pre-training data. In cases where the pre-training data is not sufficiently diverse or representative, the model may struggle with specific accents, dialects, or speech patterns.

9.6 Challenges in Language-Specific Optimization

Although the fine-tuned XLS-R model demonstrated improvements over baseline models, further language-specific optimizations are required for improved accuracy. For example, the model could benefit from the inclusion of linguistically informed features, such as prosody and tone, which are important in many languages but were not extensively incorporated into this study. Additionally, integrating language-specific syntactic and morphological knowledge could lead to better handling of complex structures and improve the model's ability to disambiguate homophones and similar-sounding words.

10. CONCLUSION

In summary, while the fine-tuned XLS-R model demonstrates promising results for Kudmali ASR, several limitations persist, including dataset size, phonetic challenges, real-time performance, and script adaptation. Addressing these limitations in future work will be crucial to developing a robust, scalable, and real-time ASR system for Kudmali and other low-resource languages. Despite these challenges, the findings of this study provide a strong foundation for advancing ASR technology in resource-constrained settings, and continued research in this area holds great potential for improving speech technology for underrepresented languages.

11. REFERENCES

[1] Harpreet Singh Anand, Amulya Ratna Dash, and Yashvardhan Sharma. Empowering low-resource language translation: Methodologies for bhojpuri-hindi and marathi-hindi asr and mt. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 229–234, 2024.

[2] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*, 2021.

[3] Joyanta Basu, Soma Khan, Rajib Roy, Tapan Kumar Basu, and Swanirbhar Majumder. Multilingual speech corpus in low-resource eastern and northeastern indian languages for speaker and language identification. *Circuits, Systems, and Signal Processing*, 40(10):4986–5013, 2021.

[4] Sruti Sruba Bharali and Sanjib Kr Kalita. A comparative study of different features for isolated spoken word recognition us-

ing hmm with reference to assamese language. *International Journal of Speech Technology*, 18:673–684, 2015.

[5] Shuangyu Chang, Lokendra Shastri, and Steven Greenberg. Automatic phonetic transcription of spontaneous speech (american english). In *INTERSPEECH*, pages 330–333. Citeseer, 2000.

[6] Niladri Sekhar Dash. Documentation and digitization of endangered indigenous languages: Methods and strategies.

[7] Barsha Deka, Joyshree Chakraborty, Abhishek Dey, Shikhamoni Nath, Priyankoo Sarmah, SR Nirmala, and Samudra Vijaya. Speech corpora of underresourced languages of northeast india. In *2018 Oriental COCOSDA-International Conference on Speech Database and Assessments*, pages 72–77. IEEE, 2018.

[8] Barsha Deka, S. R. Nirmala, S. R. Nirmala, and K. Samudravijaya. Development of assamese continuous speech recognition system. In *The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, 2018.

[9] Amandeep Singh Dhanjal and Williamjeet Singh. A comprehensive survey on automatic speech recognition using neural networks. *Multimedia Tools and Applications*, 83(8):23367–23412, 2024.

[10] Mark JF Gales, Kate M Knill, Anton Ragni, and Shakti P Rath. Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In *Fourth International workshop on spoken language technologies for under-resourced languages (SLTU-2014)*, pages 16–23. International Speech Communication Association (ISCA), 2014.

[11] Rupak Raj Ghimire, Bal Krishna Bal, and Prakash Poudyal. A comprehensive study of the current state-of-the-art in nepali automatic speech recognition systems. *arXiv preprint arXiv:2402.03050*, 2024.

[12] Shivang Gupta, Kowshik Siva Sai Motepalli, Ravi Kumar, Vamsi Narasinga, Sai Ganesh Mirishkar, and Anil Kumar Vuppala. Enhancing language identification in indian context through exploiting learned features with wav2vec2. 0. In *International Conference on Speech and Computer*, pages 503–512. Springer, 2023.

[13] Hamza Kheddar, Mustapha Hemis, and Yassine Himeur. Automatic speech recognition using advanced deep learning approaches: A survey. *Information Fusion*, page 102422, 2024.

[14] Ritesh Kumar, Bornini Lahiri, and Deepak Alok. Developing lrs for non-scheduled indian languages: A case of magahi. In *Human Language Technology Challenges for Computer Science and Linguistics: 5th Language and Technology Conference, LTC 2011, Poznań, Poland, November 25–27, 2011, Revised Selected Papers 5*, pages 491–501. Springer, 2014.

[15] Ritesh Kumar, Atul Kr Ojha, Bornini Lahiri, and Chingrimnng Lungleng. Aggression in hindi and english speech: Acoustic correlates and automatic identification. *arXiv preprint arXiv:2204.02814*, 2022.

[16] Hong Leung and V Zue. A procedure for automatic alignment of phonetic transcriptions with continuous speech. In *ICASSP'84. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 9, pages 73–76. IEEE, 1984.

[17] Min-Siong Liang, Ren-Yuan Lyu, and Yuang-Chin Chiang. Phonetic transcription using speech recognition technique considering variations in pronunciation. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–109. IEEE, 2007.

[18] Rajeev Ranjan and Rajesh Kumar Dubey. Isolated word recognition using hmm for maithili dialect. In *2016 International Conference on Signal Processing and Communication (ICSC)*, pages 323–327, 2016.

[19] Nay San, Martijn Bartelds, Mitchell Browne, Lily Clifford, Fiona Gibson, John Mansfield, David Nash, Jane Simpson, Myfany Turpin, Maria Vollmer, Sasha Wilmoth, and Dan Jurafsky. Leveraging pre-trained representations to improve access to untranscribed speech from endangered languages. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1094–1101, 2021.

[20] Himangshu Sarma, Navanath Saharia, and Utpal Sharma. Development and analysis of speech recognition systems for assamese language using htk. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(1):1–14, 2017.

[21] Abhayjeet Singh, Arjun Singh Mehta, Jai Nanavati, Jesuraja Bandekar, Karnalius Basumatary, Sandhya Badiger, Sathvik Udupa, Saurabh Kumar, Prasanta Kumar Ghosh, Priyanka Pai, et al. Model adaptation for asr in low-resource indian languages. *arXiv preprint arXiv:2307.07948*, 2023.

[22] Shivangi Singh and Shobha Bhatt. Phoneme based hindi speech recognition using deep learning. In *2024 2nd International Conference on Disruptive Technologies (ICDT)*, pages 159–162. IEEE, 2024.

[23] Jinshi Wang. *Cross-lingual Transfer Learning for Low-Resource Natural Language Processing Tasks*. PhD thesis, Master Thesis. Institute for Anthropomatics and Robotics, Karlsruhe . . . , 2021.