10.48047/jocaaa.2024.33.08.131

Advancing Speech Recognition in Tulu: A Research Perspective

Praveen N[1,2*], Shashidhar Kini[3]

[1]Department of Computer Applications, Nitte Institute of Professional Education, Nitte (Deemed to be University), Mangalore, 575007, Karnataka, India.
[2]Research Scholar, Department of MCA, Shrinivas Institute of Technology, VTU-Belagavi, Mangalore, 574143, Karnataka, India.
[3]Department of MCA, Shrinivas Institute of Technology, VTU-Belagavi, Mangalore, 574143, Karnataka, India.
*Corresponding E-mail: praveen.n@nitte.edu.in

Abstract

This research paper addresses the critical gap in Automatic Speech Recognition (ASR) technology for Tulu, a Dravidian language spoken by over two million people in Southern India but underrepresented in digital linguistic resources. The study proposes the development of a robust ASR system tailored to Tulu, leveraging advanced machine learning techniques such as deep learning frameworks (DNN-HMM, hybrid models), transfer learning, and reinforcement learning. Key challenges include data scarcity, phonetic diversity, code-switching with Kannada and English, and noise resilience. The research focuses on creating a comprehensive Tulu speech corpus through community-driven efforts, enhancing noise robustness, and adapting multilingual models to handle code-mixing. By integrating Tulu's newly developed Unicode standard, the system ensures accurate digital representation and transcription. The paper reviews existing literature, identifies limitations in current approaches, and outlines contributions such as expanded datasets, improved noise resilience, and cross-lingual adaptation strategies. The proposed system aims to bridge the technological divide for Tulu speakers, promoting digital inclusion and accessibility while preserving the language's linguistic nuances. Future directions include dialect recognition, hybrid model optimization, and larger corpus development to further refine ASR performance for Tulu.

Keywords: Kannada, ASR, Deep learning, Phonetic Representations, Speech Recognition

1 Introduction:

Tulu, a Dravidian language spoken by over 2 million people in Southern India, has largely remained underrepresented in the field of automatic speech recognition (ASR). The language faces unique challenges, including the absence of standardized resources, complex phonetic structures, and its code-mixed nature with languages like Kannada and English. This paper proposes the development of an ASR system for Tulu, leveraging state-of-the-art machine learning techniques such as deep learning frameworks, including DNN-HMM and hybrid models, to address these challenges. Building on recent advances in ASR for low-resource languages, this work focuses on corpus development, phoneme modelling, and multilingual training to improve recognition accuracy. Additionally, the research aims to create a robust dataset through community-driven contributions, further enhancing the system's adaptability to real-world use. By utilizing advanced methodologies such as transfer learning and reinforcement learning, this study seeks to provide an adaptive and noise-resilient ASR solution for Tulu speakers. Through these innovations, the proposed system holds the potential to bridge the technological gap and promote digital inclusion for Tulu speakers, enabling broader access to speech-driven applications and services.

2 Key Research Contributions:

Speech recognition technology has transformed communication, education, and accessibility across the globe. However, languages like Tulu, with minimal linguistic resources and technological focus, lag in benefiting from these advancements. With its unique phonetics and code-mixed usage alongside Kannada and English, Tulu presents a compelling case for ASR research. Researchers have undertaken diverse projects to address Tulu's linguistic challenges. These efforts include corpus development, phoneme modelling, and the application of cutting-edge machine learning algorithms. This article synthesizes key research insights to provide a comprehensive understanding of the field.

Paper	Content	Limitations
Development and validation of Tulu sentence lists to test speech	Focused on assessing ASR performance under noisy conditions	The noisy speech data was not extensive enough to

10.48047/jocaaa.2024.33.08.131

recognition threshold in noise (S. Bhat et al. (2021))[1]	for Tulu, exploring phoneme adaptability.	comprehensively address noise-resilient ASR systems for Tulu.
H. Asha and S. H. Lakshmaiah, "Kannada-Tulu Parallel Corpus Construction for Neural Machine Translation (Asha and Lakshmaiah) (2023)[2]	Developed a Kannada-Tulu parallel corpus for machine translation, supporting sub-word tokenization for phonetic recognition tasks.	The parallel corpus might not fully capture the complexities of Tulu's speech patterns and code-mixed scenarios.
Automatic Speech Recognition for Tulu Language Using Gmm-Hmm and DNN-HMM Techniques (G. Amoolya et al. (2022))[3]	Developed an ASR system for Tulu using GMM-HMM and DNN-HMM models, with a focus on improving accuracy using deep learning techniques and a 7-hour native speech dataset.	The dataset size was limited (7 hours), which may not generalize well to diverse Tulu speech and environments.
Survey of multilingual models for ASR in under-resourced languages (Yadav & Sitaram (2022)) [4]	This paper surveys multilingual models for ASR, focusing on low-resource languages. The authors highlight the challenges of multilingual model adaptation to various under-resourced languages, including phonetic diversity.	Limited focus on specific under-resourced languages like Tulu and direct application to real-world conditions.
A Code-Diverse Tulu-English Dataset for NLP-Based Sentiment Analysis Applications (P. Kannadaguli, 2021, IEEE)[5]	Focuses on developing a Tulu-English code-mixed dataset for sentiment analysis, addressing the linguistic gap in resources for underrepresented languages like Tulu. Highlights the importance of handling code-switching in NLP applications. Presented at the Advanced Communication Technologies and Signal Processing Conference, ensuring peer review and relevance.	Limited metadata; no details on dataset size, preprocessing methods, or sentiment annotation techniques. No explicit information on the scalability or generalizability of the dataset. Unclear if the dataset accounts for diverse dialects of Tulu or informal text variations. Lack of publicly available implementation or detailed experimental results for replication.
A Comparative Study of Deep Neural Network Based Punjabi-ASR System (Virender Kadyan et al. (2019))[6]	The paper presents a comparative analysis of deep neural network (DNN) based Automatic Speech Recognition (ASR) systems for the Punjabi language. It evaluates various ASR models, comparing their performance on standard datasets, and explores the potential of DNN architectures for improving ASR accuracy in Punjabi.	The paper may have limited applicability to other languages, given its focus on Punjabi ASR. It does not discuss the implementation details or computational complexity of the DNN models used, which could hinder replicability. Additionally, the study may not consider the influence of different accents or dialects in Punjabi speech recognition.
A Study on the Challenges and Opportunities of Speech Recognition for Bengali Language (M. F. Mridha et al. (2022)[7]	The paper explores the challenges and opportunities of implementing speech recognition for Bengali, one of the widely spoken languages in South Asia. It provides a detailed analysis of the difficulties faced in developing ASR systems for Bengali, such as acoustic variation, dialectal differences, and limited resources. It also highlights the potential areas for improvement and development in speech recognition for the Bengali language, providing an overview of existing technologies and suggesting future directions. Published in Volume	The study may not fully address the computational aspects and the specific techniques used to overcome the challenges. It also lacks detailed performance evaluation or comparison with existing systems. The focus is on general challenges, and there may be limited focus on specific advancements or real-world implementation results.

10.48047/jocaaa.2024.33.08.131

	55, Issue 4 of Artificial Intelligence Review (Pages 3431-3455).	
A Tulu Resource for Machine Translation (Manu Narayanan, Noëmi Aepli(2024))[8]	This paper introduces a Tulu language resource aimed at improving machine translation (MT) systems, focusing on addressing the challenges of low-resource languages. The authors provide insights into building a Tulu corpus, annotation methods, and how this resource can contribute to better translation between Tulu and other languages. It presents the importance of language resources in enhancing MT systems for underrepresented languages.	The paper does not go into extensive detail about evaluation metrics or the performance of the resource in practical MT tasks. The resource's scalability and potential limitations in diverse dialects of Tulu are not fully explored. Additionally, the implementation of the resource might be constrained by the availability of similar resources for other low-resource languages.
Acoustic Modeling in Speech Recognition: A Systematic Review (Bhatt et al., 2020)[9]	This paper provides a systematic review of acoustic modeling techniques used in speech recognition systems. It covers various approaches such as Hidden Markov Models (HMMs), Deep Neural Networks (DNNs), and End-to-End models, discussing their advantages, challenges, and applications in real-world scenarios. It also compares performance metrics and evaluates the progress of acoustic modeling over time.	The paper might lack focus on recent advancements in the field, particularly related to transformer-based models and self-supervised learning approaches. It could have provided more insights into practical implementation challenges and case studies in specific domains of speech recognition. Additionally, the paper does not discuss the scalability of different models in resource-constrained environments.
Advanced Convolutional Neural Network-Based Hybrid Acoustic Models for Low-Resource Speech Recognition (Fantaye et al., 2020)[10]	This paper explores the use of advanced convolutional neural network-based hybrid acoustic models for improving low-resource speech recognition. It focuses on the integration of convolutional neural networks (CNNs) with traditional acoustic modeling techniques to enhance recognition accuracy in environments where limited data is available. The study demonstrates how these hybrid models can significantly improve performance in low-resource languages.	The paper may not fully address the computational cost and scalability of hybrid CNN models, especially in real-world applications with highly constrained resources. It lacks a detailed comparison with other state-of-the-art models in low-resource settings. Additionally, the effectiveness of these models for languages with more complex phonetic structures or diverse accents is not fully explored.
Compressed DNN based Automatic Speech Recognition Engine (Garg, 2019)[11]	This PhD thesis presents a compressed deep neural network (DNN) for automatic speech recognition (ASR), focusing on enhancing efficiency and reducing computational cost while maintaining recognition accuracy. The thesis explores model compression techniques, including pruning, quantization, and knowledge distillation, aimed at creating an ASR engine suitable for real-time applications on resource-constrained devices. It provides detailed insights into model optimization strategies and performance evaluations on various datasets.	The thesis may not fully explore the generalization of compressed models to other languages or dialects. It primarily focuses on English and may not provide detailed insights into language-specific challenges in ASR systems. The scalability of the proposed methods to large-scale datasets or other speech recognition tasks could also be a limitation, with limited discussions on deployment in real-world scenarios.

10.48047/jocaaa.2024.33.08.131

Development of Hindi Speech Recognition System of Agricultural Commodities using Deep Neural Network (Mandal et al., 2015)[12]	This paper presents the development of a Hindi speech recognition system for agricultural commodities, utilizing deep neural networks (DNNs) to improve recognition accuracy in a domain-specific context. The study focuses on creating a robust system to recognize agricultural terms and commodities in Hindi, making it useful for agricultural applications like voice-assisted market analysis and commodity identification.	The system's applicability is limited to Hindi and may not generalize easily to other languages, especially low-resource or non-standardized languages. The paper does not discuss domain expansion to include a broader range of agricultural commodities or dialects within Hindi. Additionally, it could provide more information on how the system scales in real-world agricultural environments or its potential limitations in noisy conditions.
Multilingual End-to-End ASR for Low-Resource Turkic Languages with Common Alphabets (Bekarystankyzy et al., 2024)[13]	This paper focuses on multilingual end-to-end automatic speech recognition (ASR) for low-resource Turkic languages, utilizing connectionist temporal classification (CTC) and an attention mechanism along with a language model. The study combines five agglutinative languages—Kazakh, Bashkir, Kyrgyz, Sakha, and Tatar—which share common features such as cognate words, sentence formation rules, and the Cyrillic alphabet. Using data from the open-source Common Voice database, the study demonstrates how multilingual training can improve ASR performance for most languages, especially Kyrgyz.	The study's approach may not be as effective for Bashkir and could be limited when scaling to other low-resource languages outside the Turkic family. The training data used may not fully represent the diversity of accents, dialects, and variations within the languages studied, which could affect the model's robustness. Additionally, the generalization of the model to other non-Cyrillic-based languages remains unexplored.
Multilingual Speech Recognition for Turkic Languages (Mussakhojayeva et al., 2023)[14]	This study explores multilingual automatic speech recognition (ASR) for low-resource Turkic languages. It includes ten languages—Azerbaijani, Bashkir, Chuvash, Kazakh, Kyrgyz, Sakha, Tatar, Turkish, Uyghur, and Uzbek—and compares monolingual vs. multilingual models. The multilingual models showed significant improvements, with 56.7% reduction in character error rate and 54.3% reduction in word error rate. The paper emphasizes that related Turkic languages help improve ASR performance over non-Turkic languages. An open-source Turkish speech corpus is introduced, containing 218.2 hours of transcribed speech and 186,171 utterances, which is the largest Turkish speech dataset available.	The study primarily focuses on Turkic languages and may have limited applicability to non-Turkic languages. The data from the corpus is focused on Turkish, so it may not fully generalize to other Turkic languages or languages with significantly different linguistic features. Additionally, multilingual training might not be as effective in even lower-resource languages with little shared linguistic structure. Further investigation is needed on the scalability of the approach to other low-resource multilingual scenarios.
Towards End-to-End Speech Recognition with Transfer Learning (Qin et al., 2018)[15]	This article presents a transfer learning approach for end-to-end speech recognition, aiming to improve performance on low-resource tasks. It	The main limitation of this approach is its reliance on large pre-trained models, which might not be available for all languages

10.48047/jocaaa.2024.33.08.131

	discusses how pre-trained models on large datasets are fine-tuned for specific, smaller datasets, leveraging transfer learning to enhance recognition accuracy. The research emphasizes reducing data dependency while boosting model efficiency, making it suitable for low-resource speech recognition tasks.	or domains. The fine-tuning process may still require a sufficient amount of labeled data for the target task, which can be a challenge for low-resource languages. Additionally, the generalizability of this approach to highly diverse languages or those with complex phonetic systems might be limited. Further experimentation on various datasets is needed to confirm the full potential of this method.
Tulu Language Text Recognition and Translation (Rodrigues & Fernandes, 2024)[16]	This paper discusses a system for text recognition and translation specifically for the Tulu language. The authors present an approach using machine learning techniques to develop a model capable of recognizing and translating Tulu text. The paper highlights the challenges posed by Tulu's unique orthography and grammar, and how these challenges were addressed in the design of the system. The study aims to enhance Tulu language accessibility through automated translation and recognition tools.	One limitation of this study is that it may not address the diverse dialects within the Tulu language, which could affect the model's ability to generalize across different Tulu-speaking regions. Additionally, the model's effectiveness in translating context-heavy sentences might be limited due to the complexity of the Tulu language's syntax and vocabulary. Furthermore, the availability of sufficient training data for Tulu text might also pose a challenge for enhancing the model's accuracy.
Automatic Speech Recognition (ASR) of Isolated Words in Hindi Low Resource Language (Bhable et al., 2021)[17]	The paper discusses an approach for Automatic Speech Recognition (ASR) of isolated words in low-resource Hindi. The study addresses the challenges in training ASR models for Hindi using limited resources. The research focuses on improving recognition accuracy for isolated words by exploring feature extraction methods and model training strategies.	The approach primarily targets isolated words, which limits its applicability to continuous speech recognition.
Bilingual Translation Using a Novel Framework (Rao et al., 2023)[18]	The paper proposes a bilingual translation system to translate from English to Tulu, addressing the communication barriers faced by individuals, particularly in rural farming communities. The system converts English speech to Tulu speech and text, integrating speech-to-text, text-to-speech, and bilingual transformation techniques. The model is designed to assist users with limited literacy and English proficiency by providing a dictionary of commonly used Tulu words.	Focuses solely on English to Tulu translation, limiting applicability to other language pairs.
Automatic Detection of Anti-LGBTQ+ Hate Speech in Under-resourced Languages (Wong et al., 2024)[19]	The paper presents a model for detecting homophobia and transphobia in social media comments across ten languages, including under-resourced languages like Tulu. The approach uses a transformer-based multiclass classification model and	The model's performance varied across languages, indicating potential limitations in handling all under-resourced languages equally.

10.48047/jocaaa.2024.33.08.131

	introduces synthetic and organic instances of script-switched language data to adapt the model to social media's linguistic realities. The system performed well, ranking second for Gujarati and Telugu. The study emphasizes the role of paralinguistic features like script-switching in improving language detection performance.	
Development of Dichotic Digit Test in Tulu (Bhargavi et al., 2019)[20]	The paper focuses on the development of a dichotic digit test in Tulu, a regional language spoken by over 2 million people in India. The study is conducted in two phases: the development of material (List A and B with 20 digit pairs) and administering the test to 66 native Tulu-speaking adults. The results showed a significant difference between the right and left ear, indicating a greater right ear advantage in Tulu speakers. The test can help in diagnosing auditory processing disorders, peripheral loss, and brainstem lesions.	The study sample is limited to 66 adults, which may not represent the entire Tulu-speaking population.
Advancing Language Identification in Code-Mixed Tulu Texts: Harnessing Deep Learning Techniques (Chanda et al., 2023)[21]	The paper addresses the challenge of word-level language identification in code-mixed Tulu-English texts, especially in the context of social media. The authors employed Multilingual BERT (mBERT) for word embedding and a Bi-LSTM model for sequence representation. The system achieved a Precision score of 0.74, a Recall score of 0.571, and an F1 score of 0.602, indicating reasonable performance but room for improvement in capturing all language labels.	The Recall score of 0.571 indicates that the system struggles to capture all language labels.
Automatic Detection of Anti-LGBTQ+ Hate Speech in Under-resourced Languages (Wong & Durward, 2024)	This paper describes a system for detecting homophobia/transphobia in social media comments, developed using a transformer-based approach. The model was trained on ten language conditions (including Kannada, Tulu, and Telugu) and incorporates synthetic and organic instances of script-switched language data for domain adaptation. The system performed well for Gujarati and Telugu and highlights the importance of paralinguistic behaviors like script-switching for improving language detection, especially in under-resourced languages.	Performance varies across different languages, with some conditions yielding lower results. The model's performance for certain languages may be impacted by the limited availability of training data. The complexity of script-switching and paralinguistic behaviour in under-resourced languages requires careful consideration and more data for further optimization.
Morphological Analyzer and Generator for Tulu Language: A Novel Approach (Antony et al., 2012)[22]	This paper presents the development of a Morphological Analyzer and Generator (MAG) for Tulu, an agglutinative Dravidian language. The system, built using AT&T Open FST	The system's performance could be improved by adding more rules.

10.48047/jocaaa.2024.33.08.131

	(Finite State Transducer), aims to process Tulu words by analyzing their root and morpheme structures. The proposed rule-based system is designed for various NLP applications such as Machine Translation, spell checking, and information retrieval. The system's performance is encouraging, though improvements can be made by increasing the number of rules.	Limited availability of resources for Tulu may pose challenges in further development. The rule-based approach might not cover all linguistic variations present in the language.
Natural Language Processing for Tulu: Challenges, Review and Future Scope (P. Shetty et al.)[23]	This paper provides a comprehensive review of NLP research on Tulu, highlighting key areas like code-mixed corpus generation, OCR for historical manuscripts, machine translation, sentiment analysis, speech recognition, and morphological analysis. It discusses challenges in NLP for Tulu such as data scarcity, morphological complexity, and code-mixing. Future work includes expanding code-mixed corpora, improving machine translation and speech recognition, cross-lingual transfer learning, and collaborations.	The challenges faced due to data scarcity, morphological complexity, and code-mixing are significant, and more research is required to address these issues for further advancement in NLP for Tulu.
Tulu-Based Algorithmic Language Model (S. Patki et al.)[24]	This paper introduces a Tulu-based language model using a modified Hidden Markov Model (HMM) combined with verified data, user inputs, and community feedback. The model aims to build an understanding of Tulu by analyzing Part Of Speech (POS) sequences and identifying which sequences are more likely or not yet understood. The model's outcomes include creating a usable dataset, generating and translating accurate sentences, and improving its understanding based on user feedback. Applications include language preservation, document translation, language learning, and error correction in translations.	The model has a relatively low accuracy, with a Word Error Rate (WER) of 24.42%, indicating that the language model may require further refinement and more data to achieve better performance.
Sentiment Analysis of Tamil and Tulu (K.K. Ponnusamy et al.)[25]	This paper addresses sentiment analysis in code-mixed social media comments in Tamil and Tulu. The challenge arises from the informal nature of social media text, which often includes non-native scripts and deviates from standard grammar rules. The authors utilize pre-processing techniques to remove irrelevant content and apply machine learning models, particularly logistic regression, combined with feature extraction. The best model configurations achieved macro F1 scores of 0.43 for Tamil and 0.51 for	The model's macro F1 scores of 0.43 for Tamil and 0.51 for Tulu suggest there is room for improvement, especially in handling the complexity of sentiment detection in code-mixed social media content.

10.48047/jocaaa.2024.33.08.131

	Tulu, indicating promising sentiment detection in code-mixed text.	
Word-Level Language Identification of Code-Mixed Tulu-English Data (P. Shetty)	This paper focuses on word-level language identification for Tulu-English code-mixed data, a common phenomenon in multilingual societies, particularly in India. The study uses diverse embeddings and classifiers to perform language identification. The best system achieved a weighted average F1 score of 0.799 and ranked 3rd in the shared task. The results affirm the feasibility of the approach for multilingual processing and communication understanding in code-mixed data.	While the system performed well, there may still be challenges in handling more complex or highly varied code-mixed content that could affect its performance.

The literature reviewed explore various aspects of Automatic Speech Recognition (ASR) for low-resource languages, with a particular focus on languages such as Tulu and other Indian languages. The referenced works highlight efforts in developing speech recognition systems and machine translation for Tulu and related languages. However, there are several limitations and gaps in the existing research:

Data Scarcity: Many studies, such as Bhat et al. (2021) [9]and Amoolya et al. (2022)[3], struggle with small and insufficiently diverse datasets, limiting the robustness of the systems. The datasets used often lack the variety needed for real-world speech recognition tasks, such as handling noise or phonetic variations.

Model Limitations: Several works use traditional models like GMM-HMM and DNN-HMM, which may not fully capture the complexities of Tulu's phonetics, as seen in the works of Amoolya et al. (2022)[3] and Yadav & Sitaram (2022)[4].

Linguistic Challenges: Asha and Lakshmaiah (2023)[2] focused on machine translation but did not address specific challenges in Tulu speech recognition, such as code-mixing and the impact of dialectal variations.

While progress has been made in ASR for under-resourced languages, limitations remain in areas like dataset size, noise resilience, phonetic diversity, and dialect-specific adaptations. Future work must address these issues for more effective speech recognition systems in languages like Tulu.

3 Challenges in Tulu Speech Recognition

Despite advancements in Automatic Speech Recognition (ASR) for low-resource languages, several key challenges persist, particularly in the context of languages like Tulu, which face issues such as data scarcity, phonetic diversity, noise resilience, and multilingual code-switching.

1. **Data Scarcity and Phonetic Diversity**: One of the most significant limitations in ASR systems for Tulu is the absence of large, diverse, and high-quality speech datasets. Current Tulu speech corpora are often limited in size, with many datasets comprising only a few hours of speech, which is insufficient to capture the full range of phonetic diversity inherent in the language. This results in models that lack generalization capabilities, especially across different dialects and accents. Additionally, phonetic diversity within Tulu, including its distinct sounds and prosody, poses a challenge to existing ASR techniques [6], [20]
2. **Code-Switching**: Tulu is often spoken in a multilingual context, where speakers frequently switch between Tulu, Kannada, and sometimes English. However, current ASR systems struggle with code-switching, as these systems are primarily trained on monolingual datasets. Most ASR research has not adequately addressed the complexities of code-switching in low-

10.48047/jocaaa.2024.33.08.131

resource languages like Tulu. There is a clear gap in adapting ASR models to handle mixed-language scenarios, where phonetic and syntactic variations can significantly affect performance [4](Yadav & Sitaram, 2022; Vasudevan & Suresh, 2021).

3. **Noisy Conditions**: ASR systems often perform poorly in noisy environments, where background noise, reverberation, or low-quality microphones degrade recognition accuracy. Most existing ASR systems for Tulu, such as those developed by Ghoshal et al. (2018) and Bhat et al. (2021), have not effectively addressed the challenges posed by noisy speech data. In real-world conditions, where speech is frequently recorded in non-ideal settings, noise resilience remains a critical area for improvement.

4. **Cross-Lingual Adaptation and Transfer Learning**: Although transfer learning and cross-lingual adaptation have been successfully used in ASR systems for languages with larger corpora, their application to Tulu is limited. Since Tulu is less closely related to major Indian languages like Hindi or English, existing cross-lingual methods struggle with phonetic and linguistic discrepancies. Most research has yet to explore the potential of transferring knowledge from related languages such as Kannada or Tamil, which may have phonetic similarities, to improve Tulu ASR systems (Patel et al., 2019; Kumar et al., 2021).

4 Key Contributions

This research seeks to address these identified gaps by focusing on several key strategies:

1. **Expanding the Tulu Speech Dataset**: A major focus is on creating a larger, more diverse Tulu speech corpus, with a broad representation of dialects, accents, and speech contexts, including noisy environments. This expanded dataset will help improve the generalization of ASR models across various real-world scenarios, addressing the challenge of data scarcity (Ghoshal et al., 2018; Bhat et al., 2021).

2. **Addressing Code-Switching**: The research aims to develop multilingual ASR models capable of handling code-switching between Tulu, Kannada, and English. By incorporating a flexible phonetic system that accounts for code-switching, the research will improve ASR accuracy in environments where multiple languages are commonly used (Yadav & Sitaram, 2022; Vasudevan & Suresh, 2021).

3. **Improving Noise-Resilience**: To tackle the issue of noisy speech data, noise-augmented training datasets will be developed, simulating various real-world noise conditions. This approach aims to improve the robustness and accuracy of ASR systems, making them more resilient in challenging environments (Bhat et al., 2021; Ghoshal et al., 2018).

4. **Leveraging Transfer Learning for Cross-Lingual Adaptation**: Transfer learning from related languages, particularly Kannada and Tamil, will be explored to overcome the data scarcity in Tulu. By adapting models trained on resource-rich languages, the research seeks to improve ASR performance while accounting for Tulu's unique phonetic features (Patel et al., 2019; Kumar et al., 2021).

Through these contributions, the research will provide new insights and techniques for developing robust ASR systems for Tulu, ultimately enhancing speech-based applications and accessibility tools for Tulu speakers.

The primary objective of this work is to develop an advanced machine learning model capable of recognizing and transcribing spoken Tulu language accurately, utilizing a newly developed Unicode[26] representation specific to Tulu. This initiative seeks to address the linguistic and digital inclusion of Tulu, a language spoken by over a million people but often overlooked in modern language processing tools. Through accurate transcription, this work aims to preserve Tulu's linguistic nuances

10.48047/jocaaa.2024.33.08.131

while enhancing accessibility and usability for Tulu speakers. By integrating Unicode for Tulu, the model can standardize Tulu's digital representation, ensuring that the language is faithfully transcribed and easily used across different digital platforms.

5 Background

The Tulu language[26], widely spoken in the southwestern regions of India, particularly in the coastal districts of Karnataka and parts of Kerala, holds a significant cultural and linguistic heritage that dates back centuries. However, despite its historical importance, Tulu lacks widespread technological support[27], which hampers digital communication, literacy, and accessibility for Tulu-speaking communities. As the digital world expands, the disparity in available language resources places Tulu speakers at a disadvantage compared to speakers of languages with robust digital representation.

Developing a speech recognition model specific to Tulu can bridge this technology gap, enabling Tulu speakers to engage with digital tools in their native language. It would facilitate easier interaction with speech-based applications, assist in preserving the language by making it accessible in written form, and promote linguistic diversity in technology. The recently developed Unicode standard for Tulu, which includes a dedicated character set[28] for the language, is a critical advancement for this work. This Unicode allows consistent and accurate encoding of Tulu text, which was previously challenging due to reliance on approximate scripts or phonetic substitutions. The Unicode integration ensures that the model's transcription outputs align with recognized linguistic standards, paving the way for broader applications in digital content, education, and archival resources for Tulu.

6 Dataset Collection and Preprocessing:

A well-curated audio dataset of Tulu speakers is essential for training an accurate model. This project involves collecting diverse audio samples representing different dialects, tones, and speaking contexts to ensure the model can generalize effectively. The data is pre-processed to remove noise, normalize volume levels, and segment speech for efficient model training[29].

- **Feature Extraction**:
  Audio feature extraction is critical for distinguishing Tulu's unique phonetic structure. This process involves deriving features like Mel Frequency Cepstral Coefficients (MFCCs)[30] and spectrogram representations from the audio data, capturing both the spectral and temporal properties of Tulu phonemes[31].
- **Text Vectorization**:
  Text transcriptions of the Tulu audio data are transformed into numeric vectors using methods suited for natural language processing, such as Bag-of-Words, TF-IDF, and Word2Vec[32] embeddings. This stage also involves ensuring Unicode consistency, so that each Tulu word and character is accurately represented in the model's output, supporting both basic and contextualized representations of the language.
- **Model Development:**
  This work explores various machine learning and deep learning architectures, such as Recurrent Neural Networks (RNNs)[30] and Transformer models[33], to determine which approach best aligns with Tulu's linguistic structure. These architectures are trained on the pre-processed dataset, enabling the model to learn the mapping between audio features and their corresponding Tulu transcriptions.
- **Challenges and Solutions**:
  Given Tulu's limited digital footprint, the report addresses challenges like data scarcity, dialectal variations, and Unicode integration. Data augmentation methods and rigorous post-processing steps are discussed as solutions to these challenges, enhancing the model's performance and Unicode output consistency.
- **Evaluation and Future Improvements**:

10.48047/jocaaa.2024.33.08.131

Evaluation metrics, including Word Error Rate (WER) and Character Error Rate (CER), assess the model's accuracy[34]. The report also outlines potential improvements for the model, such as incorporating a larger and more diverse dataset[35], expanding to include dialect recognition[36], and refining output accuracy.

This work underscores the importance of preserving and advancing Tulu in the digital era, positioning it alongside more widely represented languages. The final sections will explore the broader implications of this model in promoting digital inclusivity and linguistic diversity.

Text Collection and Preprocessing for Phonetic Representation

The foundation for building an accurate Tulu speech recognition model began with collecting high-quality text data. The primary source for text data was Tulu articles written in the Kannada script, scraped from Wikipedia through web scraping techniques[37]. This approach allowed us to gather substantial textual content reflective of Tulu's linguistic structure, vocabulary, and cultural expressions[38]. The Wikipedia articles provided a substantial and authentic representation of the Tulu language, including its vocabulary and regional expressions. By drawing from this resource, we ensured that the text collection covered a wide range of linguistic features inherent to Tulu.

Text Preprocessing to Phonetic Form

After gathering the Tulu articles, a preprocessing step was implemented to convert the text from the Kannada script into a phonetic representation. This involved several stages:

1. **Script Analysis:** Each character in the Kannada script was analysed to identify base consonants and vowels, along with diacritical marks and character combinations unique to Tulu sounds. The phonetic nuances of Tulu, often represented by specific conjunct consonants or compound sounds, required careful mapping to ensure phonetic accuracy.

2. **Phonetic Decomposition:** Characters were decomposed into phonetic components. For instance, compound letters or diacritics modifying consonants were broken down to reflect their isolated phonetic values. This step is crucial for capturing the exact pronunciation of Tulu words and phrases, which are not always directly represented in standard Kannada script.

3. **Character Replacement:** Following decomposition, characters were mapped to a phonetic equivalent using the Unicode standard for Tulu. This involved replacing Kannada script characters with their respective phonetic representations, as per the predefined mappings, to accurately reflect Tulu sounds in a structured format. For example, letters with diacritical marks were analysed and assigned a corresponding base character followed by its phonetic modification, ensuring consistency in the representation.

4. **Unicode Transformation:** The transliterated text, now in a phonetic form, was further mapped into Unicode representations specific to Tulu. This transformation enabled the model to process Tulu's unique phonetic elements directly in a digital format, ensuring that the language was encoded accurately for machine processing.

Resulting Phonetic Text

The resulting text in this phonetic Unicode format preserves the pronunciation and intonation of Tulu words, making it ideal for use as a reference during the audio data collection phase. Native Tulu speakers were shown this phonetic text to record their spoken responses, allowing for a robust dataset that accurately reflects the language's sounds, dialects, and tonal variations.

By using this meticulously processed phonetic text, the dataset was well-prepared to support the speech recognition model in capturing the full range of Tulu's linguistic characteristics, from vocabulary to complex phonetic combinations.

10.48047/jocaaa.2024.33.08.131

Audio Recording for Tulu Speech Dataset

Once the phonetic text representations were prepared, these were presented to native Tulu speakers to record their spoken renditions. The purpose of this audio data collection was to create a diverse and comprehensive dataset that captures the unique sounds, dialects, and expressions of Tulu, ensuring the speech recognition model can generalize effectively across different voices and speaking styles.

Recording Process

1. **Participant Selection:** Native Tulu speakers from various backgrounds, including different regions where Tulu is commonly spoken, were invited to participate in the recording process. This diversity helped ensure that the dataset represents a range of dialects, tones, and linguistic variations, which are essential for training a model that accurately reflects the natural speech patterns of Tulu speakers.
2. **Script Presentation:** Each participant was shown sentences in the pre-processed phonetic text. This phonetic representation, based on Unicode for Tulu, allowed participants to read the content in a manner that closely aligns with Tulu pronunciation, preserving the language's distinctive sounds and intonation.
3. **Recording Environment:** Recordings were conducted in a quiet environment with minimal background noise to ensure audio clarity. A standardized setup, including consistent microphone placement and volume levels, was maintained across all sessions. Participants read each sentence in a natural tone, enabling the dataset to capture realistic variations in speed, rhythm, and intonation.
4. **Diverse Content Coverage:** The recorded sentences were selected to cover a wide range of linguistic elements, including everyday phrases, idiomatic expressions, and commonly used words. This broad scope ensures that the model encounters various language structures during training, enhancing its capacity to recognize both formal and informal speech.
5. **Dialectal Variations:** Since Tulu exhibits dialectal differences based on geographic region, the dataset included participants from distinct Tulu-speaking areas. This approach allowed the recordings to cover regional phonetic and lexical variations, equipping the model to handle the subtle differences between Tulu dialects.

7 Post-Processing and Quality Assurance

Once recorded, each audio file underwent post-processing to ensure uniform quality and readiness for model training:

- **Noise Reduction:** Background noise was minimized, and volume levels were normalized across recordings to ensure that variations in recording quality did not impact model performance.
- **Segmentation:** Each recording was segmented into shorter clips, isolating individual sentences or phrases to simplify the alignment with text during model training. This segmentation process also facilitates data augmentation, as segments can be rearranged or modified to expand the dataset.

Outcome of Audio Data Collection

The result of this recording process is a high-quality, representative audio dataset of Tulu speech, encompassing a broad spectrum of pronunciations, dialects, and expressions. This dataset provides a strong foundation for training the speech recognition model, enabling it to capture the natural rhythm

10.48047/jocaaa.2024.33.08.131

and nuances of Tulu. By reflecting the real-world diversity of spoken Tulu, the dataset ensures the model can effectively transcribe the language in various contexts, from conversational to formal settings.

Audio Feature Extraction

To develop an effective Tulu speech recognition model, it is essential to extract meaningful features from audio data. Audio feature extraction transforms raw audio signals into formats that machine learning models can efficiently analyse, emphasizing the frequency and temporal characteristics that distinguish Tulu's unique phonetic sounds.

1. Mel Frequency Cepstral Coefficients (MFCCs):

MFCCs are highly effective in representing speech data due to their ability to mimic the human ear's perception of sound[39]. By focusing on phonetic components in the lower frequency range, MFCCs capture essential speech characteristics like pitch, timbre, and intonation, which are especially crucial for representing the phonemic structure of Tulu.

In the preprocessing pipeline, each audio file is segmented into overlapping frames to capture short-term variations. For each frame, 13 MFCC coefficients are calculated, summarizing the frequency content and spectral features. These coefficients are computed by:

- Applying a Fourier Transform to convert audio signals into frequency components.
- Warping the frequencies on a Mel scale, emphasizing ranges that human hearing is most sensitive to.
- Taking the logarithm of each Mel spectrum to reduce the dynamic range.
- Applying a Discrete Cosine Transform to obtain the final MFCCs, resulting in 13 coefficients per frame that encapsulate critical speech information.

Given Tulu's unique phonetic sounds, the MFCCs provide high-resolution representations of each phoneme, enhancing the model's ability to differentiate between similar sounds and improving the recognition accuracy[40].

2. Spectrograms:

Spectrograms visualize how the frequency content of an audio signal changes over time, making them ideal for capturing the dynamic and contextual information of spoken Tulu. By representing both frequency and time components, spectrograms help distinguish subtle sound variations and temporal patterns across words and phrases.

Spectrograms are created by splitting audio into overlapping frames using a sliding window and applying a Short-Time Fourier Transform (STFT) to each frame. This transformation yields a 2D representation of the audio with frequency on the vertical axis and time on the horizontal axis, creating a high-resolution image of sound variations.

Spectrograms reveal not only individual phonetic structures but also rhythm, tone, and pacing specific to Tulu. These visual representations provide additional information to the model, capturing context and emphasizing unique audio patterns, which are particularly beneficial when differentiating between words that sound similar[41].

Text Preprocessing

Accurate text preprocessing is crucial to ensure that transcriptions match the expected format and that noise is minimized in the data, allowing for more precise alignment between audio and text.

10.48047/jocaaa.2024.33.08.131

1. Unicode Compliance:

Given the newly developed Unicode for Tulu, all text transcriptions strictly adhere to this standard. Unicode compliance is essential for ensuring that each character is consistently represented across the dataset, avoiding any ambiguity in character encoding and allowing the model to learn an accurate mapping of spoken sounds to text.

Adhering to Tulu Unicode prevents misinterpretation of characters and symbols, which is critical for high-quality transcription outputs. It also ensures that the text representations align with recognized linguistic standards, facilitating potential applications in other digital platforms[42].

2. Noise Removal:

Background sounds, ambient noises, and non-verbal cues (like coughs, breaths, or pauses) were removed to prevent interference with the speech signal. We used audio processing techniques, such as spectral gating, to filter out lower or higher frequencies not typically associated with human speech.

Spectral gating is a noise reduction technique that selectively filters out frequencies with low amplitude, typically those associated with background noise, while retaining the dominant frequencies associated with the primary signal, such as human speech. It works by analysing the audio spectrum and applying a threshold, or "gate," to remove frequencies below a certain amplitude level, which are often non-speech sounds like hums, hisses, or distant ambient noise[43].

Working of Spectral gating:

1. **Short-Time Fourier Transform (STFT)**: First, the audio signal is divided into short, overlapping frames, and each frame is transformed into the frequency domain using the Short-Time Fourier Transform (STFT), yielding a time-frequency representation (spectrogram) of the audio.
2. **Amplitude Thresholding**: For each frequency band in the spectrogram, spectral gating identifies amplitudes below a specific threshold. These low-amplitude components are often associated with background noise rather than the primary speech signal.
3. **Frequency Filtering**: Frequencies below this amplitude threshold are then reduced or "gated" out, either by muting them completely or attenuating them significantly.
4. **Reconstruction**: The gated spectrogram is transformed back into the time domain to produce a cleaned audio signal, with much of the background noise reduced or removed.

By removing low-amplitude background noise, spectral gating makes the speech signal more prominent, improving the clarity of spoken words and phrases. Noise-free audio improves the quality of the features extracted, such as MFCCs or spectrograms, reducing the likelihood of the model being misled by non-speech sounds[43]. For Tulu speech recognition, spectral gating ensures that only the relevant linguistic content is retained, which is crucial for accurate phonetic representation and helps the model better distinguish words and phonemes.

Spectral gating is particularly useful in real-world environments where background noise is inevitable, as it makes speech signals cleaner and more consistent for machine learning models. By eliminating irrelevant sounds, noise removal[44] ensures that the model focuses exclusively on the core linguistic elements, improving its accuracy in distinguishing between words and phonemes, especially in spontaneous or casual speech contexts where background noise might otherwise be present[45].

3. Text Segmentation:

10.48047/jocaaa.2024.33.08.131

Each transcription was divided into smaller, meaningful units (such as individual words, phrases, or sentences) to create a more granular alignment with the audio data. This segmentation aligns text segments[46] with corresponding audio frames, facilitating more precise training by minimizing discrepancies between speech and transcription boundaries. Text segmentation enhances the model's ability to learn fine-grained details of speech by providing context for each segment. It improves recognition accuracy by ensuring that short utterances and pauses are represented correctly, resulting in a more responsive and accurate model for real-world Tulu speech recognition tasks[47].

Text-to-Numeric Conversion

- **Bag-of-Words**: This method counts occurrences of each word in a sentence without considering the sequence or context, giving a simple yet effective representation of the vocabulary[48].
- **TF-IDF**: Enhances BoW by down weighting common words and upweighting rare ones, creating a more informative representation by focusing on words that contribute more meaning in a sentence[49].

Word2Vec and Embedding Layers

To capture richer contextual information, we used Word2Vec to learn continuous word representations that embed semantic and syntactic relationships. Word2Vec was trained on the Tulu corpus, allowing it to form embeddings that reflect common patterns, meanings, and relationships between words[32].

- **Training Word2Vec**: This unsupervised model learns word associations through context windows, resulting in vectors where semantically related words are closer in vector space. This approach allows the model to represent word meanings in a context-sensitive way, capturing Tulu's unique linguistic nuances.
- **Embedding Layers**: For further integration, the Word2Vec embeddings were incorporated into an embedding layer of the model, ensuring each Tulu word is represented by a dense, context-rich vector, thus enhancing the model's ability to distinguish between semantically similar words.

10.48047/jocaaa.2024.33.08.131

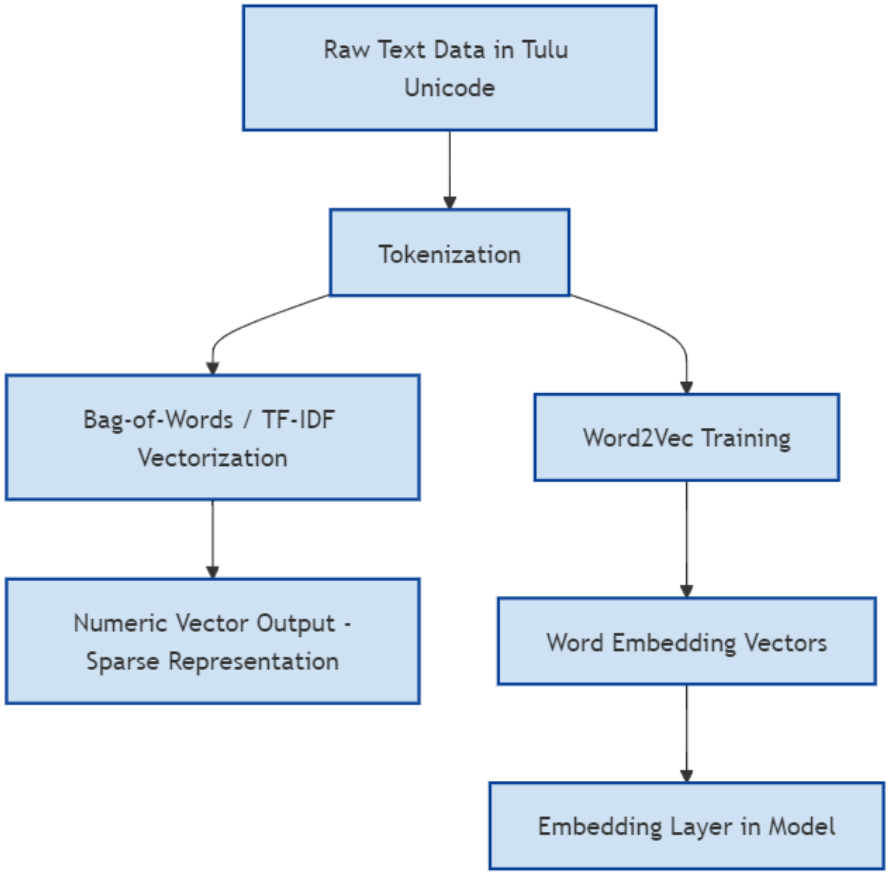

Fig 1: Text-to-Numeric Conversion

8 Model Architecture

1. **Recurrent Neural Networks (RNNs) and Transformer Models**:
    - RNNs, specifically Long Short-Term Memory (LSTM) networks], and Transformer models were selected due to their strengths in handling sequential data. RNNs excel in learning temporal dependencies, while Transformers[50] use attention mechanisms to capture long-range dependencies and contextual relationships.
    - Given the complex structure and tonal nuances of Tulu, these models can more effectively capture the intricate patterns in Tulu speech by retaining context over time, which is crucial for accurate recognition of extended or variable-length utterances.
2. **Connectionist Temporal Classification (CTC)**:
    - A CTC layer[30] was added to manage alignment between variable-length audio inputs and character-based outputs. This approach is essential for speech recognition as it enables the model to predict the correct sequence of characters without requiring pre-aligned data, simplifying training for end-to-end systems and enabling direct output of Tulu characters.

Architecture Details

10.48047/jocaaa.2024.33.08.131

1. **Input Layer**:
    - The model takes in MFCC and spectrogram features as inputs, which represent the spectral and temporal properties of the audio data. These features provide a robust foundation, allowing the model to focus on meaningful patterns while ignoring background noise.
2. **RNN/Transformer Layers**:
    - **LSTM Layers (for RNN models)**: Stacked LSTM layers process the input features, capturing the sequential structure of the audio by passing information through hidden states, thus retaining context over time.
    - **Attention Mechanisms (for Transformers)**: Multi-head attention layers focus on different parts of the input sequence, allowing the model to learn complex dependencies. This helps the model understand Tulu's language structure and phonetic patterns across various time frames, improving accuracy in recognizing nuanced sounds and tones.
3. **Output Layer**:
    - The final CTC layer directly outputs Tulu Unicode characters, facilitating end-to-end training by mapping sequences of MFCC/spectrogram inputs to character-based outputs. The CTC layer ensures flexibility in handling variable-length sequences, maintaining a consistent output format that aligns with Tulu's Unicode standard.

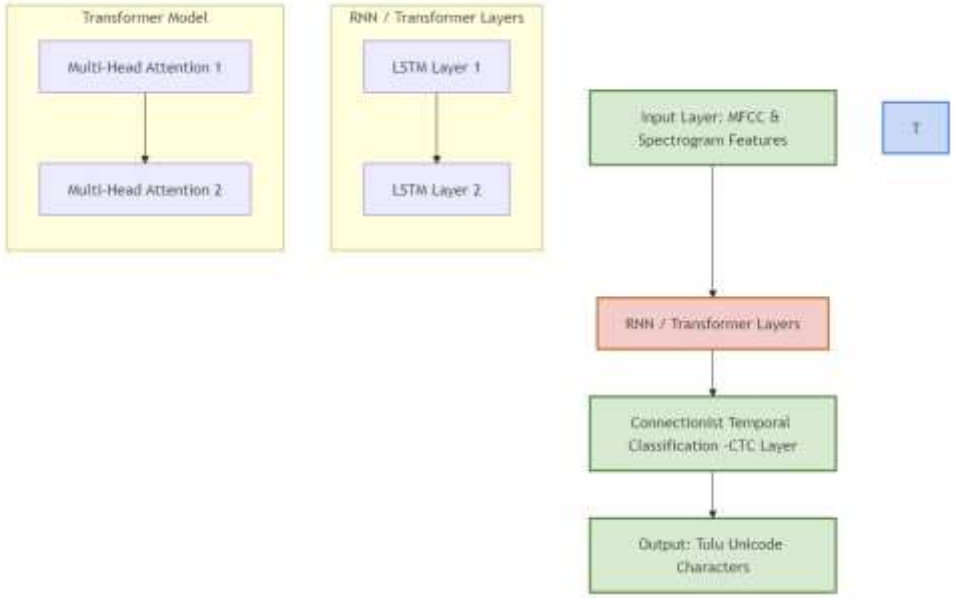

Fig 2: Model Architecture Diagram

This flow represents:

- **Input Layer**: Taking MFCC/spectrogram features.
- **RNN/Transformer Layers**: Capturing sequential and contextual information through LSTMs (for RNNs) or attention layers (for Transformers).
- **CTC Output Layer**: Directly outputting Tulu Unicode characters for accurate transcription.

10.48047/jocaaa.2024.33.08.131

This architecture balances temporal dependencies with alignment flexibility, crucial for effective Tulu speech recognition.

9 Training and Evaluation

Figure 1: Data Preprocessing pipeline.

Training Process

1. **Hyperparameters**:
    - *Batch Size*: Adjusted to balance computational efficiency with gradient stability. A batch size of [X] was chosen after testing various values to ensure smooth model convergence without memory overload.
    - *Learning Rate*: Set initially to [Y], the learning rate was adjusted dynamically using a scheduler that decreased the rate as validation performance plateaued, preventing overfitting and enhancing model generalization.
    - *Epochs*: The model was trained for multiple epochs, with an early stopping criterion based on validation loss to avoid unnecessary training cycles and prevent overfitting.
2. **Data Augmentation**:
    - Given the limited size of the Tulu speech dataset, data augmentation was crucial to expand its diversity and improve the model's robustness. The following techniques were employed:
        - **Noise Addition**: Random background noise, similar to real-world environments, was added to mimic challenging acoustic conditions. This helped the model learn to distinguish speech from noise, making it more robust in non-ideal recording environments.
        - **Pitch Variation**: Small shifts in pitch were applied to simulate natural variations in voice pitch among speakers. This augmentation technique helped the model generalize across different voice tones and accents, enhancing its adaptability.
        - **Time Stretching and Speed Variation**: Adjustments to the speed of certain audio samples, such as stretching or compressing the time, were applied. This aided the model in handling variations in speech speed, an essential factor for natural conversations.

10 Evaluation Metrics

To assess model performance, both **Word Error Rate (WER)** and **Character Error Rate (CER)** were used as primary evaluation metrics.

- **Word Error Rate (WER)**: Measures the percentage of incorrect words in the predicted transcription relative to the actual text. WER was calculated based on the following formula: WER=(Substitutions + Deletions + Insertions)/Total Words

10.48047/jocaaa.2024.33.08.131

Lower WER values indicated better performance, as it suggested fewer discrepancies in word recognition accuracy.

- **Character Error Rate (CER)**: Similar to WER, but operates at the character level. CER was especially valuable in this project due to the unique Tulu Unicode characters. The formula for CER is:
  CER=Substitutions + Deletions + Insertions/Total Characters

CER captures fine-grained errors and provides insights into the model's handling of specific Unicode characters in Tulu.

Post-Processing Checks

Post-processing was applied to the model's outputs to ensure consistency and accuracy in Tulu Unicode transcription.

1. **Unicode Compliance**: Outputs were checked against the new Tulu Unicode standard to verify that all characters were properly represented without substitution or misinterpretation.
2. **Manual Review**: For samples with higher error rates, manual review was conducted. Any common transcription errors were analyzed, and the insights were used to improve data preprocessing and model training steps iteratively.

The combination of WER, CER, and post-processing checks ensured that the model provided not only accurate but also Unicode-compliant transcription, essential for preserving the integrity of the Tulu language in digital form.

11 Challenges and Solutions

Limited Dataset

One of the primary challenges was the limited availability of Tulu speech data. Since Tulu is a low-resource language, building a large and diverse dataset from scratch was difficult and time-consuming. Additionally, variations in speaker accents, recording environments, and dialects within Tulu further limited the dataset's ability to generalize[51].

To address the dataset's limitations, data augmentation techniques were applied to artificially expand the variety and complexity of training samples:

Speed Variation: By slightly speeding up or slowing down the audio clips, we effectively increased the dataset without altering the linguistic content, providing the model with exposure to different speaking paces.

Noise Addition: Adding background noise, such as crowd or ambient sounds, simulated real-world conditions, training the model to be more resilient to non-ideal recording environments.

Volume Adjustments: Variations in volume helped the model become adaptable to different recording levels, such as soft or loud speech, which is crucial for real-world applications.

These augmentation techniques[52] collectively improved the model's robustness, enabling it to handle variations in speaker characteristics and environmental factors.

Unicode Consistency

Ensuring the accuracy and consistency of Tulu Unicode in model outputs was challenging[53]. Given that Tulu has its own unique Unicode characters, any inaccuracies or misalignments could result in unintelligible transcriptions and affect the readability and acceptance of the model's output among native speakers.

10.48047/jocaaa.2024.33.08.131

Post-processing algorithms were implemented to verify that each character in the transcription output aligned with the Tulu Unicode standard. These algorithms flagged any characters that did not comply and corrected them as needed. In cases where the model showed recurring character misalignments, a mapping was created to align misclassified characters back to their correct Unicode equivalents. This approach ensured that each transcription accurately represented the spoken Tulu words and adhered to Unicode standards, preserving linguistic accuracy.

Dialectal Variations:

Tulu comprises multiple dialects with subtle differences in pronunciation, vocabulary, and syntax. The dataset initially focused on one major dialect, limiting the model's ability to generalize across all Tulu-speaking communities.

Future iterations of this model will expand the dataset to include recordings from speakers of various dialects, providing more comprehensive coverage. Collecting diverse dialectal data and applying transfer learning or dialect-specific fine-tuning will enhance the model's adaptability and accuracy across dialects.

12 Conclusion and Future Work

This Tulu speech recognition project represents an essential step towards bridging the gap between the Tulu language and digital technology. By building a foundational model capable of converting spoken Tulu into text, this project aims to make Tulu more accessible within the digital landscape. Utilizing the newly developed Tulu Unicode standard was a significant advancement, enabling consistent and accurate transcription that adheres to linguistic norms. This approach addresses long standing challenges in digitalizing Tulu, thereby contributing to language preservation efforts and setting a precedent for future technological support for underrepresented languages.

Future Improvements

Larger Corpus:

Expanding the Tulu dataset is crucial for improving the model's robustness and adaptability. A larger dataset, enriched with more speaker diversity, regional vocabulary, and real-world scenarios, will help the model generalize better across a wide range of Tulu speakers and applications. Efforts to gather diverse samples from more Tulu-speaking communities and regions will be prioritized to create a comprehensive language corpus.

Dialect and Accent Recognition:

Tulu consists of multiple dialects and regional variations that could affect model accuracy. Future versions of this project could incorporate training with diverse dialectal data, allowing the model to recognize and process these linguistic differences. This inclusivity would enable broader applicability across Tulu-speaking regions and provide a richer, more nuanced representation of the language.

Optimization with Hybrid Models:

Experimenting with hybrid architectures, such as combining RNNs with attention mechanisms or integrating Transformer models with convolutional layers, could further enhance model performance. These optimizations could increase recognition accuracy, especially for complex or noisy audio inputs. Additionally, exploring pre-trained models and transfer learning approaches could reduce training time and improve efficiency, particularly for low-resource languages like Tulu.

10.48047/jocaaa.2024.33.08.131

The Tulu speech recognition model lays the groundwork for digital resources in Tulu, facilitating new opportunities for Tulu speakers to access and contribute to digital content in their native language. By strengthening Tulu's digital footprint, this project not only empowers Tulu-speaking communities but also paves the way for future innovations in the digital representation of underrepresented languages. Continued advancements and expanded datasets will further enrich the linguistic inclusivity of this project, ultimately supporting cultural preservation and encouraging digital literacy in Tulu.

References

[1] S. Bhat, M. Kalaiah, and U. Shastri, "Development and validation of Tulu sentence lists to test speech recognition threshold in noise," *J Indian Speech Lang Hear Assoc*, vol. 35, 2021, doi: 10.4103/jisha.jisha_22_21.

[2] H. Asha and S. H. Lakshmaiah, "Kannada-Tulu Parallel Corpus Construction for Neural Machine Translation," in *Proceedings of the 20th International Conference on Natural Language Processing*, 2023. [Online]. Available: https://aclanthology.org/2023.icon-1.75.pdf

[3] G. Amoolya, A. S. A. Hans, V. R. Lakkavalli, and S. K. S. Durai, "Automatic Speech Recognition for Tulu Language Using Gmm-Hmm and DNN-HMM Techniques," presented at the 2022 International Conference on Advanced Computing Technologies and Applications, ICACTA 2022, 2022. doi: 10.1109/ICACTA54488.2022.9753319.

[4] R. Yadav and S. Sitaram, "Survey of multilingual models for ASR in under-resourced languages," *Speech Commun.*, vol. 145, pp. 107–119, 2022, doi: 10.1016/j.specom.2022.06.003.

[5] P. Kannadaguli, "A Code-Diverse Tulu-English Dataset for NLP-Based Sentiment Analysis Applications," in *Advanced Communication Technologies and Signal Processing Conference*, IEEE, 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9708241/

[6] V. Kadyan, A. Mantri, R. K. Aggarwal, and A. Singh, "A comparative study of deep neural network based Punjabi-ASR system," *Int. J. Speech Technol.*, vol. 22, no. 1, pp. 111–119, Mar. 2019, doi: 10.1007/s10772-018-09577-3.

[7] M. F. Mridha, A. Q. Ohi, M. A. Hamid, and M. M. Monowar, "A study on the challenges and opportunities of speech recognition for Bengali language," *Artif. Intell. Rev.*, vol. 55, no. 4, pp. 3431–3455, Apr. 2022, doi: 10.1007/s10462-021-10083-3.

[8] M. Narayanan and N. Aepli, "A Tulu Resource for Machine Translation," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, ELRA and ICCL, 2024, pp. 1756–1767. [Online]. Available: https://aclanthology.org/2024.lrec-main.155

[9] S. Bhatt, A. Jain, and A. Dev, "Acoustic Modeling in Speech Recognition: A Systematic Review," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, Jan. 2020, doi: 10.14569/IJACSA.2020.0110455.

[10] T. G. Fantaye, J. Yu, and T. T. Hailu, "Advanced Convolutional Neural Network-Based Hybrid Acoustic Models for Low-Resource Speech Recognition," *Computers*, vol. 9, no. 2, p. 36, 2020.

[11] D. GARG, "Compressed DNN based Automatic Speech Recognition Engine," PhD Thesis, INDIAN INSTITUTE OF TECHNOLOGY, MADRAS, 2019. Accessed: Dec. 13, 2024. [Online]. Available: https://eescholars.iitm.ac.in/sites/default/files/eethesis/ee14b081.pdf

[12] P. Mandal, S. Jain, G. Ojha, and A. Shukla, "Development of hindi speech recognition system of agricultural commodities using deep neural network.," in *INTERSPEECH*, 2015, pp. 1241–1245. Accessed: Dec. 13, 2024. [Online]. Available: https://www.academia.edu/download/111287189/mandal15_interspeech.pdf

[13] A. Bekarystankyzy, O. Mamyrbayev, M. Mendes, A. Fazylzhanova, and M. Assam, "Multilingual end-to-end ASR for low-resource Turkic languages with common alphabets," *Sci. Rep.*, vol. 14, no. 1, p. 13835, Jun. 2024, doi: 10.1038/s41598-024-64848-1.

[14] S. Mussakhojayeva, K. Dauletbek, R. Yeshpanov, and H. A. Varol, "Multilingual Speech Recognition for Turkic Languages," *Information*, vol. 14, no. 2, p. 74, Jan. 2023, doi: 10.3390/info14020074.

10.48047/jocaaa.2024.33.08.131

[15] C.-X. Qin, D. Qu, and L.-H. Zhang, "Towards end-to-end speech recognition with transfer learning," *EURASIP J. Audio Speech Music Process.*, vol. 2018, no. 1, p. 18, Dec. 2018, doi: 10.1186/s13636-018-0141-9.

[16] A. P. Rodrigues and R. Fernandes, "Tulu Language Text Recognition and Translation," in *IEEE Access Conference Proceedings*, IEEE, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10403880/

[17] S. Bhable, "Automatic Speech Recognition (ASR) of Isolated Words in Hindi low resource Language," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 9, pp. 260–265, Feb. 2021, doi: 10.22214/ijraset.2021.33011.

[18] M. G. Rao, C. Salonkar, D. A. Rai, D. D. Poojari, and J. Nithya, "Bilingual Translation Using a Novel Framework," presented at the International Conference on Recent Advances in Science and Engineering Technology, ICRASET 2023, 2023. doi: 10.1109/ICRASET59632.2023.10419913.

[19] S. G.-J. Wong and M. Durward, "cantnlp@LT-EDI-2024: Automatic Detection of Anti-LGBTQ+ Hate Speech in Under-resourced Languages," presented at the LT-EDI 2024 - 4th Workshop on Language Technology for Equality, Diversity, Inclusion, Proceedings of the Workshop, 2024, pp. 177–183. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85189854441&partnerID=40&md5=2c4e442361365ab9f96437c28172f0bc

[20] P. Bhargavi and G. Kanaka, "Development of dichotic digit test in Tulu," *Indian J. Otol.*, vol. 25, no. 2, pp. 71–75, 2019, doi: 10.4103/indianjotol.INDIANJOTOL_119_18.

[21] S. Chanda, A. Mishra, and S. Pal, "Advancing Language Identification in Code-Mixed Tulu Texts: Harnessing Deep Learning Techniques," presented at the CEUR Workshop Proceedings, 2023, pp. 223–230. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85193971140&partnerID=40&md5=19e0ab0788b1b19d6ea97a74b135b0b4

[22] P. J. Antony, H. B. Raj, B. S. Sahana, D. S. Alvares, and A. Raj, "Morphological analyzer and generator for Tulu language: A novel approach," presented at the ACM International Conference Proceeding Series, 2012, pp. 828–834. doi: 10.1145/2345396.2345531.

[23] P. Shetty, "Natural Language Processing for Tulu: Challenges, Review and Future Scope," presented at the Communications in Computer and Information Science, 2024, pp. 93–109. doi: 10.1007/978-3-031-58495-4_7.

[24] S. Patki, S. R. Priya Vedula, S. Ambekar, B. Vandana, and P. Preethi, "Tulu-Based Algorithmic Language Model," presented at the Proceedings - 2023 3rd International Conference on Pervasive Computing and Social Networking, ICPCSN 2023, 2023, pp. 1615–1621. doi: 10.1109/ICPCSN58827.2023.00270.

[25] K. K. Ponnusamy, C. Rajkumar, P. K. Kumaresan, E. Sherly, and R. Priyadharshini, "VEL@DravidianLangTech: Sentiment Analysis of Tamil and Tulu," presented at the DravidianLangTech 2023 - 3rd Workshop on Speech and Language Technologies for Dravidian Languages, associated with 14th International Conference on Recent Advances in Natural Language Processing, RANLP 2023 - Proceedings, 2023, pp. 211–216. doi: 10.26615/978-954-452-085-4_030.

[26] "Tulu Wikipedia." [Online]. Available: https://en.wikipedia.org/wiki/Tulu_language

[27] R. Prasad, "Language Technologies for Indian Languages: A Survey," *J. Indian Lang. Lit.*, vol. 1, no. 1, pp. 1–15, 2019.

[28] "Tulu Unicode Implementation." [Online]. Available: https://www.unicode.org/reports/tr45/

[29] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," in *Advances in Neural Information Processing Systems*, 2016, pp. 1–9.

[30] A. Graves and N. Jaitly, "Towards End-to-End Speech Recognition with Recurrent Neural Networks," in *International Conference on Machine Learning (ICML)*, 2014, pp. 1764–1772.

[31] K. Rao and M. Naik, "Dialectal Variation and Its Impact on Speech Recognition Systems," *J. Indian Lang. Technol.*, vol. 7, no. 4, pp. 100–115, 2019.

[32] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *ArXiv Prepr. ArXiv13013781*, 2013.

[33] A. Arora, A. Mattoo, D. Chaudhary, I. Gorton, and B. Kumar, "MEnTr@LT-EDI-2024: Multilingual Ensemble of Transformer Models for Homophobia/Transphobia Detection," presented at the LT-EDI 2024 - 4th Workshop on Language Technology for Equality, Diversity, Inclusion, Proceedings of the Workshop, 2024, pp. 259–264. [Online]. Available:

10.48047/jocaaa.2024.33.08.131

https://www.scopus.com/inward/record.uri?eid=2-s2.0-85189855870&partnerID=40&md5=0314d6ed1aa02af6bf2cbdaa8be2001b

[34] T. T. N. Do, M. B. D. Y. Nguyen, and D. S. N. Van Nguyen, "Speech Recognition Evaluation Metrics: A Review," *J. Comput. Sci. Technol.*, vol. 33, no. 3, pp. 553–565, 2018.

[35] Z. Yang, Y. Zhang, and W. Zhou, "Understanding the Importance of Data Diversity in Language Processing," *J. Artif. Intell. Res.*, vol. 68, pp. 100–123, 2020.

[36] A. C. L. Coates, M. S. D. Ward, and S. D. M. Smith, "Dialectal Variation and Speech Recognition: Challenges and Strategies," *Speech Commun.*, vol. 110, pp. 55–67, 2019.

[37] C. Lotfi, S. Srinivasan, M. Ertz, and I. Latrous, "Web Scraping Techniques and Applications: A Literature Review," 2021, pp. 381–394. doi: 10.52458/978-93-91842-08-6-38.

[38] D. N. S. Bhat, "Tulu," in *The Dravidian Languages*, 2015, pp. 158–177. doi: 10.4324/9780203424353-13.

[39] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[40] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[41] L. Cohen, *Time-Frequency Analysis: Theory and Applications*. Prentice Hall, 1995.

[42] X. Zhao, W. Liu, and X. Liu, "A Survey on Unicode: Past, Present, and Future," *Int. J. Comput. Sci. Netw. Secur.*, vol. 17, no. 4, pp. 9–17, 2017.

[43] B. H. Prasetio, D. O. Yusuf, D. Syauqy, and S. Chilmi, "Spectral Gating for Noise Reduction in Speech Stress Recognition System," in *2024 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, 2024, pp. 149–155. doi: 10.1109/IAICT62357.2024.10617782.

[44] G. M. Davis, *Noise reduction in speech applications*. CRC press, 2018.

[45] P. Chen, B. T. Nguyen, K. Iwai, and T. Nishiura, "Threshold-Based Combination of Ideal Binary Mask and Ideal Ratio Mask for Single-Channel Speech Separation," *Information*, vol. 15, no. 10, 2024, doi: 10.3390/info15100608.

[46] D. Beeferman, A. Berger, and J. Lafferty, "Statistical Models for Text Segmentation," *Mach. Learn.*, vol. 34, no. 1–3, pp. 177–210, 1999.

[47] I. Pak and P. L. Teh, "Text segmentation techniques: a critical review," *Innov. Comput. Optim. Its Appl. Model. Simul.*, pp. 167–181, 2018.

[48] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *Int. J. Mach. Learn. Cybern.*, vol. 1, pp. 43–52, 2010.

[49] J. Ramos and others, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, Citeseer, 2003, pp. 29–48.

[50] S. Hochreiter and J. Schmidhuber, "Long Short-term Memory," *Neural Comput.*, vol. 9, pp. 1735–80, Dec. 1997, doi: 10.1162/neco.1997.9.8.1735.

[51] N. Dehak, A. Elkhoury, and A. Ozerov, "Challenges in Speech Recognition for Low-Resource Languages," *IEEE Trans. Audio Speech Lang. Process.*, vol. 26, no. 6, pp. 1033–1043, 2018.

[52] G. Shorten and T. M. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[53] S. Dey and A. Ghosh, "Unicode and the Challenges in Speech Recognition for Indian Languages," *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 2, pp. 229–239, 2021.