JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Accent Bridge: Building Inclusive Voice Recognition System for Regional Indian Accents using Acoustic Models

P. Kamakshi Thai [1], **K Vinayak** [2], **M Aparna** [3], **M Madhu sudhan** [4]
[1]Assistant Professor of Department of CSE(AI & ML), ACE Engineering College, An Autonomous Institution, Ghatkesar, Hyderabad, Telangana, India.
[2,3,4] Students of Department of CSE(AI & ML), ACE Engineering College, An Autonomous Institution, Ghatkesar, Hyderabad, Telangana, India.

ABSTRACT

Most existing voice recognition systems are optimized for Western English accents, often failing to correctly interpret Indian regional pronunciations. Project, Accent Bridge, focuses on building an inclusive voice recognition system tailored to Indian accents such as Telugu, Tamil, and Bengali. The system is trained on real voice data from Indian speakers to capture unique acoustic features. By using tools like Python, TensorFlow, and pre-trained models such as Whisper, the project ensures accurate transcription of regional speech. Unlike current systems that misinterpret accents, Accent Bridge recognizes and processes them effectively. The project promotes inclusivity in AI by making voice technology accessible to people across all regions of India.

Keywords: Voice Recognition, Indian Accents, Acoustic Models, Speech Recognition, Inclusivity in AI, TensorFlow, Regional Languages

1. INTRODUCTION

Language diversity in India presents a unique challenge for modern speech recognition systems. While global voice assistants such as Siri, Alexa, and Google Assistant demonstrate impressive accuracy for Western accents, they often fail to recognize or interpret Indian regional pronunciations correctly. This gap in inclusivity restricts effective communication and limits accessibility for millions of Indian users. To overcome these challenges, advances in Artificial Intelligence (AI) and Machine Learning (ML) have made it possible to design specialized systems that can understand diverse accents. Accent Bridge is an AI-powered voice recognition system built specifically to address this issue by focusing on regional Indian accents such as Telugu, Tamil, and Bengali. The system leverages deep learning techniques and robust acoustic models to identify and process unique sound patterns from Indian speakers. By using real voice datasets and fine-tuning pre trained models like Whisper and wav2vec 2.0, the system aims to provide higher accuracy and inclusivity in voice-based applications.

2. LITERATURE SURVEY

[1] Title: ASR One-Tuning On Limited Domain-Specific Data (2024)**.**

Authors: Franco Mak et al.

This paper “ASR One-Tuning On Limited Domain-Specific Data” by Franco Mak et al.(2024) presents a domain-specific fine-tuning approach for improving Automatic Speech Recognition (ASR) systems, particularly in low-resource and specialized speech environments. The proposed method adapts large pre-trained models using small but high-quality datasets, focusing on practical strategies such as careful data selection, alignment, and noise reduction to enhance performance efficiently. Their system demonstrates that targeted fine-tuning significantly improves recognition accuracy while avoiding the need for large-scale retraining, thereby reducing computational cost and making the approach suitable for real-world deployment. The framework also increases robustness to accent variations, pronunciation differences, and background noise, which is especially beneficial for regional and domain-specific applications. However, the study notes limitations including dependence on dataset quality and challenges in generalizing across diverse accents. Despite these constraints, the research provides a strong foundation for building efficient, scalable, and adaptive ASR solutions, directly supporting the Accent Bridge project' s strategy of fine-tuning pre-trained models such as Whisper and wav2vec 2.0 on Indian accent datasets to achieve high accuracy with limited regional data.

[2] **Title**: Improving Low-Resource ASR Using Data Augmentation (2023)

Authors: Nay San et al.

This paper "Improving Low-Resource ASR Using Data Augmentation" by Nay San et al. (2023) presents an approach to enhancing Automatic Speech Recognition (ASR) performance for low-resource languages through transfer learning and data augmentation techniques. The proposed method leverages pre-trained speech models and adapts them to new languages and accents using limited datasets, reducing the dependency on large-scale labeled data. Their framework incorporates augmentation strategies such as noise injection, speed perturbation, and pitch variation to artificially expand training data and improve robustness against pronunciation diversity and environmental noise. The study demonstrates that combining synthetic augmented data with real speech samples significantly improves recognition accuracy and helps models generalize better to unseen accents and speaking conditions. However, the authors note challenges related to maintaining data quality and avoiding overfitting to artificial patterns. Despite these limitations, the research provides a practical and cost-effective solution for building reliable ASR systems in resource-constrained settings, strongly supporting the Accent Bridge project' s objective of developing accurate speech recognition models for Indian regional accents even when large datasets are unavailable.

[3] **Title**: ASR for Noisy Code-Mixed Low-Resource Speech (2023).

Authors: Ashutosh Modi et al.

This paper "ASR for Noisy Code-Mixed Low-Resource Speech" by Ashutosh Modi et al. (2023) introduces Low-Rank Multiplicative Adaptation (LoRMA), an efficient technique designed to adapt large Automatic Speech Recognition (ASR) models to new tasks, accents, and noisy environments with minimal computational overhead. The proposed method modifies only a small subset of model parameters instead of retraining the entire network, significantly reducing memory usage, training time, and resource requirements while maintaining high performance. Their framework is particularly effective for low-resource and code-mixed speech scenarios, where traditional fine-tuning can be costly and inefficient. Experimental results demonstrate that LoRMA can match or even outperform conventional full-model fine-tuning approaches while using fewer computational resources, making it suitable for scalable real-world deployment. However, the study highlights challenges in optimizing adaptation for highly diverse linguistic conditions. Despite these limitations, the research provides a practical and scalable solution for rapid ASR customization, strongly supporting the Accent Bridge project' s objective of efficiently adapting speech models to multiple Indian regional accents without full retraining, thereby enabling inclusive and resource-efficient voice recognition systems.

[4] **Title**: Multitask Adaptation With LF-MMI for Low-Resource ASR (2021).

Authors: Srikanth Madikeri et al.

This paper "Multitask Adaptation With LF-MMI for Low-Resource ASR" by Srikanth Madikeri et al. (2021) presents a sequence-discriminative training approach for improving Automatic Speech Recognition (ASR) performance in low-resource settings using Lattice-Free Maximum Mutual Information (LF-MMI). The proposed method focuses on building robust acoustic models through multilingual and multi-genre training, where speech data from different languages, accents, and environmental conditions are combined to enhance generalization. By learning shared linguistic patterns across multiple datasets, the system becomes more adaptable to variations in speaking style, pronunciation, and background noise. Their experiments demonstrate that this strategy significantly reduces word error rates (WER) and improves recognition accuracy, especially for unseen or diverse speech types. However, the study notes that managing heterogeneous multilingual data and training complexity can be challenging. Despite these limitations, the research provides an effective framework for developing scalable and resilient ASR systems, strongly supporting the Accent Bridge project’ s objective of handling multiple Indian regional accents through shared multilingual learning while maintaining accent-specific performance.

[5] **Title**: Code-Mixed Street Address Recognition and Accent Adaptation for Voice-Activated Navigation Services (2024).

Authors: Syed Meesam Raza Naqvi et al.

This paper by Syed Meesam Raza Naqvi et al. (2024) presents a hybrid Automatic Speech Recognition (ASR) system designed for recognizing Urdu-English code-mixed speech in low-resource environments. The proposed system combines acoustic modeling and language modeling with accent adaptation techniques to improve recognition accuracy for real-world navigation applications. They use TDNN-LSTM based deep neural networks along with task-specific datasets to handle accent variations and limited data conditions. Experimental results show significant reductions in word error rate compared to generic models, highlighting the effectiveness of customized ASR systems. This work strongly supports the Accent Bridge project, as both aim to develop accent-aware and application-specific speech recognition solutions for regional users.

[6] **Title**: wav2vec 2.0: Self-Supervised Learning for Speech Recognition (2020).

Authors: Alexei Baevski et al.

Alexei Baevski et al. focused on self-supervised learning for speech recognition through the wav2vec 2.0 framework, which learns powerful speech representations from large amounts of unlabeled audio data. Their approach reduces the need for expensive labeled datasets and enables efficient fine-tuning with small domain-specific data. The study demonstrates that pre-trained models can be easily adapted to new languages and accents while maintaining high accuracy. This method is particularly useful for low-resource scenarios and improves robustness to noise and pronunciation differences. These findings directly support the Accent Bridge system, which relies on fine-tuning wav2vec 2.0 for Indian accent speech recognition.

[7] **Title**: Robust Speech Recognition via Large-Scale Weak Supervision (Whisper) (2022).

Authors: Alec Radford et al.

Alec Radford et al. developed Whisper, a large-scale multilingual speech recognition model trained on diverse and weakly labeled datasets covering multiple languages, accents, and real-world environments. Their approach leverages large foundation models to improve generalization and robustness, enabling the system to handle noisy recordings, pronunciation differences, and accent variations effectively. The model demonstrates strong zero-shot performance across tasks without requiring extensive retraining, while further fine-tuning enhances domain-specific accuracy. The study emphasizes the benefits of large pre-trained models for building scalable and inclusive speech technologies. This work is highly relevant to Accent Bridge, as Whisper serves as a powerful baseline model that can be adapted to Indian accents through targeted fine-tuning for improved regional speech recognition.

[8] **Title**: Kaldi Speech Recognition Toolkit (2011).

Authors: Daniel Povey et al.

Daniel Povey et al. introduced the Kaldi Speech Recognition Toolkit, an open-source platform widely used for developing state-of-the-art ASR systems using hybrid modeling techniques such as GMM-HMM and DNN-HMM. The toolkit provides comprehensive modules for feature extraction, acoustic modeling, pronunciation dictionaries, language modeling, and decoding, allowing researchers to build flexible and customizable recognition systems. Kaldi supports multilingual and low-resource speech applications by enabling efficient training pipelines and integration of diverse datasets. Their framework has become a standard tool for both academic research and industrial deployment of speech technologies. This toolkit is particularly useful for the Accent Bridge project, as it facilitates the implementation and experimentation of accent-specific acoustic models and hybrid architectures.

[9] **Title**: Convolutional Neural Networks for Large-Vocabular y Speech Recognition (2013).

Authors: Tara N. Sainath et al.

Tara N. Sainath et al. explored the application of deep neural network architectures, especially convolutional neural networks, for improving acoustic modeling in large-vocabulary speech recognition tasks. Their research shows that CNN-based models are capable of capturing local spectral and temporal speech features more effectively than traditional methods, leading to reduced word error rates and improved robustness to noise and speaker variability. By learning hierarchical feature representations, these models better handle pronunciation differences and accent variations across speakers. The study highlights the importance of deep learning techniques for modern ASR systems operating in diverse real-world environments. These insights are valuable for Accent Bridge, as advanced neural architectures can enhance recognition accuracy for Indian regional accents

[10] **Title**: ESPnet an end-to-end speech processing toolkit (2018).

Authors: Shinji Watanabe et al.

Shinji Watanabe et al. proposed ESPnet, an end-to-end speech processing toolkit that integrates acoustic and language modeling within a unified neural network framework. Unlike traditional hybrid systems that require separate components, their approach directly maps speech signals to text using sequence-to-sequence learning, simplifying the training and deployment process. The toolkit supports multilingual speech recognition, transfer learning, and data augmentation, making it suitable for low-resource and domain-specific applications. Their experiments demonstrate competitive performance with reduced system complexity and faster scalability. Although careful tuning is necessary for optimal accuracy, the framework provides an efficient alternative to conventional pipelines. This approach is relevant to Accent Bridge as it offers an end-to-end solution for developing adaptable and scalable speech recognition systems for multiple Indian accents.

2.1 Comparison Table

S. No	**Authors(s)**	**Title**	**Proposed Methodology**	**Findings from the Reference Paper**
1	Franco Mak et al.	ASR One-Tuning on Limited Domain-Specific Data (2024)	Fine-tuned large pre-trained ASR models using small, high-quality domain-specific datasets with data selection and noise reduction techniques.	Improved recognition accuracy with minimal retraining, reduced computational cost, effective for low-resource and specialized speech tasks
2	Nay San et al.	Improving Low-Resource ASR Using Data Augmentation (2023)	Applied transfer learning with augmentation techniques such as noise addition, speed perturbation, and pitch variation to expand training data.	Synthetic+ real data improved generalization and accuracy, suitable for low-resource languages and accents
3	Ashutosh Modi et al.	ASR for Noisy Code-Mixed Low-Resource Speech(2023)	Introduced Low-Rank Multiplicative Adaptation (LoRMA) to update only small model parameters instead of full retraining.	Achieved performance similar to full fine-tuning with lower memory and computation, enabling scalable accent adaptation.
4	Srikanth Madikeri et al.	Multitask Adaptation with LF-MMI for Low-Resource ASR(2021)	Used sequence-discri minative LF-MMI training with multilingual and multi-genre speech data for robust acoustic modeling.	Reduced Word Error Rate (WER) and improved generalization across accents and unseen speech conditions.

5	Syed Meesam Raza Naqvi et al.	Code-Mixed Street Address Recognition and Accent Adaptation for Voice-Activated Navigation Services(2024)	Developed hybrid ASR using TDNN-LSTM acoustic models and task-specific code-mixed datasets with accent adaptation.	Significant WER reduction in low-resource and code-mixed speech, proved effectiveness of customized domain-specific ASR systems.
6	Alexei Baevski et al.	wav2vec 2.0: Self-Supervised Learning for Speech Recognition (2020)	Pre-trained speech representations using large unlabeled audio and fine-tuned with small labeled datasets.	Reduced dependency on labeled data, strong performance in low-resource and multilingual environments.
7	Alec Radford et al.	Robust Speech Recognition via Large-Scale Weak Supervision (Whisper) (2022).	Trained large foundation ASR model on massive multilingual and multi-acoustic datasets using weak supervision.	High robustness to noise, accents, and real-world speech, strong zero-shot and fine-tuned performance.
8	Daniel Povey et al.	Kaldi Speech Recognition Toolkit(2011)	Provided modular open-source toolkit supporting GMM-HMM and DNN-HMM hybrid architectures with flexible training pipelines.	Enabled easy development of customizable, scalable, and research-oriented ASR systems for multiple languages.
9	Tara N. Sainath et al.	Convolutional Neural Networks for Large-Vocabular y Speech Recognition (2013)	Applied CNN-based deep learning for improved acoustic feature extraction and noise robustness.	Lower WER and better handling of speaker and accent variations compared to traditional models.
10	Shinji Watanabe et al.	ESPnet: End-to-End Speech Processing Toolkit(2018)	Developed end-to-end neural framework integrating acoustic language modeling with transfer learning.	ASR and Simplified training deployment, competitive accuracy, suitable for multilingual and low-resource speech tasks.

3. RESEARCH GAPS IN EXISTING SYSTEMS:

Based on the literature review, several important research gaps have been identified in existing speech recognition systems, particularly in relation to regional accent recognition. Although modern Automatic Speech Recognition (ASR) technologies have achieved significant improvements with the help of deep learning and large-scale datasets, many systems are still primarily designed for standard accents and widely spoken languages.

3.1 Lack of Regional Datasets

One of the major challenges in developing accurate voice recognition systems for Indian users is the lack of large and we l-structured datasets that represent regional accents. Many existing speech datasets mainly contain standard or widely spoken forms of languages, while regional variations such as Telugu, Tamil, Bengali, and other dialects are underrepresented. Because of this limitation, machine learning models cannot learn the unique pronunciation patterns and speech characteristics of different regions. As a result, speech recognition systems often produce incorrect outputs when users speak with strong regional accents. Creating diverse datasets that include speakers from multiple regions, age groups, and backgrounds is therefore essential to improve the accuracy and inclusiveness of speech recognition technologies.

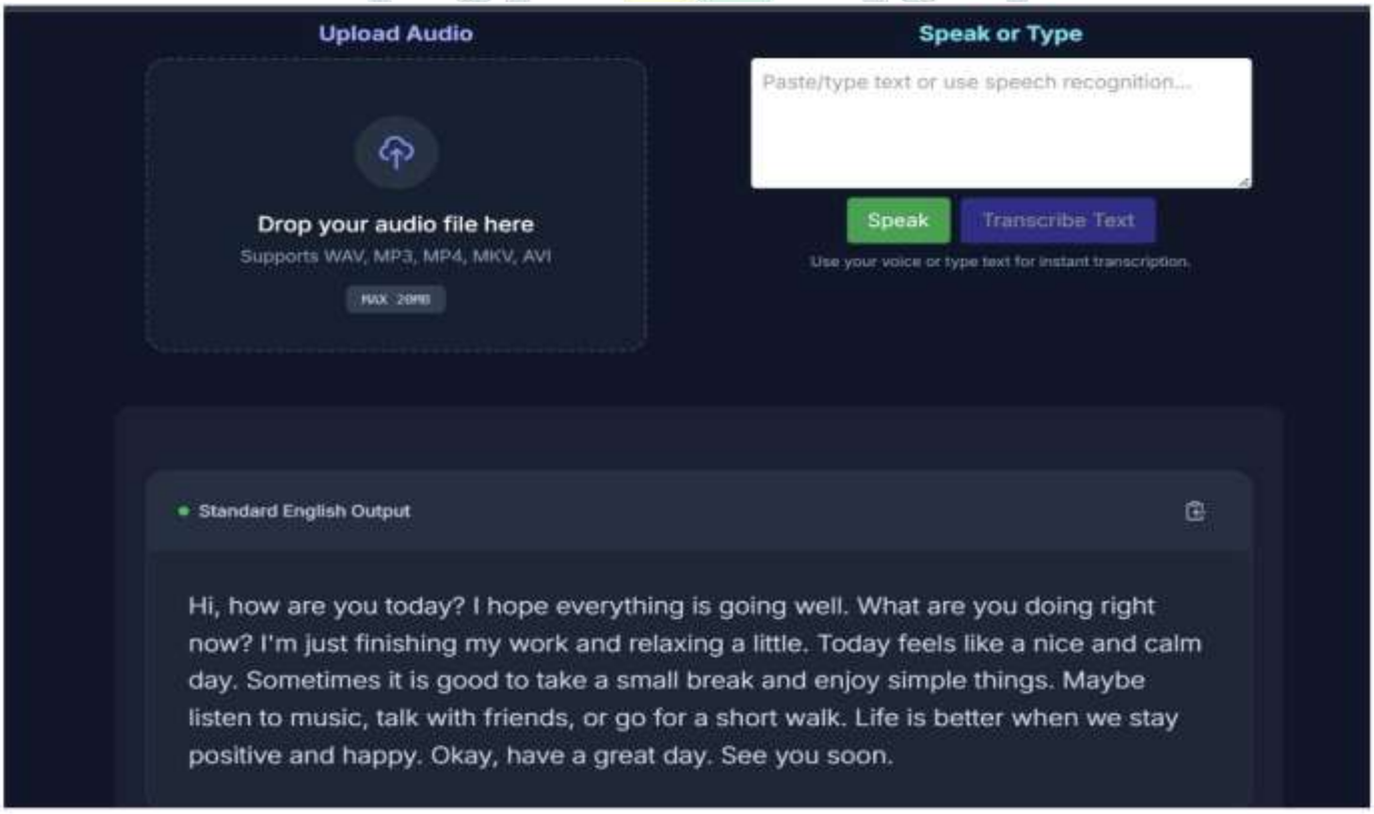
3.2 Accent Variation Within Languages

Another important research gap is the large variation in pronunciation that exists within the same language across different regions. For example, Telugu spoken in Andhra Pradesh may sound different from Telugu spoken in Telangana, and similar variations can be observed in many other Indian languages. Most existing Automatic Speech Recognition (ASR) systems are designed to recognize only a standard or neutral accent, which makes them less effective when users speak with local pronunciation patterns.

3.3 Real-World Environment Challenges

Many speech recognition models are developed and tested using clean audio recordings captured in controlled environments such as studios or quiet rooms. However, real-world conditions are very different and often include background noise, overlapping conversations, traffic sounds, and other disturbances. When users interact with voice systems in such environments, the performance of the recognition system can significantly decrease. This creates a gap between laboratory performance and practical usability. To make speech recognition systems more reliable, research needs to focus on training models with noisy and real-world audio data so that they can accurately process speech even in challenging environments.

4. BACKGROUND AND FUNDAMENTALS


Upload Audio
Drop your audio file here
Supports WAV, MP3, MP4, MKV, AVI
MAX 20MB
Speak or Type
Paste/type text or use speech recognition...
Speak
Transcribe Text
Use your voice or type text for instant transcription.
Standard English Output

Hi, how are you today? I hope everything is going well. What are you doing right now? I'm just finishing my work and relaxing a little. Today feels like a nice and calm day. Sometimes it is good to take a small break and enjoy simple things. Maybe listen to music, talk with friends, or go for a short walk. Life is better when we stay positive and happy. Okay, have a great day. See you soon.

Figure 1: System Over view of the Accent Bridge Voice Recognition System

4.1 Regional Accent Variations in Speech Recognition Systems

Regional accent diversity presents a significant challenge for the development of reliable speech recognition systems. In India, the pronunciation of words varies widely across different states and regions due to differences in dialects, linguistic backgrounds, and cultural influences. For instance, the pronunciation of Telugu words may differ between speakers from Andhra Pradesh and Telangana. Similar pronunciation variations also exist in languages such as Hindi, Tamil, and Bengali.

4.2 Acoustic Modeling for Speech Recognition

Acoustic modeling plays a key role in the functioning of speech recognition systems. It focuses on identifying the relationship between the acoustic characteristics of speech signals and the phonetic components of a language. When a user speaks into a device, the audio signal is captured and converted into a digital format. The system then processes the audio signal to extract meaningful features such as frequency components, pitch variations, and energy levels.

4.3 Speech Recognition Workflow for Accent-AwareSystems

An effective accent-aware speech recognition system fo lows a structured workflow that ensures accurate processing of spoken input. The process begins when a user provides speech input through a microphone or voice-enabled device. The captured audio signal is first preprocessed to reduce background noise and enhance speech clarity. This step is important because real-world environments often contain disturbances that can affect recognition accuracy.

5. METHODOLOGY:

5.1 System Architecture and Web Application Design

The proposed Accent Bridge system is designed using a modular architecture that integrates speech input processing, acoustic modeling, and text generation modules. The architecture consists of three major components: the user interaction layer, the speech processing layer, and the machine learning model layer. These components work together to capture speech input, analyze the audio signal, and generate accurate text output. The user interaction layer provides the interface through which users interact with the system. This interface alows users to provide voice input through a microphone and view the transcription results generated by the system. The application interface can be developed using web technologies such as HTML, CSS, and JavaScript, which provide a simple and accessible platform for users.

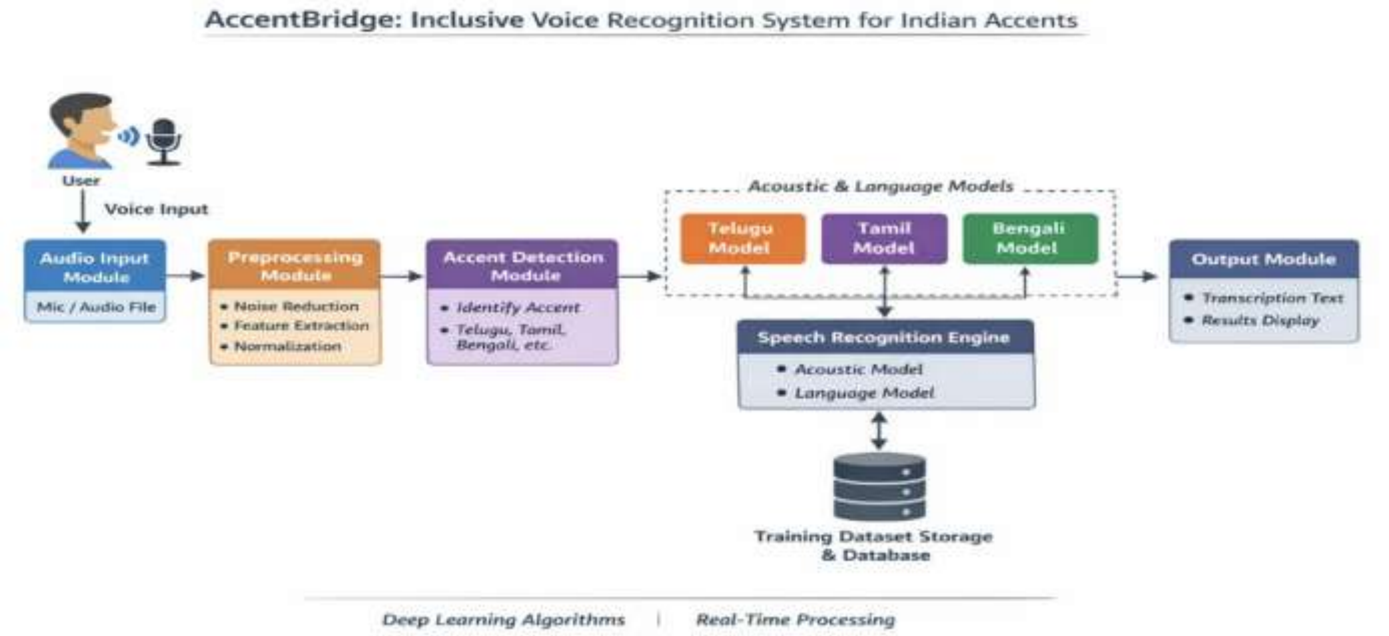


Figure 2: Proposed Architecture

5.2 Speech Data Collection and Accent Dataset Preparation

The first step in building an accent-aware speech recognition system is the collection of diverse speech datasets. In the Accent Bridge system, speech samples are collected from speakers belonging to different regions and linguistic backgrounds. These samples include variations in pronunciation, speaking speed, and intonation patterns that represent real-world accent diversity. After collecting the speech recordings, the data is carefully organized and labeled. Each audio sample is associated with its corresponding text transcript, which a lows the model to learn the relationship between spoken sounds and written words.

5.3 Acoustic Model Training Using MachineLearning

The core component of the Accent Bridge system is the acoustic model that learns the relationship between audio signals and phonetic units of language. During the training phase, the processed speech dataset is used to train machine learning models capable of identifying speech patterns and mapping them to corresponding textual outputs. Feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCC) are applied to represent speech signals in a format suitable for machine learning algorithms. These features capture important characteristics of human speech such as frequency distribution and energy levels. The extracted features are then used to train deep learning models, including neural networks or transformer-based architectures. Pretrained speech recognition models such as Whisper or wav2vec 2.0 can also be fine-tuned using regional accent datasets. Fine-tuning allows the system to adapt the pretrained model to better understand pronunciation variations found in Indian regional accents. This process improves the model’ s ability to accurately recognize speech from diverse speakers.

5.4 Speech Recognition and Accent Adaptation Workflow

The speech recognition process begins when a user provides voice input through the system interface. The audio signal is captured and passed through the preprocessing module, where noise reduction and audio enhancement techniques are applied. After preprocessing, the system extracts relevant speech features that represent the characteristics of the spoken signal. These extracted features are then fed into the trained acoustic model, which analyzes the speech patterns and predicts the most likely sequence of words. Because the model is trained using accent-diverse datasets, it is capable of recognizing different pronunciation styles and regional variations. The predicted text output is then generated and displayed to the user in real time.

6. CHALLENGES AND LIMITATIONS

6.1 Scalability and Computational Performance

The proposed Accent Bridge system relies on machine learning models and speech processing algorithms to recognize spoken language from various regional accents. While the system performs effectively for small to medium datasets, scalability becomes a challenge when handling large volumes of speech data or supporting a large number of users simultaneously. Training and fine-tuning acoustic models require high computational resources, including powerful processors and large memory capacity. As the number of speech samples and users increases, the system may experience longer processing times during model training and speech transcription. In large-scale deployments, optimizing model efficiency and using scalable cloud-based infrastructure may be necessary to maintain consistent performance.

6.2 Accent Diversity and Dataset Limitations

One of the major challenges in developing an accent-aware speech recognition system is the availability of diverse and balanced datasets. Indian languages exhibit significant regional variation in pronunciation, tone, and speech rhythm. However, publicly available speech datasets often lack sufficient representation of many regional accents. As a result, the trained models may perform well for certain accents but struggle with others that are underrepresented in the training data. This imbalance can affect the accuracy and fairness of the system. Collecting high-quality speech recordings from different regions and dialect groups remains an essential but challenging task for improving accent recognition.

6.3 Noise and Real-World Speech Conditions

Speech recognition systems often face difficulties when operating in real-world environments where background noise and disturbances are common. Factors such as traffic noise, overlapping conversations, poor microphone quality, and environmental sounds can degrade the clarity of speech signals. Although preprocessing techniques such as noise reduction and filtering can improve audio quality, they may not completely eliminate these disturbances. As a result, the system may produce incorrect transcriptions or fail to recognize certain words when speech input is captured in noisy conditions. Developing robust noise-handling techniques and training models using real-world audio data are important steps toward improving system reliability.

6.4 Adoption and Integration Challenges

The successful implementation of accent-aware speech recognition systems depends on their adoption by developers, organizations, and end users. Integrating such systems into existing applications and digital platforms may require additional technical effort, including compatibility with different software frameworks and hardware devices. Some organizations may face challenges related to infrastructure requirements, data collection, and model training. In addition, concerns related to data privacy and the co lection of voice recordings from users may influence the willingness of individuals to participate in speech data collection initiatives. Addressing these concerns through clear policies, secure data handling practices, and user-friendly system design will be essential for wider adoption of the technology.

7. CONCLUSION AND FUTURE SCOPE

The Accent Bridge system focuses on improving the inclusiveness and effectiveness of speech recognition technology by addressing the challenges related to regional Indian accents. Traditional speech recognition systems often struggle to accurately interpret speech when users speak with diverse pronunciation patterns. This limitation arises mainly due to insufficient representation of regional accents in the training datasets used by most models. The proposed system aims to reduce this gap by incorporating accent-diverse speech data and applying acoustic modeling techniques to better understand variations in pronunciation. Although the proposed Accent Bridge system provides a foundation for accent-aware speech recognition, there are several opportunities for further improvement and expansion. One possible direction for future work is the creation of larger and more diverse speech datasets representing multiple regional accents across India. Collecting speech samples from a wider population will help improve the accuracy and generalization ability of the recognition models.

8. REFERENCES

1. A. Rouhe, T. Grósz, and M. Kurimo, ‘ ‘ Principled comparisons for end to-end speech recognition: Attention vs hybrid at the 1000-hour scale,’ ’ IEEE/ACM Trans. Audio, Speech, Lang., Process., vol. 32, pp. 623– 638, 2024.

2. M. A. Ashraf, R. M. A. Nawab, and F. Nie, ‘ ‘ Tran-switch: A transfer learning approach for sentence level cross-genre author profiling on code switched English– RomanUrdu text,’ ’ Inf. Process. Manage., vol. 60, no. 3, May 2023, Art. no. 103261.

3. M. A. Hassan, A. Rehmat, M. U. G. Khan,and M. H. Yousaf, ‘ ‘ Improvement in automatic speech recognition of south Asian accent using transfer learning of DeepSpeech2,’ ’ Math. Problems Eng., vol. 2022, pp. 1– 12, Oct. 2022.

4. K. Nowakowski, M. Ptaszynski, and K. Murasaki, ‘ ‘ Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining,’ ’ Inf. Process. Man age., vol. 60, no. 2, Mar. 2023, Art. no. 103148.

5. A. Imtiaz, M. Rashid, S. Abid Syed, H. Zahid, M. Iqbal, and A. A. Khan, ‘ ‘ Automatic speech recognition on non-pathological dataset of Urdu language,’ ’ KIET J. Comput. Inf. Sci., vol. 5, no. 2, Jul. 2022.

6. A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “ wav2vec 2.0: A framework for self-supervised learning of speech representations,” Adv. Neural Inf. Process. Syst., vol. 33, pp. 12449– 12460, 2020.

7. A. Radford et al., “ Robust speech recognition via large-scale weak supervision,” Proc. Int. Conf. Mach. Learn., 2022.

8. S. Sreeram, P. Ghahremani, V. Manohar, and S. Khudanpur, “ Spoken language recognition for Indian languages using deep learning,” IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), pp. 6294– 6298, 2018.

9. N. Jain, R. Goyal, and K. Bali, “ Improving automatic speech recognition for code-mixed Indian languages using transfer learning,” Proc. Interspeech, pp. 4054– 4058, 2020.

10. K. Bali, J. Sharma, M. Choudhury, and Y. Vyas, “ Chalenges in Indian language speech recognition,” ACMComput. Surveys, vol. 52, no. 3, pp. 1– 34, 2019