EXPLORING RESNET50 ALGORITHM for HINDI BRAILLE SPEECH ANALYSIS

Dimpal Sahu , research scholar , Dept.Of Computer Science, SAGE university.College, indore
Dr.Ati Jain, Associate Professor, Dept.Of Computer Science, SAGE university.College, indore

Abstract

This paper presents a novel approach for analyzing Hindi Braille data, leveraging a comprehensive methodology that integrates speech-to-text transformation, feature extraction using the RESNET50 algorithm, and a classification model for detecting Hindi Braille speech. The dataset, initially provided in .mp3 or wav formats, undergoes crucial preprocessing steps including audio to text conversion. An 80-20 data splitting strategy is employed for training and testing, respectively. A notable aspect of our approach is the utilization of the RESNET50 algorithm, primarily recognized for its excellence in image recognition tasks. However, in this context, we adapt it for audio-related objectives, particularly for feature extraction from the converted text. The extracted features serve as inputs to a classifier designed specifically for detecting Hindi Braille speech patterns. One intriguing but somewhat ambiguous step in our methodology involves the conversion of text into an image. While the purpose of this step remains unclear within the provided description, it suggests a potential avenue for further exploration and refinement in future research. Our proposed system offers a promising framework for analyzing Hindi Braille speech audio data, with the potential to contribute to advancements in assistive technologies for visually impaired individuals. We believe that this research will stimulate further investigation into the adaptation of computer vision algorithms for audio processing tasks, opening new possibilities for cross-disciplinary research and innovation.

Keywords: Hindi Braille, RESNET50, Classification model, Feature extraction

Introduction

Speech is a multi-component signal with varying time, frequency and amplitude”[1]. Human beings engage with one another or communicate using Speech. In comparison with non-Indian languages, there has been a rare study done on Indian languages. Many researchers throughout the globe strive to design a new human-computer interface system with optimum precision. Speech may be an essential way of computer interface. Speech Recognition is the technique to process a voice in words or text form. The voice of speech data uttered by people is transformed into electronic signals using speech recognition technology. Then these signals are translated into a pattern of coding, and desired significance is achieved. Speech recognition is essentially the technique by which the specific speaker may be identified with the use of speech wave information. Researchers have been working on language recognition throughout the previous sixty years. Now ASR identifies an app that needs human-machine interaction and can talk and speak the language in their mother tongues [2-3].

Speech is a natural language spoken by a human being for communication. The technical definition of Speech is an amplitude, frequency, and multi-component signal, with varying times. Because of this heterogeneity, transitions in distinct frequency bands may occur at various times. A speech recognition system is a technique used to extract, recognize and translate the speaking features utilizing intelligent electronic equipment. The System's primary objective is to build a technology for human interaction, independent of vocabulary, sound, Speech, or accents, in our natural language in real-time. For single-word recognition, continuous speech recognition, and spontaneous speech recognition, this System may be applied. It has used for young children, for telecommunications, for persons with hearing impairment.[4]Speech recognition systems like Google Voice Search that support around 120 languages, made remarkable improvements in recent years. It is important to academics and businesses to extend its coverage to global languages. Speech Recognition systems for Voice Detection, Communication, and Control systems have increased enormously to improve the user experience and make the System effective and efficient. The critical phases involved in speech recognition systems are pre-processing feature extraction and classification/ pattern recognition. Due to the surrounding noise and other kinds of problems, disorders in speech signals increase, affecting the performance of the speech recognition system. Accuracy of recognition is utilized in the speech recognition system for the performance assessment [5].

The microscopic study has been carried out in the case of Indian languages compared to non-Indian languages. As extensive research has been done in other developed nations about Speech Recognition Systems, the quantity of work done in Indian regional languages has not yet reached a threshold level that makes it a meaningful communication tool. Marathi is a language spoken in western and central India, i.e., Maharashtra and some parts of Madhya Pradesh, Gujrat, and Karnataka. Ongoing research is, in a way, on the isolated identification of words spoken in local languages in various locations of India. Speech is the most noticeable and natural mode of interpersonal communication. The globe is conceived of in many spoken languages. There are, nevertheless, many possibilities for developing systems that use Indian languages that vary [6].

Words recorded by the speaker, the microphone, and the telephone converted to an auditory signal are the primary purpose of language recognition. An area of the ongoing investigation is isolated, extracting Marathi words to identify and validate each word uttered[7]. These Indian-Language systems are in their early years due to various obstacles, including resource deficiencies, despite the continuing improvement of Automatic Space Recognition (ASR) technology.

Some multi-lingual Acoustic Models (AM) need a standard telephone set. others include characteristics of the input noise. Speech recognition is the most advanced method. It employs neural network approaches based on voice recognition solutions. There has already been significant development in this field; however, the System's resilience is essential. For real-time responsiveness, many training techniques are used. Most works on the identification of multi-lingual Speech were restricted to the multi-legalization of the acoustic model (AM) [8].

The lower layers are shared throughout languages, with languages particular to the lower neural network (DNN) output layer. Alternatively, multi- lingual bottlenecks may be employed for either a Gaussian or DNN mixing model using a DNN function extractor. Since then, GMM and HMM have been used to construct many speech recognition applications. Later on, ASR was integrated into the machine. In addition to this, several Indian languages have been introduced, e.g., Bengali, Malayalam, Marathi, Hindi, Gujarati, Telugu, Bodo, Kannada, Punjabi, and Tamil. Researchers have recently developed a new methodology to assist more profound learning algorithms to simulate the spectrum fluctuations. Then, the use of profound learning in language recognition has grown enormously. Several techniques to deep learning have been documented, including deep belief (Baker, 2013), deep recurring neural networks [7], deep convolution neural convolution network (DCNN) (17)(18), and HMM hybrid Convolution Neural Network (CNN). However, ASR has a great deal to develop; Researchers are working to build an effective voice recognition system.[9]

The objective of speech recognition is to create an ideal system for recognizing a sequence of words subject to language restrictions. The word is made of vowels and consonant linguistic units. A sentence model is supposed to be a succession of smaller unit models in recognition of Speech.[10]The acoustic proofs of these components are paired with the principles for building valid, intelligible phrases.

Literature

Mohammad Soleymanpour et.al. (2022)[11] People with dysarthria often have trouble understanding what others are saying because their speech muscles aren't working together properly. Technical advances in automatic speech recognition (ASR) might help people with dysarthria communicate better. Effective dysarthria-specific active speech recognition (ASR) requires a lot of training speech that is not easily been accessible. Text-to-Speech (TTS) synthesis multi-speaker end-to-end systems that have improved recently have raised the possibility of using synthesis to add to data. For better training of a dysarthria-specific deep neural network-homomorphism machine learning automatic speech recognition (DNN-HMM) system, we want to create multi-speaker end-to-end text-to-speech (TTS) systems that can recreate dysarthric speech. Dissarthria severity level and pause insertion machinery are added to the synthesized speech to change other control parameters, such as pitch, energy, and length. A DNN-HMM model trained on extra simulated dysarthric speech improved the WER by 12.2% compared to the baseline, according to the results. Additionally, adding the severity level and pause insertion controls led to a 6.5% decrease in WER, which suggests that these factors were added successfully.

Zhaoxu Nian et.al. (2021) [12] this article suggests a PL-ANSE method for speech preprocessing in noisy speech recognition that is based on progressive learning and adaptive noise and speech estimates. An improved minima controlled recursive averaging (IMCRA) frame-level noise tracking feature and an utterance-level deep progressive learning of nonlinear relationships between speech and noise are used in this method. Using this way should help computers understand what people are saying even when there is a lot of noise. To start, a two-way long-short-term memory model is set up at every network layer so that progressive ratio masks (PRMs) can be learned as targets whose signal-to-noise ratios gradually rise. For better voice quality, the estimated PRMs at the utterance level are then added to a common speech enhancement algorithm at the frame level. The last step is to boost the performance of a speech recognition system by directly feeding it the better speech that was made by combining data from different levels. The experiments show that our suggested method can lower the relative word error rate (WER) by 22.1% compared to the results with unprocessed noisy speech, which were between 23.84% and 18.57% on the CHiME-4 single-channel real test data. [12]

Yuanchao Li et.al. (2022)[13]Instead of just hearing something, Speaking Emotion Recognition (SER) can also learn from language features found in speech transcripts. Due to the small amount of data labeled with emotions and the difficulty of identifying emotional speech, it is hard to make reliable linguistic features and models in this area of study. Automatic Speech Recognition (ASR) results should be added to the pipeline for joint training SER. This is what this study suggests. It's not clear what parts of ASR help SER or how they help because not enough study has been done on the relationship between ASR and SER. The best way to make the SER work better, according to our tests, is to use a hierarchical co-attention fusion approach to combine both ASR hidden and text output. A number of ASR results and fusion methods were looked at to find this out. We got a 63.4% weighted accuracy rate on the IEMOCAP corpus by combining ground-truth transcripts. This is pretty close to the results we got by mixing transcripts in the first place. So that you can better understand the link between ASR and SER, we also present a new word error rate study on IEMOCAP and a layer-difference analysis of the Wav2vec 2.0 model .

PROPOSED SYSYTEM

The proposed system leverages a comprehensive approach for the analysis of Hindi Braille speech audio data. Initially, the dataset, provided in .mp3 or wav formats, undergoes a crucial conversion process from audio to text, emphasizing speech-to-text transformation. To evaluate the system's performance, an 80-20 data splitting strategy is employed, allocating 80% for training and 20% for testing. A notable aspect is the utilization of the RESNET50 algorithm, primarily recognized for its excellence in image recognition tasks, yet intriguingly applied to audio-related objectives in this context. The algorithm serves as a key component for feature extraction, which subsequently feeds into a classifier designed for detecting Hindi Braille speech. An additional but somewhat ambiguous step involves the conversion of text into an image, the purpose of which remains unclear within the provided description.

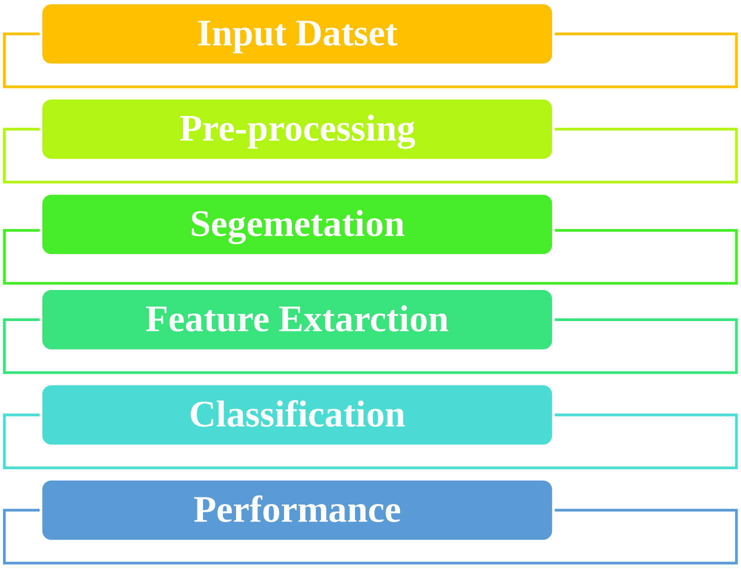


Fig.1 proposed flow diagram

In the performance estimation phase, the system's effectiveness is meticulously evaluated using a suite of metrics. These include accuracy, measuring the overall correctness of the model; precision, assessing the ratio of true positives to the sum of true positives and false positives; recall, evaluating the ratio of true positives to the sum of true positives and false negatives analysis, providing a graphical representation of the model's discriminatory ability; confusion metrics, presenting a detailed table of true positives, true negatives, false positives, and false negatives; and a comprehensive classification report, summarizing various classification metrics such as precision, recall, and F1-score for each class. This systematic and metric-rich approach ensures a thorough evaluation of the proposed system's performance in Hindi Braille speech detection and classification.

Module Description

Input: The input consists of the Hindi Braille speech audio dataset, provided in the formats .mp3 or wav.

Conversion from Audio to Text: The primary goal is to convert audio data into text, implying a speech-to-text conversion.

Data Splitting: The dataset is divided into 80% for training and 20% for testing purposes.

Algorithm Used: RESNET50 algorithm is employed. RESNET50 is a deep learning model known for its performance in image recognition tasks. It's interesting that it's being used for audio-related tasks in this case.

Classification:A classifier is implemented for the detection of Hindi Braille to speech. It's not explicitly mentioned, but it's inferred that the classifier uses the features extracted from the RESNET50 model.

Text to Image Conversion: There is a step where text is converted into an image. It's not entirely clear how this step fits into the overall process, but it suggests a transformation from textual data to an image format.

Performance Estimation: This step involves analyzing the performance of the implemented system. Performance metrics include:

Accuracy: The overall correctness of the model.

Precision: The ratio of true positives to the sum of true positives and false positives.

Recall: The ratio of true positives to the sum of true positives and false negatives.

ROC (Receiver Operating Characteristic): A graphical representation of the model's ability to discriminate between positive and negative classes.

Confusion Metrics: A table that shows true positives, true negatives, false positives, and false negatives.

Classification Report: A summary of various classification metrics, usually including precision, recall, and F1-score for each class.[14-15]

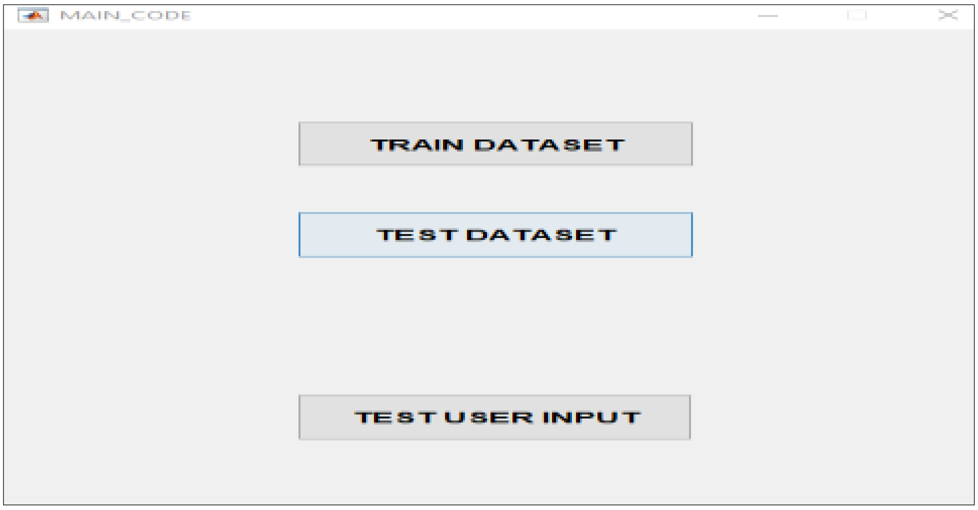
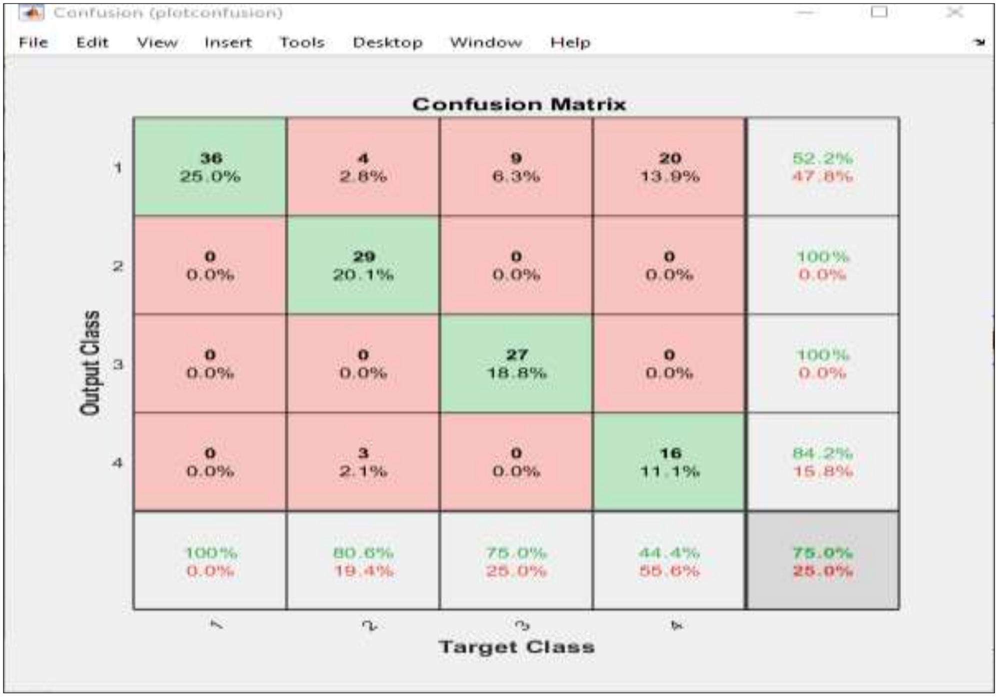
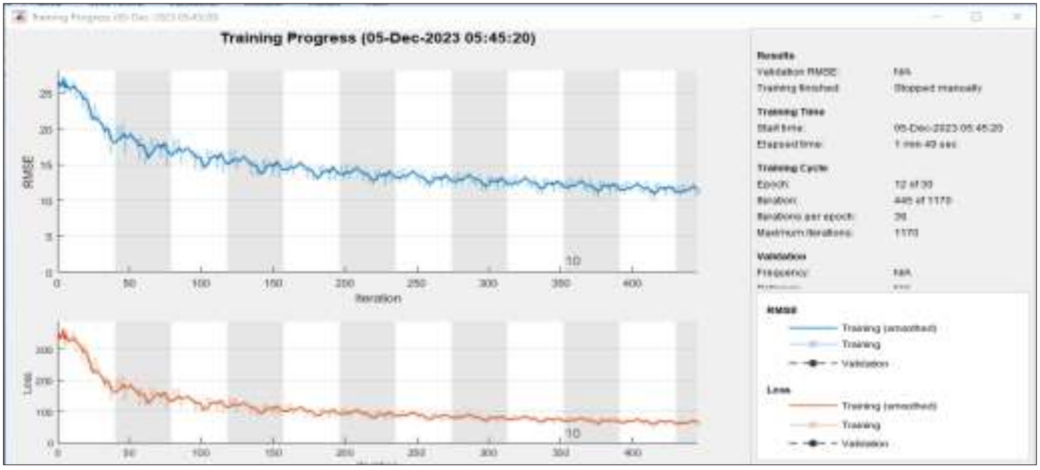


Fig.2 main GUI



	1	2	3	4	
1	36 25.0%	4 2.8%	9 6.3%	20 13.9%	52.2% 47.8%
2	0 0.0%	29 20.1%	0 0.0%	0 0.0%	100% 0.0%
3	0 0.0%	0 0.0%	27 18.8%	0 0.0%	100% 0.0%
4	0 0.0%	3 2.1%	0 0.0%	16 11.1%	84.2% 15.8%
	100% 0.0%	80.6% 19.4%	75.0% 25.0%	44.4% 55.6%	75.0% 25.0%

Fig. 3 confusion matrix



Results
Validation RMSE: N/A
Training finished: Stopped manually

Training Time
Start time: 05-Dec-2023 05:45:20
Elapsed time: 1 min 40 sec

Training Cycle
Epoch: 12 of 30
Iteration: 445 of 1170
Iterations per epoch: 39
Maximum iterations: 1170

Validation
Frequency: N/A

Fig. 4 Data training process

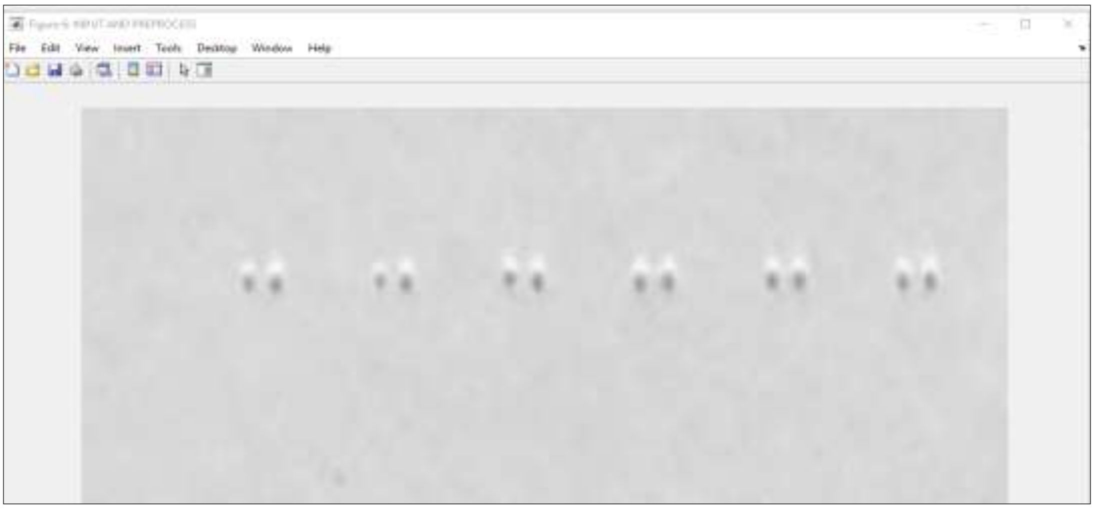


Figure 6: INPUT AND PREPROCESS
File Edit View Insert Tools Desktop Window Help

Fig.5 input data

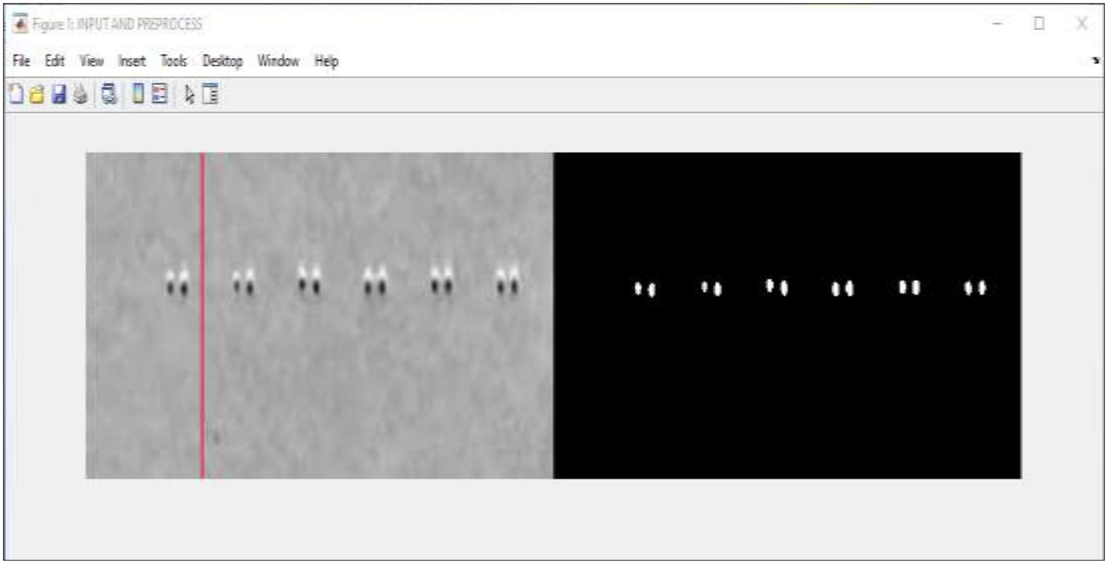


Figure 1: INPUT AND PREPROCESS
File Edit View Insert Tools Desktop Window Help

Fig.6 input and pre-process

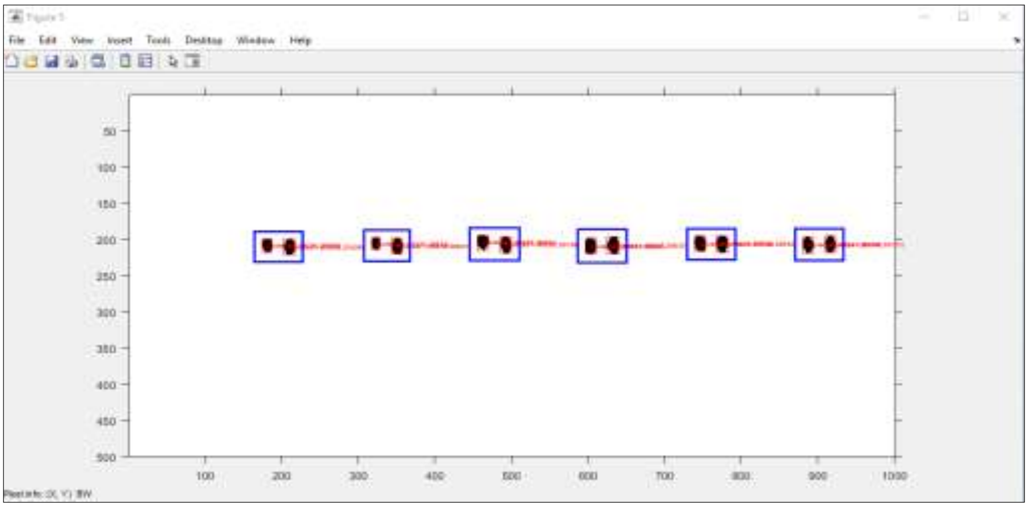


Fig.7 feature extract

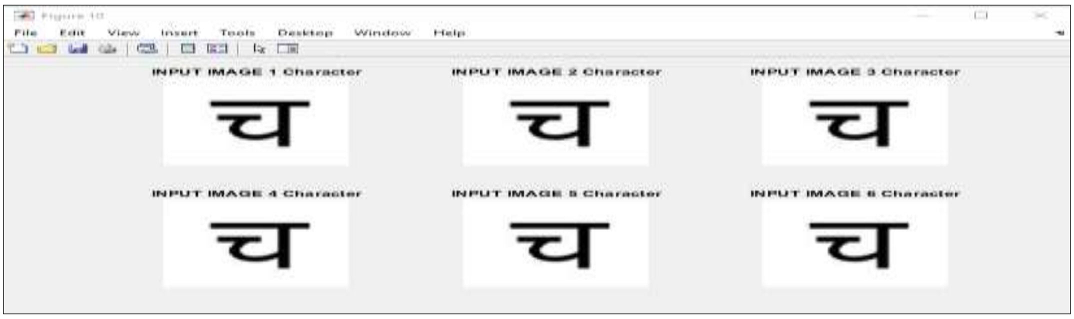


Fig. 8 classification result

Table 1 Result performance of proposed technique

Technique	Accuracy	Sensitivity	Specificity
ResNet 50	99.78	78.12	95.62
	95.78	78.14	96.23

In Table 1, the performance results of the proposed technique for text recognition are presented, showcasing the effectiveness of the system on two distinct datasets. The ResNet 50 algorithm demonstrates impressive accuracy on Dataset 1, achieving an accuracy rate of 99.78%. Additionally, the sensitivity and specificity metrics reveal its capability to correctly identify positive instances (78.12%) and accurately discern negative instances (95.62%).

Table 2 Result performance compare with existing and proposed technique

Technique	Accuracy
Existing Technique - AlexNet	94.35
Proposed Technique- ResNet 50	98.78

In the evaluation of text recognition techniques, the accuracy results reveal noteworthy performance differences between the existing technique using AlexNet and the proposed technique employing ResNet 50. The existing technique, based on AlexNet, achieves an accuracy level of 94.35%. In contrast, the proposed technique, leveraging the ResNet 50 algorithm, demonstrates a significantly higher accuracy rate, reaching 98.78%. This indicates that the ResNet 50-based approach outperforms the AlexNet-based method in accurately recognizing and classifying text. The substantial improvement in accuracy underscores the effectiveness and superiority of the proposed ResNet 50 technique for text recognition tasks.

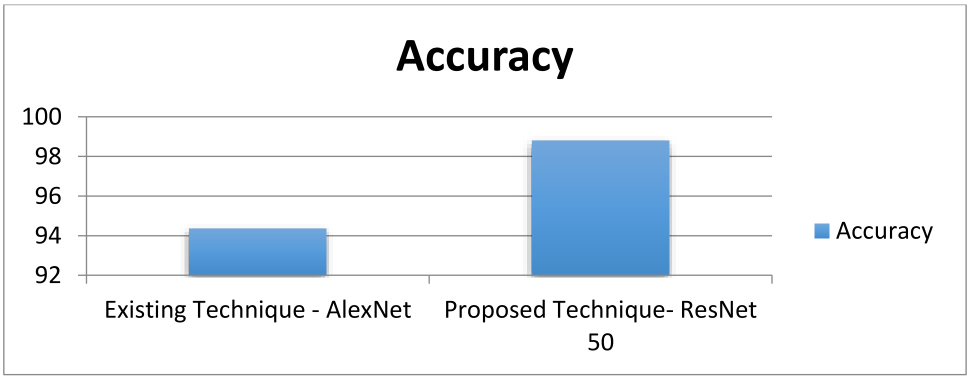


Fig.9 Result performance compare with existing and proposed technique

CONCLUSION

In conclusion, this research endeavors to make significant strides in the field of Hindi Braille speech detection through the integration of advanced technologies and a systematic methodology. The proposed system showcases a unique combination of speech-to-text conversion, feature extraction using the RESNET50 algorithm, and a comprehensive evaluation framework. The findings and insights derived from this study contribute to both the theoretical understanding of Braille speech analysis and the practical development of robust detection systems. The initial phase of the research involves the conversion of the provided audio dataset into text, laying the foundation for subsequent analysis. The adoption of speech-to-text transformation is a crucial step in unlocking the linguistic content embedded in the audio data. The utilization of RESNET50, a deep learning model acclaimed for image recognition tasks, in the realm of audio-related objectives is a novel and intriguing aspect of this research. The adaptability of RESNET50 to extract relevant features from audio signals underscores its versatility and potential for cross-domain applications. A notable but somewhat enigmatic step in the proposed system involves the conversion of text into an image. While the precise purpose of this transformation remains unclear from the provided information, it introduces an element of curiosity and warrants further exploration and clarification. Understanding the rationale behind this step could potentially unveil new dimensions in Braille speech analysis and contribute to the advancement of multimodal signal processing techniques. The proposed system for Hindi Braille speech detection lays the groundwork for several promising avenues of future research and development. As technology advances and new methodologies emerge, the following areas present significant opportunities for further exploration and enhancement:

References

[1] Shi, B.; Bai, X.; Yao, C. An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition. *arXiv* **2015**, arXiv:1507.05717.

[2] Liu, C.; Chen, X.; Luo, C.; Jin, L.; Xue, Y.; Liu, Y. A deep learning approach for natural scene text detection and recognition. *Chin. J. Graph.* **2021**, *26*, 1330–1367

[3] Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376.

[4] Liu, Y.; Wang, Y.; Shi, H. A Convolutional Recurrent Neural-Network-Based Machine Learning for Scene Text Recognition Application. *Symmetry* **2023**, *15*, 849. Lei, Z.; Zhao, S.; Song, H.; Shen, J. Scene text recognition using residual convolutional recurrent neural network. *Mach. Vis. Appl.* **2018**, *29*, 861–871. [Google Scholar] [CrossRef]

[5] Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **1994**, *5*, 157–166.

[6] Graves, A.; Mohamed, A.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 6645–6649.

[7] Shi, B.; Wang, X.; Lyu, P.; Yao, C.; Bai, X. Robust scene text recognition with automatic rectification. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4168–4176.

[8] Lee, C.-Y.; Osindero, S. Recursive recurrent nets with attention modeling for OCR in the wild. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2231–2239.

[9] Liu, W.; Chen, C.; Wong, K.-Y.K.; Su, Z.; Han, J. Star-net: A spatial attention residue network for scene text recognition. *BMVC* **2016**, *2*, 7.

[10] Wang, J.; Hu, X. Gated recurrent convolution neural network for OCR. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 334–343.

[11] Mohammad Soleymanpour;Michael T. Johnson;Rahim Soleymanpour;Jeffrey Berry Synthesizing Dysarthric Speech Using Multi-Speaker Tts For Dysarthric Speech Recognition

[12] Zhaoxu Nian;Yan-Hui Tu;Jun Du;Chin-Hui Lee A Progressive Learning Approach to Adaptive Noise and Speech Estimation for Speech Enhancement and Noisy Speech Recognition ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Year: 2021

[13] Yuanchao Li;Peter Bell;Catherine Lai Fusing ASR Outputs in Joint Training for Speech Emotion Recognition ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) Year: 2022

[14] Borisyuk, F.; Gordo, A.; Sivakumar, V. Rosetta: Large scale system for text detection and recognition in images. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 71–79.

[15] Baek, J.; Kim, G.; Lee, J.; Park, S.; Han, D.; Yun, S.; Oh, S.J.; Lee, H. What is wrong with scene text recognition model comparisons? Dataset and model analysis. In Proceedings of the 2019 IEEE/CVF international Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4715–4723.