SEYBOLD Report
ISSN: 1533 - 9211
https://seyboldreport.net/

SPEECH DIGIT RECOGNITION USING DEEP LEARNING

[1]Nivesh, [2]Shobha Bhatt
[1,2]Department of Computer Science & Engineering, Netaji Subhas University of Technology New Delhi, India,
[1]nivesh.pg21@nsut.ac.in, [2]Shobha.bhatt@nsut.ac.in

Abstract—The way individuals communicate with computers and other devices is evolving as a consequence of recent major developments in speech recognition technology. The development of comparable skills for languages like Hindi, which has a large user base, has been hindered by the reality that a large portion of research in this area has focused on English speech recognition. Because it improves voice-controlled technology and increases accessibility for Hindi speakers, recognising spoken Hindi numbers is very important. The present research uses convolutional neural networks (CNN) to show an innovative technique for Hindi speech digit recognition. In computer vision tasks, CNNs performed with remarkable performance, as they also demonstrated potential in speech recognition. CNNs are capable of learning and classifying Hindi digits with high accuracy by utilising the built-in patterns and characteristics in spoken digit audio signals. Using recordings of 300 individuals saying Hindi digits from 0 to 9, a carefully curated dataset was created to train and evaluate the CNN model. The dataset contains a range of speakers and considers consideration variances in age, gender, and regional accents to ensure the accuracy and generality of the proposed model. The model give accuracy of 96.4 in speech recognition.

Keywords- Automatic Speech Recognition (ASR), Convolutional Neural Network (CNN), Spectrogram Extraction, Epochs, Confusion Matrix.

1. INTRODUCTION

Speech recognition technology has improved human computer interaction by enabling a wide range of applications such as voice assistants, transcription systems, and voice-controlled gadgets. Pattern recognition mainly focuses on description, formalization, and identification of the patterns [1]. While significant progress has been made in English speech recognition, there is a need to investigate and create similar capabilities for languages such as Hindi, which has a big user base and a wide linguistic variety. Recognising spoken numbers in Hindi is very important because it can help advance voice-enabled gadgets while enhancing accessibility for Hindi speakers. To implement the ASR system, some obstacles may occur due to abnormality in speaking style and noises in the environment. The acoustic environment for ASR is much difficult or different than in the past [2]. The basic ambition of ASR is to handle all the challenges faced in the domain of speech recognition such as different speaking styles, uncertain environmental noise, and so on [3] [4] The development of an ASR for a local language is a difficult endeavour due to a lack of resources such as a corpus with sufficient vocabulary, dialectical diversity, and so on. Many works have been done for local languages such as in Punjabi (spoken in Pakistan and India) [5] [6] [7], Gujrati (local language of India)

[8], Urdu (national language of Pakistan and forth most widely spoken language in the world) [9] [10] [11] [12],Marathi (spoken in India) [13], Arabic (an official language of Arab and fifth widely used language in the world) [14], Bengali (spoken language of Bangladesh). The current work also includes Hindi language but very little works has been done in the development of Hindi speech recognition system [15]. In this research, we present a Convolutional Neural Networks (CNN) approach to Hindi voice digit recognition. In a variety of computer vision tasks, CNNs have done well, and they also demonstrated promise in speech recognition. CNNs can efficiently learn and classify Hindi digits with high accuracy by using the built-in patterns and features in spoken digit audio signals. Convolutional neural networks (CNNs) are successful variants of DNNs and valuable models for working with speech recognition systems. CNNs are moderately successful models for developing the speech recognition system, but this efficient architecture design is quite complicated. Their design requires prior and expert knowledge for performing the recognition process [16]. The natural visual perception paradigm of living creatures inspired this learning architecture [17] The basic aim of this research work is to develop an ASR for Hindi language by utilizing new machine learning technique such as deep learning. More particularly, the primary goal of this research study is to design Hindi isolated digit recognition system by using deep convolutional neural network (CNN). Originally, CNN is developed for image recognition and become more popular for handwritten digit recognition, however in the last few years it also used for speech recognition [3] [18] The study focuses on creating and evaluating a CNN model for Hindi spoken digit recognising. The starting point of this study is a meticulously collected dataset of recordings from individuals saying Hindi digits from 0 to 9. For the accuracy and generality of the proposed approach, the dataset includes a variety of speakers from a range of age groups, genders, and accents from different regions. Convolutional layers for feature extraction, pooling layers for dimensionality reduction, and fully connected layers for classification are just a few of the layers which make up the proposed CNN the design. Current optimisation techniques are used to enhance the model's performance while decreasing overfitting, such as dropout regularisation and batch normalisation. To evaluate the effectiveness of our approach, the dataset is divided into training and testing subsets. The model is trained using the training subset, and its performance is assessed on the testing subset. The model's accuracy, precision, recall, and F1 score are evaluated, providing comprehensive insights into its performance in recognizing spoken Hindi digits. The initial results indicate that the CNN technique is able to effectively recognising Hindi spoken digits. The achieved accuracy rate shows that applying CNNs for Hindi speech recognition tasks is feasible and shows the potential of developing accurate and efficient voice-based Hindi systems. This research has significance because it effectively bridges the gap in Hindi speech recognition technology for non-English languages. The results of this study contribute enhance Hindi speakers' speech recognition abilities and open the door for the development of voice-enabled applications for this category. This research study concludes by proposing a technique using CNN for Hindi voice digit recognition. Our research contributes to the growing field of speech recognition for Hindi languages through the use of CNNs and a well curated dataset. The results of this study offer implications for a number of applications,

such as voice-controlled systems, transcription services, and accessibility tools, which will encourage advancements in Hindi speakers' human-computer interaction.
The remaining part of the paper is structured as follows. Section 2 describes the Literature Review. Section 3 describes an overview of speech recognition technique. Section 4 describes the proposed method. Section 5 describes experimental set-up followed data processing, spectrogram extraction, CNN model and training of CNN model. Section 6 describes Result and discussion. Finally, conclusion and future work is presented in Section 7.

2. LITERATURE REVIEW

Speech recognition is a well-known topic in the fields of machine learning and natural language processing. It involves developing tools that can translate verbal communication into text. This survey of the literature on voice recognition needs to evaluate and describe the most significant findings and techniques from a number of research papers. For this review, the following papers had been taken into consideration:" Discriminatively trained continuous Hindi speech recognition using integrated acoustic features and recurrent neural network language modelling" The paper develops an integrated acoustic feature and recurrent neural network (RNN) language modelling-based Hindi speech recognition system. It explores discriminative approaches to training that increase the system's accuracy [15]" Developing a Speech Recognition System for Recognizing Tonal Speech Signals Using a Convolutional Neural Network" The design and implementation of a speech recognition system that's able to recognise tonal signals from voices is the primary subject of this research. It analyses the model's performance for tonal speech recognition tasks using a Convolutional Neural Network (CNN) architecture [19]." Discriminatively trained continuous Hindi speech recognition system using interpolated recurrent neural network language modelling": The authors provide a continuous Hindi speech recognition system using calculated RNN language modelling. In particular, for Hindi speech data, the research highlights the discriminative training method for boosting recognition accuracy [20]. The use of machine learning techniques for speech recognition is discussed in the paper" Speech Recognition Using Machine Learning IEEE Transaction". It presents an overview of the various algorithms and models that are used in the field, focusing on the way they perform when used for speech recognition tasks [21]. The paper entitled" Pashto isolated digits recognition using deep convolutional neural network" explains techniques to recognise Pashto isolated digits using a deep convolutional neural network (CNN). In the context of Pashto language recognition tasks, it evaluates the accuracy of the CNN architecture [22]. The research paper" English speech recognition based on deep learning with multiple features" is concerned with the recognition of English speech through the use of deep learning techniques with multiple features. It evaluates the application for different feature representations and their effects on the functionality of the speech recognition system [20]. The paper entitled" Persian speech recognition using deep learning" provides a technique based on deep learning for Persian speech recognition. It addresses the use of deep neural networks and investigates into the manner in which they can handle challenges of the Persian language [23]." Deep Speech: Scaling up end-to-end speech recognition" by A. Hannun et al. (2014) This study

introduces Deep speech, an end-to-end deep learning-based speech recognition system. State-of-the-art performance is achieved by using a deep neural network (DNN) with numerous layers of convolutional and recurrent layers to directly convert audio to text [24]. W. Chan et al. (2016)," Listen, Attend, and Spell" Listen, Attend, and Spell (LAS) is an attention-based end-to-end speech recognition paradigm presented in this study. To boost recognition accuracy, the LAS model employs an encoderdecoder architecture with an attention mechanism. On numerous voice recognition benchmarks, it obtains competitive performance [25]. K. He et al. (2016) published" Deep Residual Learning for Image Recognition." This study on residual learning, while not directly about speech recognition, is significant to deep learning approaches used in speech recognition. It introduces residual neural networks (ResNets), which allow for the effective training of very deep networks that can be tailored for speech recognition applications [26]. A. Graves et al. (2006) published" Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks." Connectionist Temporal Classification (CTC), a strategy for training recurrent neural networks (RNNs) without the necessity for explicit alignments between input and output sequences, is introduced in this influential paper. CTC has been frequently utilised to increase performance in speech recognition [27]. The reviewed papers highlight different aspects of speech recognition, including language-specific recognition, the use of deep learning models, feature engineering, and discriminative training techniques. These studies provide valuable insights into the development of speech recognition systems for various languages and demonstrate the effectiveness of different approaches in improving recognition accuracy. Future research can leverage these findings to enhance speech recognition systems for different languages and explore the potential of novel techniques in the field.

3. OVERVIEW OF SPEECH RECOGNITION

Speech digit identification using CNN and spectrograms is a special use of Convolutional Neural Networks (CNNs) paired with spectrogram representations to recognise and classify spoken digits. Spectrograms, which visually reflect the frequency content of audio signals over time, are used as input for CNN models. The spectrograms are processed as 2D images in this approach, and CNNs are used to learn discriminative features from the spectrogram representations. Convolutional layers are often used to extract local patterns, pooling layers for dimensionality reduction, and fully linked layers for classification in CNN models. The model is trained using a labelled dataset of spectrograms and their associated digit labels. The model learns to map spectrogram characteristics to the correct digit class by optimising parameters using techniques such as backpropagation and gradient descent. The classification of spoken digits using CNN and spectrograms is accurate and efficient, with potential applications in voice controlled systems, automated phone services, and voice biometrics. It allows for the smooth integration of spoken digit recognition into a variety of technology solutions, improving user experience and providing hands-free interactions. Figure 1 shows the basic diagram of speech Recognition given speech as input and predicted text is in output form.

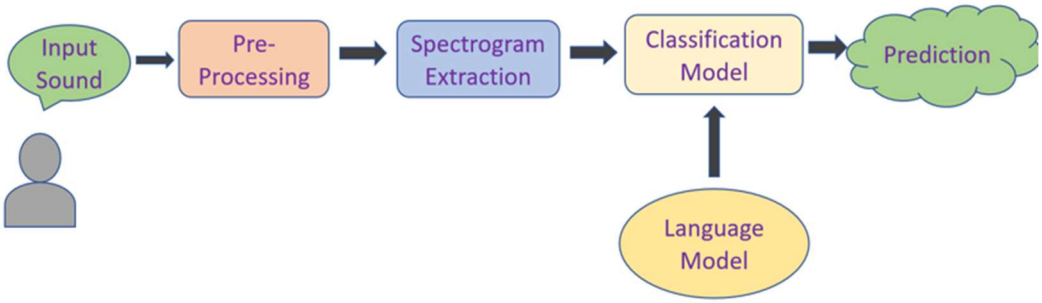

Fig 1- Basic Diagram of Speech Recognition

4. PROPOSED METHOD

The Goal of this work is to build and speech to text system for Hindi which can generate text we present speech to text that learns to recognise text from audio, using CNN model. System is trained on CNN model which reduces system complexity and keeps output good quality. The pipeline of the propose model is depicted in figure 2.

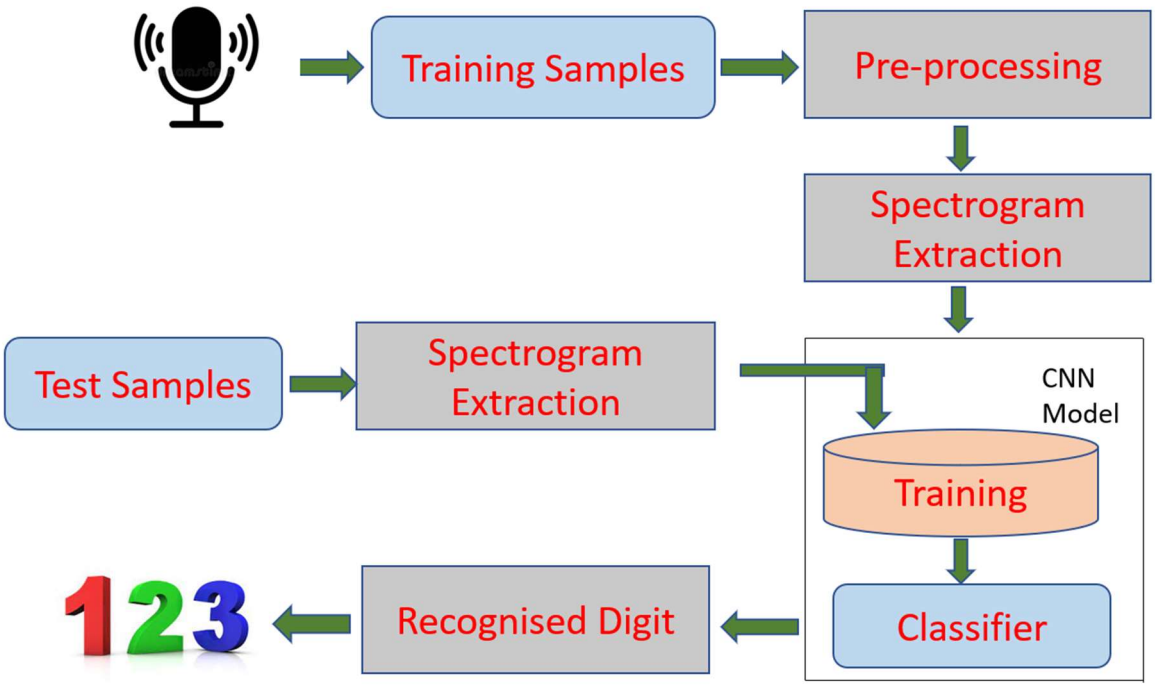

Fig 2- Speech Recognition using CNN

5. EXPERIMENTAL SET-UP

Experimental set-up for the Isolated Hindi Digit Speech recognition can be broadly classified into data preparation, spectrograms extraction, CNN Model and Training blocks. The speech corpus has been created by taking Hindi speech of 300 speakers of different age group and gender. Spectrograms has been taken out from these recordings and design a 3 layered CNN

Model. Test this Model after Train accordingly.

5.1 Data Preparation

The data preparation process includes recording speech datasets and extracting spectrograms from speech utterances.

A self-created speech dataset was used to examine the proposed methodology's impact on enhancing Hindi ASR performance. The created speech dataset is not constrained because it does not belong to any certain domain. Table 1 displays the details of the speech dataset, including recording conditions. The isolated word Hindi Speech corpus was produced with 300 speakers. The speech corpus was divided into 3000 isolated digit samples, which were used for training and testing datasets. 2400 utterances were used for training and 600 utterances were used for testing. The speech software applications Wave Surfer and Praat were used for recording and splitting. Table 2 lists the speakers who helped create the speech dataset for the studies.

Table 1. Isolated word Hindi speech corpus description

Parameter	Values
Type of speech corpus	Isolated Words
Language	Hindi
Number. of speakers	300
Speaker Accent	Different states of India
Sampling rate	48KHz
Recording Environment	Home and College

Table 2. Speaker information for speech corpus

S. NO	Number of speakers	Gender	Age	Recording Condition
1	18	Male child	6-17	Home
2	12	Female child	6-17	Home
3	161	Male	17-47	Home and College
4	109	Female	17-47	Home and College

5.2 Spectrogram Extraction

A spectrogram is a visual representation of the frequency content of a signal over time. Create the function, which takes an audio file path and a save path as input parameters. It also allows to change the spectrogram dimensions, overlap, and colormap. The method

reads the audio file and returns the sample rate and audio samples. A figure object is produced, with the size according to the spectrogram dimensions (64,64). An axis object (axe) is added to figure and its properties are configured to remove axes, ticks, and labels. On the axis object, the spectrogram function is called, with the audio samples as input. The spectrogram is computed and shown on the axis by this function. To eliminate any tick marks or labels, the x- and y-axis locators are set to null. Figure 3 shows spectrogram of zero.

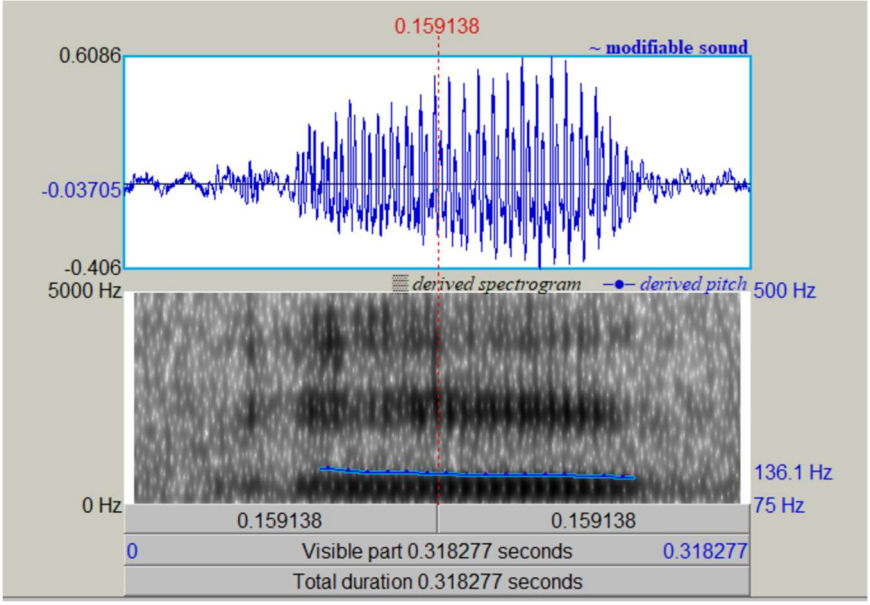

Fig 3. Spectrogram of (शून्य)

5.3 CNN Model

CNN model of Deep learning is used for recognising Hindi speech digits. The Summary of CNN Model is Given below in Table 3.

Table 3. Model Summary

Layer Type	Output Shape	Parameter
conv2d	(None, 63, 63, 32)	416
batch normalization	(None, 63, 63, 32)	128
conv2d_1	(None, 62, 62, 48)	6192
batch_normalization_1	(None, 62, 62, 48	192
conv2d_2	(None, 61, 61, 120)	23160
batch_normalization_2	(None, 61, 61, 120)	480
max_pooling2d	(None, 30, 30, 120)	0
dropout	(None, 30, 30, 120)	0
flatten	(None, 108000)	0
dense	(None, 128)	13824128
batch_normalization_3	(None, 128)	512

dropout_1	(None, 128)	0
dense_1	(None, 64)	8256
batch_normalization_4	(None, 64)	256
dropout_2	(None, 64)	0
dense_2	(None, 10)	650

The proposed CNN model is made up of various layers that are designed to perform image classification tasks. To normalise the activations, the model begins with three Conv2D layers, followed by Batch Normalisation layers. The first Conv2D layer contains 32 filters with 2x2 kernels, the second Conv2D layer contains 48 filters, and the third Conv2D layer contains 120 filters. Convolution operations are used by these layers to extract features from the input data. A Batch Normalisation layer is introduced after each Conv2D layer to increase training stability and speed up convergence. A MaxPooling2D layer with a pool size of 2x2 is also included in the model, which decreases the spatial dimensions of the feature maps. Dropout layers with a dropout rate of 0.25 are placed after the MaxPooling2D layer and the first Dense layer to prevent overfitting. During training, these layers set a percentage of the input units to zero at random. The model then includes a Flatten layer, which flattens the previous layer's output into a 1D vector. This gets the data ready for the fully connected layers. The model then includes two Dense layers, each with 128 and 64 units. To regularise the outputs, each Dense layer is followed by a Batch Normalisation layer and a Dropout layer. Finally, the model includes a Dense layer with 10 units and a softmax activation function that is appropriate for multi-class classification problems. The categorical cross-entropy loss function and the Adadelta optimizer are used to build the model.Overall, with its special layer structure, this CNN model intends to learn discriminative features from input spectrogram images for spoken digit identification tasks. The details mentioned above are shown in figure 4.

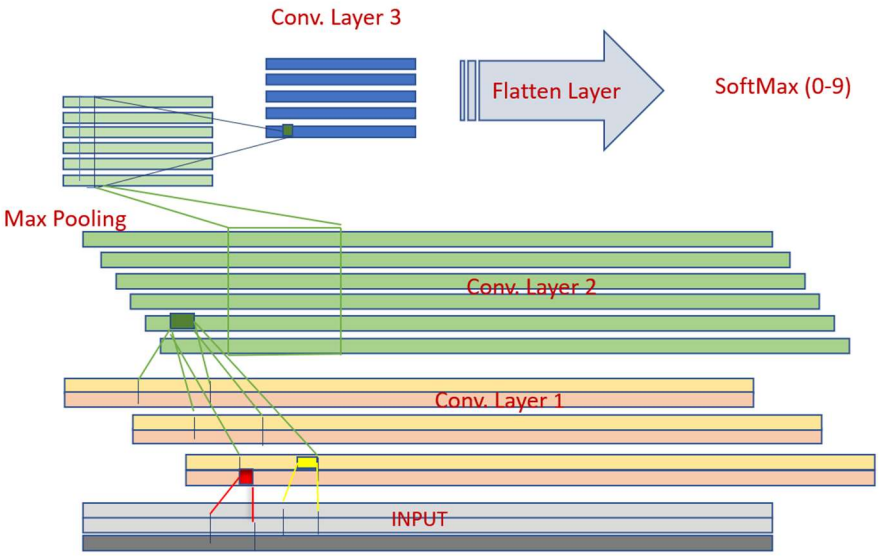

Fig 4. CNN structure

5.4 Training CNN Model

The function is used to train the model by providing the training dataset, batch size, validation split, number of epochs, and verbosity settings in the function. Batch_size: This parameter specifies how many samples are processed in each batch during training. The batch size is set to 50 in this case, which indicates that the model's parameters will be updated after analysing 50 samples at a time. validation_split: This parameter determines how much of the training data will be used for validation. In this situation, 20% of the training data will be set aside for validation, with the remaining 80% used for model training. epochs: This parameter specifies how many times the full training dataset will be iterated over during training. The model will go through 100 epochs in this scenario, which means it will see the full training dataset 100 times during the training phase. verbose: It regulates the amount of logging output generated during training. When set to 1, it displays progress updates and training metrics (such as loss and accuracy) for each epoch. The model will iterate through the training dataset during the training process, processing the data in batches of size 50. It will compute the loss between its predictions and the true labels before using an optimizer (such as stochastic gradient descent) to adjust the model's weights and biases to minimise the loss. During training, the validation split allows you to check the model's performance on unseen data. At the end of each epoch, the model will evaluate the validation dataset and return metrics such as validation loss and accuracy.

6. RESULT AND DISCUSSION

The research article describes a CNN-based speech digit recognition system. On the test dataset, the model scored a remarkable accuracy of 96.33%, indicating its ability to properly detect spoken digits as shown in figure 5. The equivalent test loss was 18.14%, indicating that the model's predictions were generally near to true values.

```
10/10 [==============================] - 0s 10ms/step - loss: 0.1814 - accuracy: 0.9633
Test Loss: 0.1814442276954651
Test Accuracy: 0.9633333086967468
```

Fig 5. Accuracy and Test loss

The graph of accuracy and epochs represented the model's learning process visually. Its accuracy most certainly grew as the number of epochs increased, demonstrating that the model was gradually improving its performance over time, as shown in figure 6. 100 epochs are used to represent the graph between epochs and accuracy.

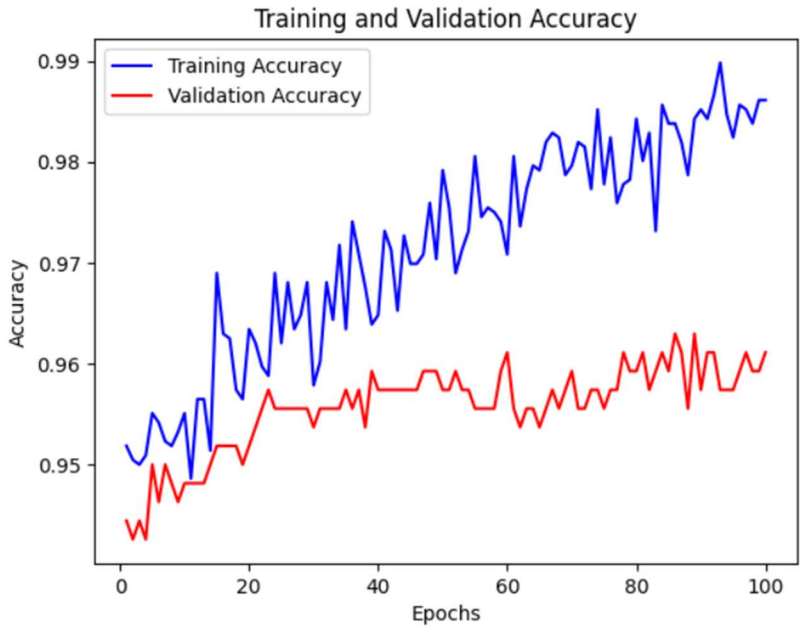

Fig 6- Graph Accuracy Vs Epochs

The confusion matrix showed useful information about the model's performance for each digit class as shown in figure 7. We can see from the matrix that the model has good precision and recall values for the majority of the digits, with only a few misclassifications. Because of its similarities to other digits or differences in speech patterns, the digit '8' exhibited the lowest precision and recall.

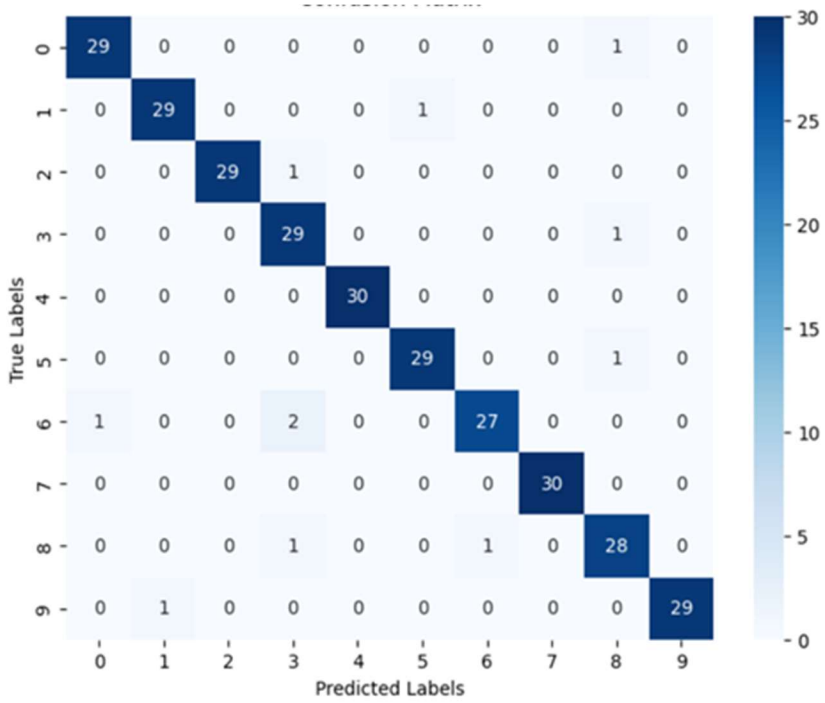

Fig 7- Confusion Matrix

Overall, the research report demonstrates the importance of the CNN model for speech digit identification, with excellent accuracy and valuable insights provided by the confusion matrix analysis. The findings demonstrate CNNs' potential for effectively recognising and classifying spoken digits, which can be useful in a variety of fields such as voice-controlled systems, automated call routing, and speech-to-text conversion.

7. CONCLUSION

The research provided a CNN-based technique for speech digit recognition that achieved an outstanding 96.33% accuracy on the test dataset. The model's ability to minimise the difference between anticipated and real digit labels is demonstrated by the low test loss of 18.14%.

The confusion matrix analysis demonstrates that the model performed effectively across all digit classes. The diagonal members of the matrix, which reflect correct predictions, demonstrated good precision and recall scores. The off-diagonal elements show some misclassification, but generally, the model performed well in properly detecting speech digits.

The accuracy and epochs graph depicts the model's training progress over time. It shows a steady rise in accuracy as the number of epochs grows, demonstrating that the model is still learning and improving.

Overall, the study paper's findings show that the CNN model is effective for voice digit recognition. The high accuracy and minimal test loss, as well as the confusion matrix analysis, indicate that the model can reliably recognise spoken digits with a high degree of precision. These findings help to enhance voice recognition technology and its potential uses in a variety of fields.

REFERENCES

[1] H. Liu, J. Yin, X. Luo, and S. Zhang, “Foreword to the special issue on recent advances on pattern recognition and artificial intelligence,” Neural Computing and Applications, vol. 29, pp. 1–2, 01 2018.

[2] l. Deng, K. Wang, A. Acero, H.-W. Hon, J. Droppo, C. Boulis, Y.-Y. Wang, D. Jacoby, M. Mahajan, C. Chelba, and X. Huang, “Distributed speech processing in mipad’s multimodal user interface,” Speech and Audio Processing, IEEE Transactions on, vol. 10, pp. 605 – 619, 12 2002.

[3] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, and G. Penn, “Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition,” 05 2012, pp. 4277–4280.

[4] ——, “Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition,” in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 4277–4280.

[5] M. Dua, R. Aggarwal, V. Kadyan, and S. Dua, “Punjabi automatic speech recognition using htk,” International Journal of Computer Science Issues (IJCSI), vol. 9, no. 4, p. 359, 2012.

[6] K. Sharma and P. Singh, “Speech recognition of punjabi numerals using synergic hmm and dtw approach,” Indian Journal of Science and Technology, vol. 8, no. 27, pp. 1–6, 2015.

[7] Y. Kumar and N. Singh, “An automatic spontaneous live speech recognition system for

punjabi language corpus," Int. J. CTA, vol. 9, no. 20, pp. 9575–9595, 2016.
[8] H. Chauhan and B. Tanawala, "Comparative study of mfcc and lpc algorithms for gujrati isolated word recognition," International Journal of Innovative Research in Computer and Communication Engineering, vol. 3, no. 2, pp. 822–826, 2015.
[9] I. Ashraf, L. T. W. Ho, and H. Claussen, "Improving energy efficiency of femtocell base stations via user activity detection," in 2010 IEEE Wireless Communication and Networking Conference, 2010, pp. 1–5.
[10] A. A. Raza, S. Hussain, H. Sarfraz, I. Ullah, and Z. Sarfraz, "An asr system for spontaneous urdu speech," the Proc. of Oriental COCOSDA, pp. 24–25, 2010.
[11] H. Ali, A. Jianwei, and K. Iqbal, "Automatic speech recognition of urdu digits with optimal classification approach," International Journal of Computer Applications, vol. 118, no. 9, pp. 1–5, 2015.
[12] M. Qasim, S. Nawaz, S. Hussain, and T. Habib, "Urdu speech recognition system for district names of pakistan: Development, challenges and solutions," in 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), 2016, pp. 28–32.
[13] S. Kayte, M. Mundada, and D. C. Kayte, "Implementation of marathi language speech databases for large dictionary," IOSR Journal of VLSI and Signal Processing (IOSR-JVSP) Volume, vol. 5, pp. 40–45e, 2015.
[14] M. Fasha, B. Hammo, N. Obeid, and J. Widian, "A hybrid deep learning model for arabic text recognition," CoRR, vol. abs/2009.01987, 2020. [Online]. Available: https://arxiv.org/abs/2009.01987
[15] A. Kumar and R. Aggarwal, "Discriminatively trained continuous hindi speech recognition using integrated acoustic features and recurrent neural network language modeling," Journal of Intelligent Systems, vol. 30, no. 1, pp. 165–179, 2021. [Online]. Available: https://doi.org/10.1515/jisys-2018-0417
[16] V. Passricha and R. K. Aggarwal, "Pso-based optimized cnn for hindi asr," International Journal of Speech Technology, vol. 22, no. 4, pp. 1123–1133, 2019. [Online]. Available: https://app.dimensions.ai/details/ publication/pub.1122084518
[17] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, and G. Wang, "Recent advances in convolutional neural networks," CoRR, vol. abs/1512.07108, 2015. [Online]. Available: http://arxiv.org/abs/1512.07108
[18] Y. Zhang, M. Pezeshki, P. Brakel, S. Zhang, C. Laurent, Y. Bengio, and A. C. Courville, "Towards end-to-end speech recognition with deep convolutional neural networks," CoRR, vol. abs/1701.02720, 2017. [Online]. Available: http://arxiv.org/abs/1701.02720
[19] S. Dua, S. S. Kumar, Y. Albagory, R. Ramalingam, A. Dumka, R. Singh, M. Rashid, A. Gehlot, S. S. Alshamrani, and A. S. AlGhamdi, "Developing a speech recognition system for recognizing tonal speech signals using a convolutional neural network," Applied Sciences, vol. 12, no. 12, 2022. [Online]. Available: https://www.mdpi. com/2076-3417/12/12/6223
[20] A. Kumar and R. Aggarwal, "Discriminatively trained continuous hindi speech recognition using integrated acoustic features and recurrent neural network language modeling," Journal of

Intelligent Systems, vol. 30, pp. 165–179, 07 2020.
[21] L. Deng and X. Li, “Machine learning paradigms for speech recognition: An overview,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 5, pp. 1060–1089, 2013.
[22] B. Zada and R. Ullah, “Pashto isolated digits recognition using deep convolutional neural network,” Heliyon, vol. 6, no. 2, p. e03372, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ S2405844020302176
[23] H. Veisi and A. Haji Mani, “Persian speech recognition using deep learning,” Int. J. Speech Technol., vol. 23, no. 4, p. 893–905, dec 2020. [Online]. Available: https://doi.org/10.1007/s10772-020-09768-x
[24] A. Y. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, “Deep speech: Scaling up end-to-end speech recognition,” CoRR, vol. abs/1412.5567, 2014. [Online]. Available: http://arxiv.org/abs/1412.5567
[25] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell,” CoRR, vol. abs/1508.01211, 2015. [Online]. Available: http://arxiv.org/abs/1508.01211
[26] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 06 2016, pp. 770–778.
[27] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, “Connection- ´ ist temporal classification: Labelling unsegmented sequence data with recurrent neural ’networks,” vol. 2006, 01 2006, pp. 369–376