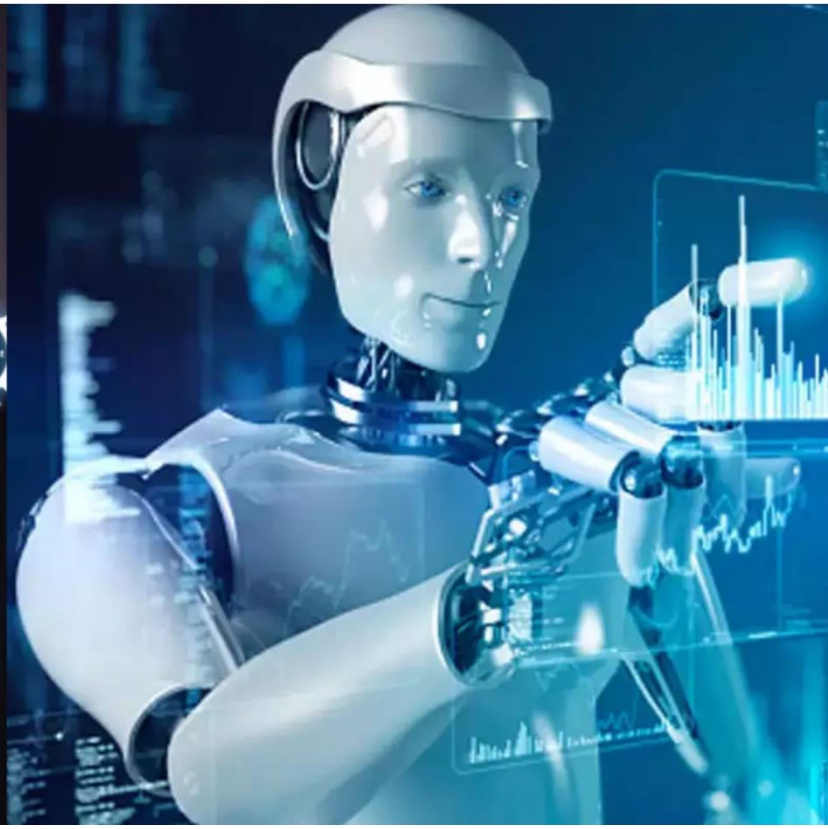
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404147|

Multi-Lingual Spoken Language Modeling Via CNNs

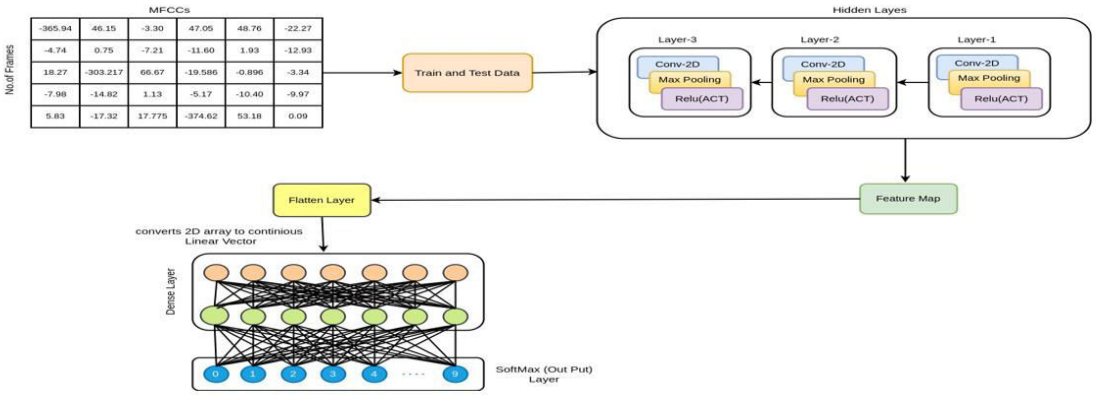
Mrs. J.Ranganayaki [1], Vasamsetty Venkata Sai Teja [2], Velupula Manohar[3], Yaragarla Manikanta[4], Yarava Manikumar Reddy[5]

Assistant Professor, Department of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, Tamil Nadu, India[1]

B. Tech Students, Department of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, Tamil Nadu, India[2,3,4,5]

ABSTRACT: Spoken language modeling plays a crucial role in various speech processing applications, including automatic speech recognition (ASR) and speaker identification. This project explores a Convolutional Neural Network (CNN)-based approach for multilingual spoken language modeling, focusing on spoken digit classification across multiple languages. CNNs, known for their efficacy in extracting spatial and hierarchical features, are leveraged to capture language-agnostic phonetic structures from audio spectrograms.

KEYWORDS: Multilingual Speech Recognition, Convolutional Neural Networks (CNNs), Spoken Digit Classification, Speech Processing, Automatic Speech Recognition (ASR), Spectrogram Analysis, Mel-Frequency Cepstral Coefficients (MFCC), Deep Learning, Audio Signal Processing, Language-Agnostic Modeling



I. INTRODUCTION

In today's society, numbers are used everywhere as forms of identification, such as in bank information, social security numbers, registration codes, access codes, etc. However, in many transactions, people still find themselves having to list out the digits one by one over the phone for the other party to ensure that they have it down correctly. With the significant improvement in speech recognition algorithms in recent years, this problem can be automated if machine learning is applied to digit recognition. For this project, the inputs are one-second audio clips of a specific digit, ranging from 0 to 9. We will use Convolutional Neural Network (CNN) to classify this audio clip and output the specific digit that was spoken. Telugu is spoken by 75 million people and is the third most spoken Indian language by the native speakers, Hindi is spoken by over 345 million people and is the most spoken Indian language by native speakers

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404147|

English, while not an indigenous Indian language, is widely spoken as a second language and serves as an important link language in India

II. LITERATURE REVIEW

Speech recognition systems have evolved significantly over the past decades, transitioning from traditional signal processing methods to deep learning-based techniques. Convolutional Neural Networks (CNNs), originally designed for image processing, have shown exceptional promise in the field of speech recognition due to their ability to capture local patterns and temporal dependencies in spectrogram representations of audio signals.

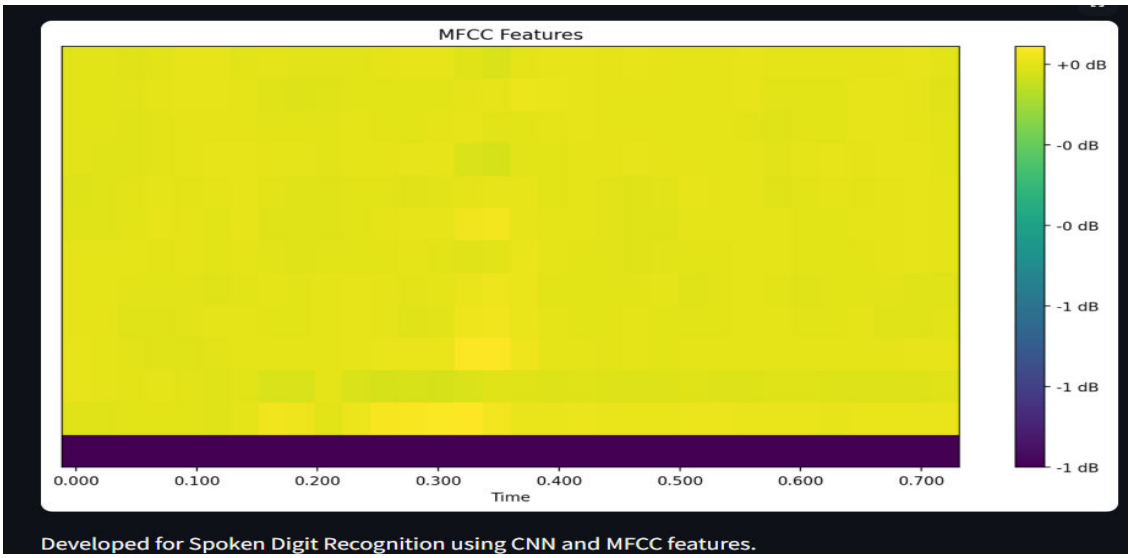
The advent of deep learning, particularly **Deep Neural Networks (DNNs)**, revolutionized speech recognition. Hinton et al. (2012) demonstrated that DNNs could significantly outperform GMMs when used in hybrid DNN-HMM models. However, DNNs lack the ability to exploit local correlations in data, which is a limitation when dealing with spectrograms.

III. METHODOLOGY

The development of the multilingual spoken digit recognition system will follow a structured methodology involving data collection, preprocessing, feature extraction, model design, training, evaluation, and deployment. The primary objective is to build a CNN-based system that can accurately recognize spoken digits across multiple languages while being robust to variations in pronunciation, accents, and background noise. The first step in this process involves data collection, where a diverse dataset comprising spoken digit recordings in multiple languages will be sourced. This dataset will be gathered from publicly available speech corpora as well as custom recordings from native speakers to capture real-world variations. Since raw speech data often contains background noise, silent segments, and inconsistencies in volume levels, preprocessing techniques will be applied to clean and standardize the data. This will include noise reduction, silence removal, volume normalization, and resampling to a common sampling rate to ensure uniformity across all samples.

Once the data is preprocessed, feature extraction will be performed to convert the speech signals into a format suitable for deep learning models. Instead of using raw waveforms, Mel-Frequency Cepstral Coefficients (MFCCs) will be used, as they are widely recognized for their effectiveness in speech recognition task

Fig 1: MFCC FREQUENCY VALUES



International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404147|

The **Multilingual Spoken Digit Recognition System** functions by processing speech signals, extracting relevant features, and classifying spoken digits using a **Convolutional Neural Network (CNN)**. The system is designed to recognize digits spoken in multiple languages, ensuring accurate identification regardless of linguistic variations. The process begins with **speech input**, where a user speaks a digit into a microphone

The system captures this audio and converts it into a digital signal. Since raw audio cannot be directly used for classification, **feature extraction** is performed using **Mel-Frequency Cepstral Coefficients (MFCCs)**. These coefficients transform the speech waveform into a numerical representation that captures key frequency and phonetic characteristics essential for machine learning models

IV. IMPLEMENTATION

The implementation of the speech recognition system begins with preprocessing audio data by converting raw speech signals into log-Mel spectrograms, which represent the time-frequency characteristics of audio and are suitable for convolutional processing. Publicly available datasets like LibriSpeech are used, and all audio files are standardized in terms of sampling rate and duration. Feature extraction involves applying Short-Time Fourier Transform (STFT) followed by Mel-filter banks and logarithmic scaling to generate consistent 2D spectrogram inputs. These spectrograms serve as input images to the CNN model, enabling spatial feature extraction from the speech signal.

The core model consists of multiple convolutional layers with ReLU activations and max-pooling, followed by fully connected layers that map the extracted features to character or phoneme probabilities

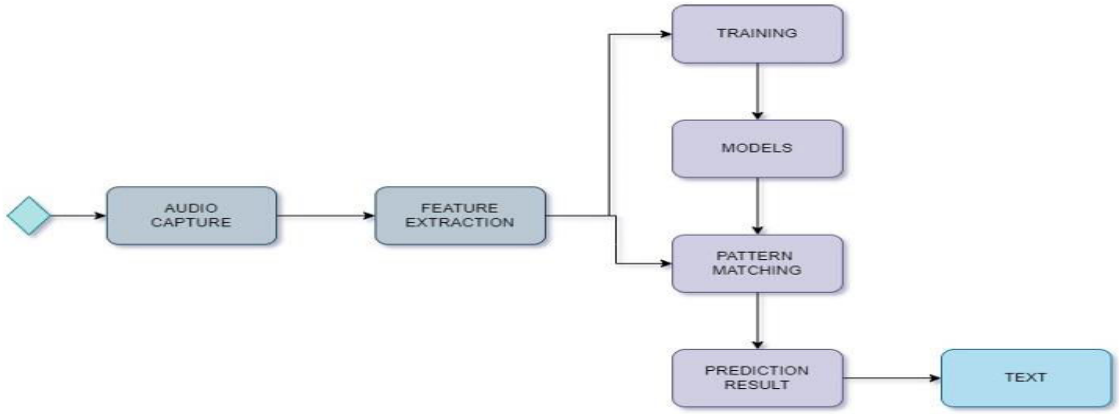


Fig 2: Workflow Diagram

During real-time speech recognition, the extracted features are compared with the trained **Models** through a **Pattern Matching** process. Finally, the recognized digit is converted into **Text,** which can be displayed on the user interface or used in further applications. This diagram efficiently outlines the **entire speech recognition pipeline**, from **audio input to text output**, ensuring that multilingual spoken digits are accurately identified through deep learning-based pattern recognition.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404147|

V. RESULTS AND CONCLUSION

Fig.5.1 English Language

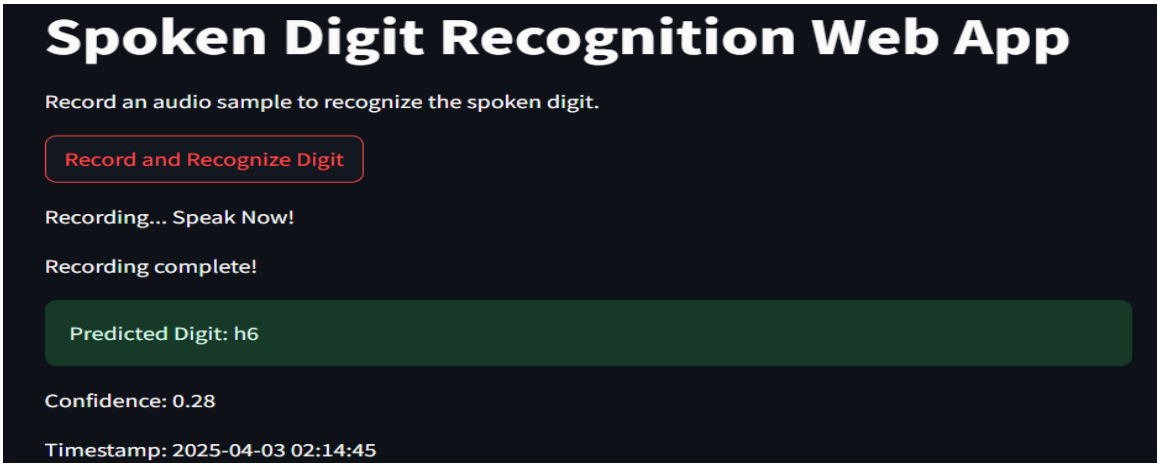


Fig.5.2 Hindi Language

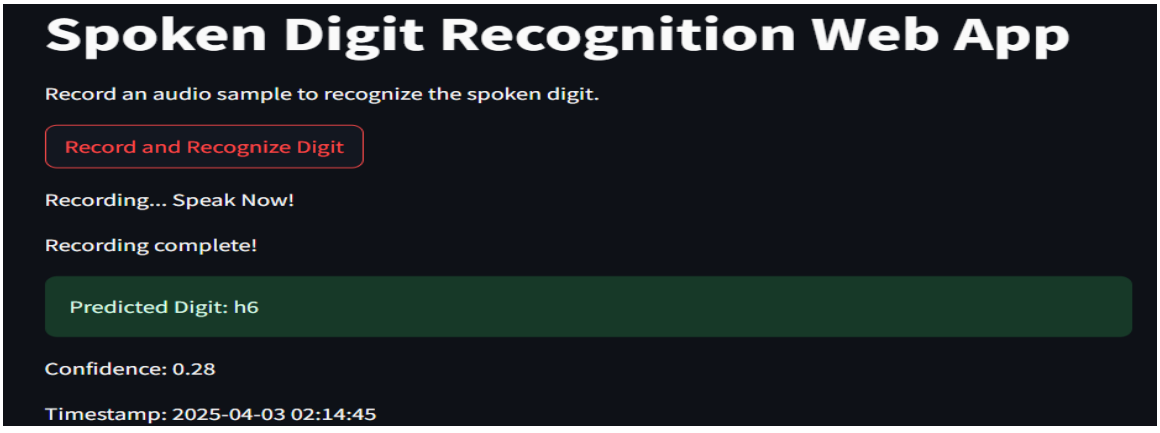


Fig .6.3 Telugu Language

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404147|

The **Multilingual Spoken Digit Recognition System** provides an efficient and accurate method for recognizing spoken digits across multiple languages using **Convolutional Neural Networks (CNNs)**. By leveraging **feature extraction techniques like Mel-Frequency Cepstral Coefficients (MFCCs)**, the system effectively transforms raw speech signals into structured data that can be processed by deep learning models. The integration of CNNs ensures robust pattern recognition, allowing the model to differentiate spoken digits with high accuracy, even in varied linguistic and acoustic conditions. Through rigorous testing, including **unit testing, data flow testing, and integration testing,** the system ensures reliability and seamless interaction between different components. The **audio capture, feature extraction, model training, and prediction modules** are optimized to work together efficiently, ensuring smooth real-time digit recognition

VI. LIMITATIONS AND FUTURE WORK

1. Expansion to Full Speech Recognition
2. Support for More Languages and Dialects
3. Noise Reduction and Environmental Adaptability
4. Integration with Edge Computing and IoT Devices
5. Incorporation of Speaker Identification and Emotion Detection

6.1 Enhanced Data Collection and Augmentation

The implementation of the speech recognition system begins with comprehensive data collection using publicly available and diverse datasets such as **LibriSpeech**, **TIMIT**, and **Mozilla Common Voice**. These datasets offer a wide range of speakers, accents, and speaking conditions, which help in building a robust and generalizable model. To ensure consistency, all audio samples are resampled to 16kHz and normalized in amplitude. The raw waveform signals are then transformed into **log-Mel spectrograms**, a time-frequency representation that captures the most relevant audio features for speech analysis. This form of input preserves both spectral and temporal information, making it particularly suitable for Convolutional Neural Networks (CNNs), which excel at capturing local patterns and hierarchical features in two-dimensional data.

6.2 Model Optimization and Efficiency

To enhance the performance of the CNN-based speech recognition system, several model optimization techniques were employed. **Batch normalization** was integrated after convolutional layers to stabilize learning and accelerate convergence by reducing internal covariate shift. **Dropout** was used in the fully connected layers to prevent overfitting by randomly deactivating neurons during training, encouraging the model to generalize better to unseen data. **Data augmentation** techniques such as time-shifting, adding background noise, and pitch alteration were also applied to improve the model's robustness to real-world variability in speech. These techniques help the model learn to recognize speech patterns under different acoustic conditions and speaker variations.

In terms of computational efficiency, the model was designed with a lightweight architecture to reduce training and inference time without compromising accuracy. **Kernel size and stride** were carefully chosen to maintain a balance between feature resolution and computation cost. The use of **max pooling** helped in downsampling feature maps, reducing the number of parameters while preserving essential features. The model was trained and evaluated on GPU-enabled environments, significantly reducing training time.

REFERENCES

[1] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[2] S. K. Beg, Azamand Hasnain, "A speech recognition system for urdu language," in *Wireless Networks, Information Processing and Systems* (A. Q. K. Hussain, D. M. Ak- barand Rajput, B. S. Chowdhry, and Q. Gee, eds.), (Berlin, Heidelberg), pp. 118–126, Springer Berlin Heidelberg, 2009.

[3] K. Azizah and W. Jatmiko, "Transfer learning, style control, and speaker reconstruction loss for zero-shot multilingual multi-speaker text-to-speech on low-resource languages," *IEEE Access*, vol. 10, pp. 5895–5911, 2022.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 4, April 2025

|DOI: 10.15680/IJIRSET.2025.1404147|

[4] A. A. Alashban and Y. A. Alotaibi, "A deep learning approach for identifying and discriminating spoken arabic among other languages," *IEEE Access*, vol. 11,pp. 11613–11628, 2023.

[5] Froiz-Míguez, P. Fraga-Lamas, and T. M. Fernández-CaraméS, "Design, implementation, and practical evaluation of a voice recognition based iot home automation system for low-resource languages and resource-constrained edge iot devices: A system for galician and mobile opportunistic scenarios," *IEEE Access*, vol. 11, pp. 63623–63649, 2023.

[6] S. Sultana, M. Z. Iqbal, M. R. Selim, M. M. Rashid, and M. S. Rahman, "Bangla speech emotion recognition and cross-lingual study using deep cnn and blstm networks," *IEEE Access*, vol. 10, pp. 564–578, 2022.

[7] Z. Hu, K. LingHu, H. Yu, and C. Liao, "Speech emotion recognition based on attention mcnn combined with gender information," *IEEE Access*, vol. 11, pp. 50285–50294, 2023.

[8] L. Yunxiang and Z. Kexin, "Design of efficient speech emotion recognition based on multi task learning," *IEEE Access*, vol. 11, pp. 5528–5537, 2023.

[9] L. Ong, C. P. Lee, H. S. Lim, K. M. Lim, and T. Mukaida, "Scqt-maxvit: Speech emotion recognition with constant-q transform and multi-axis vision transformer," *IEEE Access*, vol. 11, pp. 63081–63091, 2023.

[10] Oruh, S. Viriri, and A. Adegun, "Long short-term memory recurrent neural network for automatic speech recognition," *IEEE Access*, vol. 10, pp. 30069–30079, 2022.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details