# Deep Learning Approach for Spoken Digit Recognition in Gujarati Language

Jinal H. Tailor[1], Rajnish Rakholia[2], Jatinderkumar R. Saini[3]*, Ketan Kotecha[4]
MCA, S. S. Agrawal Institute of Management & Technology, Navsari, India[1, 2]
Symbiosis Institute of Computer Studies and Research, Symbiosis International (Deemed University), Pune, India[3]
Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune, India[4]

*Abstract*—**Speech Recognition is an emerging field in the area of Natural Language Processing which provides ease for human machine interaction with speech. Speech recognition for digits is useful for numbers oriented communication such as mobile number, scores, account number, registration code, social security code etc. This research paper seeks to achieve recognition of ten Gujarati digits from zero to nine (૦ to ૯) by using a deep learning approach. Dataset is generated with total 8 native speakers 4 male 4 female with the age group of 20 to 40. The dataset includes 2400 labeled audio clips of both genders. To implement a deep learning approach, Convolutional Neural Network (CNN) with MFCC is used to analyze audio clips to generate spectrograms. For the proposed approach three different experiments were performed with different dataset sizes as 1200, 1800 and 2400. With this approach maximum 98.7% accuracy is achieved for spoken digits in Gujarati language with 98% Precision and 98% Recall. It is analyzed from various experiments that increase in dataset size improves the accuracy rate for spoken digit recognition. No of epochs in CNN also improves accuracy to some extent.**

***Keywords—CNN; Deep learning; digit; Gujarati; speech recognition***

## I. Introduction

Speech is the most common medium used by humans for communication. Humans express their thoughts in the form of speech. Speech recognition is the emerging field in the area of Natural Language Processing [1]. Performance and Accuracy of Speech recognition process depends upon the speaker's mental state, emotions, accent, style and ease of speaking and environment. Human speech includes various parameters such as words, sentences, digits and fillers. Speaker feels more comfortable when he is communicating in his native language. In general, people have good command and control over their native language as they have learnt it from their childhood. Speech recognition with native language provides better performance and accuracy with the use of algorithms [2]. Generally the speech recognition process focuses upon identification of words and sentences. Very little work has been done by researchers regarding digits recognition in Gujarati language. Digits recognition with deep learning includes zero to nine numbers identification using optimized algorithms.

Gujarati is Indo-Aryan language and used as native language for people living in Gujarat state in India. It is spoken by 55.5 million people in India. Gujarati language includes 34 consonants, 12 vowels, 10 digits and 106 special symbols. Speech recognition related to the Gujarati language is a challenging task due to morphological structure, language barrier, dialects and limited resources [3]. This research paper mainly focuses upon digit recognition in Gujarati language using deep learning approach.

Deep learning is a subfield of machine learning that includes various methods and functions that imitates human behavior and knowledge. Deep learning algorithms apply a hierarchy of nonlinear transformation to the input values and generate statistical models as output that can predict results in future on their own. Deep learning approaches require large amounts of data for training and require more computation resources [4]. There are several forms of neural networks such as Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Artificial Neural Network (ANN) and Feedforward Neural Network (FNN). Most popular deep neural network is Convolutional Neural Network (CNN) to understand and analyze data in visual form.

In the proposed approach for digit recognition for Gujarati language CNN model was implemented. For the analysis process, dataset was prepared with 8 speakers with 2400 labeled audio clips. Different dataset sizes are used as input to CNN to analyze the impact on accuracy.

## II. Literature Review

Gunawan, A. (2010) [5] developed a digit recognition system from zero through nine using HMM and MATLAB for GUI. Spoken digits were in an isolated structure with 34 male and female speakers. They have created two different modules for isolated and continuous speech recognition. To check effect of variation of environment on accuracy they have selected noisy and clean environment. In a clean environment, in multi-speaker mode for isolated and continuous speech average accuracy found is 96%. For speaker independent mode, average accuracy is 78%. In noisy environments, in multi-speaker mode for isolated and continuous speech average accuracy found is 80%. For speaker independent mode average accuracy is 61.6%. They have concluded that recognition accuracy depends upon some factors such as noise rate and microphone quality.

Chapaneri and Jayaswal (2013) [6] have proposed a speaker-independent system for isolated digits using WMFCC. They have used SOLA based techniques and IFDTW to reduce time complexity for recognition. The overall accuracy of the proposed system for recognition 0 to 9 was 99.16%. They have

*Corresponding Author.

concluded that the system performs 22 times faster with higher accuracy and lower computational overhead achieved using WMFCC and FIFDTW.

Datta et al., (2021) [7] have done comparative analysis for methods to recognize zero to nine digits. For comparison they have analyzed cepstral and MFCC. They have concluded that the feature extraction method MFCC performed with better results compared to cepstral in command based systems.

Kurian, C., & Balakrishnan (2013) [8] have developed speaker-independent systems for connected digits for Malayalam language using Perceptual Linear Predictive (PLP) cepstral coefficient and Hidden Markov Model (HMM). For the training phase they have recorded 20 sets of continuous digits of 21 speakers from the age group of 20 to 40 years in an office environment. For the proposed system they have achieved 99.5% recognition accuracy by averaging word accuracy.

Sen O. et al. (2021) [9] created a convolutional neural network model to recognize Bangla spoken digits from 0 to 9. They have used Mel Frequency Cepstrum Coefficient (MFCC) for feature extraction and Convolutional Neural Network (CNN) for digit recognition. They have developed a proposed model using python,keras and tensorflow for 4000 spoken occurrences with parameters such as gender, accent and age. Measured accuracy with 10 fold cross validation was 96.7%.

Zerari N. et al. (2019) [10] proposed a framework for digit recognition using neural network. For feature extraction they have used Mel Frequency Cepstral Coefficient (MFCC) and Filter Banks (FB) coefficients. They have applied a proposed model on two different datasets one for spoken digits and another for spoken TV commands for different age groups for a total of 88 Arabic native speakers from which 44 were male and 44 were female. Total first digit database entries were 8800 tokens and the second TV command database includes 10000 instances. It is analyzed that the delta-delta feature used to characterize speech recognition signals gives better accuracy over 96% compared to MFCC and FB.

Dhanashri D. & Dhonde S. (2017) [11] have proposed isolated word speech recognition using HMM and DNN for acoustic modeling. Dataset for digit sequences samples were collected in a quiet environment at 20 kHz with 55 men and 57 women. System accuracy is measured for three different hidden layers such as 300,400 and 500 units. Accuracy achieved for 300 hidden layer units is 83.67%, for 400 its 85.44% and for 500 units its 86.06%. They have analyzed that proposed model accuracy can be improved by increasing the number of hidden layer units.

Wazir A. & Chuah J. (2019) [4] have proposed Arabic digits speech recognition model using Recurrent Neural Network (RNN). Feature extraction is done with MFCC and extracted features are fed into the Long Short Term Memory (LSTM) network. They have collected total 1040 data points from different countries and dialects. For the training phase 840 data points were used from 1040 data points and achieved 94% with minimum loss. In the testing phase total 200 data points were included and 65% accuracy was achieved. They have found that the most recognized digit is ‘0’ with 80% accuracy and the worst recognized digit is ‘6’ with 60% accuracy.

Hamidi M. et al. (2020)[12] have presented analysis of Amazigh digit recognition with IVR system for speaker-independent systems in noisy environments. They have used open source software Asterisk 1.6, Ekiga softphone and CMU Sphinx Tools. They have selected 22 Moroccan native speakers of both genders male and female. They have concluded that minor degradation in accuracy found for SNR as low 3 and 9 lb. Major degradation was found with decoded speech signal and best recognition rate at 3 lb.

Swedia E. et al. (2018) [13] have proposed a deep learning approach for Indonesian digit recognition using LSTM. For feature extraction they have used LPC and MFCC techniques and compared their performance. They have recorded total 7990 speech digits with 12 coefficients of MFCC and 12 for LPC. They have concluded that Indonesian speech recognition gives better performance using MFCC with 96.58% compared to LPC with 93.79%.

Swastika R. et al. (2017) [14] have performed comparative analysis for four deep learning architectures such as DBN-DNN, LSTM, TDNN and CNN. For feature extraction MFCC and LPC techniques are used. They have used TIDigits dataset which includes 226 speakers from which 111 men and 114 women each uttered 77 digit sequences. Kaldi speech recognition toolkit is used for training and decoding processes. They have concluded that CNN gives the best recognition performance in a noisy and reverberant environment compared to other three architectures. There are also research instances where the authors worked for printed Gujarati characters, for instance [15]. A comparative study of Literature review is presented in Table I.

TABLE I. COMPARATIVE STUDY OF LITERATURE REVIEW

| Authors | Year | Spoken Language | Methods / Techniques | Category of Tokens | Accuracy |
|---|---|---|---|---|---|
| Gunawan [5] | 2010 | English | HMM | Digits | 96% |
| Chapaneri, S. V., & Jayaswal, D. J. [6] | 2013 | English | WMFCC<br>FIFDTW | Isolated Digits | 99.16% |
| Datta et al. [7] | 2021 | English | Cepstral<br>MFCC<br>Vector Quantization | Digits | MFCC VQ - 93.3 %<br>Cepstral - 86.6 % |
| Kurian, C., & Balakrishnan, K. [8] | 2013 | Malayalam | PLP<br>HMM | Connected Digits | 99.5 % |
| Sen O. et al. [9] | 2021 | Bangla | MFCC<br>CNN | Digits | 96.7 % |
| Zerari N. et al. [10] | 2019 | Arabic | LSTM<br>MLP | Digits | 96 % |
| Dhanashri D. & Dhonde S. [11] | 2017 | English | HMM<br>DNN | Isolated Digits | 300 layer - 83.67 %<br>400 layer - 85.44 %<br>500 layer - 86.06 % |
| Wazir A. & Chuah J. [4] | 2019 | Arabic | MFCC<br>LSTM<br>RNN | Digits | Training - 94 %<br>Testing - 65% |
| Hamidi M. et al. [12] | 2020 | Amazigh | CMU Sphinx | Digits | 72.43 % |
| Swedia E. et al. [13] | 2018 | Indonesian | LSTM<br>MFCC<br>LPC | Digits | MFCC - 96.58 %<br>LPC - 93.79 % |
| Swastika R. et al. [14] | 2017 | English | DBN - DNN<br>LSTM<br>CNN<br>Kaldi | Digits | CNN with better performance in clean environment |

## III. METHODOLOGY

The methodology to recognize digits in Gujarati language using CNN includes 6 major steps depicted in Fig. 1.

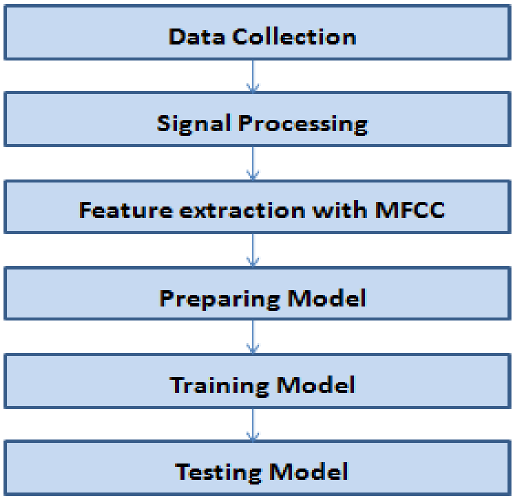


Fig. 1. Flowchart of Steps for Proposed Model.

### A. Data Collection

For data collection, 8 native speakers, comprising 4 males and 4 females, of the Gujarati language were selected. All 8 speakers were between the age group of 20 to 40. For recording purposes Audacity software was used and recording is performed in a lab environment. Total 30 occurrences of 0 to 9 digits were spoken by each speaker. Dataset consists of 0 to 9 digits in Gujarati with English language representation listed in Table II. Total 2400 occurrences (8 speakers * 10 digits * 30 occurrences) were used for the proposed speech recognition model. Collected data is split in two parts: Training and Testing. For data division 80:20 ratios are followed from total collected data. From 2400 total occurrences 1920 (80% of 2400) were used for training the model and 480 (20% of 2400) occurrences used for testing purposes. All the audio files are collected and stored in wav format.

TABLE II. DATASET OF DIGITS (0 TO 9) IN GUJARATI AND ENGLISH REPRESENTATION

| Digit in Gujarati | Pronunciation in Gujarati | Digit word in English | Digit in English numeric |
|---|---|---|---|
| ૦ | ૦- શૂન્ય | 0 | zero |
| ૧ | ૧- એક | 1 | one |
| ૨ | ૨- બે | 2 | two |
| ૩ | ૩- ત્રણ | 3 | three |
| ૪ | ૪ -ચાર | 4 | four |
| ૫ | ૫- પાંચ | 5 | five |
| ૬ | ૬- છ | 6 | six |
| ૭ | ૭- સાત | 7 | seven |
| ૮ | ૮- આઠ | 8 | eight |
| ૯ | ૯- નવ | 9 | nine |

### B. *Signal Processing*

Speech signals are non-stationary signal combination of voiced and unvoiced signals. From actual speech signals, it is required to extract voiced segments without silence and noise. Data segments with silence are not useful for feature analysis. To remove noise from signals, audacity has a provision with noise reduction option through which we can control by specifying how much dB value we want to consider as noise from 0 to 24 (maximum) dB values.

For processing, first we have removed noise from recorded audio files. Through Audacity noise reduction parameters can be set with residue data that is removed. It provides ease to select a voice sample window which contains useful information about speech signals. The dataset containing a total of 2400 occurrences of the 10 digits, from 0 to 9, were recorded by the speakers in Gujarati. Noise parameters are set to preferences and selected tracks are stored with reduced noise.

### C. *Feature Extraction with MFCC*

Raw data is not used directly in the model training process because it contains noise and other unnecessary parameters that affect accuracy of the model. Feature extraction is an important step in the speech recognition process. Extracted features are used as input to the model to generate text sequences. In the model, extracted features as input, gives better performance compared to raw audio signals. To extract spectral frames from speech signals, we have used the MFCC technique which is a widely used technique for feature extraction. First audio signal is converted from analog to digital with a sampling frequency 16 kHz. To improve model performance, pre-emphasis increases the energy in higher frequency for better phone detection.

To differentiate phones from audio signals, segments are divided into window width 25ms. From each segment, we have extracted 39 features. Hamming window is used to divide signal into segments without noise in a high frequency region. By applying Discrete Fourier Transform (DFT) to convert the time domain of a signal into frequency domain for easy analysis.

By modeling human hearing property into a model at the feature extraction step can improve the performance of the model. For mapping actual frequency with human hearing frequency, mel scale parameter is used. At lower energy humans can easily identify increased levels compared to higher energy. To adapt the same human behavior in the system, log is applied to Mel-filter.

After DFT transformation, the time domain and frequency domain are inverted. The lowest frequency is the highest frequency in the time domain. So inverse transform IDFT (inverse Discrete Fourier Transform) of log of magnitude of signal is generated which is called a cepstrum. After applying IDFT, MFCC includes 12 coefficients to extract signal sample energy as features.

In the dynamic feature generation step, with 13 features, MFCC technique generates first order derivatives and second order derivatives for each feature so total 26 features generated. Derivatives are calculated with differences between coefficients of audio signals to understand transition. Calculated 39 features generated by MFCC for each audio signal that was included as input to the speech recognition system.

### D. *Deep Learning Approach*

Deep learning includes various sub domains such as computer vision, digital signal processing, music classification and automatic speech recognition. With the popularity of virtual assistants like Alexa, Google Home and Siri, the field machine learning systems are promoted at a higher level for speech to text or action. All these systems work upon audio signal processing using deep learning concepts.

### E. *CNN*

Neural networks include a major part as Convolutional Neural Network. CNN is made up of neurons that have learnable weights and biases. It includes various application implementations such as image processing, object identification and classification. CNN can also be used for speech recognition as it can be applied on acoustic models and define convolution layers to each overlap acoustic frames. It extracts phones, pitch, frequency and gender features from acoustic frames according to weights of the network. Convolution Layer is the first layer of CNN that is the core building block and performs major computational tasks. It is a mathematical operation that merges two sets of information. CNN includes three main aspects: locality, weight sharing and pooling. Before start processing input data it is necessary to reduce non-white noise through locality. Weight sharing is used to improve robustness of model and overfitting and reduce weights. Pooling or subsampling process reduces the size. The input image is resized to optimal level to input into the convolutional layer. CNN Architecture is shown in Fig. 2.

In CNN, a filter or kernel is available that is placed in a sliding window over some of the pixels of the image depending upon kernel size to produce feature maps in different convolutional layers. In this step, through element wise multiplication each original pixel value of the image is multiplied with values in the filter. The sum of multiplication values is generated.

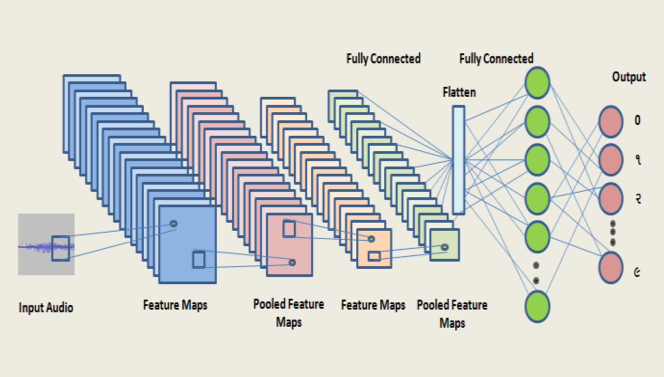


Fig. 2. Convolutional Neural Network.

If the value found is too low then there is nothing to compare in the activation map by curve detector filter. By adding more filters input data can be verified in depth of the activation map. The second layer called the Activation layer, ReLu (Rectified Linear Unit) function is applied to increase non-linearity in the CNN as images consist of objects that are not linear. In the third pooling layer, that involves downsampling of features. Common pooling layer uses 2 X 2 max filters with stride of 2 that is non-overlapping filter. Max filter returns max value in the feature from selected region. Last layer is known as a fully connected layer that pooled a full feature map matrix into a single column that was used for processing in a neural network. All the features are combined in a network through fully connected layers.

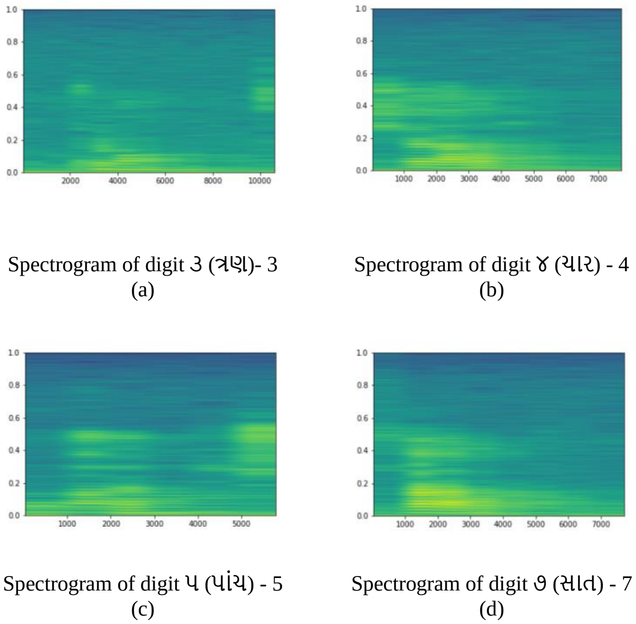
### F. *Preparing Model*

The dataset consists of a total 2400 occurrences both from male and female speakers. It consists of 10 spoken digits recording from 0 to 9. Recording of the dataset was performed using audacity software with 22050 Hz monophonic 16-bit audio files with .wav format. Google colab is used to implement and execute neural networks to get GPU and TPU as a runtime environment. It is free and easy to use. Librosa python library is used to extract spectrograms from audio files. Total dataset is splitted into training and testing data with 80:20 ratios. From a total 2400 audio files, 1920 files were used in the training session. The algorithm is written in python with TensorFlow library using Keras open source library as python interface. Each spectrogram file is stored with PNG format. Spectrogram images of different digits in Gujarati are depicted in Fig. 3. All spectrogram images belong to classes ranging from 0 to 9. Then data is split into training and testing sets. We have used sequential types to build a layer wise model in keras.

In the training phase, input images of the spectrogram are passed to a 2D convolutional layer (Conv2D) with filter size of 32. There are 2 layers as Conv2D layers. First layer consists of 64 nodes and the second layer consists of 32 nodes. With kernel size of 3, 3 X 3 filter matrix is used. For two layers activation function ReLu (Rectified Linear Activation) is used. A ‘Flatten’ layer is added between conv2D and dense layer to transform the resultant vector into 1 dimensional vector. Dense is the standard output layer used in neural networks. Mainly 10 different nodes were added in layers ranging from 0 to 9. Resultant vector is passed to a fully connected dense layer with a Softmax function that concludes the output in sum up to 1 to represent it in the form of probabilities. Based on the highest priority, prediction series can be generated. After setting all the parameters for the model, compilation is done with three parameters. Compiling process includes optimizer, loss and metrics. SGD (Stochastic Gradient Descent) optimizer is added from keras as it is an iterative method for optimizing objective function. The ‘categorical_crossentropy’ is used for loss function which indicates the model performance. Third parameter Accuracy metric is used to show accuracy of model on validation data.

### G. *Training Model*

To train a model, fit () function is used with training data and validation data as test data with number of epochs. Epochs specify the total number of model cycles for data. We can get better model performance by increasing the number of epochs up to some extent. For proposed model, 500 epochs were set to analyze neural network performance.



Spectrogram of digit ૩ (ત્રણ)- 3
(a)

Spectrogram of digit ૪ (ચાર) - 4
(b)

Spectrogram of digit ૫ (પાંચ) - 5
(c)

Spectrogram of digit ૭ (સાત) - 7
(d)

Fig. 3. Spectrogram of different Digits Spoken in Gujarati.

### H. *Testing Model*

To predict the output based upon test data, predict_generator function is used. It generates an array with 10 digit numbers. Each digit range represents the digit classification according to probability.

## IV. Result and Discussion

The experiment was executed with training and testing of data to create a neural network. The proposed approach includes a total 2400 occurrences of spoken digits 0 to 9 in Gujarati language. The model is developed with Python programming language with Keras and Tensorflow libraries. The model is executed three times with varied dataset sizes as 1200, 1800 and 2400 digits. The result analysis with various data sets is depicted in Table III. In the first iteration with 1200 dataset values, received accuracy was 58.6% with 58% precision and 58% recall. In the second iteration, accuracy achieved is 78% with 78% precision and 78% recall. In the third iteration with 2400 occurrences, 98.7% accuracy is achieved with 98% precision and 98% recalls.

It is analyzed that accuracy of recognition improves with dataset size. The result can be viewed in graphical form depicted in Fig. 4.

TABLE III. Result Analysis with Various Dataset Sizes

| Experiment No. | Data set size | Learning Rate | No. of Epochs | Batch size | Accuracy % | Precision% | Recall % |
|---|---|---|---|---|---|---|---|
| 1 | 1200 | 0.01 | 500 | 32 | 58.6 | 58 | 58 |
| 2 | 1800 | 0.01 | 500 | 32 | 78 | 78 | 78 |
| 3 | 2400 | 0.01 | 500 | 32 | 98.7 | 98 | 98 |

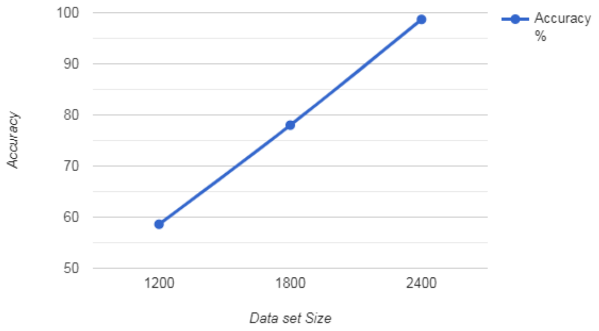


Fig. 4. Accuracy Measure for Various Dataset Sizes.

It is observed that CNN gives better performance with speech recognition compared to other related approaches described in literature review.

## V. Conclusion

Speech recognition is a very important and useful application to interact with machines using speech. This emerging field includes many challenges like dataset size, noisy environment, age, gender, dialect and mental state of the speaker and many more. Speech wav files for Gujarati digits are stored and features extracted with MFCC technique. In this research paper, digit recognition for Gujarati language is performed using CNN.

Speech is passed and processed as input into multiple convolutional layers. For model training Tensorflow and Keras are used with python. Maximum Accuracy rate of recognition out of three experiments with different datasets size for all digits 0 to 9 is found 98.7%. It is observed that accuracy for speech recognition models can be improved by increasing the size of the dataset. Deep learning approach implemented with CNN trained with a large dataset gives better performance. In the future the proposed model can be enhanced to observe results for major parameters such as age, gender and dialect of the speaker that affect the speech recognition result.

## References

[1] Daniel Jurafsky et al. Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd ed. draft). https: //web.stanford.edu/˜jurafsky/slp3. 2020.

[2] Tailor, J. H., & Shah, D. B. (2015). Review on Speech Recognition System for Indian Languages. International Journal of Computer Applications, 119(2).

[3] Tailor, J. H., & Shah, D. B. (2016). Speech recognition system architecture for Gujarati language. International Journal of Computer Applications, 138(12).

[4] Wazir, A. S. M. B., & Chuah, J. H. (2019, June). Spoken arabic digits recognition using deep learning. In 2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS) (pp. 339-344). IEEE.

[5] Gunawan, A. (2010). English digits speech recognition system based on hidden markov models “. In Proceedings of International Conference Computer.

[6] Chapaneri, S. V., & Jayaswal, D. J. (2013). Efficient speech recognition system for isolated digits. IJCSET, 4(3), 228-236.

[7] Datta RKS., Rudresh, M. D., & Shashibhushsan, G. (2021). Comparative performance analysis for speech digit recognition based on MFCC and vector quantization. Global Transitions Proceedings, 2(2), 513-519.

[8] Kurian, C., & Balakrishnan, K. (2013). Connected digit speech recognition system for Malayalam language. Sadhana, 38(6), 1339-1346.

[9] Sen, O., & Roy, P. (2021, September). A Convolutional Neural Network Based Approach to Recognize Bangla Spoken Digits from Speech Signal. In 2021 International Conference on Electronics, Communications and Information Technology (ICECIT) (pp. 1-4). IEEE.

[10] Zerari, N., Abdelhamid, S., Bouzgou, H. & Raymond, C. (2019). Bidirectional deep architecture for Arabic speech recognition. Open Computer Science, 9(1), 92-102. https://doi.org/10.1515/comp-2019-0004.

[11] Dhanashri, D., & Dhonde, S. B. (2017). Isolated word speech recognition system using deep neural networks. In Proceedings of the international conference on data engineering and communication technology (pp. 9-17). Springer, Singapore.

[12] Hamidi, M., Satori, H., Zealouk, O., & Satori, K. (2020). Amazigh digits through interactive speech recognition system in noisy environment. International Journal of Speech Technology, 23(1), 101-109.

[13] Swedia, E. R., Mutiara, A. B., & Subali, M. (2018, October). Deep learning long-short term memory (LSTM) for Indonesian speech digit recognition using LPC and MFCC Feature. In 2018 Third International Conference on Informatics and Computing (ICIC) (pp. 1-5). IEEE.

[14] Sustika, R., Yuliani, A. R., Zaenudin, E., & Pardede, H. F. (2017, November). On comparison of deep learning architectures for distant speech recognition. In 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE) (pp. 17-21). IEEE.

[15] Audichya, M.A., & Saini, J.R. (2017, September). A Study to Recognize Printed Gujarati Characters Using Tesseract OCR. International Journal for Research in Applied Science and Engineering Technology, 5(9), 1505-1510. doi: 10.22214/ijraset.2017.9219.