




A Novel Bangla Spoken Numerals Recognition System Using Convolutional Neural Network

Ovishake Sen^(✉) , Pias Roy, and Al-Mahmud

Department of Computer Science and Engineering, Khulna University of Engineering and Technology, Khulna 9203, Bangladesh
sen1607066@stud.kuet.ac.bd, mahmud@cse.kuet.ac.bd

Abstract. Automatic speech recognition (ASR) converts human speech into text or words that can be understood and classified easily. Only digits from '০-৯' (0-9) were used in the few studies on Bangla number recognition systems, which completely ignored duo-syllabic and tri-syllabic numbers. Audio samples of '০-৯৯' (0-99) Bangla spoken numbers from Bangladeshi citizens of various genders, ages, and languages were used to construct a speech dataset of spoken numbers in this work. Time shift, speed tuning, background noise mixing, and volume tuning are among the audio augmentation techniques used on the raw speech data. Then, to extract meaningful features from the data, Mel Frequency Cepstrum Coefficients (MFCCs) are used. This research developed a Bangla number recognition system based on Convolutional Neural Networks (CNNs). Our proposed dataset includes the diversity of speakers in terms of age, gender, dialects and other criteria. The proposed method recognizes '০-৯৯'(0-99) Bangla spoken numbers with 89.61% accuracy across the entire dataset. The model's efficacy was also evaluated using a 10-fold cross-validation procedure, with 89.74% accuracy for recognizing '০-৯৯' (0-99) Bangla spoken numbers across the entire dataset. This proposed method is also compared to some existing works in the field of recognizing spoken digits classes, demonstrating its dominance.

Keywords: Bangla speech recognition · Bangla spoken '০-৯৯' (0-99) numbers classification · CNN · MFCC · Cross-validation

1 Introduction

The capacity of electronic devices or systems to interpret spoken words is referred to as speech recognition. Speech recognition software uses speech data as a user interface to communicate with computers and comprehend the data. Speech recognition technology has become increasingly widespread in recent years. Because of the various benefits it brings, technology is widely employed by everyone, from corporations to people [4].

The enhanced communication by eliminating illegible hand-writing is one of the most noticeable advantages of voice recognition technology. Fast re-examination of documents, time saved due to increased efficiencies and reduced paperwork, speech recognition software can produce documents within less than half the time it takes to type them, multiple tasks, ability to share files on multiple devices, workflow visibility allows for better priorities management, turnarounds, etc. [1].

Speech recognition systems may be classified into isolated, linked, continuous, and spontaneous classifications based on the sorts of utterances they can recognize [22]. Isolated word recognizers generally demand silence on both sides of the sample window for each utterance. Connected word systems are similar to isolated words in that they allow distinct utterances to run concurrently with little wait time. Users may talk almost normally while the computer uses continuous speech recognizers to identify the content. Spontaneous speech is defined as natural-sounding and unrehearsed speaking [22].

Speech recognition systems rely largely on the speaker's age, gender, quantity, domicile, mode, and other characteristics. It may also be categorized depending on vocabulary size, accents, identification algorithms, and so on. Because each language has its unique phonemes, there are significant variations between languages. As a result, no one can guarantee that a recognition system that works well for one language, like English, would also work well for another, like Bangla [13, 28].

Despite being the world's seventh most spoken language, only a modest amount of substantial research has been done on Bangla.

'০-৯৯' numbers work as basic fundamental numbers to represent all other Bangla numbers. For example, one can utter '১২২' in Bangla as 'এক শত বাইশ'. Here '১২২' is expressed with three words where 'এক' and 'বাইশ' represents '১' and '২২' respectively, and 'শত' indicates positional value for '১'. That's why, in this paper we've proposed a Bangla spoken number recognition system that can classify '০-৯৯' Bangla spoken numbers with reasonable accuracy. The most popular deep learning and machine learning techniques for speech recognition include Artificial Neural Networks (ANN), Long Short Term Memory (LSTM), Gated Recurrent Units (GRU), Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and Backpropagation Neural Networks (BPNN) [25]. Our proposed voice recognition system is created based on the self developed dataset of Bangla spoken digits containing gender, accent, and age criteria. This research aims to develop a Convolutional Neural Network (CNN) model that can recognize Bangla numerals from the speech input.

2 Related Work

A great deal of work has been done with digits in different languages. Graves et al. [10] claimed that 88% accuracy was discovered for recognizing any digit and that using conventional pre-processing increased the accuracy to 99.4%. Graves et al. [11] claimed that Bidirectional LSTM outperformed standard RNN, Multilayer Perceptron, and other unidirectional networks.

For recognizing Hindi digits, Saxena et al. [24] created an HMM-based model. According to them, training data revealed 94.09% word-level accuracy, and testing data revealed 85% word-level accuracy. On training data, 92.82% phone level accuracy was discovered, and on testing data, 86.17% phone level accuracy was discovered. Renjith et al. [23] discovered that a well-trained database resulted in a proper recognition system. They further stated that when the number of combinations increases, the system's accuracy will grow in each state. Netshiombo et al. [18] used the MFCC and Kaldi toolkit to develop a spoken digit recognizer for the under-resourced languages of South Africa. Taufik et al. [32] created an Automated Visual Acuity Test system that used Digit Recognition technology to extract features and a Convolutional Neural Network to recognize spoken digits. They claimed to be able to recognize digits with 91.4% accuracy.

In comparison to English or other rich languages, there have been few studies on Bangla speech recognition. Again, no standard Bangla speech corpus has been discovered, complicating research on Bangla speech recognition [16]. Paul et al. [20] presented a Bangla speech recognition system utilizing Pre-emphasis filtering, speech coding, LPC, and ANN. Sultana et al. [29] proposed a method for converting Bangla speech to text using the Speech Application Programming Interface (SAPI) [8]. An XML grammar file for SAPI was created with English character combinations for each Bangla word, and the average recognition rate for repeated words and different names was 78.54% and 74.81%, respectively. Ali et al. [7] proposed a technique for recognizing spoken words in Bangla where Mel-frequency cepstral coefficients(MFCC), LPC, and GMM were used for feature extraction, template matching, and DTW were used for matching the speech signal. MFCC and DTW achieved 78% accuracy, LPC and DTW achieved 60% accuracy, MFCC and GMM yielded 84% accuracy, and MFCC, LPC, and DTW achieved 50% accuracy.

Using the Hidden Markov Modeling Toolkit (HTK) [9], Hasnat et al. [12] developed a method for constructing an isolated and continuous Bangla voice recognition system. The isolated voice recognition model had a speaker-dependent accuracy of 70% and a speaker-independent accuracy of 90%, whereas the continuous speech recognition model had an accuracy of 80% and 60%, respectively. Ahammad et al. [5] suggested a connected digit recognition system employing Backpropagation neural network and segmenting the raw audio input. They got an average of 89.87% accuracy recognizing 0 to 9 Bangla spoken digits.

Using a double-layered LSTM-RNN approach, Nahid et al. [17] developed a technique for constructing a Bangla Speech Recognition system, with a 28.7% phon detection error rate and a 13.2% word detection error rate. Using HMM CMU Sphinx and Android Text-to-Speech (TTS) API, Ahmed et al. [6] presented a voice input speech output calculator that can identify isolated and continuous speech in the Bangla language and extract mathematical equations. The accuracy of word recognition was 86.7% in this case. Sumit et al. [30] developed a strategy to recognize continuous Bangla speech for noisy situations by Aligning and segmenting the audio clips and employing a multi-layer neural network, CNN, RNN, GRU, and FC layers.

To recognize Bangla’s short speech instructions, Sumon et al. [31] presented three CNN architectures: an MFCC-based CNN model, a raw CNN model, and a pre-trained CNN model utilizing transfer learning. The MFCC model has a 74% accuracy, and the proposed raw model has a 71% accuracy, and the proposed transfer model has a 73% accuracy. Islam et al. [14] introduced a speech recognition system in the Bengali language using CNN and built RNN based method to identify the Bengali character level probability again 86.058% accuracy was found in the CNN model. Shuvo et al. [28] suggested a Bangla number recognition system from speech signals using MFCC and CNN, with a test set recognition accuracy of 93.65%.

Sharmina et al. [27] presented an in-depth learning method for identifying Bengali spoken digits using MFCC and CNN, 98.37% accuracy, 98.37% precision, 98.37% recall, and 98.37% F1-score. Paul et al. [21] suggested a Bangla numeral recognition system from speech signals utilizing MFCC and GMM, and their self-built Bangla numeral data set produced 91.7% correct predictions.

3 Preliminaries

3.1 Convolutional Neural Network (CNN)

The convolution neural network algorithm is a multi-layer perceptron that is used to recognize and classify two-dimensional data by identifying patterns in the data. An input layer, hidden layers, and an output layer make up CNN’s architecture. Convolution between the kernel and the input matrix of the layer is performed by hidden layers [15]. A basic structure of CNN is shown in Fig. 1.

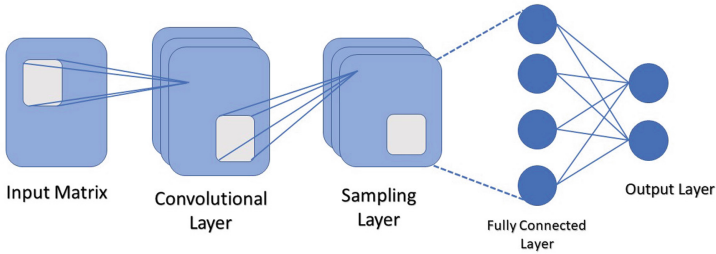


Fig. 1. Basic structure of CNN.

The two main processes of a convolutional neural network are convolution and sub-sampling. A trainable filter is used to convoluted the input feature map and then apply a bias value to it during the convolution process. The convoluted feature map(CFM) is calculated using Eq. (1). Here \sum denotes summation operation and \otimes denotes convolution operation.

$$CFM_{x,y} = f(b + \sum_{i=1}^{k_h} \sum_{j=1}^{k_w} K_{i,j} \otimes I_{x+i,y+j}) \tag{1}$$

Some procedures, such as max and average pooling, are done on a selected pooling region of CFM during the sub-sampling process. The sub-sampled feature map(SFM) is calculated using Eq. (2). Here \sum denotes summation operation.

$$SFM_{x,y} = d\left(\sum_{i=1}^R \sum_{j=1}^C CFM_{i,j}\right) \quad (2)$$

3.2 Mel Frequency Cepstral Coefficients (MFCCs)

Mel Frequency Cepstral Coefficients (MFCCs) are an often used feature in artificial speech and speaker recognition systems. Because the frequency bands on the mel-scale are evenly spaced, MFCCs can more effectively imitate the human auditory system's response [3]. The following steps are used to calculate MFCCs in general [2]:

1. Break the signal up into little frames.
2. For each frame calculate the periodogram estimation of the power spectrum.
3. Apply the mel filterbank to the power spectra.
4. Take the logarithm of all filterbank energies.
5. Calculate the log filterbank energies' DCT.
6. Save DCT coefficients 2–13 and toss the rest.

4 Proposed Method

In this study, a Bangla numerical speech recognition system has been proposed using deep learning models. Initially, a speech corpus of Bangla numbers from '০-৯৯' was created. The raw speech data is then preprocessed to remove noise and silence from the signals. The raw speech data were then subjected to various audio augmentation techniques. After that, the MFCC features are extracted from the speech data that has been processed. The dataset is then divided into three sets: train, validation, and test. The train and validation data are used to train our proposed CNN model in the training phase and got the train and validation accuracy of the dataset. The processed test data is fed into the CNN model during the testing phase of the proposed CNN model to obtain the predicted output classes and the CNN model's test accuracy for the dataset. We also used the cross-validation technique to check the effectiveness of the model's performance.

Figure 2 is showing the Basic structure of the proposed deep learning based model for numerical speech recognition.

4.1 Dataset Description

There is no public dataset available for Bangla numbers speech data. So a speech dataset of '০-৯৯' Bangla numbers has been created from 19 speakers for this research. The speakers are of various ages and come from various parts

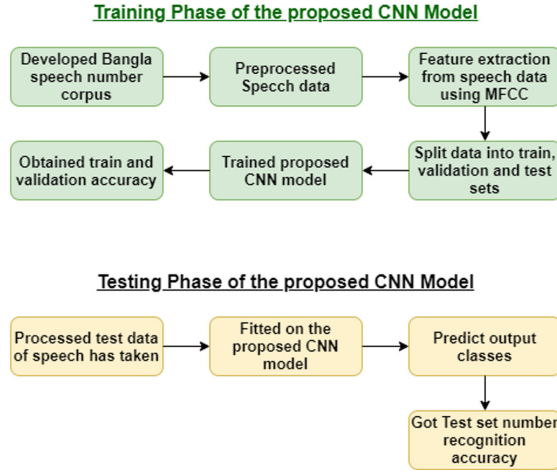


Fig. 2. Basic structure of the proposed deep learning based model for numerical speech recognition

of Bangladesh and the dataset includes contributions from both male and female speakers. Gender, age, and dialect all influence how a word is said. As a result, collecting voice recordings from persons of various Bengali dialects, genders, and age groups aided in the diversification of this dataset. Approximately 400 audio samples were recorded in both noisy and quiet environments for each number. So, a total of 40,000 audio samples were recorded. Each audio sample was recorded with a microphone using the ‘audacity’ recording software, with a sampling rate of 44 kHz and a wav audio format.

Table 1 shows the speakers description for constructing the dataset.

Table 1. Speakers description for constructing the dataset

Places of Bangladesh	No. of speakers
Bagerhat	3
Chittagong	2
Dhaka	2
Dinajpur	1
Mymensingh	1
Narayanganj	4
Rajshahi	1
Sylhet	1
Thakurgaon	4
Total	19

4.2 Preprocessing and Augmentation

The raw speech recordings were trimmed to obtain isolated number speech using the ‘audacity’ recording software. The raw speech data was subjected to audio augmentation techniques such as time shift, speed tuning, background noise mixing, and volume tuning [19]. The raw speech data are gathered from a variety of noisy environments. First, using Python’s ‘noise reduce’ library, the noise was removed from the raw speech data. The stereo audio files were then converted to mono audio files in order to improve the performance of the speech recognition model. Then, using Python’s pyDub library, silence from the audio signal was removed.

4.3 Feature Extraction

One of the most crucial steps in the development of a speech recognition system is feature extraction. The primary goal of feature extraction is to convert the speech waveform into a distinct parametric representation to extract meaningful features from the speech data. Mel Frequency Cepstrum Coefficient (MFCC) is used to extract features in this study. The MFCC features were extracted from the numerical speech data and saved into an array using Python’s Librosa library. Then, based on the speech data, the collected speech data was labelled as ‘০-৯৯’. The first 13 MFCC coefficients and their first and second delta values were used to recognize numerical Bangla speeches.

Figure 3 shows the spectrogram of mfcc features for “৯” number.

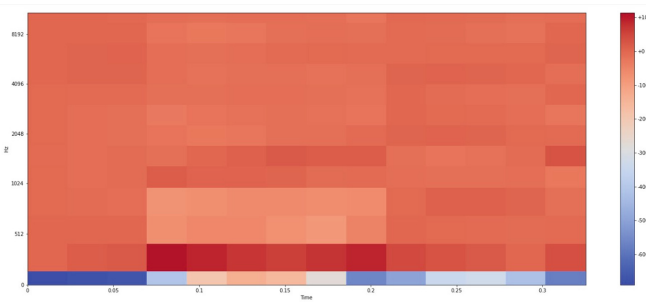


Fig. 3. Spectrogram of mfcc feature for “৯” number

4.4 Train Test Split

From the whole dataset, 64% of data are used for training purposes, 16% data used for validation purposes and the rest of 20% data are used for testing purposes. So, a total of 25600 audio samples are used for training, with 256 audio

samples for each number, a total of 6400 audio samples are used for validating, and a total of 8000 audio samples are used for testing the proposed model, with 64 audio samples for each number. We used Python’s’scikit-learn’ library for this train-test split.

4.5 Feature Learning and Classification Using CNN

The Convolutional Neural Network (CNN) architecture has been used for feature learning and classification purposes in recognizing '০-৯৯' spoken Bangla numbers. This model architecture was built with Keras and Tensorflow.

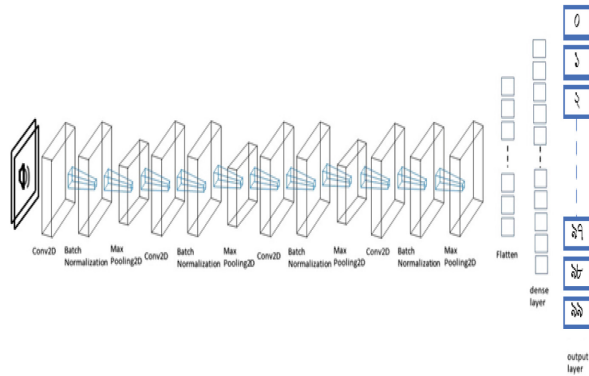


Fig. 4. Proposed Convolutional Neural Network Model Architecture

The CNN model’s input shape is (39*39*1). The preprocessed train data was fed into a 2D-convolutional layer (Conv2D) with a filter size of 24 and a kernel size of (3, 3). Again relu activation function and L2 regularizer with a regularization factor of 0.1 was also used in this input layer. The output of this layer was then passed to the BatchNormalization layer, which was then passed to the MaxPooling layer with a pool size of (2, 2). This process of combining Conv2D-BatchNormalization-MaxPooling layers was repeated three times, with 32, 64, and 128 filter sizes used in each Conv2D layer.

The flatten layer converts the output vector to a one-dimensional vector. Then, the 1D vector is connected to a fully connected layer of 128 units with a relu activation function. To avoid overfitting, a Dropout layer with a dropout rate of 0.2 was used. Finally, softmax layer is used for classification. The categorical cross entropy loss function was used. Also adam optimizer was used. A learning rate of 0.0001 was employed. The result is a prediction of the '০-৯৯' classes of Bangla numbers.

Figure 4 shows the proposed Convolutional Neural Network Model Architecture.

5 Experimental Results and Analysis

The proposed system has been tested with about 40,000 instances of the '০-৯৯' Bangla spoken numbers self-created dataset. Python, Keras, and Tensorflow are used to implement the model. The experiments were done on Google Colaboratory which is a free online cloud-based Jupyter notebook environment.

The recognition accuracy of spoken Bangla numbers are calculated using the following formula:

$$Accuracy = \frac{Correctly\ recognized\ numbers}{Recognized + Unrecognized\ Numbers} * 100\% \quad (3)$$

In this study, 4 experiments were done on the proposed method using different data sizes. Firstly, 100 instances per class were taken and 66.07% accuracy was obtained with 66.07% precision, 66.07% recall, and 66.07% f1-score. . Secondly, 200 instances per class were taken and 79.69% accuracy was obtained with 79.69% precision, 79.69% recall, and 79.69% f1-score. Thirdly, 300 instances per class were taken and 85.68% accuracy was obtained with 85.68% precision, 85.68% recall, and 85.68% f1-score. . Finally, 400 instances per class were taken and 89.61% accuracy was obtained with 89.61% precision, 89.61% recall, and 89.61% f1-score. So we can presume that even higher accuracy can be obtained with this model if more data are being fed into this model.

Table 2 shows the experiments details and results that has been performed on our proposed CNN model. Figure 5 shows the improvement of results after providing more data into the proposed CNN model.

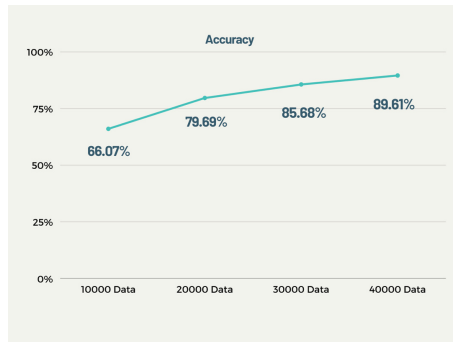


Fig. 5. Results improvement after providing more data into the proposed CNN model

At the time of working with spoken Bangla numbers, we have seen people from different areas pronounce a arbitrary number differently. For example, some people pronounce 88 by 'অষ্টআশী' and some other pronounces 88 by 'আটাশী', again some people pronounce 87 by 'সপ্তআশী' and some other pronounces 87 by 'সাতাশী',

Table 2. Experiments details and Results on proposed CNN model

Exp no.	Data Size instances	No. of epochs	Batch Size	Learning Rate	Accuracy (%)	Loss	Precision (%)	Recall (%)	F1- Score (%)	Training Time (hour)
1	10000	1000	32	0.0001	66.07	1.80	66.07	66.07	66.07	0.5
2	20000	1000	32	0.0001	79.69	1.06	79.69	79.69	79.69	1
3	30000	1000	32	0.0001	85.68	0.73	85.68	85.68	85.68	2
4	40000	1000	32	0.0001	89.61	0.52	89.61	89.61	89.61	3.5

Table 3. Recognition accuracy of each Bangla spoken numbers '০-৯৯' on proposed CNN model

Number	Accuracy (%)	Number	Accuracy (%)	Number	Accuracy (%)	Number	Accuracy (%)	Number	Accuracy (%)
0	95.83	20	83.53	40	80.0	60	83.78	80	95.65
1	93.75	21	88.42	41	97.10	61	86.91	81	94.67
2	96.10	22	91.67	42	98.65	62	87.95	82	95.95
3	95.23	23	86.75	43	86.11	63	93.33	83	93.75
4	100.00	24	96.0	44	90.29	64	90.91	84	92.0
5	92.68	25	85.71	45	92.31	65	87.65	85	87.65
6	86.59	26	92.96	46	87.06	66	92.00	86	90.47
7	81.94	27	88.46	47	90.70	67	95.18	87	95.71
8	91.67	28	90.14	48	81.01	68	89.89	88	91.76
9	92.41	29	78.57	49	88.61	69	93.15	89	88.0
10	87.5	30	79.52	50	97.47	70	94.74	90	94.52
11	81.82	31	83.10	51	75.28	71	89.74	91	74.12
12	89.16	32	94.29	52	91.78	72	92.59	92	81.15
13	96.10	33	94.25	53	95.35	73	89.02	93	86.42
14	80.82	34	85.71	54	94.29	74	96.93	94	95.18
15	80.20	35	90.29	55	91.46	75	91.83	95	84.71
16	86.84	36	85.19	56	95.12	76	97.75	96	92.77
17	84.71	37	93.90	57	80.26	77	90.36	97	84.51
18	88.89	38	97.14	58	84.06	78	93.51	98	88.16
19	85.00	39	86.91	59	89.47	79	87.84	99	90.12

again 44 is pronounced differently as ‘চওউচল্লীশ’ and ‘চুয়াল্লিশ’, again 20 is pronounced differently ‘বিশ’ and ‘কুড়ি’, again some people pronounce 60 by ‘ষাইট’ and some other pronounces 60 by ‘ষাট’, again some people pronounce 37 by ‘সায়ত্রীশশ’ and some other pronounces 37 by ‘সাত্রীশশ’, again 45 is pronounced differently ‘পয়তাল্লীশশ’ and ‘পাঁচল্লীশশ’, again 55 is pronounced differently ‘পঞ্চগন্ন’ and ‘পাঁচপান্ন’, again some people pronounce 53 by ‘তীপান্ন’ and some other pronounces 53 by ‘তীরপান্ন’, again at the time of pronouncing 29,39,49,59,79,89 some people loudly utters ‘উনো’ and some other speaks softly ‘উন’. Again some numbers are phonetically very close to other numbers when they are pronounced. Table 3 shows the recognition accuracy of each Bangla spoken numbers '০-৯৯' on proposed CNN model.

Table 4. Comparative analysis between proposed method and previous existing approaches

Article	Dataset description	Methods	Accuracy on test set
Ahmed et al. [6]	7 speakers and 2,733 instances	CMU Sphinx and Android TTS API	~86.7% for 0–9 digits.
Paul et al. [21]	1,000 instances	MFCC and GMM.	91.7% for 0–9 digits.
Ahammad et al. [5]	30 speakers and Noise-free 260 instances	Segmentation and BPNN.	89.87% for 0–9 digits.
Nahid et al. [17]	15 speakers and Noisy 2000 instances	Noise reduction, phoneme mapping and LSTM	28.7% phon detection error rate and 13.2% word detection error rate for 0–9 digits.
Shuvo et al. [28]	120 speakers and Noise free 6,000 instances	CNN	93.65% for 0–9 digits.
Sharmina et al. [27]	5 speakers and 1,230 instances	CNN	98.37% for 0–9 digits.
Ovishake et al. [26]	Nineteen talkers and 4,000 audio samples	CNN	97.1% for only 0–9 digits.
Proposed method	19 speakers from different gender, age groups and areas of Bangladesh and 40,000 noisy audio samples for '০-৯৯' Bangla numbers	MFCC, CNN and cross-validation	89.61% for recognizing '০-৯৯' numbers on CNN model and 89.74% for recognizing '০-৯৯' numbers on 10-fold cross-validation

Table 4 shows the Comparative analysis between proposed method and previous existing approaches. From Table 4, we can see that the existing research works that have been done on recognizing and classifying spoken numbers are limited to only '০-৯' spoken Bangla numbers. But in this work, we developed our model to recognize Bangla spoken '০-৯৯' numbers. We have also added variations on our dataset by considering the gender, age-groups, dialects, etc. parameters. Again we have tested our model on a huge self-created Bangla spoken numbers dataset.

6 Future Work

In this study, we attempt to categorize Bangla spoken numerals from '০-৯৯'. We have built a dataset of '০-৯৯' spoken numbers for this research because there is no standard Bangla spoken numbers dataset. The speech data was gathered from 19 people of various genders, dialects, and ages. However, this dataset does not include all dialects of the Bengali language. In the future, we hope to collect more speech data from people of all ages in different parts of Bangladesh. CNN and cross-validation provide satisfactory performance for categorizing Bengali spoken numbers; however, we will try to construct a better model in the future. In the future, this categorization model might be used to construct a speech input calculator that can execute simple arithmetic operations instantly by receiving user input as spoken numbers.

7 Conclusion

As technology advances, the use of speech recognition technologies in various aspects of life is becoming more common. The speech recognition system in Bangla is no exception. This paper aims to develop a Convolutional Neural Network (CNN) model that can recognize Bangla numerals from the speech input. In this paper, we've seen a lot of variations on spoken '০-৯৯' numbers pronunciation from different genders, age groups, and dialects of Bangladeshi people so more research works should be done in this field. The proposed method achieves an overall accuracy of 89.61%, and the model's effectiveness was tested again using 10-fold cross-validation, yielding an overall accuracy of 89.74% for recognizing '০-৯৯' Bangla spoken numbers across the entire dataset. This proposed method is also compared to some existing works in the field of recognizing '০-৯' digits classes, demonstrating its dominance.

References

1. 12 benefits of speech to text. <https://www.dataworxs.com.au/12-benefits-speech-text>. Accessed 30 June 2021
2. Mel frequency cepstral coefficient (MFCC) tutorial. <http://www.practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs>. Accessed 28 June 2021
3. Mel-frequency cepstrum. <https://en.wikipedia.org/wiki/Mel-frequencycepstrum>. Accessed 28 June 2021
4. speech recognition. <https://searchcustomerexperience.techtarget.com/definition/speech-recognition>. Accessed 30 June 2021
5. Ahammad, K., Rahman, M.M.: Connected Bangla speech recognition using artificial neural network. *Int. J. Comput. Appl.* **149**(9), 38–41 (2016)
6. Ahmed, T., Wahid, M.F., Habib, M.A.: Implementation of Bangla speech recognition in voice input speech output (viso) calculator. In: 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), pp. 1–5. IEEE (2018)
7. Ali, M.A., Hossain, M., Bhuiyan, M.N., et al.: Automatic speech recognition technique for bangla words. *Int. J. Adv. Sci. Technol.* **50** (2013)
8. Chung, T.D., Drieberg, M., Hassan, M.F.B., Khalyasmaa, A.: End-to-end conversion speed analysis of an FPT. AI-based text-to-speech application. In: 2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech), pp. 136–139. IEEE (2020)
9. Gales, M., Young, S.: The application of hidden Markov models in speech recognition (2008)
10. Graves, A., Beringer, N., Schmidhuber, J.: A comparison between spiking and differentiable recurrent neural networks on spoken digit recognition. In: The 23rd IASTED International Conference on Modelling, Identification, and Control (2004)
11. Graves, A., Schmidhuber, J.: Framework phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **18**(5–6), 602–610 (2005)
12. Hasnat, M., Mowla, J., Khan, M., et al.: Isolated and continuous Bangla speech recognition: implementation, performance and application perspective (2007)

13. Hossain, S., Rahman, M., Ahmed, F., Dewan, M.: Bangla speech synthesis, analysis, and recognition: an overview. Proc, NCCPB (2004)
14. Islam, J., Mubassira, M., Islam, M.R., Das, A.K.: A speech recognition system for Bengali language using recurrent neural network. In: 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), pp. 73–76. IEEE (2019)
15. Liu, T., Fang, S., Zhao, Y., Wang, P., Zhang, J.: Implementation of training convolutional neural networks. arXiv preprint [arXiv:1506.01195](https://arxiv.org/abs/1506.01195) (2015)
16. Muhammad, G., Alotaibi, Y.A., Huda, M.N.: Automatic speech recognition for bangla digits. In: 2009 12th International Conference on Computers and Information Technology, pp. 379–383. IEEE (2009)
17. Nahid, M.M.H., Purkaystha, B., Islam, M.S.: Bengali speech recognition: a double layered LSTM-RNN approach. In: 2017 20th International Conference of Computer and Information Technology (ICCIT), pp. 1–6. IEEE (2017)
18. Netshiombo, D., Mokgonyane, T.B., Manamela, M.J., Modipa, T.I.: Spoken digit recognition system for an extremely under-resourced language
19. Park, D.S., et al.: SpecAugment: a simple data augmentation method for automatic speech recognition. arXiv preprint [arXiv:1904.08779](https://arxiv.org/abs/1904.08779) (2019)
20. Paul, A.K., Das, D., Kamal, M.M.: Bangla speech recognition system using LPC and ANN. In: 2009 Seventh International Conference on Advances in Pattern Recognition, pp. 171–174. IEEE (2009)
21. Paul, B., Bera, S., Paul, R., Phadikar, S.: Bengali spoken numerals recognition by MFCC and GMM technique. In: Mallick, P.K., Bhoi, A.K., Chae, G.-S., Kalita, K. (eds.) *Advances in Electronics, Communication and Computing*. LNEE, vol. 709, pp. 85–96. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-8752-8_9
22. Reddy, D.R.: Speech recognition by machine: a review. Proc. IEEE **64**(4), 501–531 (1976)
23. Renjith, S., Joseph, A., Anish Babu K.K.: Isolated digit recognition for Malayalam-an application perspective. In: 2013 International Conference on Control Communication and Computing (ICCC), pp. 190–193. IEEE (2013)
24. Saxena, B., Wahi, C.: Hindi digits recognition system on speech data collected in different natural noise environments. In: International Conference on Computer Science, Engineering and Information Technology (CSITY 2015) February, pp. 14–15 (2015)
25. Sen, O., et al.: Bangla natural language processing: a comprehensive analysis of classical, machine learning, and deep learning based methods. IEEE Access **10**, 38999–39044 (2022)
26. Sen, O., Roy, P., et al.: 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 (2021)
27. Sharmin, R., Rahut, S.K., Huq, M.R.: Bengali spoken digit classification: a deep learning approach using convolutional neural network. Procedia Comput. Sci. **171**, 1381–1388 (2020)
28. Shuvo, M., Shahriyar, S.A., Akhand, M.: Bangla numeral recognition from speech signal using convolutional neural network. In: 2019 International Conference on Bangla Speech and Language Processing (ICBSLP), pp. 1–4. IEEE (2019)
29. Sultana, S., Akhand, M., Das, P.K., Rahman, M.H.: Bangla speech-to-text conversion using SAPI. In: 2012 International Conference on Computer and Communication Engineering (ICCCE), pp. 385–390. IEEE (2012)

30. Sumit, S.H., Al Muntasir, T., Zaman, M.A., Nandi, R.N., Sourov, T.: Noise robust end-to-end speech recognition for Bangla language. In: 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), pp. 1–5. IEEE (2018)
31. Sumon, S.A., Chowdhury, J., Debnath, S., Mohammed, N., Momen, S.: Bangla short speech commands recognition using convolutional neural networks. In: 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), pp. 1–6. IEEE (2018)
32. Taufik, D., Hanafiah, N.: Autovat: an automated visual acuity test using spoken digit recognition with MEL frequency cepstral coefficients and convolutional neural network. *Procedia Comput. Sci.* **179**, 458–467 (2021)