Application of Transfer Learning in Low-Resource Language Processing: A Case Study on Bangla Numeral Recognition

Abstract Speech recognition refers to the process of transforming human speech into text or other machine-readable formats. This paper presents a comprehensive exploration into Bangla Numeral Recognition from Speech Signal. This work features a combination of authentic and augmentation that significantly reduces the dependency on raw data specially for a low-resource language like Bangla. This study uses Transfer Learning and Convolutional Neural Networks (CNNs) to create a Bangla number identification system based on voice signals. The dataset includes 19,906 individual audio files for '০-১৯৯' (0-199) Bangla numerals where numerals '০-৯৯' (0-99) are taken from a pre-existing dataset, while numbers '১০১-১৯৯' (101-199) are augmented. The methods are designed around the extraction and analysis of log Mel spectrograms, capitalizing on their ability to represent complex audio structures in a manner that is both comprehensive and conducive to CNN architectures including DenseNet, ResNet and VGG. Among our experiments, ResNet performed most significantly with an accuracy of 95.58% recognizing '০-১৯৯' (0-199) Bangla spoken numbers across the entire dataset.

Keywords: Speech Processing · Transfer Learning · Bangla Numerals · Speech Recognition

1 Introduction

In everyday communication, speech is the most natural and efficient method used. The development of speech recognition research has evolved from experimental demonstrations in controlled environments to practical applications in real-world scenarios. Speech recognition is the process of converting acoustic signals into textual information or commands, allowing machines to understand and respond to spoken language. Despite significant advancements in the design of speech recognition systems, there are still challenges in achieving robust recognition in low-resource environments [3]. The categorization of speech recognition systems based on different factors, such as utterance elements and speaker modes, highlights the intricate nature of this field.

As of 2021 [26], approximately 240 million people use Bangla as native language and 41 million people use it as their secondary language. Bangla ranks fifth in most spoken first language and globally, the sixth most spoken language in terms of speakers. However, the lack of a standardized Bangla speech corpus has presented obstacles for researchers. Numerals are a significant part of it and the research in this field is unsatisfactory. Numerous approaches, including Artificial Neural Networks (ANN), Linear Predictive Coding (LPC) and Hidden Markov Models (HMM) have been suggested for Bangla speech recognition [18, 7], covering areas such as phenome, continuous speech and numeral recognition.

The motive of this study is to improve the field of Bangla speech numeral recognition through the Convolutional Neural Networks (CNNs) system utilizing Transfer Learning in audio recognition. Additionally, Audio to Image classification and reduction of reliability on physical data for a low resource language, Bangla. Unlike previous methods that mainly utilize 1D inputs such as root mean square(RMS), chroma, zero-crossing rate (ZCR), tempo, Mel-frequency cepstral coefficients (MFCCs), etc., the proposed system utilizes 2D Log-Mel spectrograms as features, which improves recognition performance significantly.

Section II contains a literature review, Section III elaborates on our proposed methodology which includes description of the dataset and the proposed architecture. Section IV presents the experimental results and analysis, Section V concludes the study by summarizing the main contributions and insights.

2 Literature Review

In comparison to well-researched languages like English, the field of Bangla speech recognition, both isolated and continuous types of speech, remains relatively under-explored.

O. Sen et al. [21] introduced a novel method for recognizing Bangla spoken digits in the range ‘০-৯৯’ (0-99) using a convolutional neural network (CNN). Their approach utilized a dataset comprising 400 samples per digit, which included both noisy and noise-free environments. Feature extraction was performed using Mel Frequency Cepstrum Coefficients (MFCCs), a popular technique in speech processing that captures the timbral aspects of audio signals.

Expanding on the realm of numeral recognition, Shuvo et al. [23] devised a convolutional neural network-based technique specifically tailored for converting spoken Bangla numerals into text. This system demonstrated significant advancements in accuracy and processing speed. Meanwhile, Rahu et al. [17] presented a strategy for recognizing Bengali spoken numerals employing a combination of MFCC and Gaussian Mixture Models (GMM). GMMs are often used for their efficacy in modeling the statistical characteristics of complex data such as audio.

Further contributions to the field include those by B. Paul et al.[22], who also applied MFCC and GMM technologies but leveraged a proprietary dataset of 10 Bangla numerals in the range ’১-১০’ (1-10). Their method achieved a notable accuracy rate of 91.7%, underscoring the potential of customized datasets in enhancing model performance.

On a different note, Ahammad et al. [1] explored connected digit recognition for Bangla. They employed a neural network, achieving an average accuracy of 89.87% in recognizing sequences of Bangla digits from zero to nine. This approach highlighted the effectiveness of neural networks in learning and predicting more complex patterns of spoken digits.

Nahid et al. [14] took a unique approach by implementing a double-layered model combining Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) for a Bangla Speech Recognition system. Their system showed phoneme and word detection error rates of 28.7% and 13.2%, respectively, indicating areas for further refinement. This approach was integrated with tools like HMM CMU Sphinx(Open source speech recognition toolkit) for speech-to-text conversion and Android TTS(text-to-speech) API for text-to-speech processes, showcasing a full-cycle speech interface.

Additionally, Islam et al. [12] developed a dual-system where a CNN-based method was used for initial vocabulary recognition, followed by an RNN-based system to calculate character-level probabilities for Bengali text. This combination harnessed the strengths of both convolutional and recurrent neural networks, enhancing the overall accuracy to 86.058% and reliability of the speech recognition process.

These studies collectively advance the field of Bangla speech recognition, each contributing unique methodologies and insights that pave the way for further innovations in processing and understanding Bangla spoken language.

3 Background Study

3.1 Audio Classification

CNN-based models have been applied for many tasks, including music genre classification [8, 5, 28], environment sound classification [9, 2, 7], and audio generation [16, 19]. Several 1-D convolution models, such as EnvNet [25] and Sample-CNN [13], have been created for processing raw audio waveforms. The majority of state-of-the-art findings were produced utilizing CNNs on Spectrograms. Many models use many models with distinct inputs and aggregate results to create predictions, complicating the design process. The research [14] employed three networks to analyze spectrograms and delta STFT coefficients, whereas the work [25] used MFCCs and mel-spectrograms as inputs for two networks. Our investigation shows that employing simple log mel-spectrograms might achieve in superior results.

3.2 Log-Mel Spectrogram

Speech is frequently characterized by a number of characteristics, including vocal harshness,variable pronunciation speeds, nasal tone, unexpected sounds, disfluencies, blurred phoneme transitions, prolonged phoneme duration, unnecessary pauses, and other difficulties such as pitch fluctuation and monotony, which reduce its intelligibility. Acoustic spectrograms, which display time-varying speech spectrum generated at brief acoustic intervals, can be used to assess speech intelligibility. These spectrograms, computed using short-time analysis, provide insights on distorted phonemes and energy differences in frequency bands. To construct the acoustic spectrogram of a voice signal $x(n)$, the following short-time analysis equation:

$$X_a(n_a, k'_a) = \sum_{l=-\infty}^{\infty} x(l) w_a(n_a - l) e^{-j \frac{2\pi k'_a}{N_a} l} \quad (1)$$

The acoustic domain is denoted by the subscript a, and the frame index n_a relates to the discrete-time index n with $n_a = \frac{n}{L_a}$, where L_a is the frame period of each sample, k'_a is the acoustic frequency bin index, and $w_a(n)$is the analysis window of N_a length.

Log-Mel spectrograms provide bio-inspired spectro-temporal representations of speech by converting acoustic spectrograms to mel-frequency and logarithmic scales. Figure 1 demonstrates the audio to spectrogram conversion process.

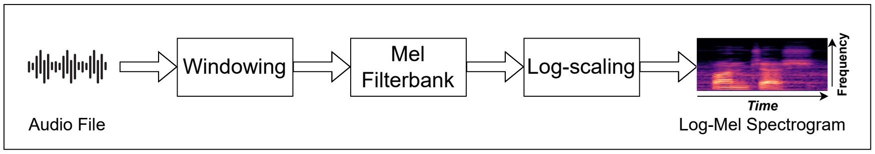

Figure 1. Step-by-Step Process of Converting Audio Signal to Log-Mel Spectrogram.

Although standard machine learning methods do not directly employ log-mel spectrograms because of probable information loss, their value in short-term analysis of speech has been demonstrated in our work, notably with CNN-based classifiers.

3.3 Transfer Learning

Transfer Learning involves extending models learned on a big dataset to extract important characteristics for a new task based on past knowledge. Deep models trained on huge datasets, such as ImageNet, have been popular for transfer learning tasks like image segmentation, and medical image analysis. C3D [4] achieved 88% performance on UCF-101 [24], whereas pre-training on ImageNet and Kinetics datasets yielded 98% performance. Our investigation shows the impact of pre-trained weights on audio classification due to the significant performance disparity.

3.4 Transfer Learning in Classification of Audio

Implementation of Transfer Learning in speech involves pre-training a model using huge audio datasets, such as AudioSet and Million Songs. This work[6] fine-tuned a simple CNN architecture fine-tuned specifically for the Million Song dataset for tasks like audio event classification and emotion prediction. In the work[11] large-scale CNN architectures such as ResNet, Inception, and VGG was used to classify audio files on the dataset, AudioSet. The models were trained using AudioSet, a popular platform for audio transfer learning [27, 15]. Previous research indicates that transfer learning in speech has mostly focuses on audio data. The architectures are huge and the characteristics are becoming more complicated. As indicated, This work [10] was one of the first works to employ ImageNet-pretrained models for audio categorization. Recent studies have utilized ImageNet-pretrained models for audio tasks. The publications did not appreciate the models' in-depth potential due to design alterations. Our research demonstrates that a single model and input characteristics can achieve state-of-the-art performance over several fields, minimizing the time and space complexity of creating audio categorization models Unlike other studies, our focus remained on transfer learning using large-picture datasets such as ImageNet.

4 Methodology

4.1 Dataset

Kaggle's ”Bangla Spoken 0-99 Numbers”[21] dataset was utilized for this experiment to recognize Bangla numerals spoken by diverse speakers. The dataset consists of 101 classes in the range '০-১০০' (0-100), each with 100 audio recordings. Each recording was in Waveform Audio File (.wav) format and each was 1s in duration. All data is made sure to contain no empty spaces. Audio samples were produced using ’audacity’ recording software at 44 KHz sampling rate. This dataset contains a total of 10,006 individual audio files. The audio was captured by 99 speakers and covers 9 different Bangladeshi dialects. The speakers include both men and women.

4.2 Data Augmentation

The existing dataset accounts for the numerals in the range'০-১০০' (0-100). Rest the of the data in the range '১০১-১৯৯' (101-199) was augmented from the same dataset. The augmentation was achieved by concatenating two existing audio files using the 'AudioSegment' module from pydub. For every numeral in the range '১০১-১৯৯' (101-199), 100 individual audio files were generated. This process resulted in a total of 19,906 audio files, approximately 100 for each numeral. Figure 2 demonstrates an example of this process.

4.3 Data Pre-processing

Logarithmic Mel-Spectrograms were generated for each audio in the collection using Python's 'Librosa Library'. The magnitude of the spectrogram is computed by dividing the time series input. Mel-Spectrograms were generated with a hop length of 128, 256 Mel-bins and a window of 2048 samples which is about

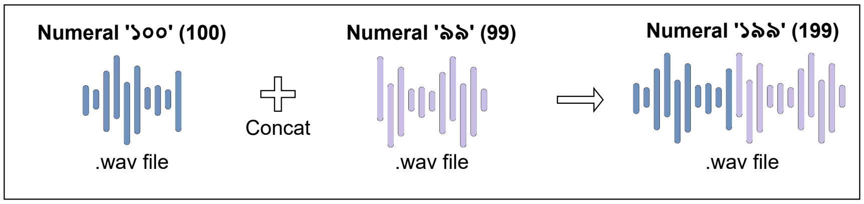

Figure 2. Illustration of Audio Data Augmentation by Concatenating Waveform Samples.

2.9ms at a 44.10KHz sampling rate. The Mel-Spectrogram is converted to a logarithmic scale to obtain the Log-Mel Spectrogram. Log Mel-Spectrograms were transformed to RGB color space to generalize. The converted photos were scaled to 224x224 pixels shape using area interpolation and converted to PNG format. Figure 3 shows a few example Log-Mel Spectrograms for audio data.

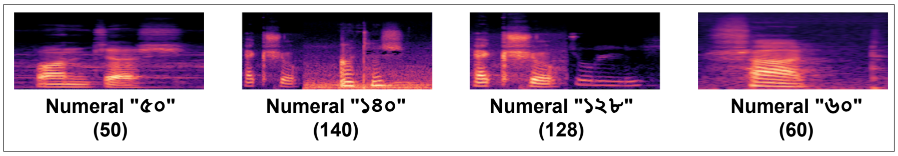

Figure 3. Examples of Log-Mel Spectrograms for Distinct Bangla Numerals.

The training dataset is shuffled and randomly divided for training, validation and testing in 80%-20%-10% ratio respectively. Dual normalization approach as implemented on log-mel spectrograms. The pixel value range of images normalized to [0,1] using 'min-max'. The pixel value for each color channel is then normalized again by 'Z-score' normalization which is substituting the mean from pixel values and dividing it by the standard deviation. The mean values of [0.485, 0.456, 0.406] and standard deviation values of [0.229, 0.224, 0.225] were employed into 3 channels respectively according to the standard procedure for training ImageNet datasets.

4.4 Method

The sophisticated architectures of pre-trained Convolutional Neural Networks (CNN) were used to optimally and correctly categorize Bangla numerals. In our technique, we used DenseNet-201, ResNet-152, and VGG-19.

VGG-19 is known for its simple and uniform architecture, using only 3x3 convolutional layers stacked on top of each other. This simplicity, combined with its depth, makes it highly effective for feature extraction. This model has been pre-trained on large datasets like ImageNet, which allows for effective transfer learning. This can significantly improve performance on tasks with limited training data, such as Bangla numeral recognition. This model has a proven track record in various computer vision tasks, and its architecture is well-understood and widely used, providing a reliable benchmark for comparing other models.

DenseNet-201 utilizes dense connections where each layer receives input from all previous layers, leading to improved information flow and gradient propagation. This helps in capturing intricate patterns in speech data. DenseNet models are parameter-efficient, meaning they require fewer parameters than traditional convolutional networks to achieve similar or better performance. This efficiency is beneficial when dealing with complex tasks like speech recognition.

ResNet-152 leverages residual learning, which allows the network to learn residual functions with reference to the layer inputs. This makes training deep networks easier and more effective. With 152 layers, ResNet-152 can capture very complex patterns and features in the data. This depth is particularly useful for nuanced tasks such as distinguishing different Bangla numerals from speech signals. Moreover, It helps mitigate the vanishing gradient problem, which is crucial for training deep networks on large and complex datasets. ResNet architectures have consistently achieved state-of-the-art results in various image and speech recognition tasks, making them a robust choice for our study.

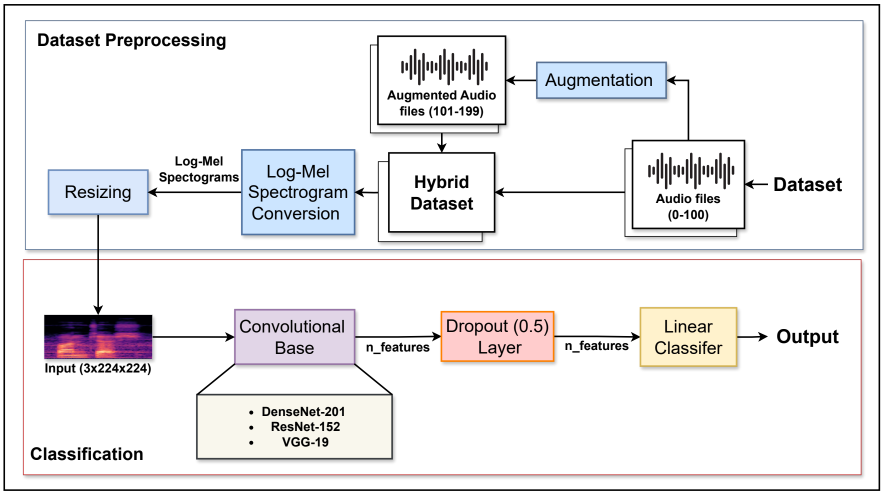

Figure 4. Pipeline Diagram of the Methodology for Bangla Numeral Recognition Using CNN Architectures and Transfer Learning.

Following careful observation and multiple trials and errors, each model's classification layer was replaced by a block containing a dropout layer with a dropout rate of 0.5 and a fully connected linear layer. The methodology is illustrated in Figure 4.

The initial learning rate was set to 5×10^{-4} alongside StepLR as a scheduler with a step size of 7 and a gamma value of 0.7 for decaying learning rate. The models were operated with a batch size of 32 for 10 epochs using SGD as the optimizer, Sigmoid as the activation function and Cross-entropy as the loss function.

5 Result analysis

5.1 Experimental Design

Our experiments were divided into three sections for comparison and evaluation purposes. The first section indicates testing on 10 Bangla numerals in the range '০-৯' (0-9) among which every numeral only includes one syllable. The second section is the testing on 100 Bangla numerals in the range '০-৯৯' (0-99) which also includes di-syllables and tri-syllables. For the first and second sections, only authentic data was used. The last section is the classification of the entire hybrid dataset of 200 numerals in the range '(০-১৯৯) ' (0-199). Each model in every instance was trained using both random weights as well as ImageNet weights to study the effects of Transfer Learning in this scenario.

5.2 Evaluation Metrics

The developed model was evaluated using a variety of criteria, including precision, recall, F1-score and accuracy. The accuracy was determined as the ratio of the macro-average to the total number of consultations (n), as specified in Equation (2), where y is the actual value of consultations (0, and 1 is the projected value.

$$Accuracy(g, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(g_i = y_i) \tag{2}$$

The the mean recall for the classes is calculated using macro-recall, which indicates how well the model can detect categories. Equation (3) defines macro-recall as follows: set of all classes is defined by L, the fraction of anticipated consultations is defined by y_i with i labels, and the true labels consultations is represented by g_i.

$$Recall = \frac{1}{|L|} \sum_{i \in L} R(y_i, g_i), \quad R(y_i, g_i) = \frac{|y_i \cap g_i|}{|g_i|} \tag{3}$$

Similarly, macro-precision computes the average precision over every class. In this situation, precision is the ratio of properly recognized positive consultations to the total number of positive consultations (Equation (4)).

$$Precision = \frac{1}{|L|} \sum_{i \in L} P(y_i, g_i), \quad P(y_i, g_i) = \frac{|y_i \cap g_i|}{|y_i|} \tag{4}$$

The macro F1-Score (F1-score) calculates the score for each class and delivers the unweighted average. The F1-score, which is known as the harmonic mean of recall and accuracy, represents the balance between the two metrics. The F1-score mathematical formula is described by Equations (5)-(6), where P, R and β are precision, recall, and weighting parameter.

$$F1{-}score = \frac{1}{|L|} \sum_{i \in L} F(y_i, g_i) \tag{5}$$

$$F(y_i, g_i) = \frac{(1 + \beta^2) \cdot P(y_i, g_i) \cdot R(y_i, g_i)}{\beta^2 \cdot P(y_i, g_i) + R(y_i, g_i)} \tag{6}$$

5.3 Results

Our study on Bangla numeral recognition employing transfer learning and convolutional neural networks (CNNs) achieved notable results, which are detailed in this section. Among the tested models, the ResNet architecture was particularly outstanding, achieving an accuracy of **95.58%** in recognizing the '০-১৯৯' (0-199) range of Bangla spoken numbers as described in Table 1. This was the highest performance observed, surpassing other models like DenseNet and VGG, which also performed well but did not reach the same levels of accuracy.

The graph included as Figure 5 in the paper visually represents the convergence of DenseNet, VGG, and ResNet across the numeral range '০-১৯৯' (0-199). It clearly illustrates that while all models benefit from the use of transfer learning, ResNet models consistently outperform others across all tested scenarios. The faster convergence on ResNet dictates the model's efficacy and suitability particular to this scenario. Not only that, The smoothness on the curves portray the stability of the models. This superiority of ResNet can be attributed to the depth and complexity of the architecture, which seems particularly suited to capturing the nuances of speech data in Bangla.

The superior performance of ResNet could also be linked to its ability to mitigate the vanishing gradient problem, a common issue in deep networks that can affect learning in complex tasks like speech recognition. Furthermore, the architectural nuances of ResNet, featuring residual learning blocks, which is more adept at handling the spectro-temporal features extracted from the speech signals, as represented by Log-Mel spectrograms.

Moreover, the utilization of ImageNet pre-trained weights has shown a approxmately 25% performance enhancement across all models. Models initialized with these weights outperformed those that started with random weights, emphasizing the effectiveness of transfer learning in this domain. This is particularly crucial for low-resource languages like Bangla, where large, annotated datasets are scarce, and the ability to leverage pre-trained networks can significantly reduce the need for extensive data collection.

Table 1. Performance Metrics of CNN Models Across Different Datasets and Weight Initialization.

Model	Dataset	Weights	Precision	Recall	F1-score	Accuracy
DenseNet-201	0-9	Random	0.7940	0.7778	0.7736	0.7921
		ImageNet	0.9661	0.9649	0.9632	0.9652
	0-99	Random	0.6927	0.6582	0.6243	0.7470
		ImageNet	0.9570	0.9580	0.9563	0.9570
	0-199	Random	0.5392	0.5703	0.6021	0.6833
		ImageNet	0.9417	0.9375	0.9365	0.9408
ResNet-152	0-9	Random	0.8225	0.8118	0.8113	0.8212
		ImageNet	0.9784	0.9765	0.9728	0.9783
	0-99	Random	0.7895	0.7758	0.7776	0.7801
		ImageNet	0.9721	0.9602	0.9623	0.9606
	0-199	Random	0.7214	0.7127	0.7125	0.7145
		ImageNet	**0.9598**	**0.9548**	**0.9550**	**0.9558**
VGG-19	0-9	Random	0.7159	0.7082	0.7084	0.7041
		ImageNet	0.9317	0.9271	0.9263	0.9325
	0-99	Random	0.6524	0.6651	0.6651	0.6692
		ImageNet	0.9123	0.9279	0.9201	0.9233
	0-199	Random	0.6218	0.6286	0.6281	0.6289
		ImageNet	0.9055	0.8955	0.8949	0.8951

In terms of other evaluation metrics, the precision, recall, and F1-score were analyzed to provide a comprehensive picture of model performance. These metrics further substantiated the robustness of the training and validation phases. Notably, the recall rates were particularly high, suggesting that the models were effective at identifying relevant features from the speech inputs, even in noisy or complex auditory environments.

The F1-score, representing the harmonic mean of precision and recall, balanced these factors and underscored the efficacy of the models in handling the diversity of the Bangla numeral dataset. This balance is critical in practical applications where both the inclusion of true positives and the exclusion of false positives are important.

Table 2. Comparison of Accuracy Percentages for Different Datasets

Dataset	Current State-of-the-art	Proposed Method (ResNet)
0-9	97.10% [21]	97.83%
0-99	89.74% [20]	96.06%
0-199	–	95.58%

Our model employs the ResNet architecture, renowned for its deep residual learning capabilities, which enhances feature extraction and gradient propagation significantly. As described in Table 2, our approach ensures robust performance across diverse audio environments. This leads to higher accuracy and

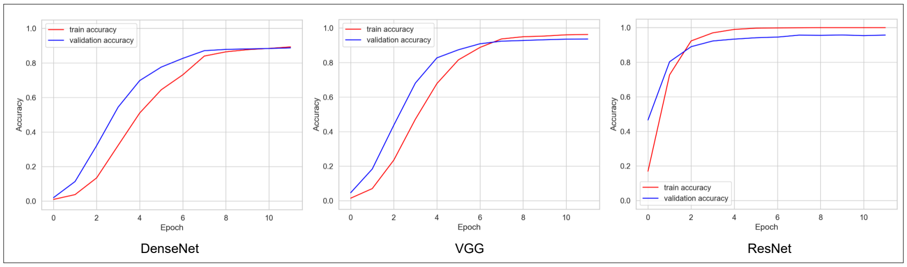

Figure 5. Comparative Convergence Performance of CNN Models Across Numeral Range '০-১৯৯' (0-199) on Pre-trained ImageNet Weights.

faster convergence than existing models, especially in complex numeral ranges, making our solution not only technologically advanced but also highly practical for real-world applications.

The results of this study not only reinforce the viability of using advanced neural network architectures for speech recognition in underrepresented languages but also demonstrate the practical benefits of transfer learning. By reducing the reliance on large-scale data collection and enabling the use of pre-existing models, significant strides can be made in making speech recognition technology more accessible and effective for a wider range of languages and applications.

6 Conclusion

This study sets a new standard for the recognition of Bangla spoken numerals, achieving a best-reported accuracy of **95.58%**. The profound impact of integrating advanced deep learning architectures, especially ResNet, with transfer learning techniques has been demonstrated, establishing a new benchmark in this field. The complexities of speech data have been effectively handled by our methodology, significantly reducing the reliance on extensive training data, a common challenge in speech recognition tasks for low-resource languages such as Bangla. The use of pre-trained models has been shown to be crucial, enhancing the system's ability to accurately recognize spoken numerals across a diverse dataset. Furthermore, the implications of these results for the development of technology that can be integrated into various interactive and automated systems have been highlighted, thereby enhancing technological accessibility for Bangla speakers. The recognition accuracy and robustness demonstrated in this study have paved the way for real-world applications, making voice-driven interfaces more reliable and user-friendly. This work lays a solid foundation for future innovations in speech recognition, driving forward the technological capabilities for Bangla and similar languages.

References

1. Ahammad, K., Rahman, M.M.: Connected bangla speech recognition using artificial neural network. International Journal of Computer Applications **149**(9), 38–41 (2016)
2. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video (2016)
3. Bhattacharjee, A., Hasan, T., Ahmad, W.U., Shahriyar, R.: BanglaNLG and BanglaT5: Benchmarks and resources for evaluating low-resource natural language generation in Bangla. In: Vlachos, A., Augenstein, I. (eds.) Findings of the Association for Computational Linguistics: EACL 2023. pp. 726–735. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023). https://doi.org/10.18653/v1/2023.findings-eacl.54, https://aclanthology.org/2023.findings-eacl.54
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset (2018)
5. Choi, K., Fazekas, G., Sandler, M., Cho, K.: Convolutional recurrent neural networks for music classification (2016)
6. Choi, K., Fazekas, G., Sandler, M., Cho, K.: Transfer learning for music classification and regression tasks (03 2017)
7. Demir, F., Abdullah, D.A., Sengur, A.: A new deep cnn model for environmental sound classification. IEEE Access **8**, 66529–66537 (2020). https://doi.org/10.1109/ACCESS.2020.2984903
8. Dong, M.: Convolutional neural network achieves human-level accuracy in music genre classification (2018)
9. Guzhov, A., Raue, F., Hees, J., Dengel, A.: Esresnet: Environmental sound classification based on visual domain models (2020)
10. Gwardys, G., Grzywczak, D.: Deep image features in music information retrieval. International Journal of Electronics and Telecommunications **60**, 321–326 (2014), https://api.semanticscholar.org/CorpusID:13752140
11. Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J., Wilson, K.: Cnn architectures for large-scale audio classification (2017)
12. 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)
13. Lee, J., Park, J., Kim, K.L., Nam, J.: Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms (2017)
14. 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)
15. Nguyen, T., Pernkopf, F.: Lung sound classification using snapshot ensemble of convolutional neural networks. vol. 2020 (04 2020). https://doi.org/10.1109/EMBC44109.2020.9176076
16. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: Wavenet: A generative model for raw audio (2016)
17. Paul, B., Bera, S., Paul, R., Phadikar, S.: Bengali spoken numerals recognition by mfcc and gmm technique. In: Advances in Electronics, Communication and Computing: Select Proceedings of ETAEERE 2020. pp. 85–96. Springer (2021)

18. Paul, D.A., Das, D., Kamal, M.: Bangla speech recognition system using lpc and ann. pp. 171–174 (02 2009). https://doi.org/10.1109/ICAPR.2009.80
19. Roberts, A., Engel, J., Raffel, C., Hawthorne, C., Eck, D.: A hierarchical latent vector model for learning long-term structure in music (2019)
20. Sen, O., Roy, P., Al-Mahmud: A Novel Bangla Spoken Numerals Recognition System Using Convolutional Neural Network, pp. 344–357 (06 2023). https://doi.org/10.1007/978-3-031-34619-4_28
21. 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)
22. Sharmin, R., Rahut, S.K., Huq, M.R.: Bengali spoken digit classification: A deep learning approach using convolutional neural network. Procedia Computer Science **171**, 1381–1388 (2020)
23. 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)
24. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild (2012)
25. Tokozume, Y., Harada, T.: Learning environmental sounds with end-to-end convolutional neural network. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2721–2725 (2017). https://doi.org/10.1109/ICASSP.2017.7952651
26. Wikipedia contributors: Bengali language (2021), https://en.wikipedia.org/wiki/Bengali$_{l}anguage, [Online; accessed 13 - April -$ 2024]
27. Xie, H., Virtanen, T.: Zero-shot audio classification based on class label embeddings. In: 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). pp. 264–267 (2019). https://doi.org/10.1109/WASPAA.2019.8937283
28. Zhang, W., Lei, W., Xu, X., Xing, X.: Improved Music Genre Classification with Convolutional Neural Networks. In: Proc. Interspeech 2016. pp. 3304–3308 (2016). https://doi.org/10.21437/Interspeech.2016-1236