Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 13, No. 3, 2025, 337-345.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Robust Persian Digit Recognition in Noisy Environments Using Hybrid CNN-BiGRU Model

Ali Nasr-Esfahani, Mehdi Bekrani* and Roozbeh Rajabi

Department of Electrical and Computer Engineering, Qom University of Technology, Iran.

Article Info

Article History:
Received 25 March 2025
Revised 25 March 2025
Accepted 29 May 2025

DOI:10.22044/jadm.2025.15932.2707

Keywords:
Spoken Digit Recognition, Data Augmentation, Convolutional Neural Network, Bidirectional Gated Recurrent Unit.

**Corresponding author: bekrani@qut.ac.ir (M. Bekrani).*

Abstract

Artificial intelligence (AI) has significantly advanced speech recognition applications. However, many existing neural network-based methods struggle with noise, reducing accuracy in real-world environments. This study addresses isolated spoken Persian digit recognition (zero to nine) under noisy conditions, particularly for phonetically similar numbers. A hybrid model combining residual convolutional neural networks and bidirectional gated recurrent units (BiGRU) is proposed, utilizing word units instead of phoneme units for speaker-independent recognition. The FARSDIGIT1 dataset, augmented with various approaches, is processed using Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction. Experimental results demonstrate the model's effectiveness, achieving 98.53%, 96.10%, and 95.92% accuracy on training, validation, and test sets, respectively. In noisy conditions, the proposed approach improves recognition by 26.88% over phoneme unit-based LSTM models and surpasses the Mel-scale Two Dimension Root Cepstrum Coefficients (MTDRCC) feature extraction technique along with MLP model (MTDRCC+MLP) by 7.61%.

1. Introduction

Artificial intelligence (AI) has become widely used in signal processing, including audio, speech recognition, image processing, and machine vision [1,2,3,4]. Among these, automatic speech recognition (ASR) has garnered extensive interest due to its vast applicability in hands-free user interfaces, voice-based authentication, and assistive technologies [5,6]. Spoken digit recognition, as a specific sub-task of ASR, demands high precision due to its frequent use in structured inputs such as phone numbers, PIN codes, and spoken commands. Even minor errors in digit recognition can lead to serious misunderstandings in critical applications.

Within ASR, digit recognition serves as a fundamental task with applications in telecommunication systems, voice-controlled banking, and automated customer service. Accurate recognition is especially critical when used in sensitive scenarios such as password entry or financial transactions.

Research on spoken number recognition has been ongoing. Homayounpour et al. [7] reviewed methods using hidden Markov models (HMMs) and MLP neural networks for Persian number recognition, achieving 99.1% accuracy for discrete and 83.7% for continuous numbers in the FARSDIGIT1 database [8]. In 2008, Hierarchical Temporal Memory (HTM) was applied to isolated digit recognition [9], and a Gaussian mixture model (GMM) classifier with and Delta-Delta Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction achieved 99.3% accuracy in Arabic digit recognition [10].

In recent years, deep neural networks (DNNs) have gained significant interest in speech recognition. Danashri et al. [11] used a DNN with a Deep Belief Network (DBN), achieving 86.06% accuracy for

English numbers from the TIDIGIT database [12]. Zada et al. [13] utilized a Convolutional Neural Network (CNN) for Pashto number recognition, achieving an accuracy of 84.17%. The model featured four deep convolutional layers and a maximum pooling layer. However, the effect of noise on recognition accuracy remains largely unexplored.

In [14], an LSTM-based network was presented that categorizes discrete numbers into groups with similar phonetic characteristics and trains an LSTM neural network for each class, independently. It achieved 91.7% accuracy in noise-free conditions but dropped to an average of 69.22% when various noises are added to the audio data.

More recently, the combination of multi-layer perceptron (MLP) and Mel-scale Two Dimension Root Cepstrum Coefficients (MTDRCC) feature extraction method was used for number recognition in noisy conditions, achieving 98.85% accuracy in noise-free conditions and 88.49% in noisy ones [15], for Persian numbers. Viriri et al. [16] combined recurrent neural network (RNN) and LSTM, achieving 99% accuracy on English numbers, while Sotisna et al. [17] used transfer learning networks, like AlexNet and GoogleNet, reporting lower recognition rates, 72% for AlexNet and 66% for GoogleNet. In 2023, a hybrid deep CNN model for Bengali digit recognition, utilizing a unique hybrid feature extraction, including MFCC, Spectral Sub-band Energy (SSE), and Log Spectral Sub-Band Energy (LSSE), achieved 98.52% accuracy [18].

In another recent study, Ramadan and Ezzat [19] proposed two approaches for spoken digit recognition using the Free-Spoken Digit Dataset (FSDD) [20]. The first approach combines Wavelet Time Scattering with a Support Vector Machine (SVM) classifier, achieving 96.67% accuracy. The second approach employs Mel-frequency spectrograms as input to a Deep Convolutional Neural Network (DCNN), leading to a small increase in recognition accuracy of 97.67%. Their results highlight that even traditional ML techniques, when paired with robust feature extraction methods, can be competitive. However, their work focuses on English digits and does not explicitly address noisy conditions or phonetically similar classes.

To address limited data for many non-English languages, data augmentation has proven effective. Lunas et al. [21] applied techniques like adding white noise and altering sound length, using a Markov model for recognition. Tom Ko et al. [22] expanded a 300-hour dataset to 900 hours with noise and room simulation techniques, employing MFCC and BiLSTM for recognition.

Yet, these methods often rely on phoneme-level segmentation and struggle with phonetic similarities among digits.

In recent years, combining different types of deep neural networks has become a promising approach to enhance model performance, especially in tasks where both spatial and temporal features are critical. Fusion strategies enable networks to leverage the strengths of individual architectures, leading to more robust and accurate systems. Hybrid CNN–RNN models have been widely applied in speech recognition and image-based sequence modeling. These models typically employ CNNs for spatial feature extraction and RNNs (e.g., LSTM, GRU) for temporal pattern modeling. For example, Mahdavi et al. [23] proposed a hybrid CNN–LSTM architecture with optimized fusion coefficients for RSS-based localization in wireless sensor networks. Their results demonstrated significant improvements in localization accuracy by intelligently combining the outputs of two deep models.

In Persian-language contexts, digit recognition presents additional challenges due to phonetic similarities among digits (e.g., /sefr/ vs. /se/, /do/ vs. /noh/) and the limited availability of large annotated speech datasets. These issues are further compounded in noisy environments, where acoustic interference can drastically degrade recognition performance.

To address these challenges, recent efforts have explored various architectures, ranging from traditional statistical models to deep learning frameworks. However, many of these approaches either lack robustness to environmental noise or fail to effectively capture temporal dependencies and linguistic nuances in Persian digits.

Motivated by these limitations, we propose a novel DNN architecture that combines Residual CNNs with Bidirectional Gated Recurrent Units (BiGRUs). The proposed method adopts word-level units instead of phoneme-level representations, enhancing robustness to noise and improving the recognition of phonetically similar digits.

This paper presents a novel Persian digit recognition approach, addressing linguistic and recognition challenges. Key aspects include:

- Data Augmentation: Improves robustness with limited data.
- MFCC Features: Extracts relevant speech characteristics.

- Word Units: Enhances accuracy over phoneme-based methods.
- Hybrid Network: Combines Residual CNN and BiGRU for superior performance.

In addition, our method investigates the effects of various noises, including monotonous horns, nature sounds, vehicle movement sounds, humming sounds, and factory sounds, on the speech recognition performance.
Although it is technically possible to generalize our architecture to broader speech recognition tasks—such as by adopting end-to-end frameworks like DeepSpeech—doing so would require substantial changes, particularly in the output layers and the training strategy. Our intention, however, is not to develop a general-purpose ASR system, but rather to build a robust and efficient model specifically for digit recognition, which is a representative command-level task commonly used in real-world applications (e.g., phone-based services, voice-controlled authentication, and user interfaces)
The paper is organized as follows: Section 2 covers the data generation method. Section 3 discusses the DNN architecture. Section 4 presents experimental results, and Section 5 concludes with key takeaways from the study.

2. Datasets

This study utilizes recordings from 51 speakers in the FARSDIGIT1 database [8], containing discrete and continuous Persian numbers (zero to nine). The recordings, captured over telephone lines (both intra- and inter-city), have an SNR of approximately 8.8 dB and a sampling rate of 11025 Hz. The dataset includes 31 male speakers (ages 12–61) and 20 female speakers (ages 14–52). Each speaker recorded numbers in one to two sessions, 7 to 30 days apart, with 10 repetitions per number, totaling 510 audio data samples per digit.
To prevent neural network overfitting, we apply the data augmentation on the dataset using:

- Sound Speed Variation [21]: Adjusting playback speed.
- Reverb Filter [22]: Simulating different acoustic environments.
- Background Noise [24]: Adding various ambient sounds.
- Hall Environment Simulation [25]: Introducing additional reverberation effects.

The data augmentation methods were applied with different probabilities: 70% for noise, 15% for speed changes, and 7.5% each for reverb and hall simulation. Noise sources included horns, nature sounds, vehicle engines, buzzing, and industrial sounds, with SNR levels of 0, 5, 10, 15, and 20 dB. These noise samples were obtained from a publicly available dataset on Kaggle [26].
These probabilities were chosen to ensure diverse audio conditions: 70% noise to introduce varied SNR levels, 15% speed changes to explore audio variations, and 7.5% each for reverb and hall simulation due to their similar shape features. This distribution supports a balanced augmentation strategy for comprehensive dataset enrichment.
Applying augmentation increased the dataset size 5x, resulting in 25,500 data samples (2,550 per digit). Figure 1 illustrates the distribution of augmented data.

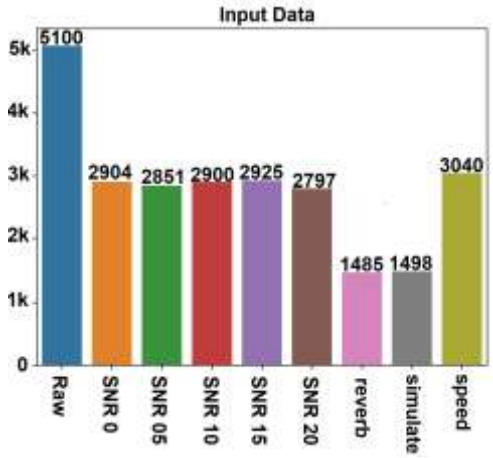

Figure 1. Input data count after data augmentation.

This expanded dataset enhances DNN training, ensuring robustness in diverse real-world noise conditions.

3. The Proposed Spoken Digit Recognition Method

The proposed method consists of three main stages:

- Data augmentation technique described in Section 2
- Speech feature extraction using the MFCC technique.
- The proposed neural network architecture.

3.1. Speech Feature Extraction Using the MFCC Technique

Feature extraction is crucial in speech recognition, as raw audio contains noise and irrelevant parameters that affect accuracy. Common feature extraction techniques include [27]:

- MFCCs: Capture both spectral and temporal speech characteristics.
- Perceptual Linear Prediction: Similar to MFCCs but models the human auditory perception with non-linearity.

- Mel-Frequency Discrete Wavelet Coefficients: Applies wavelet transform to the speech signal in the mel-frequency domain.
- Spectrogram: 2D representation of frequency changes over time.

MFCC stands as the primary choice for speech feature extraction due to its efficiency, noise resilience, and ability to emulate human auditory perception. MFCCs excel in capturing both spectral and temporal characteristics of speech, while reducing the dimensionality of the data, thus offering an efficient representation of speech signals [27].

In the proposed method, to account for and mitigate the impact of facet similarities on the recognition rate, the input data is first separated based on words and numbers and the MFCC technique is subsequently used to extract features from segmented speech signals instead of using raw audio signals.

Figure 2 illustrates the MFCC process for feature extraction. The input audio undergoes *pre-emphasis* to enhance high frequencies, followed by *framing* and *windowing* to reduce spectral leakage. *The Short-Time Fourier Transform (STFT)* computes the power spectrum, which is processed through a *triangular filter bank* to mimic human auditory perception. A *Discrete Cosine Transform (DCT)* is then applied to the log-filtered energies to obtain MFCCs, retaining only relevant coefficients. Finally, *mean normalization* ensures scale consistency across samples [23,24].

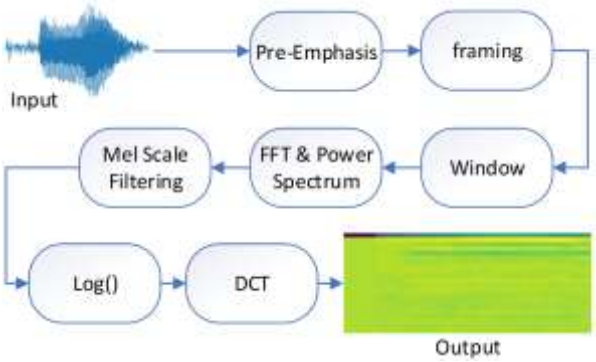

Figure 2. MFCC block diagram [27].

For this study, MFCC extraction was configured with a pre-emphasis coefficient (α) of 0.97 [27], a sampling rate of 16,000 Hz, a frame size of 25 milliseconds, and an FFT size of 512 points (NFFT), in accordance with the acoustic characteristics of the speech data. A total of 40 MFCC coefficients were extracted per frame to capture a detailed spectral representation.

Given the approximate 1-second duration of each audio sample and the corresponding frame shift, every sample was converted into an MFCC feature map of size 40 × 80, where 40 denotes the number of cepstral coefficients and 80 represents the number of time frames. To ensure consistent input dimensions for the network, all MFCC feature maps were either zero-padded or truncated to a fixed size of 40 × 80. This fixed-size representation was treated as a single input patch for the model.

3.2. The Proposed DNN Architecture

The proposed DNN for Persian digit recognition is inspired by DeepSpeech2 [28], with modifications to enhance performance. It consists of a CNN layer, residual CNN blocks, a Fully Connected (FC) layer, and Bidirectional GRU (BiGRU) blocks, as illustrated in Figure 3. Each component plays a key role in feature extraction and classification.

The CNN layer [3] extracts feature and adjusts input dimensions. Next, three residual CNN blocks [29] refine feature representation. Unlike DeepSpeech2, word units are used instead of phonemes, and the Cross-Entropy Loss function replaces Connectionist Temporal Classification (CTC). The FC layer [30] further adjusts feature dimensions before classification.

As opposed to the LSTM architecture used in DeepSpeech2, the proposed method employs BiGRU blocks [31]. This modification is motivated by GRU's advantage of having fewer parameters than that of the LSTM, resulting in a reduction in the network weights and the computational resource consumption while often achieving superior performance. In addition, the BiGRU's configuration enables the network to process each frame by considering data from both previous and next frames, leading to more accurate predictions. Five BiGRU blocks classify features, followed by two FC layers with a softmax function to determine results.

These modifications optimize DeepSpeech2 for robust Persian digit recognition in noisy environments. The following sections detail the CNN layer, residual CNN blocks, FC layer, and BiGRU blocks.

3.2.1. CNN Layer

The CNN layer transforms input audio data into feature vectors while adjusting input dimensions. A single CNN layer is used with one input channel and 32 output channels to extract diverse features from the raw audio input. A kernel size of 3 is utilized to reinforce feature extraction, while a stride of 2 enables down-sampling, reducing computational complexity while preserving essential information.

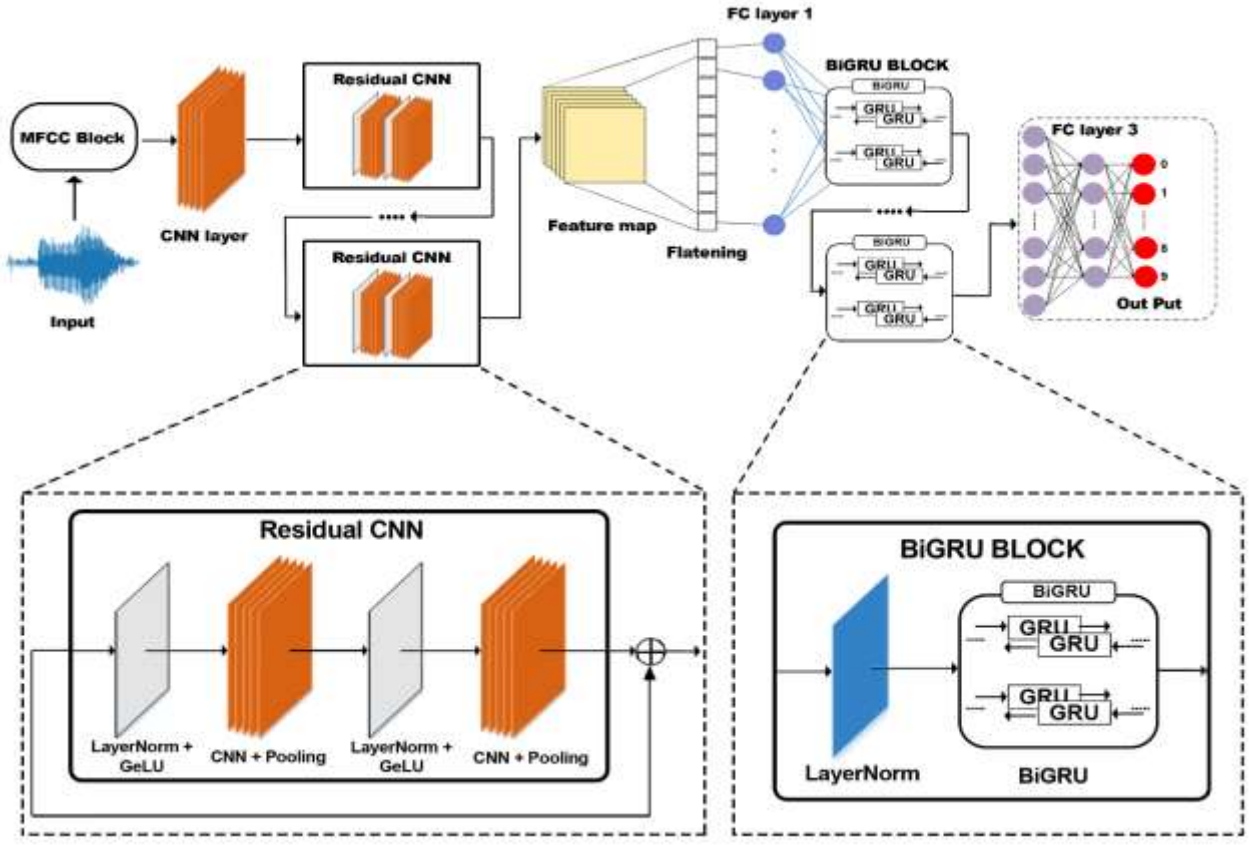

Figure 3. Block diagram of the proposed DNN.

3.2.2. Residual CNN Blocks

Three residual CNN blocks [29] are used to learn audio features, enhancing the model's depth and ability to recognize complex patterns in spoken Persian digits, even in noisy environments.

Each block includes two normalization layers (normalized shape: 20) to stabilize and accelerate the training process, two Gaussian Error Linear Unit (GELU) activation functions for contribute to the non-linearity of the model, and two CNN layers (32 input/output channels, kernel size: 3, stride: 1, padding: 1). Crucially, the output from these layers is summed with the input to form the final result of the Residual CNN block. This design choice enables the network to retain essential information from the input while incorporating the features extracted through the normalization and convolutional layers.

3.2.3. FC Layer

An FC layer [30] connects the residual CNN blocks to the BiGRU, optimizing extracted features before passing them forward. Configured with an input size of 25,600 and an output size of 512, it compresses information for efficient processing.

Additionally, two FC layers handle classification, each with dropout regularization, GELU, and softmax activation functions. The first FC layer reduces 1,024 inputs to 512 outputs, simplifying classification, while the second maps 512 inputs to 10 outputs. The final softmax activation converts outputs into probability scores across the 10 Persian digit classes, ensuring accurate recognition. This configuration allows for effective classification and prediction, contributing to the overall accuracy of our neural network in recognizing spoken Persian digits.

3.2.4. BiGRU Block

The BiGRU block [31], applied five times, captures temporal dependencies essential for recognizing spoken Persian digits. Each block includes layer normalization with a parameter of 512 to stabilize training and prevent vanishing gradient problem often associated with recurrent neural networks. A GELU activation function adds non-linearity to the model, enhancing its capacity to capture complex patterns.

BiGRU is a variant of the Gated Recurrent Unit (GRU) that processes input sequences in both forward and backward directions. Unlike unidirectional GRUs, BiGRU captures both past and future context within a sequence, making it highly suitable for speech recognition tasks where information from upcoming frames improves prediction accuracy. Its internal gating

mechanism—specifically the update and reset gates—allows the network to selectively preserve or discard information, enabling efficient learning of long-range dependencies with fewer parameters than LSTM.

At the core, the BiGRU processes input bidirectionally with an input size of 512 and a hidden layer size of 512. This design enables the network to process input sequences in both forward and backward directions, capturing contextual information from both past and future time steps.

4. Results and Discussion

The dataset is divided into training, validation, and testing sets, with frequency distributions shown in Figures 4, 5 and 6, respectively. Data shuffling is performed before applying MFCC for feature extraction.

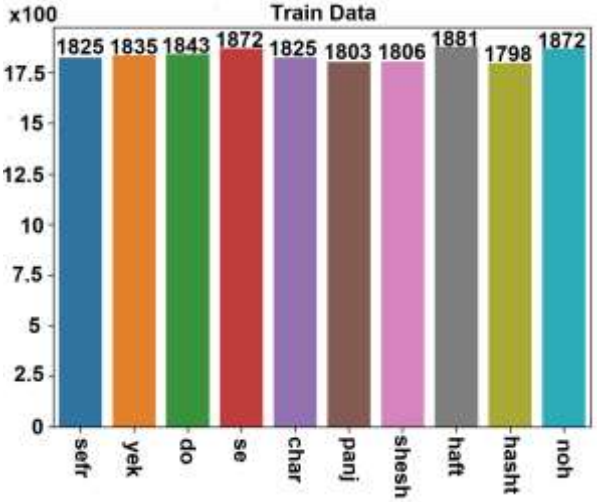

Figure 4. Frequency chart of each class in Train data.

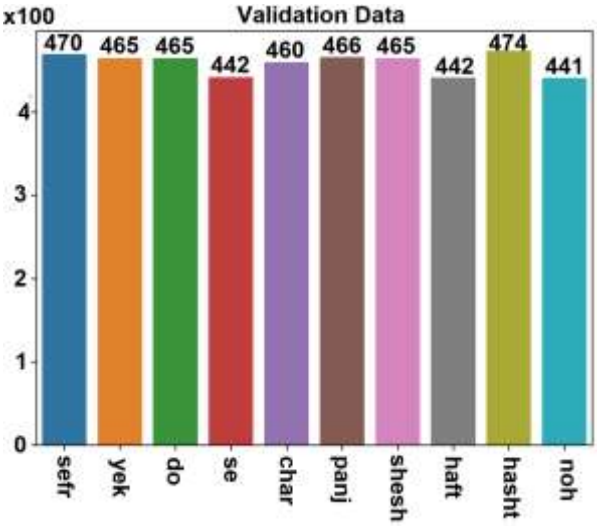

Figure 5. Frequency chart of each class in the Validation data.

Training is conducted on Google Colab, utilizing a Tesla T4 GPU (15.1 GB RAM, 78.19 GB storage, 12.68 GB processing RAM). The model is implemented in Python programming language and the library used to create and train the model is PyTorch.

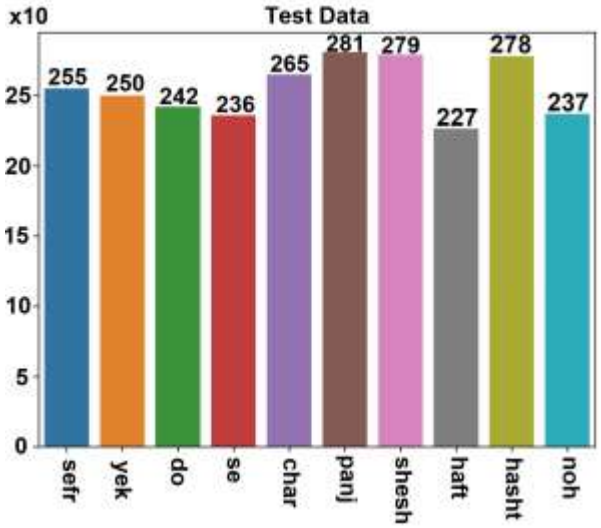

Figure 6. Frequency chart of each class in the Test data.

The training process was conducted using the Adam optimization algorithm with a learning rate of 0.005 over 25 epochs and a batch size of 1.

Initially, simpler networks like CNN and GRU are evaluated but achieve limited validation accuracies of 83.22% and 78.76%, respectively, due to their smaller sizes and struggling to effectively learn from the noisy data. The LSTM model improves upon these, achieving 87.86%.

Figure 7 illustrates the proposed model's accuracy across training, validation, and testing over 25 epochs. The final accuracy reaches 98.53% (training), 96.10% (validation), and 95.92% (testing).

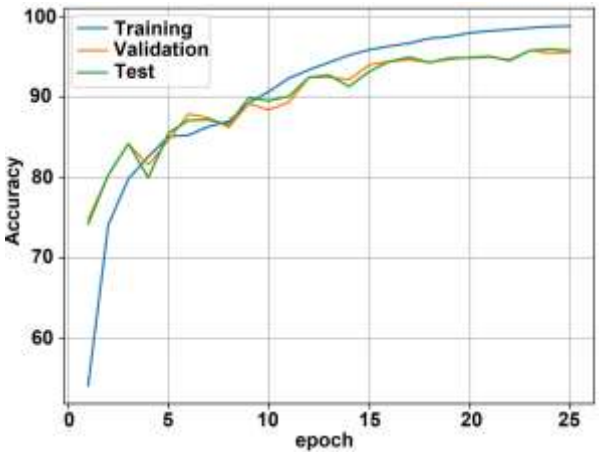

Figure 7. Recognition accuracy of the proposed CNN-BiGRU model on training, validation, and test sets over 25 epochs.

In addition to evaluating the model on the full test set (comprising both clean and noisy speech samples), an attempt was made to assess its performance using only clean data. Initially, training exclusively on clean samples (5,100 recordings) resulted in overfitting, with the model achieving 99.16% training accuracy but only 95.69% test accuracy after 25 epochs. This performance gap highlights the model's sensitivity to data scarcity, especially given its architectural complexity. To address this, data augmentation was applied to the clean dataset to increase its size and diversity. Following this enhancement, the model achieved 99.96% training accuracy and 99.84% test accuracy on clean speech samples (25500 recordings) after 25 epochs demonstrating improved generalization. These findings emphasize the importance of sufficient and varied training data—even in clean conditions—for achieving optimal performance.

Table 1 compares these results, confirming the superior performance of the proposed model.

Table 1. Classification accuracy (%) of various models on the Persian digit recognition task across training, validation, and test sets.

Network	Training	Validation	Test
LSTM	91.16%	87.45%	86.82%
GRU	76.05%	83.22%	83.49%
CNN	82.90%	80.34%	80.59%
Proposed DNN	98.53%	96.10%	95.92%

The robustness of the proposed model was further assessed by conducting three independent training runs with different random initializations. The results exhibited high consistency across both training and testing phases, yielding an average training accuracy of 98.79% ± 0.22 and a test accuracy of 95.57% ± 0.42. The low standard deviations confirm the stability and reliability of the architecture in recognizing spoken digits under noisy conditions.

Figure 8 presents the confusion matrix of the model on the test set, illustrating the classification accuracy for each digit. Diagonal elements indicate correct predictions, while off-diagonal entries reflect misclassifications.

As can be seen from Fig. 8, the model performs well across all digit classes; however, certain confusions arise, particularly between phonetically similar Persian digits. For example, the digit "0" (/sefr/) is occasionally misclassified as "3" (/se/), likely due to their shared initial consonant /s/ and short syllabic structures—especially when masked by background noise. Similarly, the digit "2" (/do/) sometimes gets confused with "9" (/noh/), which may be attributed to the similar long vowel /o/ and their open articulation patterns under noisy conditions.

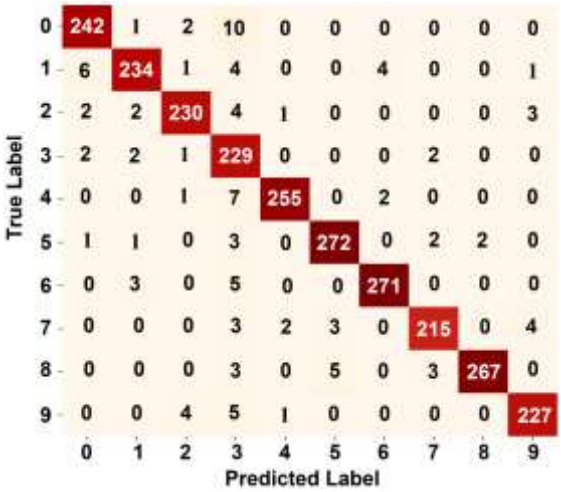

Figure 8. Confusion matrix for the test set.

Furthermore, since the majority of augmented samples include background noise resembling the acoustic pattern of the digit "3" (/se/), increased misclassifications toward "3" are observed across different classes, as reflected in Figure 8.

Table 2 presents a comparative overview of the proposed model alongside several prior approaches in speech recognition. It is important to note that the datasets used across these studies are not entirely uniform, which limits the direct comparability of the reported results. For each baseline, we attempted to reimplement the original architecture based on the descriptions provided in the respective papers. However, since our reimplementation consistently yielded lower accuracy than the original reports, we have referenced the published results for a fairer comparison.

Despite these dataset differences, the results clearly demonstrate the superior performance of our proposed model, particularly under noisy conditions. While methods based on MLP, DBN, and GMM exhibit limited robustness, and CNN- or LSTM-based architectures perform relatively better, they still fall short of the accuracy and generalization achieved by our hybrid model. Specifically, our approach outperforms the MTDRCC model using MLP by 7.61%, and improves upon the phoneme unit-based LSTM model for Persian digits by an average margin of 26.88% under noisy conditions.

Table 2. Accuracy comparison of our method with related works.

Reference	Method	Accuracy (%)	Dataset
2003 [7]	HMM + MLP (clean)	83.7 (cont. num.), 99.1 (disc. num.)	FARSDIGIT1[8]
2016 [11]	DBN	86.06 (valid.)	TI digits [12]
2020 [13]	MFCC + CNN	84.17 (valid.)	Pashto numbers [32]
2021 [14]	LSTM (clean/noisy)	91.7 (clean), 69.2 (noisy)	FarsDat
2022 [15]	MTDRCC + MLP	98.85 (clean), 88.49 (noisy)	Persian numbers
2022 [16]	RNN + LSTM (clean)	99 (valid.)	English digits [20]
2022 [17]	Transfer Learning (AlexNet, GoogleNet)	72 (AlexNet), 66 (GoogleNet)	-
Our Method	Residual CNN + BiGRU (noisy)	98.53 (train), 96.10 (valid.)	FARSDIGIT1[8]

To better understand the computational cost of the proposed model, a structural complexity analysis is presented. The architecture begins with an initial convolutional layer, followed by three residual blocks, each containing two convolutional and two normalization layers (totaling six convolutional and six normalization layers). This is succeeded by five BiGRU blocks, each comprising a bidirectional GRU layer and a normalization layer. Additionally, a dense layer connects the final residual block and the first BiGRU block, with two dense layers used at the output stage. Although the hybrid model is relatively complex to capture both spatial and temporal patterns in noisy Persian speech, it balances computational demand with significant performance gains.

Training was conducted on Google Colab with a Tesla T4 GPU, with each epoch taking about 8 minutes. While the model is more computationally demanding than simpler MLP or shallow CNN alternatives, the substantial improvement in accuracy—especially under noisy conditions—justifies the additional complexity. The trade-off between accuracy and computational cost thus favors our hybrid approach in applications where robustness is critical.

5. Conclusion

This paper presents a deep neural network (DNN) for Persian spoken digit recognition, integrating CNN, residual CNN, BiGRU, and fully connected layers. By adopting word units instead of phoneme units, the model effectively handles phonetic similarities, enhancing feature extraction and recognition, particularly in noisy conditions. Experimental results confirm that the proposed DNN surpasses phoneme-based and LSTM-based methods in accuracy. This word unit-based approach offers a robust solution for spoken digit recognition, contributing to advancements in speech recognition technology with potential real-world applications.

References

[1] A. S. Dhanjal and W. Singh, “A comprehensive survey on automatic speech recognition using neural networks,” *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 23367–23412, Mar. 2024.

[2] H. Veisi and A. H. Mani, “Persian speech recognition using deep learning,” *International Journal of Speech Technology*, vol. 23, no. 4, pp. 893–905, Dec. 2020.

[3] M. S. Zandi and R. Rajabi, “Deep learning based framework for Iranian license plate detection and recognition,” *Multimedia Tools and Applications*, vol. 81, no. 11, pp. 15841–15858, May 2022.

[4] A. Kavand and M. Bekrani, “Speckle noise removal in medical ultrasonic image using spatial filters and DnCNN,” *Multimedia Tools and Applications*, vol. 83, pp. 45903–45920, May 2024.

[5] D. Yu and L. Deng, *Automatic speech recognition: A deep learning approach*, Springer Publishing Company, 2016.

[6] M. H. Rahimi Pour, N. Rastin, and M. M. Kermani, “Persian automatic speech recognition by the use of whisper model,” in *20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP)*, Iran, Feb. 2024.

[7] M. M. Homayounpour, J. Kabudian, H. Bashiri, and Z. Ahmadpour, “Recognition of Farsi number over telephone: A comparison of statistical neural and hybrid approaches,” *Amirkabir*, vol. 14, no. 56, pp. 1045–1065, Jan. 2003.

[8] M. M. Homayounpour, "FarsDigits database," in *Technical Report, Laboratory for Intelligent Sound and Speech Processing*, Amirkabir University of Technology, 2005.

[9] J. V. Doremalen and L. Boves, "Spoken digit recognition using a hierarchical temporal memory," in *Interspeech*, 2008, pp. 2566–2569.

[10] N. Hammami, M. Bedda, N. Farah, and R. O. Lakehal-Ayat, "Spoken Arabic digits recognition based on (GMM) for e-Quran voice browsing: Application for blind category," in *International Conference on Advances in Information Technology for the Holy Quran and Its Sciences*, 2013, pp. 123–127.

[11] D. Dhanashri and S. B. Dhonde, "Isolated word speech recognition system using deep neural networks," in *International Conference on Data Engineering and Communication Technology: ICDECT* 2016, vol. 1, 2017, pp. 9–17.

[12] R. G. Leonard and G. Doddington, "TIDIGITS dataset," *Linguistic Data Consortium*, Philadelphia, 1993.

[13] B. Zada and R. Ullah, "Pashto isolated digits recognition using deep convolutional neural network," *Heliyon*, vol. 6, no. 2, Feb. 2020.

[14] S. Tabibian, "Robust Persian isolated digit recognition based on LSTM and speech spectral features," *Iranian Journal of Electrical and Computer Engineering*, vol. 86, no. 19, pp. 1–17, Oct. 2021.

[15] S. M. Hoseini, "Recognition of Persian digits from zero to nine using acoustic images based on Mel Cepstrum coefficients and neural network," *International Journal of Mechatronics, Electrical and Computer Technology*, vol. 11, no. 42, pp. 5059–5064, 2020.

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

[17] C. Amadeus, I. Syafalni, N. Sutisna, and T. Adiono, "Digit number speech recognition using spectrogram-based convolutional neural network," in *International Symposium on Electronics and Smart Devices (ISESD)*, 2022, pp. 1–6.

[18] B. Paul and S. Phadikar, "A hybrid feature-extracted deep CNN with reduced parameters substitutes an end-to-end CNN for the recognition of spoken Bengali digits," *Multimedia Tools and Applications*, vol. 83, no. 1, pp. 1669–1692, Jan. 2024.

[19] A. A. Ramadan and K. M. Ezzat, "Spoken digit recognition using machine and deep learning-based approaches," in *International Telecommunications Conference (ITC-Egypt)*, Alexandria, Egypt, 2023, pp. 592-596.

[20] Z. Jakobovski, "Free spoken digit dataset." github.com, Aug. 2020, [Online]. Available: https://github.com/Jakobovski/free-spoken-digit-dataset.

[21] K. Lounnas, M. Lichouri, and M. Abbas, "Analysis of the effect of audio data augmentation techniques on phone digit recognition for Algerian Arabic dialect," in *International Conference on Advanced Aspects of Software Engineering (ICAASE)*, 2022, pp. 1–5.

[22] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[23] F. Mahdavi, H. Zayyani, and R. Rajabi, "RSS localization using an optimized fusion of two deep neural networks," *IEEE Sensors Letters*, vol. 5, no. 12, pp. 1–4, Dec. 2021.

[24] W. Hartmann, T. Ng, R. Hsiao, S. Tsakalidis, and R. M. Schwartz, "Two-stage data augmentation for low-resourced speech recognition," in *Proc. Interspeech*, vol. 9, 2016, pp. 2378–2382.

[25] D. S. Park, W. Chan, Y. Zhang, C. C. Chiu, B. Zoph, E. D. Cubuk, Q. V. and Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint*, arXiv:1904.08779, 2019.

[26] M. Sithu, "Audio Noise Dataset," kaggle.com, *Kaggle*, 2019. [Online]. Available: https://www.kaggle.com/datasets/minsithu/audio-noise-dataset

[27] N. Dave, "Feature extraction methods LPC, PLP and MFCC in speech recognition," *International Journal for Advance Research in Engineering and Technology*, vol. 1, no. 6, pp. 1–4, July 2013.

[28] D. Amodei, et al., "Deep speech 2: End-to-end speech recognition in English and Mandarin," in *International Conference on Machine Learning*, 2016, pp. 173–182.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[30] S. H. S. Basha, S. R. Dubey, V. Pulabaigari, and S. Mukherjee, "Impact of fully connected layers on performance of convolutional neural networks for image classification," *Neurocomputing*, vol. 378, pp. 112–119, Feb. 2020.

[31] Q. Tao, F. Liu, Y. Li, and D. Sidorov, "Air pollution forecasting using a deep learning model based on 1D convnets and bidirectional GRU," *IEEE Access*, vol. 7, pp. 76690–76698, June 2019.

[32] A. Zakir, et al. "Database development and automatic speech recognition of isolated Pashto spoken digits using MFCC and K-NN," *International Journal of Speech Technology*, vol. 18, pp. 271-275, June 2015.

بازشناسی مقاوم ارقام فارسی در محیط‌های نویزی با استفاده از مدل ترکیبی CNN-BiGRU

علی نصر اصفهانی ، مهدی بکرانی * و روزبه رجبی

دانشکده مهندسی برق و کامپیوتر، دانشگاه صنعتی قم، قم، ایران.

ارسال ۲۰۲۵/۰۳/۲۵؛ بازنگری ۲۰۲۵/۰۴/۲۵؛ پذیرش ۲۰۲۵/۰۵/۲۹

چکیده:

هوش مصنوعی (AI) پیشرفت چشمگیری در کاربردهای بازشناسی گفتار داشته است. با این حال، بسیاری از روش‌های مبتنی بر شبکه‌های عصبی در مواجهه با نویز با مشکل روبه‌رو هستند و دقت آن‌ها در محیط‌های واقعی کاهش می‌یابد. این مقاله به بازشناسی جداگانه ارقام گفتاری فارسی (از صفر تا نه) در شرایط نویزی می‌پردازد، به‌ویژه برای اعداد دارای شباهت آوایی. یک مدل ترکیبی شامل شبکه‌های عصبی پیچشی باقیمانده (Residual CNN) و واحدهای بازگشتی گیتی دوطرفه (BiGRU) پیشنهاد شده است که به جای استفاده از واحدهای واجی، از واحدهای واژه‌ای برای بازشناسی مستقل از گوینده بهره می‌برد. پایگاه داده FARSDIGIT1 که با روش‌های مختلف تقویت شده است، با استفاده از ضرایب کپسترال فرکانس مل (MFCC) برای استخراج ویژگی پردازش می‌شود. نتایج آزمایش‌ها اثربخشی مدل را نشان می‌دهد، به‌گونه‌ای که به دقت‌های ٪۹۸,۵۳، ٪۹۶,۱۰ و ٪۹۵,۹۲ در مجموعه‌های آموزشی، اعتبارسنجی و آزمون دست یافته است. در شرایط نویزی، روش پیشنهادی دقت بازشناسی را نسبت به مدل‌های LSTM مبتنی بر واحدهای واجی تا ٪۲۶,۸۸ افزایش می‌دهد و همچنین نسبت به روش ترکیبی استخراج ویژگی MTDRCC همراه با مدل MLP بهبود ٪۷,۶۱ را نشان می‌دهد.

کلمات کلیدی: بازشناسی ارقام گفتاری، افزایش داده، شبکه عصبی پیچشی، واحد بازگشتی گیتی دوطرفه.