Performance Analysis of Word Recognition System Using Tensor Flow

Bittu Kumar(✉), P. Sri Ram Rahul, G. Karthikeya, T. V. Sai Nithin Vishnu, M. Srikanth, and Peta Shivani

Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Hyderabad, Telangana 500075, India
bittu.mlrit@gmail.com

Abstract. Word recognition stands as a pivotal element within natural language processing (NLP) and machine learning, crucial for diverse applications ranging from automatic speech recognition to optical character recognition and from text analysis to machine translation. This paper explores TensorFlow, a robust open-source machine learning framework, to tackle the complexities associated with word recognition. The research introduces innovative methods employing deep learning and neural networks to enhance the precision and efficiency of word recognition tasks. The evolution of word-to-text recognition technologies has witnessed transformative strides in recent years, impacting various industries and applications. This study scrutinizes the performance of a CNN-based word recognition system, evaluating accuracy and time variations with changes in hyperparameters, including the number of hidden layers, epoch size, activation function, and training-testing datasets.

Keywords: CNN · Word Recognition · RNN · HMM

1 Introduction

A word recognition system holds diverse applications across numerous domains, playing a pivotal role in language-related technologies. Its impact is prominently felt in language translation systems, where accurate and efficient word recognition facilitates seamless translation processes [1]. In speech recognition, the system aids in deciphering spoken words, contributing to enhanced communication between humans and machines. Additionally, it serves a crucial role in speech-to-text conversion, converting spoken words into written text, and text-to-speech conversion, generating audible speech from written text. The integration of word recognition systems is integral to developing virtual assistants [2–4] such as Google Assistant, Amazon Alexa, and Siri, enabling these digital entities to understand and respond to spoken commands, thereby enhancing user interaction. Furthermore, in the broader context of natural language processing, word recognition forms the foundation for extracting meaningful insights and understanding the nuances of human language, facilitating advancements in various fields.

P. Pareek et al. (Eds.): IC4S 2024, LNICST 597, pp. 103–110, 2025.
https://doi.org/10.1007/978-3-031-77075-3_9

A range of methodologies can be employed to establish an effective word recognition system, each offering unique advantages. Hidden Markov Models (HMMs) [5] serve as a foundational approach employed to model the statistical properties inherent in speech, encompassing phonemes and acoustic features. Integrating Gaussian Mixture Models (GMMs) with HMMs [6] enhances the system's capability to capture the probability distributions associated with speech features, contributing to robust recognition. Convolutional Neural Networks (CNNs) emerge as powerful tools for acoustic modeling, leveraging their capacity to analyze spectrograms and extract intricate voice features. Their convolutional layers allow for effective feature extraction, enabling a comprehensive understanding of speech patterns.

Recurrent Neural Networks (RNNs) [7, 8] introduce a temporal element, facilitating the consideration of sequential dependencies in speech data. With their ability to capture context and temporal relationships, RNNs enhance the modelling of dynamic aspects in speech recognition. Deep Neural Networks (DNNs) [9] play a multifaceted role in various components of speech recognition systems, encompassing acoustic and language modeling. Leveraging multiple hidden layers, DNNs excel at capturing complex patterns within speech data, contributing to the system's overall proficiency. The utilization of diverse methods underscores the versatility and adaptability required to address the intricacies of word recognition in different contexts and applications.

In this paper, CNN is used to perform the word recognition and its performance is analysed. CNNs can efficiently extract relevant acoustic features from raw audio spectrograms, a crucial step in speech recognition. It can enhance the representation of speech data. Notably, CNNs [10] exhibit a parallel processing capability on input data, enabling swifter training and inference times, thereby proving advantageous for real-time applications. The inherent robustness of CNNs to certain types of noise and variations in acoustic environments further underscores their significance in addressing challenges encountered in real-world speech recognition scenarios, distinguishing them from Recurrent Neural Networks (RNNs).

The subsequent sections of this paper are systematically arranged to offer a comprehensive understanding of the research presented herein. Section 2 meticulously examines and reviews the relevant body of work, providing a contextual background for the study. Section 3 delves into an intricate discussion of the methodology employed, elucidating the techniques and strategies and providing details about the database utilized. The results obtained from the research findings are meticulously presented and discussed in Sect. 4, offering insights into the empirical outcomes. Lastly, Sect. 5 dedicates itself to a thorough exploration and explanation of the derived outcomes, emphasizing the distinctive contributions made by the study to the broader field of research.

2 Literature Survey

In [11], authors explored the utilization of frequency spectral information with Mel frequency as an innovative approach to enhance speech recognition within Hidden Markov Models (HMM) framework. The conventional Mel spectrum, which forms the basis of speech recognition, was augmented by incorporating frequency spectral information. The Mel frequency approach strategically observes speech frequencies within a designated resolution, leading to overlapping resolution features and limiting recognition

capabilities. A resolution decomposition technique was employed to address this limitation in the HMM-based speech recognition system, employing a mapping approach to separate frequencies effectively [12]. The study's findings revealed notable improvements in the quality metrics of speech recognition, particularly in terms of computational time and learning accuracy within the speech recognition system.

In [13], Kavita Sharma and Prateek Hakar delved into the expansive domain of speech recognition, aiming to develop solutions that transcend the limitations of single-speaker targeting. Their work focused on technologies capable of recognizing speech patterns across diverse speakers. A key challenge addressed in their study was the inherent variability in speech patterns, exacerbated by factors such as accent, environmental noise, and co-articulation. The authors introduced an innovative approach in which the function of the basilar membrane, a crucial component of the human auditory system, was emulated in the front end of the filter bank within the speech recognition system. This emulation aimed to mimic the human auditory system more closely, believing that a finer band subdivision would improve recognition results. The filter constructed for speech recognition was evaluated to differentiate between noise and clean speech, contributing to enhanced accuracy and robustness in recognizing speech across varied speakers and challenging acoustic environments.

In [14], Puneet Kaur et al., delved into the intricacies of leveraging Hidden Markov Models (HMMs) in speech recognition. Their study focused on the application of HMMs in the development of Automatic Speech Recognition (ASR) systems, recognizing the pivotal role of these models in achieving desired outcomes. The researchers outlined the essential three-step process involved in creating an ASR system. The initial step encompassed pre-processing, where raw speech signals underwent necessary transformations to enhance their suitability for subsequent analysis. Following pre-processing, the crucial feature extraction phase ensued, aiming to distil key acoustic features from the processed signals. Finally, the recognition step involved the utilization of Hidden Markov Models, which played a central role in deciphering and interpreting the extracted features to attain the desired speech recognition results. The motivation behind this research stemmed from the ongoing quest to refine ASR systems, acknowledging the substantial progress made in digital signal processing. However, the researchers noted that the performance of computers in this domain faced challenges, particularly regarding response speed and matching accuracy. The study highlighted three approaches the research community employs to enhance ASR systems. These approaches included the acoustic-phonetic approach, which focused on understanding the acoustic properties of speech; the pattern recognition approach, aimed at identifying patterns within speech signals; and the knowledge-based approach, leveraging contextual and linguistic knowledge to enhance recognition capabilities. Through these concerted efforts, the researchers aimed to contribute to the ongoing advancements in ASR technology, addressing challenges and improving speech recognition systems' overall efficiency and accuracy.

3 Methodology

This paper centres on applying Convolutional Neural Networks (CNNs) for word recognition, utilizing a carefully curated dataset comprising eight fundamental words. The dataset encompasses one thousand audio clips for each word, each lasting one second,

resulting in a comprehensive collection of eight thousand audio clips. A portion of the dataset is allocated to facilitate model training, while the remainder serves for validation and testing.

Figure 1 shows the block diagram of the word recognition system, which is considered for the evaluation. Each audio file undergoes Short Time Fourier Transform (STFT) with a window size (frame length) of 255 and a frame step of 128. The STFT application yields distinct features for each audio clip, encapsulating essential information about the sound spectrum. Initially arranged in a one-dimensional array, the feature values undergo reshaping into a two-dimensional array. Subsequently, resizing is performed to create a more manageable dataset, ensuring optimal input for the CNN model [15].

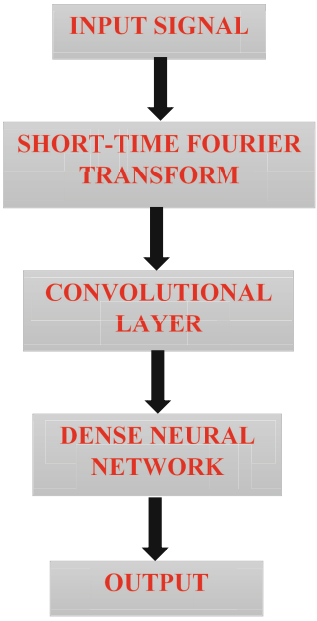

Fig. 1. Block Diagram of word recognition system

The pre-processed dataset is fed into the CNN model, where convolutional and max-pooling layers are strategically employed to iteratively reduce the data dimensions before reaching the input layer. Following these layers, a single hidden layer is introduced, featuring an activation function tailored to the model's requirements. The output layer has the 'SoftMax' activation function, providing probabilities for each class.

The model undergoes compilation and fitting to the training data, with a deliberate exploration of varying parameters such as the number of epochs, activation functions, and hidden layers. This systematic variation allows for a comprehensive analysis of the word recognition system's performance under diverse scenarios, shedding light on its adaptability and robustness.

4 Result and Discussion

This investigation delves into the application of Convolutional Neural Networks (CNNs) for the task of word recognition, aiming to provide a thorough assessment of its performance. CNNs, renowned for their capability to extract relevant acoustic features directly from raw audio spectrograms, play a pivotal role in the landscape of speech recognition. This study focuses explicitly on implementing CNNs using Tensor-Flow, a robust framework for machine learning.

To facilitate the training process, a curated dataset comprising eight distinct words is employed, each accompanied by an audio dataset consisting of 1000 recordings, each lasting one second. The abundance of data presents a comprehensive set of variations and scenarios, enriching the learning process for CNN. The extensive dataset undergoes meticulous pre-processing before being fed into the Convolutional Neural Network. The model, built using Tensor-Flow, is designed to learn intricate patterns inherent in the acoustic features of the audio data. Through this learning process, CNN becomes adept at predicting the word to which a set of features belongs. By undertaking this exploration and analysis, the study aims to provide valuable insights into the effectiveness of CNNs for word recognition tasks, shedding light on their potential contributions to achieving enhanced recognition accuracy in speech-processing applications.

Table 1. Accuracy for activation functions

Activation function	Accuracy
Relu	87.26
tanh	86.42
sigmoid	87.50
linear	86.29

Table 1 provides insights into the accuracy of Convolutional Neural Networks (CNNs) across various activation functions, aiming to identify the most effective activation function for a high-accuracy word recognition system. The results of this experimentation suggest that the sigmoid activation function yields the highest accuracy, with relu closely following. This analysis optimises the CNN model, offering valuable information for selecting the most suitable activation function to enhance word recognition accuracy.

Table 2 illustrates the accuracy and time metrics for various epoch sizes, aiming to identify the optimal configuration for achieving high accuracy in a word recognition system. The results indicate that, from this experimentation, 15 epochs yield the best accuracy when using the sigmoid activation function. Additionally, an observed trend is that as the epoch size increases, the training time also experiences an increase.

Table 3 displays the accuracy across various hidden layers, aiming to identify the optimal configuration for achieving high accuracy in a word recognition system. The results indicate that, at the first layer, CNNs attain maximum accuracy using the sigmoid activation function with an epoch size of 15 for the word recognition system.

Table 2. Accuracy and time for different Epochs

Epochs	Accuracy	Time
10	84.13	10s
15	88.83	15s
20	87.50	20s
25	87.38	25s

Table 3. Accuracy for different hidden layers

No of Layers	Accuracy
1	87.50
2	85.70
3	84.00
4	84.25

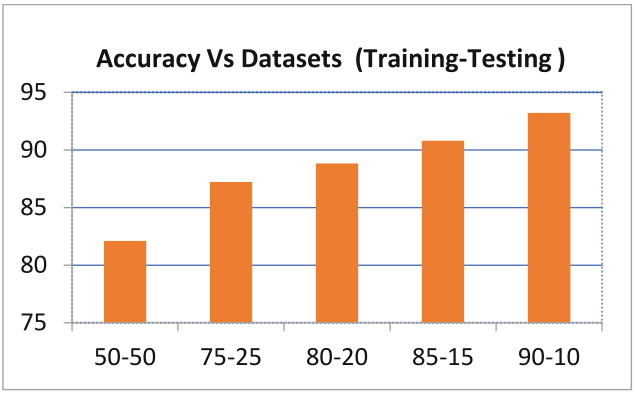

Fig. 2. Accuracy for different combination of datasets

Figure 2 illustrates the accuracy across various datasets. The y-axis represents the CNN model accuracy in percentage, while the x-axis indicates the training-to-test split (left and right sides represent the percentages allocated to training and testing sets, respectively). The observation from Fig. 2 reveals that, at the initial layer, CNNs achieve peak accuracy of 93.23% under the training dataset (90% of the total datasets) and the testing dataset (10% of the total datasets). The model utilizes the sigmoid activation function with an epoch size of 15 for the word recognition system.

5 Conclusion

In this study, we extensively explored the application of Convolutional Neural Networks (CNNs) for word recognition, achieving a notable validation accuracy of 88.83% through careful hyperparameter optimization. Both Sigmoid and ReLU activation functions were effective, with Sigmoid showing a slight accuracy advantage. The research highlights the importance of tuning parameters and considering activation function choices based on network architecture and dataset characteristics. While promising, there is room for further investigation, suggesting avenues such as exploring advanced neural network architectures and experimenting with different activation functions. The success of the CNN model in word recognition holds significant potential for real-world applications, simplifying processes in document analysis, speech recognition, and automated transcription services, leading to increased efficiency and productivity.

References

1. Nassif, A.B., et al.: Speech recognition using deep neural networks a systematic review. IEEE Access **7**, 19143–19165 (2019)
2. Kumar, B.: Comparative performance evaluation of MMSE-based speech enhancement techniques through simulation and real-time implementation. Int. J. Sp. Tech **21**(4), 1033–1044 (2018)
3. Kumar, B.: Real-time performance evaluation of modified cascaded median-based noise estimation for speech enhancement system. Fluct. Noi. Lett. **18**(4) (2019)
4. Kumar, B.: Comparative performance evaluation of greedy algorithms for speech enhancement system.Fluct. Noi. Lett. **20**(2), (2021)
5. Ananthi, S., Dhanalakshmi, P.: Speech recognition system and isolated word recognition based on Hidden Markov model (HMM) for Hearing Impaired. Int. J. Comput. Appl. **73**(20), 30–34 (2013)
6. Zhang, Y., Mike, A., Roberto, T.: Using Gaussian mixture modeling in speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. I/613–I/616. Australia (1994)
7. Graves, A., Abdel-rahman, M., Geoffrey, H.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics. speech and signal processing, pp. 6645–6649. IEEE, Canada (2013)
8. Amberkar, A., et al.: Speech recognition using recurrent neural networks. In: International Conference on Current Trends Towards Converging Technologies (ICCTCT), pp. 1–4. IEEE, Coimbatore (2018)
9. Singh, D.A., Singh, W.: A comprehensive survey on automatic speech recognition using neural networks. Mult. To. Appl. **83**(8), 23367–23412 (2024)
10. Alsaeedi, A., et al.: Arabic words recognition using CNN and TNN on a smartphone. In: 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR), pp. 57–61. IEEE, London, UK (2018)
11. Patel, I., Rao, Y.S.: A frequency spectral feature modeling for hidden markov model based automated speech recognition. In: International Conference on Web and Semantic Technology, pp. 134–143 (2010)
12. Patel, I., Rao, Y.S.: Speech recognition using hidden Markov model with MFCC-subband technique. In: International Conference on Recent Trends in Information, Telecommunication and Computing, pp. 168–172. IEEE, Kerala (2010)

13. Kumar, B., et al.: Comparative studies of single-channel speech enhancement techniques. IETE J. Res, pp.1–17 (2023)
14. Singh, B., Neha, K., Puneet, K.: Speech recognition with hidden Markov model a review. Int. J. Adv. Res. in Comp. Sci. Soft. Eng. **2**(3), 400–403 (2012)
15. Kumar, B., Varma, A.K.: FPGA implementation of dynamic quantile tracking based noise estimation for speech enhancement. J. Eng. Sci. Tech. Rev. **16**(4) (2023)