Iraqi Journal for Computer Science and Mathematics

Journal Homepage: http://journal.esj.edu.iq/index.php/IJCM
ISSN: 2788-7421

# SPEAKER IDENTIFICATION MODEL BASED ON DEEP NURAL NETWOKS

**Saadaldeen Rashid Ahmed[1,*], Zainab Ali Abbood[2], hameed Mutlag Farhan[3], Baraa Taha Yasen[3], Mohammed Rashid Ahmed[4], Adil Deniz Duru[5]**

[1]Computer Science, Tikrit University, Tikrit, 34001, Iraq
[2]Computer Tech. Engineering, AL-Esraa University, Baghdad, Iraq
[3]Electrical and Computer Engineering, Altinbas University, Istanbul, 34000, Turkey
[4]Computer Engineering, Karabuk University, Karabuk, 34000, Turkey
[5]Physical Education and Sports, Marmaraa University, Istanbul, 34000, Turkey

*Corresponding Author: Saadaldeen Rashid Ahmed

DOI: https://doi.org/10.52866/ijcsm.2022.01.01.012
Received October 2021; Accepted December 2021; Available online Januery 2022

**ABSTRACT:** This study aims is to establish a small system of text-independent recognition of speakers for a relatively small group of speakers at a sound stage. The fascinating justification for the International Space Station (ISS) to detect if the astronauts are speaking at a specific time has influenced the difficulty. In this work, we employed Machine Learning Applications. Accordingly, we used the Direct Deep Neural Network (DNN)-based approach, in which the posterior opportunities of the output layer are utilized to determine the speaker's presence. In line with the small footprint design objective, a simple DNN model with only sufficient hidden units or sufficient hidden units per layer was designed, thereby reducing the cost of parameters through intentional preparation to avoid the normal overfitting problem and optimize the algorithmic aspects, such as context-based training, activation functions, validation, and learning rate. Two commercially available databases, namely, TIMIT clean speech and HTIMIT multi-handset communication database and TIMIT noise-added data framework, were tested for this reference model that we developed using four sound categories at three distinct signal-to-noise ratios. Briefly, we used a dynamic pruning method in which the conditions of all layers are simultaneously pruned, and the pruning mechanism is reassigned. The usefulness of this approach was evaluated on all the above contact databases.

**Keywords:** Machine learning; Deep neural network; DNNs; Speaker identification

## 1. INTRODUCTION

This work focuses on Deep Neural Network (DNN)-based text-independent, closed-set speaker identification for a relatively limited number of users. Additional objectives include reducing DNN complexity and assessing the system's robustness to acoustic background noise and variability of the telephone handset. The above-stated problem and associated goals were inspired by the requirements of the NASA Johnson Space Center (JSC) for their application to the International Space Station (ISS). We are in discussion with the NASA-JSC engineers from the Human Interface Branch in Houston [1]. The ISS application requires a low-complexity solution with a low power consumption and a small footprint. DNNs can learn complicated functions by using a large number of hidden layers, providing the network "depth". Fewer layers may also be capable of learning complex functions with the same number of parameters as "deep" models in certain cases. Accordingly, deeper networks are not necessary for all applications [2]. The advantage of having multiple layers is that they can learn features at various levels of abstraction. DNNs require a considerable amount of data for training.

If the amount of data used to train the network is insufficient, the network may fail to produce reliable performance under test conditions. However, the performance of a system can be improved using a suitable algorithm for training. Applications, such as search engines and Facebook and iPhone image searching tasks, use deep learning. In these cases, providing sufficient data for training is not a problem because there are millions of users every day. [3] However, DNNs with a complex design may not be necessary for a speaker recognition application involving a closed- set of a relatively small number of users and a limited training database. In this case, a DNN with a smaller number of parameters can be used. Instead of treating a DNN as a black box, this research uses the knowledge of machine learning in configuring the DNN with just enough parameters [4]. Moreover, this research portrays the procedure of interruption identification and classification with the assistance of a few techniques that are somewhat sure in deciding the classification procedure and characterizing them to accomplish higher precision utilizing logistic processing with the clusters situated in a wireless network (WN) with low execution time. This WN technique classifies the real-time attacks in an software defined networking (SDN)-based network. Machine learning (ML) methods aim to predict the quality of images as perceived by humans without access to a reference image. Recently, deep learning methods have gained substantial attention in the research community and have proven useful for DNN. Although a previous study of DNN methods has been presented, some novelty DNN methods that are recently proposed are not summarized for DNN. In this work, we provide a survey covering various DNN methods for DNN. First, we systematically analyze the existing DNN-based quality assessment methods according to the role of DNN.

## 2. LITERATURE REVIEW

Hidden Markov Model (HMM), Support Vector Machine (SVM), and GMMM were used by earlier speaker recognition systems (Gaussian Mixture Model). SVM is one of the discriminative classification algorithms. The averaging process requires the speech utterances to be long enough to obtain the long-term averages; otherwise, it may result in loss of speaker specific information [5]. In the text-independent speaker recognition, the HMM-based speech recognizer was used for segmentation at the front end to improve the performance. However, this approach resulted in a huge computational complexity and only a minor increase in accuracy. Gaussian mixture densities were used for speaker identification [6]. Telephone handset variability in the speech database causes performance degradation in speaker recognition. The telephone channel effects are nonlinear in nature, and they are coupled with speaker specific information, making it difficult to isolate and remove them from the features, thereby degrading the system performance [7]. Reference [8] described two approaches: Sparse speaker representation approach and discriminative regularization approach to train the Universal Background Model. Figure (1)shows the manner by which deep networks work with hidden layers.

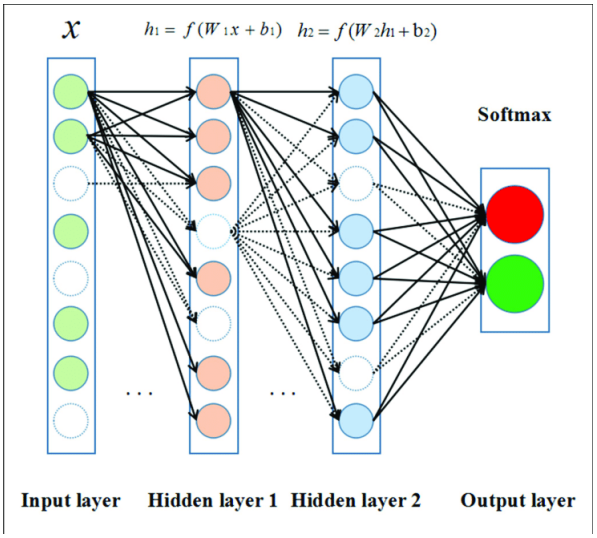


**FIGURE 1. The structure of DNN neurons are represented by circles.**

### 2.1 DNN-BASED SPEAKER RECOGNITION

Over the last few years, deep learning has gained wide acceptance in speech processing research. Several organizations, including IBM, Google, and Microsoft, have succeeded in using DNNs for acoustic modeling of speech [9]. DNN can

be used as a feature extractor or as a classifier. Among the applications used and their importance in the field of machine learning, especially in deep learning, are as follows:

- Clarify the concept of feature data and provide a corresponding processing model.
- Propose a feature data processing system to make the data fit the DNN.
- Given the exact application of the proposed system on a network traffic use case.
- Provide an effective reference for the promotion of DNN applications.

### *2.1.1. Indirect DNN approach*

The indirect method is one of the two approaches to designing a speaker classification method where the trained DNNs are used to extract the features. The features are used to train a secondary classifier for the role of speaker identification. The DNN that was equipped for a different function could be modified and used for a particular mission. For example, the DNN learned for automatic speech recognition could be used for speaker or language recognition. The posterior percentages of DNN performance and disadvantage characteristics are used as features for the secondary classifier, such as i-vector. A voice activity scanner is used to capture speech-only segments. The 600D i-vectors are derived from and are normalized in length by vectors of the stacked mean function. The average training sample is used as a target model for the language or speaker recognition assignment. Figure 1 shows the used DNN design in the indirect DNN process.

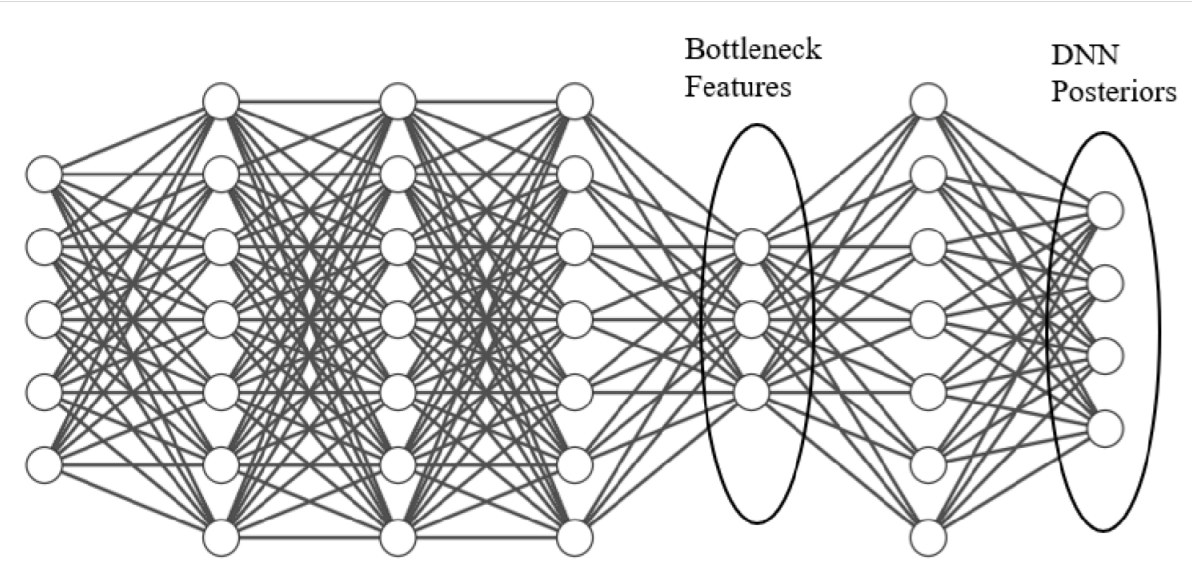


**FIGURE 2. DNN architecture used in the indirect DNN approach.**

The DNN is trained at the frame level to classify speakers. The features from the last hidden layer are extracted and used for speaker modeling. The average of these feature vectors over the training database is called as deep vector or d-vector, and it is considered as a speaker model [10]. The d-vector from a test speech utterance is extracted by passing it through the DNN layers and compared with the saved target model to perform the recognition task. The d-vector classifier provides a robust performance in acoustic background noise. The d-vector approach has been shown to outperform the i-vector approach.

### *2.1.2. Direct DNN approach*

In the direct DNN approach, the trained neural networks are used as a classifier to recognize the speaker. The input layer in the direct approach represents the dimension of the input spectral features, followed by three or more hidden layers and an output layer. The dimension of the output layer is equal to the number of speakers the system is designed to identify. The frame level DNN posteriors from the output layer must be combined by simply averaging over the test utterance [11].

A DNN can be utilized to train and predict the speaker without using a secondary classifier for a small footprint, low resource application. The use of a secondary classifier requires additional computational resource, which is not suitable for a small footprint system. The system performance may be slightly improved by increasing the number of hidden layers,

but it results in increased complexity. During testing, the frame-level aggregated DNN posteriors are averaged to produce a single decision [12]. The direct DNN architecture for a speaker recognition is shown in

The mini-batch stochastic gradient descent algorithm speeds up the training process by processing the training data in small batches. DNN learns the basic features of the speaker in the lower layers and complex features in the higher layers [13]. The text-independent speaker recognition system is a relatively complex problem because it requires all possible speech contexts for training. The system should be trained in all operating conditions, including background noise and telephone handset variability, to achieve a robust performance, as shown in the chart.

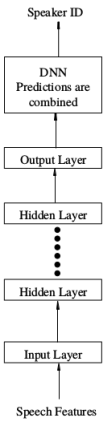


**FIGURE 3. Direct DNN approach**

The first level is an increasing chirp, the second is decreasing, the third is increasing, and the fourth is decreasing. Any two consecutive chirps will be orthogonal owing to this setup. This work is created utilizing the rationale-augmented convolutional neural network (CNN) [14]. In this research vehicle detection system from infrared images using YOLO (You Look Only Once) computational mechanism [15]. Data clustering is a crucial topic in machine learning. It can be used for a variety of purposes, one of which being image segmentation [16]. SVM, K-Nearest Neighbors (KNN), Decision tree, Logistic Regression, and Artificial Neural Network back propagation are a few of the many categorization models available. We will look at various procedures and methods for early diagnosis of glaucoma disease using the MATLAB Deep CNN [17]. Enhancement of the medical diagnostic procedures is one of the issues that has occupied many publications, with several articles beginning to elicit methods to raise the efficiency of illness diagnosis [18]. In this study, we claim an accuracy of 88.4% for the five-class grouping assignment. At the high-affectability working point, we report 92.3 percent exactness, 96.2 percent, and affectability 94.5 by 87.2 percent for the four-class grouping attempting to distinguish carcinomas. In the opinion of everyone [19]. We claim an accuracy of 88.4% for the five-class grouping assignment. At the high-affectability working point, we report 92.3 percent exactness, 96.2 percent, and affectability 94.5 by 87.2 percent for the four-class grouping attempting to distinguish carcinomas. According to various scholars [20], wireless systems can be utilized in various areas. These wireless systems require a flexible mechanism for the allocation of the available time and frequency resources. The framework time-recurrence designation is troublesome in the regular OFDM framework [21]. The process of energy efficiency improvement in any cellular network requires that the network design be densified to allow for more spatial reuse while maintaining the user quality of service. This study will analyze the combination of two densification techniques, namely, the small cell access point and the massive multiple input–multiple-output (MIMO) base [20]. The main goal of this research is to find the best tangent space central point for Tangent Space Linear Discriminant Analysis-based Motor-imagery Brain–Computer Interface [22]. We use Dedicated Short-range Communication techniques to ensure encrypted vehicular communication employing wireless controller area network performance at high node density in this work. The influence of vehicle communication parameters, such as message rate, data rate, transmission power, and carrier detection threshold, on the application performance is investigated. We present a data-rate DSR algorithm [23] based on a state-of-the-art analysis.

## 3. METHODOLOGY

The following tasks were carried out in this research with this proposed approach: (1) development of a baseline DNN system; (2) evaluation of the performance of the baseline system over the three speech databases mentioned above; (3) investigation of the complexity reduction techniques; and (4) evaluation of the performance of the final complexity-reduced system over the same three speech databases. The rest of the chapter explains the details of HTK, CNTK, and

speech databases used in this work. The training begins with the data in the first section of the chart and extracts the required characteristics from it by using HTK and CNTK. Then, we measure the performance of the results. If errors appear, then the error returns to the penultimate stage (i.e., CNTK), as shown in the chart to be retrained, and the desired results are extracted. Figure 4 a and b explain the block diagram of the proposed method

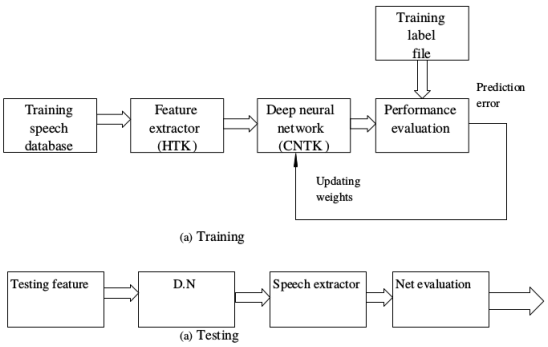


**FIGURE 4. Block diagram of the proposed method**

### 3.1 ANALYSIS OF THE SITUATION OR DISCUSSION

There is no rule for selecting the number of hidden units to use in DNN. Given that this work is focused on designing a closed-set small footprint speaker recognition system, it is our goal to design a model with just enough parameters. Hence, we conducted experiments to find the right number of hidden units per layer for our problem that are applicable to all three databases: TIMIT, HTIMIT, and noise-added TIMIT. We initially trained the network on full utterance speech data, but it failed to provide acceptable results because it is not considered an effective training approach in speech applications. We experimented with different learning rates in the exponential range of 0.1, 0.01, and 0.001 and found that the learning rate of 0.001 provides good overall performance for our training data. We conducted the following experiments using HTIMIT database.

The hidden layer units are a linear function of their inputs. The activation function transforms the linear hidden units into a nonlinear function. We initially used sigmoid as the activation function, but it resulted in the vanishing gradient problem. We switched to the ReLu activation function to address this problem.

## 4. SOLUTIONS PROPOSED

Reducing the complexity of the model is critical in a small footprint DNN-based speaker recognition system. When the available resources are limited, a simplified DNN speaker recognition system can be designed using a suitable complexity reduction technique. Complexity reduction of the model can be achieved in several ways. This research is focused on reducing the number of parameters in the DNN model without significant loss in the accuracy. Pruning is a technique of removing the parameters from the network using a suitable algorithm. The pruning technique converts the fully connected dense network into a sparse network.

As noted earlier, this research uses the CNTK deep learning library to design the DNN. However, one disadvantage of utilizing CNTK for pruning is that the CNTK parameters can only be dense when they correspond to fully connected hidden layers.

Therefore, we had to come up with a different approach to implement pruning using CNTK. Instead of removing the parameters from the network, we set the less important parameters to zero and explicitly developed a python-based function in CNTK to avoid the zero weights from being updated while training the network.

## 5. ADAPTIVE PRUNING

Pruning is a technique of removing parameters, namely, weights, from the neural network without significant loss in accuracy. In a pruning technique, any weight that is less than a certain threshold is considered less important and is discarded or zeroed out after training is completed. In the adaptive pruning technique, the standard deviation of each layer's weights multiplied by a quality factor is used as a threshold, making the threshold data-adaptive rather than fixed. The weights below this adaptive threshold are removed from the network (or zeroed out as previously mentioned), and the network is retrained. Failing to retrain will significantly affect the network's performance

**Table 1. Performance of adaptive pruning on TIMIT Data**

| Condition | Performance | Parameters left (complexity reduction) |
|---|---|---|
| Training set | 98.73% | 254K (9.5X) |
| Testing set | 95.23% | |

## 6. CONCLUSION

The complexity-reduced DNN provided comparatively similar results as the complex baseline DNN. Hence, we have achieved the goal of developing a DNN with reduced complexity for small footprint applications.

The main contributions of this research include an optimized baseline DNN configuration for a small footprint system. Baseline DNN provides error-free performance for clean speech and a robust performance under handset variability and acoustic background noise. Complexity of DNN is reduced using complexity reduction techniques. Methodology is developed to customize CNTK for implementing pruning of weights. Adaptive pruning is not as effective in nonhomogeneous database conditions. Finally, we describe some emerging challenges in designing and training a DNN-based ML, along with few directions that are worth further investigations in the future.

## ACKNOWLEDGEMENTS

The first author thanks the reviewers for their useful suggestions that improved the presentation of this paper.

## CONFLICTS OF INTEREST

The authors declare no conflict of interest.

## REFERENCES

[1] G. Salazar, A. Romero, and D. Juge, "Personal communication, 2016, Human Computer Interface," *Avionics Systems Division, NASA Johnson Space Center,Houston,TX*, 2016.
[2] B. Jimmy and R. Caruana, "Do Deep Nets really need to be Deep?," *in NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing systems*, 2014.
[3] G. Fant, "Phonetics and Phonlogy in the last 50 years," *in Dept. of Speech, Music and Hearing, KTH,Sweden*, 2004.
[4] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, "The Kaldi Speech Recognition Toolkit," *in IEEE Signal Processing Society, Hilton Waikoloa Village, Big Island, Hawaii,US*, 2011.
[5] D. Yu, A. Eversole, M. L. Seltzser, K. Yao, and B. G. Huang, *An Introduction to Computational Networks and the Computational Network Toolkit*. 2015.
[6] E. J. Bradley, K. Panagiotis, A. Zeynettin, K. Timothy, and K. Philbrick, "Toolkits and Libraries for Deep Learning," *Digit Imaging*, 2017.
[7] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Transactions on speech and Audio Processing*, 1995.
[8] D. Garcia-Romero, X. Zhang, and A. Mccree, "Improving Speaker Recognition Performance in the Domain Adaptation using Deep Neural Networks," in *IEEE Spoken Language Technology Workshop*, 2014.
[9] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohammed, N. Jaitly, A. Senior, and V. Vanhoucke, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine*, 2012.
[10] J. P. Campbell, "Speaker Recognition: A Tutorial," *Proceedings of the IEEE*, 1997.
[11] D. Reynolds, "Speaker Identification and Verification using Gaussian mixture speaker models," *Speech Communications*, 1995.
[12] F. Richardson, D. Reynolds, and N. Dehak, "Deep Neural Network Approaches to Speaker and Language Recognition," *IEEE SIGNAL PROCESSING LETTERS*, 2015.
[13] E. Variani, X. Lei, E. Mcdermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep Neural Network For Small Footprint Text-Dependent Speaker Verification," *IEEE International Conference on Acoustic, Speech and Signal Processing*, 2014.
[14] S. Ahmed, . R. Ahmed, and E. Sonuç, "Deepfake detection using rationale-augmented convolutional neural network," *Applied Nanoscience*, vol. 2021, pp. 1–9.
[15] MAHMOOD, Mohammed Thakir; AHMED, Saadaldeen Rashid Ahmed; AHMED, Mohammed Rashid Ahmed, "Detection of vehicle with Infrared images in Road Traffic using YOLO computational mechanism," *In: IOP Conference Series: Materials Science and Engineering. IOP Publishing. p. 022027.*, 2020.
[16] S. Abdulateef, . Khalid, S. Ahmed, . R. Ahmed, M. Salman, and Dawood, "A Novel Food Image Segmentation Based on Homogeneity Test of K-Means Clustering," *IOP Conference Series: Materials Science and Engineering*, pp. 32059–32059, 2020.
[17] M. Ahmed and Rashid, "An Expert System to Predict Eye Disorder Using Deep Convolutional Neural Network," *Academic Platform Journal of Engineering and Science*, vol. 9, pp. 47–52.
[18] M. . Waleed, A. S. . Abdullah, Ahmed, and R. Saadaldeen, "Classification of Vegetative pests for cucumber plants using artificial neural networks," *2020 3rd International Conference on Engineering Technology and its Applications (IICETA)*, pp. 47–51, 2020.
[19] S. R. Ahmed and Ahmed, "Breast cancer detection and image evaluation using augmented deep convolutional neural networks," *Aurum journal of engineering systems and architecture*, vol. 2, no. 2, pp. 121–129, 2018.
[20] "Universal Filtered Multicarrier (UFMC) vs. Orthogonal Frequency Division Multiplexing (OFDM)," *Journal of Physics: Conference Series*, vol. 1530, no. 1, 2020.

[21] M. Al-Qaraghuli, S. Ahmed, and M. Ilyas, "Encrypted Vehicular Communication Using Wireless Controller Area Network," *Iraqi Journal for Electrical and Electronic Engineering. sceeer*, pp. 17–24, 2020.

[22] A. S. M. Miah, S. R. A. Ahmed, M. R. Ahmed, O. Bayat, A. D. Duru, and M. K. I. Molla, "Motor-Imagery BCI task classification using riemannian geometry and averaging with mean absolute deviation," in *International Scientific Meeting of Electrical-Electronics &*, pp. 24–26, 2019.

[23] S. R. Ahmed and Ahmed, "Energy improvement using Massive MIMO for soft cell in cellular communication," *IOP Conference Series: Materials Science and Engineering*, vol. 928, 2020.