Journal of Innovative Technology, Vol. 6, No. 2, pp. 1-8, September 2024
http://doi.org/10.29424/JIT.202409_6(2).0001

An Intelligent Model for Speaker Verification System Using Pattern Recognition

Rooh Ullah Khan [1], Zainab Raheem [2], Chung-Chian Hsu [3], Maqsood Hayat [4], Aneela Kausar [5], Naveed Ishtiaq Chaudhary [6*]

ABSTRACT

Speaker recognition has evolved over nearly five decades, with speech standing out as the most intuitive mode of communication. The i-vector has long held its position as the pinnacle of technology in speaker verification. However, this proposed work introduces deep learning technology with the aim of surpassing the established i-vector in speaker verification applications. Numerous techniques have been explored in prior research to enhance speaker accuracy, but the integration of deep learning techniques marks a significant and revolutionary shift. This research aims to establish an automated deep-learning framework specifically designed to enhance the discriminative power of speaker verification representations. We conducted various experiments on the VoxCeleb-1 database to assess the performance of different deep learning methods, including the use of multiple activation functions and optimizers. These experiments were designed to evaluate the effectiveness of the algorithms, and we validated our proposed system's performance using benchmark dataset tests. Our system achieved its highest success rate by utilizing the Relu activation function, employing Stochastic gradient descent (SGD) as the optimizer, and incorporating a second layer. This resulted in a notable decrease in the Equal Error Rate (EER) from 17.6 to 9.93, representing an approximate 50% improvement in accuracy on the benchmark tests. These results clearly indicate that our automated model surpasses existing literature in this area. We anticipate that our proposed model will be a valuable asset for researchers and the academic community, facilitating further exploration and advancement in this field.

Keywords: Sounds detection, speech recognition, artificial intelligence, deep neural network, speaker recognition, signal processing, unsupervised learning.

1. INTRODUCTION

In recent decades, there has been a remarkable surge in the advancement of speaker recognition technology. In 1997, computer systems were limited to comprehending just over 1000 words, a capacity that leaped to 20,000 words by the 1980s. The evolution of speaker recognition technology has been notably spearheaded by IBM, which introduced the initial consumer-oriented product, Dragon Dictate, in the 1990s. Subsequently, 1996 marked a significant milestone in voice recognition with the introduction of a groundbreaking product by IBM, marking a pivotal moment in the field. The technological landscape further evolved with Google's introduction of its voice search app for the iPhone, followed by the launch of Apple's Siri. The past decade has witnessed a surge in the development of advanced voice recognition systems by leading technology companies. This is evident in the introduction of prominent virtual assistants like Amazon's Alexa, integrated within the Echo device, and Microsoft's Cortana. This paper explores the trajectory of speaker recognition technology, delving into its historical milestones and the transformative impact of innovative products and systems introduced by industry leaders. Efforts by researchers to develop an efficient and reliable speaker recognition system have led to various advancements. Initial efforts focused on identifying individuals, followed by the introduction of automatic systems [1]. Previous research has explored the use of features extracted from both the speaker's vocal tract (source) and the recording system (channel) for speaker recognition tasks [2]. Speaker recognition systems have primarily been utilized in facilities and network access-control applications [3]. Comparisons have been made between the best speaker-specific information and information derived from Mel-Frequency Cepstral Coefficients (MFCC) [4].

The field of speaker pattern recognition has witnessed a surge in research activity in recent years. Initial efforts relied on private datasets compiled by individual researchers, focusing on a limited range of sounds [5-6]. For instance, Woodard employed a hidden Markov model (HMM) to classify just three categories: opening/shutting wooden doors, dropping metal objects, and pour-

Manuscript received December 20, 2023; revised April 2, 2024; accepted April 17, 2024.

[1] Master Student, International Graduate School of Artificial Intelligence, National Yunlin University of Science and Technology, Taiwan (ROC)

[2] Master Student, Department of Computer Science, Abdul Wali Khan University, Pakistan.

[3] Professor, International Graduate School of Artificial Intelligence, National Yunlin University of Science and Technology, Taiwan (ROC)

[4] Professor, Department of Computer Science, Abdul Wali Khan University, Pakistan.

[5] Ph.D. Scholar, Department of Computer Science and Information Engineering, Graduate School of Engineering Science and Technology, National Yunlin University of Science and Technology, Taiwan (ROC)

[6*] Assistant Professor (corresponding author), Future Technology Research Center, National Yunlin University of Science and Technology, Taiwan (ROC). (email: chaudni@yuntech.edu.tw)

ing water [5]. A turning point came with the introduction of publicly available datasets through the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge series [7-9]. DCASE offered datasets for tasks like acoustic scene classification and sound event detection, fostering broader research interest. The 2019 DCASE challenge alone attracted 311 entries across five subtasks, highlighting the increased engagement [10].

However, a key question remains: how effectively can audio pattern recognition systems leverage large-scale datasets for training? In computer vision, the extensive ImageNet dataset has fueled the development of image classification systems [11]. Similarly, natural language processing has benefited from large text datasets like Wikipedia for building language models [12]. In contrast, training on large-scale audio datasets has been less explored [13-15].

A significant development was the release of AudioSet, a dataset boasting over 5,000 hours of audio recordings categorized into 527 sound classes. Notably, AudioSet provided pre-extracted embedding features derived from a trained convolutional neural network, rather than raw audio data [13]. While these features have been utilized by researchers to build recognition systems [13, 16-20], their efficacy as the sole representation for audio recordings might be limited, potentially hindering system performance.

Building upon existing methods, researchers have adapted training techniques like WCCN for generalized linear kernel approaches [21]. Additionally, the field has seen exploration of deep neural networks (DNNs) integrated with hidden Markov models (HMMs) for speech recognition, aiming to achieve higher recognition accuracy [22, 23]. A novel framework employing a statistical i-vector model guided by deep neural networks has been introduced, representing a significant advancement in speaker audio recognition [24]. DNNs have also been utilized for extracting high-level features from raw data, like intelligent neuro-computational approach for piezoelectric material [25], soft computing approach for wideband transducer [26], Bouc-wen hysteresis model investigation with deep intelligent network [27], and Recurrent neural networks for piezoelectric models [28], these all showing promising results of DNNs in the field of optimization and speech emotion recognition [29].

In the domain of background noise modeling, there's a shift towards acquiring functions resilient to noise effects, eliminating the need for hand-designed components [30]. Previous research explored a connectionist hidden Markov model (HMM) system for extracting noise-resistant audio features [31]. Despite recent advancements in deep learning technology, i-vectors remain at the forefront of speaker recognition. However, there's still a need for further exploration and attention to harness the full potential of deep learning in this field [32]. This work presents an automated system leveraging artificial intelligence (AI) to refine i-vectors employed in speaker verification. We propose a novel approach that transcends the limitations of traditional i-vector representation. Our system utilizes deep learning architectures, specifically multi-layer neural networks, to extract more discriminative speaker embeddings. To develop an efficient and reliable computational model for speaker verification, benchmark datasets relevant to the domain are essential. We plan to construct or acquire benchmark datasets and train deep neural network classifiers using i-vectors and speaker labels. Subsequently, testing i-vectors will undergo transformation using the trained deep neural network, followed by cosine scoring to evaluate the transformed vectors.

1.1 Contributions and innovative insights

- This research presents a significant advancement in speaker verification by leveraging deep learning techniques to surpass the limitations of traditional methods like i-vectors. Here's a breakdown of the key contributions and innovative aspects:
- The work proposes a move from established techniques like i-vectors to deep learning architectures for speaker verification. This shift has the potential to unlock superior accuracy and robustness in speaker identification tasks.
- The research focuses on developing an automated system that leverages deep learning for improved speaker verification vector accuracy. This automation streamlines the process and facilitates wider adoption.
- The study employs a systematic approach by evaluating various deep learning methods, including activation functions and optimizers, using the VoxCeleb-1 database. This exploration provides valuable insights into the impact of different deep learning configurations on speaker verification performance.
- research validates the proposed system's effectiveness through benchmark dataset tests. The significant reduction in Equal Error Rate (EER) from 17.6 to 9.93 demonstrates a substantial improvement in accuracy (approximately 50%). This surpasses existing literature in the field.
- The proposed model is expected to be a valuable asset for researchers and the academic community. It can serve as a foundation for further exploration and advancement in speaker verification research, potentially leading to even more robust and accurate systems.

1.2 Organization

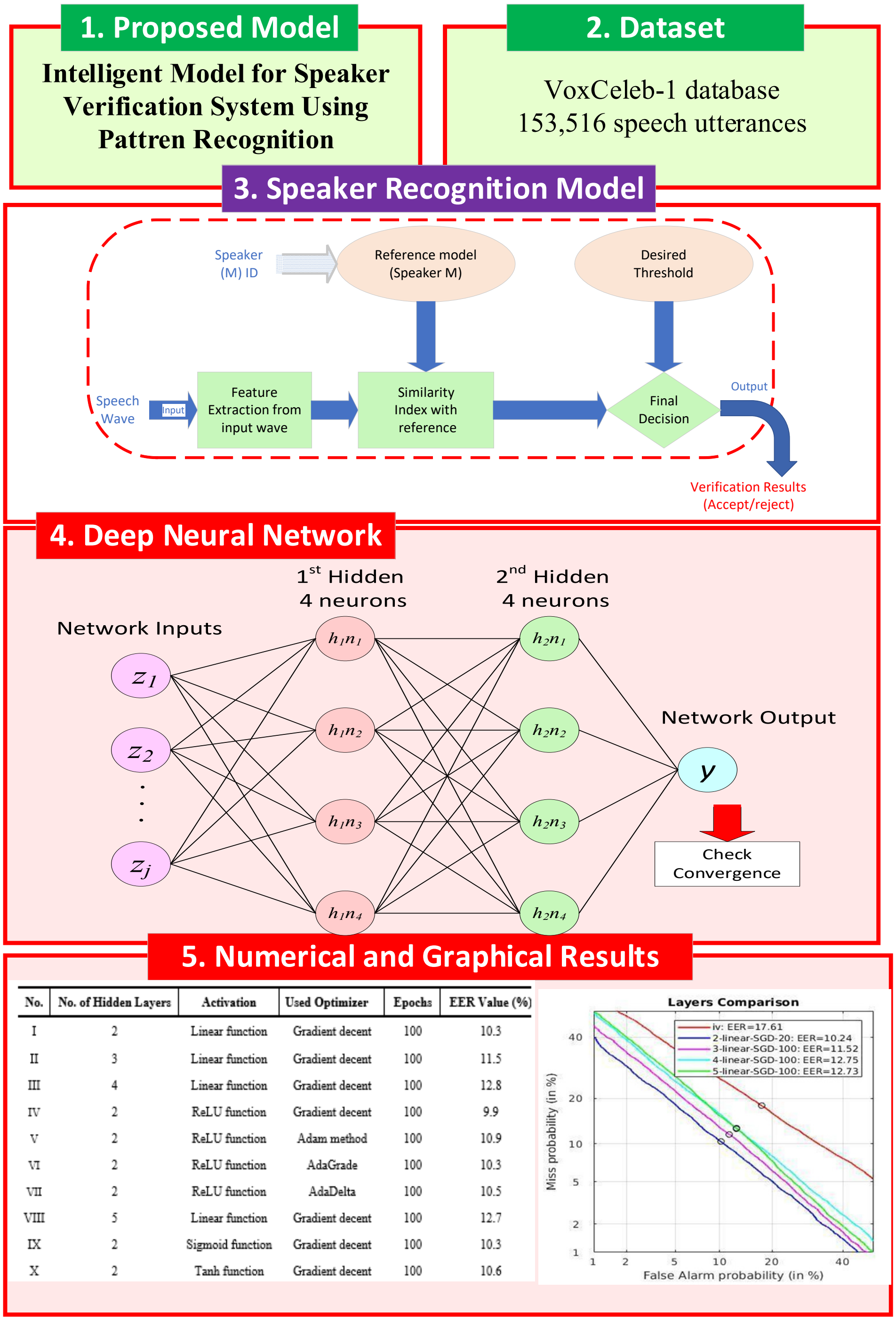
This paper is organized as follows. Section II details the proposed deep neural network (DNN)-based technique for speaker embedding extraction from i-vectors. Section III describes the experimental framework, outlining the setup and the specific database employed for evaluation. Section IV presents and analyzes the obtained results. Finally, Section V concludes the paper, summarizing the key findings and potential future directions.

2. PROPOSED METHOD

The optimal backend for i-vector implementation has been identified as Probabilistic Linear Discriminant Analysis (PLDA). However, obtaining labeled data for speakers incurs higher costs. Performance gains are minimal for short utterances but substantial for longer ones. This observation prompted researchers to explore alternative deep learning (DL) backends. Most proposed approaches leverage speaker labels from background data, yet show no significant improvement compared to PLDA [33].

NIST recently organized a speech recognition challenge addressing the achievement of comparable performance to PLDA when the development data lacks labels. One approach incorporates a specialized hybrid architecture that merges a Deep Belief Network (DBN) with a Deep Neural Network (DNN) [34]. An alternative approach involves training an end-to-end speech recognition system. This system processes data through a unified network across multiple stages, potentially offering a more streamlined and efficient approach compared to conventional, multi-stage pipelines [35]. This combination potentially leverages the strengths of both architectures to enhance speaker recognition. Various models were explored where speaker spectral features are input, yielding similarity scores as output. However, these techniques proved less competitive than other speaker embedding approaches. The whole methodology steps are presented as a graphical abstract in Fig 1.

For speaker verification investigations, a deep neural network (DNN) framework was employed. During development, the DNNs were trained to classify frame-level speakers. Our proposed speaker verification system employs a Deep Neural Network (DNN) for frame-level speaker classification. During enrollment, speaker-specific features are extracted from the final hidden layer of the trained DNNs. These features are then averaged to create a speaker model. In the evaluation phase, a D-vector is extracted for each utterance and compared to the designated speaker model for verification. Experimental results demonstrate that the DNN-based system achieves competitive performance compared to existing i-vector systems. Notably, the DNN system exhibits superior robustness and effectiveness, particularly at low false rejection rates. Overall, the proposed system surpasses the i-vector system by up to 50% in terms of Equal Error Rate (EER) for speaker verification tasks. Fig 2. Despite the block diagram of the verification system for speech recognition.


1. Proposed Model

Intelligent Model for Speaker Verification System Using Pattren Recognition

2. Dataset

VoxCeleb-1 database
153,516 speech utterances

3. Speaker Recognition Model

4. Deep Neural Network

5. Numerical and Graphical Results

No.	No. of Hidden Layers	Activation	Used Optimizer	Epochs	EER Value (%)
I	2	Linear function	Gradient decent	100	10.3
II	3	Linear function	Gradient decent	100	11.5
III	4	Linear function	Gradient decent	100	12.8
IV	2	ReLU function	Gradient decent	100	9.9
V	2	ReLU function	Adam method	100	10.9
VI	2	ReLU function	AdaGrade	100	10.3
VII	2	ReLU function	AdaDelta	100	10.5
VIII	5	Linear function	Gradient decent	100	12.7
IX	2	Sigmoid function	Gradient decent	100	10.3
X	2	Tanh function	Gradient decent	100	10.6

Layers Comparison
iv: EER=17.61
2-linear-SGD-20: EER=10.24
3-linear-SGD-100: EER=11.52
4-linear-SGD-100: EER=12.75
5-linear-SGD-100: EER=12.73
Miss probability (in %)
40
20
10
5
2
1
1
2
5
10
20
40
False Alarm probability (in %)

Fig.1 Graphical Abstract of proposed work

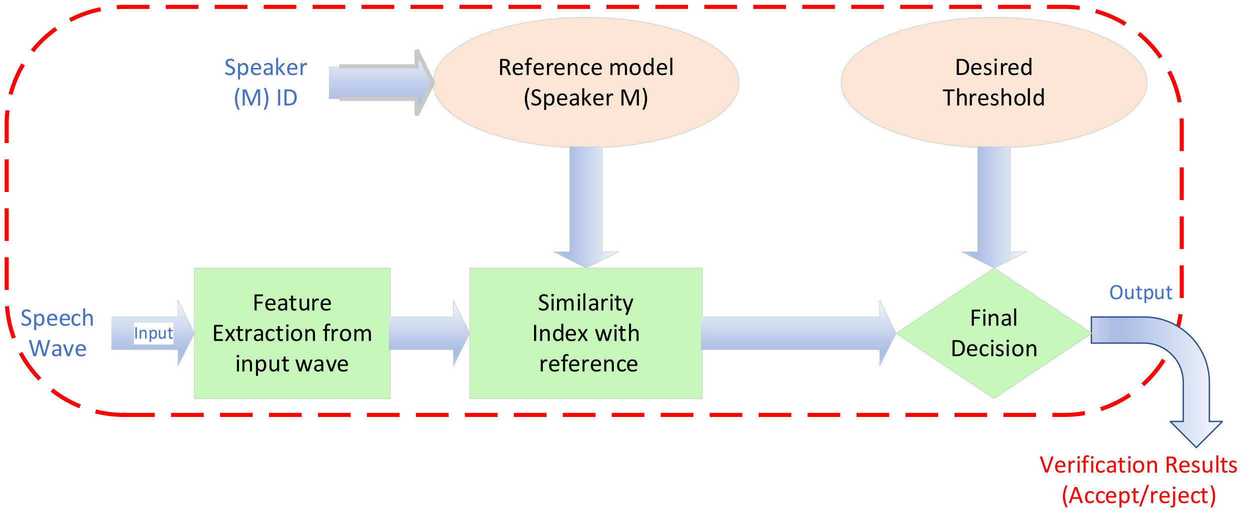


Fig.2 Block diagram of verification system for speakers

2.1 Deep neural network

In the realm of speaker recognition, deep neural networks (DNNs) have emerged as a powerful tool for identifying individuals based solely on their voice characteristics. Unlike traditional methods that rely on handcrafted features, DNNs excel in their ability to automatically learn these features directly from speech data. This section delves into the intricate workings of DNNs and how they are employed for speaker recognition. The deep neural network is composed of a sophisticated arrangement of neurons organized into various layers. The fundamental unit of a deep neural network is referred to as a neuron, the structure of DNNs is depicted in Fig 3. At the input layer of the network, i-vectors are utilized, while the output layer generates new vectors (embedded) as a result of the deep neural network processing.

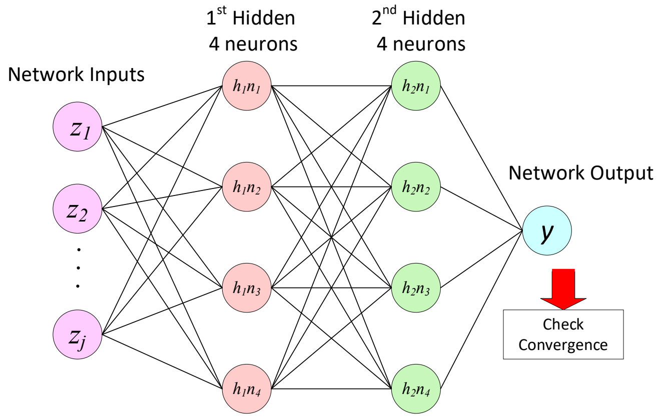


Fig.3 Basic Architecture of Deep learning Neural networks

Imagine a complex web of interconnected processing units, inspired by the structure and function of the human brain. This is the essence of a DNN. It comprises multiple layers, stacked one upon another, with each layer containing numerous artificial neurons. These artificial neurons, unlike their biological counterparts, perform simple mathematical operations on the data they receive.

The initial layer where the speech features, extracted from the audio signal, are fed into the network. These features can encompass various aspects of the voice, such as pitch, mel-frequency cepstral coefficients (MFCCs), and spectral information. The heart of the DNN, these layers are responsible for learning complex relationships between the input features and the desired output (speaker identity). Each hidden layer contains a predetermined number of artificial neurons, and the connections between these neurons carry weights that are adjusted during the training process. The final layer of the DNN produces the network's prediction. In speaker recognition, the output layer typically uses a SoftMax activation function, which assigns a probability score to each potential speaker. The speaker with the highest probability score is identified as the most likely source of the voice.

The magic of DNNs lies in their ability to learn from vast amounts of labeled speech data. This data consists of audio recordings paired with the identities of the speakers. During training, the DNN iteratively processes these recordings, adjusting the weights between its neurons to minimize the difference between the predicted speaker identity and the actual identity. The learning process is as follows:

1. Forward Pass: The speech features are fed into the input layer and propagate forward through the hidden layers. At each layer, the weighted sum of the incoming signals is passed through an activation function, introducing non-linearity and allowing the network to learn complex patterns.
2. Error Calculation: The output layer's prediction (speaker identity) is compared to the actual speaker label. The difference between the predicted and actual identity is calculated as the error.
3. Backward Pass: The error is then propagated backward through the network. This process, called backpropagation, adjusts the weights of each connection in a way that minimizes the overall error.
4. Iteration: Steps 1-3 are repeated for numerous iterations over the training data. With each iteration, the DNN progressively refines its internal representation of speaker characteristics, ultimately learning to distinguish between different voices.

2.2 Stochastic Gradient Decent optimizer (SGDO):

Stochastic Gradient Descent (SGD) is an optimization technique widely used for training deep learning models. Unlike traditional gradient descent, which considers the entire dataset for each update, SGD takes a nimbler approach. It utilizes mini-batches, smaller subsets of the training data, to calculate the gradient and update the model's parameters. This iterative process allows SGD to navigate the complex error landscape efficiently, gradually steering the model toward the optimal parameters that minimize the loss function. While SGD can introduce slight fluctuations in the learning process due to the mini-batch nature, it offers a significant advantage in terms of computational efficiency, especially for large datasets, making it a cornerstone optimizer in the realm of deep learning.

Training deep neural networks for speaker recognition involves navigating a complex landscape of errors. Stochastic Gradient Descent (SGD) acts as a powerful guide in this journey. SGD tackles the challenge by iteratively updating the network's internal parameters in small steps. During each step, it analyzes a mini-batch of training data, calculates the error associated with the current parameter settings, and adjusts the parameters in the direction that minimizes the error for that specific mini-batch. This iterative process, repeated across numerous mini-batches, gradually steers the network towards a more accurate representation of speaker characteristics. While SGD offers computational efficiency, it can take longer to converge than other optimizers. However, its effectiveness and simplicity make it a cornerstone technique for training deep learning models in speech recognition.

3. EXPERIMENTAL STEPS AND USED DATABASE

3.1 Dataset formulation

All experiments conducted in this study utilized the VoxCeleb-1 database, which comprises 153,516 speech utterances [36]. The VoxCeleb-1 database was employed for training and evaluation purposes. To maximize the available data for model development, an unlabeled learning approach was adopted. This involved utilizing both the development and test subsets during the training phase. The development section comprised 148,642 speaker utterances, representing 1,211 speakers with various accents. The test subset contained 4,874 utterances from 40 speakers and served for model assessment. Consequently, our network comprises 1,211 neurons in the classification layer.

3.2 Experiment

A total of 37,720 experiments were conducted, with the VoxCeleb-1 database's test set used for experimentation. The evaluation employed a two-part test structure, with the first half designated for non-target trials and the second half for target trials. For model training, we leveraged development data from the VoxCeleb-1 dataset. This data was used to train both the Universal Background Model (UBM) and the Total Variability (TV) matrix. Mel-Frequency Cepstral Coefficients (MFCCs) with delta features were employed throughout the process, with a consistent dimensionality of 20. Additionally, a 1024-component UBM was trained to extract 400-dimensional i-vectors. Notably, the Alize Toolkit handled all stages involving total variability matrix computation, UBM model training, and i-vector extraction [37].

The network architecture designed in this research comprises various hidden layers, each consisting of 400 neurons. The network input layer contains 400 neurons, while output layer comprises 1,211 neurons representing the 1,211 data classes. The training process employed 100 epochs with the SoftMax activation function. To optimize the network, a learning rate of 0.03 with a decay of 0.0002 and a batch size of 100 samples were chosen. These hyperparameters were selected through a separate grid search optimization process to ensure optimal training convergence. The Equal Error Rate (EER) served as the primary evaluation metric for speaker verification vector performance. The system achieved a minimum EER of 9.93% after 100 epochs, indicating a significant improvement over the baseline of 17.6%. This reduction in EER translates to a roughly 50% enhancement in speaker verification accuracy. For testing phase, various experiments were conducted to explore different parameters in the proposed Deep Neural Network (DNN) aiming to improve accuracy. Different numbers of layers were experimented with, including Layer 2, 3, 4, and 5.

4. RESULTS AND DISCUSSION

We conducted various experiments on different parameters within our proposed Deep Neural Network (DNN) to enhance accuracy. These experiments involved exploring different types of layers with varying numbers of neurons, including Layer 2, 3, 4, and 5. The Equal Error Rate (EER) exhibited a progressive increase with the addition of layers to the deep learning architecture. The EER was measured at 10.24 with 2 layers, rising to 11.52, 12.75, and 12.73 with 3, 4, and 5 layers, respectively. Based on validation accuracy, Layer 2 was determined to be the most effective (as depicted in Fig. 4), with an EER of 10.24. Consequently, Layer 2 was fixed, and further experiments were conducted on other parameters.

Next, this work explored the impact of various activation functions on the neural network's performance. We experimented with sigmoid, linear, tanh, and relu activation functions. The corresponding EER ratios obtained were: 10.34 for linear, 9.93 for relu, 10.31 for sigmoid, and 10.52 for tanh. Relu activation function demonstrated the best performance in terms of validation accuracy (as shown in Fig. 5), with an EER of 9.93. Hence, relu function was selected, and experiments continued with other parameters. Subsequently, we explored the influence of different optimizers on the network's performance, including SGD, Adam, Adagrad, and Adadelta. The resulting EER ratios were: 9.93 for SGD, 10.91 for Adam, 10.28 for Adagrad, and 10.46 for Adadelta.

SGD optimizer exhibited the highest validation accuracy (as illustrated in Fig. 6), with an EER of 9.93. Therefore, SGD optimizer was chosen, along with relu activation function and Layer 2.

After compiling all experimental results, we integrated them into a table to evaluate and compare the performance across different configurations. This integration process resulted in the following EER table, providing a comprehensive overview of the network's performance under various conditions.

By comparing this proposed work with the i-vector approach.

- Our experiments established a baseline Equal Error Rate (EER) of 17.6 using the i-vector approach.
- The proposed deep learning architecture achieved a significant reduction in EER, reaching 9.93. This translates to an approximate improvement of 44% in speaker verification accuracy compared to the i-vector baseline on the benchmark test.
- Our system's performance surpasses previously reported results in the relevant literature, demonstrating its effectiveness in speaker verification.
- In other references, the EER was 10.2; hence, this research shows enrichment in terms of accuracy with enhanced and improved results.

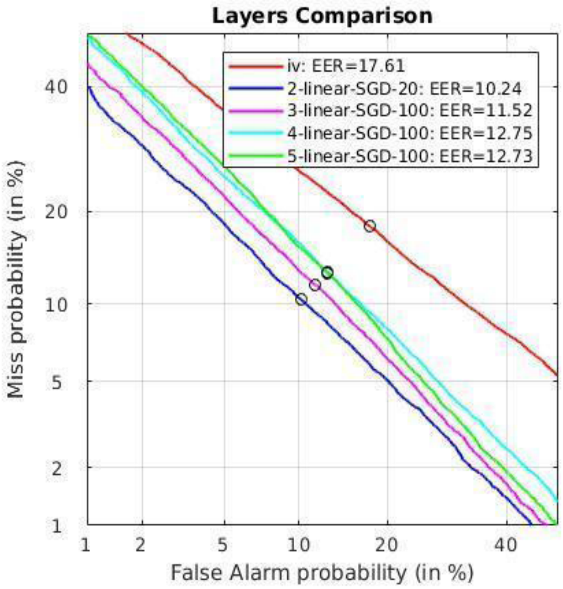


Fig.4 Layers comparison using EER

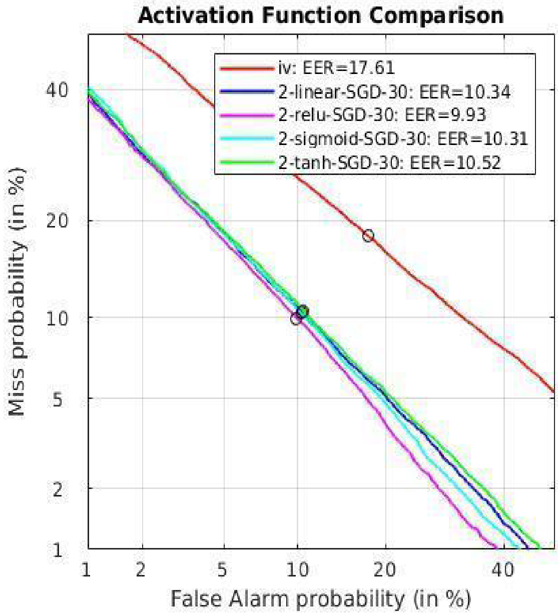


Fig.5 Activation function comparison using EER

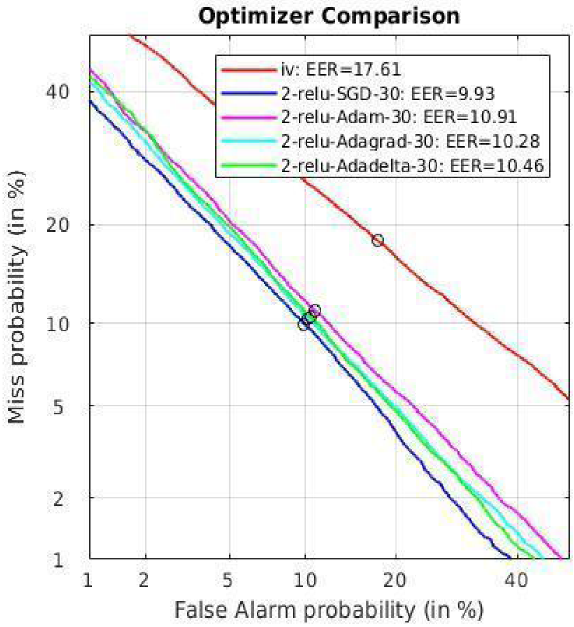


Fig.6 Optimizer's comparison using EER

Table 1 Comparison of EER Using different optimizers and activation functions

No.	No. of Hidden Layers	Activation	Used Optimizer	Epochs	EER Value (%)
I	2	Linear function	Gradient decent	1×10^{2}	10.3
II	3	Linear function	Gradient decent	1×10^{2}	11.5
III	4	Linear function	Gradient decent	1×10^{2}	12.8
IV	2	ReLU function	Gradient decent	1×10^{2}	9.9
V	2	ReLU function	Adam method	1×10^{2}	10.9
VI	2	ReLU function	AdaGrade	1×10^{2}	10.3
VII	2	ReLU function	AdaDelta	1×10^{2}	10.5
VIII	5	Linear function	Gradient decent	1×10^{2}	12.7
IX	2	Sigmoid function	Gradient decent	1×10^{2}	10.3
X	2	Tanh function	Gradient decent	1×10^{2}	10.6

Table 2 EER Comparison of proposed embedding's, baseline and i-vectors

S.No	Methods	EER Value (%)
1.	Baseline approach	17.6
2.	i-vector approach	10.2
3.	Proposed approach	9.9

5. CONCLUSION

For nearly five decades, speaker recognition has been an ongoing endeavor, primarily driven by the ubiquitous nature of speech in communication. Previous research has established i-vectors as a dominant approach in speaker recognition. However, this work explores deep learning architectures with the goal of achieving superior performance in speaker verification tasks. Through the incorporation of deep learning techniques, this research aims to significantly enhance the accuracy and precision of speaker verification. Various methodologies have been employed over the years to refine speaker verification processes, but it is the integration of deep learning methodologies that has ushered in a substantial and transformative advancement. The main objective of proposed work is to establish an automated and efficient predictive system, utilizing deep learning techniques, to refine the vectors essential for speaker verification. To achieve this goal, extensive experiments were conducted using the VoxCeleb-1 database, evaluating the performance of different deep learning techniques. These experiments involved testing various activation functions, including linear, sigmoid, tanh, and relu, as well as different optimizers such as SGD, Adam, Adagrad, and Adadelta. Our experiments revealed that the deep learning model achieved optimal performance when configured with the ReLU activation function, Stochastic Gradient Descent (SGD) optimizer, and employing a two-layer architecture.

REFERENCES

[1] Larcher, A., Bonastre, J. F., Fauve, B., Lee, K. A., Lévy, C., Li, H., ... & Parfait, J. Y. (2013, August). "ALIZE 3.0-Open source toolkit for state-of-the-art speaker recognition." In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (pp. 2768-2772).

[2] Nagrani, A., Chung, J., & Zisserman, A. (2017). "VoxCeleb: a large-scale speaker identification dataset." *Interspeech* 2017.

[3] *"British English definition of voice recognition"*. Macmillan Publishers Limited. *Retrieved February* 21, 2012.

[4] Campbell, W. M., Sturim, D. E., & Reynolds, D. A. (2006). "Support vector machines using GMM supervectors for speaker verification." *IEEE signal processing letters,* **13**(5), 308-311.

[5] Woodard, J. P. (1992). "Modeling and classification of natural sounds by product code hidden markov models." *IEEE Transactions on signal processing,* **40**(7), 1833-1835.

[6] Available: https://academiccommons.columbia.edu/doi/10.7916/D8F19821/download. [online]

[7] Stowell, D., Giannoulis, D., Benetos, E., Lagrange, M., & Plumbley, M. D. (2015). "Detection and classification of acoustic scenes and events." *IEEE Transactions on Multimedia,* **17**(10), 1733-1746.

[8] Mesaros, A., Heittola, T., Benetos, E., Foster, P., Lagrange, M., Virtanen, T., & Plumbley, M. D. (2017). "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge." *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* **26**(2), 379-393.

[9] Mesaros, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Vincent, E., ... & Virtanen, T. (2017, November). "DCASE 2017 challenge setup: Tasks, datasets and baseline system." *In DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events.*

[10] Available: http://dcase.community/challenge2019. [online]

[11] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). "ImageNet: A large-scale hierarchical image database." *In 2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248-255). IEEE Computer Society.

[12] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, pp. 4171-4186, 2018.

[13] Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Wilson, K. (2017, March). "CNN architectures for large-scale audio classification." *In ICASSP 2017-2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 131-135). IEEE.

[14] Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). "Transfer learning for music classification and regression tasks." *In 18th International Society for Music Information Retrieval Conference, ISMIR 201* (pp. 141-149). International Society for Music Information Retrieval.

[15] J. Pons, O. Nieto, M. Prockup, E. Schmidt, A. Ehmann and X. Serra, End-to-end learning for music audio tagging at scale, In *Proc. Conf. Int. Soc. Music Inform. Retrieval,* pp. 637-644, 2017.

[16] Kong, Q., Xu, Y., Wang, W., & Plumbley, M. D. (2018, April). "Audio set classification with attention model: A probabilistic perspective." *In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 316-320). IEEE.

[17] C. Yu, K. S. Barsim, Q. Kong and B. Yang, Multi-level attention model for weakly supervised audio classification, In *Proc. Detection Classification Acoust. Scenes Events,* pp. 188-192, 2018.

[18] Chou, S. Y., Jang, J. S. R., & Yang, Y. H. (2018, July). "Learning to recognize transient sound events using attentional supervision." In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (pp. 3336-3342).

[19] Wang, Y., Li, J., & Metze, F. (2019, May). "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling." In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 31-35). IEEE.

[20] Kong, Q., Yu, C., Xu, Y., Iqbal, T., Wang, W., & Plumbley, M. D. (2019). "Weakly labelled audioset tagging with attention neural networks." *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* **27**(11), 1791-1802.

[21] Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2011). "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition." *IEEE Transactions on audio, speech, and language processing,* **20**(1), 30-42.

[22] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). "Front-end factor analysis for speaker verification." *IEEE Transactions on Audio, Speech, and Language Processing,* **19**(4), 788-798.

[23] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). "Front-end factor analysis for speaker verification." *IEEE Transactions on Audio, Speech, and Language Processing,*

19(4), 788-798.

[24] Ghahabi, O., & Hernando, J. (2015, April). "Restricted Boltzmann machine supervectors for speaker recognition." *In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4804-4808). IEEE.

[25] Naz, S., Raja, M. A. Z., Kausar, A., Zameer, A., Mehmood, A., & Shoaib, M. (2022). "Dynamics of nonlinear cantilever piezoelectric–mechanical system: An intelligent computational approach." *Mathematics and Computers in Simulation (MATCOM),* **196**(C), 88-113.

[26] Abdullah, Z., Naz, S., Raja, M. A. Z., & Zameer, A. (2021). "Design of wideband tonpilz transducers for underwater SONAR applications with finite element model." *Applied Acoustics*, **183**, 108293.

[27] Naz, S., Raja, M. A. Z., Mehmood, A., Zameer, A., & Shoaib, M. (2021). "Neuro-intelligent networks for Bouc–Wen hysteresis model for piezostage actuator." *The European Physical Journal Plus,* **136**(4), 1-20.

[28] Kausar, A., Chang, C. Y., Raja, M. A. Z., Zameer, A., & Shoaib, M. (2024). "Novel design of recurrent neural network for the dynamical of nonlinear piezoelectric cantilever mass–beam model. " *The European Physical Journal Plus,* **139**(1), 16.

[29] Ghahabi, O., & Hernando, J. (2014, November). "Global impostor selection for DBNs in multi-session i-vector speaker recognition." *In Advances in Speech and Language Technologies for Iberian Languages: Second International Conference, IberSPEECH 2014, Las Palmas de Gran Canaria, Spain, November 19-21, 2014. Proceedings* (pp. 89-98). Cham: Springer International Publishing.

[30] Ghahabi, O., & Hernando, J. (2017). "Deep learning backend for single and multisession i-vector speaker recognition." *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* **25**(4), 807-817.

[31] Serrano, J. L. (2012). *Speaker diarization and tracking in multiple-sensor environments* (Doctoral dissertation, Universitat Politècnica de Catalunya (UPC)).

[32] Kenny, P., Stafylakis, T., Ouellet, P., Gupta, V., & Alam, M. J. (2014, June). "Deep Neural Networks for extracting Baum-Welch statistics for Speaker Recognition." *In Odyssey* (Vol. 2014, pp. 293-298).

[33] Khan, U., India Massana, M. À., & Hernando Pericás, F. J. (2019). "Auto-encoding nearest neighbor i-vectors for speaker verification." *In Interspeech 2019: the 20th Annual Conference of the International Speech Communication Association: 15-19 September 2019: Graz, Austria* (pp. 4060-4064). International Speech Communication Association (ISCA).

[34] Stafylakis, T., Kenny, P., Senoussaoui, M., & Dumouchel, P. (2012, June). "Preliminary investigation of Boltzmann machine classifiers for speaker recognition." *In Odysse*y (pp. 109-116).

[35] Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). "Familiar voice recognition: Patterns and parameters part I: Recognition of backward voices." *Journal of phonetics,* **13**(1), 19-38.

[36] "Voice Recognition To Ease Travel Bookings: Business Travel News". BusinessTravelNews.com. March 3, 1997. The earliest applications of speech recognition software were dictation. Four months ago, IBM introduced a "continual dictation product" designed to debuted at the National Business Travel Association trade show in 1994.

[37] Wang, S., Qian, Y., & Yu, K. (2017, August). "What does the speaker embedding encode?" *In Interspeech* (pp. 1497-1501).