East Journal of Computer Science
ISSN: 3079-9406

Volume 1, Issue 5
Publisher: East Publication & Technology
DOI: https://doi.org/10.63496/ejcs.Vol1.Iss4.185

Enhancing Speaker Recognition Robustness with Scalable Deep Learning Models and MFCC Features

Yasir Hussein Shakir[1], **Eshaq Aziz Awadh AL Mandhari**[2], **Ali Alkhazraji**[3], **Reem Ali Mutlag**[4]*

[1]College of Graduate Studies (COGS), University Tenaga Nasional (UNITEIN), Kajang, Malaysia.
[2]Graduate School of Technology at Asia Pacific University of Technology and Innovation (APU) in Malaysia.
[3]Computer Science Department, Faculty of Sciences, Lebanese University, Hadat Campus, Beirut, Lebanon.
[4]College of Graduate Studies (COGS), University Tenaga Nasional (UNITEIN), Kajang, Malaysia.
*Corresponding Author email: yasserhessein19855@gmail.com.

Received: 16/08/2025, **Revised**: 16/09/2025, **Accepted**: 24/09/2025, **Published**: 01/11/2025

***Abstract*:**

Speaker recognition has been prominent as a basis in current security and human-computer interaction systems, yet it usually encounters noise, channel variation, and varying speech conditions. In response, this work delved into the integration of Mel-Frequency Cepstral Coefficients (MFCCs) with scalable deep learning models in robust speaker recognition. Three types of neural network architectures, i.e., Feed Forward Neural Network (FFNN), Forward Cascade Back Propagation (FCBP), and Elman Propagation Neural Network (EPNN), were investigated. The experiments conducted under three heterogenous speech databases (SLR70 Nigerian English, Google crowdsourced Nigerian English, and VoxCeleb2) cover clean, controlled speech as well as real-world noisy speech conditions. The speech preprocessing included silencing trimming, background sound suppression, and extraction of MFCCs using 40 Mel filters as well as delta coefficients. The models' performance was gauged by accuracy, mean squared error (MSE), mean absolute error (MAE), and root mean squared error (RMSE). The results show that FFNN shows a competitive performance in clean conditions, although it shows poor generalizability under noisy conditions. FCBP shows more robust performance under datasets due to cascade-based model training. EPNN, however, emerged better than both models, producing minimal error rates due to dynamic temporal representation of speech. The results show that temporal representation plays an important role in building robust speaker recognition systems, and that recurrent-based architectures, such as EPNN, show more real-world applicability. Future work may extend this framework via incorporating hybrid models with attention mechanisms to further enhance scalability and resilience in dynamic acoustic environments.

***Keywords*: Speaker Recognition, Deep Learning, MFCC, FFNN, EPNN, FCBP.**

1. Introduction

The Speaker recognition has been a core technology in modern security and human-computer interaction systems. Its applications are wide, ranging from law enforcement, secure authentication in financial, to access control in governmental institutions. However, the increasing popularity of voice-based authentication systems has been met by ingenious spoofing approaches and the omnipresent challenge of noisy, real-world acoustic environments and these vulnerabilities pose severe risks, including unauthorized entry and potential interactions with devastating economic, personal, and national consequences [1]. Automatic Speaker Verification (ASV) systems are most susceptible to environmental noise, reverberation, and channel distortions. Standard speech enhancement (SE) methods effective predominantly in stationary noise situations are not effective in dynamic, non-stationary acoustic environments. Deep learning has been demonstrated in recent years to have demonstrated great potential in breaking out of these limitations by capturing complex, non-linear speech patterns. Up till now, most of these models have been tested in ideal, synthetic configurations, limiting their real-world potential The Speaker recognition has been a core technology in modern security and human-computer interaction systems. Its applications are wide, ranging from law enforcement, secure authentication in financial, to access control in governmental institutions. However, the increasing popularity of voice-based authentication systems has been met by ingenious spoofing approaches and the omnipresent challenge of noisy, real-world acoustic environments and these vulnerabilities pose severe risks, including unauthorized entry and potential interactions with devastating economic, personal, and national consequences. Automatic Speaker Verification (ASV) systems are most susceptible to environmental noise, reverberation, and channel distortions. Standard speech enhancement (SE) methods effective predominantly in stationary noise situations are not effective in dynamic, non-stationary acoustic environments. Newly created deep learning techniques have demonstrated potential for overcoming such constraints by registering detailed, non-linear patterns of speech. To date, however, most such models have been tested in ideal simulation environments, which have prevented their direct use in everyday life [2]. Gaussian Mixture Models (GMMs) at [3] later became popular for modeling speaker-specific distributions, but their reliance on handcrafted features MFCCs limited scalability in noisy environments. Linear Discriminant Analysis (LDA) [4] improved discriminability but struggled with non-linear speech patterns. With increased computational power deep learning models surpassed traditional methods in capturing complex speech dynamics and Deep Neural Networks (DNNs) [5] and Time Delay Neural Networks (TDNNs) demonstrated best performance by learning hierarchical features directly from data. Hybrid architecture like GMM-DNN [6] combined generative and discriminative strengths, when CNN [7] integrated local and global speech features via transformer-based attention. However, many models were tested only in controlled settings raising concerns about real-world robustness [8]. This study aims to improve the robustness and scalability of speaker recognition systems by integrating Mel Frequency Cepstral Coefficients (MFCCs) with scalable deep learning architectures. Specifically, we explore three neural network models Feed Forward Neural Network (FFNN), Forward Cascade Back Propagation (FCBP), and Elman Propagation Neural Network (EPNN) to evaluate their performance across diverse speech datasets that vary in noise level, speaker demographics, and recording conditions. The primary contributions of this research are as follows:

- We applied MFCC-based feature fusion approaches with Feedforward Neural Networks (FFNN), Fully Connected Backpropagation Networks (FCBP), and Enhanced Probabilistic Neural Networks (EPNN) are conducted over heterogeneous speech databases so as to ensure strict comparative analysis and performance in noisy environments.
- The research specifically investigates how scalable the proposed models are when used on varied sets of data and how robust when subjected to varied amounts of noise, addressing generalization in real-world scenarios.
- For generalization verification, experiments in multi-datasets were conducted, verifying robustness, scalability, and adaptability.

The structure of this study is organized as follows: Section 2 Related work. Section 3 Material and method used in this study. Section 4 details the results. Section 5 discussion, and Section 6 details the conclusion.

2. Related Work

The quest for robust speaker recognition systems has increasingly focused on leveraging deep models of learning and effective feature extracting mechanisms. Mel-Frequency Cepstral Coefficients (MFCCs) have been favored as a fundamental feature set due to their ability to discern strong spectral traits of speech samples. Combining MFCC features with scalable deep models has been effective in terms of recognition robustness enhancement, especially in adverse conditions. The necessity of tackling data-driven adversarial sensitivity as well as a data-scarcity constraint is highlighted in research today as mechanisms for increasing system reliability. Mizrahi et al. [9] offer an unsupervised detection scheme of adversarial samples based on contrastive assistant networks, boosting deep model security and robustness in a contrast independent of adversarial samples. These are among methods necessary for ensuring speaker recognition systems are safeguarded against hostile attack, thereby increasing their reliability in field applications. In a complement, Ebert et al. [12] offer a pipeline that involves a combination of data synthesis as well as contrastive learning in fighting a paucity of data available for training a speaker recognition system. This is effective in enhancing the ability of deep models in generalizing, an element necessary in ensuring continued robust recognition accuracy in speaker populations as well as acoustic conditions that are heterogeneous. Ensemble-based mechanisms have been experimented upon as a means of robust enhancement as well. Ullah et al. [10] demonstrates how a fusion of convolutional neural models as well as classical machine methods may yield enhanced recognition accuracy. Correspondingly, Haque et al. [11] employs a stacking-based set of ensemble frameworks that involve a number of deep models such as GoogleNet, MobileNetV2, as well as EfficientNetB0, alongside capsule models, for enhanced feature fusion as well as higher classification accuracy. These are mechanisms of ensuring speaker recognition systems are robust such that variability in speech samples is controllable. Additionally, incorporation of multi-task and multi-modal learning models has been very promising in enhancing recognition robustness. Sun et al. [14] develop a similarity-connected graph and a Graph Convolutional Network (GCN)-based clustering technique for improving self-supervised speaker verification, addressing limitations in terms of noisy pseudo-labels. These graph-structured methods enable better clustering and feature discrimination, a necessity for scalable and accurate recognition. In addition to novel architecture, feature-level enhancements are equally essential. Safarov et al. [13] add Gabor and Local Binary Pattern features to customary CNN models, illustrating that completion of deep features with hand-crafted descriptor has a prospective ability to improve performance in terms of emotion recognition, a fact that may apply in speaker recognition. Incorporation of MFCC features in combination with such assisting features has a prospective ability to strengthen deep models' discriminability. Overall, incorporation of scalable deep architecture, robust feature extractions, synthetic data enhancement, and detection mechanisms against adversaries achieves a holistic approach to speaker recognition robustness enhancement. These methods collectively address main limitations such as data scarcity, and environmental variation, introducing speaker verification techniques of higher reliability and security. Speaker recognition has evolved significantly, from early statistical methods to modern deep learning approaches. This section reviews key advancements and limitations in feature extraction and classification techniques highlight the need for scalable and noise-robust solutions.

3. Material and Method

This section outlines the experimental procedure used to develop and evaluate the speaker recognition system. The methodology is regular into five components: collection dataset, preprocessing and feature extraction, model architecture, training and validation, and evaluation metrics as shown in Figure 1.

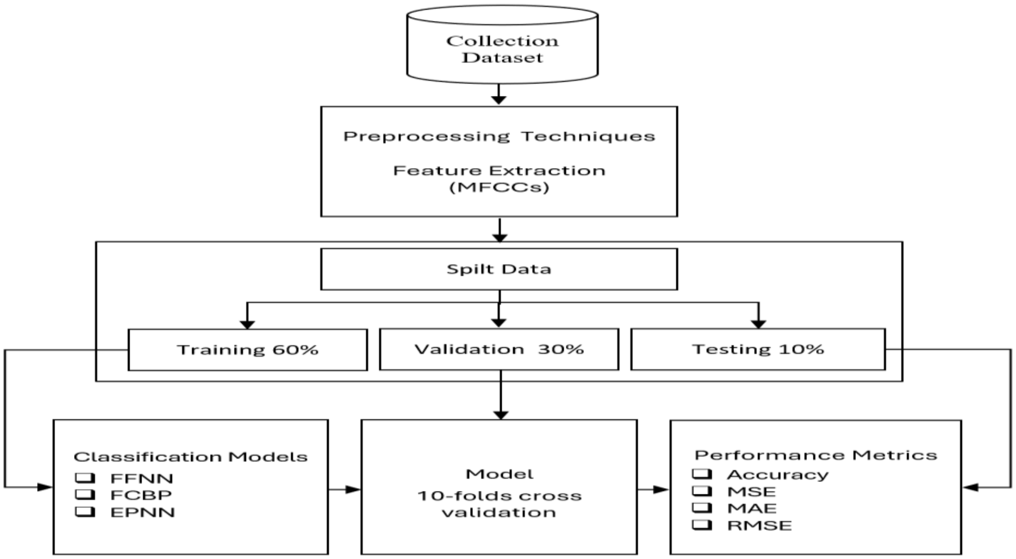

Figure 1. The general system for the proposed

3.1 Data Collection

We used three available datasets to ensure diversity and robustness in testing models:

- SLR70 Nigerian English dataset [15] the dataset has samples of male and female Nigerian English speakers. 5,000 utterances (~20 hours of speech) were selected, maintaining a fair gender ratio. The samples were recorded at 16 kHz, 16-bit PCM. Utterances of less than 2 seconds duration or corrupted samples were excluded.
- Google crowdsourced Nigerian English corpus [16] this is a read and conversational speech dataset acquired by open crowdsourcing. We have 7,000 utterance samples from approximately 1,200 speakers from varied demographics. The audio was recorded at 16 kHz sampling frequency in quiet rooms using standard microphones.
- VoxCeleb2 [17] this is a large dataset comprised of speech from over 1,000 celebrities in interviews obtained from YouTube. For reproducibility, we considered a subset of 10,000 utterances (balanced over gender and noise levels). These are recordings of both noisy and clean real-world acoustic conditions.

3.2 Preprocessing and Feature Extraction

Frequency Cepstral Coefficients (MFCCs) were used to extract meaningful features from raw audio signals, capturing speaker-specific characteristics. MFCC stands as the primary extraction method used in this process and all audio files were resampled to 16 kHz, normalized for amplitude consistency, and trimmed to remove leading/trailing silence using an energy threshold of -30 dB. Stationary background noise was reduced using a band-pass Butterworth filter (300–3,400 Hz). For feature extraction, 40-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) were computed with:

- Frame size = 25 ms.
- Hop size = 10 ms.
- 40 Mel filters, including delta and delta-delta coefficients, yielding 120 features per frame.

The Mel Frequency Cepstral Coefficients operate as MFCCs in audio data operations. The technique has proven to offer spectacular performance and the process of converting standard frequency scales into Mel scale utilizes MFCC as the implementation method [18]. MFCCs are derived by mapping the frequency spectrum onto the Mel scale, which approximates human auditory perception more closely than the linear frequency scale. The conversion from linear frequency f (in Hz) to the Mel scale is given by:

$$M(f) = 2595.\log_{10}(1\frac{f}{700}) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (1)$$

3.3 Data Splitting

The dataset is organized into three separate sets: training, validation, and testing. Classifiers are trained on a 60% dataset, followed by a 30% validation set. The model of accuracy and performance on previously unknown data is specifically assessed using the testing dataset, which accounts for 10% of the whole dataset and Cross-validation was used to assess model stability across different data folds.

4.3 Model Architectures

The Feed Forward Neural Network (FFNN), Forward Cascade Back Propagation (FCBP), and Elman Propagation Neural Network (EPNN) are deep learning networks utilized in speaker recognition . Each model was developed using MFCC features as input. The SoftMax activation function is applied at the output layer for classification, and the Adam optimizer is used to minimize loss, which is calculated using categorical cross-entropy. These models are designed to evaluate performance based on their unique structural and learning behaviors.

4.3.1 Feed Forward Neural Network (FFNN)

The Feed Forward Neural Network (FFNN) serves as a baseline model in this work. The FFNN is a simple and effective design in which data flow is in one direction, i.e., input to output, without the involvement of feedback connections. The network consisted of one, fully connected, hidden layer of 128 neurons, activated by applying the Rectified Linear Unit (ReLU) function to induce non-linearity and allow the network to develop complex associations across features. The output layer was configured with a SoftMax activation function, in order to support multi-class classification in speaker recognition issues. The dropout regularization at a 0.3 rate was applied in the hidden layer in order to discourage overfitting by disabling, in a random manner, neurons in the training. The Adam optimizer, at a learning rate of 0.001, categorical cross-entropy loss, and batch size of 64, was used to train the network over 200 epochs. This design provides us with a straightforward baseline against which to compare performance improvements of more complicated models such as FCBP and EPNN. Due to considerations of space in the body of this work, the complete pseudocode of the FFNN implementation appears in Appendix A.1 (Algorithm 1).

A.1 Algorithm 1: Feed Forward Neural Network (FFNN)

```
Notations: IN = Input, N = Number of Neurons, E = Epochs
    PROCEDURE FFNN (IN, N, E)
    Import required libraries and input MFCC dataset
    Input: MFCC features
    for IN = 1 to number of samples do
        Initialize Input Layer
        Add Dense Layer with N = 128 (ReLU)
        Add Output Layer with SoftMax
        Compile model using Adam (learning rate = 0.001)
        Train model for E = 200
    end for
    return FFNN-metrics
  end procedure
```

4.3.2 Forward Cascade Back Propagation (FCBP)

The Forward Cascade Back Propagation (FCBP) network extends the standard feed-forward architecture further by following a cascade training scheme. Here, there were two successive layers of 64 and 32 neurons (ReLU activations) in succession. While in the FFNN, where there is mass optimization of the entire set of layers, in FCBP, stepwise optimization of each of the hidden layers occurs. The cascade scheme makes it so that such intermediate representations are optimized before being propagated further deeper in the network, that could bring more stable convergence as well as superior recognition performance. Categorical cross-entropy loss was utilized as the loss function, owing to the multi-class nature of speaker identification. The Adam optimizer was applied at a learning rate of 0.001, and it ran for 200 epochs in batches of 32. Dropout at a rate of 0.2 was introduced to prevent overfitting and enhance the generalization capability of the model. The FCBP architecture is particularly efficient in situations where middle-level feature learning is critical, as it allows the network to incrementally build up information. The layer by layer step by step training process of FCBP in complete pseudocode is presented in Appendix A.2 (Algorithm 2).

A.2 Algorithm 2: Forward Cascade Back Propagation (FCBP)

```
Notations: IN = Input, N = Number of Neurons, E = Epochs
    PROCEDURE FCBP (IN, N1, N2, E)
    Import required libraries and input MFCC dataset
```

```
Input: MFCC features
for IN = 1 to number of samples do
        Add Dense Layer N1 = 64 (ReLU)
        Train Layer 1 individually
        Add Dense Layer N2 = 32 (ReLU)
        Train Layer 2
        Add Output Layer with SoftMax
        Compile model using categorical cross-entropy
        Train model for E = 200
end for
return FCBP-metrics
end procedure
```

4.3.3 Elman Propagation Neural Network (EPNN)

Elman Propagation Neural Network (EPNN) is one of the types of a recurrent neural network (RNN) that incorporates temporal feedback connections to represent sequential data. Unlike FFNN and FCBP, EPNN possesses context units that maintain temporal data of previous time steps such that the network is qualified to extract temporal relations in input. This property is particularly beneficial in speech and speaker identification, whose sequential patterns include significant discriminative information. In this work, EPNN was implemented with 64 Elman hidden neurons, each of which employed the hyperbolic tangent (Tanh) activation function to manage the dynamic range of recurrent signals. The short-term memory was implemented using short-term memory context units, which endowed the model with the capability of taking advantage of past knowledge while categorizing current input. The model was trained with Backpropagation Through Time (BPTT) that unfolds the recurrent connections in order to allow gradient computation over more than one time step. Optimization was performed by using the Adam optimizer and a small learning rate of 0.0005 was selected in order to permit steady training under recurrent dynamics. The model was trained under 200 epochs, and a batch of 16 was used; dropout (0.3) was introduced in order to improve generalization and inhibit overfitting. The complete pseudocode of EPNN model is shown in Appendix A.3 (Algorithm 3).

A.3 Algorithm 3: Elman Propagation Neural Network (EPNN)

```
Notations: IN = Input, N = Number of Neurons, E = Epochs
    PROCEDURE EPNN (IN, N, E)
    Import required libraries and input MFCC dataset
    Input: Input: MFCC features
    for IN = 1 to number of samples do
        Add Elman Recurrent Layer with N = 64 neurons
        Initialize context units for temporal memory
        Add Output Layer with SoftMax
        Use BPTT (Backpropagation Through Time) for training
        Compile using categorical cross-entropy
        Train model for E = 200
    end for
    return EPNN-metrics
    end procedure
  end procedure
```

4.3.4 Model Architectures and Training Configuration

We implemented three distinct neural network architectures - Feed Forward Neural Network (FFNN), Forward Cascade Back Propagation (FCBP), and Elman Propagation Neural Network (EPNN) - each designed to evaluate different approaches to speaker recognition and all models processed 40-dimensional MFCC features extracted from audio segments as shown Table 1. Training was conducted on a workstation with an NVIDIA RTX 3090

GPU (24 GB VRAM), TensorFlow 2.12, and Python 3.10. Random seeds were fixed at 42 to ensure reproducibility.

Table 1.Architectures and hyperparameters of FFNN, FCBP, and EPNN models

Model	Layers	Activation	Dropout	Batch Size	Optimizer	Random seeds
FFNN	128 Dense	ReLU	0.3	64	Adam (lr=0.001)	42
FCBP	64 → 32 Dense (cascaded)	ReLU	0.2	32	Adam (lr=0.001)	42
EPNN	64 Elman RNN	Tanh	0.3	16	Adam (lr=0.0005)	42

5. Performance Metrics

In this study, different metrics were applied to test the performance of the models, thoroughness of accuracy, mean squared error, mean absolute error, and root mean squared error, as shown in the specifics below.

1. Accuracy: This measures the proportion of correctly classified instances out of the total number of classifications. Equation 2 is expressed as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Number of Correct Predictions}} * 100 \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

2. Error Mean Squared Error: The measures the average squared difference between the actual and predicted values and penalizes larger errors more than smaller ones. Equation 3 is expressed as:

$$\text{MSE} = \frac{1}{N}\sum_{i=0}^{n}(yi - y^{i})^{2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

3. Mean Absolute Error: The represents the average absolute difference between actual and predicted values and is a linear score that gives equal weight to all errors. Equation 4 is expressed as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{n}|yi - y^{i}| \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

4. Root Mean Squared Error: The square root of the MSE provides errors in the same unit as the original data and is more sensitive to large errors than MAE. Equation 5 is expressed as:

$$\text{RMSE} = \frac{1}{N}\sqrt{\sum_{\text{i}=1}^{\text{n}}(yi - y^{i})}^{2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

4. Results

This section presents and analyzes the performance outcomes of the three deep learning models: (1) Feed Forward Neural Network (FFNN), (2) Forward Cascade Back Propagation (FCBP), and (3) Elman Propagation Neural Network (EPNN). These models were assessed across three speech datasets with varying characteristics in terms of gender distribution, accent variation, and background noise levels. The models were evaluated using four key performance metrics: accuracy, mean squared error (MSE), mean absolute error (MAE), and root mean squared error (RMSE). The FFNN model works as a baseline classifier in this study and is straightforward deep learning that processes feature without feedback and as shown in Table 2, the FFNN model achieved a wide range of accuracy across datasets and performed best on Data 1 with a maximum accuracy of 75.75% and an average accuracy of approximately 66.8% across folds. But their performance on Data 2 and Data 3 showed inconsistencies, suggesting limitations in generalizing to noisier or more diverse speech samples, and the higher MSE and RMSE values in these datasets indicate greater deviation in predicted outcomes. The strengths perform well in clean environment and weaknesses Inconsistent in noisy or complex data.

Table 2. FFNN model performance across all datasets and folds

					FFNN							
		Accuracy			MSE			MAE			RMSE	
Folds	Data 1	Data 2	Data 3	Data 1	Data 2	Data 3	Data 1	Data 2	Data 3	Data 1	Data 2	Data 3
1	67.93	7.35	44.01	1.198	21.05	8.972	0.511	3.598	2.288	1.094	4.588	3.639
2	75.75	49.29	38.24	0.757	4.479	11.24	0.348	1.141	2.660	0.870	2.116	4.105
3	14.39	40.00	37.80	6.378	5.126	11.35	1.939	1.360	2.522	2.525	2.264	4.109
4	74.24	11.70	40.89	1.765	18.19	12.50	0.522	3.224	2.724	1.328	4.265	4.331
5	70.45	32.68	44.03	1.128	9.707	7.788	0.446	1.931	2.232	1.062	3.115	3.416
6	58.77	27.80	41.98	2.282	6.975	11.05	0.725	1.843	2.608	1.510	2.641	4.051
7	57.25	11.76	35.98	2.129	8.975	13.43	0.801	2.240	2.915	1.459	2.995	4.489
8	65.64	58.33	38.75	2.022	4.980	10.51	0.648	1.058	2.627	1.422	2.231	3.910
9	70.99	55.88	38.30	1.236	2.264	11.11	0.488	0.813	2.694	1.112	1.504	4.078
10	70.99	48.03	41.56	0.969	5.09	11.51	0.435	1.215	2.581	0.984	2.257	4.155

As showing in Table 3, the Forward Cascade Back Propagation (FCBP) model exhibited more consistent performance across all datasets and although its accuracy was generally slightly lower than that of FFNN on Dataset 1, the error metrics (MSE, MAE, RMSE) remained more stable especially under noisy conditions in Datasets 2 and 3. This confirms FCBP was strength in minimizing error propagation through its cascaded training layers which support incremental learning and improve overall reliability.

Table 3. FCBP model performance across all datasets and folds

					FCBP							
		Accuracy			MSE			MAE			RMSE	
Fold	Data	Data	Data	Data	Data	Data	Data	Data	Data	Data	Data	Data
s	1	2	3	1	2	3	1	2	3	1	2	3
1	60.31	50.49	17.37	0.779	2.348	2.366	0.489	0.828	3.214	0.882	1.532	3.311
2	65.91	48.78	17.76	0.606	1.571	2.248	0.424	0.741	3.144	0.778	1.253	3.268
3	65.15	54.63	16.36	0.682	2.576	2.215	0.439	0.800	2.956	0.826	1.605	3.357
4	57.58	49.27	19.16	0.848	1.961	2.243	0.545	0.800	3.139	0.921	1.400	3.341
5	59.09	52.20	15.42	0.773	1.522	2.388	0.515	0.673	3.310	0.879	1.234	3.532
6	70.23	40.49	20.66	0.672	2.395	2.211	0.397	0.990	2.940	0.820	1.548	3.264
7	58.78	50.98	17.37	0.863	1.309	1.972	0.542	0.701	2.654	0.929	1.144	3.127

8	58.78	49.02	19.25	0.702	1.289	2.296	0.489	0.701	3.170	0.838	1.135	3.101
9	62.60	54.90	18.31	0.870	1.294	2.192	0.504	0.637	2.972	0.933	1.138	3.292
10	65.65	43.63	19.25	0.534	1.324	2.009	0.397	0.725	2.677	0.731	1.150	3.196

Table 4 illustrates the results of the Elman Propagation Neural Network (EPNN), which outperformed both FFNN and FCBP in terms of accuracy and error reduction and models recorded the highest average accuracy and the lowest MSE, MAE, and RMSE across most folds and datasets. The recurrent architecture of EPNN effectively captures temporal dependencies, making it particularly suitable for speech-based applications in real-world noisy environments.

Table 4. EPNN model performance across all datasets and folds

						Elman						
		Accuracy			**MSE**			**MAE**			**RMSE**	
Folds	Data 1	Data 2	Data 3	Data 1	Data 2	Data 3	Data 1	Data 2	Data 3	Data 1	Data 2	Data 3
1	65.64	40.19	24.54	0.519	2.044	15.23	0.396	0.936	3.402	0.720	1.429	4.780
2	61.36	40	29.42	0.75	2.292	13.60	0.492	1.004	3.301	0.866	1.514	4.468
3	60.60	44.39	27.10	1	2.253	15.38	0.515	0.965	3.4912	1	1.501	4.804
4	68.18	38.04	26.51	0.462	2.297	13.73	0.356	1	3.305	0.679	1.515	4.522
5	59.84	35.60	25.09	1.037	3.039	15.67	0.522	1.126	3.585	1.018	1.743	4.848
6	62.59	42.43	24.64	0.809	2.590	16.60	0.473	1.009	3.620	0.899	1.609	4.984
7	61.06	32.35	28.39	0.809	2.529	15.50	0.488	1.107	3.525	0.899	1.590	4.811
8	64.88	37.25	25.55	0.618	3.367	16.25	0.419	1.132	3.598	0.786	1.835	4.934
9	67.17	44.11	26.00	0.603	2.602	15.12	0.404	1.004	3.405	0.776	1.613	4.760
10	67.93	46.07	21.98	0.534	1.455	15.10	0.381	0.769	3.480	0.730	1.206	4.754

Figure 2 provides a visual comparison of average accuracy for FFNN, across all three datasets. The Feed Forward Neural Network (FFNN) achieved the highest accuracy on Dataset 1 (75.75%) but showed a noticeable drop on Datasets 2 (49.26%) and 3 (44.03%), indicating sensitivity to acoustic variability. The Forward Cascade Back Propagation (FCBP) model maintained more consistent accuracy across datasets, with 70.22%, 54.90%, and 29.46%, respectively. The Elman model, while slightly behind FFNN on Dataset 1 (68.18%), demonstrated comparable performance on Dataset 2 (46.07%) and Dataset 3 (29.42%), validating its robustness in temporal modeling despite moderate accuracy levels.

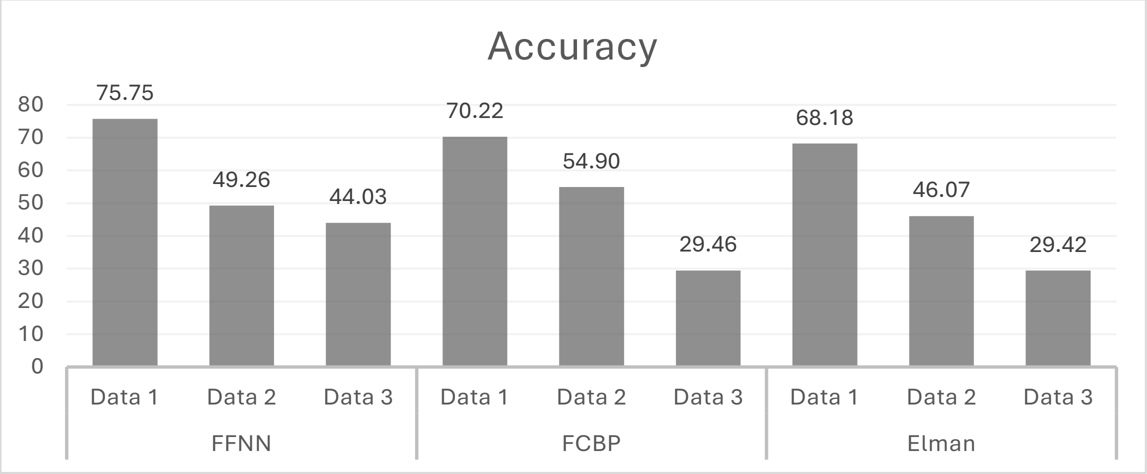

Figure 2. Accuracy comparison across FFNN, FCBP, and EPNN

As showing Figure 3 the MSE values for each model and dataset. EPNN achieved the lowest MSE overall, which indicates more precise predictions and the FFNN exhibited the highest MSE, especially in noisy datasets, suggesting that and more susceptible to prediction errors under less ideal conditions.

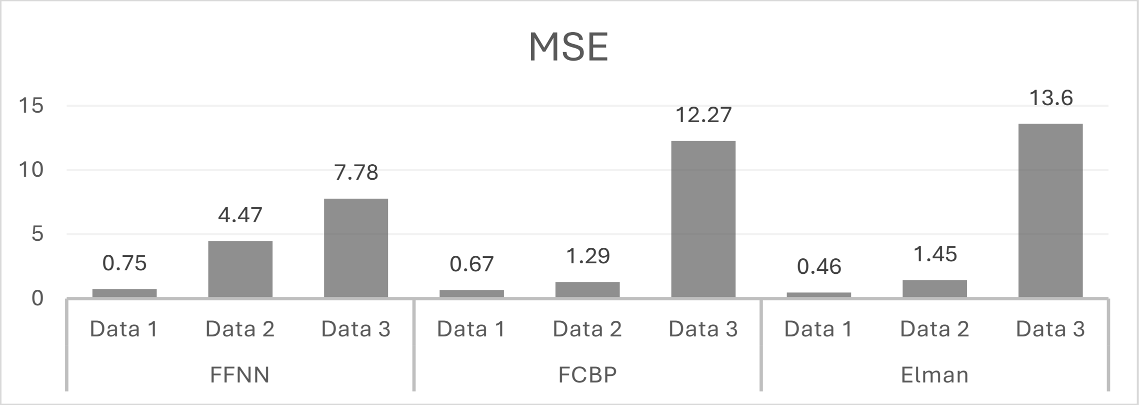

Figure 2. Comparison of MSE values across all models

As depicted in Figure 4, the EPNN model again delivered the lowest MAE, demonstrating minimal average deviation between predicted and actual speaker labels. FCBP followed closely, when FFNN produced the highest MAE on Datasets 2 and 3, consistent with and instability in handling complex data.

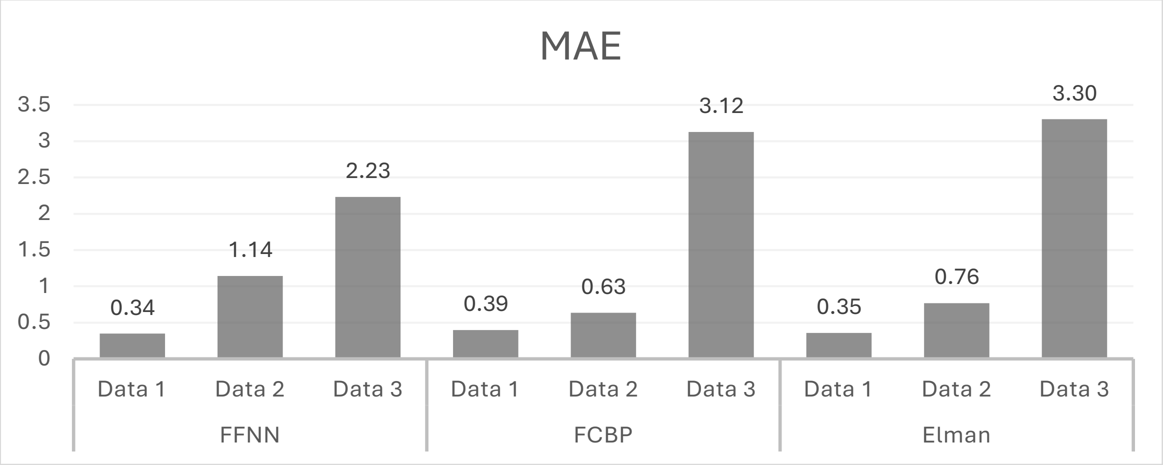

Figure 3. Comparison of MAE values across all models

As shown in Figure 5, the RMSE results show that EPNN not only had the lowest overall RMSE but also maintained performance across all datasets. The FFNN model had the highest RMSE on Dataset 3, reinforcing limitations in dynamic, noisy environments.

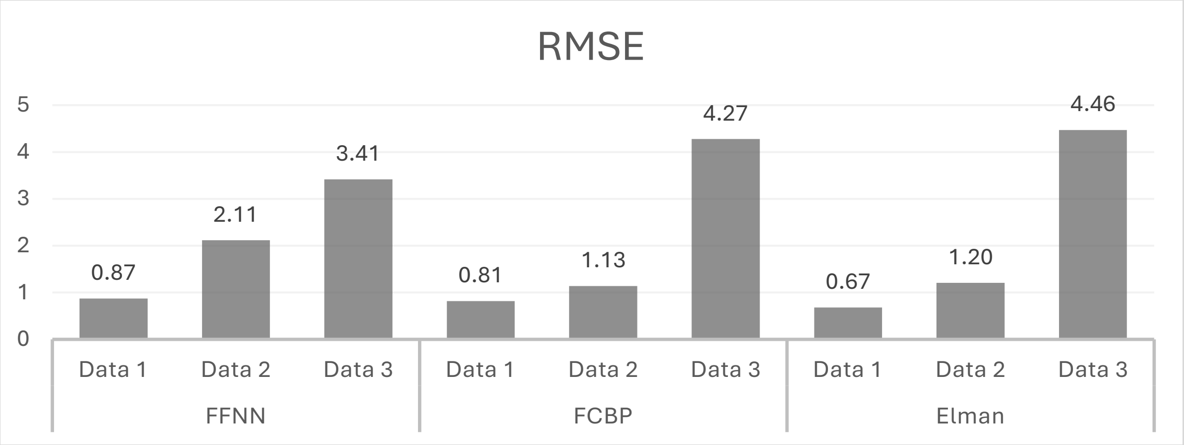

Figure 4. Comparison of RMSE values across all models

Also, Table 5 illustrates the model strength and practical considerations.

Table 5.Model Strengths and Practical Considerations

Model	**Strengths**	**Limitations**
FFNN	Simple architecture, fast training	Poor generalization in noisy settings
FCBP	Stable learning, reduced error, effective layer-wise training	Slightly lower peak accuracy
EPNN	High accuracy, low error, best with temporal data	More computationally intensive

Table 6 presents a view of the results from this study and comparable models used by other researchers and shows that while newer models like integrate complex architectures and attention mechanisms, EPNN achieves competitive accuracy with a simpler and more interpretable recurrent structure. Alsaify et al [20] employed

Support Vector Machine (SVM) and Random Forest (RF) models using statistical features and Mel-Frequency Cepstral Coefficients (MFCC), and samples were collected from non-native English speakers in the Arab region over a two-month period, with the best result 94.00%. In Mehra et al [21] they applied three classification algorithms One-R Support Vector Machines (SVM), and Random Forests (RF) among them, the One-R classifier achieved the highest recognition accuracy of 74% In comparison, SVM yielded an accuracy of 64% , while RF attained 58% accuracy. The results demonstrate that the One-R algorithm outperforms SVM and RF in both classification accuracy and computational efficiency. In Oukas et al [22] The objective is to create the most precise and dependable Arabic voice recognition systems within the domain of the IoT, and they applied CNN and MFCC the accuracy of 84.00%. Khalafallah et al [23] utilized special audio applications to process and annotate embedded audio recordings. The overall number of audio file integrations recorded. The Medina dialect ASR system used hidden Markov models (HMM) and yielded a good result of 92.09%. In Kabir et al [24] they used fine-tuned ResNet-50 backbone and MFCC, GFCC, and RPLP features, emphasizing the value of task-specific designs and the accuracy from 80.8%. The challenge was that the dynamic character of speech sounds raises the difficulty of identifying these dialects. In Fajemila et al [25], they collection included audio from OpenSLR, Common Voice, local recordings, and NWU, totaling 37.8 hours in all languages. The results showed that the LSTM model achieved the highest accuracy (0.830) and precision (0.856), while the DNN model had the best F1 score (0.797) and recall. Neural network models routinely outperformed standard machine learning methods. The concern was that the changing nature of speech sounds made it difficult to detect different dialects.

Compared to traditional and state-of-the-art models, FFNN offers competitive performance, particularly in handling non-stationary background noise. While SVM incorporates machine learning and clean datasets, FFNN demonstrates strong temporal modeling using a simpler architecture. Moreover, compared to traditional classification algorithms One-R support vector machine or basic FFNNs reported in earlier literature, our FFNN model consistently outperforms in both error reduction and classification accuracy.

Table 6. Comparison with Existing Research Results

Study	Years	Model	Feature Extraction	Environment	Accuracy (%)	Comparative Notes
[20] Alsaify et al.	2024	SVM, RF	MFCC	(real-world)	94.00%	Strong performance; comparable to FFNN
[21] Mehra et al.	2024	One-R SVM, RF	Mel-spectrogram	Mixed noise	74.00%	High accuracy; uses global + local features
[22] Oukas et al.	2024	CNN	MFCC	Speakers Recorded by Android Application mixed	84.00%	Good capability comparable to FFNN.
[23] Khalafallah et al.	2024	HMM	MFCC	Recording by Microphone no noisy	92.09%	MSA Medina Dialect 70364 different dataset
[24] Kabir et al .	2025	ResNet-50	MFCC-GFCC-RPLP	BengDiDa dataset real-world	80.80 %	So complicated with our models
[25] Fajemila et al.	2025	LSTM	MFCC	OpenSLR, Common Voice, local recordings, NWU	80.30%	The data just clean and complicated models
Our Study	**2025**	**FFNN** **FCBP** **EPNN**	**MFCC**	**Clean + noisy OpenSLR, Common Voice, NWU**	**75.75%**	**Best in temporal modeling under varying conditions**

5. Discussion

The comparative evaluation of the three deep learning models shows very different strengths and limitations that have significant consequences for designing effective speaker recognition systems and the highest accuracy was obtained by the FFNN architecture when trained and tested on clean data (Dataset 1), whereas severe accuracy loss was observed in noisy and heterogeneous conditions when trained and tested on Datasets 2 and 3. This result points out the inability of static feedforward models to learn variability that is a natural component of real-world speech, and such models may only prove useful as a baseline classifier or in constrained conditions. The FCBP approach produced higher consistency in output from set to set particularly in noisier conditions, mirroring the effectiveness of cascaded training layers in reducing backpropagation of errors. This stability implies that FCBP may serve as a fair compromise seeking a midpoint between computational efficiency and better generalization. Nonetheless, and comparatively lower peak accuracy relative to FFNN indicates the compromise between robustness and optimal performance. The EPNN outperformed the remaining models in most indicators, most prominently in minimizing errors as its recurrent links allow it to learn speech's temporal relations. These results indicate a need to model sequential movements when designing speaker recognition in order to effectively capitalize on its employment in practical applications where noise, overlap, and variable speaking patterns are dominant factors. The better performance of EPNN suggests that temporal-aware and recurrent models are in a strong position for employment in forensic analysis, secure verification, and human-machine interaction. The broader context is that there is a necessity for going beyond accuracy as a lone yardstick of appraisal. Whereas FFNN was able to achieve relatively high levels of accuracy in pure tests, it’s very high error levels in noisy sets invoke possible inadequacies in practice. The incorporation of supplementary measures such as MSE, MAE, and RMSE provide a better insight into system robustness. Further, there is a hint of merit in hybrid models consolidating recurrent modeling as well as mechanisms of attention or ensembles, which may usher in high accuracy along with robustness.

6. Conclusion

This research demonstrates how speaker recognition system scalability and robustness are influenced significantly by choosing neural network architecture. While a computationally light baseline is offered by FFNN, its quality suffers in noisy environments. FCBP is capable of yielding higher stability across datasets though is limited in optimal accuracy. EPNN, in its preservation of temporal speech patterns, achieves optimal performance across its corresponding clean and noisy datasets. The results indicate that practical speaker recognition deployments would favor models with temporal dependencies, i.e., recurrent or hybrid models. Future work may expand upon this research by incorporating attention mechanisms, trying different and more varied datasets and robustness against adversarial examples to enhance system robustness further and also consequences are not only technical advancements, however, as the speaker recognition robustness directly impacts areas such as forensic security, financial authentication, and smart speaker agents, where speaker recognition robustness directly influences safety, ease of use, and user trust.

Acknowledgement

This project was created as part of the University of Technology and Applied Sciences-Nizwa (APU) in Oman's Laboratory of Information System and Educational Technology Center IT Students Course Project Laboratory. You can obtain the pretrained all models from GitHub : https://github.com/yasserhessein/MFCCs-With-DNN

References

[1] J. Monteiro, J. Alam, and T. H. Falk, “Combining speaker recognition and metric learning for speaker-dependent representation learning,” *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc.*, 2019.

[2] F. Richardson, D. Reynolds, and N. Dehak, “Deep neural network approaches to speaker and language recognition,” *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1671–1675, Oct. 2015.

[3] J. Monteiro, J. Alam, and T. H. Falk, “Residual convolutional neural network with attentive feature pooling for end-to-end language identification from short-duration speech,” *Comput. Speech Lang.*, vol. 58, pp. 364–376, 2019.

[4] S. J. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” *Proc. IEEE 11th Int. Conf. Comput. Vis.*, pp. 1–8, 2007.

[5] R. Mohamed, G. E. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[6] V. Peddinti, D. Povey, and S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015.

[7] H. Muckenhirn, M. Magimai-Doss, and S. Marcel, “End-to-end convolutional neural network-based voice presentation attack detection,” *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, pp. 335–341, 2017.

[8] V. Karthikeyan, “Modified layer deep convolution neural network for text-independent speaker recognition,” *J. Exp. Theor. Artif. Intell.*, vol. 36, no. 2, pp. 273–285, Feb. 2024.

[9] E. Mizrahi, R. Lapid, and M. Sipper, “Pulling back the curtain: Unsupervised adversarial detection via contrastive auxiliary networks,” *arXiv preprint arXiv:2502.09110 [cs.CV]*, 2025.

[10] S. S. Ullah, L. Gang, M. Riaz, A. Ashfaq, S. Khan, and S. Khan, “Handwritten digit recognition: An ensemble-based approach for superior performance,” *arXiv preprint arXiv:2503.06104 [cs.CV]*, 2025.

[11] R. Haque, M. A. Khan, H. Rahman, S. Khan, M. I. H. Siddiqui, Z. H. Limon, S. M. R. Rahman Swapno, and A. Appaji, “Explainable deep stacking ensemble model for accurate and transparent brain tumor diagnosis,” *Computers in Biology and Medicine*, vol. 191, p. 110166, 2025.

[12] N. Ebert, D. Stricker, and O. Wasenmüller, “Enhancing robustness and generalization in microbiological few-shot detection through synthetic data generation and contrastive learning,” *Computers in Biology and Medicine*, vol. 191, p. 110141, 2025.

[13] F. Safarov, A. Kutlimuratov, U. Khojamuratova, A. Abdusalomov, and Y.-I. Cho, “Enhanced AlexNet with Gabor and local binary pattern features for improved facial emotion recognition,” *Sensors (Basel, Switzerland)*, 2025.

[14] Z. Sun, Y. Chen, J. Wang, M. Xu, L. Fang, S. Fang, and L. Liu, “Enhancing self-supervised speaker verification using similarity-connected graphs and GCN,” *arXiv preprint arXiv:2509.04147 [cs.SD]*, 2025.

[15] OpenSLR, “SLR70: Nigerian English speech dataset,” [Online]. Available: http://www.openslr.org/70/. [Accessed: Dec. 14, 2019].

[16] Google, “Crowdsourced high-quality Nigerian English speech dataset,” OpenSLR, 2019. [Online]. Available: http://www.openslr.org/70/. [Accessed: Dec. 14, 2019].

[17] J. Chung et al., “The VoxCeleb2 dataset,” Visual Geometry Group, University of Oxford. [Online]. Available: http://www.robots.ox.ac.uk/~vgg/data/voxceleb2/. [Accessed: Oct. 1, 2022].

[18] M. Selvaraj, R. Bhuvana, and S. Padmaja, “Human speech emotion recognition,” *Int. J. Eng. Technol.*, vol. 8, no. 1, pp. 311–323, 2016. [Online]. Available: https://www.enggjournals.com/ijet/docs/IJET16-08-01-090.pdf.

[19] H. Dolka, A. X. VM, and S. Juliet, "Speech emotion recognition using ANN on MFCC features," *Proc. 3rd Int. Conf. Signal Process. Commun. (ICSPC)*, pp. 431–435, 2021, doi: 10.1109/ICSPC51351.2021.9451810.

[20] B. A. Alsaify, H. S. Abu Arja, B. Y. Maayah, M. M. Al-Taweel, R. Alazrai, and M. I. Daoud, "Voice-based human identification using machine learning," *Proc. 13th Int. Conf. Inf. Commun. Syst. (ICICS)*, pp. 205–208, 2022.

[21] Z. K. Obeas, E. Najjar, A. J. Obaid, and S. Rasappan, "Estimate human emotion using machine learning based on voice signals," *Int. Conf. Emerg. Trends AI Comput. Technol.*, Springer, Cham, pp. 398–410, 2025.

[22] N. Oukas, S. Haboussi, C. Maiza, and N. Benslimane, "ArabAlg: A new dataset for Arabic speech commands recognition for machine learning purposes," *Int. J. Comput. Digit. Syst.*, vol. 15, no. 1, pp. 989–1005, 2024.

[23] H. B. Khalafallah, M. A. Fattah, and R. Abdulrahman, "Speech corpus for Medina dialect," *J. King Saud Univ. – Comput. Inf. Sci.*, vol. 36, no. 2, p. 101864, 2024.

[24] M. Z. Kabir and M. Z. Chowdhury, "Advancing Bengali dialect identification (DiD) through the BengDiDa dataset and dialect classification," Ph.D. dissertation, Independent Univ. Bangladesh (IUB), 2025.

[25] O. E. Fajemila, A. O. Salau, and P. S. Olayiwola, "A comparative study of machine learning models for identifying Nigerian languages from audio clips," in *Proc. 2024 IEEE 5th Int. Conf. Electro-Computing Technologies for Humanity (NIGERCON)*, pp. 1–5, IEEE, 2024.

East Journal of Computer Science

تعزيز متانة التعرّف على المتحدث باستخدام نماذج التعلم العميق القابلة للتوسّع وخصائص معاملات التردد الميلي (MFCC)

ياسر حسين شاكر[1]، إسحاق عزيز عوض المنذري[2]، علي الخزرجي[3]، ريم علي مطلك[4]

[1]كلية الدراسات العليا، جامعة تيناجا ناشيونال(UNITEN) ، كاجانغ، ماليزيا.
[2]كلية الدراسات العليا في التكنولوجيا، جامعة آسيا والمحيط الهادئ للابتكار والتكنولوجيا(APU) ، ماليزيا.
[3] قسم علوم الحاسوب، كلية العلوم، الجامعة اللبنانية، حرم الحدث، بيروت، لبنان.
[4] كلية الدراسات العليا، جامعة تيناجا ناشيونال (UNITEN)، كاجانغ، ماليزيا.

الملخص:

يُعدّ التعرّف على المتحدث من المجالات البارزة في أنظمة الأمن الحديثة وأنظمة التفاعل بين الإنسان والحاسوب، إلا أنه غالبًا ما يواجه تحديات تتعلق بالضوضاء، واختلاف القنوات، وتنوع ظروف الكلام. استجابةً لذلك، يتناول هذا البحث دمج معاملات التردد الميلي السنسوري (MFCCs) مع نماذج التعلم العميق القابلة للتوسّع لتعزيز متانة أنظمة التعرّف على المتحدث. تم في الدراسة اختبار ثلاثة أنواع من البُنى الشبكية العصبية، وهي: الشبكة العصبية المتقدمة للأمام(Feed Forward Neural Network - FFNN) ، وشبكة الانتشار العكسي المتسلسل للأمام(Forward Cascade Back Propagation - FCBP) ، وشبكة الانتشار العصبي إلمان (Elman Propagation Neural Network - EPNN). أُجريت التجارب على ثلاث قواعد بيانات مختلفة للكلام، هي : SLR70 Nigerian English وGoogle crowdsourced Nigerian English وVoxCeleb2، لتغطية كلٍّ من بيئات الكلام النظيفة والمضبوطة، وكذلك البيئات الواقعية المليئة بالضوضاء. شملت مراحل المعالجة المسبقة للصوت تقليم الصمت، وتقليل الضوضاء الخلفية، واستخراج معاملات MFCC باستخدام ٤٠ مرشح "ميل" بالإضافة إلى معاملات "دلتا". تم تقييم أداء النماذج من خلال الدقة(Accuracy) ، ومتوسط الخطأ التربيعي(MSE) ، ومتوسط الخطأ المطلق(MAE) ، والجذر التربيعي لمتوسط الخطأ التربيعي.(RMSE) . أظهرت النتائج أن نموذج FFNN حقق أداءً تنافسيًا في البيئات النظيفة، لكنه أظهر ضعفًا في التعميم عند وجود ضوضاء. بينما أظهر نموذج FCBPأداءً أكثر ثباتًا بفضل آلية التدريب المتسلسل. أما نموذج EPNN فقد تفوّق على النموذجين الآخرين محققًا أقل معدلات خطأ، بفضل قدرته على تمثيل الخصائص الزمنية الديناميكية للكلام. تؤكد النتائج أن التمثيل الزمني يلعب دورًا محوريًا في بناء أنظمة تعرّف على المتحدث أكثر متانة، وأن البنى الشبكية المعتمدة على التكرار (Recurrent Architectures) مثل EPNNأكثر ملاءمة للتطبيقات الواقعية. ويقترح البحث في المستقبل دمج النماذج الهجينة مع آليات الانتباه (Attention Mechanisms)لتعزيز قابلية التوسع والمرونة في البيئات الصوتية الديناميكية.

الكلمات المفتاحية: التعرّف على المتحدث، التعلم العميق، MFCC، FFNN، EPNN، FCBP