applied sciences

MDPI

Article

An Ensemble Approach for Speaker Identification from Audio Files in Noisy Environments

Syed Shahab Zarin [1], Ehzaz Mustafa [1,*], Sardar Khaliq uz Zaman [1], Abdallah Namoun [2,*] and Meshari Huwaytim Alanazi [3]

[1] Department of Computer Science, Comsats University Islamabad, Abbottabad Campus, Abbottabad 22060, Pakistan; skhaleeq@cuiatd.edu.pk (S.K.u.Z.)
[2] AI Center, Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah 42351, Saudi Arabia
[3] Computer Science Department, College of Sciences, Northern Border University, Arar 73213, Saudi Arabia; meshari.alanazi@nbu.edu.sa
* Correspondence: aizazmustafa@cuiatd.edu.pk (E.M.); a.namoun@iu.edu.sa (A.N.)

Abstract: Automatic noise-robust speaker identification is essential in various applications, including forensic analysis, e-commerce, smartphones, and security systems. Audio files containing suspect speech often include background noise, as they are typically not recorded in soundproof environments. To this end, we address the challenges of noise robustness and accuracy in speaker identification systems. An ensemble approach is proposed combining two different neural network architectures including an RNN and DNN using softmax. This approach enhances the system's ability to identify speakers even in noisy environments accurately. Using softmax, we combine voice activity detection (VAD) with a multilayer perceptron (MLP). The VAD component aims to remove noisy frames from the recording. The softmax function addresses these residual traces by assigning a higher probability to the speaker's voice compared to the noise. We tested our proposed solution on the Kaggle speaker recognition dataset and compared it to two baseline systems. Experimental results show that our approach outperforms the baseline systems, achieving a 3.6% and 5.8% increase in test accuracy. Additionally, we compared the proposed MLP system with Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) classifiers. The results demonstrate that the MLP with VAD and softmax outperforms the LSTM by 23.2% and the BiLSTM by 6.6% in test accuracy.

Keywords: speaker identification; audio; voice activity detection; deep neural network; recurrent neural network; spectrogram

Citation: Zarin, S.S.; Mustafa, E.; Zaman, S.K.u.; Namoun, A.; Alanazi, M.H. An Ensemble Approach for Speaker Identification from Audio Files in Noisy Environments. *Appl. Sci.* **2024**, *14*, 10426. https://doi.org/10.3390/app142210426

Academic Editors: Grigorios Beligiannis and Georgios A. Tsirogiannis

Received: 19 September 2024
Revised: 30 October 2024
Accepted: 4 November 2024
Published: 13 November 2024

1. Introduction

Speaker identification is a prominent research area within the speech recognition domain due to its applications in authentication, forensics, security, and biometrics. It is a subfield of Automatic Speech Recognition (ASR), an interdisciplinary domain encompassing acoustics, signal processing, machine learning, communication, and information theory. ASR can be viewed from two perspectives: speaker identification and speaker verification. The speaker identification aspect addresses the task of identifying a speaker from a group of known individuals. In contrast, speaker verification involves determining whether the suspected person is indeed the correct individual. Both aspects fall under the discipline of acoustics, which involves the study of mechanical waves such as speech sounds [1–5].

Each approach has its advantages; speaker identification is crucial for biometric security control, while speaker verification is essential for authenticating personalized interfaces tailored to each speaker. In speaker identification, the features of the input signal are compared with the stored features of all known speakers. In contrast, in speaker verification, the input signal features are compared only to the stored features of the target individual's speech [6–10].

However, there are still some significant issues with i-vectors. Moreover, some significant challenges in speaker identification systems still impact their accuracy and overall performance. The first issue is the system's robustness to noise, as noisy environments can significantly degrade identification accuracy. The second issue is the general performance of these systems, particularly in maintaining high accuracy across different conditions [11–16].

To address these challenges, we propose an ensemble approach that combines the strengths of RNNs, DNNs, and softmax. This ensemble technique enhances both noise resilience and accuracy. In contrast to previous works, we evaluate the performance of VAD, DNNs, RNNs, and softmax within the context of a speaker identification system. The accuracy of our system is compared with state-of-the-art techniques that primarily utilize Long Short-Term Memory (LSTM), Gaussian Mixture Models (GMM), and i-vector-based approaches. Similar to other studies in areas such as session compensation [1], unsupervised speaker identification [1–10,17–19], speaker/speech separation [20], text-independent identification [21–24], and natural language parsing [25], we apply our method to a widely recognized dataset and evaluate its effectiveness using well-established performance metrics. The major contributions of our work are as follows:

- We tackle the critical challenge of noise robustness in speaker identification by proposing a novel ensemble of an RNN and DNN with softmax, which significantly improves the system's accuracy in noisy environments.
- Our method is rigorously tested on a well-known dataset, demonstrating its reliability and effectiveness. The system is evaluated against standard performance measures to ensure valid and comparable results with existing approaches.
- We contribute a novel integration of our approach with existing techniques like session compensation and text-independent identification. This comprehensive solution advances the field by improving accuracy across a variety of conditions and speaker identification tasks.

2. Related Work

Several techniques have been introduced to automate the speaker identification process. Chen et al. [10] proposed an i-vector-based system for the speaker identification task. An i-vector-based method implies a low-dimensional representation of speech signals with different durations. In i-vector-based frameworks, session compensation is the first stage, before classification. Session compensation refers to methods adopted to deal with the conflation of things such as channel properties and the choice of spoken words with speaker characteristics [18]. The work in [17] proposes joint optimization of session compensation and the classifier. This work uses sparse coding (SC) for session compensation and softmax plus support vector machine (SVM) classifiers for the classification task. The system has been tested on King-ASR-010, VoxCeleb, and RSR2015 datasets. King-ASR-010 and RSR2015 are Chinese and Mandarin speech corpora that have labeled examples. This work is based on supervised learning, while that in Audeep is unsupervised learning, which is a plus in the case of unavailability of labeled data [19]. However, the work in [17] is compared with classifiers in the i-vector framework only.

The work presented in [19] was also organized in the form of a framework that was evaluated on three public datasets, the CHAINS, LapsBM1.4, and YouTube datasets. This framework uses ranked lists, which encode similarity information defined by the speaker model. As a model, Gaussian Mixture Models (GMMs), vector quantization, and i-vector techniques are used, whereas, for learning, RLSim and ReckNN algorithms are used. Tiwari [19] proposed a methodology for smart devices to recognize speakers based on a very short utterance. The work uses an i-vector and GMM-based approach which was tested on the THUYG-20 dataset and claimed to have an equal error rate (EER) of 3.21 percent. A novel GMM-based speaker identification technique is proposed in [21] which uses two statistical estimations. The novel GMM deals with noise. The system is tested on the NIST 2000 dataset and is claimed to have a 16 percent relative improvement over i-vector-based speaker identification methods.

In [22], the authors provided an exhaustive systematic review for the identification, comparison, and analysis of feature extraction methods and algorithms found in the research literature spanning the period 2011–2016. This review shows that MFCC-based feature extraction methods are applied more as compared to other methods. The importance of paralinguistic information such as gender, age group, language, accent, and identity of the speaker in a speech signal is dealt with in [23]. This work compares the Gaussian Mixture Model–Universal Background Model (GMM–UBM), GMM–Support Vector Machine (GMM–SVM), and i-vector-based approaches. This work has a very interesting finding, i.e., "For speaker recognition, error rate decreases as age increases".

Nayana et al. [23] proposed a methodology for speaker identification based on Gaussian Mixture Models (GMM) and the i-vector method with two features, PNCC (Power-Normalized Cepstral Coefficients) and RASTA PLP (Relative Spectral Perceptual Linear Prediction) coefficients. As is shown by Table 1, the most widely used features are i-vectors, whereas the most widely used classifiers are GMM, PLDA, and deep neural networks. In [26], the authors proved that Convolutional Neural Networks (CNN) outperform GMM and ResNet features and self-attention features outperform i-vectors.

Table 1. Summary of various studies on speaker identification and related features.

Study	Resolved Issues	Features Used	Speaker Identification Predictor	Dataset Used	Performance
Chen et al. [10]	Joint optimization of session compensation and the classifier	i-vectors	Softmax plus support vector machine (SVM) classifiers	King-ASR-010, VoxCeleb, and RSR2015	80% to 90%
Campos and Pedronette [18]	Unsupervised identification	speaker i-vectors	Gaussian Mixture Models (GMMs)	CHAINS, LapsBM1.4, YouTube dataset (collected for [18])	Gain of 56.29%
Tiwari et al. [19]	Speaker recognition for short-duration speech utterances	i-vectors	GMM-based Universal Background Model (GMM-UBM)	THUYG-20	92.368%
Ayadi et al. [27]	Text-independent identification	speaker i-vectors	GMM	NIST 2000	87%
Nayana et al. [23]	Comparison of text-independent speaker identification systems	i-vectors	GMM		85% to 94.7%
Ghahabi and Hernando [15]	Lack of labeled background data	i-vectors	Imposter selection algorithm, deep belief network, and deep neural network	NIST SRE 2006, NIST 2014	The proposed system fills 46% of the performance gap in terms of minDCF
Cumani and Laface [16]	Transformation of i-vectors so that they become more suitable for discriminating speakers using Probabilistic Linear Discriminant Analysis	i-vectors	Probabilistic Linear Discriminant Analysis and Linear Discriminant Analysis	NIST SRE-2010 and SRE-2012	Relative improvement of 7% and 14% of detection cost function
Zeinali et al. [28]	Text-dependent verification	speaker i-vectors	Hidden Markov Models (HMMs)	RSR2015	Reduced equal error rate (EER) by 50% and 67%. Reduced Normalized Detection Cost Function (NDCF) by 61% and 67%
Cumani and Laface [29]	Total variability i-vector treats each training segment belonging to different speakers	e-vectors	PLDA and Pair-Wise Support Vector Machine (PSVM)	NIST SRE 2010 and NIST SRE 2012	300-dimensional e-vectors for PLDA are almost equivalent to 600-dimensional i-vector PLDA

Table 1. *Cont.*

Study	Resolved Issues	Features Used	Speaker Identification Predictor	Dataset Used	Performance
Xu et al. [30]	Extensive computation run-time while calculating i-vectors	i-vectors	Baum–Weltch statistics and Subspace-Orthogonal Prior (SOP)	NIST SRE 2010	SOP approach speeds up i-vector calculation considerably as compared to standard i-vector calculation
Luo et al. [31]	Speech separation	time–frequency representation of the mixture signal	Deep learning	*Wall Street Journal* (WSJ0) Street dataset	Comparable or better performance than other state-of-the-art deep learning methods
Maghsoodi et al. [32]	Text-dependent speaker recognition with random digit strings	i-vectors	HMM	RSR2015	1.52% and 1.77% equal error rate (EER) for male and female speakers, respectively
Wang et al. [33]	Discriminant speaker embeddings for short-duration speech	i-vectors	Neural-network-based deep discriminant analysis (DDA)	SRE corpus	30% relative EER reduction

3. Proposed Framework

Our proposed methodology uses two deep neural networks: a recurrent neural network (RNN) and a multilayer perceptron, also called a deep neural network with a softmax layer as its last layer. The RNN is used for learning features from the audio signal. In contrast, the deep neural network with softmax is used to identify the speaker (i.e., for classification) based on these features [34]. The RNN is chosen because audio signal processing is a sequence processing problem. We use a multilayer perceptron with a softmax activation function for the speaker recognition task. Voice activity detection (VAD) is used to remove noise from the input signal. A BiLSTM, which can also be used for sequence processing problems, is also put in place of the RNN to compare its performance with the proposed system. A softmax classifier is used as the last layer of the deep neural network speaker identifier because it will convert the readings coming from the hidden layers into a probability distribution in which the most probable item, i.e., the speaker, will be output. For noise robustness, a framing-based noise removal method is incorporated into the preprocessing phase of the system. During this process, which is called voice activity detection, only the frames which belong to the actual speaker are preserved, while the frames belonging to the noise are discarded. This methodology uses the spectrum flatness index and energy ratio index of the signal to decide whether the frame belongs to the speaker or the noise [16]. The threshold used in this decision will be selected as a hyperparameter of the system. The layout of the proposed methodology is shown in Figure 1.

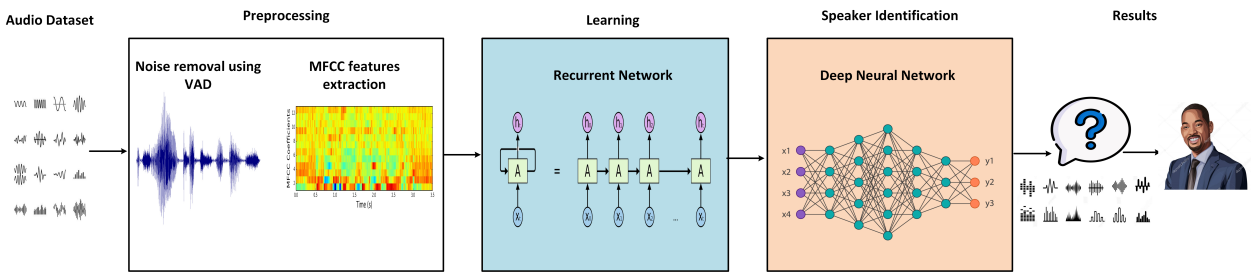

Figure 1. The proposed framework.

We use the Kaggle speaker recognition dataset. The dataset contains 16,000 samples to benchmark speaker recognition technology on single-speaker audio acquired under unconstrained conditions. The conditions represented in the dataset provide samples of five individuals in clean interviews and speech in different outdoor conditions.

3.1. Prepocessing

The purpose of preprocessing is to represent the raw audio signal in such a representation that its essence comes to the surface. In this phase, we perform a set of activities, namely, detection and extraction of voice activity. The preprocessing activities performed for the proposed system are discussed below.

Voice Activity Detection (VAD): The idea behind VAD is to take into account only actual speech frames and discard noisy frames. For this purpose, two indices are used which are (i) the spectrum flatness index and (ii) the energy ratio index. The former is a measure of noise in the signal spectrum, whereas the latter separates noise from speech frames taking into account the energy distribution along the signal spectrum.

Feature Extraction: We extract two types of features from audio files, which are MFCCs and spectrograms.

3.2. Learning

In this phase, features are learned from the preprocessed data, represented as MFCC features. An RNN is employed due to its ability to process sequential data, which, in this case, consist of audio features. The RNN autoencoder compresses MFCCs and spectrograms from a higher-dimensional space into a lower-dimensional space. To accelerate model convergence, the decoder's expected output from the previous step (i.e., the previous epoch) is fed back as input into the decoder RNN. Figure 2 shows an RNN in its simplest form. The equations for a recurrent neural network (RNN) are given below to show how it maps an input x into an output y:

$$h(t) = f_H(W_{IH}\, x(t) + W_{HH}\, h(t-1))$$
$$y(t) = f_O(W_{HO}\, h(t))$$

where W_{IH}, W_{HH}, and W_{HO} are weight matrices (I for input, O for output, and H for hidden layers).

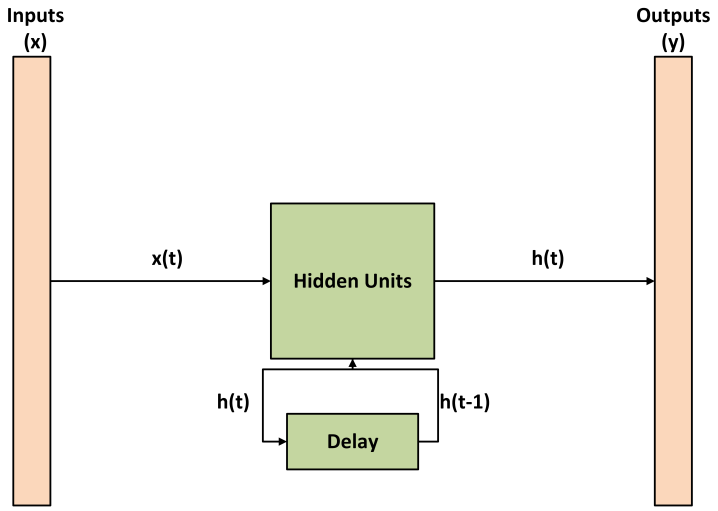

Figure 2. Illustration of recurrent neural network.

3.3. Speaker Identification

In this phase, the speaker of a given utterance is recognized based on the learned feature from the training data. A 5-fold cross-validation setup is used for the evaluation of the system for speaker identification. A deep neural network (multilayer perceptron) is trained and tested on the features learned during this phase. The optimal number of hidden layers and learning rate for this network will be identified in this work which will maximize or minimize the proposed evaluation metrics.

3.4. Softmax Classifier

A deep neural network is another name used for a multilayer perceptron. It is a perceptron with more than two layers. It is a feed-forward artificial neural network. Figure 3 shows an example of a deep neural network. A softmax classifier takes several numbers (usually the output of the hidden-layer neurons in the neural network) as input and transforms them into a probability distribution. When softmax is included in a neural network, it (softmax) is made the last layer of the network. The use of a softmax layer inside a deep neural network is needed, and it serves as the activation function of the network and transforms the outputs of the hidden layer into a probability distribution. The softmax classifier maps a number xi into probability as follows:

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)} = p_i \quad (1)$$

We evaluate our proposed work based on the following evaluation parameters.

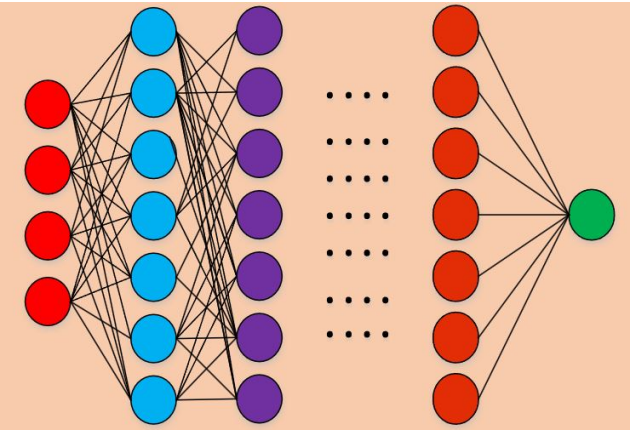

Figure 3. Illustration of deep neural network.

Accuracy: We calculate accuracy to find the ratio between the total number of input samples and the total number of correct predictions.

Root Mean Square Error: It is the measure of differences between the values predicted by a model and the observations.

4. Results and Discussion

4.1. Experimental Setup

To recognize the suspect identity, the system must consider the evidence recording both in clear and noisy environments and the recording used to match the pre-recorded sounds in different environments. Therefore, using a multilayered perceptron with a softmax output layer, we perform a set of experiments. The features used to compare the performance of the system are as follows:

i. Spectrograms.
ii. MFCC features.

The model for the proposed MLP system is shown in Figure 4. Our system is compared to two RNN classifiers including LSTM and BiLSTM. Figure 5 shows the LSTM model used in the experiments and Figure 6 shows the BiLSTM model used in the experiments.

To see the effectiveness of the spectrogram features for speaker identification, we perform an experiment as depicted in Figure 7. The effectiveness of the MFCC features is examined for the speaker identification task. Figure 8 shows this experiment. The experiments are performed on the Kaggle speaker recognition dataset. We extract the spectrogram, MFCC, LSTM, and BiLSTM features considering the following points:

- The files are variable length. To extract fixed-length spectrograms and MFCCs, we use the technique of zero padding.
- The autoencoders transform any length input into a fixed-length feature vector as output.

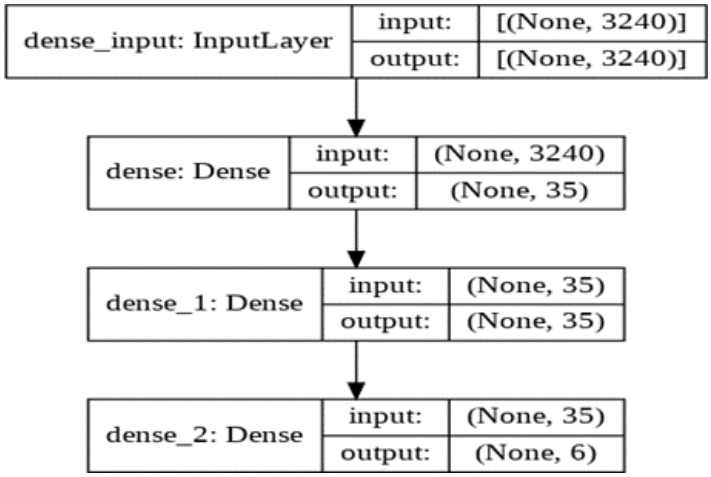

Figure 4. The proposed MLP classifier.

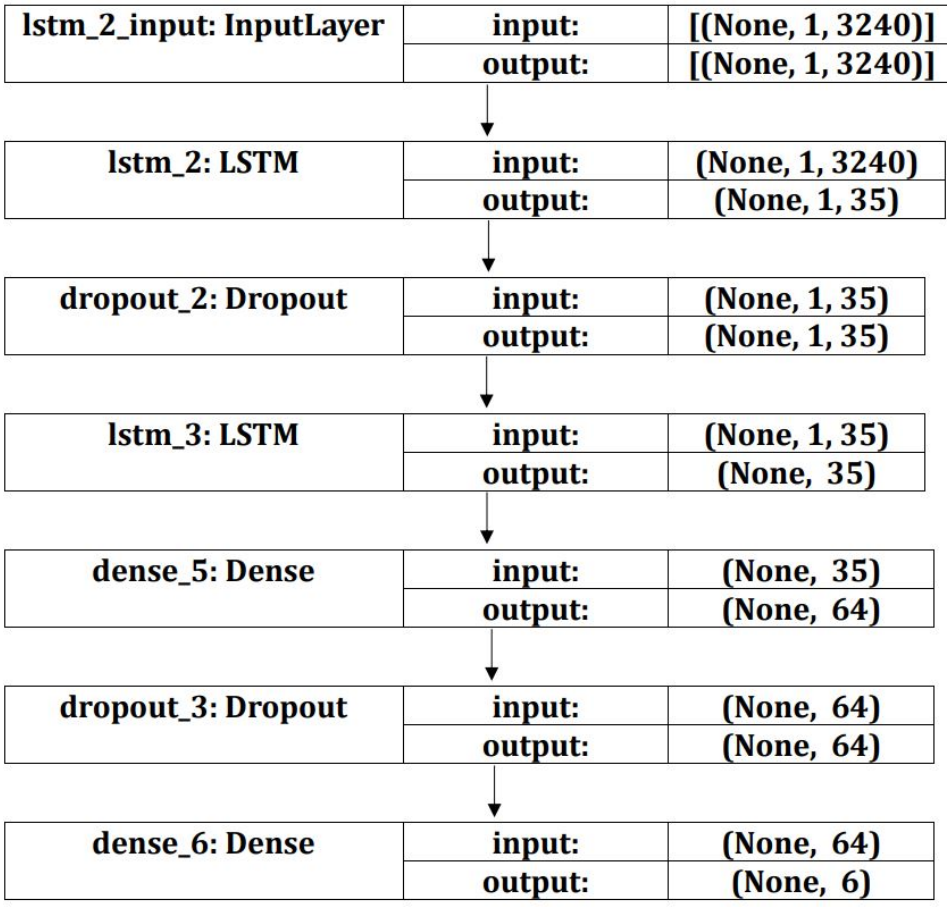

Figure 5. The LSTM network used.

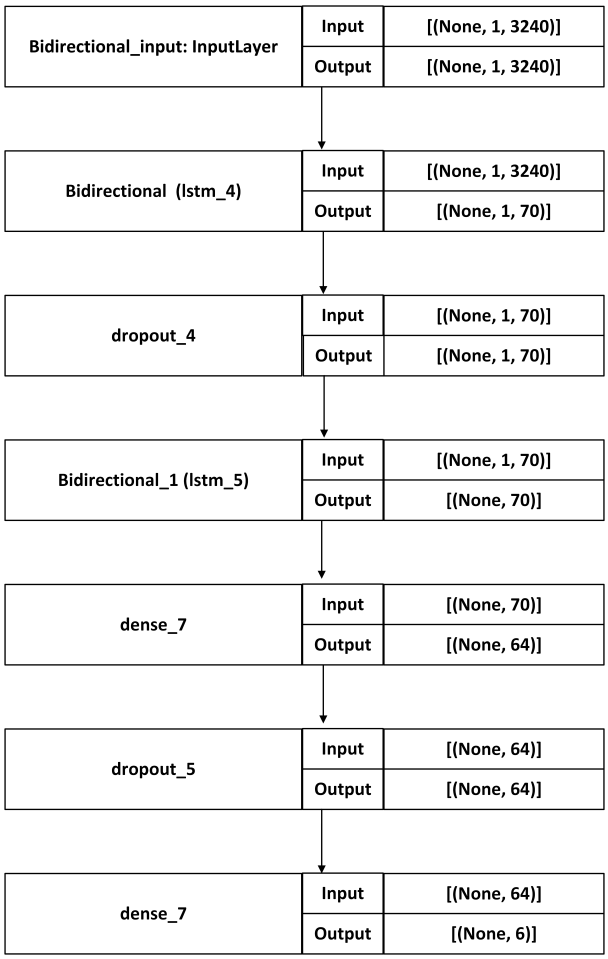

Figure 6. The BiLSTM model.

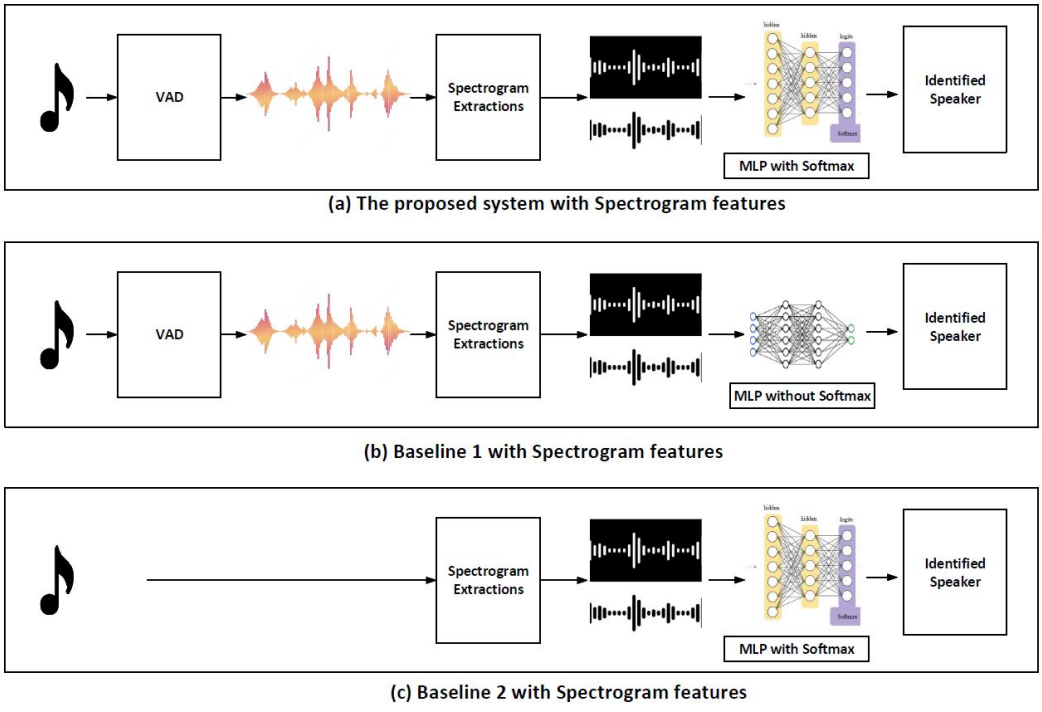

Figure 7. The proposed framework compared with baselines in terms of spectrogram features.

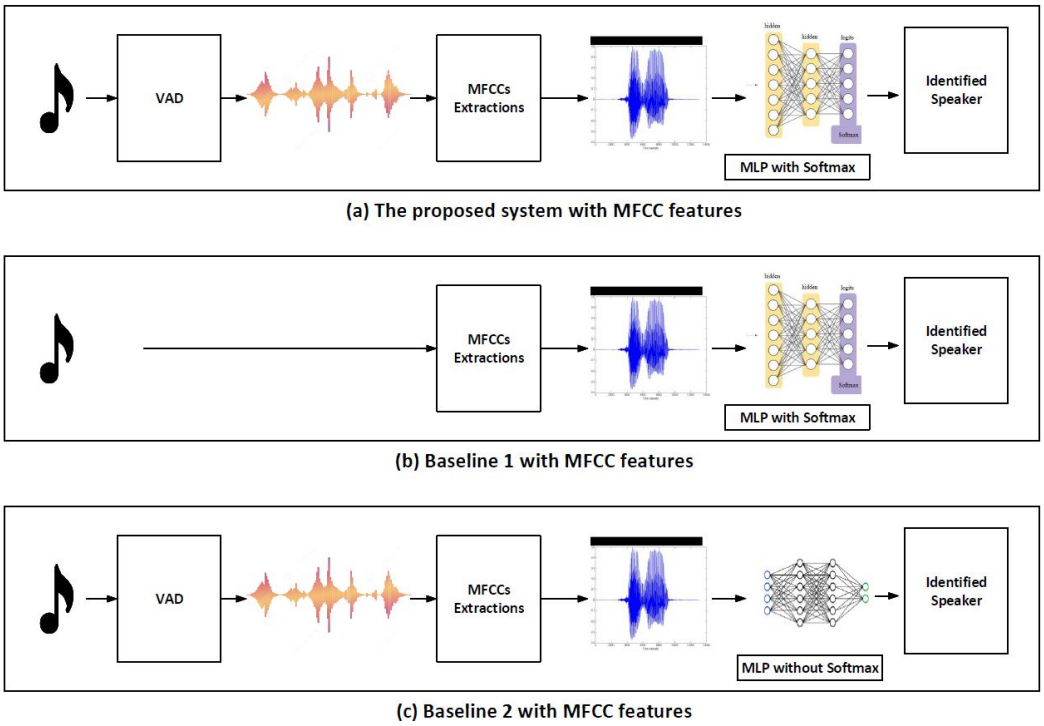

Figure 8. The proposed framework compared with baselines in terms of MFCC features.

4.2. Experimental Results

After extraction of the four types of features, these are embedded in the multilayer perceptron with a softmax activation function. Table 2 shows the selected hyperparameter settings. The results of the experiments are shown in Table 3, which shows that the proposed system outperforms both of the baseline systems with a gain in accuracy of 3.6% and 5.8%. The results suggest the best feature set for the proposed system is MFCCs with VAD. We find that, on the application of VAD, the spectrograms and MFCC feature vectors become of variable size. This is because VAD filters noisy frames from speech signals, with only speech frames filtered out. To handle the variable length of the feature vectors, i.e., to make them fixed length, we perform zero padding using Algorithm 1.

Table 2. Hyperparameter settings for the model.

Number	Hyperparameter Name	Value
1	Number of hidden layers	2
2	Number of neurons in the first hidden layer	35
3	Number of neurons in the second hidden layer	35
4	Loss function	Categorical_crossentropy
5	Optimizer	Adam

Table 3. Performance of systems using different features.

System	Spectrograms	MFCCs
Baseline1	0.829	0.958
Baseline2	0.833	0.936
The proposed system	0.806	0.994

Algorithm 1 Zero padding of the feature vectors

1: **Input:** Set of feature vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$
2: **Output:** Zero-padded feature vectors
3: Calculate the size i of the largest vector
4: **for** each feature vector $\mathbf{v}_j$ **do**
5: Calculate pad size as $i - \text{size}(\mathbf{v}_j)$
6: Pad zeros equal to $i - \text{size}(\mathbf{v}_j)$ to vector $\mathbf{v}_j$
7: **end for**

One of the effects of using VAD before feature calculation is that it reduces the dimensionality of the features (16,281 and 11,457 spectrogram features, and 3240 vs. 2280 MFCC features), which, in turn, decreases the step size, as shown in Table 4. The model loss during training with different features is presented in Figure 9. It can be observed that, when using spectrograms, the model converges more slowly compared to during the use of MFCC features. We also notice that the MLP model converges earlier in terms of training loss compared to the LSTM and BiLSTM models. The loss declines sharply after just a few epochs, which supports the idea that the MLP model rapidly learns the input data weights. LSTM and BiLSTM take longer to converge due to their complexity and the need to handle sequential data input dependencies. Additionally, the figure shows that, with the application of VAD, the noise interference in the MLP model is minimized, resulting in a smoother loss curve.

Table 4. The average step size reduction due to VAD.

MFCC Features	Without VAD	With VAD	Spectrogram Features	With VAD
	8 s 22 ms/step	3 s 6 ms/step	16 s 31 ms/step	11 s 21 ms/step

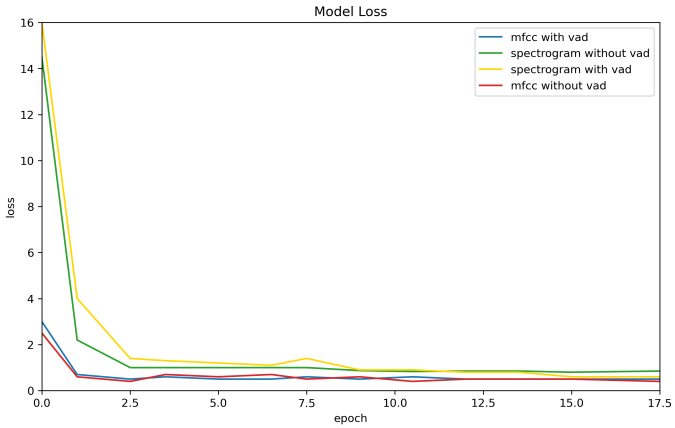

Figure 9. MLP model loss with different features.

The validation loss shown in Figure 10 also indicates that MFCC features with VAD outperform both MFCC features without VAD and spectrogram features. The validation loss curves for the MLP, LSTM, and BiLSTM models are displayed. In this experiment, the MLP consistently outperforms both LSTM and BiLSTM across all epochs, maintaining a lower validation loss. This suggests that the MLP is more likely to generalize well to unseen datasets compared to LSTM and BiLSTM, as the latter models exhibit higher fluctuations in validation loss. This is particularly true when MFCC features are combined with VAD, as VAD helps stabilize the validation performance of the MLP model.

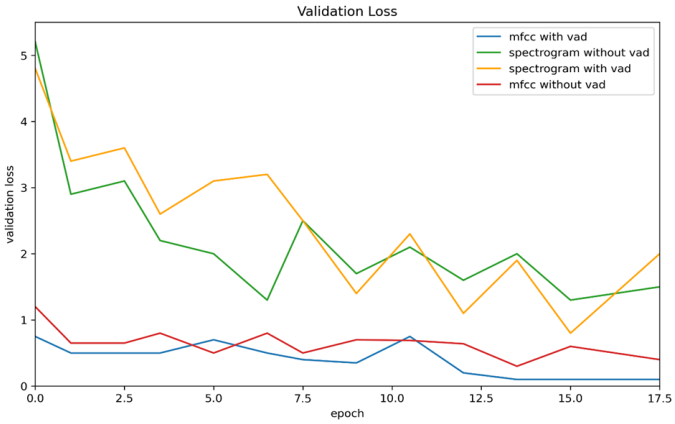

Figure 10. MLP model validation loss.

The model accuracy is shown in Figure 11. The MLP model achieves the highest training accuracy, surpassing both LSTM and BiLSTM. The prominent features learned by the MLP model contribute to a steeper rise in accuracy. Although the accuracy of LSTM and BiLSTM improves over the epochs, it remains slightly below that of the MLP model, indicating that these models require more epochs to capture sequential information but still do not reach the accuracy level of MLP. This figure reinforces the idea that using VAD with MFCCs improves model performance, particularly in noisy environments.

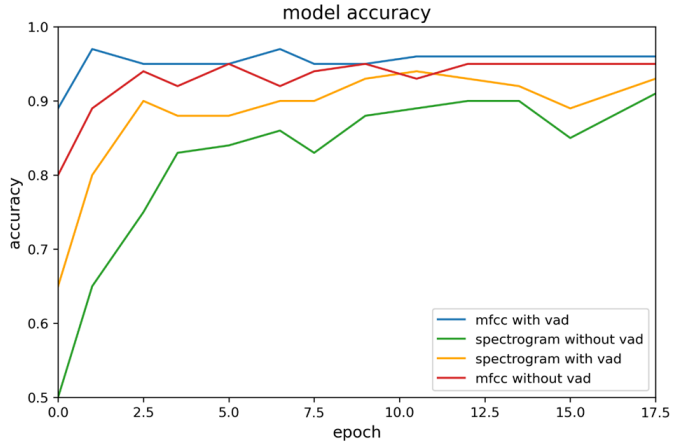

Figure 11. MLP model accuracy.

The validation accuracy is shown in Figure 12. As depicted, while the LSTM and BiLSTM models have the lowest validation accuracy, the MLP model performs significantly better. The consistency of the validation accuracy curve for the MLP suggests that it can predict speakers in the test data more reliably. Although LSTM and BiLSTM show steady improvements in accuracy, they experience greater intra-iteration fluctuations compared to the embedding-RNN case, and their rate of accuracy improvement is relatively slower. This further highlights the effectiveness of the MLP model, particularly when using MFCCs in combination with VAD, which leads to superior validation performance.

In both Figures 11 and 12, it can be observed that MFCCs with VAD outperform both MFCCs without VAD and spectrogram features with and without VAD. These results also suggest that spectrograms perform poorly for speaker identification compared to MFCC features. After identifying the best feature set with VAD, we compare our proposed model, i.e., the MLP with softmax, against LSTM and BiLSTM classifiers. The accuracy results of the MLP, LSTM, and BiLSTM classifiers are shown in Table 5.

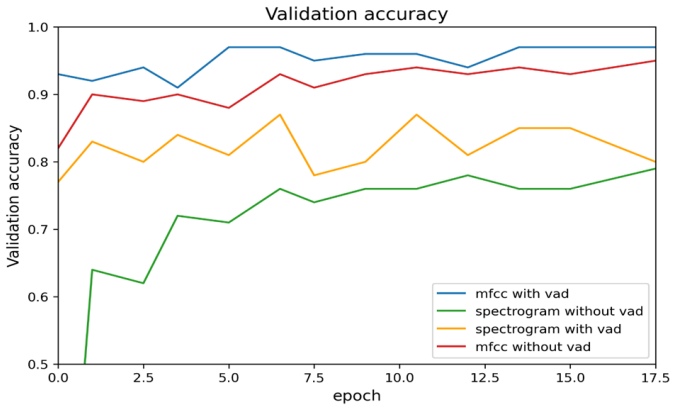

Figure 12. MLP model validation accuracy.

Table 5. Performance metrics for different models.

Metric	MLP	LSTM	BiLSTM
Test Accuracy	0.994	0.762	0.929
Test Loss	0.088	0.603	0.263
Test Mean Squared Error	0.005	0.0372	0.0218

Note that all of the MLP, LSTM, and BiLSTM models are compared on the following settings:

- VAD performed;
- MFCC features extracted;
- Softmax output activation function.

It can be seen that the proposed system outperforms the LSTM and BiLSTM classifiers. Figure 13 compares the model accuracy and validation accuracy of the MLP, LSTM, and BiLSTM models. The error rate is consistently lower in the MLP compared to the LSTM and BiLSTM models, which demonstrates more accurate speaker identification. The sharp decline in the MSE for the MLP confirms its ability to reduce prediction errors in less time than more established methods, such as those using VAD. The MSE in the LSTM and BiLSTM models remains slightly higher throughout, suggesting that their sequential nature may introduce more variability in predictions, especially when noise affects the data. Figure 14 compares the model loss and validation loss of these models, showing that the validation MSE for the MLP continues to be lower than that of LSTM and BiLSTM, proving that the MLP has a better ability to generalize successfully on unseen data. The relatively stable and lower MSE curve for the MLP supports the hypothesis that MFCC features combined with VAD enhance noise robustness. The MSE for both LSTM and BiLSTM, though slightly lower than for MLP, highlights their sensitivity to noise and variable-length data.

To evaluate the performance of our proposed system, we compare it with the following baselines.

Baseline 1: We use a multilayered perceptron without a softmax output layer and with VAD as our first baseline.

Baseline 2: We use a multilayered perceptron with a softmax output layer and without VAD as our second baseline.

We compare the proposed system against these two baseline models, as well as the two following recurrent neural network (RNN) architectures: Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM).

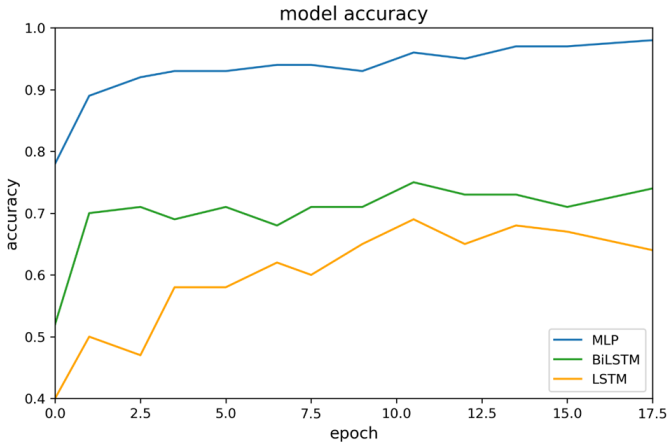

Figure 13. Model accuracy of the three models.

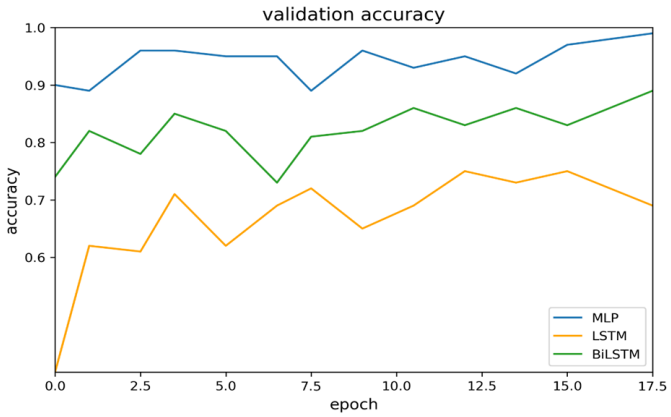

Figure 14. Validation accuracy comparison.

The results clearly show that our method outperforms the two baseline models. The first baseline, based on an MLP with VAD but without the softmax output layer (Baseline 1), provides lower accuracy. This indicates that including softmax allows the model to extract speaker information more effectively and distinguish it from noise, which is crucial in noisy environments. Baseline 2, which lacks VAD, performs the worst of all systems (and significantly worse compared to the proposed system), underscoring the importance of VAD in reducing noise and improving overall accuracy.

The proposed system also outperforms both LSTM and BiLSTM, as shown in Figures 15 and 16. This result is somewhat surprising given that RNN architectures like LSTM and BiLSTM are expected to perform well on sequential data. However, the integration of MFCCs with VAD and softmax appears to provide a more powerful feature set for the MLP, outperforming these sequentially concatenated models. Additionally, the oscillations observed in the MLP's loss may be caused by the learning rate or the complex nature of the dataset, which includes noisy environments. These oscillations are not a major concern, as the general trend shows convergence, and the model continues to improve in accuracy over time. Furthermore, the validation loss remains substantial, indicating that the model has not overfitted and is capable of generalizing to unseen inputs.

Moreover, the proposed system converges faster than the LSTM and BiLSTM models, as shown in Figures 17 and 18. A key advantage is its faster convergence; the model learns more quickly, meaning it requires less computing time and fewer resources to optimize performance. This makes the simpler MLP architecture, combined with VAD and softmax, more efficient for speaker identification tasks while still providing comparable performance to more complex LSTM and BiLSTM models. Our proposed system, leveraging an MLP with VAD and softmax, converges faster than RNN architectures and achieves higher

accuracy. This makes it an optimal choice for speaker identification tasks, especially in noisy environments. The results highlight the effectiveness of integrating VAD for noise reduction and softmax for enhancing speaker identification, providing a robust solution for real-world applications.

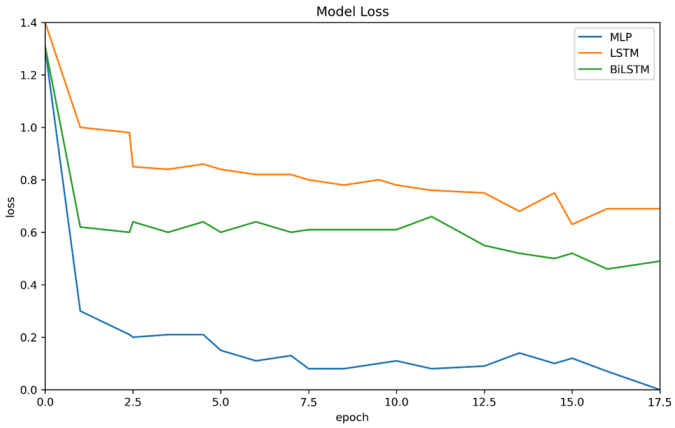

Figure 15. Model loss of the three models.

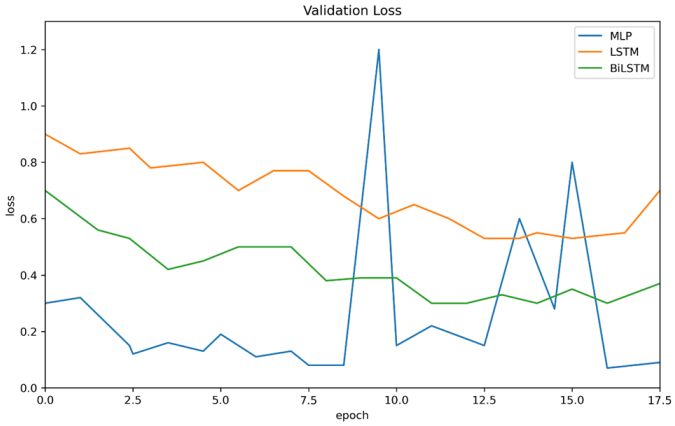

Figure 16. Validation loss of the three models.

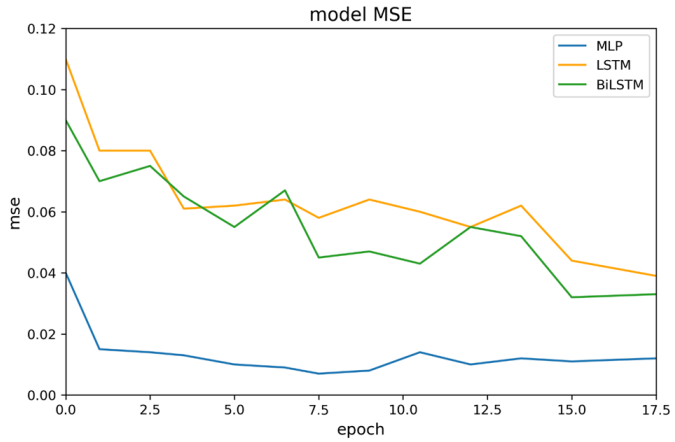

Figure 17. Model MSE of the three models.

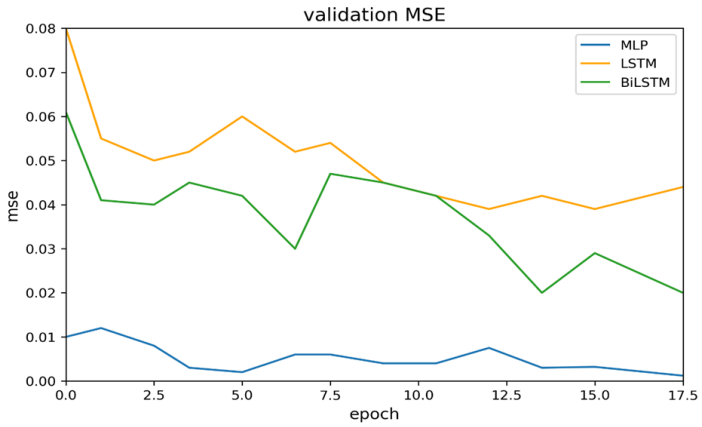

Figure 18. Validation MSE of the three models.

5. Conclusions

In this article, we present a novel approach to speaker recognition in noisy environments utilizing voice activity detection (VAD), a multilayer perceptron (MLP) classifier, and a softmax activation function. We conducted a series of experiments to demonstrate the effectiveness of these techniques in addressing the challenges of speaker recognition under noise. The results of our study show that the proposed system, which combines MFCCs with VAD and softmax, significantly outperforms the two baseline systems. Furthermore, we observed that the MLP classifier delivers superior performance compared to Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) classifiers for the speaker recognition task. Our findings indicate that using VAD, an MLP, and softmax with MFCC features provides a robust solution for speaker recognition, especially in noisy environments. This suggests that the MLP classifier, when combined with these techniques, offers a more reliable and efficient approach for handling noise in speaker recognition tasks.

Author Contributions: Conceptualization, S.S.Z.; methodology, S.S.Z.; software, E.M.; validation, A.N., M.H.A. and S.K.u.Z.; formal analysis, M.H.A. and A.N.; investigation, S.S.Z.; resources, S.S.Z., M.H.A. and E.M.; data curation, S.S.Z. and S.K.u.Z.; writing—original draft preparation, S.S.Z.; writing—review and editing, E.M. and S.K.u.Z.; visualization, M.H.A. and A.N.; supervision, M.H.A.; project administration, A.N.; funding acquisition, S.S.Z. and A.N. All authors have read and agreed to the published version of the manuscript.

Funding: The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through the project number "NBU-FFR-2024-1180-02".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data associated with this research will be made available from the corresponding authors upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

VAD	Voice Activity Detection
MLP	Multilayer Perceptron
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
ASR	Automatic Speech Recognition
DNN	Deep Neural Network
RNN	Recurrent Neural Network

GMM	Gaussian Mixture Model
SVM	Support Vector Machine
CNN	Convolutional Neural Network

References

1. Freitag, M.; Amiriparian, S.; Pugachevskiy, S.; Cummins, N.; Schuller, B. Audeep: Unsupervised learning of representations from audio with deep recurrent neural networks. *J. Mach. Learn. Res.* **2017**, *18*, 6340–6344.
2. Piczak, K.J. Environmental sound classification with convolutional neural networks. In Proceedings of the 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, USA, 17–20 September 2015; pp. 1–6.
3. Bahuleyan, H. Music genre classification using machine learning techniques. *arXiv* **2018**, arXiv:1804.01149.
4. Nassif, A.B.; Shahin, I.; Attili, I.; Azzeh, M.; Shaalan, K. Speech recognition using deep neural networks: A systematic review. *IEEE Access* **2019**, *7*, 19143–19165. [CrossRef]
5. Samudravijaya, K.; Shah, S.; Pandya, P. *Computer Recognition of Tabla Bols*; Technical Report; Tata Institute of Fundamental Research: Mumbai, India, 2004.
6. Subasi, A. EEG signal classification using wavelet feature extraction and a mixture of expert model. *Expert Syst. Appl.* **2007**, *32*, 1084–1093. [CrossRef]
7. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi speech recognition toolkit. In Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Waikoloa, HI, USA, 11–15 December 2011; IEEE Signal Processing Society: Piscataway, NJ, USA, 2011.
8. Dave, N. Feature extraction methods LPC, PLP and MFCC in speech recognition. *Int. J. Adv. Res. Eng. Technol.* **2013**, *1*, 1–4.
9. Yin, C.; Zhu, Y.; Fei, J.; He, X. A deep learning approach for intrusion detection using recurrent neural networks. *IEEE Access* **2017**, *5*, 21954–21961. [CrossRef]
10. Chen, C.; Wang, W.; He, Y.; Han, J. A bilevel framework for joint optimization of session compensation and classification for speaker identification. *Digit. Signal Process.* **2019**, *89*, 104–115. [CrossRef]
11. Biagetti, G.; Crippa, P.; Falaschetti, L.; Orcioni, S.; Turchetti, C. An investigation on the accuracy of truncated DKLT representation for speaker identification with short sequences of speech frames. *IEEE Trans. Cybern.* **2016**, *47*, 4235–4249. [CrossRef]
12. Chin, Y.-H.; Wang, J.-C.; Huang, C.-L.; Wang, K.-Y.; Wu, C.-H. Speaker identification using discriminative features and sparse representation. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 1979–1987. [CrossRef]
13. Ranjan, S.; Hansen, J.H.L. Curriculum learning based approaches for noise robust speaker recognition. *IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP)* **2018**, *26*, 197–210. [CrossRef]
14. Bisio, I.; Garibotto, C.; Grattarola, A.; Lavagetto, F.; Sciarrone, A. Smart and robust speaker recognition for context-aware in-vehicle applications. *IEEE Trans. Veh. Technol.* **2018**, *67*, 8808–8821. [CrossRef]
15. Ghahabi, O.; Hernando, J. Deep learning backend for single and multisession i-vector speaker recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 807–817. [CrossRef]
16. Cumani, S.; Laface, P. Nonlinear i-vector transformations for PLDA-based speaker recognition. *IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP)* **2017**, *25*, 908–919. [CrossRef]
17. Stolcke, A.; Kajarekar, S.S.; Ferrer, L.; Shrinberg, E. Speaker recognition with session variability normalization based on MLLR adaptation transforms. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 1987–1998. [CrossRef]
18. Campos, V.d.A.; Pedronette, D.C.G. A framework for speaker retrieval and identification through unsupervised learning. *Comput. Speech Lang.* **2019**, *58*, 153–174. [CrossRef]
19. Tiwari, V.; Hashmi, M.F.; Keskar, A.; Shivaprakash, N.C. Speaker identification using multi-modal i-vector approach for varying length speech in voice interactive systems. *Cogn. Syst. Res.* **2019**, *57*, 66–77. [CrossRef]
20. Sahidullah, M.; Saha, G. Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. *Speech Commun.* **2012**, *54*, 543–565. [CrossRef]
21. Tirumala, S.S.; Shahamiri, S.R.; Garhwal, A.S.; Wang, R. Speaker identification features extraction methods: A systematic review. *Expert Syst. Appl.* **2017**, *90*, 250–271. [CrossRef]
22. Safavi, S.; Russell, M.; Jancović, P. Automatic speaker, age-group and gender identification from children's speech. *Comput. Speech Lang.* **2018**, *50*, 141–156. [CrossRef]
23. Nayana, P.K.; Mathew, D.; Thomas, A. Comparison of text independent speaker identification systems using GMM and i-vector methods. *Procedia Comput. Sci.* **2017**, *115*, 47–54. [CrossRef]
24. Singh, N.; Khan, R.A.; Shree, R. Applications of speaker recognition. *Procedia Eng.* **2012**, *38*, 3122–3126. [CrossRef]
25. Jaf, S.; Calder, C. Deep learning for natural language parsing. *IEEE Access* **2019**, *7*, 131363–131373. [CrossRef]
26. An, N.; Thanh, N.; Liu, Y. Deep CNNs with Self-Attention for Speaker Identification. *IEEE Access* **2019**, *7*, 85259–85268. [CrossRef]
27. Ayadi, M.E.; Hassan, A.K.S.O.; Abdel-Naby, A.; Elgendy, O.A. Text-independent speaker identification using robust statistics estimation. *Speech Commun.* **2017**, *92*, 52–63. [CrossRef]
28. Zeinali, H.; Sameti, H.; Burget, L. HMM-based phrase-independent i-vector extractor for text-dependent speaker verification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1421–1435. [CrossRef]
29. Cumani, S.; Laface, P. Speaker recognition using e-vectors. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 736–748. [CrossRef]

30. Xu, L.; Lee, K.A.; Li, H.; Yang, Z. Generalizing I-vector estimation for rapid speaker recognition. *IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP)* **2018**, *26*, 749–759. [CrossRef]
31. Luo, Y.; Chen, Z.; Mesgarani, N. Speaker-independent speech separation with deep attractor network. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 787–796. [CrossRef]
32. Maghsoodi, N.; Sameti, H.; Zeinali, H.; Stafylakis, T. Speaker Recognition with Random Digit Strings Using Uncertainty Normalized HMM-Based iVectors. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1815–1825. [CrossRef]
33. Wang, S.; Huang, Z.; Qian, Y.; Yu, K. Discriminative Neural Embedding Learning for Short-Duration Text-Independent Speaker Verification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1686–1696. [CrossRef]
34. Chen, D.; Manning, C.D. A fast and accurate dependency parser using neural networks. In Proceedings of the EMNLP, Doha, Qatar, 25–29 October 2014.