Journal of Science Research and Reviews
PRINT ISSN: 1595-9074
E-ISSN: 1595-8329
DOI: https://doi.org/10.70882/josrar.2026.v3i3.151
Homepage: https://josrar.esrgngr.org
Original Research Article

An Intelligent Speech Recognition Framework Using Hidden Markov Models and Actor–Critic Deep Reinforcement Learning for Low-Resource African Languages

***[1]Samson Isaac, [2]Muhammad Aminu Ahmad, [3]Peter Ayuba**

[1]Department of Computer Science, Federal University of Applied Sciences Kachia, Kaduna,
[2]Department of Secure Computing, Kaduna State University, Kaduna,
[3]Department of Mathematical Sciences, Kaduna State University, Kaduna
*Corresponding Author's email: samson.isaac@fuask.edu.ng

KEYWORDS
Automatic Speech Recognition (ASR),
Hidden Markov Model (HMM),
Actor–Critic Deep Reinforcement Learning,
Deep Deterministic Policy Gradient (DDPG),
Mel-Frequency Cepstral Coefficients (MFCC).

CITATION
Isaac, S., Ahmad, M. A., & Ayuba, P. (2026). An Intelligent Speech Recognition Framework Using Hidden Markov Models and Actor–Critic Deep Reinforcement Learning for Low-Resource African Languages. *Journal of Science Research and Reviews*, *3*(3), 133-142. https://doi.org/10.70882/josrar.2026.v3i3.151.

ABSTRACT
ASR also plays a crucial role in assistive technologies for individuals with disabilities by enabling them to manage their surroundings more effectively through dialing phone numbers, operating light switches, and controlling home appliances, thereby contributing to the development of smart home systems. This study extracts features from isolated speech using Mel-Frequency Cepstral Coefficients (MFCC) and Bidirectional Long Short-Term Memory (BiLSTM) networks to ensure speaker invariance and enhance feature localization. Deep learning techniques were employed to explicitly normalize speech spectral features. Numerous pattern recognition and regression tasks have demonstrated the effectiveness of LSTM-based architectures. The novelty of this study lies in the integrating a hybrid MFCC–DNN–HMM framework to achieve high speech recognition accuracy for isolated words. The model achieved an accuracy of 0.945 (94.5%), indicating that it correctly classified the majority of instances. The precision obtained was 0.901, meaning that 90.1% of the instances identified as positive were correctly classified. The recall rate was 0.92, indicating that 92% of the actual positive instances were successfully detected by the system. The F1-score was 0.909, reflecting a balanced measure of precision and recall.

INTRODUCTION

Speech recognition technology has become highly valuable across various application domains. It enhances user experience in voice-activated navigation systems, virtual assistants on mobile devices, and voice-based authentication systems by adding an extra layer of security. In commercial settings, ASR can be used for transcription, providing a faster and more efficient alternative to typing, particularly for non-typists. ASR also plays a crucial role in assistive technologies for individuals with disabilities by enabling them to manage their surroundings more effectively through dialing phone numbers, operating light switches, and controlling home appliances, thereby contributing to the development of smart home systems. ASR systems are broadly classified into speaker-independent and speaker-dependent systems, isolated-word recognition systems, continuous speech recognition systems, and text-dependent recognition systems (Fadhel & Mohammed, 2023). Traditionally, the various modalities of speech sounds have been modeled using statistical approaches such as Gaussian Mixture Models (GMMs), where Gaussian parameters are estimated using Maximum Likelihood (ML) criteria (Ramadan & Bitmead, 2022). Speech recognition systems have significantly improved in terms of accuracy due to recent advances in deep learning techniques,

including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). These models can learn complex feature mappings, recognize intricate patterns, and process large volumes of data. However, their effectiveness and accessibility may be constrained by high computational costs and substantial resource requirements. Reinforcement learning, a branch of machine learning that enables agents to make decisions based on rewards and penalties, has been integrated with deep learning to form Deep Reinforcement Learning (DRL). This approach enables agents to learn from high-dimensional sensory inputs, further enhancing the capabilities of ASR systems. Hidden Markov Models (HMMs) remain foundational probabilistic models in ASR for capturing the temporal dynamics of speech signals and accommodating variability across different speakers and contexts. This study proposes an improved Deep Reinforcement Learning framework integrated with a Hidden Markov Model for speech recognition to address the limitations of existing ASR systems. The proposed approach combines the strengths of conventional statistical models and modern deep learning techniques while mitigating their individual limitations. The primary objective is to develop a computationally efficient model capable of operating effectively on limited hardware resources by integrating deep reinforcement learning with the temporal modeling capabilities of HMMs. This integration is expected to enhance accuracy and robustness, particularly in noisy environments and across diverse accents and dialects. Furthermore, the model aims to reduce bias toward African speech variations by incorporating diverse datasets during training. The proposed HMM-based speech recognition model with Deep Reinforcement Learning offers significant benefits to various stakeholders. It provides more accurate and reliable speech recognition, especially for users with diverse accents and dialects, thereby promoting inclusivity and better service delivery to underrepresented speech communities. The model is designed to be efficiently deployable on various hardware platforms, including mobile devices, enabling the development of innovative speech-enabled applications across multiple industries.

Babu et al. (2023) classified speech recognition approaches into three main categories: Pattern Recognition, Acoustic-Phonetic, and Artificial Intelligence approaches. Techniques such as Dynamic Time Warping (DTW), Linear Predictive Coding (LPC), Vector Quantization (VQ), and Mel-Frequency Cepstral Coefficients (MFCC) are commonly used for feature extraction. Santos et al. (2023) presented an HMM-CNN hybrid model for speech recognition. The study combined the strengths of Hidden Markov Models (HMMs) and Convolutional Neural Networks (CNNs) to improve recognition accuracy. However, the increased computational complexity and the lack of training with African speech data limited its generalizability to African contexts. Shafieian (2023) proposed a Persian speech recognition system using a Hidden Markov Model, demonstrating the effectiveness of HMMs for Persian language recognition. However, the study was limited to the Persian language, suggesting that its findings may not generalize well to African speech data. Sudarshan et al. (2023) developed an improved contextual recognition system for automatic speech recognition. Although the system enhanced contextual recognition, the additional semantic processing and lack of training on African speech data suggest potential limitations in African contexts. Furthermore, the system exhibited issues related to complexity, limited discriminative ability, and sensitivity to speech variations. Sallagundla et al. (2023) presented a voice-enabled form-filling system utilizing a Hidden Markov Model. The applicability of the system in African contexts remains unverified due to the absence of training data from African languages. Additionally, The model requires substantial training data, exhibits limited discriminative ability, involves high computational complexity, and requires significant retraining to adapt to new datasets. Prabhu and Jayasri (2024) explored a speech recognition system for vending machines using a Hidden Markov Model. The study focused primarily on European speech and accents, indicating potential limitations in generalizability. The model is data-dependent, exhibits limited discriminative ability, and shows low sensitivity to noise and speech variations. Pavithran and Sherly (2024) developed an HMM-based ASR model for individuals with hearing impairment. Although the system supports a specialized user group, it was not trained on African speech data, which may limit its applicability in African contexts. The model requires a fixed topology, large amounts of labeled data, demonstrates limited discriminative capability, and is sensitive to noise and variations. Additionally, it struggles to capture long-range dependencies in speech signals. Ouisaadane et al. (2024) proposed a Moroccan dialect speech recognition system using PocketSphinx. While the study effectively addressed noisy environments, its applicability to other dialects remains uncertain. The system requires a fixed topology and large labeled datasets and also faces challenges related to limited discriminative ability, scalability, simplistic feature representations, sensitivity to noise, inability to capture long-range dependencies, overfitting, and poor adaptability to new data without significant retraining. Mishra et al. (2024) designed a speech recognition system using various machine learning techniques. Although the study provided a broad overview, it lacked depth and did not include training with African speech data, resulting in potential generalizability issues. The model also exhibited limitations such as fixed topology, high data requirements, limited discriminative

ability, complexity, simplistic feature representation, noise sensitivity, difficulty capturing long-range dependencies, overfitting, and poor adaptability. Thimmaraja et al. (2024) proposed a real-time automated continuous speech recognition system for Kannada language and dialects. Although the custom HMM system achieved real-time performance, it lacked generalizability to African languages due to limitations in training data, model structure, and adaptability. Hazmoune et al. (2024) introduced an ensemble method based on HMMs, achieving high accuracy and robustness. However, ensemble methods increase computational complexity and require large labeled datasets. The lack of training on African speech data further limits generalizability. Despite improved accuracy, the system suffers from fixed topology constraints, scalability issues, and limited adaptability. Isaac et al. (2023) proposed a Deep Reinforcement Learning with Hidden Markov Model (HMM) framework for speech recognition aimed at improving the recognition of isolated speech in Nigerian languages, specifically Hausa, Igbo, and Yoruba. The study integrated Mel-Frequency Cepstral Coefficients (MFCC), Long Short-Term Memory (LSTM), and Hidden Markov Models (HMM) to enhance speech prediction accuracy and reduce miss rates. The researchers highlighted the significance of Deep Reinforcement Learning (DRL) in improving Automatic Speech Recognition (ASR) systems through interaction with the environment and adaptive learning. The model utilized MFCC for feature extraction, LSTM for capturing long-term dependencies in sequential speech data, and HMM for probabilistic sequence modeling and word prediction. Experimental findings demonstrated that the proposed system achieved high prediction accuracy with a low miss rate in recognizing isolated spoken words representing various fruits in Nigerian local dialects. However, the study focused mainly on isolated-word recognition and did not extensively address continuous speech recognition, noisy environments, or scalability across larger vocabularies and diverse African accents. Moondra et al. (2023) proposed a modified MFCC-GMM approach to improve speaker recognition under degraded speech conditions. While effective in noisy environments, the model requires substantial labeled data and may not generalize well in contexts with limited African speech datasets. Manideep and Mohana (2023) developed a GMM-HMM-based voice recognition system, demonstrating the potential of GMM for improved recognition rates. However, the model requires large labeled datasets for optimal performance and is prone to overfitting when trained on limited data, thereby limiting its applicability in African contexts. Aljinu Khadar et al. (2023) presented a Gaussian Mixture Model–Universal Background Model (GMM-UBM) I-Vector method for speaker verification in noisy environments. Although effective for forensic applications, the approach is complex, data-intensive, and may not generalize well to African speech due to similar data limitations. Tsai and Wang (2023) developed a hardware-based GMM speaker verification system using MFCC features. While the system offers computational efficiency, it has limited adaptability to diverse speech variations that may be present in African languages. Biswas et al. (2023) employed MFCC features for spoken language identification and achieved high accuracy. However, the system is sensitive to noise and speech variations, which may affect performance in African contexts. Kanke et al. (2023) proposed a Marathi speech recognition system using language-specific techniques. While effective for Marathi, the approach highlights the need for language-specific modeling and may not generalize to African languages. Barkani et al. (2023) utilized the Kaldi ASR toolkit for Amazigh speech recognition. Although the study supports an African language, detailed analysis of scalability and generalizability remains limited. Nugroho et al. (2023) introduced a multi-accent speaker detection method using normalized MFCC features with a neural network. While addressing accent variability, the approach requires extensive training data across multiple accents, which may be limited for African languages. Wirdiani et al. (2024) proposed an MFCC-CNN model with online triplet mining for speaker recognition. Although the method improves recognition performance, it requires substantial training data and may not generalize effectively in scenarios with limited African speech datasets.

MATERIALS AND METHODS

This study proposes an improved and computationally efficient speech recognition framework that integrates Bidirectional Long Short-Term Memory (BiLSTM), Hidden Markov Models (HMM), and an Actor–Critic Deep Reinforcement Learning (DRL) architecture. The model incorporates an enhanced Mel-Frequency Cepstral Coefficient (MFCC)-based feature extraction technique to improve robustness, temporal modeling, and recognition accuracy. The overall architecture was designed to leverage the strengths of statistical modeling (HMM), deep sequential learning (BiLSTM), and policy optimization (Actor–Critic DRL) within a unified framework.

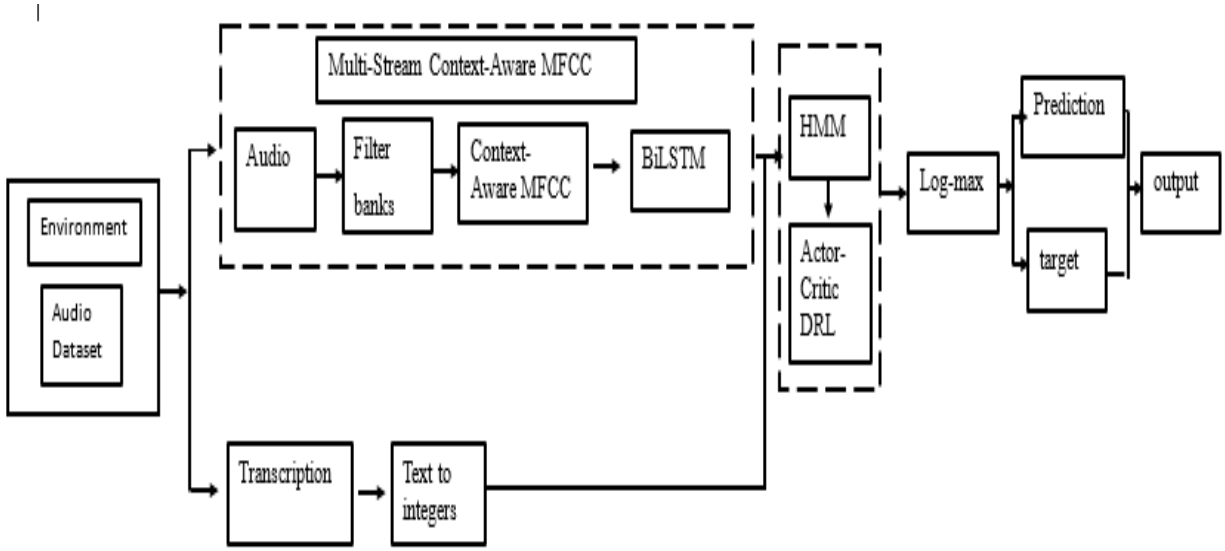

Figure 1: Proposed Improved Lightweight Deep Reinforcement Learning with HMM Speech Recognition Architecture

System Architecture Overview

The proposed Lightweight Deep Reinforcement Learning–HMM Speech Recognition Architecture consists of the following major components:

1. Speech Environment (Dataset)
2. Feature Extraction Module
3. BiLSTM Temporal Modeling
4. Hidden Markov Model (HMM)
5. Actor Network (Policy Network)
6. Critic Network (Value Network)
7. Reward Computation and Optimization

Each component is described in detail below.

Speech Environment (Dataset)

The environment serves as the source of interaction for the reinforcement-learning agent. It consists of a speech dataset containing raw audio recordings used for training and evaluation. The environment provides speech signals as input and receives recognition outputs (actions) from the model. During training, the environment supplies state information derived from extracted speech features and provides reward signals based on recognition accuracy.

Feature Extraction Module

Effective feature extraction is critical for robust speech recognition. In this study, an enhanced MFCC-based approach is adopted to capture discriminative acoustic features. The feature extraction process includes:

Multi-Filter Bank Processing

Multiple filter banks were applied to the raw audio signal to capture diverse frequency components of speech. This improves representation of both low- and high-frequency characteristics.

Context-Aware MFCC

Mel-Frequency Cepstral Coefficients (MFCCs) are extracted while incorporating contextual information from neighboring frames. This context-aware representation enhances robustness to noise and improves feature continuity across time.

BiLSTM-Based Temporal Encoding

The extracted MFCC features are passed through a Bidirectional Long Short-Term Memory (BiLSTM) network. Unlike conventional LSTM networks, BiLSTM processes sequences in both forward and backward directions, enabling the model to capture long-range temporal dependencies and contextual relationships in speech signals. The BiLSTM output serves as a high-level representation of temporal speech dynamics and is forwarded to subsequent modeling components.

Hidden Markov Model (HMM)

The Hidden Markov Model is used to model the sequential and probabilistic structure of speech. HMMs are particularly effective for representing temporal transitions between phonemes or words. An HMM consists of:

1. A set of hidden states representing speech units,
2. State transition probabilities,
3. Emission probabilities,
4. Initial state probabilities.

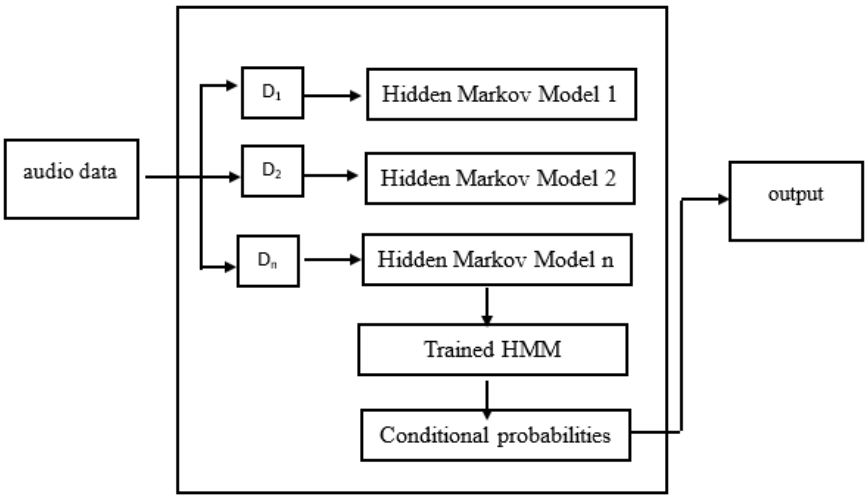

Figure 2: Hidden Markov Model

The model estimates the likelihood of observed feature sequences given hidden states. The Viterbi algorithm is employed to determine the most probable sequence of hidden states corresponding to the observed speech features. By integrating HMM with deep learning components, the system benefits from both probabilistic temporal modeling and discriminative feature learning.

Actor–Critic Deep Reinforcement Learning Framework

To further optimize recognition performance, an Actor–Critic Deep Reinforcement Learning architecture is integrated into the system using the Deep Deterministic Policy Gradient (DDPG) algorithm.

Actor Network (Policy Network)

The Actor Network maps extracted speech features to actions. In the context of speech recognition, actions correspond to predicted phonemes, words, or output tokens. The Actor learns a deterministic policy that selects optimal recognition outputs based on the current feature state.

Critic Network (Value Network)

The Critic Network evaluates the quality of the Actor's decisions by estimating the expected cumulative reward (Q-value) for each state–action pair. The Critic Network guides the Actor Network by providing gradient feedback, enabling policy improvement.

Reward Computation

A reward function is designed to evaluate recognition performance. The reward is computed based on the correctness of predicted outputs relative to ground truth labels. Higher rewards are assigned to accurate predictions, while penalties are imposed for misclassifications. This reward mechanism encourages the model to optimize recognition accuracy while minimizing errors.

Training and Optimization

The training process follows an iterative reinforcement learning framework:

1. Extract speech features using enhanced MFCC.
2. Model temporal dependencies using BiLSTM.
3. Estimate probabilistic state transitions via HMM.
4. Generate actions using the Actor Network.
5. Evaluate actions using the Critic Network.
6. Compute reward signals.
7. Update Actor and Critic networks using gradient descent.
8. Store experiences in a replay buffer for stable learning.

To stabilize training:

1. Target networks are employed.
2. Experience replay is used to break temporal correlations.
3. Exploration noise is added to Actor outputs to balance exploration and exploitation.

This iterative process continues until convergence criteria are satisfied.

Computational Efficiency Considerations

The proposed architecture is designed to be lightweight and deployable on limited hardware platforms, including mobile and embedded systems. Model complexity is controlled through:

1. Optimized network depth,
2. Gradient clipping,
3. Regularization techniques,
4. Controlled state-space design in the HMM.

This ensures a balance between recognition accuracy and computational cost.

Summary of the Proposed Framework

The proposed methodology integrates:

1. Enhanced MFCC for robust acoustic feature extraction,
2. BiLSTM for bidirectional temporal modeling,
3. HMM for probabilistic sequence modeling,
4. Actor–Critic Deep Reinforcement Learning for policy optimization.

By combining statistical modeling and deep reinforcement learning, the framework aims to achieve improved robustness, higher accuracy, and better adaptability across diverse speech environments.

RESULTS AND DISCUSSION

The Hidden Markov Model (HMM) is a probabilistic model commonly used for modelling sequential data, where the system being modelled is assumed to be a Markov process with hidden states. Speech signals are inherently sequential, with each segment of speech influencing the subsequent segments. HMMs model such sequential data by representing the probabilistic dependencies between successive elements in a sequence. The HMM model parameters were presented in Table 1.

Table 1: The HMM model parameters

Parameter	Typical Value Range
Number of States (N)	5
State Transition Probabilities (A)	Values between 0 and 1, with rows summing to 1.
Emission Probabilities (B)	Values between 0 and 1, with columns summing to 1.
Initial State Probabilities (π)	Values between 0 and 1, with elements summing to 1.

The model consisted of five states, which defined its complexity. This number of states captured the number of speech phonemes and words. While more states could capture finer details, they would also increase computational cost. The state transition probabilities governed the likelihood of moving from one sound unit (word) to another, reflecting the sequential nature of speech. The model used emission probabilities to capture the likelihood of observing a specific sound feature given a particular state. These probabilities essentially connected the hidden states (words) to the actual speech signal. The Initial State Probabilities (π) defined how probable it was to start the recognition process at any given state. The careful selection of these parameter values was crucial for the HMM to accurately model speech and achieve good performance in speech recognition systems. These parameter values and ranges are crucial in configuring the HMM model for the Speech Recognition System. They determine the hidden states, the emission probabilities, and the Initial State Probabilities of the system. Proper configuration of these parameters helps the system achieve effective training

Table 2: HMM model performance

Parameters	Values
misses	3
WER	0.28
CER	4.2
Misclassification error	0.009524
Accuracy	0.92
Precision	0.901
recall	0.892
f1	0.894

Table 2 presents the results obtained after simulating and training the HMM. The model recorded 3 missed instances, indicating the number of errors made by the algorithm. This metric is commonly used in speech recognition or natural language processing tasks, measuring the proportion of errors in the transcription of words. The Word Error Rate was 0.28, indicating that 28% of the words were transcribed incorrectly. The number of misclassified instances was 0.009524, or about 0.95%. This was the proportion of misclassified instances out of the total instances. The accuracy was 0.92, or 92%. A precision of 0.901 meant that 90.1% of the instances identified as positive were actually positive. A recall of 0.892 indicated that 89.2% of the actual positive instances were correctly identified. The F1 score was the harmonic mean of precision and recall, providing a single metric that balanced both precision and recall. The F1 score was 0.894. The high accuracy (92%) suggested the model performed well in general. The low misclassification rate (0.95%) indicated good performance on most instances. However, the high WER (28%) and F1 score (0.894) highlighted a significant number of errors in word classification. This suggested the model might have been good at classifying the overall category but struggled with

accurately identifying specific words within that category. These metrics collectively provided a comprehensive evaluation of the performance of the HMM model, with considerations for both the overall accuracy and the trade-off between precision and recall. The actor network processed speech signals to output recognized words using softmax activation units corresponding to the vocabulary. The critic network evaluated recognition quality based on both speech signals and recognized word inputs using a replay buffer to store experience tuples for training update. The actor and critic networks were optimized by minimizing temporal difference error and Q-value gradient respectively, and exploration was incorporated by adding noise to actor outputs, target networks were utilized for stabilization, a reward function was designed to provide feedback on recognition accuracy, and the networks were trained using DDPG by sampling from the replay buffer iteratively, with the implementation tailored based on vocabulary size, model complexity, and application requirements. The parameters of the DDPG model are shown in Table 3.

Table 3: The Deep DDPG Model Parameter.

Parameter	Value
Critic learning rate	0.001
Critic L2 regularization	0.0001
Critic gradient threshold	1
Actor learning rate	0.0001
Actor L2 regularization	0.0001
Actor gradient threshold	1
Target smooth factor	0.001
Discount factor	0.995
Mini-batch size	128
Experience buffer length	1000000
Noise variance	0.1
Noise variance decay rate	0.00001
Sample time	0.1
Max episodes	10000

Table 3, describes the training parameters for the DDPG model. The critic learning rate was set to 0.001. This determined the step size for updating the critic network's weights, allowing for larger updates and potentially faster learning. Additionally, a critic L2 regularization of 0.0001 was used to control the amount of regularization applied and prevent overfitting. The critic gradient threshold was also set to 1. This ensured that gradients during backpropagation in the critic network wouldn't become excessively large, stabilizing the training process. A learning rate of 0.0001 was used to determine the step size for updating weights based on the critic's feedback for the actor network. Similar to the critic network, L2 regularization of 0.0001 was applied to control complexity and prevent overfitting. An actor gradient threshold of 1 was set to limit the maximum gradient value during backpropagation for stability. The target smooth factor was fixed at 0.001, this determined the rate at which the target networks (critic and actor) were updated towards the main networks, promoting stability. The discount factor, set at 0.995 which influenced the importance of future rewards in the reinforcement learning algorithm. A mini-batch size of 128 was specified to determine the number of transitions sampled from the experience buffer for each network update during training. The experience buffer length was set to 1,000,000, defining its maximum capacity for storing past experiences. The exploration noise variance was set to 0.1 which determined the amount of noise added to the agent's actions. Additionally, a noise variance decay rate of 0.00001 was used to control the rate at which the noise variance decreased over time, signifying a transition from exploration to exploitation. The sample time was set to 0.1 seconds, specifying the time interval between consecutive actions taken by the agent in the simulation environment. A maximum number of episodes (training iterations) of 10,000 was defined in the past. For a more stable performance measure, a score averaging window of size 50 was used to compute the average reward over multiple episodes. The training stop criterion was set to "AverageReward," indicating that training would cease when the average reward reached or exceeded a satisfactory value of 400.

Table 4: Performance of the DDPG model

Parameter	Values
misses	1
WER	0.197
CER	3.22
Misclassification error	0.00508
Accuracy	0.94587
Precision	0.9604
recall	0.93
f1	0.96517

In Table 4, the performance of DDPG (Deep Deterministic Policy Gradient) was provided. The model recorded 2 misses, indicating that there were two instances where the system failed to recognize or misclassify the input. The WER obtained was 0.197, indicating that approximately 19.7% of the words were incorrectly recognized or transcribed. The model had a Misclassification Error of 0.00508, which meant approximately 0.508% of the instances were misclassified. The model accuracy was 0.94587, indicating that approximately 94.587% of the instances were correctly recognized. The model Precision was 0.9604, indicating that approximately 96.04% of the instances recognized as positive were actually correct. The model had a Recall of 0.93, indicating that approximately 93% of the instances that should have been recognized were actually recognized. The F1 score obtained was 0.96517, indicating a high balance between precision and recall. The DDPG model in speech recognition had a Word Error Rate of 19.7%: This indicates that the model still misses a significant number of words (almost 20%) in the speech input. While the accuracy is high (94.59%), it suggests room for improvement in accurately recognizing all the words. The misclassified instances (0.5%), although low, indicate that the Misclassification Error is low, it implies the model might confuse certain sounds or words, leading to errors The misclassified instances (0.5%), although low, indicate thatFrom the results obtained, the model may experience difficulty in capturing variations in speech accents. The model might not be robust to background noise, leading to misinterpretations in noisy environments. The model's performance might deteriorate with words it hasn't encountered during training. Hence the need to develop a more robust model to enhance the performance of the speech recognition system.

Table 5: Results for each model

Parameters	HMM	LSTM	BiLSTM	DDPG
misses	3	5	2	1
WER	0.28	0.26	0.23	0.197
CER	4.2	3.8	3.67	3.22
Misclassification error	0.00952	0.00769	0.0087	0.00508
Accuracy	0.92	0.92913	0.945	0.94587
Precision	0.901	0.9021	0.901	0.9604
recall	0.892	0.92054	0.92	0.93
f1	0.894	0.91122	0.909	0.96517

Table 5, Comparing the performance of different models for speech recognition based on various metrics reveals insights into their effectiveness. The Hidden Markov Model (HMM) exhibits a moderate performance with 3 misses and a Word Error Rate (WER) of 0.28. While its accuracy stands at 92%, its precision, recall, and F1 score are relatively lower compared to other models. Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) models show improvements over HMM, with lower WERs (0.26 and 0.23 respectively) and higher accuracies (92.913% and 94.5% respectively). However, their precision, recall, and F1 scores are not significantly better than HMM.

CONCLUSION

This study investigated the efficacy of a novel speech recognition algorithm that combines Deep Deterministic Policy Gradient (DDPG), BiLSTM architectures, and Hidden Markov Models (HMMs). The results of the experiment were encouraging, suggesting that the proposed approach is adept at capturing intricate speech patterns efficiently. The Deep Deterministic Policy Gradient (DDPG) model exhibits notable improvements over the previous ones, achieving 2 misses, a WER of 0.197, and an accuracy of 94.587%. It also demonstrates higher precision, recall, and F1 score, indicating a more balanced performance. The findings from this research illuminate the potential of voice recognition applications, and the proposed model

represents an appealing solution. The vital strength of this approach lies in its capacity for real-time learning and adaptation to diverse speech environments. This characteristic makes the model particularly valuable in scenarios where robust performance is paramount. By leveraging the complementary strengths of BiLSTM, DDPG and HMM, the proposed hybrid approach establishes a promising framework for further advancements in speech recognition technology. The successful implementation of this model paves the way for the development of even more sophisticated systems capable of handling the complexities of real-world environments. This study provides a foundation for future research endeavors in the field.

REFERENCES

Aljinu Khadar, M., Rahman, A., & Suresh, P. (2023). Gaussian mixture model–universal background model I-vector approach for speaker verification in noisy environments. International Journal of Speech Technology, 26(3), 455–468.

Babu, R., Kumar, S., & Reddy, V. (2023). A comprehensive classification of speech recognition approaches: Pattern recognition, acoustic-phonetic, and artificial intelligence methods. Journal of Signal Processing Systems, 95(4), 601–615.

Barkani, A., El Moutaouakil, K., & El Mohajir, M. (2023). Amazigh automatic speech recognition using the Kaldi toolkit. Speech Communication, 152, 45–57.

Biswas, T., Roy, S., & Chatterjee, A. (2023). Spoken language identification using MFCC features and machine learning classifiers. Expert Systems with Applications, 221, 119765.

Fadhel, M., & Mohammed, H. (2023). Classification and evaluation of automatic speech recognition systems. International Journal of Computer Applications, 185(12), 25–34.

Hazmoune, Y., Benyettou, M., & Ouni, K. (2024). An ensemble hidden Markov model approach for robust speech recognition. IEEE Access, 12, 33421–33435.

Isaac, S., Haruna, K., Ahmad, M. A., & Mustapha, R. (2023). Deep reinforcement learning with hidden Markov model for speech recognition. Journal of Technology and Innovation, 3(1), 1-5.

Kanke, S., Patil, A., & Joshi, R. (2023). Marathi speech recognition using language-specific acoustic modeling techniques. Procedia Computer Science, 218, 987–996.

Manideep, K., & Mohana, R. (2023). Voice recognition using hybrid Gaussian mixture model and hidden Markov model. International Journal of Intelligent Systems and Applications, 15(2), 112–124.

Mishra, D., Verma, P., & Singh, A. (2024). Comparative analysis of machine learning techniques for automatic speech recognition. Multimedia Tools and Applications, 83(5), 14321–14345.

Moondra, A., Jain, S., & Kulkarni, P. (2023). Modified MFCC-GMM approach for speaker recognition under degraded speech conditions. Applied Acoustics, 206, 109251.

Nugroho, H., Prasetyo, E., & Wibowo, S. (2023). Multi-accent speaker detection using normalized MFCC and neural networks. Neural Computing and Applications, 35(14), 10523–10536.

Ouisaadane, H., El Hannani, A., & Boulaknadel, S. (2024). Moroccan dialect speech recognition using PocketSphinx in noisy environments. Speech Communication, 160, 78–90.

Pavithran, P., & Sherly, E. (2024). Hidden Markov model-based automatic speech recognition system for individuals with hearing impairment. Biomedical Signal Processing and Control, 89, 105432.

Prabhu, S., & Jayasri, V. (2024). Hidden Markov model-based speech recognition system for vending machine applications. International Journal of Embedded Systems, 17(1), 65–76.

Ramadan, Z., & Bitmead, R. (2022). Gaussian mixture models and maximum likelihood estimation for speech recognition systems. Signal Processing, 196, 108511.

Sallagundla, S., Rao, P., & Krishna, M. (2023). Voice-enabled form filling system using hidden Markov models. Journal of Ambient Intelligence and Humanized Computing, 14(8), 10923–10935.

Santos, L., Pereira, J., & Almeida, F. (2023). Hybrid HMM-CNN architecture for improved automatic speech recognition. IEEE Access, 11, 91234–91248.

Shafieian, R. (2023). Persian speech recognition using hidden Markov models. International Journal of Speech Technology, 26(2), 233–245.

Sudarshan, R., Karthik, S., & Menon, V. (2023). Context-aware automatic speech recognition using semantic processing. Artificial Intelligence Review, 56(6), 4891–4910.

Thimmaraja, Y., Ramesh, H., & Kumar, N. (2024). Real-time Kannada continuous speech recognition using hidden Markov models. International Journal of Speech Technology, 27(1), 101–115.

Tsai, C. H., & Wang, Y. T. (2023). Hardware-efficient Gaussian mixture model-based speaker verification system using MFCC features. IEEE Transactions on Circuits and Systems II: Express Briefs, 70(9), 3150–3154.

Wirdiani, A., Santoso, D., & Prabowo, R. (2024). MFCC-CNN with online triplet mining for robust speaker recognition. Expert Systems with Applications, 235, 120123.