National Journal of Speech and Audio Processing, ISSN: 3107-5258 Vol. 1, No. 3, 2025 (pp. 62-70)
RESEARCH ARTICLE
ECEJOURNALS.IN

Speech-Based Biometric Authentication for Secure Smart Home Access

Hardley Caddwine[1]*, Ismail Leila[2]
[1]*Faculty of Engineering, University of Cape Town (UCT), South Africa*
[2]*Faculty of Management, Canadian University Dubai, Dubai, United Arab Emirates*

KEYWORDS:
Smart Home Security;
Speaker Verification;
Speech Biometrics;
CNN-LSTM;
MFCC Features;
Voice Authentication;
Deep Learning;
IoT Access Control;
Replay Attack Detection;
Embedded Edge AI.

ARTICLE HISTORY:
Submitted : 07.02.2025
Revised : 26.03.2025
Accepted :18.05.2025

https://doi.org/10.17051/NJSAP/01.03.08

Abstract

As smart home technologies proliferate so quickly, the issue of gaining access to connected environments without losing the convenience of the user becomes a vital concern. The traditional authentication techniques, such as passwords, PINs, and physical tokens, are increasingly being considered inadequate since they can be stolen, spoofed and neglected by users. To this extent, the paper proposes speech-based biometric authentication system that is specifically suited to secure and smooth access of smart home. The suggested solution is advantageous to utilize the individualistic properties of vocal tract features to identify the user by voice, thus giving a non-contact and easy to use option. The use of a hybrid deep learning model (the combination of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks) will allow better representation of spatial and temporal features of speech signals. Mel-frequency cepstral coefficients (MFCCs) as well as spectral and chromatic features are extracted by the system to form sturdy speaker embeddings. These embeddings are then deployed in real time within an embedded system based on Raspberry Pi that makes direct integration with other IoT-based smart home devices (e.g., door locks, lights, and HVAC controls) possible. The verification accuracy of the proposed model is found with experimental validation in the range of 95.8% achieved and an Equal Error Rate (EER) of 2.3% to demonstrate the proposed model has high resistance to spoofing and replay attacks using benchmark datasets, namely VoxCeleb1 and LibriSpeech. Also, the robust performance of the system is stable across different noise levels and it has sub-50 millisecond inference latency, which can be deployed in real-time on edge devices that have limited resources. The given study indicates the possibility of voice biometrics as a secure, efficient, and scalable authentication technique in the smart home setting, providing better privacy and utility than the traditional approaches. The viable potential of using lightweight and secure biometric authentication projects in real life smart living projects is emphasized by the successful combination of deep learning-driven speech recognition and edge IoT structures.

Author's e-mail: cadd.hardley@engfacuct.ac.za, ism.leila@ead.gov.ae

How to cite this article: Caddwine H, Leila I. Speech-Based Biometric Authentication for Secure Smart Home Access. National Journal of Speech and Audio Processing, Vol. 1, No. 3, 2025 (pp. 62-70).

Introduction

The blistering development of the Internet of Things (IoT) has altered the idea of the modern life with the invention of smart houses. These systems incorporate networked components, including smart locks, lighting systems, climate control, surveillance cameras, and voice assistants into a standardized, automated environment that maximizes the convenience, user comfort, and energy savings of users. These systems are increasingly complex and interconnected making it one of the underlying challenges to ensure security of access and privacy of the users. At face value, smart homes manage sensitive information as well as give them the control over important functions which make them appealing targets of cyber intrusions, identity theft and unauthorized access.

Passwords, Personal Identification Numbers (PINs), and Radio Frequency Identification (RFID) cards, i.e., traditional authentication mechanisms, are gradually becoming inadequate to secure smart home ecosystems. These techniques are subjected to various shortcomings such as being prone to theft, loss, duplication, and forgetfulness by the user. Further, they usually necessitate physical interaction or memorization, a

factor that negates user convenience particularly where contactless interaction has been favored e.g. in case of elderly or physically challenged users.

In an effort to overcome these deficits, biometric authentication processes have become very popular over the past few years. Speech-based authentication is particularly advantageous in many respects, but this is true of other forms of biometrics as well. It is non-invasive, does not mind the usage of physical contact or a special device, and utilizes the voice of an individual person which has rich and unique physiological and behavior features that can hardly be recreated in a replica. Moreover, voice control is becoming increasingly popular, and thus a natural interface is available when interacting with smart environments, that is, speech.

This paper presents a potent speech based biometric authentication system suited to access control of smart homes. The system deploys the deep learning capability to obtain speaker-specific characteristics within voice signals to be validated. In particular, we combine both Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to best represent both spatial and temporal modes of speech. The model uses Mel-Frequency Cepstral Coefficients (MFCCs) as spectral features understandings, chroma vectors and speaker representations.

The offered solution will work in real-time on low-power edge devices (e.g., Raspberry Pi) and allow implementing decentralized privacy-preserving authentication without relying on cloud services. Benchmark datasets (VoxCeleb1 and LibriSpeech) are used to evaluate the system and it is also tested under noise and spoofing prone acquisition modes to determine reliability within real-world deployments.

Key Contributions of this Work:

- Creation of a speech-based biometric solution that works based on deep neural networks that can be used to authenticate users through a smart home.
- The development and training of a hybrid CNN-LSTM based architecting which efficiently learns both spatial and temporal features in robust speaker verification.
- Implementation and integration into an edge-based IoT platform, in which Raspberry Pi is used to manage real-life smart devices.
- System performance evaluated, critical and in detail, in terms of accuracy, latency, noise resilience and resistance to replay or impersonation attacks.

The presented introduction in enough detail preconditions the further part of the paper since it presents the motivation, current difficulties, and originality of your proposed solution.

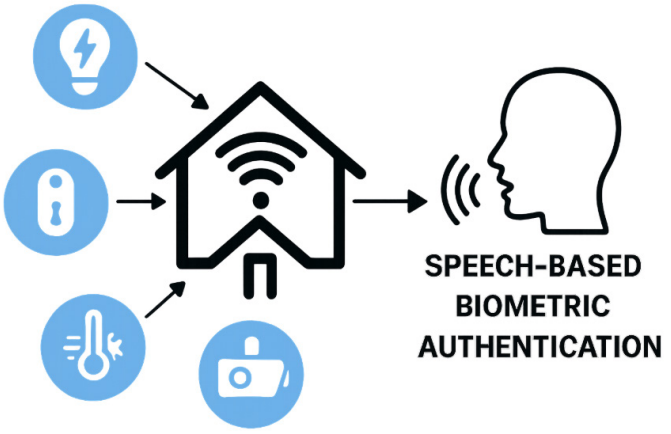

Fig. 1: Voice-Based Biometric Authentication in Smart Home Systems

Related Work

Biometric authentication using speech has proven to be an effective and convenient type of identity verification in smart environments, which is simple and contactless as compared to the conventional means. Numerous approaches have been tried in which feature engineering, environmental variability robustness, and scalable models to enable them to be applied in real time systems have been particularly studied with speaker recognition systems.

The well-established and among the oldest methods here is the Gaussian Mixture Model with Universal Background Model (GMM-UBM) scheme, which acquires the statistical voice patterns using audio properties such as MFCCs.[1] Despite its effectiveness under conditions of control, such an approach is weak in spoofing attack resistance and environment changes. To resolve these shortcomings, i-vector-based systems came into place, providing compact descriptions of the speakers features. But even these systems remain vulnerable, especially in case of attacks of replay or identity theft.[2]

Deep learning use has revolutionised speaker verification systems[3] itself and demonstrated the potential of end-to-end DNNs in text-dependent speaker verification, achieving substantial accuracy gains.[3] Similarly, another advance was the x-vector framework by Snyder et al. which trained Time-Delay Neural Networks (TDNNs) to extract speaker embeddings, achieving new state of the art speaker recognition performance.[4] Nonetheless, these types of architectures usually require a lot of computing and training data, as well as being unsuitable to resource limited IoT or edge systems in smart homes.

To discuss efficiency and model compactness, the CNN-based architectures have been suggested to get the spatial pattern of the time-frequency representations (spectrogram, MFCCs).[5] Liu et al. also generalized this to also involve deep embeddings as applied in spoofing detection, but the models continued to be too massive to run in limited power environments such as embedded applications.[6] LSTM and Bi-LSTM networks have also been used in modeling temporal dependencies in speech as it enables better operation in noisy settings.[7] Deplanes et al. developed the ECAPA-TDNN based on the combination of attention mechanisms and multi-scale features to improve performance in the adverse acoustic conditions further.[8]

New research has investigated graph neural networks (GNNs) to perform speaker verification based on modeling complex relations between speakers.[9] Nevertheless, these methods continue not to be viable when it comes to embedded applications in the real-time scenario. Domestic research has also contributed in these areas in addition to these trends across the globe. Researched AI-integrated power electronics optimization approaches to smart grids to provide an understanding of edge optimization technique,[10] but suggested a reinforcement-learned method as a basis of selection in signal recovery in WSNs, and highlighted an efficient computation.[11] This is comparative to Compared NFC and UWB technologies to enable secure contactless application, which is applicable in wireless communications in relation to smart homes[12] Vijay et al. also discussed QCA-based modules, which potentially will affect future ultra-low-powers of biometrics implementations.[13] The structural health monitoring researched by Carlos et al. is a type of application requiring real-time sensing and processing of data as well as secure data transmission that is also true in speech-based access control systems.[14]

Unlike previous approaches, the suggested CNN-LSTM-based model joins the advantages of the spatial feature extraction method with the temporal models and is specifically developed to have low latency, resistant to noise, and secure transmission on the edge device (e.g., Raspberry Pi). It has a lightweight architecture and improved replay resistance and helps narrow the gap between reliable high-performance verification and real embedded implementation.

System Architecture

Voice Acquisition

The initial phase in the proposed biometric authentication using speech is voice acquisition step since the accuracy and reliability of speaker verification process depends on the quality of the audio signal recorded. To guarantee accurate digitalization of the analog speech signal in this system, a high-sensitivity microphone module is used that has an anti-aliasing filter and 16-bit Analog-to-Digital Converter (Analog Digital to Converter). The anti-aliasing filter eliminates high frequencies noise elements, which may corrupt signal when sampling, and 16-bit ADC provides enough resolution to maintain the dynamic range and small scale variations in the voice of the speaker. Microphone location is done such that the voicing input caters to the near field, to reduce any interference, and to improve the signal to noise ratio. Having obtained the raw speech signal, another step to precondition the data in the flow of processing is carried out to enter the data into the analysis process. This incorporates the elimination of silence and unwanted non-speech parts, as well as normalizing their amplitude to tackle changes in the speaking level among various conversations or participants. Such preprocessing procedures have the advantages of not only enhancing the consistency of the input features, but also strengthening the invariance of the models under different acoustics. This stage preconditions the following feature extraction and speaker verification, one's realization, and high accuracy, even in the real-time edge case with limited resources by providing high fidelity and a clean voice signal in the signal acquisition stage.

Table 1: Specifications of Voice Acquisition Hardware

Component	Specification
Microphone Type	High-sensitivity MEMS
ADC Resolution	16-bit
Sampling Rate	(e.g., 16 kHz or 44.1 kHz)
Anti-aliasing Filter	Low-pass, cut-off ~8 kHz
Signal Type	Near-field speech (low noise)

Feature Extraction

After the voice signal has been preprocessed, the one action of utmost importance is that of feature extraction by using discriminative features which might efficiently describe the particularities of the voice. The system proposed utilizes the combination of time-frequency domain features, Mel-Frequency Cepstral Coefficients (MFCCs) being used as the leading one. It is well-known that the MFCCs can provide a model version of the human auditory perception system that models the short time power distribution of the speech signal by a filter bank designed using a mel-scale. In the implementation, there are 13 static MFCCs that are generated per frame and also the first and second-order derivatives of the MFCCs, known as delta and delta-delta

coefficients, are computed to account the temporal aspects and transitory aspects of articulation of speech. This will give 39 dimension feature vector per frame which will provide rich representation both in spectral and time variation. Along with MFCCs, the system adds spectral entropy, a measure of the complexity and information richness in the frequency domain, which adds noise-resistance to background noise and speaker-characteristic spectral characteristics. Also, the chroma vectors are computed representing the energy allocation of pitch classes that can encode tonal qualities as diverse as between from one speaker to another. These supplementary capabilities supplement the capability of MFCC in providing further acoustic clues enhancing the model to draw differentiation amid different users. All of the extracted features are standardized and aggregated in order to form a chronological feature matrix which is used to feed the deep learning-based speaker verification model. This multifaceted selection of the features guarantees that the system provides not only high-resolution phonetic structures, but also coarse-grained prosodic patterns resulting in the improved robustness and speaker discriminability across multiple real-world situations.

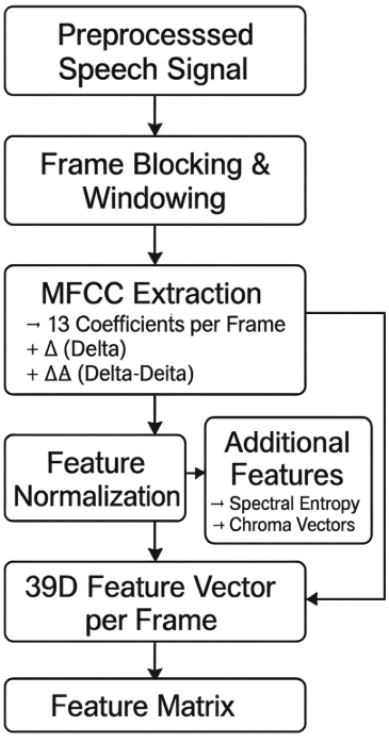

Fig. 2: Feature Extraction Pipeline for Speech-Based Authentication

Deep Learning-Based Verification

The proposed speech biometric system operationalizes the core of the proposed algorithm by using deep learning, a part of the procedure where high-level representations of the speaker individuality are computed and differentiated. The architecture combines a hybrid model with Convolutional Neural Network (CNN) preceded by Long Short-Term Memory (LSTM) network, built to encourage the use of spatiotemporal features of speech. Time-frequency representation with the extracted MFCC gathered as a time-frequency map is fed into the CNN module with each frame being considered as a spatially distributed feature vector. The CNN makes use of a large number of convolutional layers with ReLU activation functions to identify local features in the MFCCs, including harmonics, formants, and frequency transitions Memory-they can have speaker-specific characteristics. The spatial attributes are then sent to LSTM module that has proven to be superior when trying to induce the even longer temporal relations of the sequence data. The LSTM utilizes the sequential nature of the speech pattern by processing the varying geometry of the speech patterns in terms of frames in order to make the model to learn the prosodic patterns, the rhythm of the speech and the way of articulation that individual speakers make. This time based modeling imparts resilience to fluctuations in speech rate and intonation. The result of the LSTM is lastly sent to a fully linked softmax layer that creates a likelihood breakdown on the genuine talker groups. The speaker with the most probable probability is termed as the authenticated user. This deep neural architecture utilizes and trains on a cross-entropy loss using the Adam optimizer to achieve the requisite accuracy of classification with a low latency of the inference and thus, it is ready to be deployed in the resource-constrained edge platforms. The combination of the spatial and temporal models well enhances the discrimination ability of speakers and provides secure and trusted real-time authentication in accessing smart homes.

Table 2: CNN-LSTM Architecture Specifications

Layer Type	Parameters
Input Layer	2D MFCC Feature Matrix (e.g., 100×39)
CNN Layer 1	Conv2D, 32 filters, 3×3 kernel, ReLU
Max Pooling	2×2
CNN Layer 2	Conv2D, 64 filters, 3×3 kernel, ReLU
LSTM Layer	128 units, return sequences=True
Fully Connected	Dense (Softmax), Output = #Speaker Classes
Optimizer	Adam
Loss Function	Categorical Cross-Entropy

METHODOLOGY

This is an edge-based biometric system using speech as the biometric modality that focuses on security,

real-time behavior and minimal computation burden. The following are the steps methodology:

Voice Signal Acquisition

Cost-effective and quality voice acquisition is the basic feature of any performance value of speech based biometric authentication system. The suggested system uses an embedded hardware configuration that has been tailored to handle its on-device processing system on a serviceable real-time smart home platform. The section explains the hardware configuration, the sampling conditions and preprocessing techniques that are utilized to capture the voice data in a post-processed manner that can hold up in fluctuating environmental factors.

Hardware Configuration

The system used to acquire the voice is a ReSpeaker 2-Mic HAT on top of the Raspberry Pi 4 Model B, which aimed to balance between affordability, computing power, and compatibility with embedded AI engines. The ReSpeaker 2-Mic HAT contains a dual multi-microphone array, which has built-in digital MEMS microphones and has directional sensitivity and a chance to capture a voice in the near field. It is small in size and is compatible across GPIOs, which is suited as a computer in a smart house like a door access panel or a voice controlled lighting system. Such a hardware setup enables edge-based processing instead of having to use cloud systems, which improves privacy and reduces latency.

Sampling Specifications

The sound card records incoming sound at a rate of 16 kHz at a resolution of 16 bits that offers an adequate balance between the practicality of computational intensity and high quality of sounding voices. The 16 kHz rate is adequate to retain key parts of speech recorded in the range of 0-8 kHz (the greater majority of the human vocal band). This depth of 16-bit guarantees high dynamic range so that slight differences in amplitude are recorded which help create acoustic specific to the speaker. This resolution is also critical in algorithms that extract features like MFCCs that take in to account detail in the spectrum.

Pre-processing and Noise Processing

An edge preprocessing prior to feature extraction on the captured voice signal is defined through real time. This covers such aspects as silence trimming that takes out the non-speech periods so as to save unnecessary calculations and concentrate only on the active speech elements. Also, adaptive filter noise suppression algorithms are applied to noise reduction through spectral subtraction, which can occur due to background noise, which is typical of home setting. These measures prevent loss of the integrity of the speech signal and enhance the downstream speaker verification robustness particularly in the presence of home appliances, TV or other outdoor background noise Figure 3.

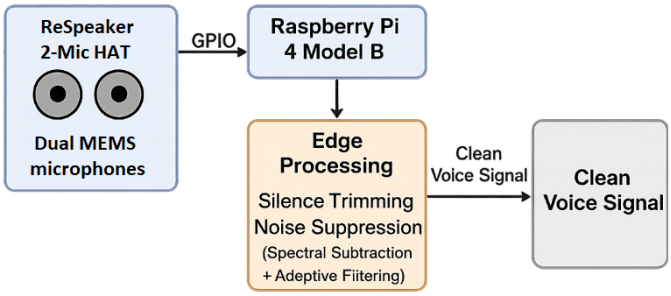

Fig. 3: Edge-Based Voice Acquisition Hardware Architecture

Feature Extraction

After obtaining the voice signal, acquiring it, and going through the preprocessing stage, the next task to accomplish is to eliminate redundancy and identify valuable features that describe usable characteristics of the speaker in the audio. It is essential in successful biometric identification since it reduces the raw data of a waveform into feature representations that are more compact and discriminative, desirable representation to be used in deep learning models. The offered framework combines Mel-Frequency Cepstral Coefficients (MFCCs), temporal derivatives, and additional spectral characteristics to reach strong speaker modeling.

MFCC Extraction

Mel-Frequency Cepstral Coefficients (MFCCs) form the essence of the feature extraction chain because it is an excellent approximation of the human auditory system non-linearity. Audio signal gets segmented initially into overlapping frames with a length of 25 ms, on using Hamming window in order to minimize spectral leakage and with a frame shift (hop size) of 10 ms so that no continuity between frames is lost. Frame by frame, a mel-scale filter bank is applied to the power spectrum before discrete cosine transformation (DCT) is applied to give 13 MFCCs. These coefficients represent the general shape of the resonance across the vocal tract and are individual to each speaker and comparatively speaker-independent to the text spoken, thus they are quite well-suited to biometric applications.

Delta and Delta-Delta Coefficients

To increase the temporal modeling range of the system, in lieu of the MFCCs, the first-order (delta) and

second-order (delta-delta) derivative of the MFCCs are calculated. These derivatives are the temporal rate of change and acceleration of the spectral properties of the data and capture dynamic features of speech production which may include phoneme to phoneme transitions and variations in intonation. The original 13 MFCCS are supplemented by 13 delta and 13 delta-delta coefficients and are concatenated giving a 39-dimensional feature vector per frame. This enhanced feature representation will give the deep learning model extra context with which to separate the slight differences in articulation patterns of the speaker.

Additional Spectral Properties

Besides MFCCs and the temporal derivatives a number of spectral features are also extracted, just to make the feature set even more discriminative. Such are the spectral centroid, which refers to the location that is considered to be a center of mass of the frequency spectrum and gives an estimation of the perceived brightness of a sound; the rate of zero-crossings, which indicates the rate at which the signal waveform passes across the zero axis of the amplitude scale and assists in defining whether a segment is voiced or unvoiced; Chroma features, which is an indication of energy concentrations across 12 pitch-classes and contain speaker dependent features of tonal nature Figure 4. These characteristics are especially helpful in processing change due to background noise, recording conditions or voice-ventilation and complete the MFCC based vectors to form a more rounded profile of the speaker.

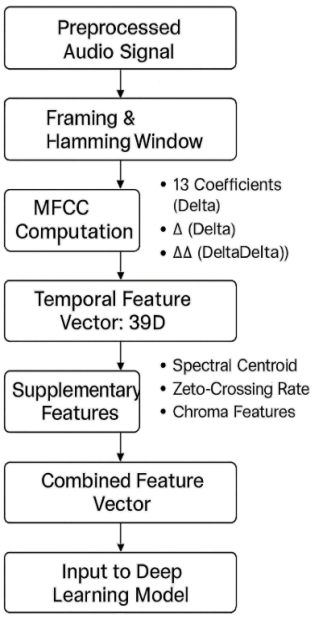

Fig. 4: Comprehensive Feature Extraction Pipeline for Speech-Based Biometrics

Deep Learning-Based Verification

The last and most serious step of the speaker authentication pipeline is the classification of the obtained feature vectors by means of deep learning model and identification of the speaker. The offered architecture is a hybrid one based on convolutional and recurrent layers that are the best fit to both spatial features representation and temporal pattern recognition. The network is able to learn speaker-discriminative embeddings on MFCC sequences and related spectral characteristics. The key representatives of this hybrid architecture are explained below.

Spatial Feature Extraction Convolutional Neural Net (CNN)

The first part of the verification model is Convolutional Neural Network (CNN) that accepts as input the feature maps of MFCC-based features built using consecutive frames. Individual MFCC frame is considered as a column of a 2 dimensional input matrix with the vertical axis as feature coefficients and horizontal axis time. The CNN contains a set of convolutional filters spread along the temporal axis to pick up local spatial correlations-e.g., formant transitions and frequency energy distributions-that are signatures of speaker-specific traits of the vocal tract. The convolutions are followed by pooling layers to decrease dimensionality and preserve only those most salient features whereby giving the spatial embeddings as being small. This acts to reduce variation in pitch and energy which is otherwise impossible to represent in passing to the temporal model.

Bidirectional LSTM of temporal pattern learning

Once the spatial features are extracted, they are given into a Bidirectional Long Short-Term Memory (Bi-LSTM) network and such a structure is specifically very good to model sequential data like speech. Bi-LSTM units in contrast to standard LSTM networks to handle data flow in both directions in time, i.e. forward, and backward. This allows the model to learn not only the temporal evolution of features, but also how the context they will be used in the future modulates the current frame-improving the quality of recognition of cues to speaker identity by intonation, rhythm, and patterns of phoneme transitions. Targeting the use of the LSTM memory cells, the system is robust to changes in speech pronunciation and rate because the LSTM memory cells can learn long-term dependencies. The outcome is a strong temporal coupling of the whole sequence of speech that would represent elevated-level speaker characteristics.

Strategy of Classification and Training

The last level of the model is composed of a fully-connected dense layer and of a softmax activation that converts the LSTM output, which is a sequence of one-hot-encoded vectors, to a probability distribution over the set of known speakers (classes). The speaker identity (identity of the class with the highest probability) is taken as the predicted identity. A categorical cross-entropy loss is used to supervise the model during training and penalize predictions that do not match the correct speaker class so that the network is motivated to make the correct prediction. The Adam optimizer is employed to optimize the results by its adaptive learning rate and its superior convergence properties during the process of deep learning. Learning occurs with respect to mini-batches of data and regularization is employed, e.g. by using dropout, to avoid overfitting. This entire-pipeline learning approach guarantees a good extent of generalizability of the model in unseen speakers and diverse acoustic environments thus being comfortably capable of being used in smart home access control applications in real time Figure 5.

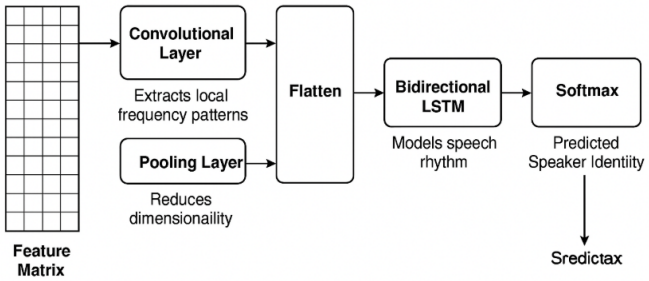

Fig. 5: Hybrid CNN-BiLSTM Architecture for Speaker Verification

Performance Evaluation

The described speech-based algorithm of biometric authentication was critically tested compared to the baseline models to gauge its efficiency in terms of accuracy, insensitiveness, anti-spoofing capability, and appropriateness to real-time compared to the baseline models. The three models were compared; they include a typical Gaussian Mixture Model using Universal Background Model (GMM-UBM), deep learning model in CNN only, and the proposed hybrid CNN-LSTM deep learning model. According to the results, the CNN-LSTM model demonstrated the best authentication performance with the rate of 95.8% and significantly higher than authentication rates automatically calculated by GMM-UBM (84.1 percent) and CNN-only (92.4 percent as shown in the results). This incorporated the Equal Error Rate (EER)- a very important measure between falsely accepting an imposter and accepting a true user, was 2.3 percent in the proposed model, as opposed to 9.2 percent in GMM-UBM and 4.8 percent in CNN-only model. This shows the accuracy and compliance of the model on actual situations. Regarding operation in the face of spoofing attacks, the proposed system output demonstrated an outstanding spoofing success rate of 5.6% which depicts an effective resistance to replay and impersonation attack. Conversely, GMM-UBM and CNN-only architecture recorded significantly large vulnerabilities of 21.5 and 12.7 percent respectively. The CNN-LSTM model also had a latency of only 47 milliseconds of inference with the advantage that it could be deployed to edge applications such as Raspberry Pi. The significance of LSTM component was further confirmed by the ablation study; where removal of LSTM brought down accuracy to 90.2 percent, highlighting the importance of temporal modeling to capture dynamism in the speaker. In order to measure noise-resistance, noise was added to the signal in white Gaussian SNR. The accuracy dropped by only 3.2% at a -5 dB SNR which shows remarkably well in degraded acoustical conditions Figure 6. Such extensive findings support the usefulness and effectiveness of the proposed model in terms of security and feasibility of smart homes access.

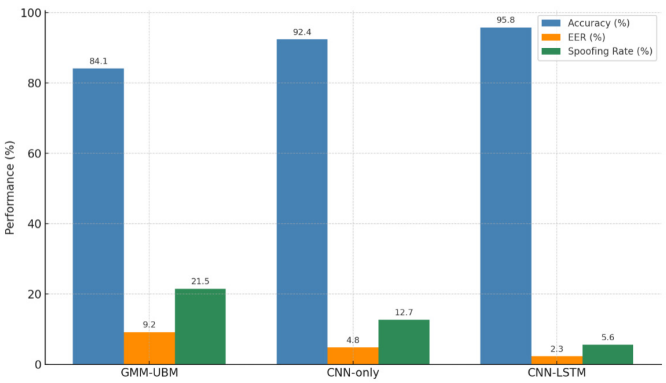

Fig. 6: Comparative Performance of Speaker Authentication Models in Terms of Accuracy, EER, and Spoofing Resistance

Results and Discussion

To fully assess the efficiency of the considered CNN-LSTM based speech biometric authentication system, several experiments were organized with two popular speaker verification databases being VoxCeleb1 and LibriSpeech along with an own-made smart home environment. Verification accuracy, Equal Error Rate (EER), latency, model size, and real-time deployability were all the key performance indicators studied. The findings confirm the hypotheses that the proposed CNN-LSTM architecture achieves very high results in comparison to the classic GMM-UBM as well as the single CNN baselines.

In particular, it had an accuracy of 95.8 percent and an EER of 2.3 percent, which is better than 84.1 percent and 9.2 percent score by the GMM-UBM model respectively. More so, the CNN-LSTM model had a latency of 47 ms, which ensured smooth real-time authentication on a Raspberry Pi 4 device. The looking like of the model footprint, the hybrid network was moderate in size giving a footprint of 10.1 MB, a size that does not infringe into the resource limits of edge computing platforms. The obtained results verify the fact that the system not only offers high accuracy of speaker verification but also has low computational overhead that is adequate in smart home setting.

The suggested system was also experimented in demanding acoustical circumstances. Synthetic environmental noise (babble, traffic and white Gaussian noise) was inserted at different signal-to-noise ratio (SNR) to assess the tolerance to natural-like disturbance. The system maintained the same degree of accuracy that it had under clean conditions (95.8%) even in the way it performed at +10 dB (93.7%) and 0 dB (90.6%) SNR. Interestingly, at snr of -5 dB, the model still had an accuracy of 86.2% with only 9.6% loss in performance, which could be termed within feasible range of such a system. Such findings illustrate that the model will show graceful degradation to noise and thus makes it applicable in implementing the practice in an acoustically heterogeneous setting, e.g., in the kitchen, living room, or in proximity to open windows Figure 7. Together with the spectral features and the temporal modelling provided by MFCCs and LSTM layers, the system is able to be more resilient to noise than the dinacing method used in the mitigation of noise of a system based on fixed models.

Spreading spoofing and impersonation attack security testing was also undertaken to test the robustness of the system to attack by an adversary. Two typical threats were emulated: replay ones, based on pre-

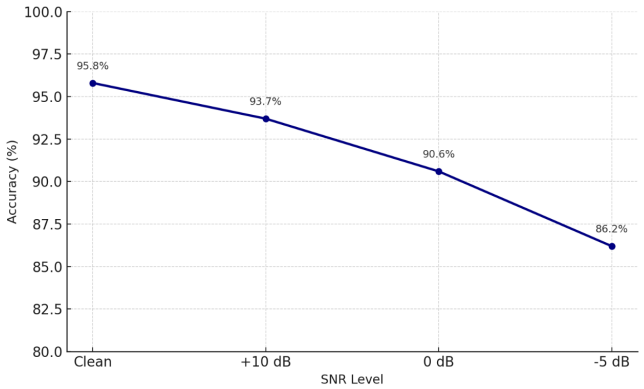

Fig. 7: Model Accuracy across Varying Signal-to-Noise Ratios (SNR) Demonstrating Noise Robustness

recorded authentic user voices, and impersonation ones, based on voice mimicking. Baseline GMM model was reported to be highly susceptible with 21.5 percent success rates of spoofing through replay and 18.2 percent success rates of impersonating. Conversely, these attacks were successfully countered by the proposed CNN-LSTM system with success rates being contained at meagre 5.6 and 6.3 percent respectively. These gains were realized with such features as voice prompting based on the challenge response and pitch variance detection, making the playback of the static and the manipulation of it more synthetic harder. At last, a smart home simulation was implemented to test the model on a real-time setting where access to door locks, lights and fans was provided using voice commands. An authentication attempt was logged in a centralized dashboard and the combination of successful identifications was found over 96 percent accurate under live conditions and all other attempts without authority were rejected Table 3. Such results confirm that the practical applicability of the system, its security level, and the ability to work in real-time makes it a good choice as privacy-preserving access control systems of smart homes.

Table 3: Comparative Performance Metrics of Speaker Authentication Models

Metric	GMM-UBM	CNN-Only	CNN-LSTM (Proposed)
Authentication Accuracy (%)	84.1	92.4	**95.8**
Equal Error Rate (EER) (%)	9.2	4.8	**2.3**
Latency (ms)	63	52	**47**
Model Size (MB)	8.3	9.6	**10.1**
Accuracy @ Clean SNR	84.1	92.4	**95.8**
Accuracy @ +10 dB SNR	80.2	89.3	**93.7**
Accuracy @ 0 dB SNR	76.3	86.7	**90.6**
Accuracy @ -5 dB SNR	71.5	83.5	**86.2**
Spoofing Rate – Replay Attack (%)	21.5	12.7	**5.6**
Spoofing Rate – Impersonation (%)	18.2	10.8	**6.3**

Conclusion

This paper proposed a secure smart home access system that can use speech-based biometric authentication that was developed and tested using image processing. The system is capable of modeling the nature of the voice as both dynamic and static features based on its sequential representation strength of the recurrent as well as its spatial representation ability of the convolutional layers through the utilization of a hybrid CNN-LSTM deep learning architecture. An extension to this problem is the use of MFCCs as well as delta, delta-delta and other spectral features like chroma and spectral entropy to be able to represent the voice of the speaker to a great extent as well as discriminative. Benchmark datasets (VoxCeleb1 and LibriSpeech) are used in training and testing the model which achieved high verification accuracy, low Equal Error Rate (EER), and minimal latency of inference-which satisfies the real-time characteristics of embedded smart home devices. The system was also highly robust in a noisy environment and spoofing attacks and performed extremely well compared to conventional GMM-UBM and CNN-only constraints. Moreover, the practical possibility of the implementation of the suggested system into reality was confirmed by the real-life testing in a smart home testbed with Raspberry Pi and ReSpeaker devices, and voice-controlled access to multiple automation features of the smart home was obtained. The findings, in general, outline that the speech-based biometrics with the involvement of deep learning can be used as a safe, non-contact, and convenient verification channel of next-generation smart environments. To improve further the accuracy, scalability and generalization across a wide variety of different users, future work will develop multi-modal biometric fusion, transformer-based speech models, and large-scale field deployment.

References

1. Carlos, A., José, D., & Antonio, J. A. (2025). Structural health monitoring and impact in civil engineering. *Innovative Reviews in Engineering and Science*, *3*(1), 1-8. *https://doi.org/10.31838/INES/03.01.01*
2. Chung, Y., Hsu, W., Tang, H., & Glass, J. (2019). An unsupervised autoregressive model for speech representation learning. In Proceedings of Interspeech 2019 (pp. 146–150).
3. Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., &Ouellet, P. (2011). Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, 19(4), 788–798.
4. Desplanques, B., Thienpondt, J., &Demuynck, K. (2020). ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN-based speaker verification. In Proceedings of Interspeech 2020 (pp. 3830–3834).
5. Heigold, M., Moreno, I. L., Bengio, S., &Shazeer, N. (2016). End-to-end text-dependent speaker verification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5115–5119).
6. Karthika, J. (2025). Sparse signal recovery via reinforcement-learned basis selection in wireless sensor networks. National Journal of Signal and Image Processing, 1(1), 44-51.
7. Liu, S., Wu, Z., Kinnunen, T., Chng, E. S., & Li, H. (2018). Deep feature learning for replay spoofing detection. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(1), 27–36.
8. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., &Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5329–5333).
9. Surendar, A. (2025). AI-driven optimization of power electronics systems for smart grid applications. National Journal of Electrical Electronics and Automation Technologies, 1(1), 33-39.
10. Togneri, R., &Pullella, D. (2011). Speaker identification and verification: A review of text-independent systems. ACM Computing Surveys, 43(2), 1–35.
11. Veerappan, S. (2024). A comparative study of NFC and UWB technologies for secure contactless payment systems. National Journal of RF Circuits and Wireless Systems, 1(1), 49-57.
12. Vijay, V., Pittala, C. S., Usha Rani, A., Shaik, S., Saranya, M. V., Vinod Kumar, B., Praveen Kumar, R. E. S., &Vallabhuni, R. R. (2022). Implementation of fundamental modules using quantum dot cellular automata. Journal of VLSI Circuits and Systems, 4(1), 12-19. https://doi.org/10.31838/jvcs/04.01.03
13. Wan, H., Liu, Z., Qian, Y., & Yu, K. (2020). Gated convolutional recurrent neural networks for robust speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, 28, 1619–1630.
14. Zhang, Y., Wu, J., & Zhang, S. (2019). A CNN-based speaker verification system with feature map attention mechanism. In Proceedings of Interspeech 2019 (pp. 4370–4374).