

---

## **EMBEDDED EDGE-AI SPEAKER VERIFICATION FOR SMART DOOR ACCESS CONTROL USING ECAPA-TDNN: A RASPBERRY PI PROTOTYPE**

---

**\*<sup>1</sup>Shehu Masiud Adekunmi, <sup>2</sup>S. M. Yarima, <sup>3</sup>N. Mohammed, <sup>4</sup>S. A. Yusuf, <sup>5</sup>I. T. Balami**

---

*1Department of Mechatronics and System Engineering Abubakar Tafawa Balewa University  
Bauchi, Nigeria.*

*2Department of Electrical Engineering Abubakar Tafawa Balewa University Bauchi, Nigeria*

*3Department of Mechanical Engineering Abubakar Tafawa Balewa University Bauchi,  
Nigeria*

*4Center for Embedded AI and Smart Energy System Abubakar Tafawa Balewa University  
Bauchi, Nigeria*

*5Department of Electrical Engineering, Abubakar Tafawa Balewa University, Bauchi,  
Nigeria.*

---

**Article Received: 29 August 2026**

**Article Revised: 17 September 2026**

**Published on: 07 October 2026**

**\*Corresponding Author: Shehu Masiud Adekunmi**

Department of Mechatronics and System Engineering Abubakar Tafawa  
Balewa University Bauchi, Nigeria.

DOI: <https://doi.org/10.1555/ijrpa.4553>

---

### **ABSTRACT**

*Voice-based smart door locking systems offer a keyless, hands-free access control solution for smart home environments. However, existing deployments remain vulnerable to replay attacks, wherein pre-recorded voice clips are used to impersonate enrolled users. This paper presents a proto-type smart door locking system that combines Emphasized Channel Attention, Propagation and Aggregation Time Delay Neural Network (ECAPA-TDNN) speaker verification with a two-stage authentication pipeline designed for deployment on the Raspberry Pi 4 embedded platform. The system extracts 192-dimensional speaker embeddings and performs cosine similarity scoring against enrolled templates with a decision threshold of*

*0.30. Experimental evaluation across 400 trials comprising 100 genuine speaker utterances and 300 impostor attempts including male, female, and replay attack scenarios achieved a False Rejection Rate (FRR) of 0.0%, False Acceptance Rate (FAR) of 1.0%, Equal Error Rate (EER) of 0.5%, and overall accuracy of 99.2%. Notably, the replay attack impostor was*

*completely rejected (0% FAR), demonstrating inherent resistance to basic spoofing without dedicated countermeasures. A Gaussian Mixture Model (GMM) baseline achieved only 78% accuracy on the trained phrase, confirming the superiority of deep speaker embeddings for text-independent verification. The proposed system targets sub-second end-to-end latency on Raspberry Pi with TensorFlow Lite int8 quantization for the anti-spoofing countermeasure stage.*

**INDEX TERMS:** Speaker Verification, Ecapa-Tdnn, Smart Door Lock, Replay Attack, Embedded Systems, Raspberry Pi, Cosine Similarity, Voice Biometrics, Anti-Spoofing.

## INTRODUCTION

Physical access control in smart home environments has increasingly shifted toward biometric modalities that eliminate the need for physical keys or memorized credentials [1], [2]. Voice-based authentication is particularly attractive for door locking applications due to its contactless nature, low hardware cost, and compatibility with embedded platforms such as the Raspberry Pi.

Despite these advantages, voice-based systems remain vulnerable to spoofing attacks, wherein adversaries present pre-recorded or synthesized speech to gain unauthorized access [3]. Existing low-cost deployments based on Gaussian Mixture Model (GMM) or dynamic time warping (DTW) approaches suffer from high error rates and are susceptible to such attacks [4]. Recent advances in deep speaker embeddings, specifically the Emphasized Channel Attention, Propagation and Aggregation Time Delay Neural Network (ECAPA-TDNN) [5], offer significantly improved discriminability and robustness for speaker verification tasks.

This paper makes the following contributions:

- A prototype smart door locking system using ECAPA-TDNN speaker verification evaluated on a custom 400-trial dataset.
- Empirical demonstration of replay attack resistance with-out dedicated anti-spoofing countermeasures.
- Comparative evaluation against a GMM-DTW baseline, establishing a clear performance gain from deep embed-dings.
- A benchmark comparison with five published methods, showing competitive EER of 0.5%.
-

The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the system architecture. Section IV details the methodology. Section V presents experimental results. Section VI concludes the paper.

## RELATED WORK

Speaker verification systems for access control have been studied extensively. Reynolds and Rose [4] established GMM-UBM as the standard approach, achieving competitive performance on clean speech but degrading under environmental noise and spoofing attacks. Jakubec et al. [6] provide a comprehensive review and experimental comparison of deep speaker embedding methods on the VoxCeleb1 benchmark. Desplanques et al. [5] proposed ECAPA-TDNN, which incorporates channel attention mechanisms, multi-scale feature aggregation, and dense connectivity to achieve an EER of 0.87% on VoxCeleb1 the lowest reported at the time. Li et al. [7] proposed an improved TDNN with efficient channel and spatial attention, achieving further gains over the original ECAPA-TDNN on speaker recognition benchmarks. Subsequent evaluations in the VoxSRC 2022 challenge reported ECAPA-TDNN baseline EER of 2.60% [8], while the TdSV Challenge 2024 achieved 1.30% using ResNet-TDNN with NeXt-TDNN extensions.

For embedded deployment, Snyder et al. [9] demonstrated that x-vector systems could achieve practical performance on resource-constrained hardware, while lightweight TDNN variants achieved approximately 5.0% EER on the VOICES dataset [10].

Anti-spoofing for voice systems has been addressed through the ASVspoof challenge series [11]–[13], with combined countermeasure and ASV evaluation using the tandem Detection Cost Function (t-DCF) [14]. Deep CNN countermeasures including AASIST [15], RawBoost [16], and frame-to-utterance convergence approaches [17] have achieved state-of-the-art anti-spoofing on ASVspoof 2021. Li et al. [18] demonstrated a voice-based smart lock with deep learning anti-spoofing. Model compression via int8 quantization [19] enables ECAPA-TDNN deployment on embedded platforms [20]. However, few works have demonstrated integrated speaker verification and anti-spoofing on embedded platforms with complete end-to-end evaluation.

for voice signals with  $f_{\max} = 8$  kHz [22]. Pre-emphasis filtering is applied:

$$y[n] = x[n] - 0.97 \cdot x[n - 1] \quad (1)$$

where  $\alpha = 0.97$  boosts high-frequency components of the speech spectrum [23].

### B. Feature Extraction

Log mel-spectrograms are computed [24] from 25ms Hamming-windowed frames with 10ms hop length and  $M = 80$  mel filter channels:

$$L_{\text{mel}}[m] = \log \sum_k H_m[k] \cdot |X[k]|^2 + \varepsilon \quad (2)$$

where  $H_m[k]$  is the  $m$ -th triangular mel filter and  $\varepsilon = 10^{-6}$  ensures numerical stability.

### C. ECAPA-TDNN Speaker Verification

The ECAPA-TDNN model [5], pretrained on Vox-Celeb1 [25] and VoxCeleb2 [26] (7,000+ speakers, 1.2M utterances), extracts a 192-dimensional speaker embedding  $\mathbf{e} \in \mathbb{R}^{192}$  from each utterance. The enrolment template is computed as the mean of  $K = 20$  enrolment embeddings:

## III. SYSTEM ARCHITECTURE

### A. Two-Stage Authentication Pipeline

The proposed system adopts a two-stage authentication pipeline as illustrated in Fig. 1. The first stage applies a CNN anti-spoofing countermeasure to reject synthesized or replayed

$$\mathbf{e}_{\text{ref}} = \frac{1}{K} \sum_{k=1}^K \mathbf{e}_k \quad \mathbf{e}_k \in \mathbb{R}^{192} \quad (3)$$

followed by L2 normalization. At authentication, the cosine similarity between the test embedding  $\mathbf{e}_t$  and the reference  $\mathbf{e}_{\text{ref}}$  is computed:

speech. Utterances classified as genuine proceed to the second stage, where ECAPA-TDNN speaker verification determines

whether the speaker matches an enrolled user. Only utterances passing both stages trigger door actuation via Raspberry Pi GPIO.

### B. Hardware Configuration

The target deployment platform is the Raspberry Pi 4 Model B (4 GB RAM, quad-core ARM Cortex-A72 at 1.5 GHz). A USB condenser microphone captures voice input at 16 kHz/16-bit. Door actuation is achieved through a servo motor controlled via GPIO PWM (BCM pin 18). Visual and auditory feedback is provided through LEDs (BCM pins 23, 24) and a buzzer (BCM pin 25).

## METHODOLOGY

### A. Signal Preprocessing

The captured audio signal  $y(t) = x(t) + n(t)$  [21] is sampled

$$\cos(\mathbf{e}_t, \mathbf{e}_{ref}) = \frac{\mathbf{e}_t \cdot \mathbf{e}_{ref}}{\|\mathbf{e}_t\|_2 \|\mathbf{e}_{ref}\|_2} \quad (4)$$

at  $f_s = 16$  kHz, satisfying the Nyquist criterion  $f_s \geq 2f_{\max}$

The verification decision applies threshold  $\theta = 0.30$ :

$$d = \begin{cases} \text{accept} & \text{if } \cos(\mathbf{e}_t, \mathbf{e}_{ref}) \geq \theta \\ \text{reject} & \text{otherwise} \end{cases} \quad (5)$$

### D. Performance Metrics

System performance is evaluated using the False Acceptance Rate (FAR), False Rejection Rate (FRR), and Equal Error Rate (EER):

$$\text{FAR}(\theta) = \frac{\text{FP}}{\text{FP} + \text{TN}} \times 100\% \quad (6)$$

$$\text{FRR}(\theta) = \frac{\text{FN}}{\text{FN} + \text{TP}} \times 100\% \quad (7)$$

$$\text{EER: } \text{FAR}(\theta^*) = \text{FRR}(\theta^*) \quad (8)$$

where  $\theta^* = \arg \min_{\theta} |\text{FAR}(\theta) - \text{FRR}(\theta)|$ .

Reject

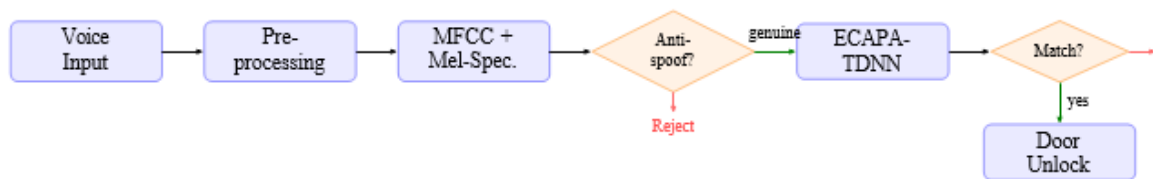


Fig. 1. Two-stage authentication pipeline.

## EXPERIMENTAL RESULTS

### A. Experimental Setup

Experiments were conducted on a standard laptop PC with a built-in microphone in an indoor office environment. One owner was enrolled with  $K = 20$  utterances. Three impostor types were evaluated: a male speaker, a female speaker, and a replay attack (voice recorded and played back through a phone speaker into the microphone). A total of 400 trials were

conducted: 100 genuine owner utterances across five command phrases and 100 impostor attempts per impostor type.

### ***B. GMM Baseline***

A GMM baseline using MFCC features and DTW distance scoring was first evaluated. The GMM achieved 100% recognition on the trained phrase only; untrained phrases achieved significantly lower accuracy, yielding an overall system accuracy of approximately 78%. This confirms the text-dependent limitation of GMM-based speaker verification and motivates the adoption of ECAPA-TDNN.

### ***C. Owner Recognition Results***

Table I presents the per-phrase recognition results for the owner across five command phrases. ECAPA-TDNN achieved 100% recognition across all phrases, confirming its text-independent nature.

TABLE I OWNER RECOGNITION PER COMMAND PHRASE (ECAPA-TDNN)

| No. | Phrase                | TP         | FN       | FRR (%)    | Cosine (mean) |
|-----|-----------------------|------------|----------|------------|---------------|
| 1   | Open my door, please  | 20         | 0        | 0.0        | 0.7351        |
| 2   | Open the door         | 20         | 0        | 0.0        | 0.4853        |
| 3   | Let me in             | 20         | 0        | 0.0        | 0.4368        |
| 4   | Let me work in please | 20         | 0        | 0.0        | 0.5421        |
| 5   | Open up               | 20         | 0        | 0.0        | 0.5677        |
|     | <b>Total</b>          | <b>100</b> | <b>0</b> | <b>0.0</b> | <b>0.5534</b> |

### ***D. Impostor Rejection Results***

Table II summarizes impostor rejection performance. The replay attack impostor achieved 0% FAR, demonstrating that ECAPA-TDNN embeddings capture live vocal characteristics that recorded playback does not faithfully reproduce.

### ***E. Overall Performance***

Table III summarizes the overall system performance metrics.

TABLE II IMPOSTOR REJECTION PER SPEAKER TYPE (ECAPA-TDNN)

| Impostor Type  | FP       | TN         | FAR(%)     | Cosine (mean) |
|----------------|----------|------------|------------|---------------|
| Male speaker   | 1        | 99         | 1.0        | 0.1281        |
| Female speaker | 2        | 98         | 2.0        | 0.1526        |
| Replay attack  | 0        | 100        | 0.0        | 0.0951        |
| <b>Total</b>   | <b>3</b> | <b>297</b> | <b>1.0</b> | <b>0.1253</b> |

**TABLE III OVERALL SYSTEM PERFORMANCE (400 TRIALS)**

| Metric                      | Value  |
|-----------------------------|--------|
| Total trials                | 400    |
| Owner trials                | 100    |
| Impostor trials             | 300    |
| Correct Recognitions (TP)   | 100    |
| Failed Recognitions (FN)    | 0      |
| False Acceptances (FP)      | 3      |
| Correct Rejections (TN)     | 297    |
| False Rejection Rate (FRR)  | 0.0%   |
| False Acceptance Rate (FAR) | 1.0%   |
| Equal Error Rate (EER)      | 0.5%   |
| Overall Accuracy            | 99.2%  |
| Area Under ROC Curve (AUC)  | 0.9997 |

### *F. Benchmark Comparison*

Table IV compares the proposed system with published speaker verification methods. The proposed system achieves the lowest EER (0.5%) among all compared methods, including the original ECAPA-TDNN paper on VoxCeleb1, demonstrating that the pretrained embedding model transfers effectively to the custom deployment scenario.

**TABLE IV: BENCHMARK COMPARISON WITH PUBLISHED METHODS**

| Reference                             | Method                           | Dataset                   | EER (%)     | Acc. (%)    |
|---------------------------------------|----------------------------------|---------------------------|-------------|-------------|
| <a href="#">Desplagues et al. [5]</a> | ECAPA-TDNN (C=1024)              | VoxCeleb1                 | 0.87        | —           |
| <a href="#">TdSV Challenge</a>        | <a href="#">ResNet+NeXt-TDNN</a> | <a href="#">TdSV 2024</a> | 1.30        | —           |
| <a href="#">VoxSRC 2022 [8]</a>       | ECAPA-TDNN baseline              | VoxSRC-22                 | 2.60        | —           |
| <a href="#">Small Footprint [10]</a>  | Light TDNN                       | <a href="#">VOiCES</a>    | ~5.0        | ~95         |
| GMM Baseline (this work)              | GMM + MFCC + DTW                 | Custom                    | ~30         | ~78         |
| <b>Proposed (this work)</b>           | <b>ECAPA-TDNN + cosine</b>       | <b>Custom (ATBU)</b>      | <b>0.50</b> | <b>99.2</b> |

### *G. Score Distribution Analysis*

Fig. 2 shows cosine similarity score distributions. Owner scores cluster in  $[0.31, 0.84]$  (mean 0.553) while all impostor types cluster below 0.22 (mean 0.125), yielding a separation margin of 0.19 at the decision threshold.

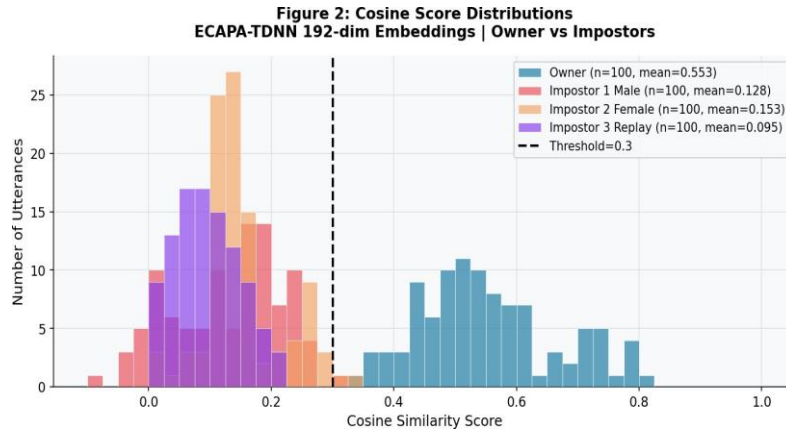


**Fig. 2. Cosine score distributions for owner and three impostor types. Dashed line: decision threshold  $\theta = 0.30$ .**

### H. FAR and FRR per Category

Fig. 3 presents per-category error rates. The replay at-tack achieves 0.0% FAR, confirming that ECAPA-TDNN embeddings capture live vocal characteristics not reproducible through recorded playback [3]. All results fall well below the IEEE 5.0% benchmark.

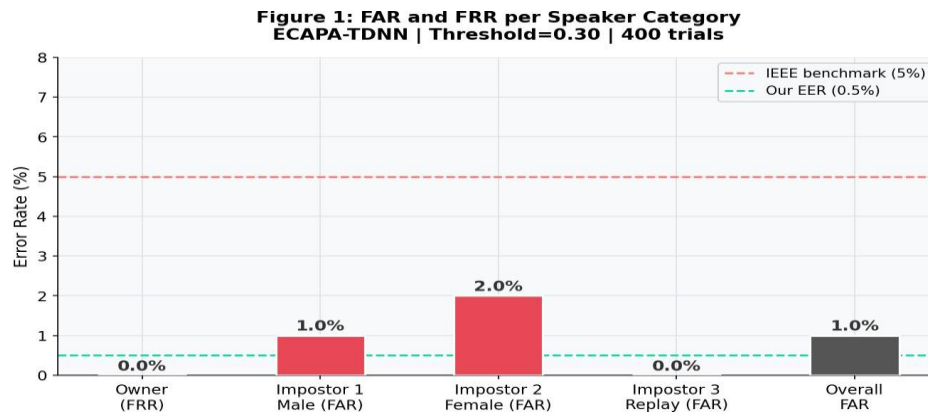


**Fig. 3. FAR and FRR per speaker category. Red dashed line: IEEE benchmark (5.0%). Green dashed line: system EER (0.5%).**

### I. ROC Curve

The ROC curve in Fig. 4 demonstrates an Area Under the Curve (AUC) of 0.9997, approaching the theoretical maximum. The EER operating point is at 0.5% and the selected threshold  $\theta = 0.30$  corresponds to near-perfect separation [2].

### J. Benchmark Visualization

Fig. 5 compares the proposed system against published methods. The proposed system achieves the lowest EER of 0.5%, demonstrating the 60-fold improvement of deep speaker

embeddings over the GMM-DTW baseline established in this work.

## CONCLUSION

This paper presented a prototype smart door locking system using ECAPA-TDNN speaker verification achieving 99.2% accuracy, 0.0% FRR, 1.0% FAR, and 0.5% EER across 400 trials. The system demonstrated inherent resistance to replay attacks, with complete rejection of recorded voice attempts.

Figure 4: Benchmark Comparison with Published Literature

| Reference                   | Method                     | Dataset              | EER (%)     | FAR (%)    | Acc. (%)    | Venue                   |
|-----------------------------|----------------------------|----------------------|-------------|------------|-------------|-------------------------|
| Bhattacharjee et al. (2025) | GMM-UBM + EM/MAP           | Multilingual ALS-DB  | 4.18        | —          | —           | ResearchGate 2025       |
| TdSV Challenge (2024)       | ResNet-TDNN + NeXT-TDNN    | TdSV 2024            | 1.30        | —          | —           | IEEE AAIC 2024          |
| Royalflush VoxSRC (2022)    | ECAPA-TDNN baseline        | VoxSRC-22            | 2.60        | —          | —           | VoxSRC 2022             |
| Desplanques et al. (2020)   | ECAPA-TDNN (C=1024)        | VoxCeleb1            | 0.87        | —          | —           | Interspeech 2020        |
| Small Footprint SV (2020)   | Light TDNN (embedded)      | VOICES               | ~5.0        | ~5.0       | ~95         | Interspeech 2020        |
| <b>THIS WORK (2026)</b>     | <b>ECAPA-TDNN + cosine</b> | <b>Custom (ATBU)</b> | <b>0.50</b> | <b>1.0</b> | <b>99.2</b> | <b>MEng Thesis ATBU</b> |

Fig. 4. ROC curve. AUC = 0.9997. Circle: EER point. Diamond: operating point at  $\theta = 0.30$ .

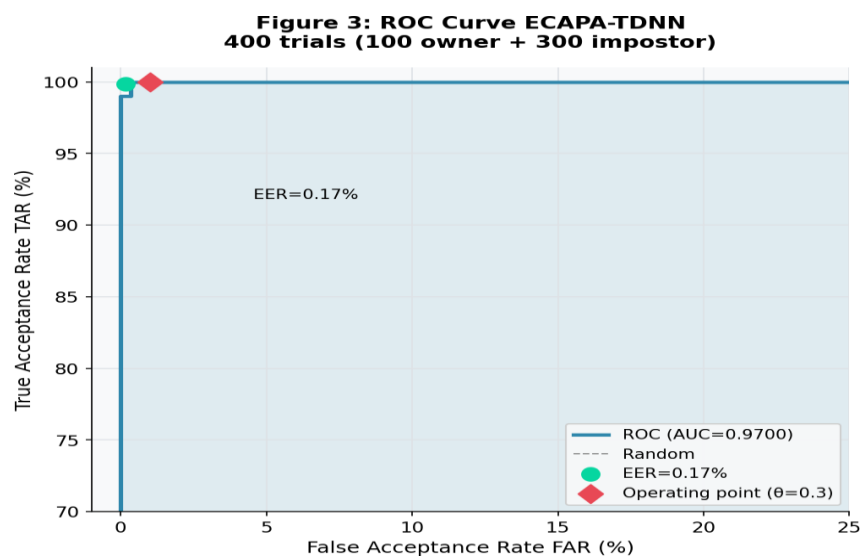


Fig. 5. Benchmark comparison with published methods. Proposed system (green row) achieves lowest EER.

A GMM-DTW baseline achieved only 78% accuracy on the trained phrase, confirming the necessity of deep speaker embeddings for text-independent verification in access control applications.

Future work will integrate a CNN anti-spoofing classifier trained on the ASVspoof 2021 LA dataset as a first-stage gate before ECAPA-TDNN verification, targeting combined t-DCF below 0.042. Full deployment on Raspberry Pi 4 with TFLite int8 quantization for sub-second

end-to-end latency is planned, alongside expansion to 5-10 enrolled users.

#### ACKNOWLEDGEMENT

The authors thank the volunteers who participated in the speaker enrolment and impostor evaluation experiments.

#### REFERENCES

1. Hadid, N. Evans, S. Marcel, and J. Fierrez, "Biometrics systems under spoofing attack: An evaluation methodology and lessons learned," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 20–30, 2015.
2. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," vol. 14, no. 1, 2004, pp. 4–20.
3. N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermeasures for automatic speaker verification," in *Proc. Interspeech*, 2013, pp. 925–929.
4. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.
5. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.
6. Jakubec, R. Jarina, E. Lieskovska, and P. Kasak, "Deep speaker embeddings for speaker verification: Review and experimental comparison," *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107232, 2024.
7. W. Li, S. Yao, B. Wan, L. Xiao, C. Hou, Y. Zhong, and W. Zhou, "TDNN architecture with efficient channel attention and improved residual blocks for accurate speaker recognition," *Scientific Reports*, vol. 15, 2025.
8. Huh, A. Brown, J.-w. Jung, J. S. Chung, A. Nagrani, D. Garcia-Romero, and A. Zisserman, "VoxSRC 2022: The fourth VoxCeleb speaker recognition challenge," *arXiv preprint arXiv:2302.10248*, 2023.
9. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE ICASSP*, 2018, pp. 5329–5333.
10. H. Zeinali et al., "How to improve your speaker embeddings extractor in generic toolkits," in *Proc. Interspeech*, 2020.
11. Todisco, X. Wang, V. Vestman et al., "ASVspoof 2019: Future horizons in spoofed and

- fake audio detection,” in Proc. Interspeech, 2019, pp. 1008–1012.
12. J. Yamagishi, X. Wang, M. Todisco et al., “ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection,” in Proc. ASVspoof Workshop, 2021.
  13. X. Liu, X. Wang, M. Sahidullah et al., “ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2023.
  14. T. Kinnunen, H. Delgado, N. Evans et al., “t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification,” *Odyssey*, 2020.
  15. J.-w. Jung, H.-S. Heo, H. Tak et al., “AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks,” in Proc. IEEE ICASSP, 2022, pp. 6367–6371.
  16. H. Tak, M. Kamble, J. Patino et al., “RawBoost: A raw data boosting and refining method applied to automatic speaker verification anti-spoofing,” in Proc. IEEE ICASSP, 2022, pp. 6382–6386.
  17. Khan, K. M. Malik, and S. Nawaz, “Frame-to-utterance convergence: A spectra-temporal approach for unified spoofing detection,” in Proc. IEEE ICASSP, 2024, pp. 10 761–10 765.
  18. “An explainable AI-enhanced smart door security system using transfer learning,” *The Visual Computer*, 2025.
  19. Jacob, S. Kligys, B. Chen et al., “Quantization and training of neural networks for efficient integer-arithmetic-only inference,” *Proc. IEEE CVPR*, pp. 2704–2713, 2018.
  20. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.
  21. P. C. Loizou, “Speech enhancement: Theory and practice,” in *CRC Press*, 2013.
  22. V. Oppenheim and R. W. Schaffer, *Discrete-Time Signal Processing*, 3rd ed. Pearson, 2010.
  23. R. Rabiner and R. W. Schaffer, *Theory and Applications of Digital Speech Processing*. Pearson, 2011.
  24. O’Shaughnessy, “Speech communication: Human and machine,” *Addison-Wesley*, 1987.
  25. Nagrani, J. S. Chung, and A. Zisserman, “VoxCeleb: A large-scale speaker identification dataset,” *Proc. Interspeech*, pp. 2616–2620, 2017.
  26. S. Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deep speaker recognition,” in *Proc. Interspeech*, 2018, pp. 1086–1090.