Ghost Recon : An Orchestration-based, Non-Intrusive, Persistent MultimodalAuthentication System*

Suhyeok Jang, Jeong Jin Yoon, Seohyun Lee, Hyelim Jung, and Ki-Woong Park†

Sejong University, Seoul, Republic of Korea
{suv240113, wlsdbs61, hyunlee03122}@naver.com, hyello13@gmail.com,
woongbak@sejong.ac.kr

Abstract

Existing one-time authentication methods, such as passwords, patterns, and facial recognition, used to detect unauthorized users have limitations in detecting session hijacking, authentication token reuse, and privilege abuse that can occur during a session. To address this, the zero-trust security model has been introduced, which adheres to the principle of "never trust, always verify." However, the high computational load and authentication delays that arise from repeatedly verifying all requests degrade the user experience. To address these issues, this paper proposes an unsuspected continuous authentication system that orchestrates the modalities of face, skeleton, and voice. By performing complementary multimodal real-time user authentication processes, the proposed system minimizes authentication computational costs while maintaining the zero-trust principle of continuous verification. Furthermore, it ensures stable operation even in the face of modality failures or environmental constraints. Furthermore, it visualizes user authentication results, reliability, and performance indicators in real time to support security managers' response to security threats.

1 Introduction

Traditional castle-and-moat security models are based fundamentally on trusting data and transactions within the network perimeter [3]. However, with the rise of security threats like LoTL (Living Off The Land) and insider attacks, there are limitations to exposing critical resources to threats. To overcome this, the Zero Trust security model emerged [7]. Zero Trust is a security paradigm based on the principle of "never trust, always verify," which considers all elements as targets for protection and continuously verifies them [9]. However, these characteristics have several limitations. Persistent authentication and authorization processes can significantly increase system resource usage, resulting in computational overhead, performance lag, and privacy violations, degrading the user experience [6]. Recently, continuous, unconscious authentication systems, which continuously verify a user's identity without the user's awareness of the authentication process, have been gaining attention as an alternative to addressing the limitations of the zero-trust model [11]. This approach aligns with the zero-trust model, maintaining high security while minimizing authentication computation overhead and degrading the user experience. However, this type of persistent, unsuspected authentication process has limitations, such as the potential for authentication continuity to be limited due to single-modality data omissions and defects. To overcome these limitations, this study proposes a multimodal authentication system combining face, skeleton, and voice, based on persistent, unsuspected

*Proceedings of the 9th International Conference on Mobile Internet Security (MobiSec'25), Article No. 69, December 16-18, 2025, Sapporo , Japan.

†Corresponding author

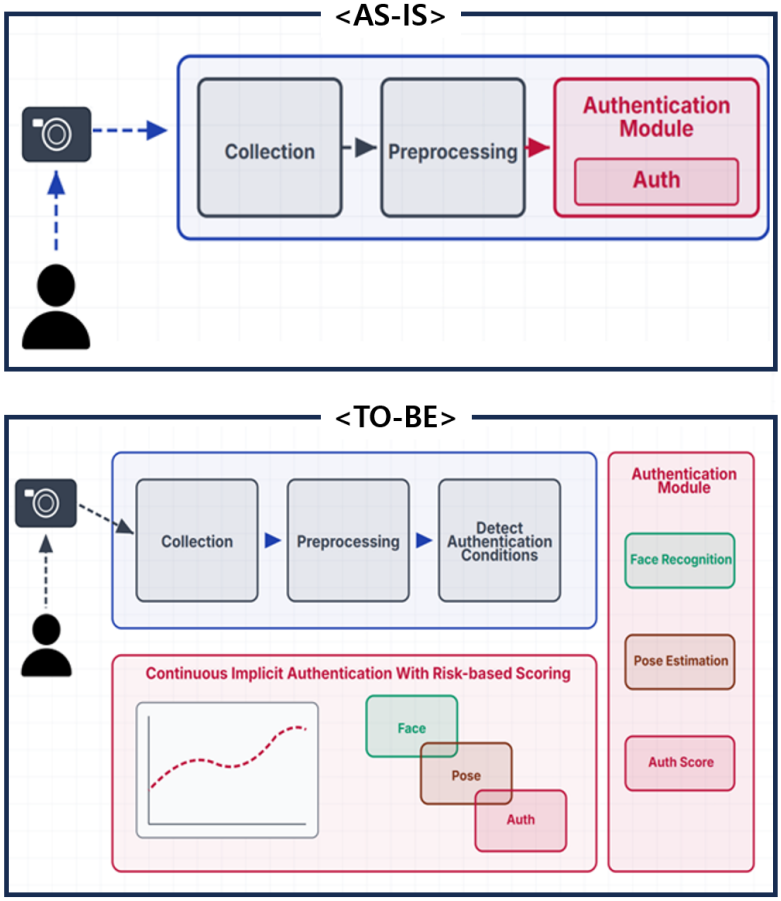

Figure 1: As-Is / To-Be Architecture for Implicit Authentication System

authentication and capable of responding to various environmental variables. The implicit continuous authentication system proposed in this paper is based on a multimodal computational approach that integrates face, skeletal, and vocal modalities. An orchestration framework is applied to dynamically adjust and distribute the roles and weights of each modality based on the situation. This reduces the computational overhead and user interaction burden inherent in existing zero-trust security models, enabling implicit authentication [8].

Figure 1 compares the existing single-modal authentication method with the implicit continuous authentication system proposed in this paper. In Figure 1 (As-Is), the authentication module collects and preprocesses user information, then performs one-time authentication, without any additional verification steps. In contrast, in Figure 1 (To-Be), the authentication module calculates authentication scores for each modality based on authentication conditions and performs continuous authentication. Therefore, the continuous, implicit multi-modal authentication proposed in this paper alleviates the repetitive authentication process required of users through a continuous, implicit authentication process.

2 Related Work

In recent security studies on authentication systems, a classification system was established from the perspective of sensor/feature/window design, and a comparison standard was established for these [4]. In a study on face-voice combination, a case was reported where the EER (Equal Error Rate) was empirically reduced through the fusion of FaceNet (face) authentication scores and GMM (voice) authentication scores [1], and in the field of fingerprint data-based authentication system research, a comprehensive review was presented that included the fusion of features/scores/decision-making levels, template protection, attack models, and evaluation indices of multimodal studies that combined fingerprint data with other biometric data as well as the method of using fingerprint data alone [10]. In a study on fingerprint-ECG (Equal Error Rate) multimodal authentication, a design was proposed that improved performance by applying stacking and channel-level fusion after embedding each biosignal using a transformer [2], and in the field of wearable-based biosignal authentication, a multimodal authentication/identification model combining ECG (Electrocardiogram) and PPG (Photoplethysmogram) was demonstrated to be capable of a high level of performance improvement compared to a single-signal-based authentication method [5].

This paper identifies two limitations of existing related research. First, a method for rapidly adjusting operational policies when anomalies occur has not been established. While the classification and systematization of continuous authentication technologies have advanced, the operational cycle, the resolution of collected data, and security system activation/shutdown guidelines (e.g., escalation according to stable/risky conditions) have been rarely presented. For example, the authentication step combining face and voice data remains at the level of research for score calculation, and research on skeleton and A/V (Audio-Visual) synchronization lacks integrated operational procedures for user-application sessions. Second, while cross-validation technology exists, an operational system for integrating and adjusting this technology at session runtime and linking it to policies has not been established. Existing authentication systems have identified modality-specific advantages (face-identity, skeleton-continuity, voice-assistant/liveness), but research on operational procedures and rules for real-time cross-validation and modality-specific weight adjustment (identity, continuity, and auxiliary) in a single pipeline environment has not been conducted.

While previous studies focused on what to combine for greater accuracy, this study applies a priority-based (1. Face 2. Skeleton 3. Voice) orchestration system and calculates a continuous confidence index C(t) that combines the similarity and quality of each modality, and uses this as a control index to expand the scope from accuracy to operation. According to C(t), the cycle, resolution, and operation status are linked to the operation rules in a step-up/downward manner, and the runtime cross-consistency of lips and audio, scenes and sound fields, and skeletons and faces is immediately reflected, thereby lowering the average computation, delay, and transmission overhead, while selectively strengthening the inspection in case of anomalies.

3 Design

The proposed implicit continuous authentication system integrates face embeddings, skeletal (pose) features, and speaker (voice) information through a late fusion strategy, thereby minimizing the limitations of any single modality. The system is designed to maintain reliable real-time authentication even under conditions where specific modalities are degraded—such as when a Head-Mounted Display (HMD) is worn, the face is partially occluded, the subject is captured from a distance, or the environment suffers from low lighting or background noise.

The overall pipeline consists of two main stages: a Training Phase and a Verification Phase. During the Training Phase, features from face, pose, and voice modalities are individually learned, after which a fusion-based classifier is trained to combine them. In the Verification Phase, the same feature extractors employed during training are used to derive facial, skeletal, and vocal representations from incoming data. These features are then normalized and passed into the trained classifier, which produces a probabilistic prediction of user authenticity. The outputs are transmitted to an operational dashboard for real-time visualization. In the event of an unauthorized intrusion, the system triggers immediate security alerts, while simultaneously logging performance metrics to enable ongoing system evaluation and analysis.

3.1 Training Phase

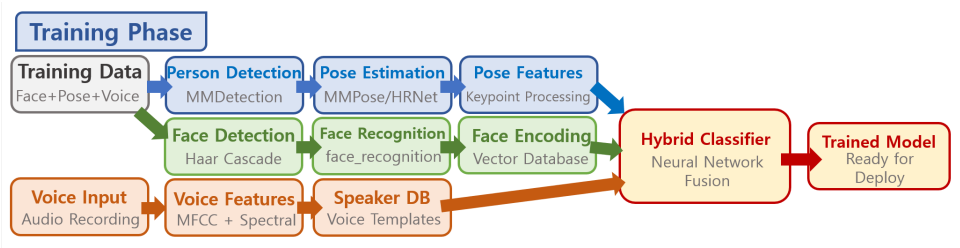

Figure 2: Training Phase Structure

Figure 2 illustrates the training procedure of the proposed implicit continuous authentication system. The training pipeline operates with the person detection module at its core, where face and pose detection are processed in parallel, and an independent voice training pathway is executed. In the diagram, blue denotes the pose pipeline, green represents the face pipeline, brown indicates the voice pipeline, red highlights the modality fusion and trained classifier, and gray corresponds to the normalization and calibration layers. The training data assumes the simultaneous collection of facial, skeletal, and vocal information, with visual and auditory samples aligned in time for each subject during preprocessing. The feature vectors extracted from the three modalities are fused and jointly trained in a hybrid classifier. The fusion input consists of face similarity, pose feature vectors, and voice similarity, which are standardized and calibrated before being passed to classifiers such as a multilayer perceptron (MLP), logistic regression, or gradient-boosted decision trees (GBDT). The classifier estimates the probability of an "authorized user" through late-fusion learning, while the training process employs cross-entropy loss and normalization techniques. Issues such as class imbalance and domain shift are mitigated by weighted loss functions, temperature scaling, and domain augmentation strategies. Finally, the training output is stored and deployed as the "Trained Model," which includes not only the learned classifier parameters but also operational configurations such as decision thresholds, temporal window lengths, and quality-gating rules for use during the verification phase.

3.2 Real-time Verification & Output/Dashboard

Figure 3 illustrates the verification phase of the proposed implicit continuous authentication system. In a real-time environment, person detection is first performed on the input camera

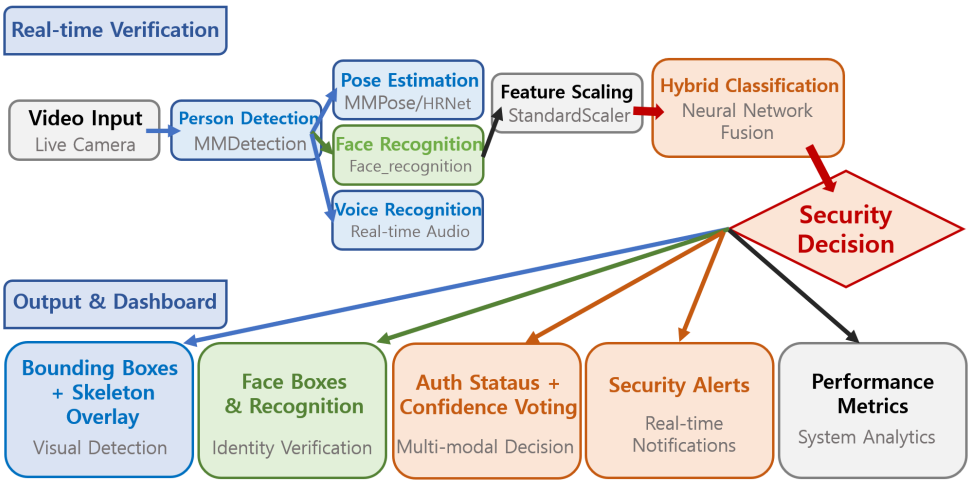

Figure 3: Real-time Verification & Output/Dashboard Structure

stream to obtain the region of interest (ROI), followed by parallel execution of pose estimation, face recognition, and speaker recognition. The face recognition module has been trained with extensive data augmentation (e.g., occlusion, resolution degradation, viewpoint variation) to enhance generalization, and the pipeline remains functional with pose-based authentication even when the face cannot be detected. When voice input is available, the captured utterances are segmented into recognizable intervals, and real-time MFCC and spectral features are extracted to compute speaker similarity. The outputs of each modality are normalized and calibrated through feature scaling to align with the training distribution, and dynamic weights are assigned based on quality indicators such as frame quality, keypoint confidence, and speech signal-to-noise ratio (SNR). The normalized features are then passed to the hybrid classifier, which performs late-fusion inference by adjusting weights according to available modalities. Temporal smoothing and sliding-window voting are applied to mitigate false rejections caused by frame-level fluctuations. The final decision is categorized into three states: accept (immediate authentication when the confidence is sufficiently high for a legitimate user), reject (denial and security alert when the confidence indicates an unauthorized user), and defer (uncertain cases where additional frames or utterance segments are collected for re-evaluation). The decision outcomes are continuously reflected on the real-time dashboard. The interface visualizes detected individuals with skeleton overlays to confirm consistency between detection and pose estimation, and the face recognition path provides matched IDs with their confidence scores. The fused decision output is aggregated as authentication states and trust trajectories, with explanatory contributions from each modality. When an intruder is detected, Security Alerts are immediately triggered, logged, and reported to the administrator. In addition, the dashboard provides a visual explanation of authentication states and security grounds, while the Performance Metrics module continuously monitors indicators such as FPS, latency, and accuracy to detect long-term performance drift at an early stage. Through this design, the system is able to maintain authentication continuity even under modality loss and supports evidence-based decision-making in real-time environments.

3.3 Differences from existing systems

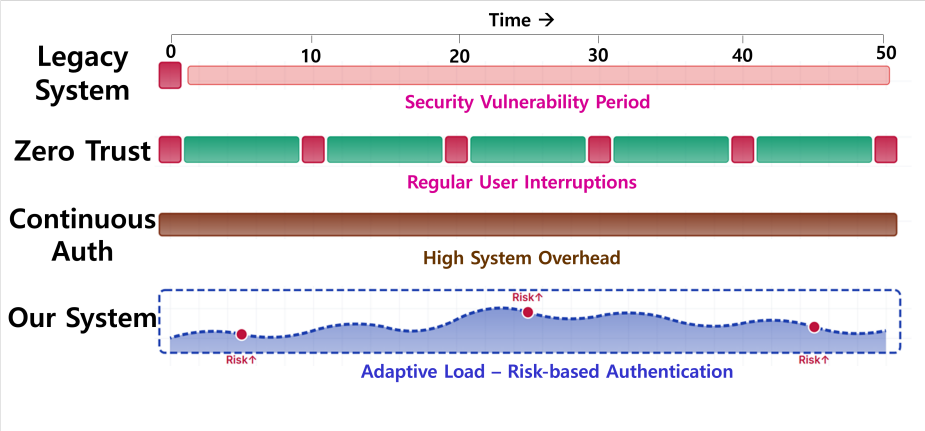

Figure 4: Differences from existing systems

Figure 4 compares the Legacy System, Zero Trust, Continuous Authentication, and the proposed implicit continuous multimodal authentication system along the time axis. In the Legacy System, authentication is performed only at the beginning of a session, which results in a long security vulnerability period thereafter. The Zero Trust model enforces verification for every request, thereby strengthening security, but frequent user interruptions degrade the user experience. Continuous Authentication eliminates security vulnerability periods by maintaining continuous verification throughout the session; however, this approach incurs excessive consumption of system resources, leading to high overhead. In contrast, the proposed system adopts a risk-based adaptive authentication strategy: during stable periods, computational and communication loads are reduced, while in periods where the risk indicator rises, the system immediately escalates the authentication intensity. This approach suppresses unnecessary computation while preserving the Zero Trust principle of "always verify," thereby achieving a balance between security, efficiency, and user experience.

3.4 Attack Scenario & Security Evaluation

In environments such as research labs, corporate security areas, smart offices, unmanned facilities, and testing centers, unattended and persistent authentication is essential. This is because unauthorized access or identity forgery in these environments can result in serious damage. Relying solely on single-signal authentication presents numerous practical vulnerabilities. For example, attackers can deceive facial authentication by playing back images or recorded videos on the screen, bypass voice authentication by playing back previously recorded voices, and attempt to simultaneously deceive multiple sensors using deepfakes that precisely synchronize synthetic video and voice with each other temporally. Furthermore, network replay attacks, which re-inject known video and audio streams into the system, or physically or electronically disrupt camera and microphone inputs, pose real threats. Therefore, to protect these environments, we must design a method that assumes the possibility of single-modal forgery and continuously integrates multiple signals to verify temporal and cross-modal consistency.

To achieve this, we designed and implemented a multimodal system that orchestrates three complementary modalities: face, skeleton (gait and sequential body movements), and voice. The system is designed to proactively detect inconsistencies between modalities, rather than relying solely on any one modality. For example, even if an attacker can temporarily obtain a pass rating for face and voice by simultaneously playing high-definition video and recorded voice, overall reliability will decline if the gait pattern or sequential upper and lower body movements observed in the skeleton data deviate from the target's normal behavior. Conversely, even attacks using only the skeleton will raise suspicion if subtle facial expression changes or vocalization timing are inconsistent. While perfectly replicating a single signal may be relatively easy, synthesizing three modalities in a temporally coherent manner is much more challenging. Therefore, orchestration itself significantly increases the attacker's difficulty and cost.

We implemented this coordinated approach and adopted a flexible continuity-based operational policy to prevent immediate and complete blocking due to a single modality's temporary anomaly. The system responds incrementally, referencing recent modal history and cross-modal consistency, by temporarily reducing privileges, requiring additional verification, or blocking access. This adaptive policy reduces false positives while remaining sensitive to sophisticated synchronization attacks. Consequently, this multimodal orchestration effectively balances security and user convenience, reducing the likelihood of successful replay, forgery, and deepfake attacks in real-world environments, such as the aforementioned labs and unmanned facilities.

4 Conclusion

In this paper, we proposed an implicit continuous authentication system designed to effectively address the trade-off between the "always verify" principle of Zero Trust and the resulting performance degradation and user experience issues. The proposed system learns facial, skeletal, and voice features during the training phase, and in the real-time verification phase, it continuously identifies users through quality-based dynamic weighting and a hybrid classifier. In addition, the system enhances reliability by providing visualized evidence of authentication decisions and security status through a dashboard interface. This implicit continuous authentication design overcomes the limitations of single-modality authentication schemes and is engineered to operate robustly even under modality loss or degraded quality in real-world environments. Future work will focus on scalability experiments in large-scale user environments, the integration of additional modalities (ex : behavioral patterns), and the application of privacy-preserving learning techniques to further advance the system.

5 Acknowledgments

This research was supported by the Future Challenge Defense Technology Research and Development Project (9150921) hosted by the Agency for Defense Development Institute in 2023.

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) (Project No. RS-2023-00228996, 30%; RS-2024-00438551, 20%; IITP-2025-RS-2021-II211816, 10%), the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2025 (Project Name: Training Global Talent for Copyright Protection and Management of On-Device AI Models, Project Number: RS-2025-02221620, Contribution Rate: 20%), and the National Research Foundation of Korea (NRF) grant funded by the Korean Government

(Project No. RS-2023-00208460, 20%).

References

[1] Bayan Alharbi and Hanan S Alshanbari. Face-voice based multimodal biometric authentication system via facenet and gmm. *PeerJ Computer Science*, 9:e1468, 2023.

[2] Nassim Ammour, Yakoub Bazi, and Naif Alajlan. Multimodal approach for enhancing biometric authentication. *Journal of Imaging*, 9(9):168, 2023.

[3] Muhammad Ajmal Azad, Sidrah Abdullah, Junaid Arshad, Harjinder Lallie, and Yussuf Hassan Ahmed. Verify and trust: A multidimensional survey of zero-trust security in the age of iot. *Internet of Things*, 27:101227, 2024.

[4] Dutliff Boshoff and Gerhard P Hancke. A classifications framework for continuous biometric authentication (2018–2024). *Computers & Security*, 150:104285, 2025.

[5] Yue Che, Lingyan Du, Guozhi Tang, and Shihai Ling. A biometric identification for multi-modal biomedical signals in geriatric care. *Sensors*, 24(20):6558, 2024.

[6] Hongzhaoning Kang, Gang Liu, Quan Wang, Lei Meng, and Jing Liu. Theory and application of zero trust security: A brief survey. *Entropy*, 25(12):1595, 2023.

[7] Chunwen Liu, Ru Tan, Yang Wu, Yun Feng, Ze Jin, Fangjiao Zhang, Yuling Liu, and Qixu Liu. Dissecting zero trust: Research landscape and its implementation in iot. *Cybersecurity*, 7(1):20, 2024.

[8] Swimpy Pahuja and Navdeep Goel. Multimodal biometric authentication: A review. *AI Communications*, 37(4):525–547, 2024.

[9] Simone Rodigari. Performance analysis of zero trust in cloud native systems. 2023.

[10] U Sumalatha, K Krishna Prakasha, Srikanth Prabhu, and Vinod C Nayak. A comprehensive review of unimodal and multimodal fingerprint biometric authentication systems: Fusion, attacks, and template protection. *IEEE Access*, 12:64300–64334, 2024.

[11] Nida Zeeshan, Makhabbat Bakyt, Naghmeh Moradpoor, and Luigi La Spada. Continuous authentication in resource-constrained devices via biometric and environmental fusion. *Sensors*, 2025.