An Architecture for Voice-Based Authentication and Authorization with Deepfake Detection

Fabian Martins Da Silva, Baladithya Balamurugan and John Hakim
Amazon Web Services, USA

fabisilv@amazon.com
baladitia@amazon.com
johnhak@amazon.com

Abstract: Voice biometrics offer a convenient and secure authentication method, but the rise of sophisticated deepfake technology presents a significant challenge. This work presents an architecture for voice-based authentication and authorization that integrates deepfake detection to mitigate this risk. This paper explores the design of this cloud-native architecture, leveraging Amazon Web Services (AWS) services for orchestration and scalability. The system combines cutting-edge AI models for robust voice-printing and real-time deepfake analysis. We discuss multi-factor authentication (MFA) strategies that provide layered defense against unauthorized access. Two specific use cases are explored: identity verification and secure approval of banking transactions. This paper addresses key considerations for real-world deployment, including system resiliency, cost-effectiveness, and the efficiency of the AI models under varying conditions. We evaluate the architecture's suitability as a two-factor authentication (2FA) solution, focusing on the accuracy of deepfake detection and the rates of false negatives and false positives.

Keywords: Voice biometrics, Deepfake detection, Authentication and authorization, AI-Driven security, Biometric security

1. Essentials of Access Control Systems

Access control is a fundamental security requirement for systems where resources or processes must be accessible only to authorized entities (individuals, systems, or devices) (Bishop, 2003). These systems employ tiered access levels (permissions) based on user profiles, evident in applications ranging from streaming services where parental controls restrict content based on age, to high-security environments like nuclear facilities, where critical functions (e.g., missile launch) necessitate synchronized physical actions (e.g., multi-person authorization) to mitigate risk. The sensitivity and criticality of the resource or process directly correlates with the stringency of the access control measures employed.

Every access control system comprises three core processes: identification, authentication, and authorization (NIST, 2013):

- *Identification*: This process encompasses the activities required to ascertain the identity of a user, either during initial enrollment or in subsequent interactions. Enrollment often involves an authoritative validation of the user's identity, establishing a unique record within the system linked to an identifier (e.g., username, government-issued ID). In future interactions, the user presents this identifier. However, this alone does not guarantee the user's authenticity.
- *Authentication (AuthN*): This process establishes confidence that the entity presenting the identifier is the legitimate owner. This confidence can be based on objective (quantitative, measurable) or subjective (qualitative) assessments of the evidence provided.
- *Authorization (AuthZ):* Once identity is confirmed through identification and authentication, the system grants access to resources or permissions based on the user's profile and any contextual factors, such as resource availability.

Historically, passwords (or passcodes) established during enrollment have been the most prevalent authentication factor, falling under the "something you know" category. These, due to poor definition (weak passwords), reuse, and potential risk of exposure to malware, tend to be a weak mechanism for authentication. However, physical tokens or keys ("something you have") are also common. Banks started working with physical tokens in the early 1990s, combining username for identification, password as the first-factor for authentication, and an OTP (one-time password/passphrase, or OTC for one-time passcode) generated by the token as the second-factor for authentication. Increasingly, systems leverage biometric traits ("something you are") captured during enrollment for authentication. Depending on the criticality of the access request, systems may require multiple authentication factors from different categories (Bonneau et al., 2012).

2. Biometrics as a Factor for AuthN and AuthZ

Biometrics serves both identification and authentication purposes. Consider the scenario of a friend arriving at your door. Their voice, as they greet you, acts as an initial identifier. You then visually verify their identity through the peephole, your brain performing a real-time comparison with stored memories. This process, however, is subject to errors, such as mistaking an identical twin or being deceived by a disguise. This analogy demonstrates the dual role biometrics can play, identifying through vocal characteristics and authenticating through facial recognition.

While biometrics can be used for identification, they are more commonly employed as authentication factors, often in conjunction with other methods. Biometric authentication relies on unique biological traits that, while subject to change over time, are generally difficult to replicate. Fingerprint recognition, dating back to the 17th century (Bhattacharyya et al., 2009), was among the first biometrics integrated into modern smartphones. Other less obvious biometric traits, such as gait (manner of walking), can also be used for authentication.

Traditionally, biometric authentication involved trained specialists visually comparing a live sample with a stored record, as with fingerprints or handwritten signatures. However, security experts recognized the vulnerability of these methods to forgery. Fingerprints could be lifted and replicated, and signatures could be forged. Similarly, facial recognition systems could be tricked with high-quality masks or images.

To counter these threats, “liveness detection” mechanisms emerged. These aim to ensure that the biometric data is captured from a live individual in real-time. For fingerprint scanning, this might involve measuring temperature, heartbeat patterns, or vein distribution. Facial recognition systems may require specific head movements or expressions that are difficult to mimic with a static image or mask. Voice authentication often involves prompting the user to utter a dynamically generated phrase.

In most cases, biometric authentication is used as a secondary factor, following an initial authentication step, such as password entry. This constitutes Multi-Factor Authentication (MFA), with 2FA being a specific instance using two factors. Common MFA combinations include “something you know” (password), “something you have” (a device), and “something you are” (biometric trait). Modern systems, especially in high-security contexts like banking, may employ three or more authentication factors. For example, when using a banking app on a smartphone, you might log in with a username and password, then provide a fingerprint or facial scan, and finally, receive a one-time dynamically generated code via text message to your registered phone number.

Secure storage of biometric data is paramount. Ideally, systems should utilize encryption vaults or Trusted Platform Modules (TPMs) (Trusted Computing Group, 2011) that allow for secure comparison without permitting data retrieval. The recent miniaturization and cost reduction of TPMs have facilitated their integration into smartphones, enabling secure biometric authentication on these devices.

Unlike deterministic methods like passwords, which result in a binary “match” or “no match” outcome, biometric authentication involves a probabilistic assessment. Due to natural variations in biometric traits, environmental factors, and sensor limitations, a perfect match is rarely achieved. Instead, systems compute a match score, indicating the degree of similarity between the presented sample and the stored template. This inherent uncertainty necessitates careful consideration of thresholds and error rates in biometric systems.

2.1 Deepfake

Deepfakes represent a sophisticated application of deep learning and neural networks, capable of generating synthetic media—images, videos, or audio—that convincingly mimic real individuals and events. While the term "deepfake" emerged in 2017 amid a proliferation of face-swapped videos on social media, the technological foundation for this capability was established decades earlier. The early 1990s saw crucial advancements in facial recognition and motion tracking, but the transformative development came in 2014 with Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). GANs operate through a sophisticated interplay between two neural networks: a generator that creates synthetic data and a discriminator that evaluates the output's authenticity. Through millions of training cycles, this adversarial process produces increasingly sophisticated generators capable of creating highly realistic synthetic content, while simultaneously developing discriminators that can detect such artificial content with remarkable accuracy.

The evolution of deepfake detection has paralleled these advancing creation capabilities. Early detection methods relied primarily on spectral analysis techniques, which offered computational efficiency but proved inadequate against sophisticated modern deepfakes that could replicate natural speech characteristics with

high fidelity. The field progressed to traditional machine learning classifiers, which improved detection rates but struggled to adapt to rapidly evolving generation techniques. The current state of the art centers on transformer-based architectures, which have demonstrated superior capability in identifying emerging deepfake technologies across varying audio quality levels. These modern approaches excel particularly in capturing long-range dependencies within audio signals, enabling the detection of subtle artifacts that betray synthetic generation.

What began as an entertainment technology quickly revealed more concerning applications. Malicious actors have exploited deep-fake capabilities to bypass authentication systems, create fraudulent recordings, and manipulate evidence. The impact extends far beyond immediate financial losses, creating ripples of psychological distress, social disruption, and potential political instability (Westerlund, 2019). This expanding threat landscape underscores the critical importance of robust deepfake detection systems in modern security frameworks.

3. Use Cases for Voice-Based Access Control

Voice-based access control systems serve diverse domains, from home automation to high-security environments. While generally used as a secondary factor, voice biometrics can be primary in certain scenarios.

Physical Access in Medical Institutions: In healthcare settings, where maintaining hygiene is critical, voice-based access control can reduce the need for physical contact with surfaces like door handles. Hospitals and clinics often employ badge-based access control as a primary authentication mechanism. Voice recognition can complement this by enabling touchless door operation, providing a secondary authentication factor for critical areas, and verifying the identity of the badge holder. While facial recognition offers similar benefits, voice-based authentication may provide a more seamless user experience, as it doesn't require precise positioning in front of a camera.

Real-Time Voice-Based Identification on Voice Calls: Voice biometrics can enhance security and mitigate fraud in telecommunications. By enrolling voice prints of phone number owners, telecom providers can provide real-time identification during calls, displaying confidence levels to the recipient. This can help combat spam and spoofing attacks. Furthermore, incorporating deepfake detection can alert users to potential manipulation attempts. Speaker diarization, which separates individual voices in a multi-party conversation, allows for identifying participants, verifying their authenticity, and even detecting potential collusion attempts in settings like correctional facilities, where preventing unauthorized communication between inmates is crucial.

Veracity Validation of Digital Evidence: Audio and video recordings are frequently used as evidence in legal proceedings. Voice-based identification, coupled with deep-fake detection, can assess the authenticity of such recordings, providing valuable insights into their reliability and integrity.

Transaction Approvals: Hands-free transaction authorization is vital when manual input is impractical or unsafe, as while driving. A secure implementation might involve a banking app sending a dynamic one-time passphrase (OTP) to the user’s phone, which then prompts the user to speak the code for voice-based verification. This combines "something you have" (phone) with "something you are" (voice). The system can then perform voice identification and deepfake detection to provide data for the transaction approval.

Figure 1 describes how the multiple components of this voice-based transaction approval platform could interact with each other to implement this voice-based transaction approval flow. Let’s dive deeper into these interactions to extract the rationale for the architecture implementation.

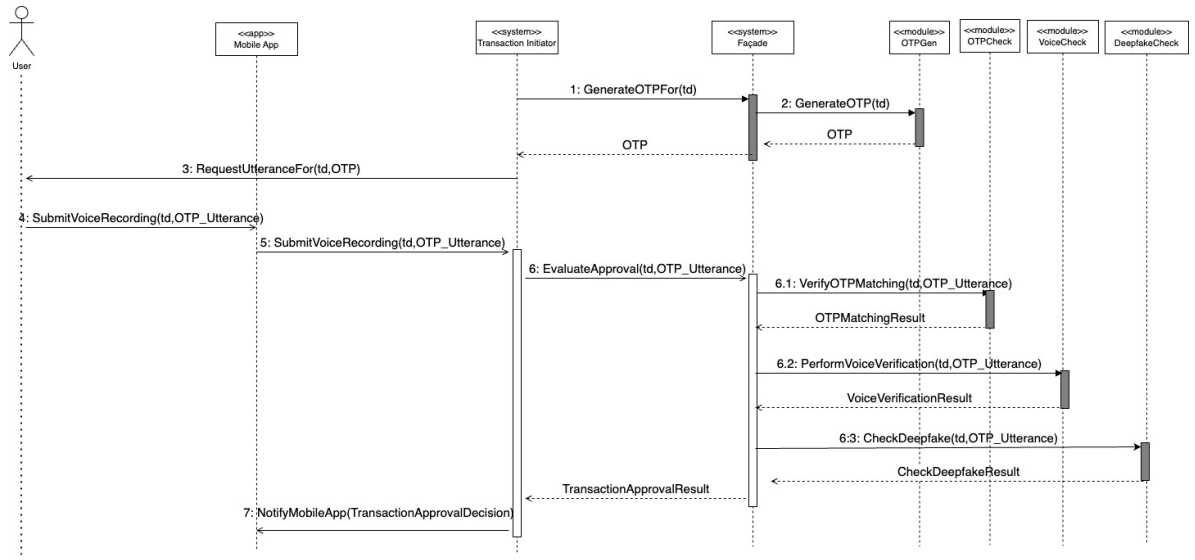

Figure 1: Sequence diagram for a voice-based transaction approval

In the figure, the "Transaction Initiator" is any external system that wants to leverage the platform. Here is how the process happens:

- *Request 1 – GenerateOTPFor(td*) : It initiates the interaction by asking the platform to generate an OTP based on the transaction data (td). Let's consider that td contains a user identifier (uid), a transaction identifier (ti) and a string that represents that specific transaction, containing only the information that the Transaction Initiator wants to provide to the platform.
- *Request 2 – GenerateOTP(td)* : This request is forwarded to the OTPGen module, which generates a phrase, the OTP tied to that specific transaction, which we expect the user to utter so that the Transaction Initiator can capture that utterance and submit it for evaluation. "*I approve the transaction 1234 with the code 1D3A*" is an example of a possible OTP, which contains the transaction identification 1234 (which can be part of it) and the code 1D3A can be a hash-based encoding of the provided transaction data, embedded into td.
- *Request 3 – RequestUtteranceFor(td,OTP)* : The generated passphrase is then sent to the user, for example through a text message to a registered phone number, to request the user to provide that phrase's utterance. Making it this way we are adding another authentication factor, using something that the user has.
- *Request 4 – SubmitVoiceRecording(td,OTP_Utterance)* : Some mechanism on the user's side, for example a mobile app, should capture the voice recording of the user uttering that passphrase.
- *Request 5 – SubmitVoiceRecording(td, OTP_Utterance)* : The mobile app submits the recorded utterance to the Transaction Initiator, so it can request the voice checking.
- *Request 6 – EvaluateApproval(td, OTP_Utterance)* : The Transaction Initiator will then submit that utterance to the platform, so it can process the voice recording. Three tasks can be executed in parallel then:

  - *Request 6.1 – VerifyOTPMatching(td, OTP_Utterance):* The OTP validation, which consists in performing a transcription of the voice recording and comparing it with the message associated with that transaction;
  - *Request 6.2 – PerformVoiceVerification(td, OTP_Utterance)*: The voice verification, which consists in responding with the level of similarity between the submitted recording and the corpus associated with that specific user;
  - *Request 6.3 – CheckDeepFake(td, OTP_Utterance)*: The deepfake detection, which consists in responding with the probability level with what that voice recording can be deepfake-generated.

The combination of these three results is used by the Transaction Initiator to decide if the transaction should be approved or not, based on the risk management strategy implemented. The platform, itself, does not decide, only informs.

The design and implementation of this architecture will be explored in detail in the subsequent section.

4. A Serverless Architecture for Voice-Based AuthN and AuthZ

This section presents an architecture for voice-based transaction approvals, focusing on authentication. Transaction management and user data handling are assumed to reside within external systems, promoting flexibility and privacy compliance.

The architecture provides two primary functions:

- Registration: This function enables external systems to register voice recordings associated with specific users, identified by unique user identifiers (uid). The architecture assumes that the external system has established the legitimacy of the provided voice recording and its association with the uid. The architecture's primary responsibility during registration is to ensure the recording's authenticity by employing deepfake detection.
- Verification: This function verifies whether a given voice recording corresponds to a registered user.

The architecture supports optional One-Time Passwords (OTPs) for diverse use cases.

4.1 Core Modules

Three core modules comprise the architecture:

- *OTP Generation*: This module generates OTPs by hashing a combination of transaction data (provided as a string by the requesting system), transaction identifier (tid), user identifier (uid), and request timestamp. The last four characters of the hash serve as the OTP. This module is optional.
- *OTP Verification*: This module uses an AI model responsible for transcribing the audio sent by the user, and to compare the transcription against the OTP sent to the user. This module will only be used if the OTP is provided.
- *Voice Verification*: This module employs an AI model trained to identify voiceprints. During registration, a voice recording and the corresponding uid are used to create a voiceprint for the user. Subsequently, this module compares new voice recordings against the registered voiceprint for verification.
- *Deepfake Detection*: This module utilizes an AI model trained to detect voice forgery. It is employed during registration to prevent the enrollment of fraudulent recordings and during verification to ensure the authenticity of the presented sample.

The architecture leverages serverless Amazon Web Services (AWS) services for scalability and cost-effectiveness. Figure 2 illustrates the implementation and will be detailed in the following section.

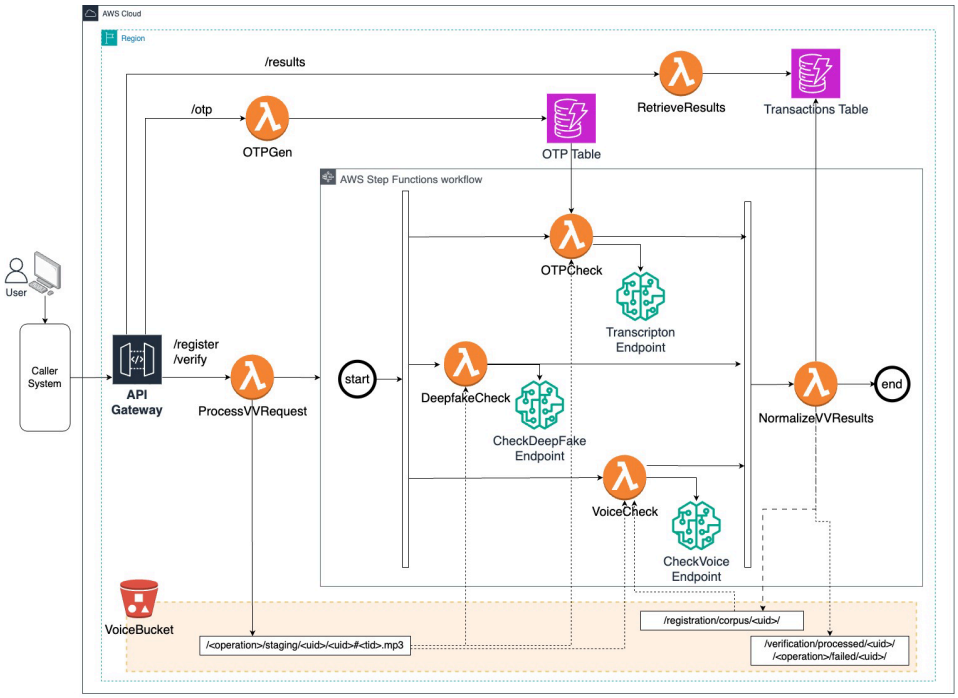

Figure 2: The architecture implemented using AWS services

4.2 Architecture Description

This section details our architecture's components and their interactions, including explanations of the AWS services used to help readers unfamiliar with cloud infrastructure.

4.2.1 System interface layer

Amazon API Gateway is a managed service that provides a REST API interface, acting as the façade for external system interactions. It processes incoming requests through the following endpoints:

- /otp: Generates One-Time Passwords (OTPs) based on transaction data
- /register: Handles user registration with the following parameters:
    - Transaction ID (tid)
    - User ID (uid)
    - Voice recording
    - Optional OTP
- /verify: Performs voice verification using:
    - Transaction ID (tid)
    - User ID (uid)
    - Voice recording
    - Optional OTP
- /results: Retrieves transaction data using tid as parameter

4.2.2 Processing layer

AWS Lambda Functions – represented by the circles with the Lambda symbol inside them – are serverless computing resources that execute code on-demand in response to events. They handle various processing tasks and can interact with databases, storage services, and external systems through multiple protocols. Our architecture employs the following functions:

1. Request Processing
    a. ProcessVVRequest:
        i. Receives API Gateway requests
        ii. Extracts Base64-encoded voice recordings
        iii. Manages storage in VoiceBucket
        iv. Triggers appropriate state machine

2. Authentication Functions
    a. OTPGen: Creates and stores transaction-specific OTPs
    b. OTPCheck: Validates OTPs through audio transcription
    c. DeepfakeCheck: Performs deepfake detection
    d. VoiceCheck: Conducts voice verification
3. Results Management
    a. NormalizeVVResults:
        i. Combines authentication results
        ii. Manages database storage
        iii. Handles voice recording file management

b. RetrieveResults: Provides transaction result access

4.2.3 AI/ML infrastructure

Amazon SageMaker is a cloud-based machine learning platform that enables the creation, training, and deployment of ML models. It supports both cloud and edge device deployments, providing: Serverless ML model hosting; first-call higher latency characteristics, with subsequent low-latency response; flexible model deployment options. In the diagram, the ML models are the Transcription, CheckDeepfake, and CheckVoice Endpoints.

4.2.4 Workflow orchestration

AWS Step Functions, which is represented in the diagram by a box around the authentication functions, is a serverless orchestration service that coordinates multiple service interactions. It manages:

- Parallel execution of verification checks (OTPCheck, DeepfakeCheck, and VoiceCheck)
- Results aggregation
- Error handling and recovery
- Process flow management.

4.2.5 Data storage layer

The architecture utilizes two primary storage services:

1. Amazon Simple Storage Service (S3) – VoiceBucket - A cloud-scale, high-performance serverless object storage service that stores files - the voice recordings, in our case, in logical containers called "buckets", and organizes them using a folder-like structure.
2. Amazon DynamoDB - A cloud-scale serverless NoSQL database delivering single-digit-millisecond latency, which manages:

- OTP Table: Stores and retrieves active OTP records.
- Transaction Table: Maintains transaction results.

4.2.6 Performance metrics

- Processing time: Under five seconds per request
- Cost efficiency: US$0.012 per transaction
- Scalability: Automatic based on demand

This architecture balances performance, scalability, and costs while ensuring security and reliability. The serverless nature of the components ensures efficient resource utilization and automatic scaling based on demand.

5. Model Selection and Technical Implementation

Selecting and implementing machine learning models for voice-based authentication requires balancing performance requirements, regulatory compliance, and system integration capabilities. Our approach builds upon recent advances in speech processing while addressing the practical constraints of real-world deployment in security-critical applications.

5.1 Core Model Selection

The system's functionality relies on three primary machine learning components, each selected through rigorous comparative analysis of current voice authentication methodologies. This analysis revealed distinct tradeoffs among existing approaches. Traditional GMM-UBM and i-vector systems (Dehak et al., 2011) demonstrate reliability in controlled environments but struggle with real-world variability and demand significant computational resources for training. More recent end-to-end deep learning models (Snyder et al., 2018) achieve impressive accuracy but require substantial computational resources and extensive training datasets, making them impractical for dynamic deployment scenarios.

For voice verification, our analysis led us to implement the Resemblyzer framework, which represents an optimal balance between performance and resource utilization. This framework employs zero-shot audio embedding with Mel-frequency Linear Prediction Cepstral Coefficients (MLFCC) for speech feature extraction, building upon recent advances in speaker verification technology (Wan et al., 2017). The zero-shot approach

offers remarkable flexibility in enrollment scenarios and reduced computational overhead, enabling authentication without extensive pre-enrollment training—a crucial advantage for rapid deployment and dynamic user bases.

While proprietary solutions like Amazon Voice ID offer comprehensive functionality, their tight integration with specific service stacks (in this case, Amazon Connect) would introduce unnecessary architectural constraints. Our implementation of Resemblyzer prioritizes modularity and flexibility, simplifying integration while maintaining high performance standards. This approach particularly excels in low-resource deployment scenarios, where computational efficiency and adaptability are paramount.

The deepfake detection component employs Meta's Wav2vec2 transformer-based model (Baevski et al., 2020; Conneau et al., 2020), which we adapted for audio classification through fine-tuning on the ASVSpoof19 and ASVSpoof21 datasets. This implementation achieves 92.86% accuracy in our validation tests, a crucial metric given the rapidly evolving nature of deepfake technology. As Westerlund (2019) notes, the continuous advancement of synthetic voice generation necessitates an adaptable detection architecture. Our implementation allows for rapid model updates as new attack vectors emerge, maintaining system security in the face of evolving threats.

For transcription functionality, we selected OpenAI's Whisper model after extensive comparative testing. Our analysis revealed significant performance advantages over the current implementation of traditional cloud-based solutions:

- Whisper-base achieves average latency of 2.3 seconds
- Whisper-medium performs at 4.6 seconds
- Amazon Transcribe batch processing requires 7.2 seconds

These metrics reflect end-to-end processing time including network overhead, making them particularly relevant for real-world deployment scenarios.

5.2 Technical Implementation Framework

Our implementation sets technical standards that ensure consistent performance across diverse deployment scenarios. The system standardizes on 16kHz sampling rate, aligning with both common practice in audio machine learning research and typical microphone capabilities in modern devices. Voice sample processing accommodates durations between five to 15 seconds, with support for longer samples through a sliding window analysis approach. This duration range, supported by recent research in biometric system usability (Bonneau et al., 2012), optimizes the balance between user experience and model accuracy requirements.

Performance validation across diverse user demographics revealed several critical factors affecting system accuracy:

- Recording Equipment Variation: Microphone quality and characteristics significantly impact feature extraction reliability.
- Environmental Conditions: Background noise and acoustic environments affect signal clarity.
- User Characteristics: Accent variations and language differences influence transcription accuracy. (Chen et al., 2022)
- Signal Processing: Sample rate conversion and audio compression artifacts can affect model performance.

We addressed these variations through threshold management in the normalization phase, extending Bhattacharyya et al.'s (2009) work on biometric system adaptation. This approach allows for demographic-specific tuning while maintaining overall system security.

5.3 Infrastructure and Compliance Framework

The architecture aligns with current regulatory requirements for biometric data handling. Our deployment achieves consistent sub-five second latency under normal operating conditions, with maximum latency of 60 seconds during cold starts. This performance profile aligns with Nielsen's usability guidelines (Nielsen, 1993), which emphasize the importance of responsive feedback in maintaining user engagement and trust.

General data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA) compliance requirements significantly influenced our architectural decisions, particularly regarding biometric data handling. The implementation includes:

- Secure Identity Management: Universally Unique Identifier (UUID)-based user profiles separate biometric data from personally identifiable information.
- Data Protection: AWS-managed encryption provides security for data at rest.
- Transmission Protection: Secure Sockets Layer (SSL) encryption ensures data protection during transit.
- Access Control: Comprehensive audit logging and strict access controls maintain data integrity.
- This framework provides a foundation for secure biometric authentication while maintaining the flexibility required for integration with existing security infrastructure. Performance metrics demonstrate that these security measures introduce minimal overhead, preserving the system's responsive user experience while ensuring regulatory compliance.

5.4 Further Improvements

While our current architecture demonstrates robust performance in controlled environments, implementation experience and recent advances in speech processing technology suggest several promising avenues for enhancement. Analysis of industry trends and benchmarks reveals three key areas for development to expand system capabilities and resilience.

5.4.1 Speaker diarization integration

Our single-speaker optimized architecture has limitations in multi-speaker scenarios and noisy environments. Recent advances in speaker diarization, particularly the development of joint training approaches for speaker diarization and speech separation (Kalda et al., 2024), offer promising solutions to these challenges. Building upon the PyAnnote framework's neural building blocks for speaker diarization (Bredin et al., 2020), we propose an enhanced system incorporating real-time speaker-isolation capabilities.

This enhancement would introduce three integrated capabilities to address current limitations. The system would employ advanced voice activity detection to isolate speech segments from background noise, enabling sustained accuracy in noisy environments. Speaker segmentation would facilitate the identification and separation of individual speakers within continuous audio streams. Finally, real-time speaker embedding would enable dynamic speaker profile matching, crucial for maintaining security in multi-speaker environments.

To validate this enhancement, we will employ standard diarization benchmarks including the DIHARD III Challenge dataset and the AMI Meeting Corpus. Our implementation targets a Diarization Error Rate (DER) below 5% in mixed-speaker environments, while maintaining the system's current sub-5-second latency requirements. Performance evaluation will encompass speaker confusion rate, missed speech rate, and false alarm rate metrics.

5.4.2 Continuous authentication through stream processing

We propose stream processing capabilities to enable continuous authentication beyond discrete events. This enhancement transforms our current model into a continuous verification system, particularly valuable for extended secure sessions and ongoing transaction sequences.

The proposed streaming architecture comprises three key components working in concert. A sliding window analysis approach with configurable overlap enables continuous monitoring while preserving computational efficiency. Real-time voice feature extraction and verification provides immediate detection of unauthorized speakers or security breaches. Dynamic speaker embedding clustering facilitates adaptive speaker profile updates while maintaining security integrity.

Validation of this enhancement will utilize the VoxCeleb2 dataset for continuous speaker verification, with specific performance targets including a Maximum Latency Threshold (MLT) of 500ms for speaker switching detection and an Equal Error Rate (EER) below 1%. We will establish resource utilization benchmarks through AWS CloudWatch metrics to ensure sustained cost-effectiveness at scale.

5.4.3 Global fraud detection framework

The growing sophistication of voice deepfake technology (Westerlund, 2019) requires a comprehensive fraud detection approach. Our proposed global fraud detection framework leverages collective intelligence across deployment instances while maintaining strict privacy controls.

The framework incorporates three innovative elements that work together to enhance security. A secure corpus of failed verification attempts, anonymized and encrypted to comply with GDPR and HIPAA

requirements, enables pattern recognition across attack attempts. Advanced attack pattern profiling facilitates the identification of emerging threats and attack methodologies. A cross-platform threat intelligence sharing mechanism coordinates responses to new attack vectors while maintaining deployment isolation.

Framework effectiveness will be measured using the ASVspoof 2021 dataset combined with real-world attack patterns from deployment instances. Key performance indicators include an Attack Detection Rate (ADR) above 99% and False Alarm Rate (FAR) below 0.1%, with ongoing evaluation against emerging attack vectors from recent ASVspoof challenges.

5.4.4 Integration considerations

These enhancements must fit within existing architectural constraints while maintaining performance metrics. Success requires careful attention to four critical factors: maintaining low latency response times for primary authentication flows, ensuring regulatory compliance across all new data collection and processing activities, preserving the system's modular architecture for selective feature deployment, and optimizing resource utilization to maintain cost-effectiveness at scale.

Through careful implementation of these improvements, we can significantly enhance the system's capabilities while preserving its core strengths in security, scalability, and user experience. Each enhancement builds upon our existing foundation while addressing emerging challenges in voice-based authentication and authorization.

6. Conclusions

This paper presents an architecture for voice-based authentication that addresses deepfake attacks while maintaining usability. Our implementation demonstrates several key contributions to the field of biometric security:

First, the integration of state-of-the-art models for voice verification, deepfake detection, and transcription achieves high accuracy (99.73% for deepfake detection) while maintaining low latency (sub-5-second processing). This performance level makes the system suitable for real-world deployment across various use cases, from healthcare access control to financial transaction authorization.

Second, our serverless architecture provides a scalable, cost-effective solution for voice authentication. The implementation leverages AWS services to achieve both operational efficiency (US$0.012 per transaction) and system reliability. The modular design allows for independent scaling of components and seamless integration with existing authentication frameworks.

Third, our approach to regulatory compliance demonstrates that robust security measures can coexist with strict privacy requirements. The implementation of UUID-based user profiles, encrypted storage, and comprehensive audit logging provides a framework for handling biometric data while maintaining GDPR and HIPAA compliance.

Future improvements in speaker diarization, stream processing, and global fraud detection will expand capabilities while preserving core strengths. These enhancements would address emerging challenges in voice authentication while preserving the architecture's fundamental benefits of scalability, security, and cost-effectiveness.

As voice-based authentication continues to gain prominence in security applications, the importance of protecting against deepfake attacks will only increase. Our architecture provides a foundation for building secure, scalable voice authentication systems that can adapt to evolving threats while maintaining usability and performance. Future work will focus on implementing the proposed improvements and expanding the system's capabilities to address emerging security challenges in voice biometric authentication.

References

Baevski, A., Zhou, H., Mohamed, A. and Auli, M. (2020). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. [online] arXiv.org. doi: https://doi.org/10.48550/arXiv.2006.11477.

Bhattacharyya et Al, (2009). Bhattacharyya, Debnath & Ranjan, Rahul & Alisherov, Farkhod & Minkyu, Choi. (2009). Biometric Authentication: A Review. International Journal of Science and Technology.

Bishop, M. (2003). *Computer security : art and science*. Boston, Mass. ; London: Addison-Wesley.

Bonneau, J., Herley, C., Oorschot, P.C. van and Stajano, F. (2012). *The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes*. [online] IEEE Xplore. doi: https://doi.org/10.1109/SP.2012.44.

Bredin, H., Yin, R., Juan Manuel Coria, Gelly, G., Pavel Korshunov, Lavechin, M., Fustes, D., Hadrien Titeux, Wassim Bouaziz and Gill, M.-P. (2020). Pyannote.Audio: Neural Building Blocks for Speaker Diarization. *HAL (Le Centre pour la Communication Scientifique Directe)*. doi: https://doi.org/10.1109/icassp40776.2020.9052974.

Chen, X., Li, Z., Setlur, S. and Xu, W. (2022). Exploring racial and gender disparities in voice biometrics. *Scientific Reports*, [online] 12(1), p.3723. doi: https://doi.org/10.1038/s41598-022-06673-y.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A. and Auli, M. (2020). Unsupervised Cross-lingual Representation Learning for Speech Recognition. *arXiv (Cornell University)*. doi: https://doi.org/10.48550/arxiv.2006.13979.

Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P. and Ouellet, P. (2011). Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*, [online] 19(4), pp.788–798. doi: https://doi.org/10.1109/tasl.2010.2064307.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative Adversarial Networks. *arXiv (Cornell University)*, 1. doi: https://doi.org/10.48550/arxiv.1406.2661.

Grassi, P.A., Garcia, M.E. and Fenton, J.L. (2017). Digital identity guidelines: revision 3. *NIST Special Publication 800-63-3*. [online] doi: Joonas Kalda, Clément Pagés, Marxer, R., Tanel Alumäe and Bredin, H. (2024). PixIT: Joint Training of Speaker Diarization and Speech Separation from Real-world Multi-speaker Recordings. *arXiv (Cornell University)*, pp.115–122. doi: https://doi.org/10.21437/odyssey.2024-17.

Nielsen, J. (1993). *Usability engineering*. [online] Cambridge, Mass.: Ap Professional. Available at: https://dl.acm.org/citation.cfm?id=2821575.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. and Khudanpur, S. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. doi: https://doi.org/10.1109/icassp.2018.8461375.

Trusted Computing Group (n.d.). *TPM 1.2 Main Specification*. [online] Trusted Computing Group. Available at: https://trustedcomputinggroup.org/resource/tpm-main-specification/.

Wan, L., Wang, Q., Papir, A. and Moreno, I. (2017). Generalized End-to-End Loss for Speaker Verification. *Arxiv (Cornell University)*. doi: https://doi.org/10.48550/arxiv.1710.10467.

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, [online] 9(11), pp.39–52. doi: https://doi.org/10.22215/timreview/1282.