2023

5(1), 96-114

A Review of Recent Machine Learning Approaches for Voice Authentication Systems

Zuhal Can*[1], Emrah Atılgan[2]

Keywords

Voice authentication
Audio signal processing
Machine learning
Security attacks

Received
May 11, 2023
Accepted
June 26, 2023
Published
June 28, 2023
Article Type
Review Paper

Abstract

Voice authentication systems are a comfortable way of protection since users do not need to remember passwords or carry identification cards. As a unique identifier for all individuals, voice is a practical tool to authenticate people into security services, including online banking and phone-based customer or computer services. Single-model voice authentication systems refer to voice recognition systems that utilize a single voice model to verify the identity of individuals based on their unique vocal characteristics, such as pitch, tone, and other speech patterns. For multi-model voice authentication systems, additional biometric factors like facial recognition or electroencephalogram data are included in the voice authentication process to enhance security. This paper reviews recent single-modal and multimodal voice authentication studies with an explanation of underlying feature extraction and classification methods. This paper also discusses security attacks on voice authentication systems, including random attacks, mimicry attacks, replay attacks, voice synthesizing attacks, counterfeit attacks, and hidden voice command attacks.

Sesli Kimlik Doğrulama Sistemleri için Makine Öğrenimi Yaklaşımlarının İncelenmesi

Anahtar Sözcükler

Sesli kimlik doğrulama
Ses sinyali işleme
Makine öğrenme
Güvenlik saldırıları

Gönderim Tarihi
11 Mayıs 2023
Kabul Tarihi
26 Haziran 2023
Yayın Tarihi
28 Haziran 2023
Makale Türü
Derleme Makalesi

Öz

Sesli kimlik doğrulama sistemleri, kullanıcıların parolaları hatırlamaları veya kimlik kartları taşımaları gerekmediği için rahat bir koruma yöntemidir. Tüm bireyler için benzersiz bir tanımlayıcı olarak ses, çevrimiçi bankacılık ve telefon tabanlı müşteri veya bilgisayar hizmetleri dahil olmak üzere güvenlik hizmetlerinde kişilerin kimliğini doğrulamak için pratik bir araçtır. Tek modelli ses kimlik doğrulama sistemleri, bireylerin kimliğini perde, ton ve diğer konuşma kalıpları gibi benzersiz ses özelliklerine dayalı olarak doğrulamak için tek bir ses modeli kullanan ses tanıma sistemlerini ifade eder. Çok modelli ses kimlik doğrulama sistemleri için, güvenliği artırmak amacıyla ses kimlik doğrulama sürecine, yüz tanıma veya elektroensefalogram verileri gibi ek biyometrik faktörler dahil edilir. Bu makale, yakın zamandaki tek modelli ve çok modelli ses kimlik doğrulama çalışmalarını, özellik çıkarma ve sınıflandırma yöntemlerinin altında gözden geçirmektedir. Bu makale aynı zamanda rastgele saldırılar, taklit saldırıları, tekrarlama saldırıları, ses sentezleme saldırıları, sahte saldırılar ve gizli sesli komut saldırıları dahil olmak üzere ses kimlik doğrulama sistemlerine yönelik güvenlik saldırılarını tartışmaktadır.

***Atıf*:** Can, Z. & Atılgan, E. (2023). Sesli kimlik doğrulama sistemleri için makine öğrenimi yaklaşımlarının incelenmesi, *Bilgi ve İletişim Teknolojileri Dergisi, 5*(1), 95-113. https://doi.org/10.53694/bited.1296035

***Cite*:** Can, Z. & Atilgan, E. (2023). A review of recent machine learning approaches for voice authentication systems, *Journal of Information and Communication Technologies*, *5*(1), 95-113. https://doi.org/10.53694/bited.1296035

***Sorumlu Yazar/Corresponding Author**:
[1] Eskişehir Osmangazi University, Faculty of Engineering and Architecture, Department of Computer Engineering, Eskişehir, Turkey, zcan@ogu.edu.tr, https://orcid.org/0000-0002-6801-1334
[2] Eskişehir Osmangazi University, Faculty of Engineering and Architecture, Department of Computer Engineering, Eskişehir, Turkey, emrah.atilgan@ogu.edu.tr, https://orcid.org/0000-0002-0395-9976

Introduction

Biometrics is the unique and measurable characteristics of people used to describe and distinguish individuals that have multiple application areas in various domains such as security systems (Beranek, 2013; Boubchir & Daachi, 2021), forensics (Tistarelli & Champod, 2017), health care (Bhalla, 2020), online banking (Goode, 2018) and customer care services. As demonstrated in Figure 1, in the literature, biometrics are divided into physiological biometrics and behavioral biometrics (Dargan & Kumar, 2020; Gayathri, Malathy, & Prabhakaran, 2019). Physiological biometrics are gained genetically, such as DNA, face, iris, ear shape, electrocardiogram (EEG), palm vein, and hand geometry. Behavioral biometrics are formed by human activity patterns, such as voice keystroke dynamics, signature, and gait. Some of these biometrics can be only used in special cases, such as DNA matching in criminal cases, etc., and are not practical for real-time personal authentication.

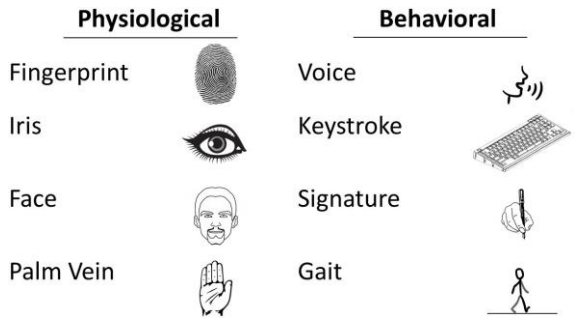

Figure 1. Classification of Common Biometrics

The human voice is in the category of behavioral biometrics. Behavioral patterns and vocal components (lungs and articulators) form a unique and distinctive voice for all individuals that can be distinguished in terms of pitch, volume, timbre, accent, rhythm, or tone (Yoshida, 2012). The usage of vocal components such as vocal vibrations, breathing noise, and articulatory gestures creates a personal ID similar to other biometric identifiers for people.

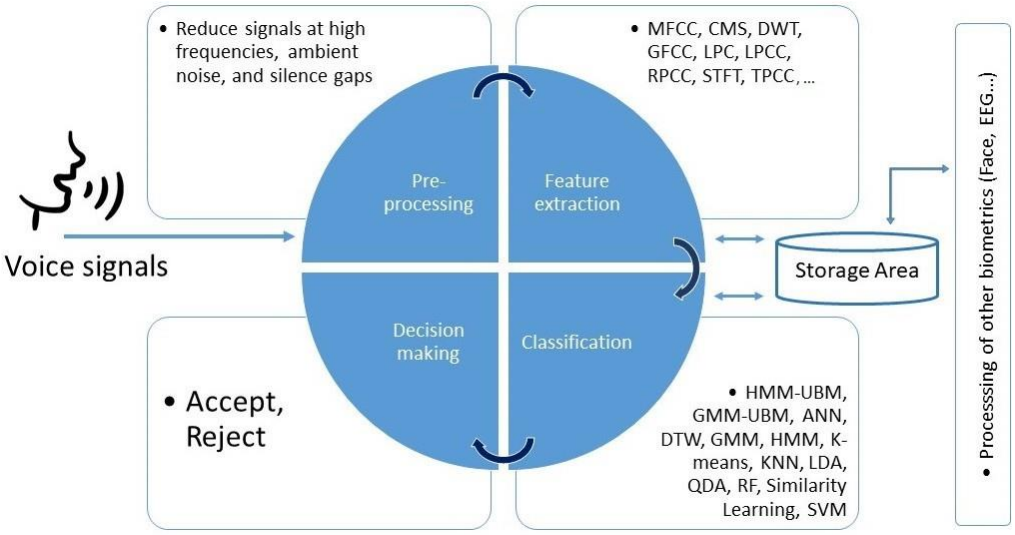

Figure 2. General Framework of Voice Authentication Systems

Voice authentication enables people to access secure services using their voices. In voice authentication systems, the user's speech is acquired by system microphones as voice signals and processed to verify the user's authentication. For example, by voice authentication integrated online banking systems, the previously recorded voice of a user is stored in the server for further access (Fairhurst, Li, & Da Costa-Abreu, 2017). When the customer wants to make a new transaction in the future, it is checked whether it matches the voice recorded in the system, and authentication is performed. As with all authorized entry systems, security systems are needed for voice authentication.

From signal acquisition to voice verification, voice signals go through pre-processing, feature extraction, classification, and decision phases, as demonstrated in Figure 2. A single-modal biometric authentication system processes only one type of biometrics data, whereas a multimodal biometrics system processes different biometric data together for user authentication and improves security comparatively.

In this study, recent single-modal and multimodal voice authentication works are reviewed, and security attacks for voice authentication systems are explained. This study focuses on traditional machine learning approaches for voice authentication systems and does not include deep learning studies on this subject. In the following sections, recent machine learning studies on voice authentication systems are discussed based on their feature extraction and classification methods.

Method

Voice-recognition systems have taken place in our lives with the technological development in voice-assisted technologies, such as Google Assistant, Google Home, Apple Siri, Microsoft Cortana, and Amazon Alexa. According to Juniper Research, the number of devices using digital voice assistants will reach 8 billion in 2023 (Juniper, 2022). Based on Statistica.com, the global market of voice recognition will reach 27.16 billion USD in 2026 (Statistica.com, 2021).

With technology development, many transactions go through the web or smartphones without arriving at the institutions personally. Remembering usernames and passwords becomes challenging as the number of authenticated websites increases. Writing down passwords for fear of forgetting them brings up significant security risks. Therefore, researchers have focused on developing user-friendly and safe security methods such as authentication through voice or voice as a part of multi-factor authentication.

In recent years, voice authentication has gained popularity in many areas, such as e-commerce, online banking, and healthcare. For example, banks provide services through customer voice confirmations, and ATM devices have started to operate by listening to the voices of their customers without the necessity of swiping their cards (Kaman, Swetha, Akram, & Varaprasad, 2013). For many years, voice recognition and speech-to-text methods have been used in the healthcare industry. Physicians record their reports about their patients audibly, and computers convert them to text to save time and energy. On the other hand, voice authentication is a new application for the health sector as in other sectors. Due to the importance of health information security, authentication is a critical process in the healthcare industry. While it is necessary to be in the institution personally for other biometric verification processes (face, iris, fingerprint, etc.), patients do not need to go to the hospital or doctor's office for voice authentication, which is especially convenient for elderly or disabled patients.

Yet, consumers are concerned about confidence due to low verification accuracy and vulnerability to malicious attacks (de Barcelos Silva et al., 2020; Smallman, 2020). Moreover, voice verification accuracy is affected by several factors, including ambient noise, utterance length, sampling frequency, and hardware quality. Moreover, change in voice due to diseases, accidents, aging, and environmental factors affect the verification accuracy. The security and accuracy of voice authentication systems are enhanced by developing multi-model systems using various biometrics. Underlying techniques of voice authentication systems are discussed in many literature papers under speech identification, verification, and recognition topics (Tirumala, Shahamiri, Garhwal, & Wang, 2017; Dişken, Tüfekçi, Saribulut, & Çevik, 2017; Singh, Nath, & Kumar, 2018; Jayanna, Mahadeva Prasanna, & Prasanna, 2009; Jayanna et al., 2009; Todkar, Babar, Ambike, Suryakar, & Prasad, 2018; Lawson et al., 2011; Sidorov, Schmitt, Zablotskiy, & Minker, 2013).

Traditional voice authentication systems consist of four phases; pre-processing, feature extraction, classification, and decision-making. In the pre-processing phase, voice signals are processed before the feature extraction phase. In this phase, voice signals are separated from unnecessary noises to improve system accuracy. Human perceives sounds at frequency bands ranging between 20 and 20 kHz, and human voice frequency is lower than 3400 Hz. In the pre-processing phase, signals at high-frequency bands (for example, 20 kHz and more) are removed from acquired voice signals. Moreover, ambient noise and silence gaps are reduced from acquired voice data. Since ambient noise is at the frequency ranges of a human voice, reducing ambient noise from voice signals is not always practically possible; therefore, voice signals may still include ambient noise after the pre-processing phase. In the feature extraction phase, various temporal (e.g., energy levels, zero-crossing rate, amplitude levels) and spectral features (e.g., frequency components, spectral energies, spectral densities) are extracted from voice signals.

In the classification phase, extracted features of previously recorded voice signals are classified and stored in the system as voice templates. The next phase after the classification phase is the decision phase. The decision phase aims to find a match between the user voice and the previously stored voice templates. According to the matching ratio, authentication is accepted or rejected. The system is evaluated based the performance metrics such as Accuracy (ACC), True Positive Rate (TPR), False Reject Rate (FRR), and Equal Error Rate (ERR).

In voice authentication systems, selecting appropriate feature extraction and classification methods is essential for improving system accuracy. This study reviews voice authentication systems in the literature under single-model and multi-model systems with several details, including feature extraction and classification methods, assisted biometrics, number of participants, and system performance. The feature extraction and classification methods applied to these systems are listed and explained in detail in this study.

Findings

Single-modal and multimodal studies are sorted by year and listed in Table 1. A single-model voice authentication system uses only voice as the biometric data. Several single-model studies are focusing on voice data for human recognition in literature. In the AVA (Z. Meng et al., 2020) study, an average windowed-based equal error rate of 3-4%, depending on the model, is achieved based on 25 speaker records. Voicelive (L. Zhang et al., 2016) presents a liveness detection system for voice authentication on smartphones and reaches 99% detection accuracy based on

Table 1. Single-modal and Multimodal Voice Authentication Systems

Study	Feature Extraction Methods	Classification Methods	Other Biometrics	Number of Users	Performance
EchoVib (Abhishek Anand et al., 2021)	MFCC, Time-Frequency Analysis	SVM, LR, RF, RT	Vocal Vibrations	30	ACC: 90%
Moreno-R. et al. (Moreno-Rodriguez, Ramirez-Cortes, Atenco-Vazquez, & Arechiga-Martinez, 2021)	MFCC	HMM-UBM, ANN, MAP	EEG	30	ACC: 83.43%
WearID (Shi, Wang, Chen, & Saxena, 2021)	Time-Frequency Analysis	Similarity Learning	Speech Vibrations	10	TPR: 99%
VocalPrint (Li et al., 2020)	GFCC[1], RPCC, TPCC, DCT, MFCC, LPC, LPCC	GMM-UBM, SVM, HMM, MAP	Vocal Vibrations	41	ACC: 96%
Zhang et al. (X. Zhang, Cheng, Jia, Dai, & Xu, 2020)	MFCC, LBP, DWT	VAD, GMM, MAP	Face	102	FRR (Voice): 10.98% FRR (Face): 1.22% FRR (Fusion): 0%
LVID (Wu, Yang, Zhou, Chen, & Wang, 2019)	MFCC, DCT	VAD, GMM	Lip Movements	104	ACC: 95%
AVA (Z. Meng, Altaf, & Juang, 2020)	MFCC	K-means Clustering, MAP, MVE, HMM	NA	25	'EER(window size of 1.01 s): 5%
Abozaid et al. (Abozaid, Haggag, Kasban, & Eltokhy, 2019)	Cepstral and statistical coefficients, PCA, Eigenfaces	GMM, ANN, SVM	Face	100	EER (Voice): 2.98% EER (Face): 1.43% EER (Fusion): 0.62%
VoicePop (Q. Wang et al., 2019)	STFT, GFCC[2], PCA, DCT	HMM, SVM	Breathing Noise	18	ACC: 93.5%
Olazabal et al. (Olazabal et al., 2019)	MFCC	KNN	Face	27	EER (Fusion): 8.04%
LipPass (Lu et al., 2019)	DNN	SVM, SVDD	Mouth Movements	48	ACC: 90.2%
Gofman et al. (M. Gofman et al., 2018)	MFCC, HOG	LDA, QDA, SVM, RF, KNN	Face	27	EER (Voice): 54.21% EER (Face): 47.51% EER (Fusion): 20.59%
VoiceGesture (L. Zhang, Tan, & Yang, 2017)	STFT, DWT	Similarity Learning	Articulatory Gestures	21	ACC: 99.34%
VAuth (Feng, Fawaz, & Shin, 2017)	MFCC	SVM	On-body Vibrations	18	ACC: 97%
Voicelive (L. Zhang, Tan, Yang, & Chen, 2016)	MFCC	HMM	NA	12	EER: 1%
Gofman et al. (M. I. Gofman, Mitra, Cheng, & Smith, 2016)	MFCC, HOG	LDA, Similarity Learning	Face	54	EER (Voice): 34.72% EER (Face): 4.29% EER (Fusion): 2.14%
Yan and Zhao (Yan & Zhao, 2016)	MFCC, CMS	DTW	NA	15	ACC: 80.6%
Gofman et al. (M. Gofman, Mitra, Cheng, & Smith, 2015)	MFCC, Fisherfaces	HMM	Face	54	EER (Face): 18.70% EER (Voice): 22.42% EER (Fusion): 14.56%

Please look at Table 2 and Table 3 for expansions of abbreviations
NA: Not Applicable ACC: Accuracy TRP: True Positive Rate FRR: False Positive Rate
EER: Equal Error Rate WEER: Window-based Equal Error Rate

12 participant data. Another single-model voice authentication study (Yan & Zhao, 2016) achieves 80.6% accuracy based on speech recognition and speech synthesis on vocalized verification codes that change each time.

A multi-model voice authentication system aims to improve security and accuracy by integrating various biometrics. There are several multi-model voice authentication systems in the literature. Moreno-Rogriguez et al. (Moreno-Rodriguez et al., 2021) achieved 83.43% accuracy with their Electroencephalography (EEG) integrated voice authentication model over a study of 50 people. WearID (Shi et al., 2021) exploits speech vibrations, VocalPrint (Li et al., 2020) and EchoVib (Abhishek Anand et al., 2021) exploit vocal vibrations, and VAuth (Feng et al., 2017) exploits on-body vibrations for the voice authentication systems and achieves more than 90% accuracy. LVID (Wu et al., 2019) uses lip movements, LipPass (Lu et al., 2019) uses mouth movement, and VoicePop (Q. Wang et al., 2019) uses breathing noise in addition to voice biometrics, and they achieve accuracy rates of more than 90%, as well. Moreover, there are several multimodal voice authentication systems based on face biometrics achieving high accuracy rates (X. Zhang et al., 2020; Abozaid et al., 2019; Olazabal et al., 2019; M. Gofman et al., 2018; M. I. Gofman et al., 2016; M. Gofman et al., 2015).

Feature Extraction Methods

Table 2. Feature Extraction Methods for Voice Authentication Systems

Feature Extraction Methods
CMS: Cepstral Mean Subtraction (Yan & Zhao, 2016)
DCT: Discrete Cosine Transform (Li et al., 2020; X. Zhang et al., 2020; Wu et al., 2019)
DNN: Deep Neural Network (Lu et al., 2019; Vincent, Larochelle, Bengio, & Manzagol, 2008)
DWT: Discrete Wavelet Transform (X. Zhang et al., 2020; L. Zhang et al., 2017; Tzanetakis, Essl, & Cook, 2001)
Eigenfaces (Abozaid et al., 2019)
Fisherfaces (M. Gofman et al., 2015)
GFCC[1]: Glottal Flow Cepstrum Coefficients (Li et al., 2020)
GFCC[2]: Gammatone Frequency Cepstral Coeff. (Q. Wang et al., 2019)
HOG: Histogram of Oriented Gradients (M. Gofman et al., 2018; M. I. Gofman et al., 2016)
LBP: Local Binary Pattern (Dargan & Kumar, 2020)
LPC: Linear Predictive Coefficients (Li et al., 2020)
LPCC: Linear Predictive Cepstral Coefficients (Li et al., 2020)
MFCC: Mel Frequency Cepstral Coefficients (Abhishek Anand et al., 2021; Suman, Harish, Kumar, & Samrajyam, 2015)
PCA: Principal Component Analysis (Abozaid et al., 2019; Q. Wang et al., 2019)
RPCC: Residual Phase Cepstrum Coefficients (Li et al., 2020)
STFT: Short-Time Fourier Transform (Q. Wang et al., 2019; L. Zhang et al., 2017)
TPCC: Teager Phase Cepstrum Coefficients (Li et al., 2020)

Table 2 lists various vocal feature extraction methods in the course of voice authentication studies. MFCC is a well-known speaker recognition technique that measures pitches on a frequency scale similar to the human hearing range. MFCC-based feature extraction approaches were reported as the most commonly used and successful approaches for speaker identification (Tirumala et al., 2017). Working relatively to human voice perception by describing the signal characteristics similar to the vocal track properties is an advantage of the MFCC technique. Another common feature extraction approach is LPC. Both LPC and LPCC are vocal feature extraction method that detects emotional information. LPC-based feature extraction approaches are commonly used in automatic speaker recognition.

DNN is a neural network feature extraction method that consists of an auto-encoder network to abstract the input features. TPCC is a vocal feature extraction method that captures speakers' excitation characteristics based on the Teager energy model. RPCC characterizes the phase information of an excitation waveform, and GFCC[1] captures speakers' excitation characteristics. GFCC[2] is a vocal feature extraction method to detect pop noises, and CMS is to reduce the influence of background noise. DWT is a voice denoising method to decompose voice signals into different wavelet coefficients. DCT is an energy compaction method for calculating the cepstral coefficients. Last but not least, STFT is a signal processing method that divides the signal into shorter segments and computes the Fourier transform on each of these shorter segments.

More feature extraction methods are used to recognize other biometrics for multimodal voice authentication systems, as listed in Table 2. PCA is a statistical feature extraction method for image processing. HOG is an image processing method that counts the occurrences of gradient orientations in each image cell. LBP is an image processing method to extract image texture features. Finally, Eigenfaces and Fisherfaces are face recognition methods based on principal components of linear discriminant analyses.

Classification Methods

Table 3. Classification Methods for Voice Authentication Systems

Classification Methods
ANN: Artificial Neural Networks (Jain & Mao, 1996)
DTW: Dynamic Time Warping (Muda, Begam, & Elamvazuthi, 2010)
GMM: Gaussian mixture model (Reynolds & Rose, 1995)
HMM: Hidden Markov Model (Juang & Rabiner, 1991)
K-means Clustering (Hajarolasvadi & Demirel, 2019)
KNN: k-Nearest Neighbor (Olazabal et al., 2019; M. Gofman et al., 2018)
LDA: Linear Discriminant Analysis (M. Gofman et al., 2018; M. I. Gofman et al., 2016)
MAP: Maximum A Posteriori (Gauvain & Lee, 1994)
MVE: Minimum Verification Error (Z. Meng et al., 2020)
QDA: Quadratic Discriminant Analysis (Kwon, Chan, Hao, & Lee, 2003)
RF: Random Forests (Abhishek Anand et al., 2021; M. Gofman et al., 2018)
RT: Random Trees (Abhishek Anand et al., 2021)
LR: Logistic Regression (Abhishek Anand et al., 2021)
Similarity Learning (M. I. Gofman et al., 2016; Shi et al., 2021)
SVDD: Support Vector Domain Description (Lu et al., 2019)
SVM: Support vector machines (Campbell, Sturim, & Reynolds, 2006)
UBM: Universal Background Model (Moreno-Rodriguez et al., 2021; Li et al., 2020)
VAD: Voice Activity Detection (X. Zhang et al., 2020; Wu et al., 2019)

In Table 3, classification methods are listed based on the authentication works listed in Table 1. There are several classification methods (Abdulrahman, Khalifa, Roushdy, & Salem, 2020). ANN is a classification method containing one input and output and several hidden computational layers for speech recognition. While UBM is a classification method to detect human-specific features, SVDD is a classification method that rejects imposter components. A statistical classification method, SVM, separates different feature sets by building a hyperplane.

Another method, HMM, is a statistical Markov model used as a speech recognition classifier based on state transition probabilities. GMM is a probabilistic classification method based on Gaussian distributions of components. GMM-based classification methods are efficient for text-dependent voice identification (Tirumala & Shahamiri, 2016; Snyder, Garcia-Romero, Povey, & Khudanpur, 2017) K-means Clustering is a vector quantization method for feature classification. KNN is a classification method for storing feature vectors using a distance function. While LDA is a linear classification method, QDA is a quadratic classification method for classifying features into related sets.

Other widespread classification methods are similarity learning and RF. Similarity learning is a feature classification method based on feature comparison and scoring. And RF is a classification method that combines and corrects the outputs of several decision trees.

Classification methods are developed by various methods to improve voice recognition accuracy rates, such as DTW, MAP, MVE, and VAD. DTW is a time series analyzing method for measuring the similarity between vocal samples at different speeds. The MAP method reduces noise components from speech signals. MVE minimizes the speaker verification error. And VAD detects the presence or absence of human speech.

Discussion and Conclusion

With technological advances, biometrics is used as a tool to identify people and collect information about their emotional and mental states and physical and behavioral characteristics (Sae-Bae, Wu, Memon, Konrad, & Ishwar, 2019). Biometrics authentication is a leading solution for security concerns in diverse application domains, including online banking (Shah & Kanhere, 2019), smart homes (Edu, Such, & Suarez-Tangil, 2021; Ponticello, Fassl, & Krombholz, 2021), smartphones (Abuhamad, Abusnaina, Nyang, & Mohaisen, 2021; Alzubaidi & Kalita, 2016; Mahfouz, Mahmoud, & Eldin, 2017; Mahfouz et al., 2017), wearable devices (Blasco, Chen, Tapiador, & Peris-Lopez, 2016), the internet of things (Obaidat, Traore, & Woungang, 2019; Duraibi, 2020; Ongun et al., 2018), and health care (Mohsin et al., 2018).

From a security point of view, all authorization systems may have some security vulnerabilities. Ideally, these vulnerabilities should be eliminated or minimized. In recent years, biometric characteristics of people have been integrated into authentication systems and become a part of authorization systems. High-security systems focus on adopting multi-factor authentication systems to strengthen their security systems against attackers (Sinigaglia, Carbone, Costa, & Zannone, 2020). Researchers are studying integrating biometrics into traditional knowledge-based authentication systems to achieve secure multi-factor authentication mechanisms, focusing on several authentication factors, including universality, uniqueness, permanence, collectability, performance, acceptability, and spoofing (Y. Meng et al., 2018; Ometov et al., 2018).

Vocal biometrics can be stolen or imitated in various ways. Voice authentication enables remote authentication, unlike other biometrics. Despite its advantages, voice authentication systems have security vulnerabilities against many attacks. Voice verification may not be as accurate and reliable as other biometric methods (e.g., fingerprint, face recognition, iris recognition). Thus, voice authentication systems have been persistently developed against security leaks (Edu et al., 2021; Y. Meng et al., 2018; Rui & Yan, 2019). During sound verification, liveness detection is required to confirm that the sound comes from a living speaker and, not a recording. The background noise in the environment affects the verification of the sound and, consequently, the matching performance. As

explained below, attackers steal or imitate a speaker's biometrics to trick the authentication systems in several ways.

Random Attacks

Although the human voice is unique, current voice assistants and authentication technologies are inadequate to distinguish all voices. Without having information about a speaker's voice, attackers have a chance to authenticate successfully using their voices (Shi, Wang, Chen, Saxena, & Wang, 2020). Further developments in voice assistant and authentication technologies are essential to prevent such kinds of random attacks.

Mimicry Attacks

Attackers imitate a speaker's articulatory movements and articulatory gestures (Vestman, Kinnunen, Hautamäki, & Sahidullah, 2020). Facial mimics, lips, tongue, or jaw movements, accent, pitch, tone, speed of speech, and word preferences are examples of articulatory gestures. Such attacks can be prevented by multimodal authentication methods, such as verifying both the voice and face of the speaker.

Replay Attacks

Attackers record vocal commands/passwords of a speaker, then replay these recordings for spoofing authentication systems (Ren, Fang, Liu, & Chen, 2019). Such attacks can be prevented by living detection methods (L. Zhang & Yang, 2021; Shang, Chen, & Wu, 2018) and multimodal authentication methods.

Voice Synthesizing Attacks

Attackers collect speech samples of a person and then generate authentication commands/passwords by speech synthesizing techniques (Ning, He, Wu, Xing, & Zhang, 2019). Then attackers play these generated commands/passwords for spoofing authentication systems.

Counterfeit Attacks

Attackers imitate the speaker's voice authentication methods, using records of vocal commands/passwords, simulations of vocal cord vibrations, or other stolen biometric features, such as fingerprints (Li et al., 2020). Such attacks can be prevented by live detection, multimodal authentication, and other security methods.

Hidden Voice Command Attacks

Attackers convert voice commands into sound signals and send them to loudspeakers. Even though humans cannot hear and realize these commands, voice verification models receive and process these commands. These commands can be hidden and embedded in the audio channel of a video to operate the device maliciously (C. Wang et al., 2019; G. Zhang et al., 2017; Blue, Abdullah, Vargas, & Traynor, 2018; Carlini et al., 2016).

Voice authentication is a promising way of developing fast, user-friendly, and reliable multi-factor authentication mechanisms. In this paper, machine learning approaches adapted to voice authentication methods are clarified and discussed with the enlightenment of recent works and trends in the literature.

This paper discusses current single-model and multimodal voice authentication systems and various security attacks, and several feature extraction and classification methods are concisely explained. Voice Authentication is becoming a widely used form of authentication instead of using strings of textual passwords and numbers that are hard to remember. Voice authentication systems are convenient to adapt to multimodal biometric authentication

systems. Also, the remote authentication opportunity of voice authentication systems is convenient for the elderly and disabled.

Together with several benefits, voice authentication approaches have some limitations. One limitation affecting security is the inadequacy of user awareness. Users need to be well-informed about the security risks and concerned about their safety and privacy. Another limitation is the computational cost and energy consumption of the authentication methods on the target device. Devices in the application domain of voice authentication systems, such as smart home appliances, may not have enough computational and energy resources. After technological developments and improvements, more robust solutions, such as deep learning-based voice authentication methods, can be adapted to these devices.

Security is the main challenge of voice authentication systems. Voice authentication systems suffer from several attacks, including random attacks, mimicry attacks, replay attacks, voice synthesizing attacks, counterfeit attacks, and hidden voice command attacks. Enhancing accuracy is another challenge; collected voice data is always affected by some noise level that affects the system's accuracy. Authentication accuracy is also affected by diseases, accidents, and aging. One way to improve security and accuracy is by expanding authentication by multimodal authentication systems. Future advances in integrating multimodal biometric authentication systems into current authentication systems will strengthen existing security and accuracy solutions.

Research Ethics

The authors declare that the research does not have an unethical problem, and they observe the topic of research and publication ethics.

Contribution Rate of Researchers

Author1 and Author2 participated in literature research. The authors read and approved the final version of the paper.

Conflict of Interest

The authors declare that the study has no conflicts of interest.

Genişletilmiş Özet

Giriş

Biyometri, güvenlik sistemleri (Beranek, 2013; Boubchir & Daachi, 2021), adli (Tistarelli & Champod, 2017), sağlık hizmetleri (Bhalla, 2020), online bankacılık (Goode, 2018) ve müşteri hizmetleri gibi çeşitli alanlarda birden çok uygulama alanına sahip bireyleri tanımlamak ve ayırt etmek için kullanılan, kişilerin benzersiz ve ölçülebilir özellikleridir. DNA, yüz, iris, kulak şekli, elektroensefalografi (EEG), avuç içi damarı ve el geometrisi gibi fizyolojik biyometri genetik olarak kazanılır. Davranışsal biyometri, ses tuş vuruşu dinamikleri, imza ve yürüyüş gibi insan aktivite modellerinden oluşur. Bu biyometriklerden bazıları yalnızca ceza davalarında DNA eşleştirmesi gibi özel durumlarda kullanılabilir ve gerçek zamanlı kişisel kimlik doğrulama için pratik değildir.

İnsan sesi davranışsal biyometri kategorisinde yer almaktadır. Davranış kalıpları ve ses bileşenleri (akciğerler ve artikülatörler), perde, ses, tını, vurgu, ritim veya ton açısından ayırt edilebilen tüm bireyler için benzersiz ve ayırt edici bir ses oluşturur (Yoshida, 2012). Ses titreşimleri, nefes alma sesi ve artikülasyon hareketleri gibi ses bileşenlerinin kullanımı, insanlar için diğer biyometrik tanımlayıcılara benzer bir kişisel kimlik oluşturur.

Sesle kimlik doğrulama, insanların seslerini kullanarak güvenli hizmetlere erişmelerini sağlar. Sesli kimlik doğrulama sistemlerinde, kullanıcı konuşması sistem mikrofonları tarafından ses sinyalleri olarak alınır ve kullanıcının kimlik doğrulamasını doğrulamak için işlenir. Örneğin, çevrimiçi bankacılık sistemlerine entegre sesli kimlik doğrulama ile, bir kullanıcının önceden kaydedilmiş sesi, daha fazla erişim için sunucuda saklanır (Fairhurst, Li, & Da Costa-Abreu, 2017). Müşteri ileride yeni bir işlem yapmak istediğinde sistemde kayıtlı sesle eşleşip eşleşmediği kontrol edilir ve kimlik doğrulaması yapılır. Tüm yetkili giriş sistemlerinde olduğu gibi sesli kimlik doğrulama için de güvenlik sistemlerine ihtiyaç duyulmaktadır.

Ses sinyalleri, sinyal alımından ses doğrulamasına kadar ön işleme, özellik çıkarma, sınıflandırma ve karar aşamalarından geçer. Tek modelli bir biyometrik kimlik doğrulama sistemi, yalnızca bir tür biyometrik veriyi işlerken, çok modelli bir biyometrik sistem, kullanıcı kimlik doğrulaması için farklı biyometrik verileri birlikte işler ve güvenliği karşılaştırmalı olarak artırır.

Bu çalışmada literatürde yer alan ve son zamanlarda yapılan tek modelli ve çok modelli ses doğrulama çalışmaları gözden geçirilmiştir. Ses doğrulama sistemleri üzerine yapılan son makine öğrenimi çalışmaları, öznitelik çıkarma ve sınıflandırma yöntemlerine dayalı olarak ele alınmaktadır. Ayrıca bu çalışmada sesli kimlik doğrulama sistemlerine yönelik güvenlik saldırıları tartışılmıştır.

Yöntem

Geleneksel sesli kimlik doğrulama sistemleri dört aşamadan oluşur; ön işleme, özellik çıkarma, sınıflandırma ve karar verme. Ön işleme aşamasında, ses sinyalleri öznitelik çıkarma aşamasından önce işlenir. Bu aşamada, sistem doğruluğunu iyileştirmek için ses sinyalleri gereksiz gürültülerden ayrılır. İnsan, sesleri 20 ile 20 kHz arasında değişen frekans bantlarında algılar ve insan sesinin frekansı 3400 Hz'den düşüktür. Ön işleme aşamasında, yüksek frekans bantlarındaki (örneğin 20 kHz ve üzeri) sinyaller, alınan ses sinyallerinden çıkarılır. Ayrıca, elde edilen ses verilerinden ortam gürültüsü ve sessizlik boşlukları azaltılır. Ortam gürültüsü insan sesinin frekans aralığında olduğundan, ses sinyallerinden ortam gürültüsünü azaltmak her zaman pratik olarak mümkün değildir; bu nedenle ses sinyalleri, ön işleme aşamasından sonra da ortam gürültüsü içerebilir. Özellik çıkarma aşamasında, ses

sinyallerinden çeşitli zamansal (örn. enerji seviyeleri, sıfır geçiş oranı, genlik seviyeleri) ve spektral özellikler (örn. frekans bileşenleri, spektral enerjiler, spektral yoğunluklar) çıkarılır.

Sınıflandırma aşamasında, önceden kaydedilmiş ses sinyallerinin çıkarılan özellikleri sınıflandırılır ve ses şablonları olarak sistemde saklanır. Sınıflandırma aşamasından sonraki aşama karar aşamasıdır. Karar aşaması, kullanıcı sesi ile daha önce saklanan ses şablonları arasında bir eşleşme bulmayı amaçlar. Eşleşme oranına göre kimlik doğrulama kabul edilir veya reddedilir. Sistem, doğruluk, gerçek pozitif oranı, yanlış reddetme oranı ve eşit hata oranı gibi performans ölçütlerine göre değerlendirilir.

Bu çalışma, literatürdeki ses doğrulama sistemlerini, tek modelli ve çok modelli sistemler altında, özellik çıkarma ve sınıflandırma yöntemleri, ikincil biyometri, katılımcı sayısı ve sistem performansı dahil olmak üzere çeşitli ayrıntılarla incelemektedir. Bu sistemlere uygulanan öznitelik çıkarma ve sınıflandırma yöntemleri bu çalışmada listelenmiş ve ayrıntılı olarak açıklanmıştır.

Bulgular

Literatürdeki 18 ses doğrulama sistemi, bu sistemlerde kullanılan özellik çıkarma ve sınıflandırma yöntemleri ile birlikte inclenmiştir. Bu çalışmalardan üç çalışma tek model olarak geliştirilmişken diğerleri ses titreşimleri, konuşma titreşimleri, vücut titreşimleri, nefes alma sesi, yüz şekli, dudak hareketleri, ağız hareketleri, eklem hareketleri gibi ek biyometrik özelliklerle kimlik tanıma modelini geliştirerek çok modelli sistemler olarak oluşturulmuşlardır. Sesle kimlik doğrulama sistemlerinde, sistem doğruluğunu iyileştirmek için uygun öznitelik çıkarma ve sınıflandırma yöntemlerinin seçilmesi esastır. Bu çalışmalarda çeşitli sayıda katılımcılardan alınan veriler üzerinde yapılan incelemelerde geliştirilen modellerin yüksek başarım gösterdikleri açıklanmıştır.

Tartışma ve Sonuç

Bu makale, mevcut tek modelli ve çok modelli sesli kimlik doğrulama sistemlerini ve çeşitli güvenlik saldırılarını tartışmaktadır ve çeşitli özellik çıkarma ve sınıflandırma yöntemleri açıklanmaktadır.

Sesli Kimlik Doğrulama, hatırlanması zor metinsel parola ve sayı dizileri kullanmak yerine, yaygın olarak kullanılan bir kimlik doğrulama biçimi haline gelmektedir. Sesli kimlik doğrulama sistemleri, çok modelli biyometrik kimlik doğrulama sistemlerine uyum sağlamak için uygundur. Ayrıca sesli kimlik doğrulama sistemlerinin uzaktan kimlik doğrulayabilme imkanı yaşlılar ve engelliler için uygundur.

Çeşitli faydalarının yanı sıra, sesle kimlik doğrulama yaklaşımlarının bazı sınırlamaları vardır. Kullanıcıların güvenlik riskleri hakkında iyi bilgilendirilmeleri gerekir. Diğer bir sınırlama, hedef cihazdaki kimlik doğrulama yöntemlerinin hesaplama maliyeti ve enerji tüketimidir. Bu sistemlerin uygulanacağı cihazlar yeterli hesaplama ve enerji kaynaklarına sahip olmayabilir.

Tüm yetkilendirme sistemlerinde güvenlik açıkları olabilir. Son yıllarda kişilerin biyometrik özellikleri kimlik doğrulama sistemlerine entegre edilerek yetkilendirme sistemlerinin bir parçası haline gelmiştir. Yüksek güvenlikli sistemler, güvenlik sistemlerini saldırganlara karşı güçlendirmek için çok faktörlü kimlik doğrulama sistemlerini benimsemeye odaklanır (Sinigaglia, Carbone, Costa, & Zannone, 2020). Araştırmacılar, evrensellik, benzersizlik, kalıcılık, toplanabilirlik, performans, kabul edilebilirlik ve sahtekarlık dahil olmak üzere çeşitli kimlik doğrulama faktörlerine odaklanarak güvenli çok faktörlü kimlik doğrulama mekanizmaları elde etmek için

biyometriyi geleneksel bilgi tabanlı kimlik doğrulama sistemlerine entegre etmeyi araştırıyorlar (Y. Meng ve diğerleri, 2018; Ometov ve diğerleri, 2018).

Vokal biyometri, çeşitli şekillerde çalınabilir veya taklit edilebilir. Ses doğrulaması sırasında, sesin bir kayıttan değil, canlı bir hoparlörden geldiğini doğrulamak için canlılık tespiti gerekir. Ortamdaki arka plan gürültüsü, sesin doğrulanmasını ve dolayısıyla eşleştirme performansını etkiler. Aşağıda açıklandığı gibi, saldırganlar kimlik doğrulama sistemlerini birkaç şekilde kandırmak için bir konuşmacının biyometriğini çalar veya taklit eder. Teknolojik gelişmeler ve iyileştirmeler sonrasında derin öğrenme tabanlı ses doğrulama yöntemlerinin de gelişmesiyle daha sağlam çözümler üretilecektir.

Sesle kimlik doğrulama sistemleri, rastgele saldırılar, taklit saldırıları, tekrar saldırıları, ses sentezleme saldırıları, sahte saldırılar ve gizli sesli komut saldırıları dahil olmak üzere çeşitli saldırılardan muzdariptir. Doğruluğu artırmak başka bir zorluktur; toplanan ses verileri her zaman sistemin doğruluğunu etkileyen bazı gürültü seviyelerinden etkilenir. Kimlik doğrulama doğruluğu ayrıca hastalıklardan, kazalardan ve yaşlanmadan da etkilenir. Güvenliği ve doğruluğu artırmanın bir yolu, çok modelli kimlik doğrulama sistemleriyle kimlik doğrulamayı genişletmektir. Çok modelli biyometrik kimlik doğrulama sistemlerini mevcut kimlik doğrulama sistemlerine entegre etme konusundaki gelecekteki gelişmeler, mevcut güvenlik ve doğruluk çözümlerini güçlendirecektir.

References/Kaynakça

Abdulrahman, S. A., Khalifa, W., Roushdy, M., & Salem, A. B. M. (2020). Comparative study for 8 computational intelligence algorithms for human identification. *Computer Science Review*, *36*, 100237. https://doi.org/10.1016/j.cosrev.2020.100237

Abhishek Anand, S., Liu, J., Wang, C., Shirvanian, M., Saxena, N., & Chen, Y. (2021). EchoVib: Exploring voice authentication via unique non-linear vibrations of short replayed speech. *ASIA CCS 2021 - Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, 67–81. https://doi.org/10.1145/3433210.3437518

Abozaid, A., Haggag, A., Kasban, H., & Eltokhy, M. (2019). Multimodal biometric scheme for human authentication technique based on voice and face recognition fusion. *Multimedia Tools and Applications*, *78*(12), 16345–16361.

Abuhamad, M., Abusnaina, A., Nyang, D., & Mohaisen, D. (2021). Sensor-based continuous authentication of smartphones' users using behavioral biometrics: A contemporary survey. *IEEE Internet of Things Journal*, *8*(1), 65–84. https://doi.org/10.1109/jiot.2020.3020076

Alzubaidi, A., & Kalita, J. (2016). Authentication of smartphone users using behavioral biometrics. *IEEE Communications Surveys and Tutorials*, *18*(3), 1998–2026. https://doi.org/10.1109/COMST.2016.2537748

Beranek, B. (2013). Voice biometrics: success stories, success factors and what's next. *Biometric Technology Today*, *2013*(7), 9–11. https://doi.org/10.1016/s0969-4765(13)70128-0

Bhalla, A. (2020). The latest evolution of biometrics. *Biometric Technology Today*, *2020*(8), 5–8.

Blasco, J., Chen, T. M., Tapiador, J., & Peris-Lopez, P. (2016). A Survey of wearable biometric recognition systems. *ACM Computing Surveys*, *49*(3), 1–35. https://doi.org/10.1145/2968215

Blue, L., Abdullah, H., Vargas, L., & Traynor, P. (2018). 2ma: Verifying voice commands via two microphone authentication. *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, 89–100.

Boubchir, L., & Daachi, B. (2021). Recent advances in biometrics and its applications. *Electronics*, Vol. 10, pp. 1–2. Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/electronics10091097

Campbell, W. M., Sturim, D. E., & Reynolds, D. A. (2006). Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, *13*(5), 308–311.

Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., … Zhou, W. (2016). Hidden voice commands. *25th $\{$USENIX$\}$ Security Symposium ($\{$USENIX$\}$ Security 16)*, 513–530.

Dargan, S., & Kumar, M. (2020). A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities. *Expert Systems with Applications*, *143*. https://doi.org/10.1016/j.eswa.2019.113114

de Barcelos Silva, A., Gomes, M. M., da Costa, C. A., da Rosa Righi, R., Barbosa, J. L. V., Pessin, G., … Federizzi, G. (2020). Intelligent personal assistants: A systematic literature review. *Expert Systems with Applications*, *147*. https://doi.org/10.1016/j.eswa.2020.113193

Dişken, G., Tüfekçi, Z., Saribulut, L., & Çevik, U. (2017). A review on feature extraction for speaker recognition under degraded conditions. *IETE Technical Review*, *34*(3), 321–332.

Duraibi, S. (2020). Voice biometric identity authentication model for IoT devices. *International Journal of Security, Privacy and Trust Management (IJSPTM) Vol*, *9*.

Edu, J. S., Such, J. M., & Suarez-Tangil, G. (2021). Smart home personal assistants: A security and privacy review. *ACM Computing Surveys*, *53*(6), 1–35. https://doi.org/10.1145/3412383

Fairhurst, M., Li, C., & Da Costa-Abreu, M. (2017). Predictive biometrics: A review and analysis of predicting personal characteristics from biometric data. *IET Biometrics*, *6*(6), 369–378. https://doi.org/10.1049/iet-bmt.2016.0169

Feng, H., Fawaz, K., & Shin, K. G. (2017). Continuous authentication for voice assistants. *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, 343–355.

Gauvain, J.-L., & Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, *2*(2), 291–298.

Gayathri, M., Malathy, C., & Prabhakaran, M. (2019). A Review on various biometric techniques, its features, methods, security issues and application areas. *International Conference On Computational Vision and Bio Inspired Computing*, 931–941.

Gofman, M. I., Mitra, S., Cheng, T.-H. K., & Smith, N. T. (2016). Multimodal biometrics for enhanced mobile device security. *Communications of the ACM*, *59*(4), 58–65.

Gofman, M., Mitra, S., Cheng, K., & Smith, N. (2015). Quality-based score-level fusion for secure and robust multimodal biometrics-based authentication on consumer mobile devices. *Proc. Int. Conf. Softw. Eng. Adv.(ICSEA)*, 274–276.

Gofman, M., Sandico, N., Mitra, S., Suo, E., Muhi, S., & Vu, T. (2018). Multimodal biometrics via discriminant correlation analysis on mobile devices. *Proceedings of the International Conference on Security and Management (SAM)*, 174–181.

Goode, A. (2018). Biometrics for banking: best practices and barriers to adoption. *Biometric Technology Today*, *2018*(10), 5–7. https://doi.org/10.1016/s0969-4765(18)30156-5

Hajarolasvadi, N., & Demirel, H. (2019). 3D CNN-based speech emotion recognition using k-means clustering and spectrograms. *Entropy*, *21*(5), 479. https://doi.org/10.3390/e21050479

Jain, A. K., & Mao, J. (1996). Artificial neural networks: A Tutorial. *Computer*, *29*(3), 31–44.

Jayanna, H. S., Mahadeva Prasanna, S. R., & Prasanna, S. R. M. (2009). Analysis, feature extraction, modeling and testing techniques for speaker recognition. *IETE Technical Review*, *26*(3), 181–190. https://doi.org/10.4103/0256-4602.50702

Juang, B. H., & Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics*, *33*(3), 251–272. https://doi.org/10.1080/00401706.1991.10484833

Juniper. (2022). Digital voice assistants in use to triple to 8 billion by 2023, driven by smart home devices.

Juniper Research. Retrieved from https://www.juniperresearch.com/press/digital-voice-assistants-in-use-to-8-million-2023

Kaman, S., Swetha, K., Akram, S., & Varaprasad, G. (2013). Remote user authentication using a voice authentication system. *Information Security Journal*, *22*(3), 117–125. https://doi.org/10.1080/19393555.2013.801539

Kwon, O.-W., Chan, K., Hao, J., & Lee, T.-W. (2003). Emotion recognition by speech signals. *Eighth European Conference on Speech Communication and Technology*.

Lawson, A., Vabishchevich, P., Huggins, M., Ardis, P., Battles, B., & Stauffer, A. (2011). Survey and evaluation of acoustic features for speaker recognition. *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5444–5447.

Li, H., Xu, C., Rathore, A. S., Li, Z., Zhang, H., Song, C., … Xu, W. (2020). VocalPrint: Exploring a resilient and secure voice authentication via mmWave biometric interrogation. *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 312–325.

Lu, L., Yu, J., Chen, Y., Liu, H., Zhu, Y., Kong, L., & Li, M. (2019). Lip reading-based user authentication through acoustic sensing on smartphones. *IEEE/ACM Transactions on Networking*, *27*(1), 447–460.

Mahfouz, A., Mahmoud, T. M., & Eldin, A. S. (2017). A survey on behavioral biometric authentication on smartphones. *Journal of Information Security and Applications*, *37*, 28–37. https://doi.org/10.1016/j.jisa.2017.10.002

Meng, Y., Wang, Z., Zhang, W., Wu, P., Zhu, H., Liang, X., & Liu, Y. (2018). Wivo: Enhancing the security of voice control system via wireless signal in IoT environment. *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 81–90.

Meng, Z., Altaf, M. U. Bin, & Juang, B.-H. F. (2020). Active voice authentication. *Digital Signal Processing*, *101*, 102672.

Mohsin, A. H., Zaidan, A. A., Zaidan, B. B., Ariffin, S. A. B., Albahri, O. S., Albahri, A. S., … Hashim, M. (2018). Real-time medical systems based on human biometric steganography: A systematic review. *J Med Syst*, *42*(12), 245. https://doi.org/10.1007/s10916-018-1103-6

Moreno-Rodriguez, J. C., Ramirez-Cortes, J. M., Atenco-Vazquez, J. C., & Arechiga-Martinez, R. (2021). EEG and voice bimodal biometric authentication scheme with fusion at signal level. *2021 IEEE Mexican Humanitarian Technology Conference (MHTC)*, 52–58.

Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *ArXiv Preprint ArXiv:1003.4083*.

Ning, Y., He, S., Wu, Z., Xing, C., & Zhang, L. J. (2019). A review of deep learning based speech synthesis. *Applied Sciences (Switzerland)*, *9*(19), 1–16. https://doi.org/10.3390/app9194050

Obaidat, M. S., Traore, I., & Woungang, I. (2019). *Biometric-based physical and cybersecurity systems*. Springer.

Olazabal, O., Gofman, M., Bai, Y., Choi, Y., Sandico, N., Mitra, S., & Pham, K. (2019). Multimodal biometrics for enhanced iot security. *2019 IEEE 9th Annual Computing and Communication Workshop and Conference*

(CCWC), 886–893.

Ometov, A., Bezzateev, S., Mäkitalo, N., Andreev, S., Mikkonen, T., & Koucheryavy, Y. (2018). Multi-factor authentication: A Survey. *Cryptography*, *2*(1). https://doi.org/10.3390/cryptography2010001

Ongun, T., Oprea, A., Nita-Rotaru, C., Christodorescu, M., Salajegheh, N., Spohngellert, O., … Salajegheh, N. (2018). The house that knows you: User authentication based on IoT data. *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2255–2257. Retrieved from http://arxiv.org/abs/1908.00592

Ponticello, A., Fassl, M., & Krombholz, K. (2021). Exploring authentication for security-sensitive tasks on smart Home voice assistants. *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*.

Ren, Y., Fang, Z., Liu, D., & Chen, C. (2019). Replay attack detection based on distortion by loudspeaker for voice authentication. *Multimedia Tools and Applications*, *78*(7), 8383–8396.

Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, *3*(1), 72–83.

Rui, Z., & Yan, Z. (2019). A Survey on biometric authentication: Toward secure and privacy-preserving identification. *IEEE Access*, *7*, 5994–6009. https://doi.org/10.1109/access.2018.2889996

Sae-Bae, N., Wu, J., Memon, N., Konrad, J., & Ishwar, P. (2019). Emerging NUI-based methods for user authentication: A new taxonomy and survey. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *1*(1), 5–31. https://doi.org/10.1109/tbiom.2019.2893297

Shah, S. W., & Kanhere, S. S. (2019). Recent trends in user authentication - a survey. *Ieee Access*, *7*, 112505–112519. https://doi.org/10.1109/Access.2019.2932400

Shang, J., Chen, S., & Wu, J. (2018). Defending against voice spoofing: A robust software-based liveness detection system. *2018 IEEE 15th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, 28–36.

Shi, C., Wang, Y., Chen, Y., & Saxena, N. (2021). Authentication of voice commands by leveraging vibrations in wearables. *IEEE Security and Privacy*, (December), 83–92. https://doi.org/10.1109/MSEC.2021.3077205

Shi, C., Wang, Y., Chen, Y., Saxena, N., & Wang, C. (2020). WearID: Low-effort wearable-assisted authentication of voice commands via cross-domain comparison without training. *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, 829–842. ICST. https://doi.org/10.1145/3427228.3427259

Sidorov, M., Schmitt, A., Zablotskiy, S., & Minker, W. (2013). Survey of automated speaker identification methods. *2013 9th International Conference on Intelligent Environments*, 236–239.

Singh, A. P., Nath, R., & Kumar, S. (2018). A survey: Speech recognition approaches and techniques. *2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, 1–4. IEEE. https://doi.org/10.1109/UPCON.2018.8596954

Sinigaglia, F., Carbone, R., Costa, G., & Zannone, N. (2020). A survey on multi-factor authentication for online banking in the wild. *Computers and Security*, *95*. https://doi.org/10.1016/j.cose.2020.101745

Smallman, M. (2020). Good call: the hybrid answer to voice authentication. *Biometric Technology Today*, *2020*(4),

10–12.

Snyder, D., Garcia-Romero, D., Povey, D., & Khudanpur, S. (2017). Deep neural network embeddings for text-independent speaker verification. *Interspeech*, *2017-Augus*, 999–1003. https://doi.org/10.21437/Interspeech.2017-620

Statistica.com. (2021). Voice recognition market size worldwide in 2020 and 2026. *Statista Research Department*. Retrieved from https://www.statista.com/statistics/1133875/global-voice-recognition-market-size/

Suman, M., Harish, K., Kumar, K. M., & Samrajyam, S. (2015). Speech recognition using MFCC and VQLBG. *International Journal of Advances in Applied Sciences*, *4*(4), 151. https://doi.org/10.11591/ijaas.v4.i4.pp151-156

Tirumala, S. S., & Shahamiri, S. R. (2016). A review on deep learning approaches in speaker identification. *Proceedings of the 8th International Conference on Signal Processing Systems*, 142–147.

Tirumala, S. S., Shahamiri, S. R., Garhwal, A. S., & Wang, R. (2017). Speaker identification features extraction methods: A systematic review. *Expert Systems with Applications*, *90*, 250–271. https://doi.org/10.1016/j.eswa.2017.08.015

Tistarelli, M., & Champod, C. (2017). *Handbook of biometrics for forensic science*. Springer.

Todkar, S. P., Babar, S. S., Ambike, R. U., Suryakar, P. B., & Prasad, J. R. (2018). Speaker recognition techniques: A review. *2018 3rd International Conference for Convergence in Technology (I2CT)*, 1–5.

Tzanetakis, G., Essl, G., & Cook, P. (2001). Audio analysis using the discrete wavelet transform. *Proc. Conf. in Acoustics and Music Theory Applications*, *66*.

Vestman, V., Kinnunen, T., Hautamäki, R. G., & Sahidullah, M. (2020). Voice mimicry attacks assisted by automatic speaker verification. *Computer Speech \& Language*, *59*, 36–54.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th International Conference on Machine Learning*, 1096–1103.

Wang, C., Anand, S. A., Liu, J., Walker, P., Chen, Y., & Saxena, N. (2019). Defeating hidden audio channel attacks on voice assistants via audio-induced surface vibrations. *Proceedings of the 35th Annual Computer Security Applications Conference*, 42–56.

Wang, Q., Lin, X., Zhou, M., Chen, Y., Wang, C., Li, Q., & Luo, X. (2019). Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones. *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, 2062–2070.

Wu, L., Yang, J., Zhou, M., Chen, Y., & Wang, Q. (2019). LVID: A multimodal biometrics authentication system on smartphones. *IEEE Transactions on Information Forensics and Security*, *15*, 1572–1585. https://doi.org/10.1109/tifs.2019.2944058

Yan, Z., & Zhao, S. (2016). A usable authentication system based on personal voice challenge. *2016 International Conference on Advanced Cloud and Big Data (CBD)*, 194–199.

Yoshida, M. (2012). The articulatory system and the consonants of North American English.

Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., & Xu, W. (2017). Dolphinattack: Inaudible voice commands. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 103–117.

Zhang, L., Tan, S., & Yang, J. (2017). Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 57–71. Dallas, TX, USA: Association for Computing Machinery. https://doi.org/10.1145/3133956.3133962

Zhang, L., Tan, S., Yang, J., & Chen, Y. (2016). Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 1080–1091.

Zhang, L., & Yang, J. (2021). A continuous liveness detection for voice authentication on smart devices. *ArXiv Preprint ArXiv:2106.00859*.

Zhang, X., Cheng, D., Jia, P., Dai, Y., & Xu, X. (2020). An efficient android-based multimodal biometric authentication system with face and voice. *IEEE Access*, *8*, 102757–102772.