Informatyka techniczna i sztuczna inteligencja
Technical informatics and Artificial Intelligence
Engineer of XXI Century'2024

Evaluation of deep learning models and digital signal processing techniques in voice biometrics

Volodymyr Khoma [1], Ivan Opirskyy [2], Dmytro Sabodashko [3,*]

[1] *Opole University of Technology, Professor of Automatic Control and Robotics Department; v.khoma@po.edu.pl*
[2] *Lviv Polytechnic National University, Head of Information Protection Department, iopirsky@gmail.com*
[3] *Lviv Polytechnic National University, Senior Lecturer of Information Protection Department, dmytro.v.sabodashko@lpnu.ua*
* *Corresponding author: dmytro.v.sabodashko@lpnu.ua*

Abstract: This article explores the limitations of traditional authentication methods, such as passwords, which are often unreliable due to vulnerabilities like loss, theft, and weak resistance to attacks. Biometric authentication, particularly voice-based systems, is a promising alternative due to its high security and user convenience. Voice is a unique personal trait, making it difficult to forge or steal. However, voice biometric systems face challenges such as variations in voice due to health, emotions, or environmental factors. This study compares modern deep learning models with traditional digital signal processing (DSP) methods for voice-based authentication. Text-dependent DSP methods (MFCC and LPC) and text-independent deep learning models (ECAPA-TDNN and ResNet) are evaluated to measure their effectiveness in speaker recognition. The experiment involved developing biometric systems using these methods and testing them on a specialized dataset. The findings highlight the strengths and weaknesses of both approaches in the context of voice authentication.

Keywords: biometric technologies; voice authentication; digital signal processing; Mel-frequency cepstral coefficients (MFCC); linear predictive coding (LPC); deep learning; neural networks.

Ewaluacja modeli głębokiego uczenia i technik cyfrowego przetwarzania sygnałów w biometrii głosu

Volodymyr Khoma [1], Ivan Opirskyy [2], Dmytro Sabodashko [3,*]

[1] *Politechnika Opolska, Profesor Katedry Automatyki i Robotyki; v.khoma@po.edu.pl*
[2] *Lwowski Narodowy Uniwersytet Politechniczny, Kierownik Katedry Ochrony Informacji, iopirsky@gmail.com*
[3] *Lwowski Narodowy Uniwersytet Politechniczny, Starszy Wykładowca Katedry Ochrony Informacji dmytro.v.sabodashko@lpnu.ua*
* *Corresponding author: dmytro.v.sabodashko@lpnu.ua*

Streszczenie: W artykule omówiono ograniczenia tradycyjnych metod uwierzytelniania, takich jak hasła, które często okazują się zawodne z powodu podatności na utratę, kradzież oraz słabą odporność na ataki. Biometryczne systemy uwierzytelniania, szczególnie te oparte na głosie, stanowią obiecującą alternatywę ze względu na wysoki poziom bezpieczeństwa i wygodę użytkowania. Głos jest unikalną cechą osobistą, co sprawia, że jest trudny do podrobienia lub kradzieży. Jednak systemy biometryczne oparte na głosie napotykają wyzwania, takie jak zmiany w głosie spowodowane stanem zdrowia, emocjami lub czynnikami środowiskowymi. W niniejszym badaniu porównano nowoczesne modele głębokiego uczenia z tradycyjnymi metodami cyfrowego przetwarzania sygnałów (DSP) stosowanymi do uwierzytelniania głosu. Metody DSP zależne od tekstu (MFCC i LPC) oraz modele głębokiego uczenia niezależne od tekstu (ECAPA-TDNN i ResNet) zostały ocenione pod kątem ich skuteczności w rozpoznawaniu mówcy. Eksperyment obejmował opracowanie systemów biometrycznych z wykorzystaniem tych metod oraz ich testowanie na specjalistycznym zbiorze danych. Wyniki podkreślają mocne i słabe strony obu podejść w kontekście uwierzytelniania głosu.

DOI: https://doi.org/10.53052/9788367652292.26

Słowa kluczowe: technologie biometryczne; uwierzytelnianie głosowe; cyfrowe przetwarzanie sygnałów; współczynniki mel-cepstralne; liniowe kodowanie predykcyjne; głębokie uczenie; sieci neuronowe.

1. Wstęp

Obecnie Internet jest potężną platformą, która znacząco wpłynęła na komunikację i procesy biznesowe we współczesnym świecie. Liczba jego użytkowników przekracza 2,4 miliarda, co przyczynia się do znacznego wzrostu popularności handlu online, wymiany wiedzy i sieci społecznościowych. Jednak wraz z tym wzrostem rośnie również zapotrzebowanie na niezawodne środki cyberbezpieczeństwa i ochrony prywatności.
Statystyki z 2023 roku pokazują, że Facebook pozostaje najpopularniejszą siecią społecznościową narażoną na ataki hakerskie, z ponad 68 000 kontami przejmowanymi co miesiąc[1]. Często jest to wynikiem niedostatecznej uwagi użytkowników na kwestie cyberbezpieczeństwa. To podkreśla znaczenie trzech kluczowych aspektów bezpieczeństwa: identyfikacji, uwierzytelniania i autoryzacji. Identyfikacja to proces określania tożsamości podmiotu, którym może być człowiek, maszyna lub inny zasób, taki jak oprogramowanie. W kontekście bezpieczeństwa uwierzytelnianie i autoryzacja określają, kto ma dostęp do zasobów informacyjnych w sieci.
„Weryfikacja", „uwierzytelnianie" i „autoryzacja" to kluczowe pojęcia stanowiące podstawę technologii bezpieczeństwa systemów informatycznych. Weryfikacja polega na przekazaniu identyfikatora do systemu. Następnie, przed uwierzytelnianiem, podmiot dostarcza systemowi swój identyfikator (np. login lub adres e-mail), a monitor potwierdza tożsamość poprzez proces uwierzytelniania (np. za pomocą hasła). Uwierzytelnianie to proces, w którym podmiot potwierdza swoją tożsamość, aby monitor mógł upewnić się, że rzeczywiście jest on osobą, za którą się podaje. Na koniec autoryzacja określa uprawnienia przyznane użytkownikowi.
Systemy uwierzytelniania odpowiadają na kluczowe pytania: „czy użytkownik jest naprawdę tym, za kogo się podaje?". Tym samym uwierzytelnianie jest jednym z najbardziej obiecujących środków zwiększania zaufania i bezpieczeństwa w zastosowaniach komercyjnych. Gwarantuje także zapewnienie tożsamości wspomnianych podmiotów. Jedną z najlepszych metod zapewnienia bezpiecznego uwierzytelniania użytkownika jest wykorzystanie uwierzytelniania biometrycznego. W tym kontekście pojawia się potrzeba badania optymalnej metody realizacji systemu biometrycznego uwierzytelniania.

Postawienie problemu.
Klasyczne metody uwierzytelniania, takie jak używanie haseł, okazują się niewystarczająco niezawodne ze względu na liczne potencjalne podatności, takie jak utrata lub kradzież haseł, ich słaba odporność oraz trudności w zarządzaniu. Metody biometrycznego uwierzytelniania oparte na cechach fizycznych, takich jak głos, są obiecującym rozwiązaniem, ponieważ mogą zapewnić wysoki poziom bezpieczeństwa i wygody dla użytkowników. Jednak istnieją wyzwania dotyczące precyzji i niezawodności tych systemów, które wymagają dalszych badań i udoskonaleń. W tym kontekście szczególnie istotne i efektywne są badania nad nowoczesnymi modelami głębokiego uczenia oraz ich poprzednikami opartymi na cyfrowym przetwarzaniu sygnałów, które są stosowane do biometrycznego uwierzytelniania za pomocą głosu.

Analiza ostatnich badań i publikacji.
Globalny rynek technologii biometrycznych został wyceniony na 34,27 miliarda USD w 2022 roku i przewiduje się, że będzie się rozwijał w średniorocznym tempie wzrostu (CAGR) wynoszącym 20,4% w latach 2023–2030. Zapotrzebowanie na technologie biometryczne jest napędzane rosnącą adopcją systemów biometrycznych w elektronice użytkowej oraz przemyśle motoryzacyjnym. Kluczowymi czynnikami wpływającymi na wzrost rynku są rozszerzające się zastosowania technologii biometrycznych w różnych sektorach przemysłu oraz rosnące zapotrzebowanie na rozwiązania w zakresie uwierzytelniania, identyfikacji oraz bezpieczeństwa i nadzoru w wielu obszarach aplikacyjnych [2]. Ludzki głos, jako wynik złożonej interakcji pomiędzy narządami jamy ustnej a strunami głosowymi, cechuje się unikalną strukturą akustyczną. Indywidualne cechy anatomiczne i behawioralne wpływają na kształtowanie unikalnego wzorca sygnału akustycznego. Więc głos jako marker biometryczny wyróżnia się specyficznym składem widmowym, cechami czasowymi oraz innymi parametrami, które umożliwiają niezawodną identyfikację osób.
Systemy biometryczne oparte na głosie są efektem podejścia interdyscyplinarnego, łączącego inżynierię z naukami biologicznymi, co prowadzi do skutecznych i niezawodnych metod identyfikacji [3]. Innowacyjnym rozwiązaniem

w tej dziedzinie jest wprowadzenie systemów, które umożliwiają jednoczesną weryfikację kilku cech biometrycznych. Przykładem może być podniesienie poziomu bezpieczeństwa systemów kontroli dostępu poprzez zastąpienie tradycyjnych kluczy elektromagnetycznymi zamkami, które identyfikują użytkowników na podstawie rozpoznawania głosu lub odcisków palców [4]. W ostatnich latach zwrócono uwagę na problem nierówności rasowych i płciowych w systemach biometrycznych. W istniejących systemach, takich jak biometria twarzy, badania wykazały, że różnice te mogą prowadzić do poważnych uprzedzeń lub innych problemów społecznych, gdy systemy biometryczne są wdrażane na szeroką skalę [5].
Najnowsze osiągnięcia w dziedzinie głębokiego uczenia się wykazały znaczący potencjał w rozwiązywaniu problemów związanych z osiąganiem wysokiej dokładności i odporności w systemach uwierzytelniania głosowego. Czynniki takie jak zmienność głosu oraz szumy środowiskowe stanowią istotne wyzwania, co podkreśla konieczność opracowania niezawodnych systemów zabezpieczających urządzenia osobiste i transakcje finansowe. W ostatnim badaniu przeanalizowano wydajność zaawansowanych modeli głębokiego transfer learningu, w tym Vision Transformers (ViT), VGG16 oraz dostosowanej konwolucyjnej sieci neuronowej (CNN), wykorzystując zbiór danych składający się z 3000 próbek głosowych równomiernie podzielonych na głosy męskie i żeńskie. Wyniki badania wskazały, że model VGG16 wykorzystujący transfer learning osiągnął najwyższą precyzję na poziomie 95%, przewyższając ViT oraz dostosowany model CNN. Wyniki te podkreślają potencjał głębokiego uczenia się, w szczególności technik transfer learningu, w poprawie dokładności i niezawodności systemów rozpoznawania głosu[6].
Dodatkowym wyzwaniem dla systemów uwierzytelniania głosu są technologie klonowania głosu. Potrafią one stworzyć model głosu na podstawie zaledwie kilku minut autentycznego nagrania, co umożliwia odtwarzanie dowolnego tekstu wybranym głosem. Stanowi to poważne zagrożenie dla bezpieczeństwa biometrycznych systemów opartych na identyfikacji głosowej [7]. W obliczu tych wyzwań, przyszłe badania nad systemami uwierzytelniania głosu powinny skupiać się na opracowaniu technologii, które nie tylko precyzyjnie identyfikują mówcę, ale również potrafią wykrywać syntetyczne nagrania dźwiękowe.

Celem artykułu jest przeprowadzenie kompleksowej analizy porównawczej, mającej na celu ocenę skuteczności nowoczesnych modeli głębokiego uczenia w zastosowaniach do biometrycznego uwierzytelniania głosowego. Badania skupią się na porównaniu tych modeli z tradycyjnymi metodami opartymi na cyfrowym przetwarzaniu sygnałów, a także na identyfikacji optymalnych technik przetwarzania wstępnego nagrań audio, które przyczynią się do zwiększenia dokładności i niezawodności systemów uwierzytelniania głosowego.

Główne zadania artykułu to:
przegląd najnowszych osiągnięć w dziedzinie biometrycznego uwierzytelniania, ze szczególnym uwzględnieniem metod opartych na analizie głosu;
dokładna analiza istniejących metod przetwarzania danych w kontekście uwierzytelniania głosowego;
porównanie wydajności systemów uwierzytelniania głosowego zależnych od tekstu i niezależnych od tekstu.

2. Przegląd nowoczesnych metod uwierzytelniania biometrycznego

Koncepcja uwierzytelniania biometrycznego jest oparta na analizie unikalnych cech fizycznych, behawioralnych lub psychologicznych osoby[8]. Obecnie systemy biometryczne znajdują szerokie zastosowanie w różnych dziedzinach, w tym w smartfonach, systemach bankowych i systemach bezpieczeństwa budynków. Przegląd metod uwierzytelniania biometrycznego pozwala lepiej zrozumieć ich zalety i ograniczenia:

Rozpoznawanie twarzy – systemy rozpoznawania twarzy są szeroko stosowane, zwłaszcza w projektach związanych z bezpieczeństwem narodowym i w interakcji człowiek-komputer. Działają na zasadzie wstępnego przetwarzania obrazu, wyodrębniania twarzy z tłumu i ich dalszej analizy w celu identyfikacji. Podstawowe metody rozpoznawania twarzy dzielą się na dwa typy: metody oparte na cechach i metody oparte na wyglądzie zewnętrznym. Metody oparte na cechach wykorzystują informacje o strukturze twarzy. Z kolei metody oparte na wyglądzie zewnętrznym traktują zadanie rozpoznawania twarzy jako problem rozpoznawania wzorców na podstawie uczenia statystycznego. W ostatnich latach zaproponowano różne algorytmy rozpoznawania twarzy o wysokiej dokładności, z których każdy ma swoje zalety i wady, uwzględniane przy ich ocenie [9]. Jednym z popularnych podejść jest rozpoznawanie twarzy na podstawie wyodrębniania i klasyfikacji cech, gdzie najpierw wyodrębniane są cechy twarzy, a następnie klasyfikator

identyfikuje twarz. Inne podejście wykorzystuje głębokie uczenie i sieci neuronowe, co zapewnia wysoką dokładność rozpoznawania, szczególnie przy pracy z dużymi zbiorami danych. Algorytmy rozpoznawania twarzy mogą automatycznie identyfikować twarze na dużą skalę, co czyni je przydatnymi do celów bezpieczeństwa, kontroli dostępu i monitoringu wideo. Jednakże algorytmy te mają również pewne ograniczenia, takie jak podatność na zmiany oświetlenia i kąta. Ponadto, ich użycie wiąże się z kwestiami prywatności i etyki;

Rozpoznawanie odcisków palców – identyfikacja za pomocą odcisków palców, oparta na stabilnych cechach linii papilarnych, może zapewnić niezawodną metodę identyfikacji. Proces uzyskiwania obrazów odcisków palców obejmuje wyodrębnianie ciemnych linii papilarnych i jasnych przestrzeni, co może być utrudnione przez czynniki środowiskowe i zachowanie użytkownika, często wymagając metod poprawy obrazu. Wykorzystanie algorytmów przetwarzania obrazów pomaga zwiększyć przejrzystość i kontrast odcisków palców, co zapewnia dokładniejsze określenie unikalnych cech. Ponadto nowoczesne systemy identyfikacji za pomocą odcisków palców wykorzystują metody uczenia maszynowego, aby poprawić dokładność rozpoznawania i dostosować się do różnych warunków rejestracji.

Rysunek 1. Informacyjny charakter odcisku linii papilarnych

Rysunek 1 przedstawia informacyjne cechy linii papilarnych stosowane w identyfikacji za pomocą odcisku palca. Informacje identyfikacyjne dotyczące odcisków palców dzielą się na trzy poziomy: makroszczegóły (poziom 1), szczegółowe cechy (poziom 2) i cały zestaw atrybutów wymiarowych (poziom 3). Poziom 1 obejmuje globalne wzory odcisków palców, które nie są unikalne, podczas gdy poziom 2 ma wystarczającą zdolność rozróżniania dzięki drobnym szczegółom wzoru. Poziom 3, obejmujący cechy strukturalne, jest stały i naprawdę unikalny.
Algorytmy identyfikacji odcisków palców można podzielić na trzy grupy: metody oparte na połączeniach, metody oparte na szczegółach oraz metody oparte na niejasnych cechach. Metody oparte na połączeniach, takie jak generatywne sieci kontradyktoryjne (ang. GAN – Generative Adversarial Networks), autoenkodery warstwowe oraz sieci głębokiego wnioskowania, ustanawiają połączenia między małymi punktami odcisków palców a ich otaczającymi cechami [9]. Metody oparte na szczegółach wykorzystują techniki przechwytywania i porównywania drobnych cech, na przykład ograniczonej maszyny Boltzmanna (ang. RBM - Restricted Boltzmann Machine), rekurencyjnych sieci neuronowych (ang. RNN - Recurrent Neural Network) oraz sieci radialnych funkcji bazowych (ang. RBFN - Radial Basis Function Network);

Skanowanie siatkówki i rozpoznawanie tęczówki oka – to zautomatyzowana metoda biometrycznej identyfikacji, która wykorzystuje algorytmy matematyczne do analizy obrazów wideo jednej lub obu tęczówek oka. Złożone wzory tęczówki są unikalne, stabilne przez całe życie i mogą być rozpoznawane z pewnej odległości. Zdolność rozróżniania technologii biometrycznych jest określana przez ilość entropii, którą mogą zakodować i wykorzystać do dopasowania [10]. Rozpoznawanie tęczówki oka jest szczególnie skuteczne w tym zakresie, minimalizując możliwość błędnych

dopasowań nawet w dużych populacjach. Główne ograniczenie tej technologii polega na tym, że uzyskanie wysokiej jakości obrazu z odległości większej niż jeden-dwa metry lub bez aktywnej współpracy osoby może być trudne. Jednak technologia rozwija się, a rozpoznawanie tęczówki oka jest już możliwe z odległości do 10 metrów lub za pomocą kamer w czasie rzeczywistym.

Rozpoznawanie mówcy – to proces identyfikacji osoby na podstawie jej unikalnych cech głosowych. Unikalność głosu wynika z indywidualnych cech anatomicznych aparatu głosowego, takich jak rozmiar krtani i inne organy. Oprócz cech fizycznych, każda osoba ma swój własny styl mówienia, wzorce wymowy i indywidualny dobór słownictwa, co umożliwia wykorzystanie głosu jako parametru biometrycznego do uwierzytelniania [11]. Zadania rozpoznawania mówcy dzielą się na weryfikację i identyfikację. Weryfikacja polega na określeniu, czy osoba, która twierdzi, że jest sobą, faktycznie nią jest. Identyfikacja mówcy to proces określenia, kto mówi, spośród znanych mówców. Większość systemów rozpoznawania mówców wykorzystuje mel-częstotliwościowe współczynniki cepstralne (ang. MFCC – Mel-Frequency Cepstral Coefficients) i liniowe kodowanie predykcyjne (ang. LPC – Linear Predictive Coding), które reprezentują cechy aparatu głosowego. Ogólnie każda z tych technologii biometrycznych ma swoje specyficzne zalety i obszary zastosowań. Połączenie kilku metod może zapewnić zwiększoną niezawodność i dokładność identyfikacji, co czyni systemy biometryczne kluczowym elementem współczesnych rozwiązań w zakresie bezpieczeństwa i wygody. Badanie [12] analizuje podatność systemów automatycznej weryfikacji mówcy na ataki typu spoofing i podkreśla znaczenie opracowania środków przeciwdziałania, które pozwalają odróżnić autentyczną mowę od sfałszowanego dźwięku. Połączenie automatycznej weryfikacji mówcy z systemami przeciwdziałania tworzy nowoczesne rozwiązania uwierzytelniania głosowego, zapewniające niezawodne mechanizmy kontroli dostępu. Aby sprostać wyzwaniom związanym z identyfikacją klientów, badanie analizuje różnorodne cechy sygnału audio i proponuje innowacyjne podejście wykorzystujące biometrię głosową. Metoda ta przewyższa alternatywne algorytmy, osiągając wyższą dokładność rozpoznawania nawet przy ograniczonych danych głosowych. W badaniu oceniono również skuteczność systemów biometrii głosowej z dwóch kluczowych perspektyw: agentów call center oraz klientów. Wnioski zebrane od obu grup wskazują na potencjał biometrii głosowej do poprawy efektywności operacyjnej i doświadczeń użytkowników. Jednocześnie podkreślono, że szerokie wdrożenie tej technologii znajduje się na wczesnym etapie. W podsumowaniu stwierdzono, że zastosowanie biometrii głosowej w call center stanowi przełom technologiczny o istotnym znaczeniu dla różnych branż na całym świecie. Dodatkowo badanie [13] przygląda się technologii głosowych deepfake'ów, zauważając jej dwoisty charakter. Z jednej strony oferuje ona nowe możliwości w zakresie dostępności i zaangażowania użytkowników, z drugiej zaś stwarza poważne zagrożenia dla bezpieczeństwa, prywatności i zaufania. Przekonująco imitowane głosy zagrażają autentyczności, umożliwiają szerzenie dezinformacji i wykorzystywanie osób. Przeanalizowane przypadki ukazują zarówno kreatywne zastosowania, jak i złośliwe wykorzystanie tej technologii, co podkreśla potrzebę wprowadzenia skutecznych środków obronnych. Obecne strategie, takie jak uwierzytelnianie głosowe i wykrywanie deepfake'ów, mają swoje ograniczenia wynikające z szybkiego rozwoju technologii oraz trudności w jej adaptacji przez użytkowników. Aby ograniczyć te zagrożenia, badanie postuluje wdrożenie wielowarstwowych systemów uwierzytelniania, zaawansowanych narzędzi detekcji, wytycznych etycznych oraz kompleksowych ram regulacyjnych. Podkreślono również znaczenie podnoszenia świadomości publicznej i promowania kompetencji cyfrowych jako kluczowych kroków w efektywnym radzeniu sobie z ryzykiem i możliwościami, jakie niesie technologia deepfake. Autorzy [14] badają wykorzystanie Funkcji Fizycznych Niezclonecznych (PUF) jako środków zaradczych wobec deepfake'ów w urządzeniach IoT. Poprzez połączenie PUF z Technikami Ochrony Szablonów Biometrycznych (BTP) na jednym układzie scalonym, zwiększają bezpieczeństwo biometryczne poprzez odnawialne odciski palców generowane przez urządzenia FPGA. Podejście to, mimo swojej wydajności, wymaga korekcji błędów i obróbki postprocesowej. Badanie bada również biometrię anulowalną, stosując transformacje w celu zachowania bezpieczeństwa cech biometrycznych. Uczenie maszynowe i głębokie uczenie są uznawane za kluczowe do ekstrakcji cech i ochrony danych.

Metody fuzji to techniki wykorzystywane do łączenia danych pochodzących z wielu źródeł lub modalności w celu osiągnięcia lepszych wyników niż przy wykorzystaniu pojedynczych źródeł. W kontekście biometrii i uczenia maszynowego metody fuzji są często stosowane do integracji informacji z różnych typów cech biometrycznych, takich jak odciski palców, rozpoznawanie twarzy, rozpoznawanie głosu czy inne modalności, takie jak sygnały ECG. Badanie [15] analizuje wyzwania związane z zestawem danych XM2VTS w kontekście uwierzytelniania biometrycznego, wynikające z różnic w oświetleniu, wyrazach twarzy i cechach głosowych. W celu ich rozwiązania zaproponowano zastosowanie T-norm, czyli podejścia opartego na logice rozmytej, które pozwala radzić sobie z niepewnością. Poprzez

integrację rozpoznawania twarzy i głosu badanie wykazało, że T-normy zwiększają dokładność uwierzytelniania dzięki multimodalnemu konsensusowi. Takie podejście poprawia zarówno bezpieczeństwo, jak i adaptacyjność systemu, oferując obiecujące rozwiązanie dla złożoności tego zestawu danych w systemach biometrycznych. Badanie [16] analizuje ograniczenia systemów biometrycznych opartych na jednym typie danych (unimodalnych), które są bardziej narażone na błędy i luki w zabezpieczeniach. Podkreśla zalety systemów biometrycznych multimodalnych, które łączą różne cechy biometryczne, aby poprawić dokładność, zmniejszyć liczbę błędów i zwiększyć zasięg populacji. Systemy te wzmacniają również ochronę prywatności i integralności danych dzięki integracji zróżnicowanych cech biometrycznych. W badaniu zaproponowano multimodalny system biometryczny wykorzystujący uczenie głębokie w celu zwiększenia bezpieczeństwa uwierzytelniania. Przeanalizowano zastosowanie modelu VGG-16, architektury uczenia głębokiego, do połączenia rozpoznawania twarzy z sygnałami elektrokardiogramu (ECG). Dzięki wykorzystaniu wysokorozdzielczych filtrów konwolucyjnych model uchwycił złożone wzorce cech twarzy i fal ECG, zapewniając dokładną identyfikację osób oraz wysoki poziom bezpieczeństwa.

3. Porównanie metod uwierzytelniania głosowego zależnych od tekstu i niezależnych od tekstu

Uwierzytelnianie głosowe jest ważną technologią, która umożliwia identyfikację użytkowników na podstawie ich unikalnych cech głosowych. Technologia ta może być realizowana za pomocą dwóch głównych metod: zależnych od tekstu i niezależnych od tekstu. Każda z tych metod ma swoje specyficzne cechy, zalety i ograniczenia, które należy uwzględniać podczas ich wdrażania w systemach bezpieczeństwa.

3.1. Metody uwierzytelniania głosowego zależne od tekstu

Metody zależne od tekstu wymagają, aby użytkownik wypowiedział wcześniej zdefiniowaną frazę lub hasło, które system porównuje z wcześniej zapisanym wzorcem głosu [17]. Podejście to pozwala systemowi skupić się na określonych wzorcach dźwiękowych, co zwiększa dokładność rozpoznawania. Implementacja takich metod obejmuje kilka kluczowych etapów technicznych, z których każdy wymaga szczególnej uwagi w celu zapewnienia dokładności i niezawodności uwierzytelniania.

Na pierwszym etapie, podczas rejestracji użytkownika, osoba wypowiada zdefiniowaną frazę kilkukrotnie. Nagrania te są przetwarzane w celu stworzenia wzorcowego profilu głosu, który jest przechowywany w bazie danych. Aby zapewnić wysoką jakość nagrania, należy używać mikrofonów o wysokiej rozdzielczości oraz stosować techniki wstępnego przetwarzania sygnału, takie jak usuwanie szumów i normalizacja poziomu głośności. Pozwala to uzyskać czyste i wyraźne nagranie głosu, które może być używane do dalszego porównania.

Kolejnym istotnym krokiem jest ekstrakcja cech z sygnału głosowego. Proces ten ma kluczowe znaczenie dla dokładności systemu, ponieważ określa, które cechy głosu będą używane do porównania. Jedną z najczęściej stosowanych metod ekstrakcji cech jest wykorzystanie mel-częstotliwościowych współczynników cepstralnych pozwalających wydobyć cechy głosu odporne na zmiany barwy i głośności. Inne metody, takie jak liniowe kodowanie predykcyjne, są wykorzystywane do modelowania traktu głosowego i ekstrakcji jego cech. Można także tworzyć unikalne odciski głosu (voiceprint) na podstawie cech spektralnych.

Po ekstrakcji cech system przechodzi do porównania nagranej frazy użytkownika z wzorcem. Wykorzystywane są różne algorytmy, wśród których dynamiczne dopasowywanie czasu jest jednym z najskuteczniejszych. Metoda ta pozwala wyrównywać różnice czasowe między wzorcem a nowym nagraniem, zapewniając poprawne porównanie. Ponadto stosuje się metody uczenia maszynowego, takie jak maszyny wektorów nośnych (ang. SVM – Support Vector Machine) czy sieci neuronowe, do klasyfikacji cech głosu. Analiza korelacyjna może być również stosowana do mierzenia podobieństwa między dwoma sygnałami głosowymi na podstawie współczynników korelacji. Podejścia te zapewniają bardziej wysoką odporność na różnorodne zniekształcenia i zmiany w sygnale głosowym co owocuje poprawą dokładności i niezawodności systemu.

Zalety:

1. Wysoka dokładność rozpoznawania: ponieważ system porównuje głos użytkownika ze stałym wzorcem, prawdopodobieństwo błędnego odrzucenia lub zaakceptowania znacznie się zmniejsza.
2. Prosta implementacja: algorytmy uwierzytelniania zależne od tekstu są zazwyczaj mniej skomplikowane i wymagają mniejszych zasobów obliczeniowych.

Wady:

1. Wrażliwość na ataki z użyciem nagrań: intruzi mogą nagrać głos użytkownika i użyć tego nagrania do uzyskania nieautoryzowanego dostępu.
2. Ograniczona elastyczność: użytkownicy muszą wypowiadać tą samą frazę, co w niektórych sytuacjach może być niewygodne.

3.2. Metody uwierzytelniania głosowego niezależne od tekstu

Metody niezależne od tekstu nie wymagają od użytkownika wypowiadania konkretnej frazy, czyli system analizuje unikalne cechy głosu niezależnie od treści wypowiadanych słów [18]. To podejście zapewnia dużą elastyczność i wygodę użytkowania, jednak wymaga uwzględnienia kilku skomplikowanych aspektów technicznych, aby zagwarantować niezawodność i dokładność uwierzytelniania.

Na etapie rejestracji użytkownik wypowiada kilka różnych fraz. Te nagrania są przetwarzane w celu stworzenia wzorcowego profilu głosu, który odzwierciedla unikalne cechy użytkownika. Aby zapewnić wysoką jakość nagrania, używa się mikrofonów o wysokiej rozdzielczości. Ważne jest również wstępne przetwarzanie sygnału, które obejmuje usuwanie szumów, normalizację poziomu głośności oraz redukcję zbędnych artefaktów dźwiękowych. Te kroki gwarantują, że próbki są czyste i gotowe do dalszej analizy.

Kolejnym krokiem jest ekstrakcja cech, która jest kluczowa, ponieważ na jej podstawie odbywa się identyfikacja użytkownika. W metodach niezależnych od tekstu stosowane są zaawansowane algorytmy ekstrakcji cech, które pozostają stabilne niezależnie od treści wypowiadanych słów. Jedną z najpopularniejszych metod są mel-częstotliwościowe współczynniki cepstralne, które pozwalają wydobyć podstawowe cechy głosu. Inna ważna metoda to konkatenacja spektralna (ang. spectral concatenation), które zapewnia wysoką dokładność w analizie sygnałów głosowych. Dodatkowo, liniowe kodowanie predykcyjne jest stosowane do modelowania traktu głosowego i ekstrakcji jego cech. W niektórych przypadkach wykorzystuje się metody głębokiego uczenia, które automatycznie wydobywają i optymalizują cechy z danych głosowych bez wyszczególnienia etapu ekstrakcji cech.

Metody niezależne od tekstu wymagają skomplikowanych algorytmów porównywania próbek głosu. Jednym z kluczowych podejść jest wykorzystanie modeli głębokiego uczenia, takich jak rekurencyjne sieci neuronowe (RNN) lub splotowe sieci neuronowe (ang. CNN –Convolution Neural Network). Te modele są trenowane na dużych ilościach danych i potrafią rozpoznawać złożone wzorce w sygnałach głosowych. Ponadto do porównywania nowych nagrań ze wzorcami stosuje się kwantyzację wektorową oraz algorytmy klasteryzacji, takie jak K-means.

Istotnym aspektem niezależnego od tekstu uwierzytelniania jest ustalenie wartości progowej, która określa poziom pewności klasyfikatora co do przynależności próbki do konkretnego użytkownika. Wartość ta jest dostosowywana w taki sposób, aby zrównoważyć prawdopodobieństwo fałszywych odrzuceń (False Negatives) i fałszywych akceptacji (False Positives). Wartość progowa może być dostosowana w zależności od wymagań systemu bezpieczeństwa oraz dopuszczalnego poziomu ryzyka.

Zalety:

1. Elastyczność użycia: użytkownicy mogą wypowiadać dowolny tekst, co sprawia, że proces uwierzytelniania jest mniej uciążliwy.
2. Wyższy poziom bezpieczeństwa: metody niezależne od tekstu są bardziej odporne na ataki z użyciem nagrań, ponieważ opierają się na unikalnych właściwościach głosu, które są trudne do sfałszowania.

Wady:

1. Spadek dokładności: zmienność mowy i trudność w analizie różnych wypowiedzi mogą obniżać dokładność rozpoznawania.
2. Złożoność implementacji: algorytmy niezależnego od tekstu uwierzytelniania są bardziej skomplikowane i wymagają dużych zasobów obliczeniowych do przetwarzania.

Niektóre etapy, takie jak ustalenie wartości progowej pewności klasyfikatora, są analogiczne dla obu kategorii metod uwierzytelniania głosowego. Wartość progowa określa, przy jakim poziomie pewności system uznaje użytkownika za autentycznego. Wartość progowa może być dostosowana do wymagań systemu bezpieczeństwa w celu zminimalizowania liczby fałszywych alarmów i fałszywych odrzuceń.

Głos użytkownika może z czasem się zmieniać, dlatego ważne jest wprowadzenie mechanizmów adaptacji systemu do podobnych zmian. Okresowe aktualizowanie próbek głosu pozwala uwzględnić naturalne zmiany w cechach

głosowych. Wykorzystanie algorytmów uczenia maszynowego, które mogą adaptować się do nowych danych, utrzymują dokładność systemu uwierzytelniania nawet przy zmianach w głosie użytkownika.
Dla niniejszego badania wybrano metody zależne od tekstu (mel-częstotliwościowe współczynniki cepstralne, predykcyjne kodowanie liniowe) oraz metody niezależne od tekstu (ECAPA-TDNN, ResNet), aby porównać ich skuteczność w zadaniach biometrycznego uwierzytelniania głosowego. Eksperyment polegał na opracowaniu systemów biometrycznego uwierzytelniania opartych na każdej z opisanych metod oraz ocenie ich skuteczności na specjalnie zebranym zestawie danych. Porównanie przeprowadzono według następujących kryteriów: False Accept Rate (FAR), False Reject Rate (FRR), Equal Error Rate (EER) i Accuracy.

4. Metody ekstrakcji cech w systemach uwierzytelniania głosowego

4.1. Metoda mel-częstotliwościowych współczynników cepstralnych

Zastosowanie mel-częstotliwościowych współczynników cepstralnych jest jednym z najbardziej popularnych standardowych podejść w systemach uwierzytelniania głosowego [19]. Typowo wykorzystuje się około 20 współczynników MFCC do analizy danych, choć w niektórych przypadkach wystarcza 10-12 współczynników. Główną wadą tej metody jest jej wrażliwość na zakłócenia, ponieważ zależy ona od kształtu widma sygnału. Aby zniwelować ten problem, można stosować metody, które uwzględniają informacje zawarte w okresowości sygnałów mowy.
Skala częstotliwości mel jest liniowa dla częstotliwości poniżej 1 kHz i logarytmiczna dla częstotliwości powyżej 1 kHz. Wynika to z faktu, że ludzki układ słuchowy staje się mniej wrażliwy na zmiany częstotliwości w zakresie powyżej 1 kHz. Charakterystyka amplitudowo-częstotliwościowa jest odpowiedzią filtra logarytmiczno-normalnego. Aby je obliczyć, najpierw wyznacza się energię logarytmu wyjść filtrów za pomocą dyskretnej transformaty Fouriera (DFT), a następnie oblicza się dyskretną transformata cosinusową logarytmicznej energii, aby uzyskać współczynniki MFCC.
Ludzki układ słuchowy jest wrażliwy na zmiany widma w czasie, dlatego często stosuje się analizę temporalnych zmian tych współczynników. Aby uchwycić te zmiany, oblicza się pierwsze i drugie współczynniki różnicowe, które następnie są łączone z współczynnikami statycznymi, tworząc ostateczny zestaw cech reprezentujących dany sygnał mowy.

4.2. Kodowanie liniowo-predykcyjne

Kodowanie liniowo-predykcyjne jest jednym z najbardziej zaawansowanych narzędzi w analizie sygnałów mowy oraz skuteczną metodą kodowania mowy o niskiej przepływności (szybkości transmisji bitów). Zasada działania LPC polega na modelowaniu obecnej próbki mowy jako liniowej kombinacji poprzednich próbek mowy. LPC opiera się na modelu filtrowania, który integruje funkcje różne komponenty układu głosowego człowieka do modelu filtra wielobiegunowego, imitujący akustykę toru głosowego.
Podczas analizy LPC minimalizowana jest suma kwadratów różnic pomiędzy rzeczywistym sygnałem mowy a sygnałem szacowanym na danym odcinku czasowym, co pozwala na uzyskanie predykcyjnych współczynników dla każdego przedziału czasowego sygnału mowy, zazwyczaj o długości 20 ms. Model ten uwzględnia również decyzje o tym, czy dana ramka mowy jest dźwięczna czy bezdźwięczna, poprzez algorytmy analizy tonu, które modyfikują parametry częstotliwości tonu [20].

4.3. ECAPA-TDNN

Model Emphasized Channel Attention, Propagation, and Aggregation Time Delay Neural Network (ECAPA-TDNN) [21], opracowany w 2020 roku, integruje tradycyjną architekturę TDNN z mechanizmami uwagi, zwracając szczególną uwagę na kanały propagacji i agregację cech kontekstowych. ECAPA-TDNN znacząco poprawia ekstrakcję cech, szczególnie w zadaniach uwierzytelniania głosowego.
ECAPA-TDNN jest zbudowany na mel-akustycznych spektrogramach, które przedstawiają sygnał w domenie czasowej i są intuicyjnym narzędziem do rozpoznawania typów sygnałów dźwiękowych. Ponadto, informacje częstotliwościowe oferują większą precyzję i są mniej podatne na zakłócenia niż czasowa reprezentacja sygnału. Architektura ECAPA-TDNN uwzględnia dodatkowe warstwy kontekstowe, aby poprawić rozpoznawanie mówców.

4.4. ResNet

Architektura ResNet (skrót od Residual Network, czyli sieć szczątkowa) [16], która odniosła sukces w dziedzinie komputerowego rozpoznawania obrazów, znalazła również zastosowanie w systemach uwierzytelniania głosowego. Wykorzystuje ona spektrogramy i mel-częstotliwościowe współczynniki cepstralne jako dane wejściowe do sieci neuronowej. W systemach uwierzytelniania głosowego ResNet jest odpowiedzialny za wyodrębnianie złożonych cech sygnałów audio i przekształcanie ich w reprezentacje wektorowe, co pozwala na dokładniejsze rozpoznawanie głosu.
Główną zaletą ResNet w systemach głosowych jest zdolność do generalizacji i wydobywania złożonych wzorców dzięki swojej głębokiej strukturze. Pozwala to na precyzyjne uczenie się na dużych zbiorach danych, co jest kluczowe dla dokładnego rozpoznawania mówców.

5. Metodologia eksperymentu i jego rezultaty

5.1. Opis zestawu danych

W ramach tego badania został zebrany specjalny zestaw danych o wielkości 245 MB, składający się z nagrań 10 anglojęzycznych celebrytów czytających audiobooki. Każdy mówca ma 20 nagrań o długości do 16 sekund. Łącznie zestaw danych zawiera 200 nagrań o łącznym czasie trwania około 30 minut.
Każde nagranie w zestawie danych to jednokanałowe audio z częstotliwością próbkowania 16 kHz w formacie .wav. Format .wav wykorzystuje kodowanie PCM (Pulse Code Modulation) bez kompresji, co zapewnia wysoką jakość i precyzję plików dźwiękowych.
Zestaw danych został zaprojektowany tak, aby był różnorodny, obejmując mówców różnej płci i wieku, co zapewnia kompleksową ocenę modeli uwierzytelniania głosowego.

5.2. Metryki oceny skuteczności

False Accept Rate (FAR), czyli współczynnik fałszywych akceptacji określa odsetek nieuprawnionych użytkowników, którzy zostali błędnie zaakceptowani przez system jako uprawnieni. Ta metryka jest kluczowa w ocenie niezawodności systemu biometrycznej autentyfikacji w kontekście zapobiegania dostępowi nieautoryzowanemu:

$$FAR = \frac{FN}{N}, \tag{1}$$

gdzie:

FN - liczba fałszywych akceptacji,

N - całkowita liczba prób dostępu przez nieuprawnionych użytkowników.

False Reject Rate (FRR) lub współczynnik fałszywego odrzucenia określa odsetek uprawnionych użytkowników, którzy zostali błędnie odrzuceni przez system. Ta metryka ma znaczenie dla oceny wygody korzystania z systemu przez legalnych użytkowników:

$$FRR = \frac{FP}{N}, \tag{2}$$

gdzie:

FP - liczba fałszywych odrzuceń,

N - całkowita liczba prób dostępu przez uprawnionych użytkowników.

Equal Error Rate (EER) lub wskaźnik równych błędów, to punkt, w którym FAR i FRR są równe. EER jest ogólnym wskaźnikiem skuteczności systemu biometrycznego, gdzie niższe wartości EER oznaczają lepszą ogólną wydajność systemu.
Accuracy lub dokładność klasyfikacji określa ogólną proporcję poprawnych decyzji systemu, obejmując zarówno uprawnionych, jak i nieuprawnionych użytkowników:

$$Accuracy = \frac{TP+TN}{N}, \tag{3}$$

gdzie:

TP - liczba poprawnych akceptacji (True Positive),

TN - liczba poprawnych odrzuceń (True Negative),

N - całkowita liczba prób dostępu.

Te metryki razem dostarczają kompleksowej oceny skuteczności i niezawodności systemów biometrycznej autentyfikacji.

5.3. Przebieg eksperymentu

Eksperyment polegał na rejestracji klas w systemie biometrycznej autentyfikacji oraz na wprowadzeniu do systemu próbek głosu zarówno zarejestrowanych (uprawnionych) użytkowników, jak i osób obcych (nieuprawnionych). Dla każdego podejścia została opracowana oddzielna aplikacja uwierzytelniająca, która dokonuje analizę wprowadzonych próbek głosu. Celem eksperymentu było sprawdzenie skuteczności modeli weryfikacji użytkowników oraz ocena zdolności systemów do odróżniania uprawnionych użytkowników od potencjalnych złośliwych podmiotów.
Eksperyment wykorzystał cztery różne metody ekstrakcji cech z próbek głosowych, które zostały szczegółowo opisane w sekcji 4. Metody oparte na Mel-Frequency Cepstral Coefficients (MFCC) oraz Linear Frequency Cepstrum (LFC) polegały na obliczaniu i przechowywaniu surowych zestawów cech dla każdego zarejestrowanego użytkownika. Z kolei metody ECAPA-TDNN i ResNet wykorzystywały sieci neuronowe do generowania wektorów osadzeń (embeddingów) z próbek głosowych. Dla każdego użytkownika te osadzenia były uśredniane na podstawie wielu nagrań, aby stworzyć reprezentatywny wektor osadzenia głosowego.
Proces klasyfikacji opierał się na pomiarze kosinusowej odległości między próbką testową głosu a wzorcem głosowym (voiceprint) zarejestrowanym w systemie. Proces przebiegał w następujących fazach:

Faza rejestracji:
Dla każdego zarejestrowanego użytkownika tworzono wzorcowy głos. W przypadku metod ECAPA-TDNN i ResNet był to uśredniony wektor cech uzyskany na podstawie wielu nagrań, natomiast w przypadku MFCC i LFC przechowywano bezpośrednio surowe zestawy cech.

Faza kalibracji progu klasyfikacji:
Próg klasyfikacji został skalibrowany w celu osiągnięcia optymalnej równowagi między precyzją a czułością systemu. Miało to na celu minimalizację współczynników fałszywego przyjęcia (FAR) i fałszywego odrzucenia (FRR), zapewniając skuteczne rozróżnianie pomiędzy poprawną autoryzacją użytkowników a odrzucaniem prób dostępu osób nieautoryzowanych.

Faza identyfikacji:
Podczas testów próbka głosu była przetwarzana w celu wyodrębnienia wektora cech (embeddingu). Jeśli kosinusowa odległość była mniejsza od zdefiniowanego progu, system klasyfikował próbkę testową jako należącą do odpowiedniego zarejestrowanego użytkownika. W przeciwnym razie próbka była uznawana za pochodzącą od osoby nieuprawnionej.
Dla każdego użytkownika przeprowadzono 20 prób autoryzacji: 10 z wykorzystaniem poprawnych próbek ich głosu oraz 10 z użyciem losowych próbek głosów innych użytkowników zarejestrowanych w systemie. Na podstawie uzyskanych wyników obliczono dokładność klasyfikacji systemu.
Wyniki tych eksperymentów pozwoliły na ocenę skuteczności poszczególnych systemów w zadaniu biometrycznej autentyfikacji. Na podstawie zebranych danych stworzono Tabelę 1, która prezentuje wydajność systemów w różnych scenariuszach. To umożliwiło porównanie efektywności i niezawodności opracowanych modeli, oceniając je pod kątem zdolności do poprawnej autentyfikacji użytkowników oraz odrzucania prób złośliwego dostępu.

Tabela 1. Wyniki eksperymentu

Nazwa metody/modelu	FAR(%)	FRR(%)	EER(%)	Accuracy (%)
MFCC	5	5	5	90,0
LPC	5	2,5	3,75	95,0
ECAPA-TDNN	0,2	0,2	0,2	97,6
ResNet	0,016	0,016	0,016	98,3

6. Wnioski

Badanie metod głosowego uwierzytelniania biometrycznego jest aktualnie obecnie i ma istotne znaczenie z uwagi na rosnące zapotrzebowanie na niezawodne i wygodne metody identyfikacji tożsamości w nowoczesnym świecie cyfrowym. Autentyfikacja głosowa, oferując naturalny i wygodny bezdotykowy sposób weryfikacji tożsamości, staje się coraz bardziej popularna, zwłaszcza w zastosowaniach takich jak urządzenia mobilne, inteligentne głośniki oraz systemy Internetu Rzeczy. Postęp w dziedzinie sztucznej inteligencji i uczenia maszynowego przyczynia się do zwiększenia dokładności i odporności systemów biometrycznych opartych na głosie, co umożliwia skuteczniejsze odróżnianie prawdziwych użytkowników od fałszywych lub syntetycznych głosów.

W systemach biometrii głosowej etap wstępnego przetwarzania ma na celu modyfikację sygnału mowy w taki sposób, aby był bardziej odpowiedni do ekstrakcji cech analitycznych. Proces ten pozwala na wyodrębnienie sygnałów mowy spośród innych dźwięków oraz tworzenie wektorów cech. Zastosowanie odpowiednich technik przetwarzania wstępnego, takich jak usuwanie szumów, normalizacja energii, okienkowanie czy filtracja wzmacniająca, pozwala na skuteczną ekstrakcję cech i podnosi wydajność systemów. Po zakończeniu etapów wstępnego przetwarzania i ekstrakcji cech, następuje proces klasyfikacji, w którym klasyfikator analizuje każdą próbkę i przypisuje ją do odpowiedniej klasy na podstawie zdefiniowanych wcześniej cech.

Celem badań było sprawdzenie skuteczności czterech różnych metod uwierzytelniania użytkowników oraz ocena ich zdolności do odróżniania uprawnionych użytkowników od potencjalnych złośliwych podmiotów. Dla każdego użytkownika stworzono specjalne pary nagrań: wewnątrz tej samej klasy (dla prawdziwych użytkowników) oraz między różnymi klasami (dla osób zewnętrznych). Ocenę wydajności modeli biometrycznej weryfikacji dokonano w oparciu o tradycyjne metryki FAR, FRR, EER i Accuracy.

Eksperymenty przeprowadzone w ramach niniejszych badań wykazały poprawę skuteczności systemów biometrycznej autentyfikacji głosowej dzięki połączeniu metod opartych na analizie widmowej sygnału głosowego oraz technologii nauczania maszynowego. Najlepsze rezultaty uzyskano dzięki zastosowaniu sieci ResNet, która osiągnęła dokładność na poziomie 98,3%, a wskaźniki FAR, FRR, EER – 0,016%. Wskazuje to na wysoką niezawodność tej metody i jej potencjalne zastosowanie w rzeczywistych warunkach, szczególnie tam, gdzie bezpieczeństwo danych oraz łatwość użycia są kluczowe.

Więc biometryczne uwierzytelnianie głosowe ma ogromny potencjał w kontekście zwiększającego się zapotrzebowania na bezpieczne, łatwe w użyciu oraz niezawodne systemy identyfikacji użytkowników, szczególnie w transakcjach finansowych, dostępie do danych poufnych oraz innych krytycznych aplikacjach, gdzie tradycyjne metody oparte na hasła lub kody PIN, mogą być niewystarczająco bezpieczne lub wygodne. Jednym z kluczowych wyzwań pozostaje jednak rozwój mechanizmów, które skutecznie zapobiegną atakom z wykorzystaniem syntetycznych głosów oraz klonowania głosu. Dlatego przyszłe badania powinny skupiać się na opracowywaniu systemów, które nie tylko rozpoznają mówców, ale również będą w stanie wykrywać i neutralizować zagrożenia wynikające z zaawansowanych technologii manipulacji dźwiękiem. Umożliwi to detekcję oszustw i ograniczanie ryzyko nieautoryzowanego dostępu. Stąd postępy te mogą przyczynić się do jeszcze większej adopcji technologii uwierzytelniania głosowego w codziennym życiu.

Reference

1. VPNCentral. Study reveals the most hacked accounts in 2023. VPNCentral. Available at: https://vpncentral.com/most-hacked-accounts-study/ Accessed on: 19 November 2024.
2. Grand View Research, *Biometrics Industry Report*, San Francisco, CA, 2024. Available at: https://www.grandviewresearch.com/industry-analysis/biometrics-industry. Accessed on: November 19, 2024.
3. Amjad Hassan Khan, M. K., & Aithal, P. S. Voice Biometric Systems for User Identification and Authentication – A Literature Review. International Journal of AppliedEngineering and Management Letters (IJAEML), 2022, 6(1), 198-209. DOI: https://doi.org/10.5281/zenodo.6471040
4. Abe, B. C., Araromi, H. O., Shokenu, E. S., Idowu, P. O., Babatunde, J. D., Adeagbo, M. A., & Oluwole, I. H. Biometric Access Control Using Voice and Fingerprint. Engineering And Technology Journal, 2022, 7(7), 1376–1382. https://doi.org/10.47191/etj/v7i7.08
5. Chen, X., Li, Z., Setlur, S., & Xu, W. Exploring racial and gender disparities in voice biometrics. Scientific Reports, 2022, 12(1). https://doi.org/10.1038/s41598-022-06673-y
6. Uddin Prince, N., Al Masum, A., Abdullah, S. M., & Bhuiyan, T. (2024). Voice recognition by deep transfer learning and vision transformers to secure voice authentication. *World Journal of Advanced Research and Reviews*, 23(3), 1365-1377. https://doi.org/10.30574/wjarr.2024.23.3.2781

7. Inamdar, F. M., Ambesange, S., Mane, R., Hussain, H., Wagh, S., & Lakhe, P. Voice Cloning Using Artificial Intelligence and Machine Learning: A review. Journal of Advanced Zoology, 2023, 44(S7), 419–427. https://doi.org/10.17762/jaz.v44is7.2721
8. A. Habeeb. Comparison between physiological and behavioral characteristics of biometric system. Journal of Southwest Jiaotong University, 2019, 54(6). https://doi.org/10.35741/issn.0258-2724.54.6.43
9. K. Win, K.Li, J. Chen, P. Viger, Fingerprint classification and identification algorithms for criminal investigation: A survey, Future Generation Computer Systems, 2020 , vol. 110, pp. 758–771, doi: 10.1016/j.future.2019.10.019
10. J. Daugman, "How iris recognition works," Proceedings. International Conference on Image Processing, Rochester, NY, USA, 2002, pp. I-I, doi: 10.1109/ICIP.2002.1037952.
11. Poddar, Arnab; Sahidullah, Md; Saha, Goutam. Speaker verification with short utterances: a review of challenges, trends and opportunities. IET Biometrics. 2017, 7 (2). Institution of Engineering and Technology (IET): 91–101. doi:10.1049/iet-bmt.2017.0065. ISSN 2047-4938.
12. A. Khan Mk & S. Aithal. Identification of customer through voice biometric system in call centres, *International Journal of Intelligent Systems and Applications (IJISA)*, 2024, vol. 16, no. 5, pp. 68–78, doi: 10.5815/ijisa.2024.05.06.
13. A. Hery, J. Oluwaseyi, F. Olaoye, & H. Luz. Audio deepfakes: threats to voice assistants and voice-activated systems,2024.
14. J. C. Bernal-Romero, J. M. Ramírez-Cortés, & J. de J. Rangel-Magdaleno. (2024). Unbreakable biometrics: How physical unclonable functions are revolutionizing security. *IEEE Instrumentation & Measurement Magazine*, 27(2), 71–78. https://doi.org/10.1109/MIM.2024.10472986.
15. L. Hellal, N.-E. Boukezzoula, M. Cheniti, & K. Chenni. Enhancing authentication security: A fusion of face and voice recognition, 2023.
16. S. Madduluri, & T. Kishorekumar. Multimodal biometric authentication system for military weapon access: Face and ECG authentication, *International Journal of Computational and Experimental Science and Engineering*, 2024, vol. 10, no. 4, doi: 10.22399/ijcesen.565.
17. Liu, Y., Qian, Y., Chen, N., Fu, T., Zhang, Y., & Yu, K. Deep feature for text-dependent speaker verification. Speech Communication, 2015 , 73, 1–13. https://doi.org/10.1016/j.specom.2015.07.003
18. Heigold, G., Moreno, I., Bengio, S., & Shazeer, N. End-to-end text-dependent speaker verification. 2016 , https://doi.org/10.1109/icassp.2016.7472652
19. Xu, M., Duan, L. Y., Cai, J., Chia, L. T., Xu, C., & Tian, Q. HMM-Based Audio Keyword Generation. In Lecture notes in computer science, 2004, pp. 566–574. https://doi.org/10.1007/978-3-540-30543-9_71
20. Wijoyo, S. Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot. 2011, http://fportfolio.petra.ac.id/user_files/97-031/E091%20full%20paper-Thiang%20-%20ICIEE%202011.pdf
21. Desplanques, B., Thienpondt, J., & Demuynck, K. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. 2020 , https://doi.org/10.21437/interspeech.2020-2650
22. M. Jakubec, E. Lieskovska and R. Jarina, Speaker Recognition with ResNet and VGG Networks, 2021 31st International Conference Radioelektronika, Brno, Czech Republic, 2021, pp. 1-5, doi: 10.1109/RADIOELEKTRONIKA52220.2021.9420202.