Jakub GAŁKA, Mariusz MĄSIOR, Michał SALASA

AGH Akademia Górniczo-Hutnicza, Wydział Informatyki, Elektroniki i Telekomunikacji, Katedra Elektroniki

System kontroli dostępu oparty na biometrycznej weryfikacji głosu

Streszczenie. *Artykuł przedstawia koncepcję głosowego, biometrycznego systemu dostępowego zrealizowanego jako system wbudowany. Zaprezentowano najważniejsze wymagania dotyczące systemów kontroli dostępu oraz wynikające z nich założenia projektowe. Opisano architekturę utworzonego systemu, jego funkcjonalność oraz zastosowane metody weryfikacji mówcy wraz z omówieniem podstawowych metod optymalizacji czasowej implementacji. Całość poprzedzona jest zarysem zagadnienia biometrii głosu oraz automatycznego przetwarzania mowy.*

Abstract*. The paper presents the concept of embedded solution for voice biometric access system. The most important requirements for access control systems are presented, as well as the resulting design intent. The architecture of the created system, its functionality and the methods used to verify the speakers is described along with a discussion of basic time-optimization methods of implementation. The entirety is preceded by an outline of the issues of voice biometrics and automatic speech processing.* ***The concept of embedded solution for voice biometric access system***

Słowa kluczowe: rozpoznawanie mówcy, biometria, system wbudowany, kontrola dostępu.
Keywords: speaker recognition, biometry, embedded systems, access control.

doi:10.12915/pe.2014.11.63

Wstęp

Artykuł przedstawia podstawowe założenia projektowe dla realizacji automatycznego zamka do drzwi, który do weryfikacji tożsamości i przydzielania dostępu wykorzystuje biometrię głosu ludzkiego. Spełnia ponadto założenia normy PN-EN 50133-1:1996 dotyczącej systemów kontroli dostępu [1]. Przedstawione urządzenie zrealizowane zostało jako system wbudowany oparty o mikrokontroler ze rdzeniem Cortex-M4F.

W kolejnych rozdziałach omówione zostaną: wymagania normatywne dla systemów kontroli dostępu, podstawowe pojęcia z zakresu biometrii głosowej, sposoby przetwarzania mowy w systemach wbudowanych, architektura zrealizowanego systemu, rejestracja i weryfikacja sygnału mowy, implementacja i optymalizacja algorytmów, dyskusja i podsumowanie.

System Kontroli Dostępu

Systemem kontroli dostępu nazywamy system obejmujący wszystkie składniki konstrukcyjne i organizacyjne niezbędne do sterowania dostępem [1]. Chronionym obszarem może być pojedyncze pomieszczenie lub też kilka pomieszczeń stanowiących zdefiniowaną strefę. Najważniejszym, z punktu widzenia bezpieczeństwa, elementem systemu kontroli dostępu jest przetwarzanie danych, których wynikiem jest decyzja o stanie kontrolowanego przejścia. To przetwarzanie danych realizowane jest w obrębie większego procesu zwanego procedurą przyznawania dostępu, której zadaniem jest rozpoznanie i zweryfikowanie użytkownika ubiegającego się o dostęp. Od jakości metod wykorzystywanych w tej procedurze zależy skuteczność i bezpieczeństwo całego systemu.

Norma PN-EN 50133-1:1996 podaje cztery klasy rozpoznania, które opierają się na poziomie wiarygodności identyfikacji uprawnionych użytkowników. „Klasa 0”, o najniższym stopniu bezpieczeństwa, nadawana jest urządzeniom, które przydzielają dostęp na podstawie zwykłego zapytania o dostęp, bez podania tożsamości (np. zainstalowane detektory ruchu). Urządzenia „klasy 1” przydzielają dostęp do chronionego obszaru po podaniu informacji zapamiętanej, którą może być np. ustalone hasło albo numer identyfikacyjny. „Klasa 2” dotyczy urządzeń realizujących identyfikację bazującą na danych zawartych w identyfikatorze (klucz, karta elektromagnetyczna) lub danych biometrycznych (odcisk palca, barwa głosu).

Tabela 1. Klasy rozpoznania zdefiniowane przez normę PN-EN 50133-1:1996.

System Kontroli Dostępu	
Klasa rozpoznania	Wymagania
Klasa 0	Brak wymagań, swobodny dostęp
Klasa 1	Informacje zapamiętane (kod pin, hasło)
Klasa 2	Identyfikator lub dane biometryczne
Klasa 3	Identyfikator i dane biometryczne

W artykule zaprezentowano projekt systemu kontroli dostępu „klasy 3”, który do identyfikacji wykorzystuje informację zapisaną na karcie zbliżeniowej, a proces weryfikacji użytkownika odbywa się następnie na podstawie biometrii ludzkiego głosu.

Biometria głosowa

Biometria to nauka o metodach mierzenia biologicznych i behawioralnych cech organizmów żywych, w tym ludzi. Wiele z tych cech i ich kombinacji jest unikalnych dla każdego pojedynczego organizmu, co pozwala na odróżnianie ich od siebie. Teza ta jest prawdziwa również w odniesieniu do ludzi. Z tego powodu od dłuższego czasu następuje znaczny rozwój technik i metod biometrycznych stosowanych do identyfikacji i weryfikacji tożsamości ludzi [2].

Obecnie upowszechniło się kilka opisanych niżej metod biometrycznej identyfikacji i (częściej) weryfikacji tożsamości [3]. Daktyloskopia to najczęściej kojarzona z biometrią metoda polegająca na analizie cech charakterystycznych linii papilarnych opuszków palców. Zasadniczo metoda ta opiera się na analizie obrazu odcisku palca [4].

Metoda *finger-vein* polega na analizie obrazu rozkładu naczyń krwionośnych wybranego palca uzyskanego za pomocą prześwietlenia wykonanego w zakresie bliskiej podczerwieni. Jest to obecnie jedna z metod zyskujących znacznie na popularności, również komercyjnie [5].

Często spotykanym sposobem identyfikacji i weryfikacji są metody wykorzystujące rozpoznawanie obrazu twarzy oraz kształtu głowy i uszu. Metody te rozwijają się bardzo szybko, głównie z powodu dostępności dużej ilości danych zgromadzonych np. w sieciach społecznościowych [6,7].

Inne metody biometryczne, wykorzystujące np. obraz tęczówki lub obraz dna oka, analiza kodu DNA czy analiza uzębienia, mimo swojej stosunkowo wysokiej skuteczności, z powodu wysokiej ceny wymaganej aparatury i stopnia skomplikowania wykorzystywane są najczęściej jedynie w wyspecjalizowanych, profesjonalnych zastosowaniach – medycznych, sądowych itp.

Wspomniane wyżej rozwiązania biometryczne posiadają zarówno zalety jak i wady. Każda z metod oferuje różny poziom skuteczności działania oraz zróżnicowany poziom wygody jej użytkowania. Nie bez znaczenia w ich przypadku jest także konieczność stosowania wyspecjalizowanych sensorów (czytniki linii papilarnych, skanery naczyń krwionośnych itp.) w miejscu przeprowadzanej weryfikacji. Jest to stosunkowo kosztowne i często wymaga rozbudowy niezbędnej infrastruktury.

Biometria głosu, jest typem biometrii, który do weryfikacji lub identyfikacji tożsamości wykorzystuje cechy osobnicze obecne w głosie człowieka. Związane są one bezpośrednio z budową anatomiczną traktu głosowego człowieka (cechy niskopoziomowe, widmowe itp.) oraz sposobu mówienia (cechy behawioralne: prozodyczne, fonotaktyczne, językowe) [8,9,10]. Duża liczba rozpatrywanych w praktyce cech, ich zmienność i liczba ich kombinacji powoduje, że głos jest bardzo skuteczną metodą biometrycznej weryfikacji i identyfikacji ludzi [11]. Dodatkową zaletą systemów biometrii głosowej jest także możliwość zdalnej weryfikacji tożsamości za pomocą rozwiązań takich jak połączenie telefoniczne czy aplikacje mobilne, bez zastosowania dodatkowych (poza mikrofonem telefonu) specjalizowanych urządzeń. Biometria głosu jest także bardzo wygodną i naturalną dla człowieka formą rozpoznawania tożsamości, co także jest zaletą przedstawianego w niniejszej pracy rozwiązania.

Zastosowania biometrii głosowej przewidują 3 podstawowe scenariusze działania.

W weryfikacji na podstawie stałego tekstu (ang. *text-dependent speaker verification*), deklarowana tożsamość potwierdzana jest zawsze za pomocą ustalonego wcześniej *hasła głosowego* (ang. *pass-phrase*). Aby zapewnić dostateczne bezpieczeństwo, ale jednocześnie szybkość i wygodę weryfikacji, hasło takie powinno trwać od 1 do 3 sekund i powinno być łatwe i naturalne do wypowiedzenia dla użytkownika danego języka. Jest to metoda o bardzo wysokiej skuteczności i dużym potencjale praktycznym [12]. Taka metoda zastosowana jest w opisywanym urządzeniu.

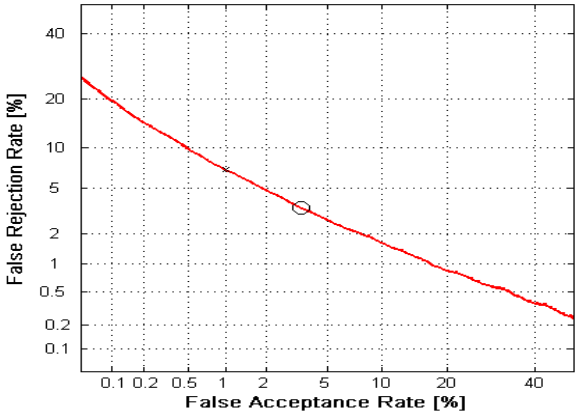
Weryfikacja na podstawie zmiennego tekstu (ang. *text-prompted verification*) polega na każdorazowym zapytaniu osoby weryfikowanej o nowe, zmienne w czasie, hasło głosowe. Metoda ta, mimo niższej bezwzględnej skuteczności biometrycznej ma swoje zalety związane m. in. z ochroną systemu biometrii przed atakami za pomocą różnych środków technicznych (np. nagrań z podsłuchu) [14].Często stosowana jest jako uzupełnienie poprzedniej metody weryfikacji.

Weryfikacja z dowolnym tekstem (ang. *text-independent verification*) wykorzystuje do analizy swobodne wypowiedzi weryfikowanej osoby. Wydaje się, że jest to najwygodniejsza forma weryfikacji, jednak konieczność zgromadzenia wystarczająco dużej ilości danych zarówno do treningu modeli mówcy (efektywnie ponad 30 sekund mowy) jak i do samej weryfikacji (przynajmniej 5 sekund) powoduje, że nie wszędzie znajduje ona zastosowanie. Jest to jednak obecnie najczęściej badana i najszybciej rozwijana przez środowisko naukowe forma weryfikacji mówców [13,15].

Głosowa identyfikacja tożsamości może być realizowana jako wielokrotnie przeprowadzana weryfikacja, z tego powodu zagadnienie to nie będzie dalej omawiane.

Podobnie jak w większości innych systemów biometrycznej weryfikacji tożsamości, weryfikacja głosowa może być opisana jako system detekcji, którego skuteczność można zmierzyć różnymi metodami. Do najważniejszych z nich należą wyrażane procentowo skalarne wskaźniki: poziom błędnych akceptacji FPR (ang. *False Positive Rate*) dla przyjętego punktu pracy (czułości systemu), poziom błędnych odrzuceń FRR (ang. *False Rejection Rate*) dla przyjętego punktu pracy, równoliczny błąd detekcji EER (ang. *Equal Error Rate*), czyli wartość, dla której EER=FPR=FRR, i która syntetycznie opisuje zdolność dyskryminacyjną systemu weryfikacji. Na potrzeby testów NIST zdefiniowano także funkcję kosztu detekcji DCF (ang. *Detection Cost Function*), która uwzględnia zmienną wagę błędnych akceptacji i błędnych odrzuceń w przyjętym punkcie pracy [11].

Do całościowej oceny jakości systemu weryfikacji, niezależnie od wybranego punktu pracy (czułości), stosuje się wykresy DET (ang. *Detection Error Tradeoff*), które wiążą i wizualizują zależność wartości FPR i FRR dla różnych punktów pracy [16]. Na rysunku 1. przedstawiono krzywą DET uzyskaną dla omawianego w niniejszym artykule systemu weryfikacji mówców. Krzywa ta uzyskana została po wykonaniu 192496 prób włamań oraz 35598 prób uprawnionych weryfikacji. Na wykresie zaznaczono również punkt EER o wartości 3,4%.



Rys.1. Krzywa DET uzyskana dla wykorzystywanego systemu weryfikacji mówców wraz z zaznaczonym punktem EER o wartości 3,4%

Przetwarzanie mowy w systemach wbudowanych

Mowa, jako metoda komunikacji ludzi charakteryzuje się konkretnymi wymaganiami dotyczącymi parametrów fizycznych sygnału oraz specyficzną ergonomią. W szczególności, bardzo istotne jest zachowanie krótkich i stałych czasów reakcji w systemach komunikacji głosowej, oraz naturalności tej komunikacji, np. poprzez dopuszczenie wariantowości i zapewnienie komunikatywności dialogu. Podobnie, jak w systemach kodowania i transmisji mowy, zakłada się, że parametry fizyczne cyfrowego sygnału mowy muszą spełniać minima jakościowe, tj. pasmo sygnału nie węższe niż 300Hz-3,7kHz, co w sposób naturalny wymusza typową częstotliwość próbkowania 8 kHz, oraz rozdzielczość bitową sygnału nie gorszą niż 12 bitów/próbkę. W praktyce oznacza to zastosowanie 16-bitowej reprezentacji sygnału mowy [17].

Istotne wymagania dotyczą także małego i stałego czasu odpowiedzi (postrzeganego opóźnienia w dialogu). Dialog postrzegany jest jako naturalny, jeśli średni czas odpowiedzi nie przekracza 0,5 sekundy i nie bywa większy niż 1 sekunda. W przeciwnym wypadku człowiek uczestniczący w dialogu odczuwa dyskomfort i często następuje silna potrzeba powtórzenia zapytania lub też następuje wrażenie błędu w pracy systemu [18][19]. Kwestie te są istotnym przedmiotem badań nad automatycznym dialogiem i powinny być uwzględnione również w głosowych systemach wbudowanych. Skutkuje to

specyficznością omawianego zagadnienia i koniecznością właściwego doboru wykorzystanych metod jak i architektury rozwiązań oraz właściwej implementacji algorytmów przetwarzania mowy. Poniżej przedstawiono przegląd rozwiązań głosowych systemów wbudowanych realizujących funkcjonalności automatycznego rozpoznawania mowy lub biometrii głosowej. Świadomie pominięto bardzo obszerną grupę zagadnień związanych z kodowaniem i kompresją mowy dla celów transmisji, jako nie dotyczącą bezpośrednio niniejszego projektu.

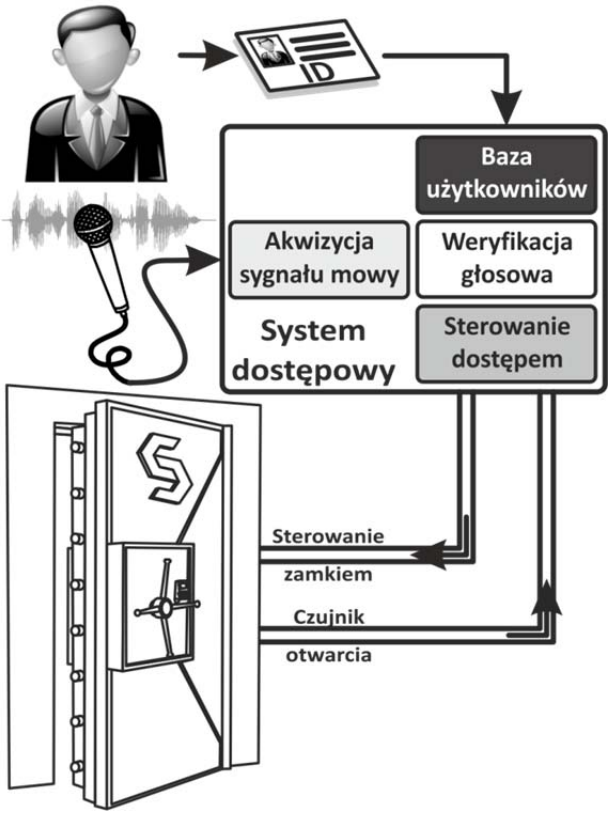
P. Mao oraz J. Liu wykonali system przeprowadzający weryfikację mówcy niezależną od wypowiadanego tekstu. Wykorzystali w tym celu 16-bitowy mikrokontroler z 16-bitowym koprocesorem o taktowaniu 60MHz [20]. Platforma ta wyposażona była dodatkowo w 32KB pamięci SRAM oraz 1MB programowalnej pamięci flash. Do procesu weryfikacji wykorzystano modelowanie GMM-UBM, którego istota przybliżona zostanie w kolejnych rozdziałach. W pracy tej zwrócono głównie uwagę na autorską metodę normalizacji wyników, która pozwoliła przyspieszyć obliczenia i jednocześnie znacząco zmniejszyć poziom wskaźnika EER. Nagrania testowe, które były wykorzystywane przez autorów pochodziły z bazy nagrań *„National 863 Program Office for Intelligent Computing Topics”*. Do adaptacji modeli wykorzystano 10 wypowiedzianych zdań, a do testów 2 zdania. Niestety w artykule nie ma informacji o ostatecznym czasie wykonywania weryfikacji.

Y. S. Moon, C. C. Leung i K. H. Pun zaproponowali implementację weryfikacji mówcy w mobilnym systemie wbudowanym (iPAQ H3600), która po optymalizacji pozwoliła na 37-krotne przyspieszenie wykonywanych obliczeń [21]. Weryfikacja mówcy w przypadku tego rozwiązania również oparta była o modelowanie GMM-UBM. Proces optymalizacji polegał głównie na konwersji typu danych, na których były wykonywane obliczenia (ze zmiennoprzecinkowego na stałoprzecinkowy) oraz zwiększeniu szybkości obliczania funkcji takich jak log, exp, sin i cos. Zabiegi ten pozwoliły na zmniejszenie czasu weryfikacji z 79,6 do 2,16 sekund i utrzymaniu wskaźnika EER na poziomie ok. 8%.W kolejnej pracy [22] opisali także propozycję modyfikacji algorytmu weryfikacji mówcy, która pozwalała zmniejszyć czas wymaganych operacji o 20% zachowując pierwotną skuteczność systemu.

Praca I. Krambergera, i in. przedstawia realizację głosowej weryfikacji mówcy realizowanej przez platformę składającą się z systemu wbudowanego wyposażonego w 32-bitowy mikrokontroler o częstotliwości pracy 160MHz i pamięć SDRAM o wielkości 16MB oraz zewnętrzny serwer z bazą użytkowników [23]. Autorzy wykorzystali prosty system rozpoznawania mowy do identyfikacji użytkownika i modelowanie GMM-UBM do jego weryfikacji. Proces identyfikacji w tym przypadku oparty jest o imię i nazwisko użytkownika. System ASR oparty na niejawnych modelach Markowa konwertuje wypowiadaną kwestię na tekst, porównuje go z bazą i w przypadku prawidłowej identyfikacji uruchamia proces weryfikacji. Proces identyfikacji oraz weryfikacji przebiegają na zewnętrznym serwerze, do którego docierają odpowiednie dane z zainstalowanego przy drzwiach domofonu. Takie podejście nie jest jednak zgodne z wymaganiami stawianymi przez polską normę, ponieważ zgodnie z jej zapisami, cały proces przydzielania dostępu powinien być realizowany bezpośrednio w urządzeniu dostępowym. Eksperymenty przeprowadzone przez autorów wykazały, że ich system weryfikacji mówcy cechuje się, w najkorzystniejszym przypadku, wskaźnikiem EER na poziomie 4.8%.

Architektura omawianego systemu

Architektura omawianego systemu zaprojektowana została zgodnie z normą PN-EN 50133-1:1996 dla systemów kontroli dostępu [1]. Zakłada ona, iż identyfikacja i weryfikacja biometryczna odbywają się bezpośrednio w urządzeniu kontroli dostępu (w zamku biometrycznym). Oznacza to, że urządzenie jest jednostką autonomiczną. Opisywany system składa się z kilku modułów, których zadaniem jest spełnienie określonych wymagań normy. Najważniejszym z nich jest *Centralka Kontroli Dostępu*, która odpowiada za przetwarzanie wszystkich danych docierających do systemu. Pozostałe moduły to: *Moduł Komunikacji* z zewnętrznymi systemami, *Moduł Bazy Użytkowników* oraz *Moduł Rejestracji Danych*.



Rys.2. Działanie biometrycznego systemu kontroli dostępu

Głównym elementem projektowanego systemu jest 32-bitowy mikrokontroler z rdzeniem rodziny ARM Cortex-M4F o maksymalnej częstotliwości pracy 168MHz. Mikrokontroler ten wyposażony jest w 1MB pamięci flash i 96kB pamięci RAM dostępnej dla użytkownika. Procesor Cortex-M4F, dzięki wbudowanemu koprocesorowi zmiennoprzecinkowemu, pozwala na wydajne przetwarzanie sygnałów cyfrowych. Mikrokontroler, wraz z zaimplementowanym systemem operacyjnym czasu rzeczywistego FreeRTOS [24], steruje pracą wszystkich urządzeń peryferyjnych, przetwarza gromadzone informacje i dane biometryczne oraz umożliwia komunikację z zewnętrznym urządzeniem znajdującym się w lokalnej sieci komputerowej.

System wyposażony jest w dodatkową pamięć znajdującą się na karcie SD. Mikrokontroler komunikuje się z nią za pomocą wbudowanego interfejsu SDIO oraz modułu czytnika kart SD. Głównym zadaniem karty SD jest przechowywanie bazy użytkowników, w której zawarte są wszystkie informacje dotyczące ich uprawnień oraz dane biometryczne – odciski głosu, które pozwalają na przeprowadzenie procesu weryfikacji. Kompletny odcisk głosu jednego użytkownika zajmuje 36,46 KB pamięci,

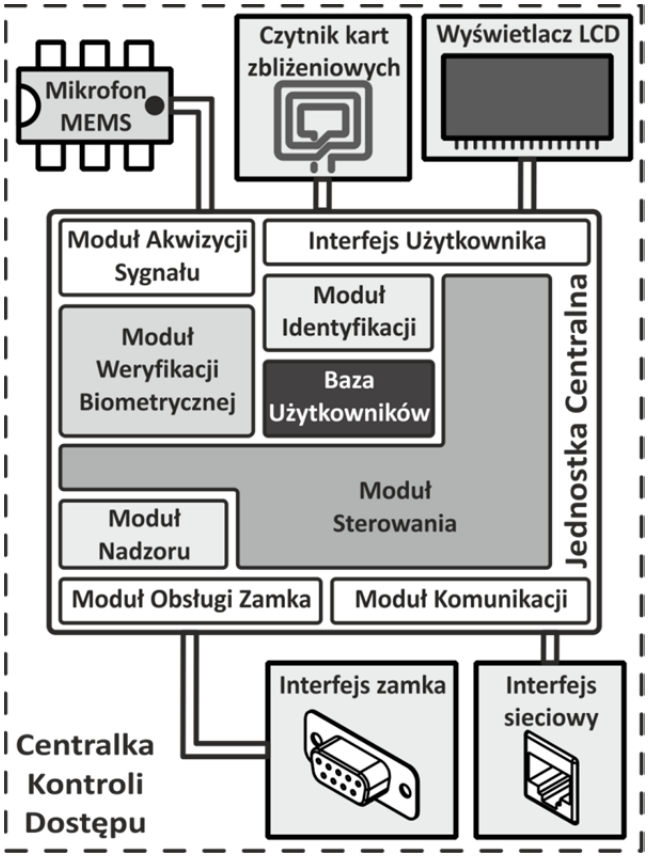
a minimalna liczba odcisków wymagana do skutecznego działania systemu wynosi 30 (są one wymagane do procesu T-Normalizacji, który opisany jest w kolejnych rozdziałach). Poza archiwizacją bazy użytkowników, pamięć karty SD wykorzystywana jest do gromadzenia nagrań pochodzących z procedury przyznawania dostępu oraz do przechowywania pliku konfiguracyjnego, który ładowany jest do pamięci tymczasowej w momencie uruchomienia urządzenia.

Konfigurowanie systemu kontroli dostępu, zarządzanie bazą użytkowników oraz prowadzenie dziennika zdarzeń realizowane jest poprzez aplikację przeznaczoną dla komputera PC, która komunikuje się z urządzeniem za pomocą protokołu TCP/IP. Zastosowanie w projekcie systemu operacyjnego czasu rzeczywistego FreeRTOS (uzyskanie wielowątkowości) pozwoliło na wygodniejszą implementację stosu TCP/IP, dzięki któremu mikrokontroler, poprzez moduł Ethernet DP83848, zdolny jest do wymiany danych. System, wykorzystując protokół DHCP, zdolny jest do połączenia się z siecią lokalną i nawiązania kontaktu z innymi urządzeniami.

Proces przyznawania dostępu do obszaru chronionego zamkiem rozpoczyna się od identyfikacji użytkownika ubiegającego się o dostęp. Identyfikacja polega na porównaniu klucza (numeru identyfikacyjnego) znajdującego się na karcie zbliżeniowej z kluczem przechowywanym w bazie użytkowników. Do odczytu danych z karty zbliżeniowej wykorzystywany jest, przedstawiony na Rys.3., moduł SL018, który pozwala zarówno na przeglądanie zawartości jak i programowanie kart zaprojektowanych w standardzie Mifare [25]. Moduł ten wyposażony jest w odpowiednią antenę oraz posiada własny mikrokontroler, który wykonuje polecenia wysyłane do niego za pomocą interfejsu I2C. Tą samą drogą odbierane są dane odczytane z karty.

Rys.3. Czytnik SL018 oraz brelok funkcjonujący jako karta zbliżeniowa

Do rejestracji sygnału akustycznego, niezbędnego do głosowej weryfikacji tożsamości użytkownika, posłużono się wszechkierunkowym, cyfrowym mikrofonem MP45DT02, wytworzonym w technologii MEMS [26]. Cechą szczególną tego przetwornika jest fakt, że do rejestracji sygnału analogowego wykorzystuje on modulację gęstości impulsów (PDM) [27], dzięki czemu nie jest wymagane wykorzystywanie dodatkowego przetwornika ADC. Konwersja zarejestrowanego sygnału na użyteczną modulację PCM wykonywana jest przez mikroprocesor i polega ona na odpowiedniej cyfrowej filtracji dolnoprzepustowej sygnału. Mikrokontroler STM32, korzystając z interfejsu I2S, wysyła do mikrofonu sygnał taktujący o określonej częstotliwości oraz odbiera rejestrowane przez niego dane. W opisywanym rozwiązaniu w wyniku przetwarzania otrzymujemy sygnał cyfrowy o częstotliwości próbkowania 8000 Hz i 16-bitowej rozdzielczości.



Rys.4. Architektura systemu kontroli dostępu

Rejestracja i przetwarzanie sygnału mowy

W procesie przydzielania dostępu, po prawidłowo zakończonej identyfikacji użytkownika jest on proszony przez urządzenie o wypowiedzenie określonego hasła głosowego (np. „Używam mojego głosu jako klucza"). Wypowiedź ta musi zostać zarejestrowana i trafić do wbudowanej pamięci RAM. Mikrofon, po wypełnieniu swojego wewnętrznego bufora, wysyła do mikrokontrolera sygnał będący źródłem przerwania informującego o konieczności odbioru tych danych w jak najkrótszym czasie.

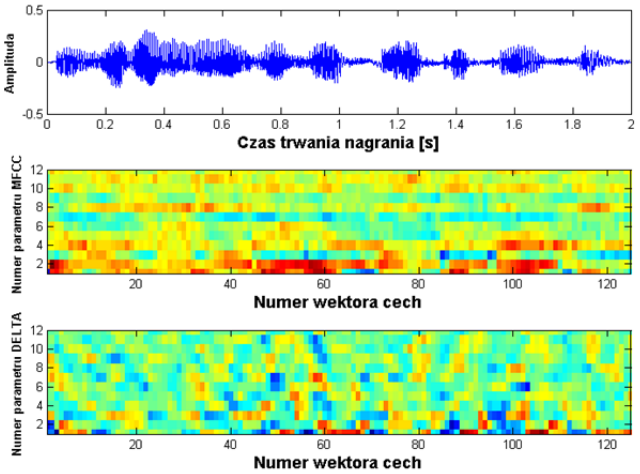
Sygnał akustyczny, zawierający analizowaną wypowiedź hasła, cechuje się dużą, z punktu widzenia biometrycznej weryfikacji głosu, nadmiarowością reprezentacji informacji. Aby zapewnić wydajną obliczeniowo i nieredundantną reprezentację mowy stosowana jest parametryzacja sygnału. W opisywanym rozwiązaniu do parametryzacji zastosowano metodę MFCC (ang. *Mel-Frequency Cepstral Coefficients*) [28].

Problemem, który pojawia się podczas wykorzystywania mikrokontrolera do przetwarzania sygnału akustycznego, jest ograniczona ilość dostępnej pamięci RAM. Sygnał akustyczny próbkowany z częstotliwością 8 kHz, o rozdzielczości 16 bitów i trwający 3 sekundy, co jest typową długością sygnału rejestrowanego przez omawiany system, ma rozmiar 47 kB. Z uwagi na ograniczoną do 96 kB pamięć RAM zastosowanego mikrokontrolera oraz konieczność przechowywania w niej, oprócz sygnału mowy, także tymczasowych zmiennych wielu zadań, zdecydowano się na parametryzację sygnału akustycznego zaimplementowaną jako operację czasu rzeczywistego, dla każdej kolejnej ramki sygnału.

Parametryzacja realizowana jest w czasie rzeczywistym z wykorzystaniem cyklicznego bufora audio. Bufor ten służy do przechowywania tylko jednej (bieżącej) ramki próbkowanego sygnału. Gdy zostanie on wypełniony danymi, jego zawartość jest kopiowana do tymczasowego

wektora, który poddawany jest parametryzacji MFCC. Aby nie doszło do utraty danych, następującej w wyniku zbyt szybkiego nadpisania nieprzeanalizowanego wektora nowymi wartościami, parametryzacja jednej 256 elementowej ramki, przy próbkowaniu z częstotliwością 8 kHz próbek i 50% zakładkowaniem ramki, musi trwać mniej niż 16 ms. Warunek ten został spełniony dzięki wysokiej wydajności rdzenia Cortex-M4F zapewnionej przez wbudowany koprocesor zmiennoprzecinkowy oraz zastosowanie w implementacji instrukcji SIMD. W omawianym rozwiązaniu parametryzacja jednej ramki trwa ok. 6 ms, a rozmiar uzyskanej ostatecznie parametrycznej reprezentacji całej 3-sekundowej wypowiedzi wynosi zaledwie 17 kB.

Uzyskana macierz MFCC poddawana jest dalszemu przetwarzaniu, które ma na celu usunięcie ramek niezawierających użytecznego sygnału mowy (ramek ciszy/szumu). Do tego celu zastosowano, oparty o logarytm energii sygnału, adaptacyjny detektor z histerezą progu detekcji. Kolejne etapy to wyznaczenie zmian parametrów MFCC w czasie (parametry Delta-MFCC) oraz normalizacja histogramów wektorów cech za pomocą metody *Feature-Warping*, która znacząco poprawia odporność systemu na addytywny szum akustyczny i zniekształcenia liniowe związane z warunkami rejestracji sygnału mowy [28]. Przykładowy efekt końcowy wykonania parametryzacji sygnału mowy przedstawiony jest na Rys.3.



Rys.5. Zarejestrowane nagranie hasła „Używam mojego głosu jako klucza" wraz z parametrami MFCC i *delta*-MFCC

Biometryczna weryfikacja mówcy

Weryfikacja polega na sprawdzeniu autentyczności tożsamości użytkownika ubiegającego się w systemie o dostęp. Jej rezultatem jest potwierdzenie lub odrzucenie deklarowanej tożsamości. Dla systemów weryfikacji *text-independent* najnowsze i najskuteczniejsze metody wykorzystują obecnie podejście *i-vectors* oraz często nieliniową klasyfikację, np. SVM [10][29]. W przypadku systemów *text-dependent*, bezpośrednie zastosowanie metody *i-vectors* nie jest obecnie rozpoznane a wynika to z charakteru samego algorytmu. Autorzy niniejszego artykułu pracują obecnie nad zastosowaniem dyskryminatora SVM w przestrzeni Fishera (zmienności wartości pochodnej log-prawdopodobieństwa modelu HMM względem parametrów tego modelu) [38], które dają obecnie bardzo obiecujące wyniki, jednak w praktyce duży rozmiar zbioru wielowymiarowych (ok. 8 tysięcy wymiarów) wektorów wspierających uniemożliwia zastosowanie wprost tego rozwiązania w prezentowanym układzie. Jest to przedmiotem dalszych rozwojowych prac, które mają na celu adaptację tego podejścia dla systemu wbudowanego. Z przedstawionego powodu, w bieżącej, rozwojowej wersji projektowanego systemu, do weryfikacji w trybie ze stałym hasłem głosowym (*text-dependent*) wykorzystywane są metody opierające się na stochastycznym modelu użytkownika i uniwersalnego modelu tła (GMM-UBM, ang. *Gaussian Mixture Models – Universal Background Model*) oraz niejawnych modelach Markowa (HMM) [30]. Wedle wiedzy autorów takie podejście nie było jak dotąd opisywane w literaturze w kontekście głosowych weryfikacyjnych systemów wbudowanych. Na Rys.4. przedstawiono ogólny schemat działania treningu i testowania modeli mówcy.

Podrozdział - Tworzenie modeli mówcy

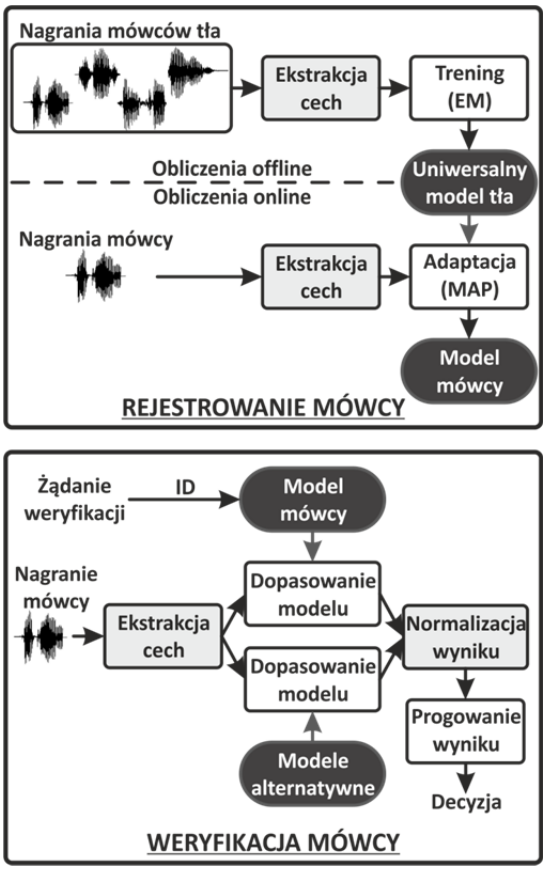
W ogólności, odcisk głosu mówcy reprezentowany jest jako zbiór parametrów (μ, Σ, ω) kombinacji liniowej normalnych rozkładów prawdopodobieństwa - GMM

(1) $$p(\boldsymbol{x}/\lambda) = \sum_{i=1}^{M} \omega_i \, p_i(\boldsymbol{x}),$$

w której każdy komponent opisywany jest jako

(2) $$\mathrm{p_i(x)} = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left(\frac{-1}{2}(\mathrm{x}-\mu_i)'(\Sigma_i)^{(-1)}(\mathrm{x}-\mu_i)\right),$$

gdzie $\boldsymbol{x}$ jest wektorem obserwacji (połączonym wektorem cech MFCC oraz Delta-MFCC), ω_i jest wagą komponentu, μ_i jest wektorem oczekiwanych wartości cech, a $\boldsymbol{\Sigma}_i$ jest diagonalną macierzą kowariancji *i*-tego komponentu.

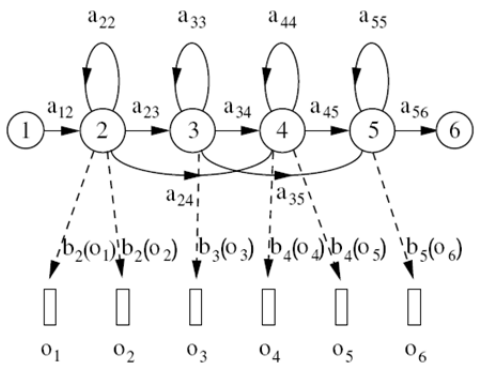


Rys.6. Schemat działania systemu dla faz treningu i weryfikacji mówcy

Aby uwzględnić temporalną strukturę wykorzystywanego hasła biometrycznego, jest ono modelowane jako sekwencja kolejnych stanów modelu HMM, odpowiadających kolejnym głoskom wypowiedzi.

Model HMM składa się z określonej, na podstawie przeciętnej długości hasła głosowego, liczby stanów

emitujących, w których prawdopodobieństwo emisji każdorazowo opisane jest za pomocą przedstawionego wyżej rozkładu GMM. W praktyce, na każde dwie sylaby hasła, przyjmuje się średnio 3 stany emitujące. Prawdopodobieństwa przejść między stanami modelu Markowa opisane są przez kwadratową macierz przejść, która definiuje jednocześnie topologię modelu. Tutaj przyjmuje ona postać typu *left-to-right* (Rys.7.) [31].



Rys.7. Topologia przykładowego modelu HMM typu *left-to-right* z pięcioma stanami emitującymi [31]

Stworzenie modelu głosu użytkownika polega na utworzeniu, specyficznych dla niego, modeli HMM-GMM. Proces ten wykorzystuje uniwersalny model tła (UBM) oraz 3 zgromadzone wcześniej treningowe (rejestracyjne) wypowiedzi danej osoby. Trening przebiega na zewnętrznym urządzeniu, które obsługiwane jest przez administratora systemu kontroli dostępu. Trening modeli mówców opiera się na adaptacji metodą *Maximum A-Posteriori* (MAP) uniwersalnych modeli tła (UBM) do nagrań treningowych danego mówcy [32]. Uniwersalne modele tła tworzone są wcześniej z nagrań reprezentujących mówcę uniwersalnego za pomocą klasycznego iteracyjnego algorytmu *Expectation-Maximization* (EM) z dużego zbioru, specyficznych dla hasła głosowego, nagrań inicjalizacyjnych. W przypadku modelowania HMM-GMM, trening MAP dla modeli HMM realizowany jest jako następująca sekwencja działań:

1) Utworzenie modelu mówcy jako kopii modelu UBM.
2) Wyznaczenie prawdopodobieństwa emisji przez każdy stan modelu HMM mówcy, dla każdej ramki obserwacji pochodzącej z nagrań treningowych.
3) Realizacja algorytmu MAP GMM dla każdego stanu, poprzez adaptację średnich poszczególnych komponentów GMM, proporcjonalnie do wyznaczonego wcześniej prawdopodobieństwa.
4) Uaktualnienie macierzy prawdopodobieństw przejść między stanami, na podstawie nowych wyliczonych parametrów GMM i nagrań treningowych, wykorzystując algorytm EM.
5) Kilkukrotne powtórzenie wykonania kroków od 2 do 4, aż do osiągnięcia zadowalającej adaptacji modelu.

Podrozdział - Weryfikacja mówcy

W trakcie weryfikacji użytkownika, dla całej sekwencji zarejestrowanych wektorów cech wyznaczany jest stosunek logarytmicznego prawdopodobieństwa (inaczej: *Score*) wygenerowania danej wypowiedzi przez model HMM-GMM weryfikowanego mówcy oraz model UBM. Wartość ta wyznaczana jest w oparciu o algorytm *Forward* [33] składający się z trzech etapów:

- Inicjalizacja:

(3) $$\log\left(\alpha_i(j)\right) = \log(\pi_i) + \log\left(b_j(o_1)\right),$$

- Indukcja:

(4) $$\log\left(\alpha_i(j)\right) = \log\left(\sum_{j=1}^{N} \alpha_t(i)\alpha_{ij}\right) + \log\left(b_j(o_{t+1})\right),$$

- Zakończenie:

(5) $$\log\left(P(O/\lambda)\right) = \log\left(\sum_{i=1}^{N} \alpha_T(i)\right),$$

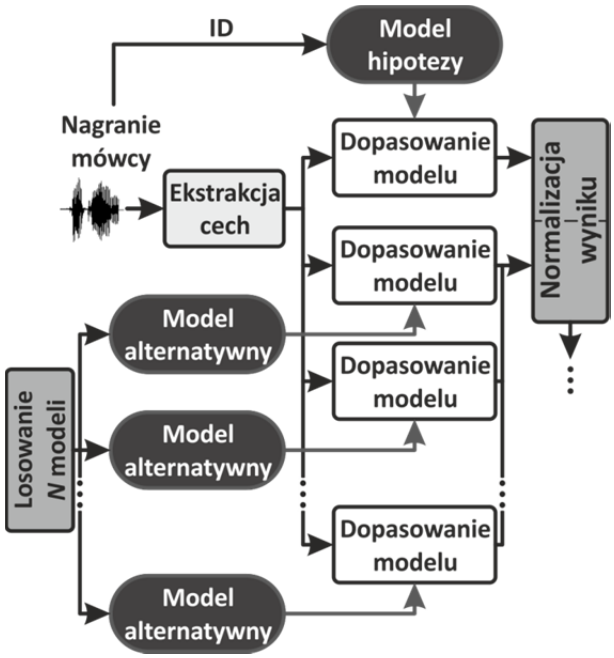
(6) $$\text{Score} = \frac{\log\left(P(O/\lambda)\right)}{\log\left(P(O/\lambda_{UBM})\right)}.$$

gdzie α_i jest prawdopodobieństwem częściowym, π_i to prawdopodobieństwo początkowe *i*-tego stanu, b_i to prawdopodobieństwo wygenerowania wektora obserwacji o_i, opisane zależnością (1), a $P(O/\lambda)$ oznacza prawdopodobieństwo wygenerowania całej obserwacji O przez model λ.

Bardzo ważną dla wysokiej skuteczności i odporności metody na zmienne warunki weryfikacji jest procedura normalizacji. W omawianym przypadku zdecydowano się zastosować, dającą stosunkowo najlepsze wyniki, metodę T-normalizacji [34]. Polega ona na wyznaczeniu scoringów dla określonej, dostatecznie dużej liczby modeli losowo wybranych użytkowników tła, znajdujących się w bazie i normalizacji rozkładu uzyskiwanych wartości zgodnie z zależnością:

(7) $$\text{score}_t = \frac{\text{score} - \mu}{\sigma},$$

gdzie μ oznacza wartość średnią scoringów dla wszystkich modeli z T-normalizacji, a σ jest ich odchyleniem standardowym.



Rys.8. Schemat działania T-normalizacji (losowanie N modeli normalizujących)

Implementacja i optymalizacja algorytmów

Opisane powyżej metody weryfikacji mówcy stanowią duże wyzwanie dla mikroprocesorów wykorzystywanych w systemach wbudowanych. Bardzo szybko rosnąca, wraz z rozmiarem modeli statystycznych, złożoność obliczeniowa wykorzystywanych algorytmów powoduje, że przeprowadzenie wszystkich czynności niezbędnych do wykonania właściwej weryfikacji zajmuje mikroprocesorowi znaczną ilość czasu. Największy wpływ na liczbę obliczeń mają: długość analizowanej wypowiedzi (liczba ramek sygnału), liczba stanów HMM i komponentów GMM, z których składa się model mówcy oraz liczba modeli w zbiorze kohorty mówców wykorzystywanych

do T-normalizacji. Aby umożliwić szybką, zgodną z ergonomią interfejsów głosowych, realizację procedury przyznawania dostępu konieczna jest optymalizacja implementacji tych algorytmów pod kątem wydajności czasowej.

W zależności od dostępnej architektury sprzętowej optymalizacja może przebiegać na kilka różnych sposobów. Jeśli istnieje dostęp do stosunkowo dużej ilości wbudowanej pamięci można zaimplementować tablice LUT (*Lookup Table*), dzięki którym obliczanie wartości takich funkcji jak logarytm czy funkcja eksponencjalna może zostać przyspieszone. Istotna redukcja szybkości działania algorytmów możliwa jest również poprzez konwersję typu danych ze zmiennoprzecinkowego na stałoprzecinkowy (opisana dokładnie we wcześniej cytowanym artykule [21]). Poza tym, mikroprocesor może być wyposażony w dodatkowe moduły, takie jak koprocesor zmiennoprzecinkowy, które wspomagają wykonywanie obliczeń i zwiększają wydajność całego systemu.

W odróżnieniu od optymalizacji wykorzystującej charakterystykę używanego sprzętu istnieją również techniki ingerujące w przebieg procesu weryfikacji mówcy (wykorzystanie zmodyfikowanych modeli mówców, zmiana algorytmów wyliczających scoring) czy też zmniejszające ilość danych wykorzystywanych do analizy. Tego typu modyfikacje są możliwe do przeprowadzenia na wykorzystywanej architekturze i te najbardziej obiecujące zostaną przedstawione w kolejnych akapitach.

Optymalizacja implementacji algorytmów realizujących opisane w poprzednim podejście GMM-UBM jest możliwa dzięki temu, że modele wykorzystywane do weryfikacji mówcy są adaptowane metodą *Maximum A-Posteriori* [35]. Takie rozwiązanie powoduje, że wartość prawdopodobieństwa wygenerowania pojedynczej wybranej ramki analizowanego sygnału jest zdominowana przez zaledwie kilka komponentów modelu GMM, niezależnie od jego całkowitego rozmiaru. Dodatkowo, jeśli modele GMM wybrane do T-normalizacji zostały adaptowane z tego samego modelu UBM, to komponenty dominujące log-prawdopodobieństwo dla danej ramki będą takie same dla każdego modelu z kohorty. Eksperymentalnie wykazano, że liczba dominujących komponentów wynosi 5 [36]. Uwzględnienie tego zjawiska w implementacji pozwala istotnie zredukować liczbę przeprowadzanych obliczeń. Dla każdej analizowanej ramki należy wykonać pełne obliczenia prawdopodobieństwa wygenerowania jej przez model UBM, zapamiętując przy tym udział każdego komponentu. Następnie należy odszukać 5 komponentów, które dominują końcowy wynik i pozostałe obliczenia scoringu wykonać już tylko dla nich.

Efektywność zastosowanej optymalizacji w głównej mierze zależy od liczby komponentów Gaussa, z których składa się model użytkownika oraz liczby modeli, które wykorzystywane są w T-normalizacji. Dla przykładu: optymalizacja systemu wykorzystującego modele GMM składające się z 64 komponentów i 30 elementowej kohorcie wykorzystanej w T–normalizacji pozwala zredukować czas wykonywania obliczeń dla jednego wektora cech 3,5 krotnie [36].

W przypadku zastosowania modelowania HMM-GMM, optymalizacja implementacji opiera się na podobnym rozumowaniu. Zwiększenie wydajności uzyskuje się poprzez odrzucanie z procedury obliczeniowej tych komponentów, których wpływ na prawdopodobieństwo wygenerowania badanego wektora cech jest znikomy. W tym przypadku jest to o tyle bardziej skomplikowane, że nie można założyć, iż te same komponenty danego modelu (np. UBM) będą odpowiadały tym samym komponentom modelu innego. Z tego powodu proces odrzucania najmniej istotnych komponentów musi odbywać się kolejno dla wszystkich stanów każdego modelu wchodzącego w skład T-normalizacji. Ta metoda optymalizacji, nazywana *Dynamic Gaussian Selection*, w najlepszym wypadku pozwala skrócić obliczenia o ok. 30% [37].

Oprócz modyfikowania algorytmów GMM-UBM i HMM możliwe jest również zmniejszenie ilości danych wykorzystywanych do weryfikacji. Najczęściej osiąga się to poprzez ograniczenie liczby analizowanych wektorów cech. W metodzie *Variable Frame Rate* przeprowadza się decymację strumienia wektorów obserwacji (ramek) wykorzystując określoną miarę (najczęściej jest to metryka euklidesowa) do oceny podobieństwa kolejnych wektorów cech [35]. Jeżeli dwa, lub więcej kolejnych wektorów są do siebie dostatecznie podobne, to w dalszych obliczeniach biorą udział jedynie wektory reprezentujące dany fragment wypowiedzi (w metodzie GMM) lub też prawdopodobieństwa uzyskane na podstawie danego wektora są powielane dla pozostałych ramek (w metodzie HMM-GMM). Skuteczność takiego rozwiązania zależy głównie od stopnia decymacji i efektywności metody porównawczej. VFR z metryką euklidesową pozwala zmniejszyć ilość wykorzystywanych wektorów cech o 50% bez istotnego spadku wskaźnika EER [35].

Podsumowanie

Przedstawiony projekt systemu kontroli dostępu spełnia wymaganą funkcjonalność opisywaną przez normę PN-EN 50133-1:1996 i realizuje procedurę weryfikacji mówcy na podstawie biometrycznych cech jego głosu. Urządzenie, wykorzystując zaprezentowany algorytm weryfikacji mówcy, cechuje się wysoką skutecznością (odznaczającą się wskaźnikiem EER o wartości 3,4%), która przewyższa skuteczność systemów w cytowanych artykułach.

Przedstawione w ostatnim rozdziale metody optymalizacyjne, właściwie zaimplementowane, zapewnią wydajną pracę urządzenia i pozwolą na naturalną komunikację systemu z użytkownikiem.

Praca współfinansowana przez Narodowe Centrum Badań i Rozwoju w ramach Programu Badań Stosowanych - projekt nr PBS1/B3/1/2012, pt. „Biometryczna Weryfikacja i Identyfikacja Głosu”.

LITERATURA

[1] Norma PN-EN 50133-1:1996 + AC:1998 + A1:2002, Systemy alarmowe – Systemy Kontroli Dostępu – Część 1: Wymagania Systemowe

[2] Jain A. K., An Introduction to Biometric Recognition, *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 14 (2004), no. 1

[3] ITL Biometrics Overview, http://www.nist.gov/itl/biometrics/

[4] Maltoni D., Maio D., Jaiin A. K., Prabhakar S., Handbook of Fingerprint Recognition, IEEE *NIST Fingerprint Evaluations and Developments,* vol. 94, no. 11 (2006)

[5] Wang J., Yau W., Suwandy A., Sung E., Person Recognition by Fusing Palmprint and Palm Vein Images Based on ‘Laplacianpalm’ Representation*, Pattern Recognition*, vol. 41, issue5 (2008), 1514-1527

[6] Taigman Y., Yang M., Ranzato M., Wolf L., Closing the Gap to Human-Level Performance in Face Verification*, Conference on Computer Vision and Pattern Recognition CVPR* (2014)

[7] Kumar N., Berg A. C., Belhumeur P. N., Nayar S. K., Attribute and Simile Classifiers for Face Verification*, ICCV* (2009)

[8] Reynolds D., Rose R., Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models, *IEEE Transactions on Speech and Audio Processing,* vol. 3, no. 1 (1995)

[9] Kinnunen T., Li H., An Overview of Text-Independent Speaker Recognition: From Features to Supervectors, *Speech Communication,* 52 (2010)*,* 12-40
[10] Campbell J., Speaker Recognition: A Tutorial, *Proceedings of the IEEE,* vol. 8, no. 9 (1997)
[11] Martin A., Greenberg C., The NIST 2010 Speaker Recognition Evaluation, *INTERSPEECH 2010,* p. 2726-2729
[12] Hebert M., Text-Depentent Speaker Recognition, *Springer Handbook of Speech Processing* (2008), 743-762
[13] Greenberg C., Martin A., Barr B., Report on Performance in the NIST 2010 Speaker Recognition Evaluation, *INTERSPEECH 2011*, 261-264
[14] Petrovska-Delacretaz D., Hennebert J., Text-Prompted Speaker Verification Experiments with Phoneme Specific MLP's, *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing* (1998), vol. 2, p. 777 - 780
[15] NIST Speaker Recognition Evaluation 2012, http://www.nist.gov/itl/iad/mig/sre12results.cfm
[16] Martin A., The DET Curve in Assessment of Detection Task Performance, *Eurospeech (1997),* vol. 4, p. 1899-1903
[17] Benesty J., Sondhi M., Huand Y., Springer Handbook of Speech Processing (2008)
[18] Heldner M., Edlund J., Pauses and Overlaps in Conversations, Journal of Phonetics (2010), vol. 38, issue 4, 555-568
[19] Robert F., Alexander L., Francis B., Morgan M., The Interaction of Inter-turn Silence with Prosodic Cues in Listener Perceptions of 'Trouble' in Conversation, *Speech Communication* 48 (2006), 1079-1093
[20] Mao P., Liu J., A Novel Embedded Speaker Verification on System on Chip, *Sixth International Conference on Fuzzy Systems and Knowledge Discovery* (2009)
[21] Moon Y. S., Leung C. C., Pun K. H., Fixed-point GMM-based Speaker Verification over Mobile Embedded System, *WBMA* (2003), Berkeley, California
[22] Leung C. C., Moon Y. S., Meng H., A Pruning Approach for GMM-Based Speaker Verification in Mobile Embedded Systems, *Lecture Notes in Computer Science* (2004), vol. 3072, p. 607 - 613
[23] Kramberger I., Grasic M., Rotovnik T., Door Phone Embedded System for a Voice-Based User Identification and Verification Platform, *IEEE Transactions on Consumer Electronics* (2011)*,* vol. 57, issue 3
[24] FreeRTOS™ Project Homepage, www.freertos.org
[25] SL018 User Manual, *StrongLink Homepage,* www.stronglink-rfid.com
[26] MP45DT02 – MEMS audio sensor omnidirectional digital microphone Datasheet, www.st.com
[27] PDM Audio Software Decoding on STM32 Microcontrollers, *Application Note AN3998,* www.st.com
[28] Kinnunen T., Li H., An Overview of Text-Independent Speaker Recognition - From Features To Supervectors, *Speech Communication 52* (2010), 12-40
[29] Kenny P., Boulianne G., Dumouchel P., Eigenvoice Modeling With Sparse Training Data, *IEEE Transactions On Speech and Audio Processing* (2005), Vol. 13, No. 3
[30] Munteanu D. P., Toma S. A., Automatic Speaker Verification Experiments using HMM, *8th International Conference on Communications* (2010), p. 107 - 110
[31] Young S., Evermann G., Gales M., Hain T., Kershaw D., Liu X., et al, The HTK Book, *Cambridge University Engineering Department* (2009)
[32] Reynolds D., Gaussian Mixture Models, *Encyclopedia of Biometrics* (2009), 659-663
[33] Rabiner L., A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE (1989),* vol. 77, no. 2
[34] Auckenthaler R., Carey M., Lloyd-Thomas H., Score Normalization for Text-Independent Speaker Verification Systems, *Digital Signal Processing 10* (2000)*,* 42-54
[35] McLaughlin J., Reynolds D., Gleason T., A Study of Computation Speed-ups of the GMM-UBM Speaker Recognition System, *Sixth European Conference on Speech Communication and Technology, EUROSPEECH (1999),* Budapest, Hungary
[36] Mohammadi S., Saeidi R., Efficent Implementation of GMM-Based Speaker Verification Using Sorted Gaussian Mixture Models, *14th European Signal Processing Conference, EUSIPCO 2006,* Florence, Italy
[37] Cai J., Bouselmi G., Fohr D., Dynamic Gaussian Selection for Speeding Up HMM-Based Continuous Speech Recognition, *International Conference on Acoustics and Signal Processing (2008),* Las Vegas, USA
[38] Wan V., Renals S., Speaker Verification Using Sequence Discriminant Support Vector Machines, *IEEE Transactions on Speech and Audio Processing* (2005), Vol. 13, Issue 2, 203-210

Autorzy*: dr inż. Jakub Gałka, Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Wydział Informatyki, Elektroniki i Telekomunikacji, Katedra Elektroniki, al. Mickiewicza 30, 30-059 Kraków, E-mail: jgalka@agh.edu.pl;*
mgr inż. Mariusz Mąsior, Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Wydział Informatyki, Elektroniki i Telekomunikacji, Katedra Elektroniki, al. Mickiewicza 30, 30-059 Kraków, E-mail: masior@agh.edu.pl;
inż. Michał Salasa, Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie, Wydział Informatyki, Elektroniki i Telekomunikacji, Katedra Elektroniki, al. Mickiewicza 30, 30-059 Kraków, E-mail: salasa@student.agh.edu.pl.