POZNAN UNIVERSITY OF TECHNOLOGY

Streszczenie rozprawy doktorskiej

# Rozpoznawanie mówcy na podstawie transkodowanej mowy do interfejsów człowiek-maszyna


mgr inż. Radosław Weychan

Instytut Automatyki i Robotyki
Wydział Informatyki
Politechnika Poznańska

Promotor: Prof. dr hab. inż. Adam Dąbrowski
Promotor pomocniczy: dr inż. Tomasz Marciniak



Poznań, 2017

# 1. Cel pracy

Interfejs człowiek-maszyna (HMI, human-machine interface) jest elementem zapewniającym użytkownikowi dostęp do funkcjonalności danego systemu.

W przypadku systemów automatyki, dostęp ten jest możliwy poprzez standardowy, dotykowy panel operatora lub nowoczesne rozwiązania bezdotykowe, takie jak sygnał mowy. Jest to naturalna, intuicyjna forma wydawania rozkazów i typowo niewymagająca długotrwałego treningu. [Tade1988, Rogo2012]. Może być także wykorzystana dodatkowo w celach biometrycznych np. do rozpoznawania mówcy [Pire2007]. Zagadnienie to jest istotne w systemach automatyki ze względu na częste przypisanie operatora do konkretnej maszyny, wysoką trudność obejścia zabezpieczenia zwłaszcza w systemach zdalnego dostępu, a także ze względu na prostotę założeń (brak haseł, fizycznych tokenów itp.). Dzięki wykorzystaniu rozpoznawania mówcy możliwe jest zapewnienie braku reakcji urządzeń na przypadkowe komendy postronnych osób, co daje pełną kontrolę nad systemem.

W literaturze prezentowanych jest wiele rozwiązań systemów automatyki sterowanych głosem, także z możliwością dostępu zdalnego [Brea2013, Gian2005, Jawa2007, Jian2000, Saue2006, Yuks2006]. Jednym z głównych problemów badawczych związanych z zagadnieniem rozpoznawania mówcy jest obniżenie skuteczności poprawnego działania systemu ze względu na długość wypowiedzi, techniki akwizycji i transmisji oraz związane z tym kodowanie sygnału (w przypadku sterowania zdalnego). Dodatkowym problemem jest zmniejszenie precyzji obliczeń w energooszczędnych systemach stałoprzecinkowych w stosunku do typowo stosowanych systemów zmiennoprzecinkowych.

Rozprawa „Speaker recognition based on transcoded speech for human-machine interfaces” („Rozpoznawanie mówcy na podstawie transkodowanej mowy do interfejsów człowiek-maszyna”) prezentuje rezultaty badań dotyczących rozpoznawania mówcy z sygnału mowy obniżonej jakości w zastosowaniach automatyki. Celem badań była analiza możliwości rozszerzenia, sterowanych za pomocą głosu, interfejsów człowiek-maszyna (human-machine interfaces, HMI) o funkcjonalność identyfikacji osoby wydającej polecenie głosowe.

Wspomnianym istotnym czynnikiem mającym wpływ na skuteczność rozpoznawania mówcy (weryfikację lub identyfikację) jest jakość transmisji / nagrania sygnału mowy. W przypadku publicznej komutowanej sieci telefonicznej (public

switched telephone network, PSTN) oraz modulacji szerokości impulsu (pulse-code modulation, PCM), skuteczność rozpoznawania mówcy opisana w literaturze wynosi do 95 %. Jednakże transmisja sygnału mowy w sieci komórkowej czy sieci internet z wykorzystaniem algorytmów stratnego kodowania obniża tę skuteczność, dlatego konieczne jest opracowanie metod uwzględniających kodowanie stratne mowy.

Propozycją autora niniejszej rozprawy jest opracowanie i zastosowanie metod poprawy skuteczności rozpoznawania mówcy na podstawie krótkich wypowiedzi. Zaproponowane i przetestowane w rozprawie rozwiązania pozwoliły zwiększyć skuteczność identyfikacji mówcy w przypadku, gdy sygnał mowy transmitowany jest przez sieć GSM lub internet, a tym samym kodowany stratnie.

## 2. Teza rozprawy

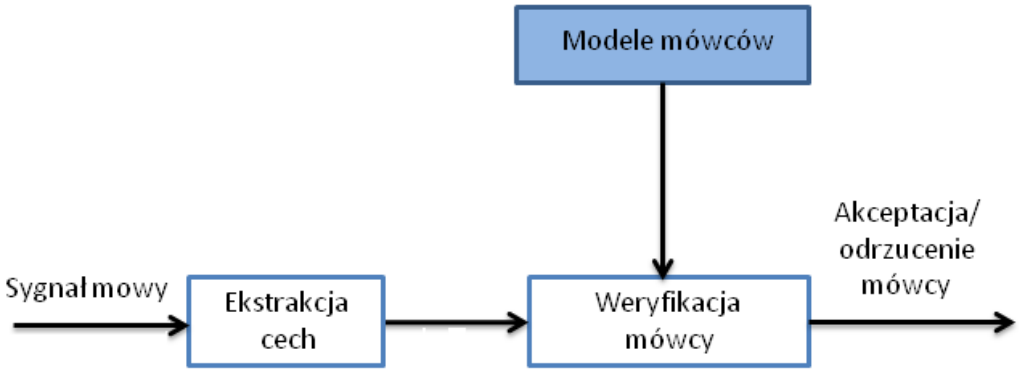
Biorąc pod uwagę opisane powyżej problemy badawcze i możliwe sposoby ich rozwiązania, teza rozprawy doktorskiej sformułowana została następująco:

*Efektywność automatycznego rozpoznawania mówcy z sygnału GSM o małej przepustowości może zostać zwiększona poprzez wykorzystanie opracowanych algorytmów detekcji kodowania oraz wyboru modelu mówcy. Zaproponowane techniki dają możliwość realizacji stałoprzecinkowego wydajnego systemu wbudowanego o niskim poborze mocy.*

## 3. Stan wiedzy

Typowe systemy rozpoznawania mówcy działają w dwóch trybach: treningu, podczas którego generowane są modele mówców, oraz testowania, podczas którego sygnał mowy, po odpowiednim przetworzeniu, porównywany jest z modelami mówców z bazy danych. W następnej kolejności podejmowana jest decyzja o rozpoznaniu/nierozpoznaniu osoby testowanej [Beig2011]. Wspólną częścią obu trybów jest ekstrakcja cech mówcy. Zwykle wykorzystuje się do tego współczynniki mel-cepstralne (mel-frequency cepstral coefficients, MFCC) [Mola2001]. W trybie treningu ze współczynników tych generowany jest model, do wyznaczenia którego najczęściej wykorzystuje się algorytmy kwantyzacji wektorowej (VQ, vector

quantzation) [Lind1980] lub mieszaniny Gaussa (Gaussian mixture models, GMM) [Reyn1995], na których oparty algorytm GMM-UBM (Gaussian mixture models – universal background models) [Reyn2000]. Z kolei w zadaniu testowania wyznaczane jest podobieństwo zestawu cech do każdego z modeli z wykorzystaniem odległości euklidesowej lub logarytmicznego stosunku prawdopodobieństwa (log-likelihood ratio). Schemat typowego systemu rozpoznawania mówcy, ilustrującego tryb testowania, przedstawia Rys. 1.




Rys. 1. Schemat typowego systemu rozpoznawania mówcy

Systemy oparte na powyższym schemacie mają jednakże kilka podstawowych wad. Rezultat rozpoznania jest zależny od długości sygnału, który zgodnie z literaturą powinien wynosić min. 20 sekund. Jest to wartość niemożliwa do osiągnięcia w przypadku systemów sterowania opartych na krótkich zwrotach sterujących. Inną z wad typowego systemu rozpoznawania jest, wspomniana wcześniej, wrażliwość na jakość sygnału mowy. Czynnikiem obniżającym tę jakość w nowoczesnych systemach zdalnego sterowania jest kodowanie stratne, związane z transmisją sygnału mowy przez sieć komórkową GSM lub internet. Analiza literatury dotyczącej systemów sterowania sygnałem mowy pokazuje, iż problem ten jest wciąż niewystarczająco przeanalizowany, brakuje badań uwzględniających aktualnie wszystkie wykorzystywane kodery GSM, badania te dotyczą rozpoznawania mowy, pokazują jedynie wpływ kodowania bez wskazania rozwiązania problemu, proponują rozwiązanie niemożliwe fizycznie do zaimplementowania ze względu na brak dostępu do sygnału przed zdekodowaniem, lub też pokazują skomplikowane rozwiązania trudne do implementacji w systemie czasu rzeczywistego [Deby2010, Dunn2001, Gras2000, Jani2011, Lill1996, McLa2013, Nema2010, Wang2016].

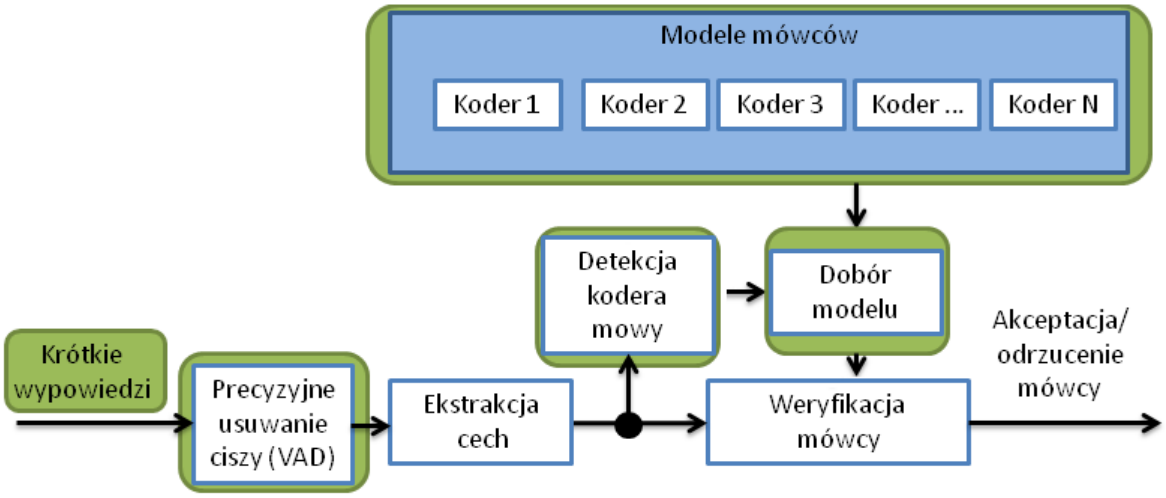
Brakuje także kompleksowego rozwiązania, wykraczającego poza badania symulacyjne. Skuteczność rozpoznawania mówcy we wbudowanych autonomicznych systemach sterowania obniża się w związku z implementacją wykorzystującą stałoprzecinkowe mikroprocesory. Umożliwiają one zmniejszenie zużycia energii przez system wbudowany oraz redukują jego koszty, jednakże w przeciwieństwie do układów wykorzystujących arytmetykę zmiennoprzecinkową, powodują generowanie większych błędów obliczeniowych. Stąd konieczność opracowania dodatkowych technik zmniejszających te błędy w celu uniknięcia obniżenia skuteczności rozpoznawania mowcy.

## 4. Koncepcja zaawansowanego systemu automatycznego rozpoznawania mówcy

Pierwszym opisanym problemem badawczym jest wpływ długości wypowiedzi na skuteczność rozpoznawania mówcy. W celu zwiększenia stopy poprawnej identyfikacji, autor zaproponował wykorzystanie algorytmów detekcji aktywności mówcy, czego rezultatem maksymalizacja zawartości informacyjnej w analizowanym sygnale. Zbadane zostały algorytmy bazujące na energii sygnału, amplitudzie, różnicach wyższych rzędów oraz dynamice przejść przez zero.

Drugim opisanym problemem badawczym jest wpływ kodowania na skuteczność rozpoznawania mówcy. Etap ten podzielony został na dwa powiązane obszary: dotyczący detekcji kodowania w sygnale mowy, oraz wpływu doboru modelu mówcy na skuteczność rozpoznawania z transkodowanej mowy. Schemat zawierający proponowane ulepszenia systemu w stosunku do typowego rozwiązania prezentowany jest na Rys. 2.

Trzecim etapem badań były aspekty implementacyjne dotyczące poprawy skuteczności rozpoznawania mówcy w przypadku wykorzystania arytmetyki stałoprzecinkowej na procesorze sygnałowym TMS320C5515. Przedstawiono także implementację zmiennoprzecinkową opartą na procesorze o architekturze ARM.




Rys. 2. Zastosowane rozszerzenia systemu automatycznego rozpoznawania mówcy

## 4.1 Wykorzystane bazy mówców i algorytmy modelowania

W badaniach autor wykorzystał dwie bazy danych mówców. Pierwsza z nich, baza mówców TIMIT [Grafo1993], opracowana przez Texas Instruments (TI) oraz Massachusetts Institute of Technology (MIT), jest bazą 630 mówców wypowiadających 10 różnych krótkich sekwencji w języku angielskim.

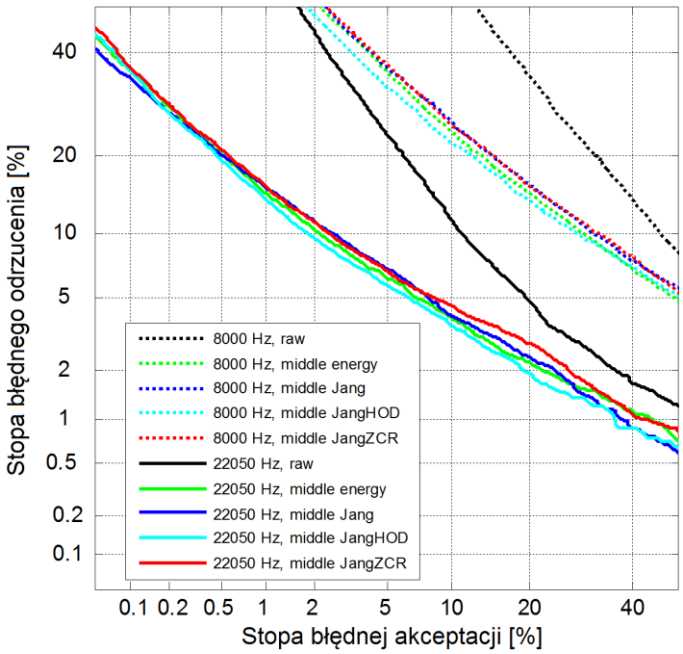
Druga baza, przygotowana przez autora rozprawy, zawiera nagrania 40 mówców wypowiadających 5 krótkich sekwencji w języku polskim 30 razy, po 10 powtórzeń na jedną sesję nagraniową. Odstęp pomiędzy sesjami nagraniowymi wynosił od 1 do 4 tygodni. Łącznie baza zawiera 7200 nagrań krótkich wypowiedzi, nagranych w specjalnie przygotowanym stanowisku w komorze bezechowej.

W trakcie prowadzenia badań autor wykorzystał trzy algorytmy modelowania sygnału mowy – kwantyzację wektorową (VQ), mieszaniny Gaussa (GMM), oraz mieszaniny Gaussa oparte na wspólnym modelu mowy całej populacji (GMM-UBM). Porównane zostały skuteczności rozpoznawania oraz czasy niezbędne zarówno na generację modelu mówcy, jak i na porównywanie modelu z wypowiedzią testową. Najwyższa skuteczność rozpoznawania uzyskana została dla algorytmów GMM oraz GMM-UBM, przy czym w implementacji czasu rzeczywistego wykorzystany został algorytm GMM, ze względu na mniejsze skomplikowanie oraz możliwość szybkiej modyfikacji utworzonej bazy modeli mówców.

## 4.2 Wpływ detekcji aktywności mówcy na skuteczność rozpoznawania mówcy z krótkich wypowiedzi

Pierwszą propozycją autora rozprawy jest wykorzystanie algorytmów detekcji aktywności mówcy (VAD, voice activity detection) w celu maksymalizacji zawartości informacyjnej w sygnale [Marc2010]. Jest to istotny etap przetwarzania wstępnego mowy, zaproponowany w związku z koniecznością minimalizacji obniżenia skuteczności rozpoznawania mówcy z krótkich sekwencji sterujących. Usunięcie z sygnału fragmentów ciszy bądź innych mających zbyt małą energię lub charakter niskopoziomowego szumu wpływa na ekstrakcję tylko tych cech sygnału mowy, które determinują poprawne rozpoznanie na etapie testu.

Autor zbadał zmianę skuteczności rozpoznawania dla czterech algorytmów detekcji aktywności: opartego na energii sygnału, algorytmu Jang'a opartego na amplitudzie sygnału, algorytmu wykorzystującego różnice wyższych rzędów (Jang HOD, high-order differences), oraz bazującego na szybkości zmian znaku w reprezentacji cyfrowej sygnału (Jang ZCR, zero-crossing rate) [Marc2011]. Algorytmy te zostały przetestowane w dwóch konfiguracjach – znajdywanie ciszy tylko na początku i na końcu wypowiedzi, oraz w jej całości. Autor zbadał także czas analizy detekcji aktywności mówcy dla każdego z algorytmów tak, aby możliwe było zastosowanie go w systemie wbudowanym czasu rzeczywistego. Uzyskana poprawa skuteczności, mierzona z wykorzystaniem współczynnika EER (equal error rate), wynosi do 19 %. Wyniki w postaci wykresu FAR/FRR (false acceptance rate / false rejecton rate) przedstawione są na Rys. 3. Dotyczy on przypadku gdzie zastosowano rozszerzone przez autora algorytmy detekcji aktywności mówcy, obejmujące nie tylko początek i koniec zwrotu, lecz także jego środkową część, stąd w nazwie człon „middle". Największą poprawę skuteczności zanotowano dla szybkości próbkowania 8 kSps, co jest istotne z punktu widzenia przeznaczenia systemu.

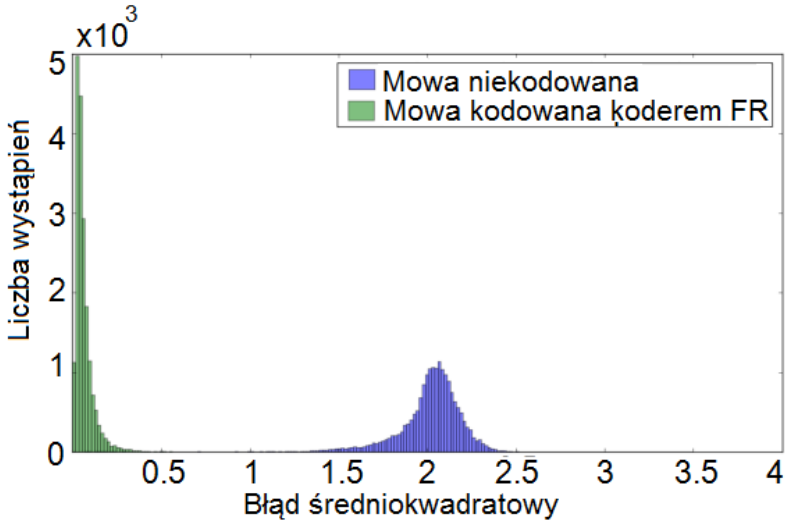



Rys. 3. Wykres FAR/FRR wpływu doboru algorytmu detekcji aktywności mówcy na skuteczność rozpoznawania.

## 4.4 Detekcja kodowania GSM w sygnale mowy oraz typu kodera GSM

Kolejne badania przeprowadzone przez autora dotyczą poprawy skuteczności rozpoznawania mówcy z kodowanych stratnie wypowiedzi. W pierwszej kolejności autor przeprowadził eksperymenty mające na celu sprawdzenie możliwości detekcji kodowania GSM w sygnale mowy. Autor rozprawy wykorzystał fakt, iż kolejne kodowania sygnału mowy wpływają na jego jakość coraz mniej, zatem różnica pomiędzy sygnałem niekodowanym a transkodowanym jednokrotnie będzie większa, niż pomiędzy transkodowanym jednokrotnie oraz dwukrotnie [Dabr2008]. Do badań wykorzystano bazę mówców TIMIT w wersji nieprzetworzonej oraz transkodowaną koderem GSM pracującym w trybie FR (full rate). Badania przeprowadzono dla sygnału mowy podzielonego na ramki o długości 125 ms, 250 ms, 500 ms, 1 s oraz 2 s [Weyc2010]. Im dłuższa ramka sygnału, tym część wspólna uzyskanych przedziałów błędów dla sygnałów kodowanych i niekodowanych jest coraz mniejsza, przy czym już dla ramki o długości 125 ms jedynie 4% ramek zostało niepoprawnie zakwalifikowanych. Dla ramki o długości 2 s błąd wyniósł 0.5 %. Ze względu na zastosowanie algorytmu do detekcji kodowania z krótkich wypowiedzi optymalnym

rozwiązaniem był wybór ramki 1-sekundowej z błędem rozpoznania 1.2 %. Rozkład błędów średniokwadratowych dla tego przypadku przedstawia Rys. 4.

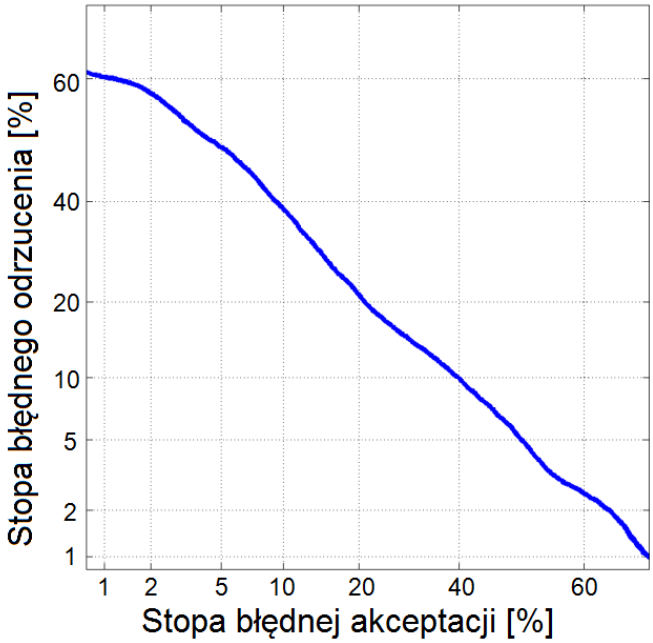



Rys. 4. Rozkład błędów średniokwadratowych dla sekwencji niekodowanej oraz kodowanej koderem GSM FR.

Wyniki badań jednoznacznie dowodzą zatem możliwości detekcji kodowania GSM w sygnale mowy. Autor sprawdził także czy możliwe jest także określenie rodzaju wykorzystanego kodera. Zbadano 4 typy koderów GSM: FR (full rate), EFR (enhanced full rate), HR (half rate) oraz AMR (adaptive multi-rate). Przeprowadzone zostały w tym celu eksperymenty pokazujące, w jaki sposób każdy z koderów wpływa na generowane współczynniki mel-cepstralne [Weyc2012] oraz na jakość mowy, przy czym do pomiaru jakości wykorzystano stosunek SNR (signal to noise ratio). Wyniki eksperymentów wykazały zróżnicowanie we wpływie koderów GSM na poszczególne współczynniki mel-cepstralne dowodząc iż możliwa jest detekcja typu kodera.

Autor zaproponował budowę modelu z wykorzystaniem algorytmu GMM-UBM. Model każdego z koderów składa się z czterech elementów związanych z każdym z koderów GSM, gdyż sygnał wejściowy musi zostać transkodowany każdym z czterech koderów GSM. Łącznie w bazie jest zatem 16 modeli, po 4 na każdy koder. Przykładowo, jeśli sygnał wejściowy kodowany był koderem FR, transkodowanie go czterema koderami tworzy zbiór rezultatów – FR-FR, FR-HR, FR-AMR, FR-EFR. Model kodera FR składa właśnie się z modeli dla transkodowania FR-FR, FR-HR, FR-AMR oraz FR-EFR. Jeśli sygnał testowany był kodowany koderem FR, natomiast najbliższym modelem okazał się nie model FR-FR a FR-

EFR, wynik porównania wskaże nadal poprawny koder - FR. Taka budowa modelu zwiększa prawdopodobieństwo poprawnej detekcji kodera. Rezultatem eksperymentów, w których wykorzystano bazę mówców TIMIT transkodowaną 16-krotnie w każdej możliwej konfiguracji koder-koder (4 kodery × 4 kodery), jest zrównoważona stopa błędów (EER) na poziomie 20 %. Otrzymana krzywa FAR/FRR prezentowana jest na rys. 5.




Rys. 5. Krzywa FAR/FRR detekcji typu kodera GSM.

## 4.4 Wpływ doboru modelu mówcy na skuteczność rozpoznawania z transkodowanych krótkich wypowiedzi

Dalsze badania dotyczące poprawy skuteczności rozpoznawania mówcy z kodowanych stratnie wypowiedzi związane były z kolejną propozycją autora rozprawy, polegającą na wykorzystaniu dedykowanej, rozszerzonej bazy mówców. W tej bazie każdy mówca posiada kilka modeli związanych bezpośrednio z rodzajem kodera, którym kodowany był sygnał mowy. W tym etapie wzięto pod uwagę zarówno kodery GSM, jak i kodery audio ogólnego przeznaczenia wykorzystywane do transmisji głosu przez internet – MPEG1 Layer 3 (MP3), OGG, WMA (Windows media audio) oraz kodery z rodziny AAC (advanced audio coding), w tym także najbardziej zaawansowane kodery wykorzystywane w technologii DAB (Digital audio broadcasting) do cyfrowej transmisji radia – HE-AAC v2 (high efficiency advanced

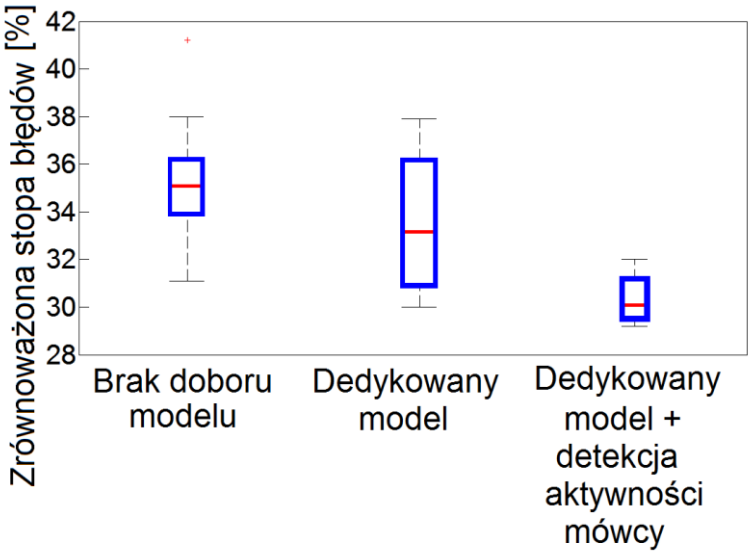
audio coding v2). Eksperymenty przeprowadzone zostały w warunkach dopasowania i niedopasowania koderów (matched- and mismatched conditions). Warunki dopasowania oznaczają, iż jeśli sygnał wejściowy kodowany był koderem A, model mówcy wykorzystany do rozpoznawania także wyznaczony został z sygnału kodowanego koderem A. Rezultat eksperymentów powinien być w tym przypadku lepszy niż dla warunków niedopasowania (sygnał wejściowy kodowany koderem A, model mówcy wyznaczony z sygnał kodowanego koderem B).

Do badań z wykorzystaniem koderów GSM użyto dwóch algorytmów detekcji aktywności mówcy – opartego na energii sygnału oraz różnicach wyższych rzędów (HOD). Eksperymenty przeprowadzono dla bazy mówców TIMIT oraz SPU. Modelowanie mówcy zrealizowano z wykorzystaniem algorytmu GMM. Rezultaty dla sygnałów niekodowanych były porównywalne dla algorytmów VQ oraz GMM, jednakże dla sygnałów kodowanych skuteczność rozpoznawania dla algorytmu VQ i dużych baz mówców była dużo niższa w porównaniu do algorytmu GMM.

Sygnał mowy transkodowany był koderami GSM 1-krotnie oraz 4-krotnie celem potwierdzenia poprawności przyjętej metodologii. Łącznie wykonanych zostało w tej części 280 eksperymentów. Zbadano także czas przetwarzania dla detekcji aktywności mówcy, ekstrakcji cech, generacji modelu oraz identyfikacji. Zwrócono uwagę na fakt, iż całkowity czas przetwarzania z wykorzystaniem algorytmu HOD jest do 3 razy dłuższy, stąd do dalszych eksperymentów wybrano metodę opartą na wyznaczeniu energii sygnału. Koniecznym przy tym zaznaczenia jest fakt, iż rezultaty rozpoznawania dla obu algorytmów VAD były bardzo zbliżone.

Wyniki wykazały jednoznacznie, iż w warunkach dopasowania zrównoważona stopa błędów EER zmniejsza się o 9 %, gdy nie zostały wykorzystane algorytmy usuwania ciszy, natomiast gdy sygnał mowy został przetworzony przez algorytmy detekcji aktywności mówcy – nawet o 15 % [Marc2012, Weyc2013].

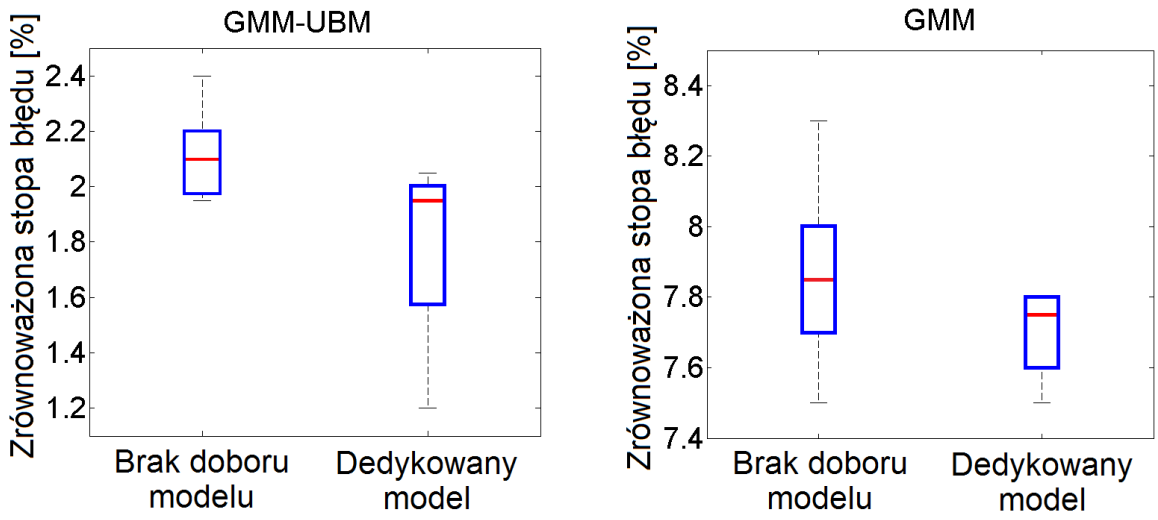
Zbiorcze zestawienie otrzymanych rezultatów zastosowania doboru modelu mówcy oraz detekcji aktywności mówcy przedstawione jest na Rys. 6. w postaci wykresu pudełkowego. Czerwona linia oznacza medianę, niebieski prostokąt obejmuje 50 % wyników (25 % poniżej i powyżej mediany), zaś poziome linie ograniczające od dołu i góry obejmują zakres 99,3 % wszystkich wyników. Zestawienie to potwierdza poprawność przyjętej tezy.




Rys. 6. Zestawienie rezultatów wpływu doboru modelu oraz detekcji aktywności mówcy na skuteczność rozpoznawania mówcy z mowy kodowanej koderami GSM.

W przypadku koderów ogólnego przeznaczenia metodologia przeprowadzenia eksperymentów była taka sama, z tym że sygnał mowy transkodowany był jednokrotnie. Jako algorytm detekcji aktywności mówcy wykorzystano metodę opartą na energii sygnału. Model mówcy wyznaczony został algorytmami GMM oraz GMM-UBM. Eksperymenty przeprowadzone zostały dla bazy mówców TIMIT.

Ze względu na fakt, iż kodery ogólnego przeznaczenia zapewniają dużo wyższą jakość sygnału niż kodery GSM (wyższa szybkość próbkowania, modele psychoakustyczne i inne), wynik błędnego rozpoznawania mierzony za pomocą zrównoważonej stopy błędów (EER) był dużo niższy, zwłaszcza dla algorytmu GMM-UBM. Ze względu na zastosowanie modelu psychoakustycznego, różnica między wynikami dla warunków dopasowania i niedopasowania koderów również była niższa. Eksperymenty wykazały jednoznacznie, iż w warunkach dopasowania oraz z wykorzystaniem algorytmu VAD skuteczność rozpoznawania mierzona zrównoważoną stopą błędów zwiększa się o maksymalnie 2 %, przy czym skuteczność rozpoznawania wynosi ok. 7 % dla algorytmu modelowania GMM, oraz 2 % dla algorytmu modelowania GMM-UBM. Łącznie na tym etapie, celem wykazania poprawności przyjętej metodologii, przeprowadzono 77 niezależnych eksperymentów z uwzględnieniem różnych wartości parametru bitrate koderów. Zbiorcze zestawienie otrzymanych wyników prezentuje Rys. 7.




Rys. 7. Zestawienie rezultatów wpływu doboru na skuteczność rozpoznawania mówcy z mowy kodowanej koderami ogólnego przeznaczenia.

## 4.5 Realizacja na zmiennoprzecinkowym procesorze ARM oraz stałoprzecinkowym procesorze sygnałowym TMS320C5515

Istotną częścią rozprawy były badania realizacji rozpoznawania mówcy za pomocą systemu wbudowanego działającego w czasie rzeczywistym. W pierwszej kolejności opracowane zostało oprogramowanie na komputerze stacjonarnym, umożliwiające rozpoznawanie mówcy w czasie rzeczywistym z audycji radiowych. Oprogramowanie wraz z graficznym interfejsem użytkownika, zostało przygotowane w środowisku Matlab [Weyc2015]. Wykorzystano algorytm modelowania GMM oraz detekcję aktywności mówcy opartą o pomiar energii sygnału. Oprogramowanie zostało wykorzystane do analizy nagrań debat radiowych, dla których dokonywano pomiaru czasu wypowiedzi każdego z uczestników debaty. Audycja była transmitowana przez sieć Internet i kodowana koderem MP3 o przepływności 128 kbps. Jest to zdaniem autora bardzo ciekawe zastosowanie proponowanego systemu. Do przedstawienia zróżnicowania modeli mówców wykorzystano algorytm ISOMAP [Tene2000].

Na podstawie tej implementacji opracowany został system rozpoznawania czasu rzeczywistego w języku Python, co umożliwiło zaprojektowanie systemu wbudowanego [Weych2015], opartego o zmiennoprzecinkowy mikroprocesor w architekturze ARM (Advanced RISC Machine), wykorzystywany w wielu

urządzeniach multimedialnych. Główną zaletą tych procesorów jest zebranie cech procesorów ogólnego przeznaczenia wraz ze wsparciem dla technik przetwarzania sygnałów, będącą domeną procesorów DSP. Analiza zapotrzebowania energetycznego, mocy obliczeniowej oraz dostępności wskazała na procesory z rodziny Cortex-A5, A7, A8 oraz A9. Ze względu na relatywnie niskie zapotrzebowanie energetyczne oraz zdecydowanie najlepszą wydajność w stosunku do poboru energii wybrany został procesor czterordzeniowy Cortex-A7, dostępny w m.in. platformie Raspberry-PI 2. Dostępność systemu operacyjnego opartego na systemie Linux umożliwiła w tym przypadku wykorzystanie języka Python i jego bibliotek zorientowanych na przetwarzanie sygnałów.

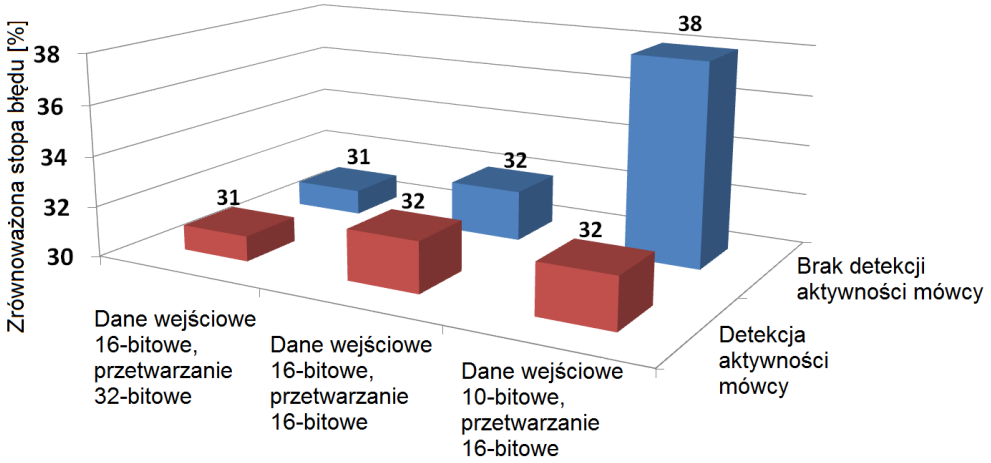
Opracowane oprogramowanie rozpoznawania mówcy (wykorzystano modelowanie algorytmem GMM) przetestowane zostało pod względem szybkości przetwarzania w funkcji szybkości próbkowania, rozdzielczości szybkiej transformaty Fouriera (FFT) oraz liczby mieszanin Gaussa dla algorytmu GMM. Przeprowadzone zostały badania, w których dobrano optymalne parametry przetwarzania sygnału audio, przy czym kryterium optymalności zdefiniowano jako funkcję szybkości przetwarzania oraz skuteczności rozpoznawania. W wyniku doboru parametrów uzyskano skuteczność rozpoznawania dla niekodowanych sygnałów 4.1 % EER, gdzie szybkość próbkowania ustalono na 16 kSps, 256 prążków FFT oraz liczbę mieszanin Gaussa równą 32. Dla tak zdefiniowanych parametrów szybkość przetwarzania wyniosła 144 ms dla 1-sekundowej ramki sygnału, przy czym zajętość obliczeniowa procesora nie przekraczała 10 %. Wartym podkreślenia jest fakt, iż parametry czasowe uzyskano dla sygnału kodowanego koderami ogólnego przeznaczenia.

Dalsze, bardziej istotne eksperymenty związane były z implementacją czasu rzeczywistego opartą o stałoprzecinkowy procesor sygnałowy Texas Instruments C5515. Procesor ten został również wybrany w wyniku analizy dostępnych rozwiązań. Pod uwagę zostały wzięte następujące parametry:

- Wydajność [MIPS]
- Zużycie energii
- Wsparcie dla przetwarzania audio (koprocesor FFT, przetwornik A/C)
- Podstawowe operacje DSP
- Architektura minimalizująca błędy przetwarzania stałoprzecinkowego

Jak wspomniano wcześniej, arytmetyka stałoprzecinkowa powoduje generowanie większych błędów i zmniejsza skuteczność rozpoznawania mówcy. Jednym z argumentów decydujących o wykorzystaniu procesora C5515 jest sprzętowe wsparcie dla FFT (Fast Fourier Transform) niezbędnej przy wyznaczaniu cech mówcy, oraz 40-bitowy akumulator minimalizujący błędy obliczeń stałoprzecinkowych. W przeprowadzonych eksperymentach sprawdzony został wpływ reprezentacji stałoprzecinkowej na skuteczność rozpoznawania mówcy dla algorytmów modelowania VQ oraz GMM. Dodatkowo przeanalizowany został wpływ rozdzielczości kwantyzacji wejściowego sygnału audio, oraz rozdzielczości przetwarzanych danych. Wykorzystanie algorytmów usuwania ciszy umożliwia redukcję rozdzielczości przetwarzania do 16 bitów i rozdzielczości kwantyzacji sygnału wejściowego nawet do 10 bitów bez znaczącego zmniejszenia skuteczności rozpoznawania [Weych2013, Marc2014]. Wyniki eksperymentów przedstawione są na rys. 8.

Możliwe jest więc zastosowanie wewnętrznych 10-bitowych przetworników analogowo-cyfrowych w procesorach sygnałowych, bez konieczności wykorzystania zewnętrznych przetworników, co prowadzi do dodatkowego zmniejszenia zużycia energii i minimalizacji kosztów systemu wbudowanego.




Rys. 8. Skuteczność rozpoznawania mówcy w systemie stałoprzecinkowym z uwzględnieniem techniki detekcji aktywności mówcy oraz obniżoną rozdzielczością akwizycji i przetwarzania sygnału.

## 5. Wnioski

W rozprawie zaproponowano nowatorskie metody poprawy skuteczności algorytmów rozpoznawania mówcy w zadaniach zdalnego sterowania systemami automatyki. Czynniki obniżające skuteczność dotyczyły niewystarczającej ilości danych do modelowania mówcy ze względu na krótką (<5 s) długość wypowiedzi (zwrotów sterujących), oraz zmniejszenia zawartości informacyjnej sygnału w wyniku kodowania stratnego koderami GSM oraz koderami ogólnego przeznaczenia, przy transmisji sygnału mowy w sieci GSM oraz sieci internet. Zaproponowane w rozprawie rozwiązania i przeprowadzone eksperymenty pokazują, iż możliwe jest zwiększenie stopy poprawnej detekcji w systemach rozpoznawania mówcy z krótkich, kodowanych wypowiedzi.

Przedstawione ulepszenia typowego systemu rozpoznawania mówcy o:

- algorytmy detekcji aktywności mówcy
- detekcję kodowania GSM w sygnale mowy
- detekcję typu kodera
- dobór modelu mówcy na podstawie typu kodera

pozwoliły na zwiększenie skuteczności rozpoznawania mówcy z kodowanych, krótkich wypowiedzi o:

- 15% dla kodowania GSM
- 2% dla koderów ogólnego przeznaczenia

Zastosowane rozwiązania przy implementacji stałoprzecinkowej pozwoliły na otrzymanie tej samej skuteczności (30% dla kodowania GSM) przy jednoczesnej redukcji rozdzielczości akwizycji (z 16 do 10 bitów) oraz przetwarzania (z 32 do 16 bitów). Wykazane zostało iż możliwe jest wykorzystanie wbudowanych w procesory przetworników A/C o rozdzielczości 10 bitów.

## 6. Bibliografia


[Beig2011] H. Beigi, *“Fundamentals of speaker recognition”*. Springer Science & Business Media”, 2011.

[Brea2013] C. Breazeal, N. DePalma, J. Orkin, S. Chernova, i M. Jung, *“Crowd sourcing human-robot interaction: New methods and system evaluation in a public environment,"* Journal of Human-Robot Interaction, vol. 2, no. 1, s. 82-111, 2013.

[Dabr2008] A. Dąbrowski, S. Drgas, i T. Marciniak, *“Detection of GSM speech coding for telephone call classification and automatic speaker recognition,"* ICSES '08. International Conference on Signals and Electronic Systems, s. 415-418, Sept. 2008.

[Deby2010] M. Debyeche, A. Krobba, i A. Amrouche, *“Effect of GSM speech coding on the performance of speaker recognition system,"* Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference, s. 137-140, 2010.

[Dunn2001] R. B. Dunn, T. F. Quatieri, D. A. Reynolds, i J. P. Campbell, *“Speaker recognition from coded speech in matched and mismatched conditions,"* A Speaker Odyssey-The Speaker Recognition Workshop, 2001.

[Gian2005] T. Giannakopoulos, N.-A. Tatlas, T. Ganchev, i I. Potamitis, *“A practical, real-time speech-driven home automation front-end,"* Consumer Electronics, IEEE Transactions, vol. 51, no. 2, s. 514-523, May 2005.

[Grafo1993] J. S. Garofolo and et al., *“Timit acoustic-phonetic continuous speech corpus,"* 1993, linguistic Data Consortium, Philadelphia, http://catalog.ldc.upenn.edu/LDC93S1.

[Gras2000] S. Grassi, L. Besacier, A. Dufaux, M. Ansorge, i F. Pellandini*, “Influence of GSM speech coding on the performance of text-independent speaker recognition,"* Tampere, Finland, September 4-8 2000, s. 437-440.

[Jani2011] A. Janicki i T. Staroszczyk, “*Speaker recognition from coded speech using support vector machines,"* Text, Speech and Dialogue. Springer, 2011, s. 291-298.

[Jawa2007] N. Jawarkar, V. Ahmed, i R. Thakare, *“Remote control using mobile through spoken commands,"* Signal Processing, Communications and

Networking, 2007. ICSCN '07. International Conference on, Feb 2007, s. 622-625.

[Jian2000] H. Jiang, Z. Han, P. Scucces, S. Robidoux, i Y. Sun, *"Voice-activated environmental control system for persons with disabilities,"* in Bioengineering Conference, 2000. Proceedings of the IEEE 26th Annual Northeast, 2000, s. 167-168.

[Lill1996] B. Lilly i K. Paliwal, *"Effect of speech coders on speech recognition performance,"* vol. 4, Oct 3-6 1996, s. 2344-2347.

[Lind1980] Y. Linde, A. Buzo, i R. Gray, *"An algorithm for vector quantizer design,"* Communications, IEEE Transactions on, vol. 28, no. 1, s. 84-95, Jan 1980.

[Marc2010] T. Marciniak, R. Weychan, A. Dąbrowski, i A. Krzykowska, *"Speaker recognition based on short Polish sequences,"* IEEE SPA: Signal Processing Algorithms, Architectures, Arrangements, and Applications Conference Proceedings, s. 95-98, 2010.

[Marc2011] T. Marciniak, R. Weychan, A. Dąbrowski, i A. Krzykowska, *"Influence of silence removal on speaker recognition based on short Polish sequences,"* IEEE SPA: Signal Processing Algorithms, Architectures, Arrangements, and Applications Conference Proceedings, s. 159-163, 2011.

[Marc2012] T. Marciniak, A. Krzykowska, i R. Weychan, *"Speaker recognition based on telephone quality short Polish sequences with removed silence,"* Przegląd Elektrotechniczny, no. 06/2012, s. 42-46, 2012.

[Marc2014] T. Marciniak, R. Weychan, A. Stankiewicz, i A. Dąbrowski, *"Biometric speech signal processing in a system with digital signal processor,"* Bulletin of the Polish Academy of Sciences. Technical Sciences, vol. Vol. 62, nr 3, s. 589-594, 2014.

[McLa2013] M. McLaren, V. Abrash, M. Graciarena, Y. Lei, i J. Pesan, *"Improving robustness to compressed speech in speaker recognition."* INTERSPEECH, 2013, s. 3698-3702.

[Mola2001] S. Molau, M. Pitz, R. Schluter, i H. Ney, *"Computing Mel-frequency cepstral coefficients on the power spectrum,"* Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference, vol. 1, 2001, s. 73-76.

[Nema2010] S. Nemat i A. Kader, *“Effect of GSM system on text independent speaker recognition,"* Journal of Theoretical and Applied Information Technology, s. 442-449, June 2010.

[Pire2007] J. N. Pires, *“Industrial robots programming: building applications for the factories of the future”.* Springer Science & Business Media, 2007.

[Reyn1995] D. Reynolds, *“Robust text-independent speaker identification using Gaussian mixture speaker models,"* IEEE Trans. Speech Audio Proc., vol. 3, no. 1, s. 72-83, 1995.

[Reyn2000] D. A. Reynolds, T. F. Quatieri, i R. B. Dunn, *“Speaker verification using adapted Gaussian mixture models,"* Digital signal processing, vol. 10, no. 1, s. 19-41, 2000.

[Rogo2012] A. Rogowski, *“Analiza i synteza systemów sterowania głosowego w zautomatyzowanym wytwarzaniu”.* Oficyna wydawnicza Politechniki Warszawskiej, 2012.

[Saue2006] P. Sauer, W. Waliszewski, M. Michalski, D. Pazderski, i P. Jeziorek, *“Asystent-control system assisting surgeon in laparoscopy,"* Biocybernetics and Biomedical Engineering, vol. 26, no. 4, s. 55-70, 2006.

[Tade1988] R. Tadeusiewicz, *„Sygnał mowy”,* WKiŁ. 1988, rozdział 5. „Sygnał mowy w automatyce”

[Tene2000] J. B. Tenenbaum, V. D. Silva, i J. C. Langford, *“A global geometric framework for nonlinear dimensionality reduction,"* Science, vol. 290, no. 5500, s. 2319-2323, 2000.

[Wang2016] N. Wang, “*Robust speaker recognition based on multi-stream features",* Proceedings of ICCE-China, 2016.

[Weyc2010] R. Weychan, T. Marciniak, i A. Dąbrowski, *“Influence of signal segmentation in GSM coding detection,"* Elektronika, no. 5, s. 94-98, 2010.

[Weyc2012] R. Weychan, T. Marciniak, i A. Dąbrowski, *“Analysis of differences between MFCC after multiple GSM transcodings,"* Przeglad Elektrotechniczny, no. 6/2012, s. 24-29, 2012.

[Weyc2013] R. Weychan, A. Stankiewicz, T. Marciniak, i A. Dąbrowski, *“Improving of speaker identification from mobile telephone calls,"* Multimedia

Communications, Services and Security, seria Communications in Computer and Information Science, 2014, vol. 429, s. 254-264.

[Weych2013] R.Weychan, A. Stankiewicz, T. Marciniak, i A. Dąbrowski, *„Analysis of the impact of data resolution on the speaker recognition effectiveness in embedded fixed-point systems,"* IEEE SPA: Signal Processing Algorithms, Architectures, Arrangements, and Applications Conference Proceedings, s. 327-331, 2013.

[Weyc2015] R. Weychan, T. Marciniak, A. Stankiewicz, i A. Dąbrowski, *"Real time recognition of speakers from internet audio stream,"* Foundations of Computing and Decision Sciences, vol. 40, no. 3, s. 223-233, 2015.

[Weych2015] R. Weychan, T. Marciniak, i A. Dąbrowski, *"Implementation aspects of speaker recognition using python language and raspberry pi platform,"* IEEE SPA: Signal Processing Algorithms, Architectures, Arrangements, and Applications Conference Proceedings, s. 162-167, 2015.

[Yuks2006] B. Yuksekkaya, A. Kayalar, M. Tosun, M. Ozcan, i A. Alkar, *"A GSM, Internet and speech controlled wireless interactive home automation system,"* Consumer Electronics, IEEE Transactions on, vol. 52, no. 3, s. 837-843, Aug 2006.