Wojskowa Akademia Techniczna

Wydział Elektroniki

ROZPRAWA DOKTORSKA

SYSTEM AUTOMATYCZNEGO ROZPOZNAWANIA MÓWCY WSPARTY BEHAWIORALNYMI CECHAMI SYGNAŁU MOWY

por. mgr inż. Dominik Mały

Promotor

prof. dr hab. inż. Andrzej P. Dobrowolski

Promotor pomocniczy

mjr dr inż. Kamil Kamiński

WARSZAWA 2025

Niniejszą rozprawę dedykuję mojej ukochanej Żonie Ewelinie,
w podziękowaniu za ogromne wsparcie, wyrozumiałość i cierpliwość w czasie jej powstawania.

Spis treści

Spis treści **3**

Indeks skrótów **6**

Streszczenie **8**

Abstract **9**

Przedmowa **10**

1. Wprowadzenie **13**

1.1. Model generacji sygnału mowy 15
1.2. Klasyfikacja systemów automatycznego rozpoznawania mówcy 17
1.2.1. Identyfikacja i weryfikacja mówcy 19
1.2.2. Treść wypowiedzi w systemach rozpoznawania mówcy 20
1.2.3. Cechy sygnału mowy 20
1.3. Krytyczny przegląd literatury 21
1.4. Podsumowanie 30

2. Proces przetwarzania sygnału mowy w aspekcie automatycznego rozpoznawania mówcy **31**

2.1. Wstępne przetwarzanie mowy 32
2.1.1. Proces rejestracji i konwersji sygnału mowy 32
2.1.2. Przetwarzanie analogowo-cyfrowe 33
2.1.3. Standaryzacja sygnału mowy 35
2.1.4. Wycinanie ciszy 35
2.1.5. Segmentacja sygnału mowy 36
2.2. Cechy zawarte w sygnale mowy 37
2.2.1. Geneza pochodzenia cech fizycznych 38
2.2.2. Przykłady cech fizycznych 39
2.2.3. Geneza pochodzenia cech behawioralnych 40
2.2.4. Przykłady cech behawioralnych 41
2.3. Potencjalne możliwości systemów opartych na cechach fizycznych i behawioralnych 42
2.4. Baza głosów zastosowana do badań – SLR-12 44
2.5. Podsumowanie - Uzasadnienie tezy rozprawy 45

3. System odniesienia oparty na cechach fizycznych **46**

3.1. Wstępne przetwarzanie sygnału mowy 49
3.1.1. Normalizacja sygnału 49
3.1.2. Wycinanie ciszy i filtracja sygnału 49
3.1.3. Segmentacja sygnału i selekcja ramek 50
3.2. Metody parametryzacji sygnału mowy 52
3.2.1. Analiza widmowa 52
3.2.2. Analiza cepstralna 53

3.2.3. Analiza melcepstralna 55
3.3. Cechy fizyczne wykorzystane w systemie 56
3.4. Modele mieszanin gaussowskich jako metoda klasyfikacji 58
3.5. Podsumowanie 60

4. System oparty na cechach behawioralnych 61

4.1. Wstępne przetwarzanie sygnału mowy na potrzeby systemu ASR 62
4.2. Definicje i przykłady wykorzystanych w systemie cech behawioralnych 63
4.2.1. Całkowity czas mowy w analizowanej ramce oraz unormowany czas trwania ramki mowy 64
4.2.2. Skala głosu 64
4.2.3. Procent ramek o niskiej energii 64
4.2.4. Liczba ramek dźwięcznych 64
4.2.5. Tempo przejść przez zero 65
4.2.6. Liczba znaczących zjawisk impulsowych 65
4.2.7. Wartość średnia amplitudy w oktawach i tercjach 67
4.2.8. Pasmo częstotliwości zawierające 95% całkowitej mocy sygnału 69
4.2.9. Miara płaskości widma 70
4.2.10. Środek ciężkości widma i jego rozproszenie 70
4.3. Metody ewolucyjne w zastosowaniu do selekcji zestawu cech 72
4.4. Zastosowany klasyfikator 75
4.4.1. Metoda najbliższej średniej w zadaniu klasyfikacji 77
4.5. Podsumowanie 78

5. Fuzja systemów 80

5.1. Analiza rodzajów wykorzystywanych fuzji danych 81
5.2. Metody fuzji danych wsparte technikami sztucznej inteligencji 84
5.2.1. Bagging, boosting i stacking 85
5.3. Metoda preselekcji najbliższej liczby sąsiadów – fuzja na poziomie decyzji 88
5.3.1. Wyniki działania fuzji na poziomie decyzji 89
5.4. Metoda połączenia zestawów cech – fuzja na poziomie cech 90
5.4.1. Wyniki działania fuzji na poziomie cech 91
5.5. Podsumowanie 93

6. Badania eksploatacyjne systemu ASR 95

6.1. Bazy głosów rozszerzające podstawowy zbiór SLR-12 96
6.1.1. Baza głosów NIST 2002 SRE 96
6.1.2. Baza stworzona na podstawie audiobooków 96
6.1.3. Baza głosów autorstwa dr inż. Eweliny Majdy-Zdancewicz 96
6.1.4. Baza głosów autorstwa mgr. inż. Daniela Posiadały 97
6.1.5. Baza głosów autorstwa mgr. inż. Michała Bojszy 97
6.2. Sposób prezentacji wyników testów 98
6.3. Testy opracowanych rozwiązań w warunkach bezszumowych 99
6.4. Testy opracowanych rozwiązań w zróżnicowanych warunkach szumowych 99
6.4.1. Badania opracowanych rozwiązań dla zakłóceń w postaci szumu białego 101
6.4.2. Badania opracowanych rozwiązań dla zakłóceń w postaci nagrania tłumu 103
6.4.3. Badania opracowanych rozwiązań dla zakłóceń w postaci nagrania ulewnego deszczu 105
6.5. Testy opracowanych rozwiązań w warunkach zróżnicowanej jakości nagrań 106
6.6. Testy opracowanych rozwiązań przy wykorzystaniu zróżnicowanych torów akustycznych 108

6.7. Oszczędność obliczeniowa 110
6.8. Podsumowanie 111

7. Konkluzja 112

Literatura 114

Załączniki:

Z. 1. Zbiór definicji wszystkich cech behawioralnych występujących w systemie ASR

Z. 2. Dokumentacja badawcza ewolucji projektowanego systemu ASR

Z. 3. Pełne wyniki badań opracowanych rozwiązań fuzji danych

Z. 4. Opis aplikacji

Indeks skrótów

AANN (ang. *Auto-Associative Neural Network*) – autoasocjacyjna sieć neuronowa

ANN (ang. *Artificial Neural Network*) – sztuczna sieć neuronowa

ASR (ang. *Automatic Speaker Recognition*) – automatyczne rozpoznawanie mówcy

CMN (ang. *Cepstral Mean Normalization*) – cepstralna normalizacja średniej

DCF (ang. *Decision Cost Function*) – koszt funkcji decyzyjnej

dCNN (ang. *deep Convolutional Neural Network*) – głęboka konwolucyjna sieć neuronowa

DCT (ang. *Discrete Cosine Transform*) – dyskretna transformacja kosinusowa

DFT (ang. *Discrete Fourier Transform*) – dyskretna transformacja Fouriera

DNN (ang. *Deep Neural Network*) – głęboka sieć neuronowa

DTW (ang. *Dynamic Time Warping*) – dynamiczne dopasowanie czasowe

DWT (ang. *Discrete Wavelet Transform*) – dyskretna transformacja falkowa

EM (ang. *Expectation Maximization*) – maksymalizacja oczekiwań

ERR (ang. *Equal Error Rate*) – stopa błędu zrównoważonego

FMMNN (ang. *Fuzzy Min–Max Neural Network*) – rozmyta sieć min–max

FSVM (ang. *Fuzzy Support Vector Machine*) – rozmyta sieć wektorów nośnych

FT (ang. *Fourier Transform*) – transformata Fouriera

FFT (ang. *Fast Fourier Transform*) – szybka transformata Fouriera

GMM (ang. *Gaussian Mixture Model*) – model mieszanin Gaussowskich

HLDA (ang. *Heteroscedastic Linear Discriminant Analysis*) – heteroskedastyczna liniowa analiza dyskryminacyjna

HMM (ang. *Hidden Markov Model*) – ukryty model Markowa

IR (ang. *Identification Rate*) – współczynnik identyfikacji

JFA (ang. *Joint Factor Analysis*) – zespolona analiza czynnikowa

LP (ang. *Linear Prediction*) – predykcja liniowa

LPCC (ang. *Linear Prediction Cepstral Coefficients*) – cepstralne współczynniki predykcji liniowej

MAP (ang. *Maximum a posteriori*) – maksimum a posteriori

MD (ang. *Minimum Divergence*) – minimalna rozbieżność

MFCC (ang. *Mel–Frequency Cepstral Coefficients*) – melowe współczynniki cepstralne

ML (ang. *Maximum Likelihood*) – maksymalna wiarygodność

MLP (ang. *Multi–Layer Perceptron*) – wielowarstwowa perceptronowa sieć neuronowa

NDSF (ang. *Normalized Dynamic Spectral Features*) – znormalizowane, dynamiczne cechy widmowe

NIST (ang. *National Institute of Standards and Technology*) – Narodowy Instytut Standaryzacji i Technologii Stanów Zjednoczonych

PSW (ang. *Pitch Synchronous Windowing*) – okienkowanie zsynchronizowane ze skokiem

ResNet (ang. *Residual Neural Network*) – resztkowa sieć neuronowa

SFM (ang. *Spectral Flatness Measure*) – miara płaskości widma

SRE (ang. *Speaker Recognition Evaluation*) – ocena rozpoznawania mówcy

SVM (ang. *Support Vector Machine*) – sieć neuronowa wektorów nośnych

SWT (ang. *Stationary Wavelet Transform*) – stacjonarna transformacja falkowa

TDCT (ang. *Temporal Discrete Cosine Transform*) – czasowa dyskretna transformacja kosinusowa

TDNN (ang. *Time Delay Neural Network*) – sieć neuronowa z opóźnieniem czasowym

TTESBCC (ang. *Temporal Teager Energy Based Subband Cepstral Coefficients*) – czasowe podpasmowe energetyczne współczynniki cepstralne Teagera

VQ (ang. *Vector Quantization*) – kwantyzacja wektorów

WOCOR (ang. *Wavelet Octave Coefficients of Residues*) – współczynniki falkowe reszt oktawowych

Wojskowa Akademia Techniczna

Wydział Elektroniki

Streszczenie

SYSTEM AUTOMATYCZNEGO ROZPOZNAWANIA MÓWCY WSPARTY BEHAWIORALNYMI CECHAMI SYGNAŁU MOWY

Autor: por. mgr inż. Dominik Mały

Promotor: prof. dr hab. inż. Andrzej P. Dobrowolski

Promotor pomocniczy: mjr dr inż. Kamil Kamiński

Celem niniejszej rozprawy było opracowanie i implementacja w istniejącym systemie ASR zbioru cech behawioralnych umożliwiających zwiększenie liczby poprawnych identyfikacji tożsamości mówców, w szczególności w utrudnionych warunkach rejestracji.

W skład rozprawy wchodzą trzy części obrazujące kolejno realizowane etapy badań. Część pierwsza (rozdział 1) to wprowadzenie w tematykę biometrii głosu, począwszy od opisu procesu generacji ludzkiego głosu, poprzez ogólną charakterystykę struktury systemów automatycznego rozpoznawania mówcy oraz ich klasyfikację, na przeglądzie literatury kończąc. Efektem tego etapu badań było określenie dalszego kierunku badań.

W części drugiej (rozdziały 2-4) przedstawiono powszechnie wykorzystywane architektury systemów ASR oraz implementowane w nich algorytmy. Następnie przedstawiono kolejne etapy analizy sygnału mowy od momentu rejestracji do procesu segmentacji. Największą uwagę poświęcono aspektom związanym z genezą pochodzenia cech możliwych do wyekstrahowania z sygnału mowy oraz sposobom ich ekstrakcji. Zagadnienia te zostały dokładnie scharakteryzowane w kontekście właściwości powstałego w WAT systemu *ARMiA* [60], opartego na cechach fizycznych oraz autorskiego algorytmu automatycznego rozpoznawania mówcy opartego na cechach behawioralnych.

W ostatniej części rozprawy (rozdziały 5-7) zamieszczono wnioski płynące z badań prowadzonych w trakcie projektowania i optymalizacji rozwiązania integracji systemu *ARMiA* [60] z autorskim systemem opartym na cechach behawioralnych. Pokazano sposób implementacji fuzji systemów, a następnie zaprezentowano wyniki testów tego rozwiązania w warunkach istnienia zewnętrznych zakłóceń oraz zastosowania różnych torów akustycznych służących do rejestracji i transmisji sygnału mowy. Otrzymane rezultaty potwierdzają postawioną w rozprawie tezę, tzn. potwierdzają, że ***możliwe jest zdefiniowanie i wyekstrahowanie z sygnału mowy takich cech behawioralnych, które w połączeniu z cechami fizycznymi zaimplementowanymi w istniejącym rozwiązaniu znacząco poprawią jego skuteczność w obecności zewnętrznych zakłóceń***.

Słowa kluczowe: rozpoznawanie mówcy, cechy behawioralne, fuzja danych, algorytm genetyczny, selekcja cech dystynktywnych

Dominik Mały

Warszawa, 01.04.2025 r.

Military University of Technology

Faculty of Electronics

Abstract

AUTOMATIC SPEAKER RECOGNITION SYSTEM SUPPORTED BY BEHAVIORAL FEATURES OF SPEECH SIGNAL

Author: Lt Dominik Mały MSc Eng.

PhD student supervisor: Prof. Andrzej P. Dobrowolski, Ph.D., D.Sc., Eng.

Assistant supervisor: Maj. Kamil Kamiński, Ph.D., Eng.

The aim of this dissertation was to develop and implement a set of behavioral features within an existing ASR system to enable an increase in the number of accurate speaker identity identifications, particularly under challenging recording conditions. The dissertation comprises three parts, each illustrating successive stages of the research.

The first part (Chapter 1) introduces the field of voice biometrics, beginning with a description of the human voice generation process, followed by an overview of the structure and classification of automatic speaker recognition systems, and concluding with a literature review. The outcome of this research stage was the determination of the subsequent research direction.

The second part (Chapters 2-4) presents commonly used ASR system architectures, and the algorithms implemented within them. Subsequently, it outlines the successive stages of speech signal analysis from recording to segmentation. The primary focus is placed on aspects related to the origin of features that can be extracted from the speech signal and the methods for their extraction. These issues are thoroughly characterized in the context of the properties of the *ARMiA* system [60] developed at MUT, based on physical features, as well as an original algorithm for automatic speaker recognition based on behavioral features.

The final part of the dissertation (Chapters 5-7) presents conclusions drawn from the research conducted during the design and optimization of the integration of the *ARMiA* system [60] with a proprietary system based on behavioral features. The implementation of system fusion is demonstrated, followed by test results of this solution under conditions of external interference and the use of various acoustic pathways for voice recording. The obtained results confirm the hypothesis proposed in the dissertation, demonstrating that **it is possible to define and extract behavioral features from the speech signal which, when combined with physical features implemented in the existing solution, significantly improve its effectiveness in the presence of external disturbances**.

Key words: speaker recognition, behavioral features, data fusion, genetic algorithm, distinct feature selection

Dominik Mały

Warsaw, 01.04.2025.

Przedmowa

Wytwarzanie, przetwarzanie oraz analiza sygnału mowy jest dla każdego człowieka codziennością. To co subiektywnie i bez głębszego zastanowienia oceniamy jako pospolite i banalne dopiero po dokładniejszej analizie okazuje się procesem niezwykle skomplikowanym. Próba zastosowania tych operacji w obszarze techniki ukazuje w całej okazałości złożoną naturę powstawania dźwięków i mechanizmów słyszenia. System artykulacyjny człowieka to doskonały „instrument" wykształcony w długim procesie ewolucji, co powoduje, iż powszechnie uważa się, że artykulacja sama w sobie to łatwy i naturalny proces. Trakt głosowy jest tak skonstruowany, że przy współpracy narządów mowy (warg, języka oraz fałd głosowych) powstający dźwięk jest zbiorem różnorodnych informacji. Wprawny detektor jakim jest ludzkie ucho jest w stanie znaczną ich część „wyłowić" [107]. Odpowiednia interpretacja mowy przez ludzki umysł jest na porządku dziennym i dopiero próba przeniesienia możliwości analizy na komputer ukazuje wcześniej wspomnianą skalę złożoności.

Możliwość wydobycia i interpretacji danych osobniczych zawartych w sygnale mowy (poprzez przetwarzanie go za pomocą różnorodnych rozwiązań technicznych, takich jak urządzenia i oprogramowanie) jest jedną z możliwości współczesnej techniki informacyjnej, które określa się mianem systemów *automatycznego rozpoznawania mówcy – ASR* (ang. *Automatic Speaker Recognition*). Dźwięk generowany podczas artykulacji niesie rozmaite informacje, takie jak *język, dialekt* czy też *emocje*, które towarzyszą mówcy [60], a wspomniana wcześniej analiza pozwala na ekstrakcję m.in. *behawioralnych cech dystynktywnych* mówcy, na których skoncentrowana jest niniejsza rozprawa.

W dzisiejszych czasach potrzeba stosowania systemów ASR staje się coraz większa. Powodem tego jest nieustanny wzrost poziomu interakcji między człowiekiem i komputerem. Dzięki coraz częstszemu stosowaniu technik biometrycznych w procesie autoryzacji w systemach wymagających wysokiego stopnia bezpieczeństwa, systemy ASR znajdują zastosowanie w branży zarówno wojskowej, jak i cywilnej. Głos jako jedna z unikatowych cech dystynktywnych każdego człowieka, pozwala na identyfikację osoby bez potrzeby posiadania dodatkowych atrybutów, które mogą ulec zniszczeniu lub zgubieniu. Stosunkowo łatwy dla człowieka proces rozpoznawania osób przy pomocy narządu słuchu, dla maszyny staje się złożonym obliczeniowo wyzwaniem. Choć w ostatnich latach dokonano wielkiego postępu w dziedzinie badań nad automatycznym rozpoznawaniem mówcy, pozostało jeszcze wiele problemów czekających na lepsze rozwiązanie.

Niniejsza rozprawa doktorska prezentuje opracowany przez autora skuteczny system wspomagający rozpoznawanie mówcy w oparciu o behawioralne cechy dystynktywne głosu wyekstrahowane z wykorzystaniem analizy czasowej, częstotliwościowej, czasowo-częstotliwościowej i cepstralnej sygnału mowy. Autor skoncentrował się na dostosowaniu systemu do warunków rzeczywistych i uodpornieniu go na wpływy zróżnicowanych torów akustycznych, maseczek, przeziębień itp. Z przeprowadzonych przez autora wstępnych badań w zakresie tematyki pracy doktorskiej wyniknęła

następująca teza rozprawy: **„możliwe jest zdefiniowanie i wyekstrahowanie z sygnału mowy takich cech behawioralnych, które w połączeniu z cechami fizycznymi zaimplementowanymi w istniejącym rozwiązaniu znacząco poprawią jego skuteczność w obecności zewnętrznych zakłóceń".**

Wykazanie słuszności postawionej tezy rozprawy, opiera się na wykorzystaniu behawioralnych charakterystyk głosu, które stanowić będą wsparcie dla istniejącego już systemu automatycznego rozpoznawania mówcy opartego na cechach fizycznych. Cechy behawioralne powinny charakteryzować się wysoką zawartością informacji dystynktywnej oraz dużą odpornością na zakłócenia zewnętrzne. Całość zrealizowana została w formie rozwiązania fuzji danych, a w szczególności poprzez integracje danych na poziomie cech i decyzji pochodzących z dwóch osobnych systemów ASR.

W skład rozprawy wchodzą trzy części obrazujące kolejno realizowane etapy badań. W części pierwszej (rozdział 1) przedstawiono wprowadzenie w tematykę biometrii głosu, począwszy od opisu procesu generacji ludzkiego głosu, poprzez ogólną charakterystykę struktury systemów automatycznego rozpoznawania mówcy oraz ich klasyfikację, na przeglądzie literatury kończąc. Zwieńczeniem tej części jest krytyczny przegląd literatury przeprowadzony w oparciu o szereg publikacji z dziedziny cyfrowego przetwarzania sygnałów (w tym także głosu) oraz określenie dalszego kierunku badań.

W części drugiej (rozdziały 2-4) przedstawiono powszechnie wykorzystywane architektury systemów ASR oraz implementowane w nich algorytmy. Następnie przedstawiono kolejne etapy analizy sygnału mowy od momentu rejestracji do procesu segmentacji. Największą uwagę poświęcono aspektom związanym z genezą pochodzenia cech możliwych do wyekstrahowania z sygnału mowy oraz sposobom ich ekstrakcji. Zagadnienia te zostały dokładnie scharakteryzowane w kontekście właściwości powstałego w WAT systemu *ARMiA* [60], opartego na cechach fizycznych oraz autorskiego algorytmu automatycznego rozpoznawania mówcy opartego na cechach behawioralnych.

W ostatniej części rozprawy (rozdziały 5-7) zamieszczono wnioski płynące z badań prowadzonych w trakcie projektowania i optymalizacji rozwiązania integracji systemu *ARMiA* [60] z autorskim systemem opartym na cechach behawioralnych. Pokazano sposób implementacji fuzji systemów, a następnie zaprezentowano wyniki testów tego rozwiązania w warunkach istnienia zewnętrznych zakłóceń oraz zastosowania różnych torów akustycznych służących do rejestracji głosu. Otrzymane rezultaty potwierdzają postawioną wyżej tezę rozprawy.

Podziękowania

Pragnę podziękować moim promotorom Panu prof. dr. hab. inż. Andrzejowi Dobrowolskiemu oraz Panu mjr. dr. inż. Kamilowi Kamińskiemu za niezastąpione wsparcie, życzliwość i nieustanną gotowość do pomocy, niezależnie od pory dnia. Ich zaangażowanie, szeroka wiedza oraz pasja do nauki stanowią dla mnie nie tylko inspirację, lecz także wzór do naśladowania. Panowie promotorzy, pomimo licznych obowiązków zawodowych oraz realizacji skomplikowanych projektów, zawsze znajdowali czas, by udzielić cennych rad, podzielić się swoją wiedzą, a nawet wesprzeć w trudnych momentach. Ich umiejętność motywowania była kluczowa w osiągnięciu mojego celu, jakim jest ukończenie niniejszej rozprawy doktorskiej.

Szczególne wyrazy wdzięczności składam moim Rodzicom, Agnieszce i Tomaszowi za ich nieoceniony trud, wyrzeczenia oraz przekazywane wartości, które ukształtowały mnie jako człowieka. Dzięki ich wychowaniu i wsparciu stałem się osobą zdolną do podejmowania wyzwań i wytrwałego dążenia do celu. Dziękuję za ich cierpliwość, troskę oraz słowa otuchy w chwilach zwątpienia, które wielokrotnie pozwoliły mi odnaleźć siłę do dalszej pracy.

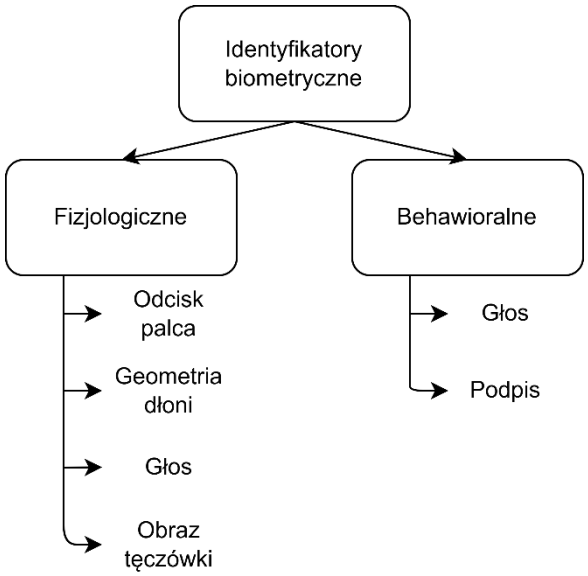
Dominik Mały

Warszawa, 01.04.2025 r.

1.

Wprowadzenie

Terminy *biometria* lub *identyfikacja biometryczna* to synonimy odnoszące się do dziedziny nauki zajmującej się ustalaniem tożsamości osoby na podstawie wyróżniających ją cech. Pierwsze udokumentowane strategie datuje się na połowę XIX wieku. Opierały się one na analizie struktury geometrycznej odciśniętej dłoni czy porównania cech antropometrycznych twarzy. Następnym krokiem było „przyjęcie" na początku XX wieku metody porównawczej odcisków palców jako oficjalnego sposobu ustalania tożsamości danego osobnika [64]. Wraz z postępem technologicznym zauważono, iż możliwa jest identyfikacja osób w oparciu o inne charakterystyczne właściwości obiektów badanych. Aktualnie rozróżnia się dwa podstawowe zbiory charakterystyk. Na rys. 1.1 pokazany jest ich podział wraz z najpowszechniej używanymi przykładami. Pierwszą grupę stanowią *biometryki fizjologiczne*, takie jak odciski palców czy geometria dłoni, a drugą *behawioralne*, których przedstawicielem jest głos bądź własnoręczny podpis [103].



Rys. 1.1. Podstawowe identyfikatory biometryczne [103]

Grupa pierwsza niesie informacje o cechach fizycznych badanego obiektu mierzonych w pewnej chwili czasu. Z kolei identyfikatory behawioralne odnoszą się do sposobu wykonywania danej czynności i są zależne od stopnia jej opanowania czy też aktualnego stanu umysłu, a ponadto biometryki te podatne są na zamierzone przez osobnika zmiany. Istnieją również takie identyfikatory biometryczne, których cechy można zaszeregować jednocześnie do fizycznych i behawioralnych. Przykładem takiej biometryki jest głos.

Podstawowym aspektem podczas wyboru i wykorzystania danej grupy identyfikatorów, i w konsekwencji konkretnych cech, jest ich skuteczność. Inaczej mówiąc, kryterium decydującym o przydatności charakterystyki na polu identyfikacji biometrycznej jest możliwość prawidłowej weryfikacji tożsamości. W literaturze wyróżnia się zbiór warunków, które muszą zostać spełnione, aby dana biometryka mogła zostać wykorzystana w efektywny sposób [63], [64]. Są to:

- uniwersalność – każdy osobnik z wyjątkiem nielicznej grupy posiada daną biometrykę;
- unikatowość – wybrana cecha przybiera indywidualną formę u każdego z osobników;
- trwałość – uodpornienie danej biometryki, a dalej całego procesu identyfikacji osoby na rozmaite, zewnętrzne procesy;
- mierzalność – możliwość ilościowego scharakteryzowania wybranej cechy;
- akceptowalność – stopień akceptacji wykorzystania biometryki przez społeczeństwo;
- odporność na podszycie (ang. *spoofing*) – trudność w pobraniu i sfałszowaniu próbki danych.

Niemożliwy jest wybór takiej biometryki, aby wszystkie z powyższych kryteriów były w stu procentach spełnione. Jedynie na drodze kompromisu możliwy jest wybór jednej z nich. Jednak istnieją dodatkowe wymagania, które ułatwiają decyzję, co do wykorzystywanej cechy. W tabeli 1.1 przedstawione są poszczególne cechy wykorzystywane w biometrii wraz ze stopniem spełnienia opcjonalnych kryteriów.

Tab. 1.1. Porównanie popularnych biometryk [63], [104]

Rodzaj biometryki	Uniwersalność	Unikatowość	Mierzalność	Trwałość	Akceptacja użytkowników	Odporność na podszycie
Twarz	Wysoka	Niska	Średnia	Średnia	Wysoka	Średnia
Tęczówka	Wysoka	Wysoka	Średnia	Średnia	Średnia	Wysoka
Odcisk palca	Średnia	Wysoka	Średnia	Średnia	Niska	Średni
Siatkówka oka	Wysoka	Wysoka	Średnia	Średnia	Niska	Wysoka
Geometria dłoni	Średnia	Średnia	Średnia	Średnia	Średnia	Średnia
Podpis	Wysoka	Wysoka	Niska	Niska	Wysoka	Niska
Głos	Średnia	Niska	Średnia	Średnia	Wysoka	Wysoka

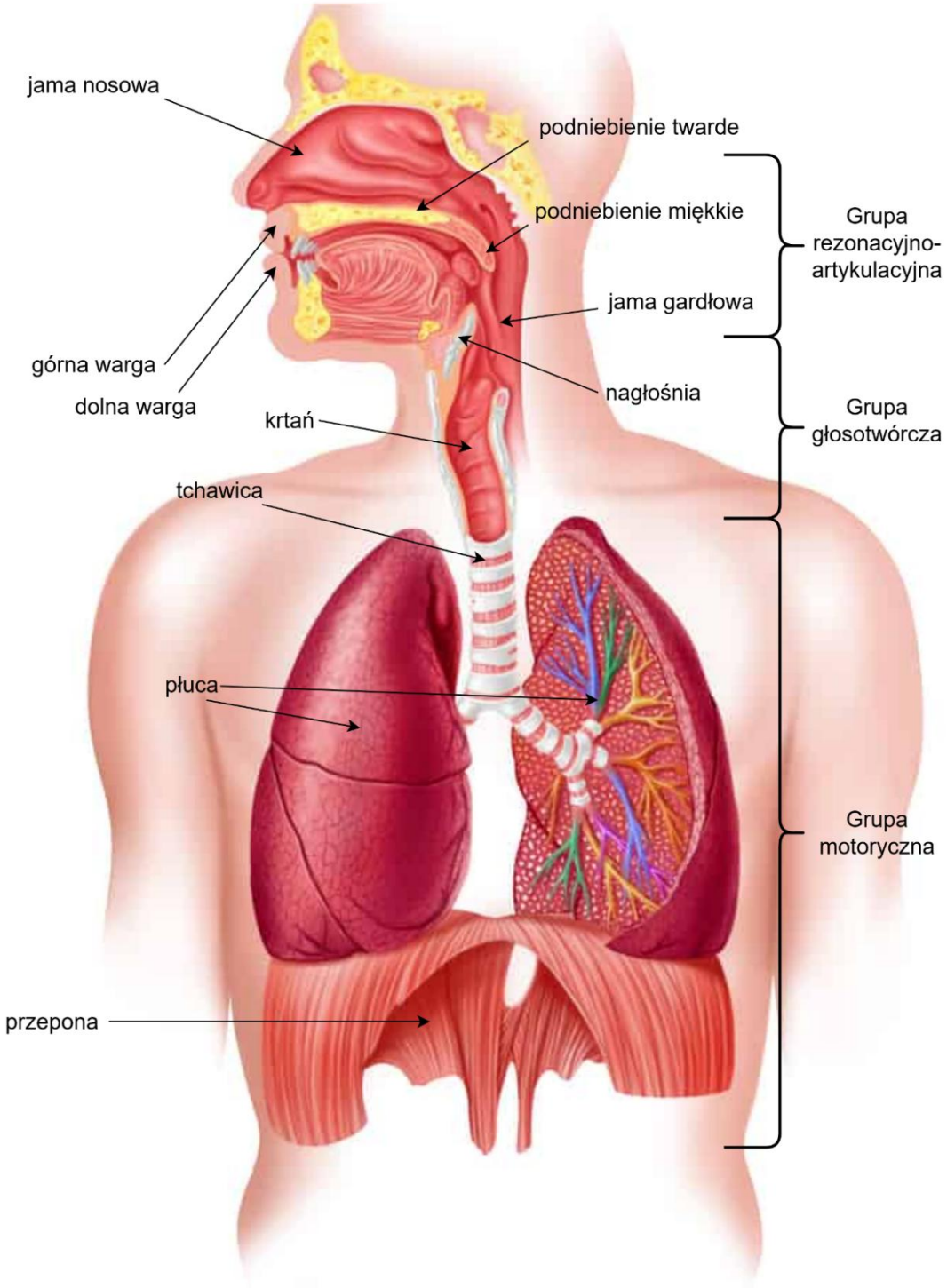
Porównując podstawowe identyfikatory biometryczne pod względem możliwości spełnienia wszystkich opisanych warunków (także tych dodatkowych), można dojść do wniosku, że to właśnie głos i jego analiza wydają się najlepszym rozwiązaniem w kwestii identyfikacji osób.

1.1. Model generacji sygnału mowy

Aby móc w pełni zrozumieć zagadnienia rozpoznawania mówców należy poznać budowę aparatu głosowego oraz proces generacji mowy. Wszelkie dostępne opisy bogate są w anatomiczne oraz biologiczne terminy, które z technicznego punktu widzenia nie niosą wielu przydatnych informacji. Charakter niniejszej pracy oraz poruszane zagadnienia jednoznacznie wskazują, iż charakterystyka traktu głosowego i procesów wytwarzania sygnału mowy pozbawiona będzie wspomnianej wcześniej terminologii.

Na rys. 1.2 przedstawiono ogólną budowę aparatu mowy człowieka. Do jego zasadniczych składowych należą m.in. *tchawica, krtań, język, podniebienie, wargi,* a także *żuchwa* i *policzki* [107]. Powyższe elementy zaszeregować można do trzech głównych grup [66]:

- Grupa motoryczna – zalicza się do niej organy znajdujące się wewnątrz klatki piersiowej oraz mięśnie brzucha. Ich zadaniem jest odpowiednie „zasilanie" traktu głosowego poprzez stały dopływ powietrza o odpowiednim, stabilnym ciśnieniu.
- Grupa głosotwórcza – narządy wchodzące w skład tej grupy odpowiadają za proces generacji fali dźwiękowej, inaczej zwany *fonacją,* poprzez wprawienie w ruch więzadeł głosowych. Są to wspomniane wcześniej *więzadła głosowe*, *krtań*, a także pozostałe *mięśnie* i *nerwy* łączące krtań z innymi narządami wchodzącymi w skład traktu głosowego.
- Grupa rezonacyjno-artykulacyjna – to elementy artykulacyjne takie jak *wargi, język, zęby* i *żuchwa,* a także *jama gardłowa, jama ustna* oraz *podniebienie twarde i miękkie.* Zadaniem tej grupy jest odpowiednia modulacja i wzmocnienie wygenerowanej wcześniej fali akustycznej za pomocą *artykulacji* wykorzystującej zjawiska *rezonansów*.



Rys. 1.2. Ogólna budowa aparatu mowy [66]

Zgodnie z przedstawionym wyżej podziałem i opisem poszczególnych składowych, a także rozważanymi w niniejszej pracy zagadnieniami, za jeden z najważniejszych elementów całego traktu głosowego uznać należy „generator", którym jest *krtań.* Tworzą ją szkielet chrzęstny, mięśnie oraz więzadła. Wśród mięśni występuje rozróżnienie na *zewnętrzne* i *wewnętrzne*. Obydwie grupy odpowiadają za ruch narządu, przy czym rola mięśni zewnętrznych skupia się na ustalaniu położenia krtani (ruch we wszystkich płaszczyznach). Z kolei mięśnie wewnętrzne, a precyzyjniej *mięśnie głosowe* uczestniczą w procesie generacji dźwięków [74]. Jest to część środkowa krtani, a jej budowa

przedstawiona została na rysunku 1.3. Zasadniczą część stanowią więzadła (fałdy) głosowe. Przestrzeń pomiędzy nimi to tchawica zwana zastępczo głośnią bądź szparą głośni.

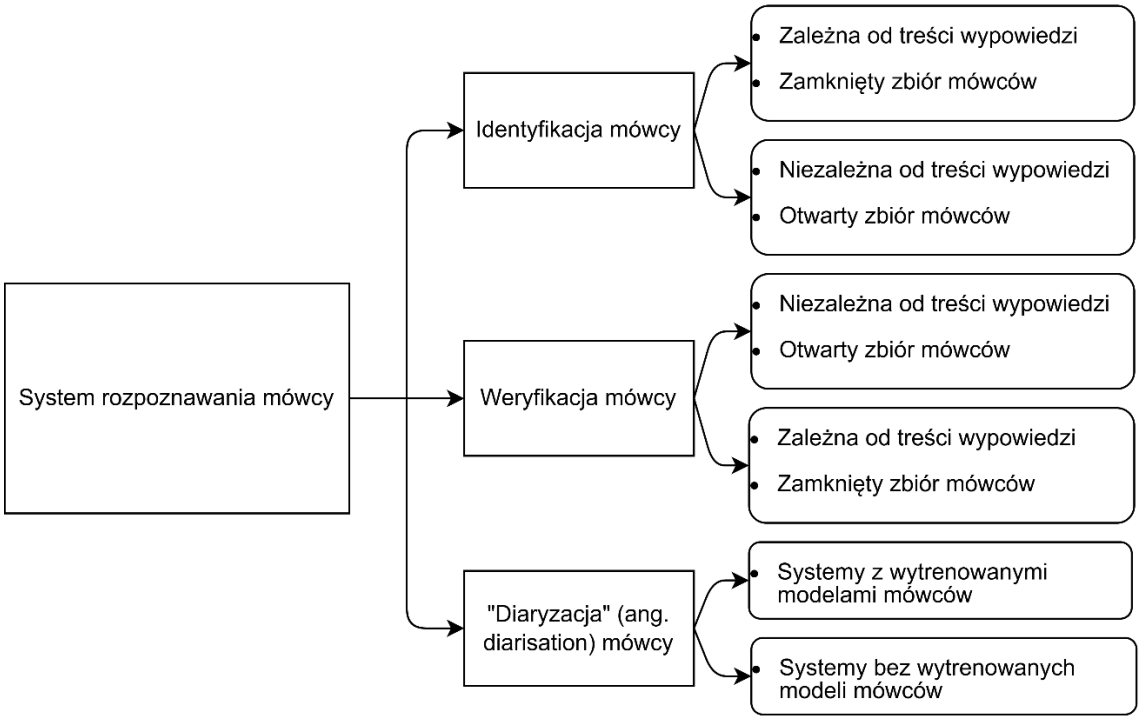
Rys. 1.3. Usytuowanie elementów krtani podczas trzech stanów komunikacji [136]

Struny (więzadła) głosowe zdolne są do zmiany swojego położenia. Mogą się otwierać i zamykać, co zmienia rozmiary tchawicy. Powiązane jest to ze zmianą stopnia przepływu powietrza z płuc. Podczas wydostawania się powietrza z płuc, fałdy głosowe wykonują bardzo szybkie ruchy otwierania i zamykania, co łudząco przypomina drgania. W ten sposób dochodzi do udźwięczniania mowy, czyli *fonacji* [36].

1.2. Klasyfikacja systemów automatycznego rozpoznawania mówcy

Automatyczne rozpoznawanie mówcy jako dziedzina zajmująca się identyfikacją rozmówcy w oparciu o jego głos ma swoje początki już w latach 60. XX wieku [57]. Wraz z ewolucją techniki komputerowej rozwijały się również systemy ASR. Jednak wszystkie wyniki opierały się na prywatnych zbiorach głosów, co uniemożliwiało ich skuteczne porównanie.
W 1996 r. *Narodowy Instytut Standaryzacji i Technologii USA* (ang. *National Institute of Standards and Technology – NIST*) zaczął przeprowadzać coroczną ocenę wszelkich dostępnych rozwiązań z dziedziny rozpoznawania mówcy pod akronimem *SRE* (ang. *Speaker Recognition Evaluation*). Ujednolicone procedury oraz wymagania, które narzucił NIST umożliwiły sprawne dokumentowanie postępów oraz rozwoju systemów ASR.

Systemy automatycznego rozpoznawania mówcy można podzielić według kilku kryteriów. Podstawowy podział zakłada rozróżnienie pod względem rodzaju systemu. Jest on widoczny na rysunku 1.4. Dokładniejsze sposoby klasyfikacji opierają się m.in. na otwartości zbioru mówców czy uwzględnianiu treści wypowiedzi [31].



Rys. 1.4. Podział systemów ASR [104]

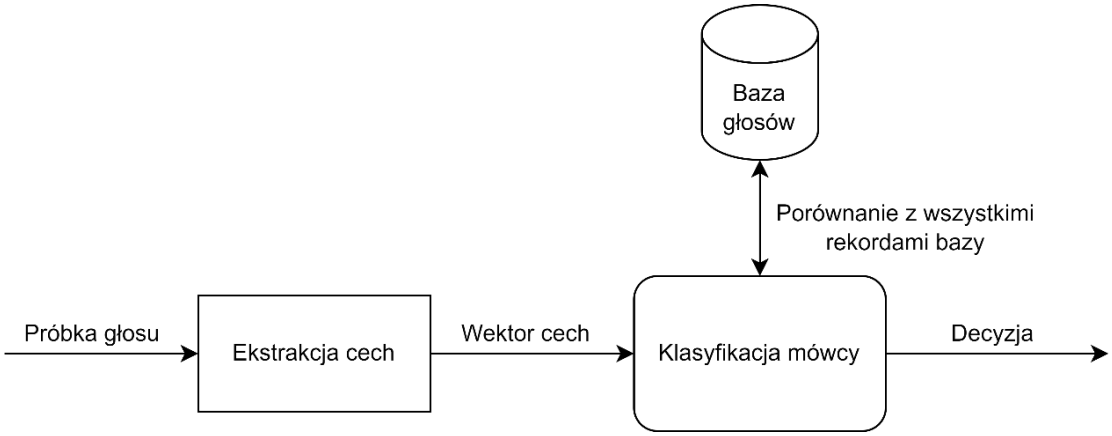
Zadania realizowane przez główne rodzaje systemów rozpoznawania mówcy przedstawione na rys 1.4., są następujące [104]:

- *Weryfikacja mówcy* – system decyzyjny, którego zadaniem jest dopasowanie nieznanego głosu do danego mówcy i decyzja co do jego uwierzytelnienia.
- *Identyfikacja mówcy* – jej zadaniem jest przyporządkowanie nieznanego mówcy do jednego z mówców zarejestrowanego w bazie wzorców.
- *Diaryzacja mówców* (ang. *speaker diarisation*) – jej zadaniem jest podział nagrania audio na segmenty, które odpowiadają poszczególnym mówcom (odpowiedź na pytanie "Kto mówi i kiedy?"). W systemach diaryzacji nie jest konieczna wcześniejsza znajomość tożsamości mówców. W ich skład wchodzą następujące moduły:
    - *Detekcja mówcy* – polega na poprawnym oznaczeniu, które części nagrania zawierają mowę, a które są hałasem lub ciszą.
    - *Segmentacja mówcy* – umożliwia znalezienie momentów czasowych w strumieniu audio, w których następuje zmiana mówców, a następnie na tej podstawie podział sygnału na mniejsze fragmenty.
    - *Grupowanie mówcy* – odpowiada za prawidłowe zaklasyfikowanie, które segmenty należą do tego samego mówcy.
    - *Przypisanie mówcy (opcjonalnie)* – jeśli system ma dostęp do uprzednio wytrenowanych modeli mówców, to w tym module może nastąpić przypisanie poszczególnych wypowiedzi do modeli mówców.

1.2.1. Identyfikacja i weryfikacja mówcy

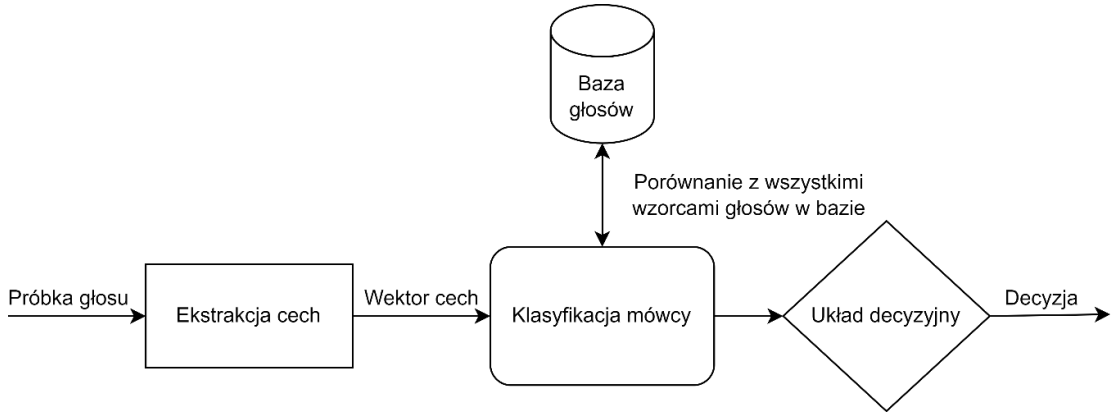
Najpopularniejszym podziałem systemów ASR jest podział uwzględniający sposób uwierzytelniania mówcy. Na tym polu wyróżnia się identyfikację oraz weryfikację głosu.

Identyfikacja to złożony proces ustalania tożsamości nieznanego mówcy poprzez porównanie jego próbki głosu z próbkami głosów zebranymi w bazie danych. Ważnym aspektem jest fakt, czy identyfikowana próbka głosu znajduję się (bądź nie) w deklarowanym wcześniej zbiorze uczącym. W związku z tym identyfikacja polega na ekstrakcji cech, a następnie przyporządkowaniu ich do najbardziej odpowiadającego wzorca. Mówi się, iż jest to identyfikacja w zamkniętym zbiorze mówców (rys. 1.5). Wadą tego rozwiązania jest to, iż mówca zawsze zostanie „rozpoznany", tzn. jego głos w każdym przypadku zostanie przyporządkowany, ale niekoniecznie zawsze w sposób poprawny.



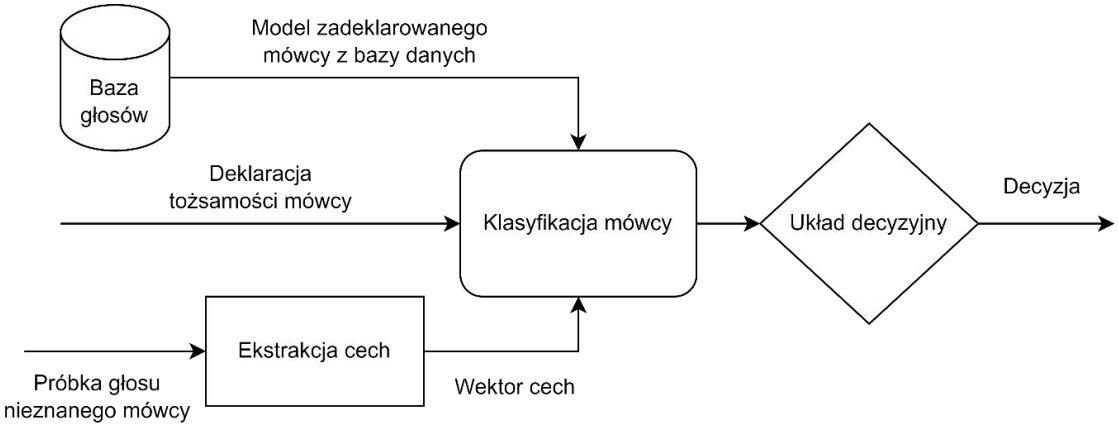
Rys. 1.5. Proces identyfikacji w zamkniętym zbiorze mówców

Dołączenie na końcu tego procesu pewnej reguły decyzyjnej, która umożliwia odrzucenie badanego obiektu jest rozwiązaniem wyżej wspomnianego problemu. Jest to system *identyfikacji* w *otwartym zbiorze mówców*. Stanowi on rozwinięcie wcześniej opisywanego systemu o zbiorze zamkniętym, gdyż w pierwszym etapie pracy następuje identyfikacja mówcy w oparciu o porównanie próbki z wzorcami. Jednak na końcu występuję układ decyzyjny, który określa czy najbardziej podobny z głosów jest w „wystarczającym" stopniu podobny, aby można było poprawnie zidentyfikować mówcę (rys. 1.6). Wynikiem tego działania jest decyzja czy identyfikacja przebiegła poprawnie i czy głos został rozpoznany.



Rys. 1.6. Proces identyfikacji w otwartym zbiorze mówców

Weryfikacja z kolei jest procesem, w którym następuje decyzja o potwierdzeniu bądź odrzuceniu deklarowanej przez mówcę tożsamości. W przypadku, gdy deklarowana próbka głosu pokrywa się w wystarczającym stopniu z autentyczną, system decyduje, że badany obiekt jest faktycznie osobą, której tożsamość deklaruje. W przeciwnym przypadku wynik weryfikacji jest negatywny, a mówca nie zostaje uwierzytelniony. Całość procesu widoczna jest na rys. 1.7.



Rys. 1.7. Proces weryfikacji mówcy

1.2.2. Treść wypowiedzi w systemach rozpoznawania mówcy

Wewnątrz procedur identyfikacji oraz weryfikacji można wyróżnić dokładniejszy podział, który opiera się na kontekście treści wypowiedzi badanego mówcy. W tym przypadku dostępne są systemy ASR *niezależne od treści wypowiedzi* (ang. *text-independent*), w których wypowiedź może być dowolna i nie ma znaczenia oraz *zależne od treści wypowiedzi* (ang. *text-dependent*), w których do tworzenia modeli głosów wykorzystywana jest wypowiedź o ściśle określonej treści [69]. W systemach zależnych od treści wypowiedzi procesy uczenia oraz testowania opierają się na tej samej frazie wypowiadanej przez mówców, co jednoznacznie przekłada się na skrócenie czasu procesu uczenia, gdyż posiadamy skończony zbiór danych wejściowych do weryfikacji. Z kolei przy systemach niezależnych od treści wypowiedzi nie występują ograniczenia co do frazy, którą badany obiekt musi wypowiedzieć. Przekłada się to na wydłużenie procesu uczenia wewnątrz systemu. Jednak zaletą takiego rozwiązania jest większa wygoda dla mówców, gdyż nie muszą wypowiadać tych samych fraz wielokrotnie [84], [121], [122].

1.2.3. Cechy sygnału mowy

Analizowany przez systemy ASR głos ludzki niesie wiele informacji przydatnych w procesie uwierzytelniania mówcy. Ekstrakcja cech z sygnału mowy ludzkiej musi spełniać kilka kryteriów [129]. Cechy te powinny:

- być odporne na szumy oraz zakłócenia;
- być trudne do naśladowania;
- być łatwe do wyekstrahowania z sygnału mowy;
- być naturalnie oraz okresowo występujące w mowie;
- być niezależne od stanu zdrowia mówcy;
- cechować się dużą różnorodnością między mówcami.

Ponadto ważne jest, aby liczba cech była relatywnie mała. Z fizycznego punktu widzenia ekstrahowane są one jako [121]:

- krótkoczasowe cechy widmowe,
- cechy widmowo-czasowe,
- cechy wysokopoziomowe (będące efektem złożonego przetwarzania).

Grupa pierwsza to charakterystyki o krótkim czasie trwania (około 20-30 ms), których właściwości zależne są od uwarunkowań fizjologicznych obiektu, m.in. takich jak budowa *traktu głosowego*.

Cechy widmowo-czasowe trwają setki milisekund i dotyczą *intonacji wypowiadanej frazy, iloczasu* czy *rytmu*. Często charakterystyki te nazywa się prozodycznymi bądź prozodyjnymi
(w nawiązaniu do brzmieniowych właściwości cech).

Ostatnia grupa to tzw. cechy wysokiego poziomu. Opierają się one na czynnikach socjoekonomicznych, które miały i mają wpływ na mówcę, np. *status społeczny, miejsce urodzenia, używany język* czy *osobowość*. Są to z reguły cechy behawioralne.

W tabeli 1.2 przedstawione są grupy cech wraz z ich zaletami i wadami [121]:

Tab. 1.2. Zestawienie cech ekstrahowanych z głosu ludzkiego [121]

Grupa	Charakterystyki	Zalety	Wady
Krótkoczasowe cechy widmowe	• widmo, • właściwości rytmu.	• mała ilość danych potrzebnych do poprawnego działania systemu, • niezależność od treści wypowiedzi, • możliwe rozpoznawanie w czasie rzeczywistym.	• podatne na szumy i zakłócenia, • duże ryzyko błędnego uwierzytelnienia mówcy.
Cechy widmowo-czasowe	• wysokość tonu, • rytm wypowiedzi, • iloczas, • akcent.	• częściowa odporność na szumy i zakłócenia.	• trudne do ekstrakcji, • wymagana duża ilość danych uczących.
Cechy wysokiego poziomu	• głoski, • idiolekt, • semantyka, • akcent, • wymowa.	• odporność na szumy i zakłócenia.	• trudne do ekstrakcji, • wymagana duża ilość danych uczących.

1.3. Krytyczny przegląd literatury

W starszych rozwiązaniach (lata 1999-2008) do ekstrakcji cech widmowych wykorzystywało się transformację Fouriera (ang. *Fourier Transform – FT*) oraz wszelkie jej rozszerzenia takie jak dyskretna transformacja Fouriera (ang. *Discrete Fourier Transform – DFT*) czy szybka transformacja Fouriera (ang. *Fast Fourier Transform* – FFT). Poza podstawowymi metodami analizy widmowej niektórzy z twórców systemów ASR korzystali z technik *predykcji liniowej* [70], [115] (ang. *Linear Prediction – LP)* oraz jej odmiany cepstralnej (ang. *Linear Prediction Cepstral Coefficients – LPCC)* [143]. Wraz z biegiem czasu odkryto, iż wykorzystanie opracowanej w latach 80. XX wieku metody *MFCC* (ang. *Mel-Frequency Cepstral Coefficients – MFCC*) jest często skuteczniejsze w procesach rozpoznawania mówcy niż metody wcześniej wymienione

[132]. Polega ona na filtracji podpasmowej sygnału mowy przy użyciu filtrów pasmowo-przepustowych, które rozłożone są w sposób równomierny w melowej skali częstotliwości [129].

Grupa druga ekstrahowanych cech (cechy widmowo-czasowe) nie jest możliwa do wyznaczenia wprost z przetworzonego wcześniej sygnału mowy. Dlatego charakterystyki prozodyczne uzyskuje się przy pomocy *predykcji liniowej*, a następnie *filtracji odwrotnej* (ang. *Inverse Filtering*) [130]. Dopiero po zastosowaniu tych technik można wykorzystać metodę MFCC, aby otrzymać pochodne cech dystynktywnych pierwszego i drugiego rzędu (ang. *delta features, delta-delta features*) [131], [134]. W niektórych podejściach wykorzystuje się również dyskretną transformację kosinusową (ang. *Discrete Cosine Transform – DCT)* [134] oraz jej rozwinięcie czasowe (ang. *Temporal Discrete Cosine Transform – TDCT)* [133].

Cechy wysokiego poziomu są zazwyczaj opisywane przez badaczy w najmniejszym stopniu [129]. Krytykują oni zarówno skomplikowany proces ich ekstrakcji, jak i niski stopień rozwoju systemów rozpoznawania mówcy działających w oparciu o nie. Podkreślają także, iż modele mieszanin gaussowskich są jednym z głównych filarów podczas procesu rozpoznawania głosu w tej grupie cech [11]. Dodatkowo często wymienia się bardzo małą liczbę udokumentowanych rozwiązań w porównaniu do dwóch poprzednich zbiorów charakterystyk.

Kolejnym krokiem w rozwoju systemów ASR, było wprowadzenie procesu detekcji aktywności głosowej (ang. *Voice Activity Detection – VAD*) w celu wykrywania okresów ciszy [137], normalizacji cech (wykorzystując cepstralną normalizację średnich – CMN czy filtrację RASTA) [110] oraz redukcji wymiaru otrzymanego sygnału przy użyciu heteroskedastycznej liniowej analizy dyskryminacyjnej (ang. *Heteroscedastic Linear Discriminant Analysis – HLDA)* [67].

Wraz z postępem technologicznym nastąpił rozwój szeroko pojętych metod uczenia maszynowego, a także sieci neuronowych. W dziedzinie rozpoznawania mówców również znalazły one swoje zastosowanie. Przewarzającym przedstawicielem sztucznych sieci neuronowych w systemach ASR była sieć wektorów nośnych (ang. *Support Vector Machine – SVM)* [141], [142].

Poza charakterystyką metod parametryzacji sygnału mowy ważnym aspektem jest proces modelowania głosu człowieka. Poniżej znajduję się spis pięciu najczęściej wykorzystywanych metod opisu głosu. Badacze traktują je jako swoiste „epoki" w dziedzinie automatycznego rozpoznawania mówcy [95]:

- „era GMM" – modele mieszanin rozkładów Gaussa,
- „era JFA" – zespolona analiza czynnikowa jako rozszerzenie poprzedniej metody,
- „era i-Vector" – metoda i-wektorów jako uproszczenie metody JFA,
- sieć neuronowa z wykorzystaniem *PLDA* (ang. *Probabilistic Linear Discriminant Analysis* – rozszerzenie LDA) oraz i-wektorów,
- sieć neuronowa z wykorzystaniem PLDA oraz x-wektorów.

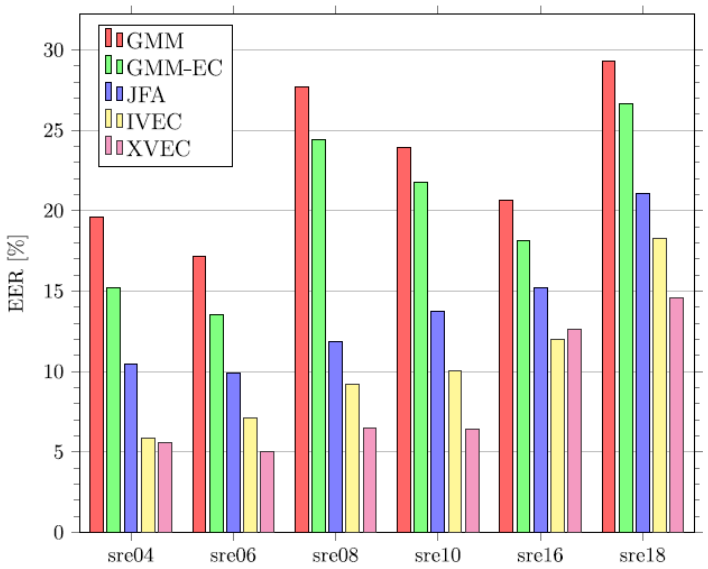
Zgodnie z teorią, technika mieszanin gaussowskich opiera się na reprezentacji parametrycznych funkcji gęstości prawdopodobieństwa jako sum *C* rozkładów Gaussa [147]. Do utworzenia ich korzysta się ze zbioru cech melcepstralnych (MFCC) otrzymanych poprzez zastosowanie filtrów melowych z bardzo krótkim (20 ms) oknem

pomiarowym i przesunięciem równym połowie tego czasu. Dodatkowo podział na ramki wiązał się z usuwaniem okresów ciszy po wcześniejszej detekcji poziomu głośności. Stosując GMM skupiono się na wykorzystaniu cech z grupy pierwszej i drugiej (krótkoczasowe cechy widmowe i prozodyczne) [60]. Następnie przy pomocy estymacji największej wiarygodności (ang. *Maximum Likelihood – ML)* i szacowania zgodnie z algorytmami maksymalizacji oczekiwań (ang. *Expectation Maximization – EM)* i minimalnej rozbieżności (ang. *Minimum Divergence – MD) t*worzone były odpowiednie modele głosów [67].

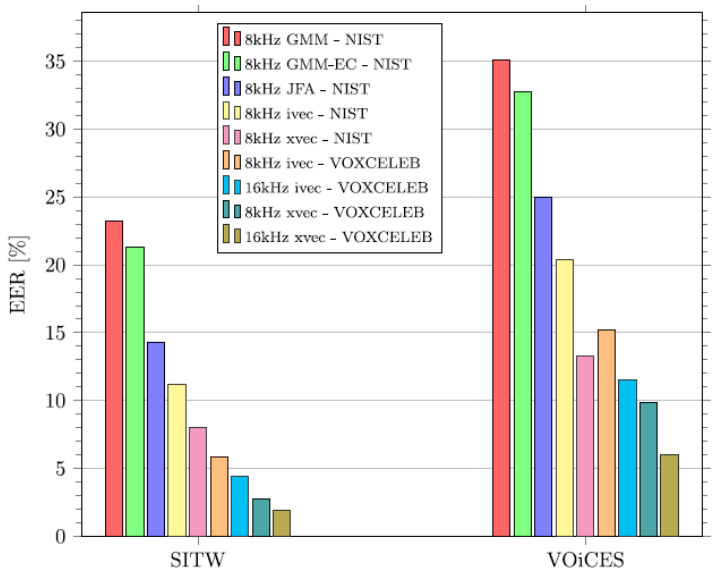
Na podstawie tego zasadniczego sposobu tworzenia bazy głosów opracowane zostały kolejne rozwiązania takie jak zespolona analiza czynnikowa (ang. *Joint Factor Analysis – JFA*) [92] czy metoda i-wektorów [78]. Techniki te różnią się jedynie sposobem reprezentacji czynników mówcy oraz kanału jako jednego bądź oddzielnych komponentów. Nawet wykorzystanie nowoczesnych sieci neuronowych oraz x-wektorów przeznaczonych do systemów ASR (ang. *ASR DNN with Kaldi Toolkit – Automatic Speech Recognition Deep Neural Network with Kaldi Toolkit*) opiera się na cechach wykorzystywanych przy GMM z tą różnicą, że długości segmentów danych uczących sięgają aż do trzech sekund [27]. Normalizacja wyników (ang. *Score Normalization*) [96] oraz rozszerzenie danych (ang. *Data Augmentation*) [89], to dodatkowe procesy poprawiające działanie wybranych algorytmów.

Możliwe jest zestawienie wyników przeprowadzonych badań oraz porównanie poszczególnych algorytmów przy użyciu *stopy błędu zrównoważonego (*ang. *Equal Error Rate – ERR)* [95]. Dzięki niej można dokonać łatwej oceny jakości wykorzystywanych algorytmów klasyfikacji mówcy. Im niższa wartość współczynnika EER tym wyższa jakość systemu ASR.

Rys. 1.8 oraz 1.9 reprezentują graficzne przedstawienie wyników uzyskanych przy różnych algorytmach generacji modeli mówców. W obu przypadkach potwierdza się zależność, iż współczesna metoda x-wektorów przoduje w dokładności rozpoznawania głosów.



Rys. 1.8. Zestawienie poziomów parametru ERR dla różnych algorytmów tworzenia modeli mówców i baz głosów podczas kolejnych edycji SRE [95]



Rys. 1.9. Porównanie metod modelowania głosów w dwóch niezależnych bazach: SITW oraz VOiCES [95]

Poza opisanymi wyżej metodami modelowania głosu istnieje również inny podział, przedstawiony poniżej [104]:

- *Generative models* – modele, które opierają się na statystycznym opisie rozkładu cech:
  - *Parametric models* – w skład tej grupy wchodzą GMM oraz HMM (ang. *Hidden Markov Model*).
  - *Non-parametric models* – modele, które operują na wzorcu zawartym w wektorze cech. Zalicza się do nich metody dynamicznego dopasowania czasowego (ang. *Dynamic Time Warping – DTW*) oraz kwantyzacji wektorów (ang. *Vector Quantization – VQ).*
- *Discriminative models* – w jej skład wchodzą sieci neuronowe takie jak *ANN* (ang. *Artificial Neural Network*) czy SVM.

W przypadku operowania na otwartym zbiorze mówców, każdy z opisanych wyżej modeli do prawidłowego działania wymaga danych uczących mówców należących do bazy głosów i spoza niej. Aktualnie termin modelu generatywnego stosuje się w ujęciu do metod tworzenia syntetycznych (sztucznych) próbek głosu.

Często poruszane są zagadnienia dotyczące wyzwań stojących przed systemami rozpoznawania mówcy z punktu widzenia systemu komputerowego. Wskazuję się tutaj dwie główne kwestie jakimi są zmienność oraz niewystarczająca ilość danych do poprawnego działania opracowanego rozwiązania. Ponadto wpływ szumów tła czy także problemów po stronie rozmówcy bądź samego procesu identyfikacji jest również znaczący.

Zgodnie z przedstawionym wyżej podziałem metod modelowania wykorzystywane są różne sposoby parametryzacji głosu mówcy. Pierwszym z nich jest dyskretna transformacja falkowa (ang. *Discrete Wavelet Transform – DWT*) zastosowana do detekcji głosu w języku czeskim [93]. Twórca tej metody użył modeli mieszanin gaussowskich wraz z siecią neuronową MLP (ang. *Multi-Layer Perceptron – MLP*), a w dalszych etapach LPCC. Skuteczność identyfikacji z wykorzystaniem tego rozwiązania to około 98%. Kolejna metoda to często wykorzystywane MFCC wraz z klasyfikatorem GMM oraz rozmytą siecią

min-max (ang. *Fuzzy Min-Max Neural Network – FMMNN*) [82]. W późniejszym czasie użyto sieci FSVM (ang. *Fuzzy Support Vector Machine*), czyli rozmytej sieci wektorów nośnych [146]. Tutaj współczynnik identyfikacji (ang. *Identification Rate – IR)* wahał się od 94% do prawie 99%. Trzecia z metod to wykorzystanie TTESBCC (ang. *Temporal Teager Energy Based Subband Cepstral Coefficients*) do detekcji mowy szeptanej w zamian za FMMNN [81]. W tym rozwiązaniu porównywano jeszcze dwie dodatkowe techniki otrzymując IR na poziomie 98,6% oraz 55,8% dla szeptów. Współczynnik identyfikacji w zakresie 98-100% uzyskano przy pomocy znormalizowanych cech widmowych (ang. *Normalized Dynamic Spectral Features – NDSF*) [123]. Dwie ostatnie metody opierają się na wcześniej opisanym współczynniku ERR. Pierwsza z nich korzysta z krótkoczasowych spektrogramów w głębokich sieciach konwolucyjnych (ang. *deep Convolutional Neural Network – dCNN*) [53]. Ostatnia opiera się na jednej z najnowszych technik jaką są x-wektory w sieciach TDNN (ang. *Time Delay Neural Network)* oraz ResNet (ang. *Residual Neural Network – resztkowa sieć neuronowa*) [55]. Wśród tych rozwiązań osiągnięto wartości odpowiednio 3,95% i 1,40% stopy błędu zrównoważonego. W tabeli 1.3. zestawiono najważniejsze metody biometrii głosu wraz z parametrami, które je opisują. Mimo wykorzystania tych samych miar sprawności systemów nie można jednoznacznie stwierdzić, który działa najefektywniej, gdyż wszystkie z opisywanych algorytmów opierają się na różnych bazach danych uczących.

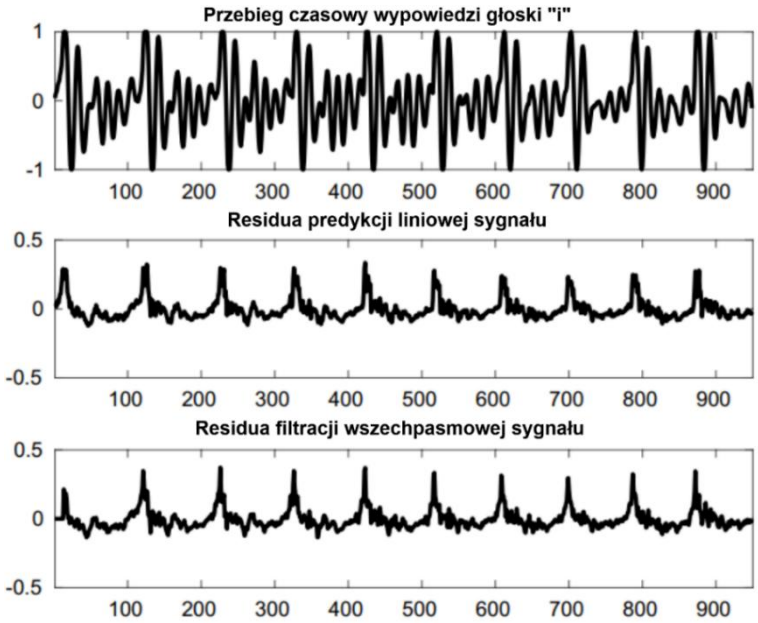
Tab. 1.3. Zestawienie metod ekstrakcji cech oraz technik modelowania głosów [104]

Rok opracowania	Metody ekstrakcji cech	Techniki modelowania bazy głosów	Wykorzystana baza głosów	Liczba mówców	Dokładność
2010	DWT	MLP	*Corpora* – baza utworzona przez twórców	Corpora 1 – 10 Corpora 2 - 50	Dla falki Sym20 – 98% IR
2011	MFCCs	FMMNN	baza utworzona przez twórców	50	99,9% IR
2012	13MFCCs + 13Δ + 13ΔΔ	FSVM	King speech	51	98,76% IR
2013	TTESBCC	GMM	baza utworzona przez twórców	25	55,8% IR - mowa szeptana 98,6% IR – mowa zwykła
2015	NDSF	-	baza utworzona przez twórców + *MVSR-IITG*	100	98-100% IR
2018	krótko-czasowe spektrogramy	CNN (ResNet)	VoxCeleb2, VoxCeleb1	6000, 1251	Dla ResNet50 – 3,95% ERR
2019	MFCCs	CNN (dCNN)	baza utworzona przez twórców + *SRL82*	50	71% IR dla *SRL82* 75% IR dla pozostałych
2019	x-wektory	E-TDNN	NIST SRE18, VAST	4500, 7000	4,95% ERR dla *NIST SRE18* 11,1% ERR dla *VAST*
2019	x-wektory	DNN	GBR-ENG	534	1,4% ERR

Istnieją również systemy ASR działające w oparciu o krótkie wypowiedzi oraz współczynniki transformacji falkowej reszt predykcji liniowej zsynchronizowanej z wysokością dźwięku [139]. Przytoczone w tym rozwiązaniu „poziomy mowy" są

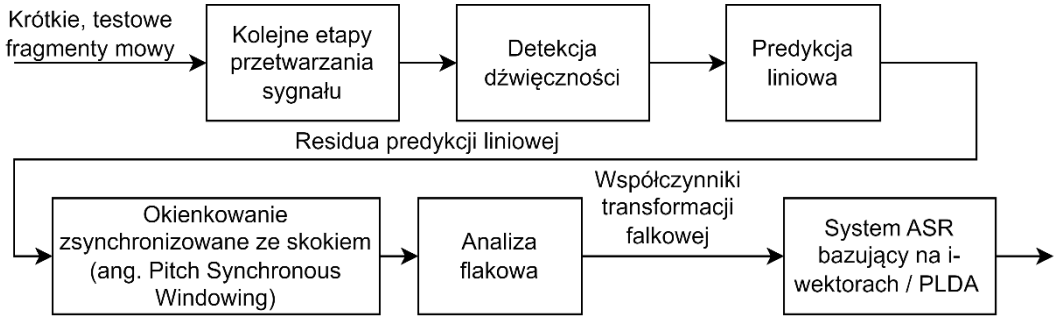
analogiczne do opisanych w pierwszym rozdziale cech sygnału mowy. Parametryzacja głosu mówcy opiera się na cechach fizycznych, otrzymanych z dźwięku mowy ludzkiej po przejściu przez trakt głosowy. Mogą być one reprezentowane poprzez residua predykcji liniowej oraz pochodne tonu krtaniowego.

Uzyskuje się je z odpowiedniej długości danych do uczenia (2,5 minuty) i testowania (od 1 do 2,5 minuty). W tym rozwiązaniu wykorzystano cepstralne współczynniki predykcji liniowej LPCC oraz filtrację odwrotną, co w efekcie pozwoliło otrzymać reszty (residua) predykcji liniowej zawierające informację o wcześniej wspomnianych cechach fizycznych [139]. Efekt przetwarzania sygnału głoski „i” przedstawiony jest na rysunku 1.10.



Rys. 1.10. Sygnał czasowy wypowiadanej głoski „i” wraz z jej współczynnikami reszt LP oraz AP [139]

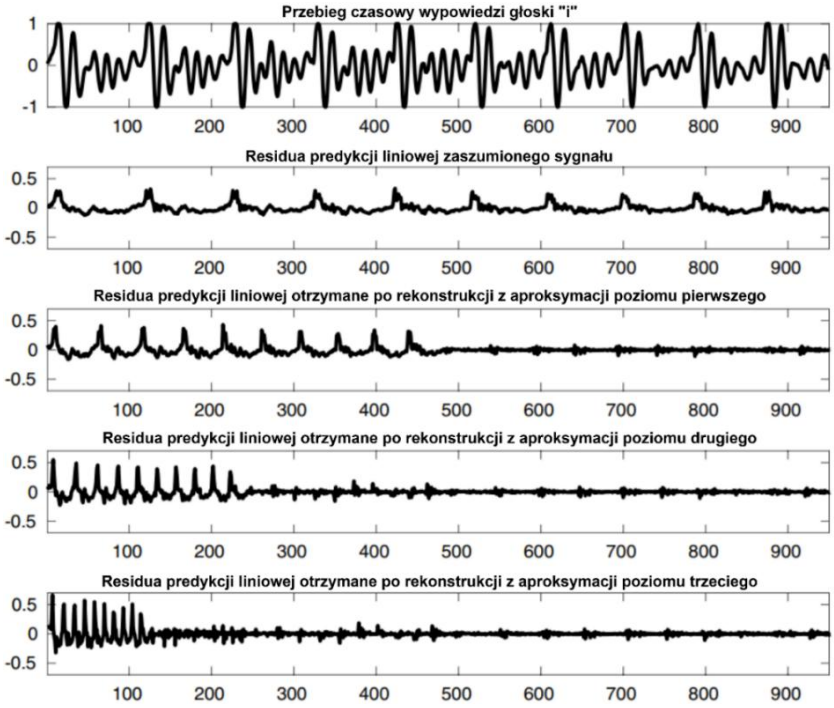
Nowatorskie podejście będące fuzją współczynników LPCC wraz z dyskretną transformacją falkową (DWT) umożliwia uzyskanie zupełnie nowego wektora cech zwanego współczynnikami falkowymi reszt oktawowych (ang. *Wavelet Octave Coefficients of Residues – WOCOR)* [117]. Dodatkowo wyróżniona jest metoda wykorzystująca residua predykcji liniowej wraz z auto-asocjacyjną siecią neuronową (ang. *Auto-Associative Neural Network – AANN)* [114]. Na rysunku 1.11 pokazana jest część struktury opisywanego rozwiązania służąca do testowania.



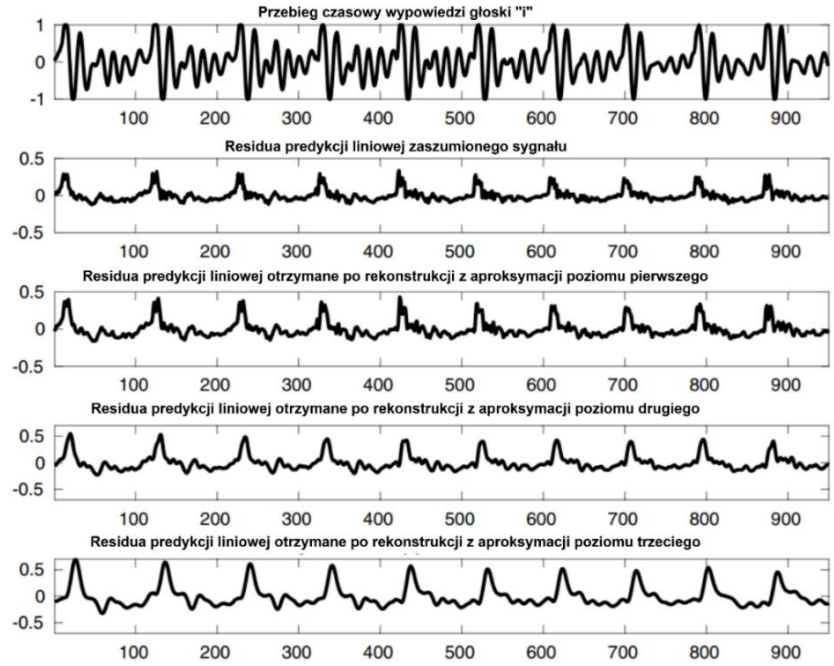
Rys. 1.11. Schemat blokowy fazy testowania systemu [139]

Na potrzeby systemu ASR wprowadzono pojęcia *downsamplingu* oraz *upsamplingu.* Pierwsze z nich dotyczy redukcji częstotliwości próbkowania sygnału poprzez usunięcie

próbek, tj.: downsampling o **2** polega na zmniejszeniu liczby próbek **2-krotnie**. Z kolei upsampling to operacja odwrotna. Zwiększona zostaje częstotliwość próbkowania sygnału poprzez dodanie do niego próbek. Analogicznie do downsamplingu o **2**, w przypadku zabiegu odwrotnego liczba próbek zostaje **2-krotnie** zwiększona. Pojęcia te są kluczowe, gdyż dzięki nim istota dekompozycji DWT (ang. *Discrete Wavelet Transform – DWT)* oraz SWT (ang. *Stationary Wavelet Transform – SWT)* staje się prostsza w zrozumieniu. Rysunki 1.12 oraz 1.13 stanowią swoiste porównanie wyników tych dwóch transformacji. Twórcy tego rozwiązania jednoznacznie wskazują na wyższość stacjonarnej odmiany przekształcenia w opracowanym systemie.

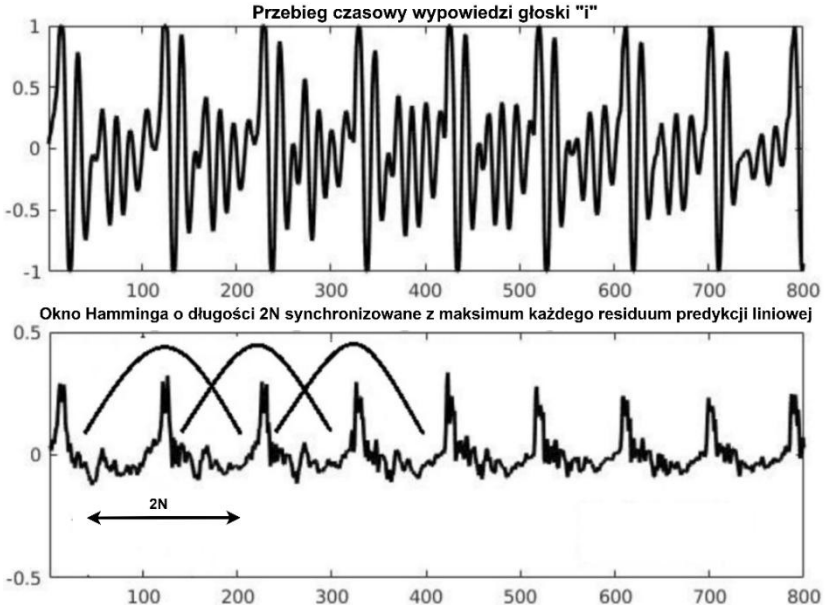


Rys. 1.12. Rekonstrukcja reszt predykcji liniowej przy użyciu DWT wymawianej głoski „i” [139]



Rys. 1.13. Rekonstrukcja reszt predykcji liniowej przy użyciu SWT wymawianej głoski „i” [139]

Nowym pojęciem jakie zostało wprowadzone jest „okienkowanie zsynchronizowane ze skokiem" (ang. *Pitch Synchronous Windowing – PSW)* [139]. Proces ten polega na dopasowywaniu okna w taki sposób, aby pik sygnału był mniej więcej w środku, przy zachowaniu cech widmowych. Pokazano to na rysunku 1.14.



Rys. 1.14. Graficzne przedstawienie procesu PSW [139]

Otrzymane wyniki wskazują, iż transformacja DWT nie jest odporna na zakłócenia w odróżnieniu od SWT. W tabeli poniżej zestawiono otrzymane rezultaty. Widać, że system oparty na przekształceniu stacjonarnym charakteryzuję się mniejszą wartością współczynnika EER oraz DCF (ang. *Decision Cost Function – DCF*), co jednoznacznie podkreśla jego wyższość nad rozwiązaniem bazującym na dyskretnej transformacji falkowej. W badaniu wykorzystano 10-sekudnowe fragmenty mowy do uczenia i testowania.

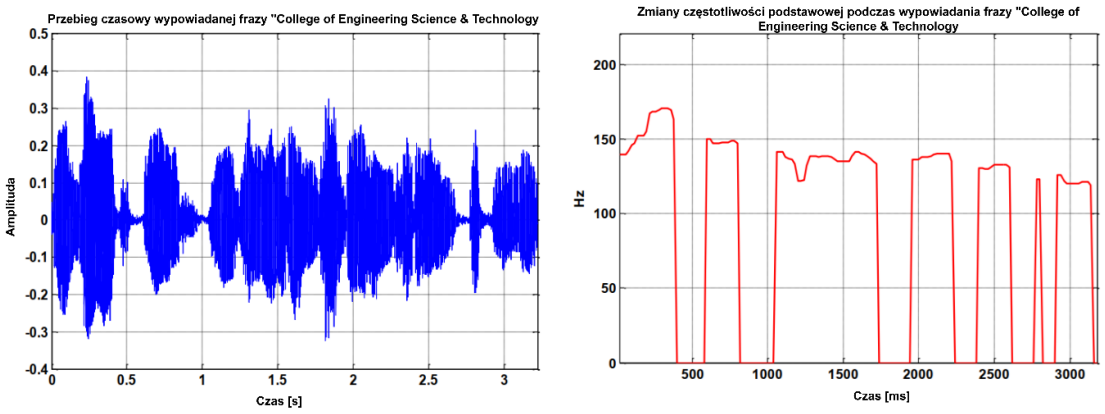
Tab. 1.4. Zestawienie wyników działania systemu na bazie danych NIST SRE 2010 [139]

Wektor cech	Poziom dekompozycji	EER (%)	DCF
DWT	1	45	0,4326
DWT	2	51	0,4816
DWT	3	53	0,4710
SWT	1	44	0,4173
SWT	2	43	0,4162
SWT	3	40	0,3959

Autorzy wskazują, że dla krótkich wypowiedzi system ten jest skuteczny, a 10-sekundowe fragmenty mowy wykorzystywane do uczenia i testowania są w pełni wystarczające. Ponadto wykorzystanie transformacji SWT daje obiecujące rezultaty, zwłaszcza w odniesieniu do zaszumionych sygnałów, z którymi przekształcenie DWT nie było w stanie sobie poradzić. Dalszy kierunek rozwoju tego systemu to fuzja wszystkich trzech poziomów cech przy tym samym, a nawet wyższym poziomie skuteczności.

Aktualnie coraz częściej dostrzega się, że cechy behawioralne są nieodłącznym elementem mowy ludzkiej, niosąc przy tym ze sobą pewną dozę unikatowych informacji przy zachowaniu zadowalającej odporności na zakłócenia. Zazwyczaj systemy

rozpoznawania mówcy opierają się na wykorzystaniu cech widmowych bądź prozodycznych, które podatne są na szumy pochodzące z zewnątrz i zdarzenia w kanale transmisyjnym [8], [119]. Pierwszą z często analizowanych cech behawioralnych jest intonacja, czyli zmiany częstotliwości podstawowej generowanego dźwięku w zależności od wypowiadanego tekstu oraz cech osobniczych badanej osoby [54]. Poniżej przytoczono przykład wypowiadania frazy: *„College of Engineering Science & Technology”* przez mężczyznę. Na rys. 1.15. pokazano dwa przebiegi obrazujące wyżej wymienioną frazę. Pierwszy z nich to wykres w dziedzinie czasu, a drugi to charakterystyka zmian częstotliwości podstawowej w czasie wypowiedzi – można z niej odczytać okresy występowania ciszy podczas konwersacji, a w zasadzie przedziały czasowe pozbawione mowy dźwięcznej.



Rys. 1.15. Zobrazowanie wypowiedzianej frazy w dziedzinie czasu oraz zmiany częstotliwości w czasie wypowiadania [119]

Kolejną cechą behawioralną, którą przytaczają badacze jest „stres lingwistyczny” zwany popularnie akcentem [119]. Akcent to umiejętność wyróżniania poszczególnych głosek/sylab przy pomocy większej siły (akcent dynamiczny), wyższego tonu (akcent prozodyczny) czy dłuższego czasu wypowiadania (akcent iloczasowy). Jest to jedna z charakterystycznych właściwości każdego z mówców, gdyż zależy od języka wypowiedzi oraz od usposobienia człowieka, a ponadto zaakcentowana część wyrazu jest bardzo szybko wychwytywana. Rytm wypowiedzi jako kolejna z cech fizjologicznych odnosi się do czasu trwania danej frazy. Każda osoba ma własny wzorzec rytmiki głosu, który jest wychwytywany nawet przez dzieci [119]. Ważnym aspektem wpływającym na ton, szybkość czy akcent wypowiedzi mają emocje. Wykrywania emocji dokonuje się za pomocą modeli mieszanin gaussowskich przy współpracy z sieciami neuronowymi. Za przykład może służyć połączenie wykorzystania głębokich sieci neuronowych i metody MFCC, jako sposobu na ekstrakcję cech wysokopoziomowych badanego obiektu z cech krótko-czasowych [85].

Ciekawym aspektem jest zastosowanie fuzji danych, a konkretniej cech wysokiego poziomu z cechami prozodycznymi, a także metod uczenia sieci SVM wykorzystanej do uczenia różnych systemów ASR. Aktualnie istnieją dwa sposoby łączenia otrzymanych charakterystyk mówcy [119]. Pierwszy z nich to suma ważona wyników klasyfikatorów użytych do ekstrakcji cech. Drugi opiera się na przedziałach ufności uzyskiwanych za pomocą łączenia reszt współczynników otrzymanych po transformacji falkowej ze współczynnikami MFCC [128]. Opis procesu uczenia sieci neuronowej skoncentrowany jest na podejściu z dużym marginesem separacji (ang. *Large Margin Approach*). W pracy

[119] wykorzystano ukryte modele Markowa (ang. *Hidden Markov Model – HMM)* oraz bazy TIMIT oraz HTIMIT z 48 fonemami, dzieląc je na 3 grupy zawierające odpowiednio 500, 100 oraz 3093 próbki głosów. Dane uczące to próbkowana z szybkością 8 kS/s baza TIMIT. Wyniki z eksperymentu zaprezentowane zostały w tabeli 1.5 wraz z porównaniem do dwóch innych prac z wykorzystaniem podobnej metodyki działania.

Tab. 1.5. Prezentacja wyników eksperymentu [119]

Autor metody z wykorzystaniem cech behawioralnych	Skuteczność rozpoznania mówcy – próbki poniżej 10 ms (% IR)	Skuteczność rozpoznania mówcy – próbki poniżej 20 ms (% IR)	Skuteczność rozpoznania mówcy – próbki poniżej 30 ms (% IR)	Skuteczność rozpoznania mówcy – próbki poniżej 40 ms (% IR)
Baza TIMIT próbkowana z szybkością 16 kS/s				
Singh [119]	76,5	90,77	96,44	99,23
Brugnara [35]	75,3	88,9	94,4	97,1
Hosom [56]		92,6		
Baza TIMIT próbkowana z szybkością 8 kS/s				
Singh [119]	84,23	94,21	97,12	99,1
Baza HTIMIT próbkowana z szybkością 8 kS/s				
Singh [119]	72,5	89,68	95,89	97,4

Badacze często dokonują analizy powszechnie wykorzystywanych metod, które zostały opisane już wcześniej w niniejszej pracy (x-wektory, GMM, PLDA czy DNN) [139]. Wynikiem tego przeglądu jest wniosek, że użycie mieszanin modeli Gaussa (GMM) oraz współczynników melcepstralnych (MFCC) przy projektowaniu systemów ASR pozwala na osiąganie wysokiej skuteczności identyfikacji mówców. Dodatkowym atutem takich rozwiązań jest ich uniwersalność, która pozwala na ich dalszą, szeroką rozbudowę.

1.4. Podsumowanie

Niniejszy rozdział to zbiór podstawowych wiadomości dotyczących modelu generacji głosu oraz związanych z nim systemów rozpoznawania mówcy. Omówiono w uproszczony sposób wytwarzanie głosu ludzkiego oraz skorelowane z tym procesem cechy, które wykorzystuje się w systemach ASR. Przedstawiono zasadniczy podział oraz struktury najważniejszych rozwiązań w dziedzinie automatycznego rozpoznawania mówcy. Ponadto – w oparciu o światową literaturę – dokonano przeglądu aktualnego stanu wiedzy w obszarze rozpoznawania mówcy.

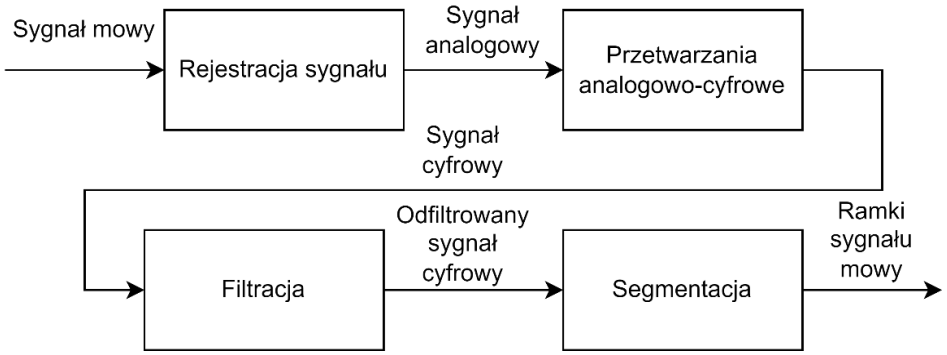
Zgodnie z wnioskami wynikającymi z dokonanego przeglądu literatury, aktualnie istniejące rozwiązania w dziedzinie systemów automatycznego rozpoznawania mówcy opierają się na cechach dystynktywnych wynikających z budowy traktu głosowego mówcy (analiza dźwięku krtaniowego), zwanych cechami fizycznymi głosu. Powszechne wykorzystanie cech z tej grupy spowodowało spowolnienie możliwości dalszego rozwoju systemów ASR. Autor proponuje przełamanie impasu poprzez analizę i przetwarzanie behawioralnych cech sygnału mowy.W rozdziale 2 scharakteryzowano poszczególne moduły wchodzące w skład systemów automatycznego rozpoznawania mówcy, a także wskazano potencjalne obszary zastosowań charakterystyk behawioralnych.

2.

Proces przetwarzania sygnału mowy w aspekcie automatycznego rozpoznawania mówcy

Systemy ASR wykorzystywane przez organy porządkowe, branże wojskową czy cywilną różnią się między sobą strukturą, gdyż przeznaczone są do wykonywania innych zadań. W związku z tym niektóre etapy przetwarzania sygnału mowy będą odmienne w każdym z rozwiązań. Kryminalistyczne systemy wykorzystywane przez Policję operują na sygnałach, które zapisane są na nośnikach danych, a sam proces przetwarzania nie odbywa się w czasie rzeczywistym. Z kolei rozwiązania komercyjne wykorzystujące biometrię głosu czy sterowanie głosem (np. niektóre systemy bankowe) opierają się na obróbce sygnału, który generowany jest przez użytkownika w bieżącej chwili, a czas

odpowiedzi takiego systemu powinien być jak najkrótszy. Niezależnie od przeznaczenia systemu ASR, pierwszym etapem całego procesu przetwarzania sygnału mowy ludzkiej jest jego rejestracja. Ucho ludzkie do którego docierają dźwięki odbiera je jako sygnał analogowy. Z kolei w urządzeniach ten sam sygnał przyjmuje charakter cyfrowy w następstwie przetwarzania analogowo-cyfrowego. Kolejny etap obróbki mowy ludzkiej to odpowiednia filtracja sygnału, a następnie jego segmentacja. Na rys. 2.1 przedstawiono uogólniony schemat wstępnego przetwarzania sygnału mowy w projektowanym systemie ASR.



Rys. 2.1. Uogólniony schemat wstępnego przetwarzania sygnału mowy

2.1. Wstępne przetwarzanie mowy

2.1.1. Proces rejestracji i konwersji sygnału mowy

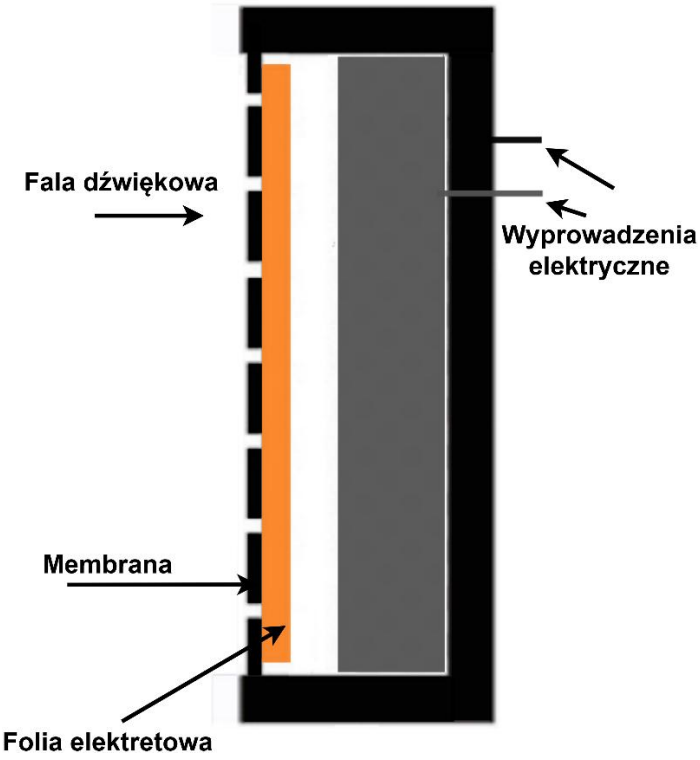
Niezależnie od zastosowanego systemu, w jego architekturze musi znajdować się element, który zamienia energię fali dźwiękowej na energię sygnału elektrycznego. Urządzenie takie nosi nazwę *mikrofonu* [148]. w zależności od zasady konwersji energii rozróżnia się następujące rodzaje mikrofonów [22]:

- stykowe (węglowe),
- dynamiczne (magnetyczne z ruchomą cewką),
- elektromagnetyczne,
- piezoelektryczne,
- pojemnościowe.

Dodatkowo ze względu na sposób poruszania membrany wyróżniamy:

- mikrofony ciśnieniowe
- mikrofony gradientowe.

Współcześnie najpowszechniej wykorzystywane są mikrofony pojemnościowe, a konkretniej rozwiązania z membraną wykonaną z folii elektretowej potocznie zwane *mikrofonami elektretowymi*. Ich popularność wzrosła za sprawą wprowadzania małych zniekształceń tłumieniowych, dużej skuteczności oraz dobrej stabilności działania [22]. Na rys. 2.2. przedstawiono konstrukcję mikrofonu elektretowego.



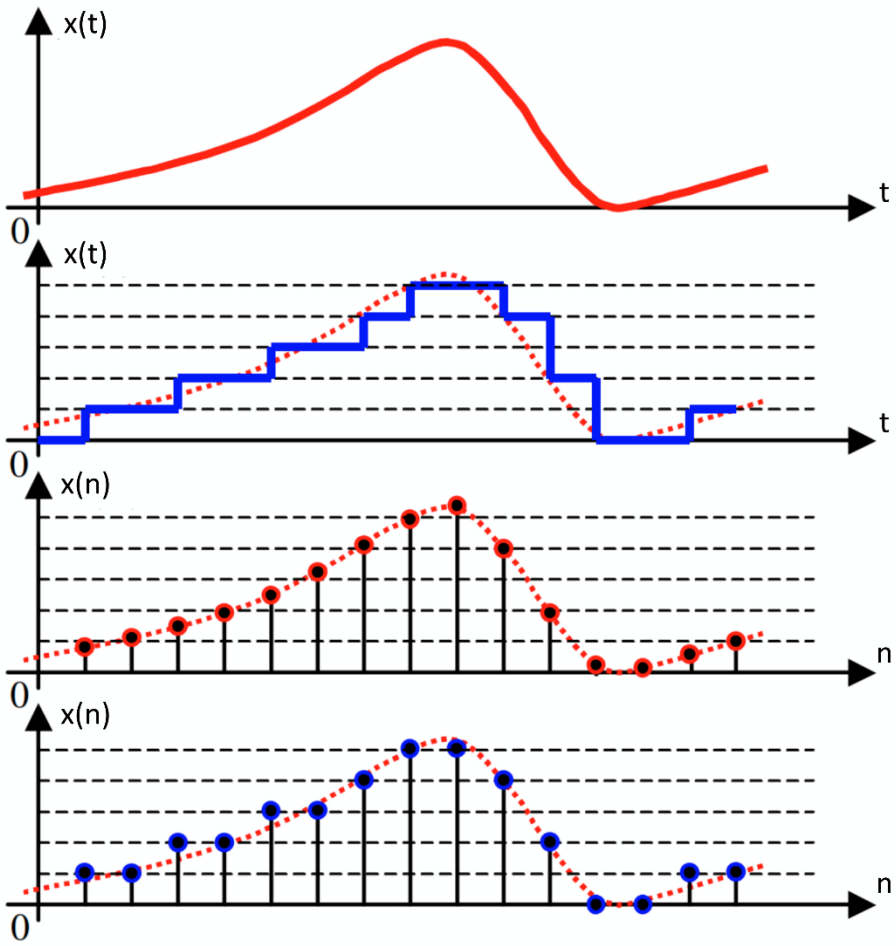
Rys. 2.2. Budowa mikrofonu elektretowego [153]

Zasada działania takiego urządzenia opiera się na wykorzystaniu zmian ciśnienia akustycznego. Ruchoma membrana wraz ze sztywną płytką wewnątrz tworzą kondensator (są jego okładzinami). Drgania membrany wywołane falą akustyczną docierającą do mikrofonu, wpływają na zmiany pojemności kondensatora, a w wyniku tego proporcjonalnie zmienia się wartość napięcia wyjściowego mikrofonu. Dodatkowo w celu zwiększenia wartości otrzymanego napięcia stosuje się wzmacniacze mikrofonowe.

Poza mową ludzką bądź zadanym dźwiękiem do mikrofonu dostają się różnego rodzaju zakłócenia, dlatego ważnym aspektem podczas pracy z takim urządzeniem jest minimalizacja wpływu czynników zewnętrznych. Kolejnym etapem po rejestracji i konwersji sygnału mowy jest jego zamiana z postaci analogowej na postać cyfrową.

2.1.2. Przetwarzanie analogowo-cyfrowe

W celu wykorzystania procedur cyfrowego przetwarzania sygnału i dalej eksploracji danych, sygnał elektryczny otrzymany na wyjściu mikrofonu musi zostać przetworzony za pomocą *przetwornika analogowo-cyfrowego* do postaci cyfrowej. Uogólniony schemat przetwarzania analogowo-cyfrowego opiera się na trzech etapach: próbkowaniu (przyporządkowanie wartości sygnału w ściśle określonych chwilach czasu – dyskretyzacja w czasie), kwantyzacji (sprowadzenie nieskończonego zbioru wartości sygnału do skończonego podzbioru – dyskretyzacja w wartościach) i kodowaniu. Na rys. 2.3 zobrazowano poszczególne rodzaje sygnałów w procesie przetwarzania A/C. Pierwsza z krzywych to sygnał analogowy (ciągły w czasie i wartościach). Kolejne to odpowiednio sygnały spróbkowany (dyskretny w czasie), skwantowany (dyskretny w wartościach) i cyfrowy (dyskretny w czasie i wartościach).



Rys. 2.3. Zobrazowanie sygnałów podczas procesu przetwarzania analogowo-cyfrowego.

Jednym z ważniejszych aspektów procesu konwersji sygnału analogowego do postaci cyfrowej jest dobór odpowiedniej szybkości próbkowania. Problem ten można przedstawić w formie pytania: *„Jak często należy „pobierać" (próbkować) wartości sygnału analogowego, aby jego dyskretna reprezentacja była jak najdokładniejsza, a co za tym idzie zapewniała możliwość pełnej rekonstrukcji sygnału analogowego w oparciu o dyskretne próbki?"* [2]. Odpowiedź na to pytanie zawarta jest w *twierdzeniu o próbkowaniu* zamiennie nazywanym *twierdzeniem Kotielnikowa-Shannona-Nyquista*. Zgodnie z nim sygnał ciągły w czasie i wartościach może być wiernie odtworzony z ciągu dyskretnych próbek wówczas, gdy szybkość próbkowania jest co najmniej dwa razy większa od częstotliwości najwyższej składowej harmonicznej występującej w próbkowanym sygnale.

Projektowany system ASR przeznaczony jest do zastosowania w warunkach powszechnie używanej transmisji telefonicznej. Trakt telefoniczny zaprojektowany został tak, aby możliwa była transmisja mowy przy jednoczesnym zachowaniu wysokiej wierności zapisu oraz oszczędności pasma. Szerokość pasma częstotliwości wykorzystywanej w telefonii do transmisji głosu wynosi 3 kHz. Powszechnie uważa się, iż mowa ludzka jest wystarczająco zrozumiała, gdy zawiera się ona w przedziale od 500 do 4000 Hz [83], [98]. Mając na względzie powyższe wartości w opracowywanym systemie przyjęto szybkość próbkowania równą 8 kS/s. Wartości sygnału reprezentowane są na 16 bitach w zapisie stałoprzecinkowym.

2.1.3. Standaryzacja sygnału mowy

Kolejny krok w procesie przetwarzania sygnału mowy, to jego standaryzacja, która polega na sprowadzeniu wartości średniej sygnału do zera, a odchylenia standardowego do jedności.

Powyższe operacje opisać można przy pomocy następujących równań:

$$x_i = \frac{x_i - \overline{x}}{\sqrt{s}},$$
$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad i = 1,2,...,N \tag{2.1}$$
$$s = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

gdzie: x_i to *i*-ta próbka sygnału, $\overline{\mathbf{x}}$ jest wartością średnią z próbek sygnału, a *s* jest wariancją z próby.

Konieczność usunięcia składowej stałej (wartości średniej) z sygnału związana jest niedoskonałością procesu rejestracji mowy ludzkiej przez mikrofon. Z fizycznego punktu widzenia sygnał mowy ludzkiej to następujące po sobie zmiany ciśnienia akustycznego o zerowej wartości średniej. Podczas konwersji analogowo-cyfrowej parametr ten jest prawie zawsze różny od zera, gdyż przetwarzane są fragmenty mowy o skończonej długości *N* [98].

Sprowadzenie wartości odchylenia standardowego do jedności można przyrównać do operacji skalowania, która eliminuje większy wpływ próbek o wartościach odstających od reszty (ang. *outliers*). Operacja ta realizowana jest zarówno podczas wyznaczania zbiorów uczącego, jak i testującego. W tym przypadku dla każdej cechy z osobna wyznaczane są wartości średnie i odchylenia standardowe.

2.1.4. Wycinanie ciszy

Następnym, nieodzownym elementem procesu wstępnego przetwarzania sygnału mowy jest wycinanie ciszy, która występuje w każdej wypowiedzi. Z punktu widzenia behawioralnych charakterystyk mowy ludzkiej cisza niesie ze sobą pewien zasób informacji dystynktywnej. Przykładowo, na podstawie relacji między czasem wypowiedzi, a ciszy w ramce możliwe jest wyznaczenie szybkości artykulacji mowy danej osoby. Dlatego w niniejszym systemie zdecydowano się na wycinanie tylko fragmentów ciszy dłuższych niż 2 sekundy. Sam proces detekcji fragmentów niezawierających mowy odbywa się przy pomocy kryterium energetycznego oraz rozproszenia widmowego (ang. *Audio Spectral Spread*). W pierwszym kroku dla sygnału mowy wyznaczana jest krótkoczasowa wartość energii (ang. *Short-Term Energy*) według zależności:

$$E_i = \sum_{j=0}^{N-1} x_j^2, \qquad j = 0,1,2,...,N-1 \tag{2.2}$$

Powyższy wzór opisuje ciąg operacji, gdzie przetwarzany sygnał mowy dzielony jest na nienakładające się na siebie ramki, a następnie w każdej z nich wyznaczana jest energia. Termin krótkoczasowa odnosi się krótkiego czasu trwania powyższych fragmentów, który w niniejszym systemie wynosi 30 ms.

Kolejny krok podczas wycinania ciszy to wyznaczenie wartości rozproszenia widmowego w każdej z ramek pierwotnego sygnału (dokładny opis tego parametru znajduję się w rozdz. 4).

Rozproszenie widmowe charakteryzuje rozkład mocy w widmie względem jego częstotliwości środkowej [127]. Dla *i*-tego fragmentu sygnału mowy, oblicza się je zgodnie z zależnością [2]:

$$SS_i = \sqrt{\frac{1}{P_c}\sum_{k=1}^{N}\left(f_k - SC_i\right)^2 \cdot P_k} \tag{2.3}$$

gdzie f_k to częstotliwość *k*-tej składowej widma, *SC* to wartość środka ciężkości widma, a P_k i P_c stanowią odpowiednio wartość mocy *k*-tej składowej widma i moc całkowitą w widmie.

Otrzymane wartości parametrów porównywane są z odpowiednimi progami E_t oraz SS_t i w przypadku przekroczenia obu progów dana ramka klasyfikowana jest jako zawierająca mowę, a fragmenty niespełniające tego kryterium są usuwane.

2.1.5. Segmentacja sygnału mowy

Segmentacja polega na podziale sygnału mowy na krótkie fragmenty, które noszą nazwę *ramek*. Mowa ludzka z punktu widzenia dziedziny przetwarzania sygnałów traktowana jest jak sygnał losowy, a dokładniej niestacjonarny proces stochastyczny. Niestacjonarność mówi o znaczącej zmianie właściwości sygnału w czasie. Podział takiej klasy sygnałów na krótkie, skończone fragmenty umożliwia ich przetwarzanie jako procesów quasi-stacjonarnych [98]. Dzięki temu każda z ramek traktowana jest jako odrębny sygnał charakteryzujący danego mówcę.

Operacja segmentacji mowy sprowadza się do *okienkowania,* czyli iloczynu próbek analizowanego sygnału z próbkami okna [135]. Okno czasowe $w(n)$ opisywane jest za pomocą szerokości Δt, która determinuje liczbę próbek podlegających mnożeniu, oraz przesunięcia τ, które określa sposób przesunięcia okienka na osi czasu. Ogólny wzór operacji okienkowania można opisać następująco:

$$x_w(n) = x(n) \cdot w(n), \qquad n = 0,1,2,\ldots,N-1 \tag{2.4}$$

Dodatkowo parametry charakteryzujące ramkę można wyrazić jako:

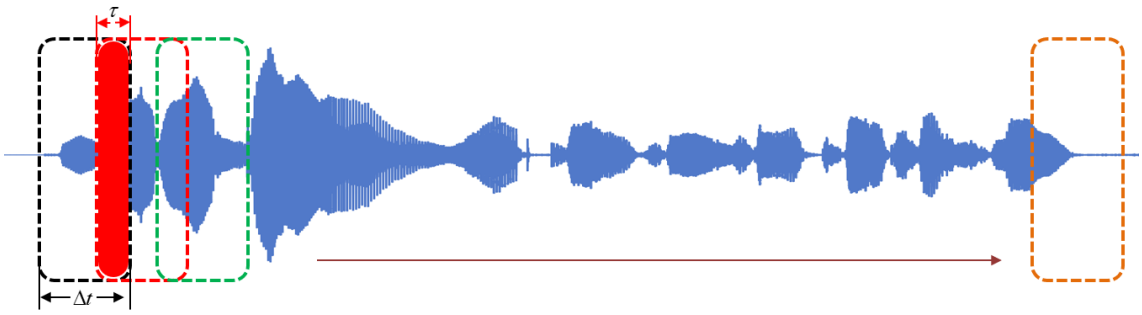
$$\begin{cases} \Delta t = T_p \cdot N_w \\ \tau = T_p \cdot N_s \end{cases} \tag{2.5}$$

gdzie T_p to częstotliwość próbkowania analizowanego sygnału, N_w to szerokość okna wyrażona w próbkach, a N_s stanowi wartość przesunięcia okna wyrażoną w próbkach.

Analiza mowy w dziedzinie czasu wiąże się ze znaczną nadmiarowością otrzymanych informacji. Aby ograniczyć niezbędne dane zawarte w sygnale mowy często stosuję się analizę częstotliwościową. Wspomniane wcześniej mnożenie próbek sygnału z próbkami okna w dziedzinie czasu jest tożsame ze splotem ich widm w dziedzinie częstotliwości. W związkuz tym, ważnym aspektem staje się wybór takiego okna czasowego, którego widmo charakteryzuję się odpowiednimi, pożądanymi parametrami. Najczęściej wykorzystywanym oknem jest okno prostokątne, które cechuje się wąskim „listkiem

głównym", ale niestety wysokim poziomem „listków bocznych". Inne, często stosowane okna, takie jak *Hanna, Blackmana* czy *Hamminga*, to funkcje będące złożeniem odpowiednich przebiegów harmonicznych [2], które stanowią swego rodzaju kompromis pomiędzy szerokością listka głównego i poziomem listków bocznych.

Poza wyborem rodzaju okna ważny jest również dobór jego szerokości i sposobu przesuwania wzdłuż sygnału. Systemy, których przeznaczeniem jest rozpoznawanie mowy charakteryzują się zmienną szerokością okna, zależną od zawartości semantycznej mowy. Wynika to z potrzeby analizy różnych składowych mowy ludzkiej takich jak: *fonemy, sylaby* czy *słowa.* W aspekcie systemów automatycznego rozpoznawania mówcy zazwyczaj stosuje się segmentację równomierną, tj. o stałej długości okna. W rozwiązaniach, w których głos mówcy modeluje się za pomocą HMM, szerokość okna wynosi 25 ms [124]. Dodatkowo często wykorzystywane jest nakładanie się sąsiadujących ze sobą fragmentów analizowanego sygnału, co pozwala uniknąć utraty cennych danych. Na rysunku 2.4 przedstawiono sposób segmentacji mowy.



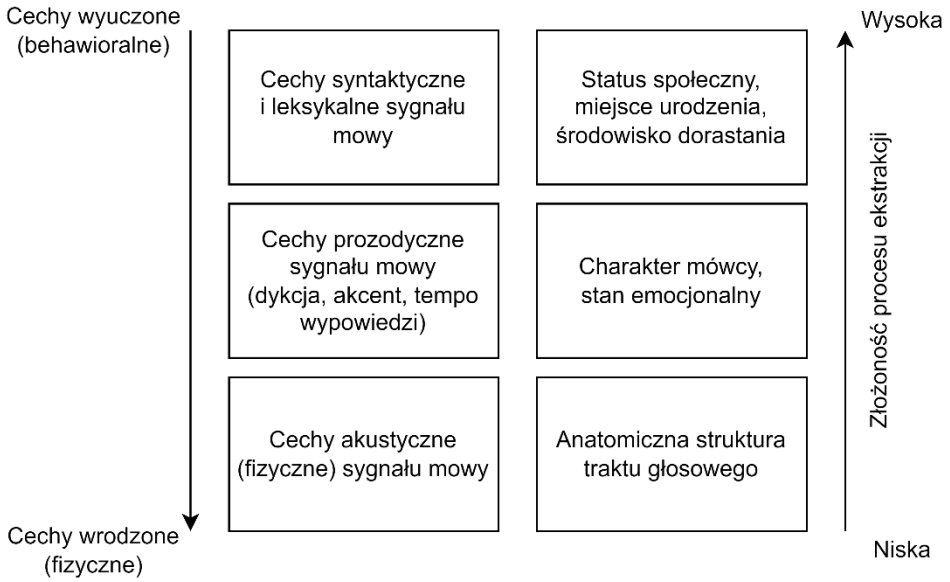
Rys. 2.4. Segmentacja analizowanego sygnału mowy wraz z nakładkowaniem.

Ustalenie wartości długości ramki oraz przesunięcia było jednym z pierwszych zadań podczas opracowywania systemu. Jako że długość fonemów jest zmienna, a możliwości detekcyjne ucha ludzkiego są ograniczone (możliwość klasyfikacji ramek o długości około 20 ms przez człowieka jest znikoma [79]) dobór szerokości okna nie należy do najprostszych. Małe wartości długości ramki oraz przesunięcia dają większą liczbę analizowanych fragmentów, co przekłada się na dokładniejszy opis sygnału mowy, ale jednocześnie większą ilość danych do przetworzenia, a to wydłuża czas obliczeń. Wartości długości ramki oraz przesunięcia zostały wyznaczone w procesie optymalizacji i opisane są w rozdziale 4.

2.2. Cechy zawarte w sygnale mowy

W rozdziale 1.2.3 krótko opisano podział cech sygnału mowy wykorzystywanych w systemach automatycznego rozpoznawania mówcy. Na rys 2.5 pokazano diagram, ilustrujący zbiór możliwych do ekstrakcji cech zawartych w mowie wraz z odpowiadającym im poziomem ekstrakcji. Po lewej stronie diagramu widać cechy związane z wypowiedzią mówcy i są to m.in.: szybkość artykulacji, intonacja oraz akcent. Z kolei prawa część skorelowana jest z samym mówcą. Znajdują się tu takie charakterystyki jak: wiek, płeć, status społeczny czy typ osobowości. Cechy najłatwiejsze do wyekstrahowania, związane są z brzmieniem głosu, co wynika ze struktury (budowy anatomicznej) traktu głosowego mówcy. Cały zbiór tych deskryptorów opisywany jest jako *cechy fizyczne*. Z kolei wyższe, a co za tym idzie trudniejsze w ekstrakcji

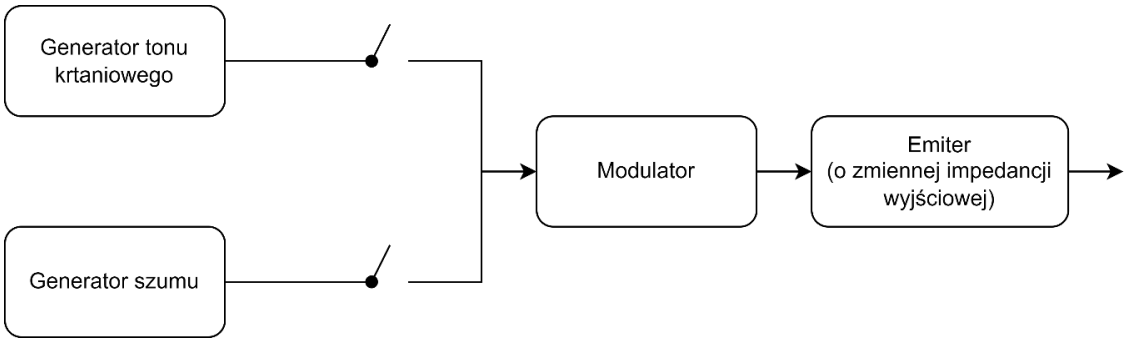
charakterystyki kształtują się w trakcie procesu dorastania i edukacji. Ich pochodzenie nie jest w żaden sposób związane z fizycznymi uwarunkowaniami mówcy, a z wpływem czynników socjoekonomicznych. W literaturze cechy te nazywane są *cechami wysokiego poziomu* bądź *cechami behawioralnymi*.



Rys. 2.5. Poziomy ekstrakcji cech zawartych w mowie ludzkiej [31]

2.2.1. Geneza pochodzenia cech fizycznych

Jako że cechy fizyczne związane są z budową traktu głosowego mówcy, znajomość anatomii oraz czynności narządów głosotwórczych są niezbędne do zrozumienia procesu ich generacji i analizy. Opis o charakterze biologicznym zawarty został w podrozdziale 1.1. Opierając się na nim można stworzyć model zastępczy traktu głosowego pokazany na rysunku 2.6 [107].



Rys. 2.6. Model zastępczy traktu głosowego

Sygnał mowy generowany jest przez generator tonu krtaniowego bądź szumu (naprzemiennie) w zależności od charakteru artykułowanej głoski (dźwięczna bądź bezdźwięczna). Następnie sygnał modulowany jest poprzez trakt głosowy, aby w końcowym etapie nastąpiła emisja dźwięku poprzez emiter o zmiennej impedancji [107]. Technicznie powstawanie głosu można przyrównać do procesu filtracji sygnału krtaniowego przez kanał głosowy o zmiennej transmitancji. Obciążeniem takiego kanału są usta, które bardzo często spełniają rolę głośnika niskotonowego. Charakterystyka tonu krtaniowego zależy od właściwości fałd głosowych takich jak ich rozmiar, napięcie,

sprężystość i masa. Parametry te są regulowane przez działanie mięśni krtaniowych, zwłaszcza mięśnia głosowego [31]. W swojej podstawowej formie, dźwięk z krtani jest zazwyczaj cichy i mało wyrazisty. Zyskuje na sile i barwie dopiero po przejściu przez trakt głosowy, który obejmuje takie struktury jak język, podniebienie, zęby, wargi, żuchwa oraz policzki. Niebagatelne znaczenie odgrywa również wielkie pudło rezonansowe jakim są płuca. Wnęki traktu głosowego działają jak wnęki rezonansowe. Widmo przepływającego przez głośnię powietrza modyfikowane jest przez trakt głosowy. Charakterystyka amplitudowo-częstotliwościowa traktu głosowego cechuje się kilkoma maksimami zwanymi formantami. Częstotliwości tych maksimów odpowiadają chwilowym częstotliwościom rezonansowym traktu głosowego, które zależne są od aktualnego stanu artykulacji. Przy wytwarzaniu dźwięków nosowych, zamknięta jama ustna służy jako bocznik akustyczny, a dźwięk jest emitowany przez jamę nosową i nozdrza. Przy odpowiednim ustawieniu podniebienia miękkiego jama nosowa pełni rolę stałego rezonatora.

Każda wypowiedź wytworzona zgodnie z powyższym opisem, jest nośnikiem informacji o wewnętrznej strukturze jej źródła (w sensie całego traktu głosowego). Odpowiednia analiza sygnału mowy pozwala zatem na ekstrakcje cech fizycznych skorelowanych z budową narządów mowy. Te osobnicze informacje odzwierciedlające indywidualne cechy głosu mówcy są wynikiem różnic w budowie anatomicznej różnych osób.

2.2.2. Przykłady cech fizycznych

Literatura naukowa w dziedzinie systemów ASR w przeważającej większości opisuje rozwiązania bazujące na cechach fizycznych. Tak samo jest w przypadku opracowanych wcześniej w WAT algorytmów rozpoznawania mówcy [31], [60]. Studiując powyższe publikacje możliwe jest przedstawienie kilka przykładowych charakterystyk związanych ze strukturą narządu głosu mówcy.

Opisany wyżej proces generacji sygnały mowy z punktu widzenia ekstrakcji cech osobniczych można traktować jako splot pobudzenia krtaniowego z odpowiedzią impulsową toru modelującego formanty. Separacja tych składników i odpowiednie, dalsze przetwarzanie skutkuje uzyskaniem pożądanych deskryptorów. W związku z multiplikatywnym charakterem tych dwóch składowych klasyczne metody wyznaczania widma nie znajdują zastosowania w procesie ekstrakcji cech. Dzieje się tak, gdyż spektrum sygnału mowy składa się z przebiegów wolnozmiennych związanych z traktem głosowym i szybkozmiennych impulsów dźwięku krtaniowego (niezbyt fortunnie *dźwięk krtaniowy* zwyczajowo nazywa się *tonem krtaniowym*). Uniemożliwia to jednoznaczne określenie miejsca rozpoczęcia składowych niosących informacje tylko o tonie krtaniowym. Zamiana multiplikatywnego charakteru splecionych elementów na addytywny sprawia, iż znacznie łatwiejsza staje się ich późniejsza separacja, a co za tym idzie ekstrakcja pożądanych parametrów. Możliwe jest to poprzez zastosowanie analizy *cepstralnej*. Jej istota – w kontekście przetwarzania sygnału mowy – została dokładniej opisana w rozdziale 3. Cechy dystynktywne uzyskiwane przy pomocy cepstrum, to m.in.: częstotliwość podstawowa mówcy będąca odwrotnością pierwszego maksimum cepstrum oraz kolejne maksima cepstrum rzeczywistego.

Jednym z najstarszych sposobów opisu sygnału mowy za pomocą deskryptorów liczbowych jest zastosowanie percepcyjnej predykcji liniowej (PLP – ang. *Perceptual*

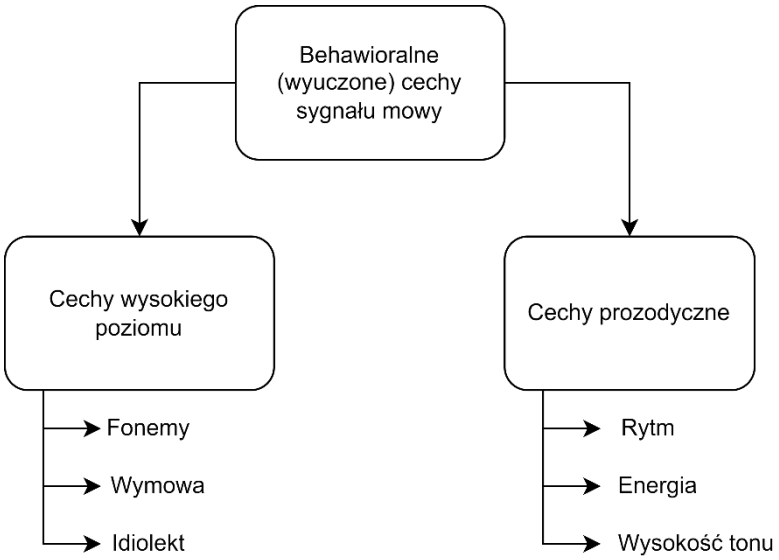
Linear Predictive) [42]. W procesie wyznaczania cech wykorzystuje się takie operacje jak preemfaza, segmentacja czy wyznaczanie wartości funkcji autokorelacji w ramce. Jednym z ostatnich etapów jest wyznaczenie współczynników predykcji liniowej LP, będących deskryptorami [19].

Dotychczas najpopularniejszą metodą parametryzacji sygnałów mowy jest operacja filtracji podpasmowej przy pomocy filtrów pasmowo-przepustowych, równomiernie rozłożonych na melowej skali częstotliwości. W literaturze naukowej metoda ta nosi nazwę *MFCC*. Zastosowanie filtrów o nakładających się pasmach pozwala na skuteczne rozróżnianie poszczególnych częstotliwości zgodnie z czułością narządu słuchu człowieka. Zmysł słuchu wykazuję możliwości rozróżniania dźwięków w zakresie od 20 do 16000 Hz, jednak największa jego czułość skupiona jest w paśmie od 1 do 5 kHz. Skala mel pozwala na relatywnie do ludzkiego odczucia liniowe odwzorowanie niskich częstotliwości. Wynikiem zastosowania tej metody jest otrzymanie *N*-1 deskryptorów stanowiących wartości współczynników melcepstralnych, gdzie *N* to liczba zastosowanych filtrów [60], [80].

Kolejna grupa parametrów możliwa do ekstrakcji to pierwsze oraz drugie pochodne cech wyznaczonych metodą MFCC. Noszą one nazwę *delta* (Δ) i *delta-delta* (Δ^2) *coefficients*. Są to różnice między sąsiednimi wektorami współczynników cepstralnych, ale w dziedzinie czasu [129]. Autor niniejszej rozprawy zaklasyfikował te parametry jako cechy fizyczne, dlatego nie były one przedmiotem badań.

2.2.3. Geneza pochodzenia cech behawioralnych

Już w późnych latach 90. XX wieku opisywano głos ludzki jako sygnał pochodzący z dwóch źródeł: fizycznego i wyuczonego (nabytego) [51]. Pierwsze z nich opisane zostało w podrozdziale 2.2.1. Przeważająca większość systemów ASR bazuje na cechach związanych ze strukturą aparatu mowy człowieka. Rozwiązania te cechują się wysoką skutecznością rozpoznawania mówcy przy zapewnionych odpowiednich warunkach oraz akceptowalnym poziomie szumu [111]. W przeciwnym przypadku sprawność zaczyna spadać. Związane jest to z podatnością cech fizycznych na negatywny wpływ szumu oraz jakość kanału akustycznego. Pokonanie tych przeciwności możliwe jest poprzez analizę i przetwarzanie behawioralnych cech sygnału mowy. Są to między innymi wyuczone nawyki czy styl mówienia [1]. Dokładniejszy podział pokazany jest na rys. 2.7. Deskryptory behawioralne można podzielić na niosące informacje o języku, dialekcie czy emocjach towarzyszących mówcy, ale także o rejonie i środowisku dorastania, kształcenia czy pracy.



Rys. 2.7. Podział cech behawioralnych [116]

2.2.4. Przykłady cech behawioralnych

Każdy mówca charakteryzuje się podobną strukturą traktu głosowego (różnice w tej strukturze są właśnie źródłem dystynktywnych cech fizycznych), ale wraz z biegiem lat człowiek kształtuje swój indywidualny sposób artykulacji. Zbiór nawyków i przyzwyczajeń w zakresie mowy stanowi cenny zbiór informacji dystynktywnych, które z powodzeniem mogą zostać wykorzystane w opracowywaniu systemów automatycznego rozpoznawania mówcy. Ten wyuczony zestaw zachowań można scharakteryzować za pomocą pokazanych wyżej cech.

Efektem końcowym artykulacji mowy jest powstanie najmniejszej, udźwiękowionej formy wypowiedzi zwanej *głoską*. W zależności od reguł gramatycznych i jej sąsiedztwa reprezentacja akustyczna może się znacząco różnić. Wrażenia dźwiękowe przyporządkowane fragmentom mowy o uniwersalnym znaczeniu noszą nazwę *fonemów* [98]. W języku polskim liczba fonemów nie jest jednoznacznie określona, a w zależności od źródła różnice są znaczące. Zgodnie z encyklopedią PWN, jest ich 32. Z kolei Panowie Esling i Gaylord wykazali istnienie 60 fonemów [47]. Tak duże rozbieżności związane są z przyjętym stopniem złożoności opisu języka, a także z nieusystematyzowanym podziałem fonemów na pomniejsze grupy. Wyróżnia się 2 główne podgrupy w ujęciu językowym, *spółgłoski* oraz *samogłoski*. Dalsze, bardziej szczegółowe podziały związane są z ułożeniem poszczególnych składowych traktu głosowego [98]. Z punktu widzenia cech behawioralnych charakteryzujących mówcę wyróżnić można główne cechy dystynktywne fonemów, takie jak *częstotliwość środkowa* czy *czas trwania*.

Zgodnie z encyklopedią PWN *idiolekt* to sposób wypowiedzi człowieka w konkretnym, aktualnie rozpatrywanym okresie rozwoju. Często zamiennie nazywany mową jednostkową, gdyż stanowi odmianę języka unikalną dla danej jednostki [152]. Idiolekt zawiera takie elementy jak słowa, konstrukcje gramatyczne czy odmienną wymowę. W związku z tym wykorzystanie tych składowych jako cech dystynktywnych może służyć do rozpoznawania mówcy. W swojej pracy G. Doddington zaproponował wykorzystanie *N-gramów,* czyli ustalonych sekwencji występujących w zasobach językowych [38].

Określa się je jako *unigramy, bigramy* bądź *trigramy*. Za cechę dystynktywną uznano wartość *entropii* określonego wcześniej *N*-gramu. Z kolei autorzy innego dzieła [13], na podstawie słów i głosek opisali kolejne parametry:

- wartość logarytmu z ilorazu liczby ramek przypadających na konkretne słowo,
- wartość logarytmu z liczby głosek przypadających na konkretne słowo,
- średnia długość dźwięcznych fragmentów w wyrazach,
- średnia długość bezdźwięcznych fragmentów w wyrazach.

Takie parametry mowy ludzkiej, jak jej głośność, rytm czy wysokość tonu również stanowią informację dystynktywną charakteryzującą mówcę. Na podstawie maniery wymowy Peskin, Navratil i inni stworzyli zbiór kilkunastu cech zaszeregowanych do dwóch grup [13]. Pierwsza z nich to deskryptory związane z długością oraz częstotliwością pauz:

- częstotliwość długich pauz,
- częstotliwość pauz średniej długości trwania,
- unormowana wartość logarytmu z długości średnich pauz względem długich pauz.

Druga grupa odnosi się do cech prozodycznych i zawiera takie parametry, jak:

- wartość logarytmu z wartości średniej częstotliwości podstawowej, uśrednionej w zbiorze wszystkich dźwięcznych ramek,
- wartość logarytmu z maksimum częstotliwości podstawowej w zbiorze wszystkich dźwięcznych ramek podzielonego przez liczbę wszystkich dźwięcznych ramek,
- wartość logarytmu z minimum częstotliwości podstawowej w zbiorze wszystkich dźwięcznych ramek podzielonego przez liczbę wszystkich dźwięcznych ramek,
- wartość logarytmu z różnicy maksimum i minimum dla przedziału częstotliwości podstawowej w zbiorze wszystkich dźwięcznych ramek podzielonego przez liczbę wszystkich dźwięcznych ramek.

2.3. Potencjalne możliwości systemów opartych na cechach fizycznych i behawioralnych

Systemy automatycznego rozpoznawania mówcy są co raz powszechniej wykorzystywane
w obszarach, gdzie wymagany jest wysoki stopień bezpieczeństwa przy jednoczesnym zachowaniu krótkiego czasu identyfikacji tożsamości człowieka. Jednak i te rozwiązania nie są pozbawione podatności. Główne zagrożenia, z którymi muszą się one mierzyć to:

- podszywanie się,
- występowanie głosów o dużym stopniu podobieństwa w korpusach[1],
- zmiany chorobowe narządów głosu,
- niewystarczająca ilość danych wynikająca z krótkich nagrań głosowych,
- negatywny wpływ toru akustycznego służącego do transmisji,
- zakłócenia występujące w procesie akwizycji mowy,

[1] Korpus (ang. *corpus*) – duży zbiór plików dźwiękowych zawierających nagrania języka mówionego zazwyczaj
z transkrypcją. Wykorzystywany jako zbiór danych w procesach przetwarzania języka naturalnego [154][1] Poprzez rozwiązanie referencyjne rozumieć należy system, z którym porównywana będzie autorska aplikacja i do którego zostanie zastosowane wsparcie z wykorzystaniem cech behawioralnych.

- stosowanie przedmiotów modyfikujących właściwości głosu, np. maseczek.

W zależności od wymagań, którym systemy ASR muszą sprostać, ich konstruktorzy stawiają na różnorakie podejścia projektowe. Systemy oparte na cechach fizycznych charakteryzują się wysoką skutecznością już przy relatywnie niewielkich zbiorach danych uczących i testujących. Z kolei algorytmy bazujące na deskryptorach behawioralnych wymagają dłuższych nagrań głosu, aby uwydatniły się informacje dystynktywne. Wszelkie zmiany chorobowe dotykające krtań silnie wpływają na własności cech fizycznych, przy jednoczesnym słabym oddziaływania na biometryki wyuczone. Dodatkowym atutem cech wysokiego poziomu, w odróżnieniu od tych związanych z budową traktu głosowego [50], jest ich odporność na szumy oraz zakłócenia, które występują w kanale transmisyjnym. Celowe zmiany parametrów głosu, takie jak szeptanie, podwyższony ton czy przyspieszone tempo artykulacji mają wpływ na sprawność systemów ASR. To samo tyczy się niecelowych deformacji charakteru mowy związanych z czynnikami stresowymi (np.: zaciśniecie szczęki) [97]. W tabeli 2.1 zaprezentowano w skondensowanej formie porównanie wpływu różnych zagrożeń na systemy oparte na fizycznych oraz behawioralnych cechach mowy.

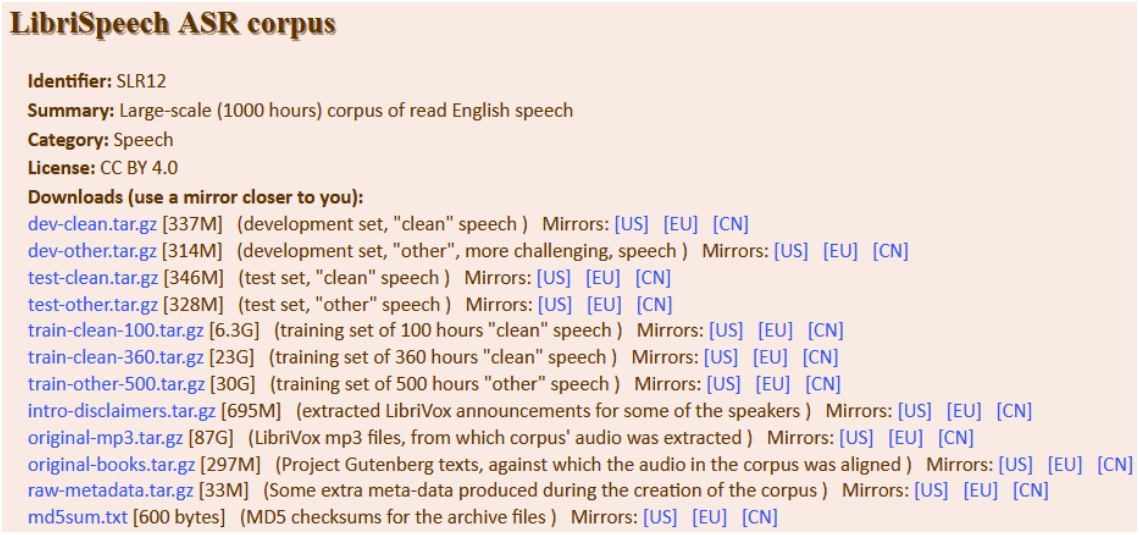
Tab. 2.1. Zestawienie wpływu zagrożeń na sprawność systemów ASR.

Rodzaj zagrożenia	System ASR oparty na cechach fizycznych	System ASR oparty na cechach behawioralnych	System ASR będący fuzją deskryptorów fizycznych i behawioralnych
Podszywanie się	Umiarkowany wpływ	Umiarkowany wpływ	Umiarkowany wpływ
Duży stopień podobieństwa głosów	Wysoki wpływ	Niski wpływ	Umiarkowany wpływ
Choroby dotykające krtani	Wysoki wpływ	Niski wpływ	Umiarkowany wpływ
Niska ilość danych uczących i testujących	Niski wpływ	Wysoki wpływ	Umiarkowany wpływ
Wpływ właściwości toru akustycznego	Wysoki wpływ	Niski wpływ	Umiarkowany wpływ
Zakłócenia i szumy	Wysoki wpływ	Umiarkowany wpływ	Umiarkowany wpływ
Czynniki zewnętrzne modyfikujące właściwości głosu	Umiarkowany wpływ	Umiarkowany wpływ	Umiarkowany wpływ

Biorąc pod uwagę tylko kryterium podatności na zagrożenia, systemy bazujące na behawioralnych cechach sygnału mowy są bardziej odporne niż systemy oparte na cechach fizycznych. Jednak uwzględniając również warunek łatwości przetwarzania mowy oraz długoczasowej stabilności ekstrahowanych cech, rozwiązania oparte na biometrykach wynikających ze struktury traktu głosowego wyglądają korzystniej. Dobór odpowiedniego sposobu parametryzacji mowy, a co za tym idzie całej konstrukcji systemu ASR zależy od stawianych wymagań. Nie jest możliwe stworzenie modeli głosów pozwalających na identyfikację mówcy z blisko 100% skutecznością przy jednoczesnym uodpornieniu ich na zewnętrzne zakłócenia, z wykorzystaniem tylko jednego zestawu cech. Kompromisem w tym względzie może być algorytm, uwzględniający fuzję deskryptorów fizycznych oraz behawioralnych. Dzięki takiemu podejściu możliwe jest uodpornienie systemu ASR na wpływ czynników zewnętrznych przy jednoczesnym zachowaniu wysokiej skuteczności działania.

2.4. Baza głosów zastosowana do badań – SLR-12

W dziedzinie systemów automatycznego rozpoznawania mówcy ważnym aspektem jest wybór odpowiedniego zbioru głosów. Taka baza danych musi być odpowiednio liczna, aby zapewnić dużą różnorodność mówców. Większość standaryzowanych korpusów jest licencjonowana, a co za tym idzie płatna. Z kolei warianty darmowe takie jak *Mozilla Common Voice* [150] są ogólnie dostępne, bardzo liczne (najnowszy zbiór to około 3300 unikalnych głosów), ale nie zawsze wystarczająco reprezentatywne. W bazie tej nagranie głosu każdego mówcy nie przekracza 10 sekund. Tak mała ilość danych może nie być wystarczająca do zaprojektowania skutecznego systemu automatycznego rozpoznawania mówcy. Dlatego też autor postanowił znaleźć taki korpus, który będzie wystarczająco zasobny w informacje przy jednoczesnym zapewnieniu swobodnego dostępu. W sieci istnieje witryna udostępniająca darmowe zasoby służące do tworzenia i testowania oprogramowania przetwarzającego mowę [157]. Autor wykorzystał zbiór danych oznaczony jako *SLR12*, a dokładniej *LibriSpeech ASR Corpus* (rys. 2.8).



LibriSpeech ASR corpus

Identifier: SLR12
Summary: Large-scale (1000 hours) corpus of read English speech
Category: Speech
License: CC BY 4.0
Downloads (use a mirror closer to you):
dev-clean.tar.gz [337M] (development set, "clean" speech) Mirrors: [US] [EU] [CN]
dev-other.tar.gz [314M] (development set, "other", more challenging, speech) Mirrors: [US] [EU] [CN]
test-clean.tar.gz [346M] (test set, "clean" speech) Mirrors: [US] [EU] [CN]
test-other.tar.gz [328M] (test set, "other" speech) Mirrors: [US] [EU] [CN]
train-clean-100.tar.gz [6.3G] (training set of 100 hours "clean" speech) Mirrors: [US] [EU] [CN]
train-clean-360.tar.gz [23G] (training set of 360 hours "clean" speech) Mirrors: [US] [EU] [CN]
train-other-500.tar.gz [30G] (training set of 500 hours "other" speech) Mirrors: [US] [EU] [CN]
intro-disclaimers.tar.gz [695M] (extracted LibriVox announcements for some of the speakers) Mirrors: [US] [EU] [CN]
original-mp3.tar.gz [87G] (LibriVox mp3 files, from which corpus' audio was extracted) Mirrors: [US] [EU] [CN]
original-books.tar.gz [297M] (Project Gutenberg texts, against which the audio in the corpus was aligned) Mirrors: [US] [EU] [CN]
raw-metadata.tar.gz [33M] (Some extra meta-data produced during the creation of the corpus) Mirrors: [US] [EU] [CN]
md5sum.txt [600 bytes] (MD5 checksums for the archive files) Mirrors: [US] [EU] [CN]

Rys. 2.8. Strona internetowa użytego korpusu

Baza jest pochodną projektu *LibriVox*, gdzie udostępnionych zostało 8000 audiobooków dających około 1000 godzin mowy czytanej w języku angielskim. Podczas tworzenia korpusu autorzy stosowali kilkuetapową normalizację dźwięku wraz z odpowiednim wyborem i segmentacją danych [138]. Do uczenia oraz testowania systemu będącego tematem tej rozprawy posłużono się podzbiorami oznaczonymi jako *train-clean-100.tar.gz* oraz *train-clean-360.tar.gz*. Pobrane zasoby to sumarycznie około 460 godzin nagrań, które po odpowiednim przetworzeniu utworzyły zbiór minimum minutowych próbek głosu 1172 unikalnych mówców. Proces przygotowania danych do wykorzystania w systemie ASR wspartym behawioralnymi cechami mowy obejmował następujące etapy:

- konwersję plików z *.flac na *.wav,
- standaryzację sygnałów dla każdego mówcy z osobna,
- przepróbkowanie sygnału do szybkości próbkowania 8 kS/s,
- połączenie nagrań jednego mówcy w ramach tych samych rozdziałów,
- usunięcie długich fragmentów ciszy,
- selekcja nagrań w taki sposób, aby uzyskać co najmniej minutę naturalnej mowy.

2.5. Podsumowanie - Uzasadnienie tezy rozprawy

W niniejszym rozdziale przedstawiono poszczególne etapy procesów akwizycji oraz wstępnego przetwarzania sygnału mowy. Opisano składowe występujące w większości systemów automatycznego rozpoznawania mówcy poczynając od rejestracji analogowego sygnału mowy, a na segmentacji ramek mowy kończąc. Dodatkowo scharakteryzowano cechy wykorzystywane do parametryzacji sygnału mowy, co z punktu widzenia tejże rozprawy jest jednym z ważniejszych aspektów. Przeważająca liczba rozwiązań w dziedzinie systemów rozpoznawania mówcy opiera się na cechach dystynktywnych wynikających z budowy traktu głosowego mówcy (analiza dźwięku krtaniowego), zwanych cechami fizycznymi głosu. Powszechne wykorzystanie tych charakterystyk spowodowało spowolnienie możliwości dalszego rozwoju systemów ASR. Pokonanie tych przeciwności możliwe jest poprzez analizę i przetwarzanie behawioralnych cech sygnału mowy [1].

W konsekwencji autor stawia tezę rozprawy w brzmieniu: ***„możliwe jest zdefiniowanie i wyekstrahowanie z sygnału mowy takich cech behawioralnych, które w połączeniu z cechami fizycznymi zaimplementowanymi w istniejącym rozwiązaniu znacząco poprawią jego skuteczność w obecności zewnętrznych zakłóceń”*.**

Udowodnienie tezy przedstawione zostanie w kolejnych rozdziałach rozprawy, kolejno poprzez:

- opis systemu odniesienia opartego na cechach fizycznych;
- propozycję samodzielnego systemu ASR opartego wyłącznie na cechach behawioralnych (w celu przebadania szerokiego zbioru cech i dokonania ich wstępnej selekcji);
- rozważenie różnych rodzajów fuzji obu systemów;
- badania eksploatacyjne oraz wskazanie najlepszego wariantu pracy systemu ASR i eksperymentalne potwierdzenie słuszności postawionej tezy.

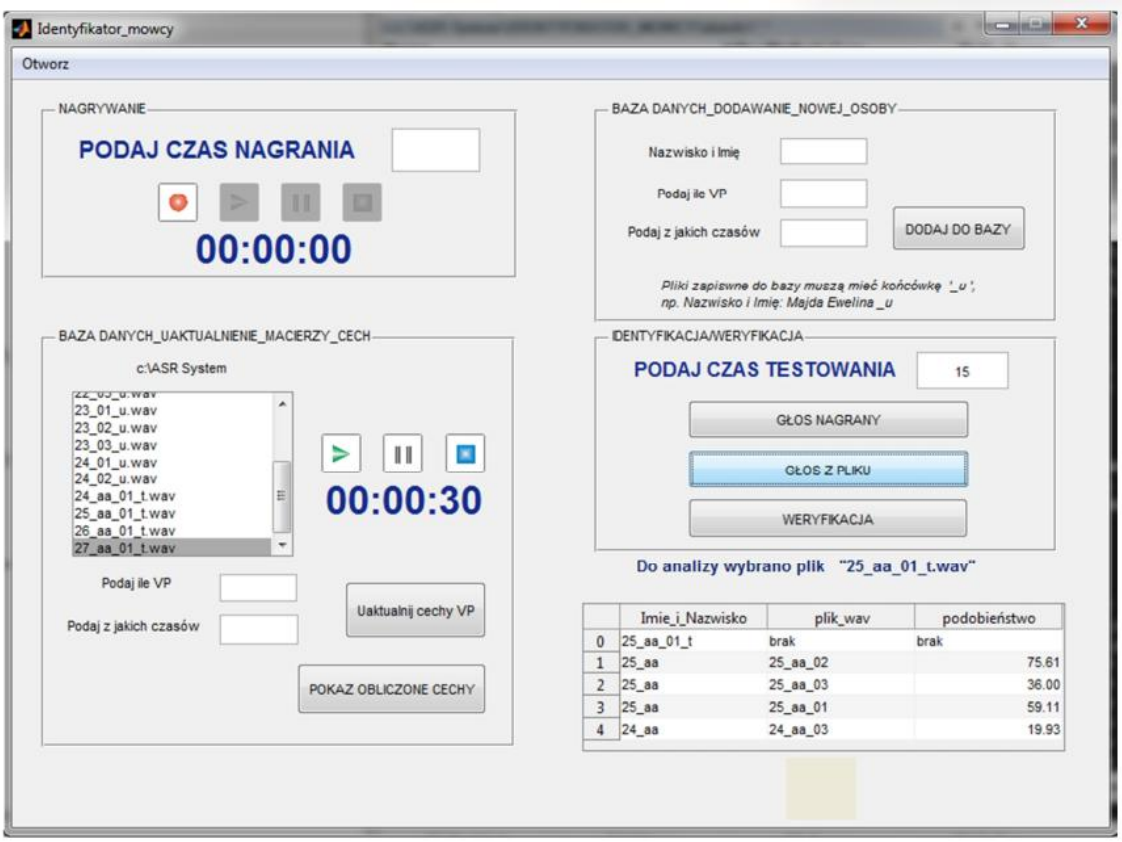
3.

System odniesienia oparty na cechach fizycznych

Aktualnie wiele zespołów badawczych zajmujących się zagadnieniami biometrii głosu opracowuje własne rozwiązania w dziedzinie systemów automatycznego rozpoznawania mówcy. W 2013 roku w Wydziale Elektroniki Wojskowej Akademii Technicznej powstał system ASR o nazwie ***Identyfikator mówcy***, którego twórcą jest dr inż. Ewelina Majda-Zdancewicz [31]. Okno główne programu pokazane jest na rysunku 3.1. Rozwiązanie to wyposażone zostało w kilka głównych modułów:

- nagrywania oraz rejestracji nowych użytkowników do bazy danych systemu,
- identyfikacji mówcy,
- weryfikacji mówcy,

- generacji modelu mówcy w postaci *odcisku głosu* – w systemie nosi on nazwę *Voice Print.*



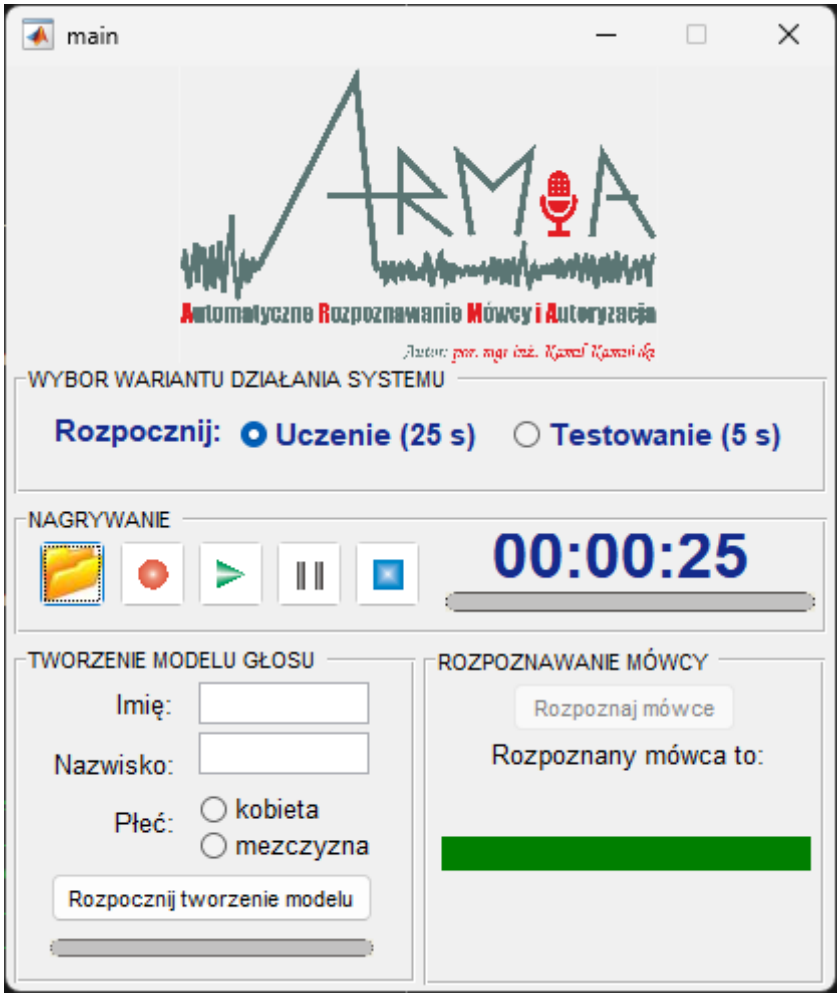
Rys. 3.1. Okno główne programu „Identyfikator mówcy" [31]

Algorytm ekstrakcji cech w głównej mierze bazuje na analizie cepstralnej sygnału mowy. System docelowo działa w konfiguracji 90/15, tj.: 90 sekund nagrań głosu stanowi segment uczący, a 15 sekund dane testujące. Możliwe jest również skrócenie ww. czasów. Autorka opracowała nowatorską możliwość wyboru metody identyfikacji. Umożliwia to przyszłemu użytkownikowi dokonanie wyboru między *identyfikacją binarną*, a *identyfikacją rankingową*. Ta pierwsza jednoznacznie określa zwycięską klasę, natomiast podejście rankingowe poszerza spektrum identyfikowanych klas do 4 najbardziej podobnych. Sprawność przy domyślnej ilości danych dostarczonych do systemu pracującego w trybie identyfikacji binarnej waha się w granicach 98% poprawnych identyfikacji tożsamości dla bazy 50 mówców [31].

Kolejnym osiągnięciem naukowców z WAT w dziedzinie automatycznego rozpoznawania mówcy jest opracowany w 2018 roku system ASR o akronimie ***ARMiA*** autorstwa mjr. dr. inż. Kamila Kamińskiego. ***A**utomatyczne **R**ozpoznawanie **M**ówcy **i A**utoryzacja* to aplikacja umożliwiająca identyfikację użytkownika w oparciu o próbkę jego głosu. Obie omawiane aplikacje powstały w wyniku realizacji prac doktorskich realizowanych pod opieką naukową prof. dr. hab. inż. Andrzeja Dobrowolskiego, który jest promotorem również niniejszej rozprawy, promotorem pomocniczym jest jednocześnie autor systemu *ARMiA*, mjr dr inż. Kamil Kamiński. Niniejsza rozprawa jest więc bezpośrednią kontynuacją prac rozpoczętych na Wydziale Elektroniki WAT w 2010 r.

Okno główne programu *ARMiA* pokazane jest na rys 3.2. Rozwiązanie może pracować w dwóch trybach pracy:

- *Uczenie* – w tym trybie na podstawie dostarczonej próbki głosu tworzony jest model głosu nowego mówcy, który następnie dodawany jest do bazy danych zawierającej modele głosu wszystkich dotychczas dodanych mówców. Do poprawnego utworzenia modelu głosu niezbędne jest dostarczenie 25 sekund nieprzerwanej mowy.
- *Testowanie* – pozwala na wskazanie mówcy z bazy danych, którego model głosu jest najbardziej podobny do głosu aktualnie badanego mówcy. Wykorzystywane jest tutaj 5 sekund mowy.



Rys. 3.2. Okno główne programu „*ARMiA*” [60]

Wymagające badania i żmudny proces optymalizacji parametrów systemu pozwoliły autorowi na uzyskanie wyników na poziomie 97% skutecznej identyfikacji w zbiorze 1160 mówców.

ARMiA to rozwiązanie stanowiące podstawę do prowadzenia dalszych badań w dziedzinie automatycznego rozpoznawania mówcy na Wydziale Elektroniki Wojskowej Akademii Technicznej. W dalszej części tego rozdziału przedstawiono dokładną charakterystykę najważniejszych modułów wchodzących w jego skład. Ich znajomość jest podstawą do zrozumienia dalszych rozważań i badań przeprowadzonych przez autora niniejszej rozprawy.

3.1. Wstępne przetwarzanie sygnału mowy

3.1.1. Normalizacja sygnału

Zgodnie z opisem przedstawionym w rozdziale drugim, w każdym systemie automatycznego rozpoznawania mówcy wyróżnić można moduły odpowiedzialne za przygotowanie sygnału mowy do dalszego przetwarzania. Często subtelne różnice w procesie wstępnego przetwarzania danych mogą znacząco wpłynąć na końcowy kształt opracowanego rozwiązania i osiągnięte rezultaty. Pierwszy etap obróbki sygnału w systemie *ARMiA* po jego rejestracji i konwersji A/C to normalizacja, czyli usunięcie wartości średniej oraz skalowanie [60]. Opisany we wcześniejszych rozdziałach proces standaryzacji jest podobny, ale różnica pojawia się w drugim kroku obróbki sygnału. Usunięcie wartości średniej realizowane jest w systemie zgodnie z poniższą zależnością

$$x(n) = x(n) - \frac{1}{N}\sum_{n=0}^{N-1} x(n) \tag{3.1}$$

Drugi krok normalizacji sygnału to jego skalowanie, czyli przeprowadzenie takich operacji, aby wartości próbek mieściły się w przedziale [0, 1]. Pozwala to na uwydatnienie cicho nagranych fragmentów mowy względem reszty sygnału. Operacja ta w rozwiązaniu *ARMiA* realizowana jest poprzez dzielenie każdej próbki przez wartość maksymalną sygnału (3.2) [60]

$$x(n) = \frac{x(n)}{\max\limits_{0<n<N-1}(|x(n)|)} \tag{3.2}$$

Autor dokonuje normalizacji również w trakcie przetwarzania ramek sygnału mowy, a konkretnie podczas ich selekcji. W tym przypadku wartość maksymalna wyznaczana jest tylko z próbek wchodzących w skład analizowanej ramki, a nie całego sygnału.

3.1.2. Wycinanie ciszy i filtracja sygnału

Kolejny krok wstępnego przetwarzania sygnału w systemie autorstwa dr. Kamińskiego to etap wycinania ciszy. Z punktu widzenia cech fizycznych ciche fragmenty mowy nie niosą ze sobą żadnej informacji dystynktywnej, dlatego zasadnym jest jak najdokładniejsze pozbycie się ich z analizowanego sygnału. W rozwiązaniu *ARMiA* realizowane jest to przy pomocy kryterium energetycznego. Energia wyznaczana jest w każdej ramce sygnału, a następnie otrzymana wartość jest skalowana względem ramki o największej energii [60], wg wzoru

$$E_i = \frac{1}{E_{\max}}\sum_{n=0}^{N-1} x^2(n) \tag{3.3}$$

Tak otrzymana wartość unormowanej energii sygnału jest porównywana z eksperymentalnie wyznaczonym progiem p_r który pomnożony przez średnią wartość energii z wszystkich ramek stanowi kryterium detekcji aktywności głosowej mówcy. Jako ramki zawierające sygnał mowy przyjmowane są ramki spełniające zależność

$$E_i > p_r \bar{E} \tag{3.4}$$

Dodatkowo w celu ograniczenia składowych o najniższych częstotliwościach (dźwięk niesłyszalny przez człowieka) autor stosował operację filtracji górnoprzepustowej.

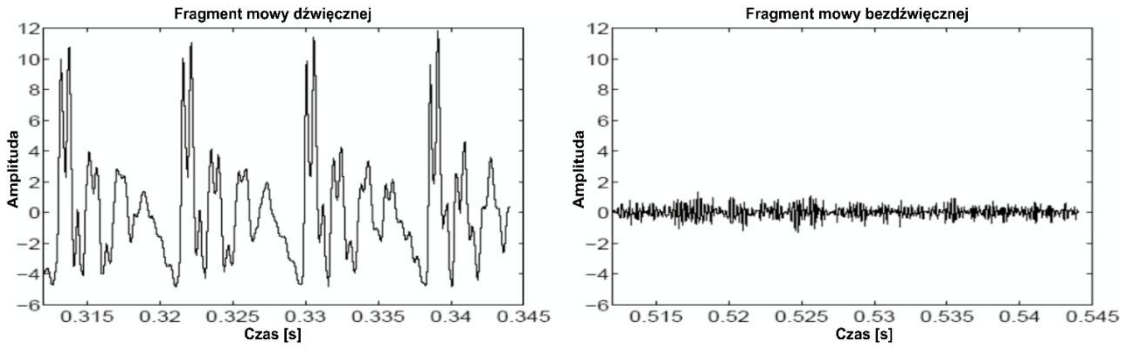
3.1.3. Segmentacja sygnału i selekcja ramek

Segmentacja sygnału mowy jest jednym z najważniejszych aspektów przetwarzania. Dokładne uwarunkowania dotyczące stacjonarności mowy ludzkiej i związane z tym ograniczenia zostały opisane w rozdziale 2. Z pojęciem segmentacji nieodłącznie kojarzy się termin okienkowania. Poprzez okienkowanie rozumiemy wymnożenie próbek sygnału przez próbki okna czasowego [135]. Iloczyn próbek w dziedzinie czasu tożsamy jest operacji splotu w dziedzinie częstotliwości. W związku z możliwością wystąpienia tzw. przecieków widma autor zdecydował się zminimalizować negatywny wpływ tego zjawiska poprzez zastosowanie okna Hamminga wyrażonego wzorem

$$w(n) = 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N}\right), \qquad gdzie \quad n = 0,1,2,...,N-1 \tag{3.4}$$

Ostatnim etapem wstępnego przetwarzania sygnału w systemie *ARMiA* jest proces selekcji ramek, czyli wyboru fragmentów najistotniejszych z punktu widzenia rozpoznawania mówcy. Pierwszy etap doboru ramek realizowany był poprzez funkcję wycinania ciszy, która eliminowała z dalszej analizy ciche fragmenty mowy. W końcowej wersji opracowanego algorytmu autor zawarł dodatkową, trójstopniową selekcję ramek [60].

Na początku dokonywana jest detekcja dźwięczności analizowanych fragmentów sygnału mowy. Wiążę się to z założeniem, że fizyczne cechy dystynktywne zawarte są w tonie krtaniowym, który występuje tylko w udźwięcznionych częściach mowy. Dźwięczność z punktu widzenia analizy w dziedzinie czasu można scharakteryzować jako występowanie pewnych okresowych wibracji, natomiast przebieg mowy bezdźwięcznej przypomina sygnał szumowy, co zilustrowano na rys. 3.3.



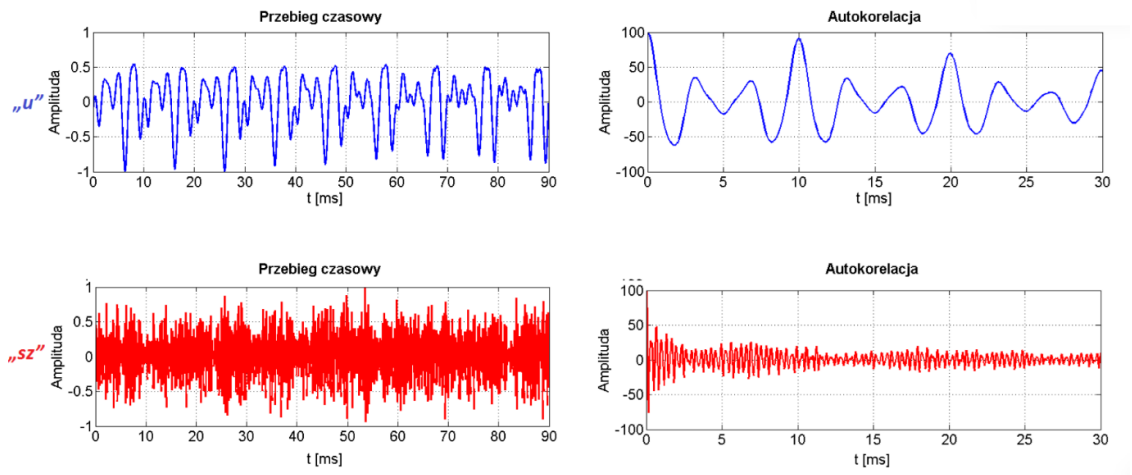
Rys. 3.3. Fragmenty mowy dźwięcznej i bezdźwięcznej [76]

Aby możliwa była klasyfikacja dźwięczności analizowanych ramek sygnału autor zastosował funkcję autokorelacji [135] opisaną zależnością

$$r(i) = \sum_{n=0}^{N-1} s(n)s(n+i), \qquad gdzie \quad i = 0,1,2,.... \tag{3.5}$$

gdzie *s(n)* to wartości próbek w ramce sygnału mowy.

Na rys. 3.4 pokazane zostały przebiegi czasowe głoski dźwięcznej „u” oraz bezdźwięcznej „sz” wraz z ich funkcjami autokorelacji [60]



Rys. 3.4. Przebiegi czasowe i funkcje autokorelacji wypowiedzi głoski „u” oraz „sz” [60]

Okresowe występowanie maksimów r_{max} funkcji autokorelacji umożliwia określenie dźwięczności badanej ramki. Najwyższa wartość funkcji autokorelacji widocznej na rys. 3.5. występuje przy zerowym przesunięciu ramki. Jednak to ekstremum związane jest z energią sygnału, dlatego podczas detekcji udźwięcznionych fragmentów mowy należy poszukiwać drugiego maksimum. Autor systemu *ARMiA* po znalezieniu powyższego ekstremum porównuje je z eksperymentalnie wyznaczonym progiem dźwięczności (analogicznie jak na etapie wycinania ciszy) [60].

Drugim etapem selekcji ramek jest ponowna detekcja występowania mowy w sygnale głosowym. Jednak w odróżnieniu od procesu wycinania ciszy tutaj usuwane są znacznie krótsze fragmenty sygnału (równe długości ramki stosowanej do ekstrakcji cech osobniczych). W tym przypadku autor wykorzystał nierówność (3.6) z empirycznie wyznaczonym progiem dla kryterium mocowego

$$P_i > p_p P_s \tag{3.6}$$

gdzie P_i to moc analizowanej i-tej ramki sygnału, p_p jest empirycznie wyznaczoną wartością, a P_s statystyczną wartością średniej mocy ramki [60].

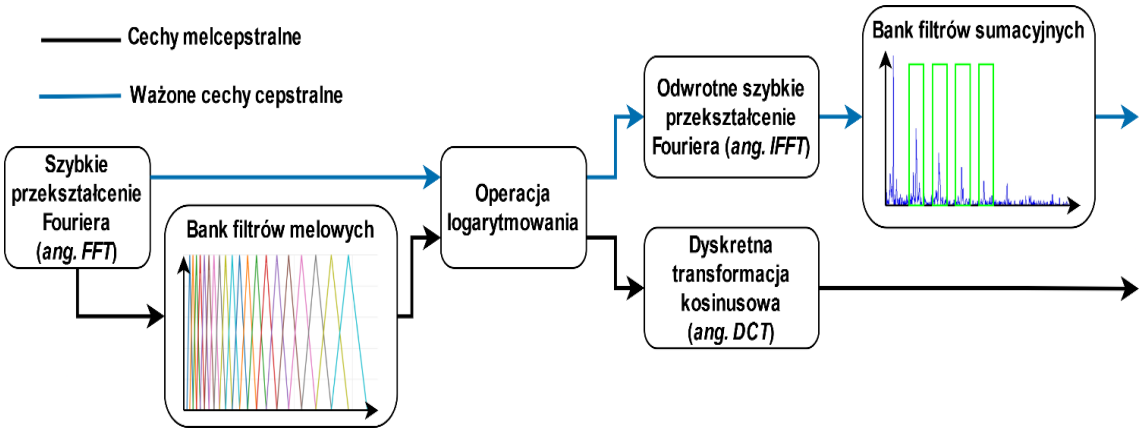
Ostatni z trzech etapów selekcji ramek to proces sprawdzenia ich zaszumienia. Odbywa się to poprzez porównanie wartości częstotliwości podstawowej wyznaczonej za pomocą metod autokorelacyjnej oraz cepstralnej. W związku z różnym poziomem odporności tych metod na szumy (wartość częstotliwości podstawowej wyznaczona metodą autokorelacyjną jest dokładniejsza, ale bardziej podatna na wpływ szumu), różnicę między otrzymanymi wartościami F_0 traktować można jako parametr liczbowy. Przy porównaniu go z wartością progową, całe wyrażenie staje się kryterium jakościowym badanej ramki. Opisane kryterium można przedstawić za pomocą nierówności

$$\left|F_{0c} - F_{0ac}\right| \le p_f \min\left(F_{0c} - F_{0ac}\right) \tag{3.7}$$

gdzie p_f to empirycznie wyznaczony próg.

3.2. Metody parametryzacji sygnału mowy

Poprzez parametryzację sygnału mowy rozumieć należy proces generacji cech osobniczych mówcy, czyli deskryptorów numerycznych, które możliwie najdokładniej reprezentują jego głos. Z punktu widzenia rozpoznawania mówcy proces parametryzacji mowy to przekształcenie przebiegu czasowego mowy w możliwie jak najmniejszą reprezentację liczbową zawierającą najistotniejsze informacje opisujące głos badanego mówcy. Dodatkowym wymaganiem jest zapewnienie jak najmniejszej wrażliwości na zmienność sygnału, tzn., że należy w jak największym stopniu uniezależnić ekstrahowane cechy od treści wypowiedzi czy wpływu toru akustycznego. Autor w trakcie tworzenia systemu *ARMiA* bazował na analizie częstotliwościowej sygnału mowy, która podlegała dalszym modyfikacjom. W końcowej formie algorytmu do parametryzacji mowy zastosowane zostały dwa zestawy deskryptorów numerycznych, tj. *ważone cechy cepstralne* oraz *cechy melcepstralne* [60]. Na rys. 3.5 przedstawiono ogólny schemat procesu generacji cech dystynktywnych.



Rys. 3.5. Uogólniony schemat generacji cech w systemie *ARMiA* [60]

3.2.1. Analiza widmowa

Wynikiem operacji opisanych w rozdziałach 2. oraz 3.1, jest sygnał w postaci cyfrowej $s(n)$, dla $n = 0, 1, 2, \ldots, N\text{-}1$, gdzie N to długość omawianej wcześniej ramki. Tak otrzymane fragmenty mowy należy przekształcić do dziedziny częstotliwości. Aby było to możliwe niezbędne jest zastosowanie dyskretnej transformacji Fouriera (ang. *DFT*) wyrażonej wzorem [2]:

$$S(k) = \sum_{n=0}^{N-1} s(n) e^{-j2\pi \frac{kn}{N}}, \qquad gdzie \;\; k,n = 0,1,2,\ldots,N-1 \tag{3.8}$$

Wynikiem tego przekształcenia jest wektor wyjściowy $S(k)$ będący częstotliwościową reprezentacją sygnału $x(n)$. W związku z powszechnym wykorzystaniem algorytmu szybkiej transformacji Fouriera (ang. *FFT*) w praktyce dąży się do uzyskania długości sygnału (liczby próbek) będącej całkowitą wielokrotnością liczby 2. Pozwala to na znaczącą redukcję liczby operacji wykonywanych w procesie transformacji [2], [135]. Badania prowadzone przez dr. Kamińskiego wykazały, iż wydobycie cech dystynktywnych głosu na podstawie samej obwiedni widma jest trudnym zadaniem. Dlatego zdecydował on się na rozszerzenie analizy sygnału poprzez transformację do tzw. cepstrum [60].

3.2.2. Analiza cepstralna

W latach 60. ubiegłego wieku Tukey, Bogert i Healy opracowali technikę przekształcenia sygnałów w pewne przestrzenie wektorowe, gdzie operacje mnożenia zastępowane są dodawaniem [12]. Rozwiązanie to należy do rodziny technik homomorficznych, czyli takich, które usiłują rozdzielać nieliniowo połączone sygnały w strukturę, która powstałaby przy połączeniu tych sygnałów w systemie liniowym. Twórcy metody cepstralnej stworzyli również specyficzną terminologię związaną z technikami homomorficznymi. Nowe słowa powstały na skutek odwrócenia kolejności pierwszych liter wybranych pojęć powszechnie używanych i związanych z analizą widmową. W tabeli 3.1 pokazane są powyższe zamiany.

Tab. 3.1. Nomenklatura stosowana w dziedzinach analizy widmowej i cepstralnej

Terminologia w dziedzinie analizy widmowej	Terminologia w dziedzinie analizy cepstralnej
SPECTRUM	**CEPS**TRUM
FREQUENCY	**QUEF**RENCY
HARMONICS	**RAH**MONICS
FILTERING	**LIF**TERING

Cepstrum zespolone wyrazić można jako wynik odwrotnego przekształcenia Fouriera logarytmu z widma sygnału czasowego otrzymanego poprzez prostą transformację Fouriera (3.9).

$$c(t) = IDFT\left\{\ln\left(DFT\left[x(t)\right]\right)\right\} \tag{3.9}$$

Powyższa zależność zakłada obliczenie logarytmu zespolonego, co wiąże się z koniecznością zapewnienia ciągłości fazy [2]. W przypadku mowy zasadnicze informacje zawarte są w amplitudzie widma. Dlatego do zastosowań rozpoznawania mówcy stosuje się tzw. *cepstrum rzeczywiste*. Dla sygnałów dyskretnych sprowadza się ono do postaci

$$c(n) = IDFT\left\{\ln\left(\left|DFT\left[x(n)\right]\right|\right)\right\} \tag{3.10}$$

Mając na uwadze funkcje okna *w(n)* oraz jądra przekształcenia Fouriera, ostatecznie cepstrum rzeczywiste można zapisać jako

$$c(n) = \frac{1}{N}\sum_{n=0}^{N-1} C(k) e^{j2\pi\frac{kn}{N}} = \frac{1}{N}\sum_{k=0}^{N-1} \ln\left(\left|\sum_{n=0}^{N-1} x(n)w(n)e^{-j2\pi\frac{kn}{N}}\right|\right) e^{j2\pi\frac{kn}{N}} \tag{3.11}$$

Jako że jądro transformaty Fouriera jest okresowe, logarytm z modułu widma amplitudowego *C(k)* również jest okresowy i spełnia zależność

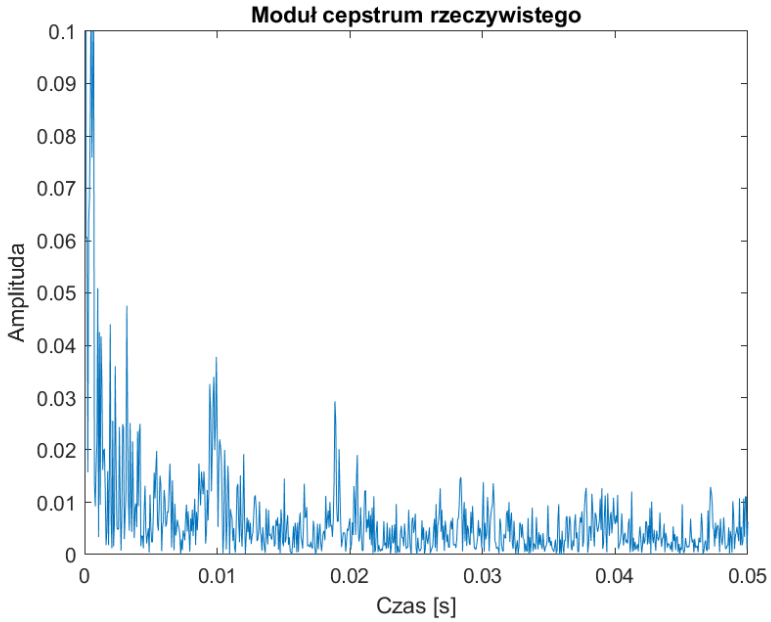
$$C(k) = C(-k) = C(N-k) \tag{3.12}$$

Z powyższego wyrażenia wynika, że jest to funkcja parzysta, więc w jej rozwinięciu występują tylko składowe kosinusoidalne. Dzięki temu możliwa jest interpretacja cepstrum rzeczywistego jako widma zlogarytmowanego spektrum amplitudowego [7].

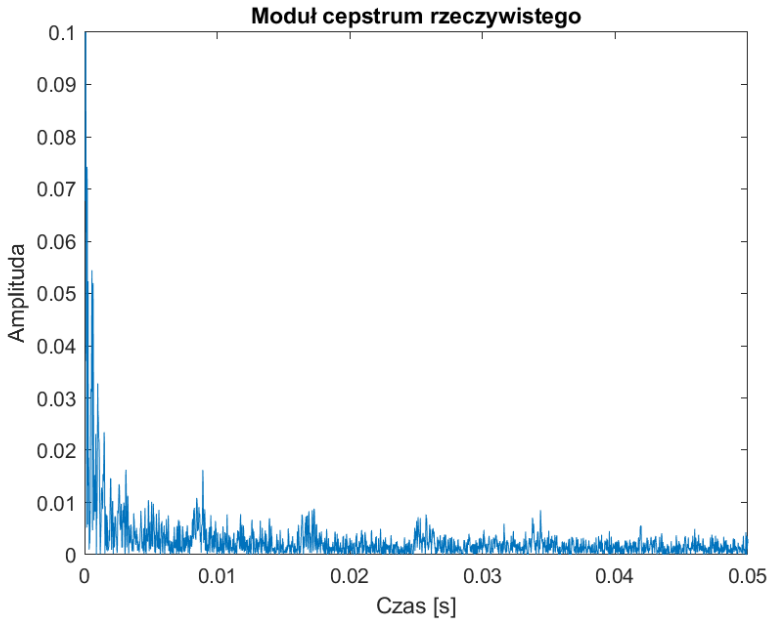
Analizując widmo sygnału mowy można wysnuć wniosek, że złożone jest ono ze składowych szybko i wolnozmiennej. Pierwsza z nich wynika z pobudzenia, a druga to czynnik modulujący amplitudę poszczególnych impulsów powstałych w wyniku tego pobudzenia. Poprzez logarytmowanie widma amplitudowego możliwe jest

przekształcenie zależności multiplikatywnych tych składowych na związki addytywne. Przez to wolnozmienne przebiegi związane z transmitancją traktu głosowego położone są w pobliżu zera na osi *pseudoczasu* (pseudoczas stanowi dziedzinę analizy cepstralnej), a impulsy pochodzące od dźwięku krtaniowego mają swój początek w okolicach jego okresu i cyklicznie powtarzają się co okres.

Analiza cepstralna sygnału mowy ułatwia detekcje dźwięcznych i bezdźwięcznych fragmentów mowy. Związane jest to z występowaniem tonu krtaniowego przy artykulacji głosek dźwięcznych, co na przebiegu modułu cepstrum rzeczywistego widoczne jest w postaci charakterystycznych maksimów (rys. 3.6). W przypadku bezdźwięcznych fragmentów mowy nie obserwuje się dźwięku krtaniowego, czyli „pików” (rys. 3.7).



Rys. 3.6. Moduł cepstrum rzeczywistego wypowiadanej głoski „a”

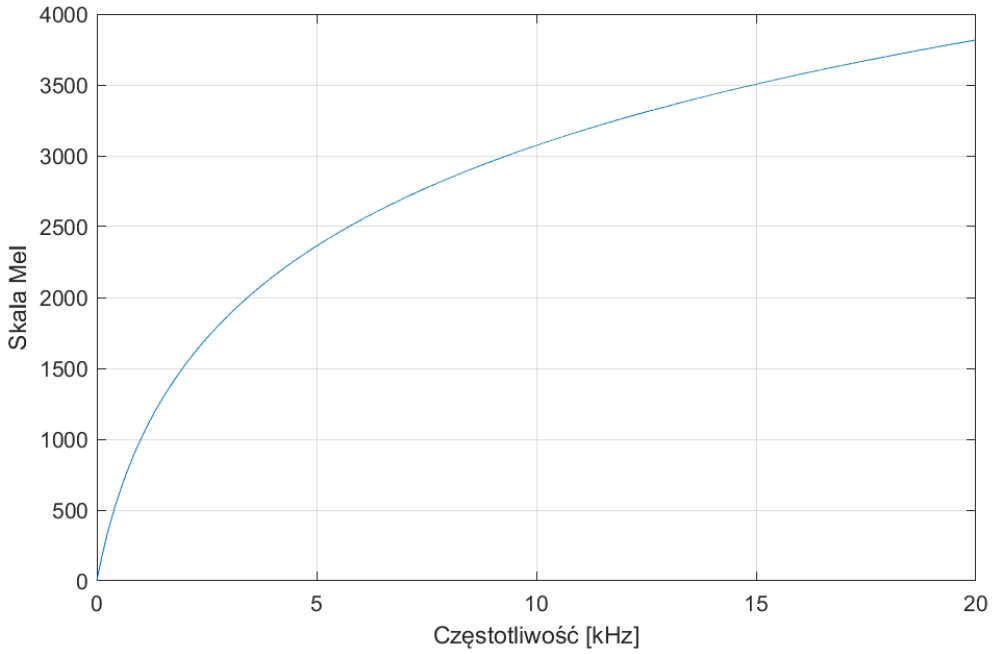


Rys. 3.7. Moduł cepstrum rzeczywistego wypowiadanej głoski „t”

3.2.3. Analiza melcepstralna

Ostatni sposób parametryzacji mowy w systemie *ARMiA* to zastosowanie metody MFCC [60]. Jest to proces filtracji podpasmowej sygnału za pomocą filtrów pasmowo-przepustowych równomiernie rozłożonych w skali Mel [129]. Podyktowane jest to naśladowaniem mechanizmów rozpoznawania głosu przez zmysł słuchu człowieka, który analizuje w sposób nieliniowy widmo sygnału. Dzieje się tak, gdyż ucho ludzkie wykazuje możliwości rozróżniania dźwięków w zakresie od 20 do 16000 Hz, jednak największa jego czułość skupiona jest w paśmie od 1 do 5 kHz. Melowa skala częstotliwości pozwala na liniowe odwzorowanie częstotliwości niskich, natomiast wysokie odwzorowuje w sposób logarytmiczny [2]. W skali tej jednostką częstotliwości jest *mel (*ang. ***mel****ody)*, którego związek z liniową skalą częstotliwości określa równanie (3.13) [25]. Graficzne przedstawienie tej zależności przedstawiono na rys. 3.8.

$$f_{mel}\left(f_{Hz}\right)=2595\log_{10}\left(1+\frac{f_{Hz}}{700}\right) \tag{3.13}$$



Rys. 3.8. Zależność skali melowej i skali liniowej

Wykorzystane w metodzie MFCC filtry mają trójkątne (rzadziej trapezowe) charakterystyki amplitudowe i zdefiniowane są w dziedzinie częstotliwości co umożliwia ich wymnożenie z widmem analizowanego sygnału mowy. Takie przekształcenie spektrum realizowane jest przy pomocy banku *M* filtrów. Następuje transformacja *N*-punktowego widma ramki w *M*-punktowe, przefiltrowane widmo według poniższego wzoru

$$X_i=\sum_{k=1}^{N-1}\left(\left|S(k)\right|\cdot F_i(k)\right), \qquad i=1,2,\ldots,M \tag{3.14}$$

gdzie $F_i(k)$ to funkcja filtru w *i*-tym podpaśmie.

Następnie amplituda widma przeliczana jest ze skali liniowej do logarytmicznej (3.15), co pozwala na separacje splecionych ze sobą pobudzenia krtaniowego i odpowiedzi impulsowej traktu głosowego.

$$X_{Li} = \log X_i, \qquad i = 1,2,...,M \tag{3.15}$$

Ostatni etap przetwarzania w metodzie MFCC to proces dekorelacji otrzymanych cech przy pomocy dyskretnej transformacji kosinusowej (ang. *DCT*), zgodnie z poniższą zależnością

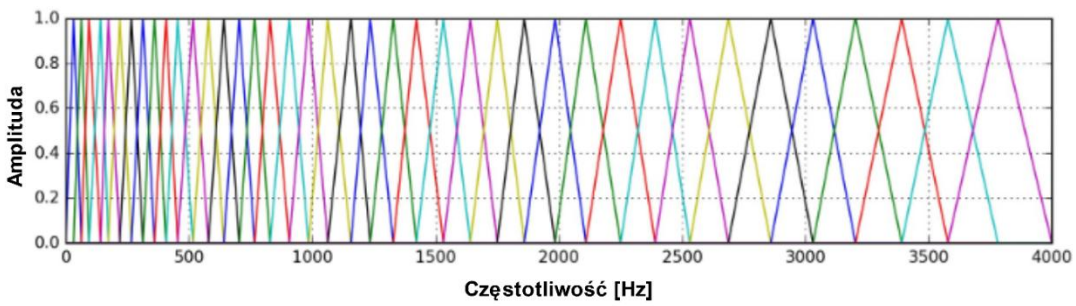
$$MFCC_j = \sum_{i=1}^{M} X_{Li} \cos\left[j\left(i - \frac{1}{2} \right) \frac{\pi}{M} \right], \qquad j = 0,1,2,...,J \tag{3.16}$$

gdzie $MFCC_j$ to *j*-ty współczynnik melcepstralny, *M* stanowi liczbę filtrów melowych, a *J* liczbę współczynników MFCC. W rozwiązaniu *ARMiA* zarówno *M* jak i *J* równe są 30 [60]. Idea wykorzystania współczynników melcepstralnych w tym systemie opiera się na szczegółowej procedurze określania zakresów pasm (3.17) oraz częstotliwości środkowych projektowanych filtrów melowych (3.18).

$$\Delta f = \frac{f_g - f_d}{M + 1} \tag{3.17}$$

$$f_{sr_i} = f_d + i \cdot \Delta f, \qquad i = 1,2,...,M-1 \tag{3.18}$$

gdzie f_d i f_g stanowią odpowiednio częstotliwości dolną i górną sygnału mowy, a *M* oznacza liczbę filtrów melowych. Autor zaprojektował filtry w taki sposób, aby ich pasma nakładały się na siebie w pobliżu częstotliwości środkowej filtru. Inaczej mówiąc każdy kolejny filtr rozpoczyna się na częstotliwości środkowej filtru poprzedniego [60]. Na rys. 3.9 zobrazowano przykładowy rozkład pasm filtrów melowych.



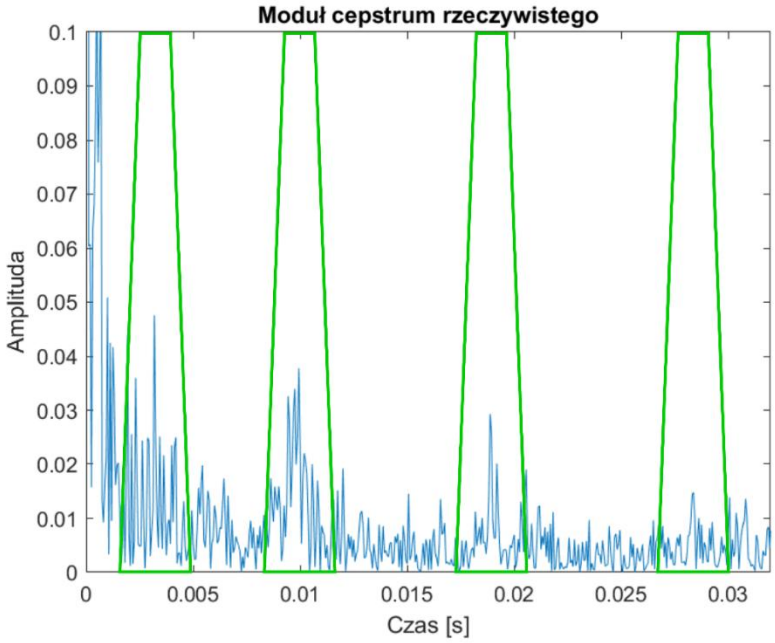
Rys. 3.9. Rozkład pasm filtrów melowych [60]

3.3. Cechy fizyczne wykorzystane w systemie

Autor w zaprojektowanym rozwiązaniu wykorzystuje także *ważone cechy cepstralne*, będące modyfikacją metody MFCC. Jest to zestaw deskryptorów w przestrzeni cepstrum powstałych na skutek zastosowania w poszczególnych podpasmach filtrów sumacyjnych. Poprzez taką filtracje rozumieć można sumowanie wszystkich punktów znajdujących się w paśmie filtru z pewną wagą [60]. Następna operacja to normalizacja otrzymanych sum względem wartości sumy wokół pierwszego ekstremum, które odpowiada częstotliwości podstawowej. W systemie *ARMiA* wykorzystano trapezową funkcję wagową, co pokazano na rys. 3.10 [60]. Sposób wyznaczania ważonych cech cepstralnych opisać można za pomocą następujących zależności

$$\begin{cases} c_{w_1} = \sum_{k=-dp}^{dp} c_{p(1,k)} \\ c_{w_i} = \dfrac{\sum_{k=-dp}^{dp} c_{p(i,k)}}{c_{w_1}} \end{cases}, \qquad i = 2,3,\ldots,J \tag{3.19}$$

gdzie c_{wi} to *i*-ta ważona cecha cepstralna, $c_{p(i,k)}$ stanowi wartość cepstrum rzeczywistego w sąsiedztwie *i*-tego maksimum w obszarze sumowania, *dp* to zakres sumowania w obrębie *i*-tego maksimum, a *J* jest liczbą ważonych cech cepstralnych.



Rys. 3.10. Graficzne zobrazowanie sposobu wyznaczania ważonych cech [60]

Kolejne z przykładowych deskryptorów, które próbowano zaimplementować w systemie *ARMiA* to cechy związane z prozodią głosu [60]. Parametry uzyskiwane przy pomocy opisanych w rozdziale 2. quasi-stacjonarnych ramek nie niosły ze sobą informacji zawartych we fragmentach mowy takich jak sylaby, słowa czy zdania. Takie dane, powiązane z brzmieniem głosu i relacjami między poszczególnymi segmentami nazwać można parametrami „dynamicznymi". W swoim rozwiązaniu dr Kamiński scharakteryzował następujące cechy prozodyczne:

- fluktuacja wartości częstotliwości podstawowej [60], [91], która pozwala na określenie melodyjności badanego głosu,
- czas trwania wypowiedzi [60], [91],
- odchyłka energii sygnału w ramce [60], [91], będąca sumą kwadratów próbek analizowanych ramek,
- pochodne cepstralnych, melcepstralnych oraz ważonych cech dystynktywnych pierwszego i drugiego rzędu [58], [60].

Ostatni z przykładowych parametrów dynamicznych to pierwsze oraz drugie pochodne cech dystynktywnych. W literaturze naukowej oznacza się je jako Δ (*delta*) i Δ^2 (*delta-delta*) *features*. Inaczej nazywane również współczynnikami różnicowymi między sąsiednimi ramkami sygnału [129].

3.4. Modele mieszanin gaussowskich jako metoda klasyfikacji

W dziedzinie rozpoznawania mówców prekursorem wśród klasyfikatorów są modele mieszanin gaussowskich (ang. *GMM*) [28], [95], [135]. Stanowią one liniową kombinację *C* rozkładów Gaussa [63]. Ich wykorzystanie podyktowane jest możliwością tworzenia oszczędnych pamięciowo modeli głosów, które bogate są w informacje dystynktywne [60].

Każdy z rozkładów λ charakteryzuje się trzema parametrami: wartością oczekiwaną (μ_i), macierzą kowariancji (Σ_i) oraz wagą rozkładu (w_i), których wartości mogą być dobierane dwojako, tj. pseudolosowo bądź w sposób zdeterminowany. Rozkład Gaussa o *d*-wymiarach można zapisać jako funkcję *N*

$$N(x_k,\mu_i,\Sigma_i)=\frac{\exp\left\{-\frac{1}{2}\left(x_k-\mu_i\right)^{'\Sigma_i^{-1}}\left(x_k-\mu_i\right)\right\}}{\left(2\pi\right)^{\frac{d}{2}}\left|\Sigma_i\right|^{\frac{1}{2}}} \tag{3.20}$$

Następnie podczas tworzenia modelu głosu wyznaczana jest postać funkcji gęstości prawdopodobieństwa występowania wektorów cech osobniczych zawartych w danych uczących konkretnego mówcy. Można przybliżyć ją do ważonej sumy *C* rozkładów Gaussa i zapisać jako

$$p(x_k\,|\,\lambda)=\sum_{i=1}^{C}N(x_k,\mu_i,\Sigma_i)\cdot w_i \tag{3.21}$$

Dysponując *K* wektorami obserwacji, proces uczenia mieszanin zrealizować można przy pomocy estymacji największej wiarygodności (ang. *ML*), która zapewnia jak najlepsze dopasowanie parametrów GMM do danych uczących. Opisuje się ją jako iloczyn funkcji gęstości prawdopodobieństwa dla *K* wektorów obserwacji

$$p(X\,|\,\lambda)=\prod_{k=1}^{K}p(x_k\,|\,\lambda) \tag{3.22}$$

W związku z nieliniowością wyrażenia (3.22) oraz licznie występującymi ekstremami, niemożliwa jest jego bezpośrednia maksymalizacja. Dlatego proces ten odbywa się iteracyjnie, a parametry modelu szacowane są poprzez algorytm *maksymalizacji oczekiwań* (ang. *EM*) [67]. Aby było to możliwe w praktyce, należy wyznaczyć funkcje pomocniczą w postaci [28]

$$Q\left(\lambda,\bar{\lambda}\right)=\sum_{i=1}^{C}\sum_{k=1}^{K}p(i\,|\,x_k,\lambda)\cdot\log\left[\bar{w}_i\cdot N(x_k,\bar{\mu}_i,\bar{\Sigma}_i)\right] \tag{3.23}$$

gdzie $p(i|x_k,\lambda)$ jest prawdopodobieństwem a posteriori wystąpienia *i*-tego rozkładu w modelu λ, dla wektora cech x_k. W założeniach algorytmu EM istnieje warunek, że w przypadku spełnienia nierówności $Q\left(\lambda,\bar{\lambda}\right)\geq Q\left(\lambda,\lambda\right)$ równanie $p(X\,|\,\bar{\lambda})\geq p(X\,|\,\lambda)$ również jest spełnione. Wspomniana wyżej iteracyjność procesu maksymalizacji oczekiwań odnosi się do dwóch następujących kroków, tj. kroku *estymacji* oraz kroku *maksymalizacji*. Pierwszy z nich odnosi się do oszacowania wartości prawdopodobieństwa $p(i|x_k,\lambda)$. Drugi natomiast, to już sama maksymalizacja, której wynikiem jest zbiór parametrów nowego modelu $\bar{\lambda}=(\bar{w}_i,\bar{\mu}_i,\bar{\Sigma}_i)$, maksymalizujących

przytoczoną wcześniej funkcję pomocniczą (5.4). Etapy te są powtarzane do momentu osiągnięcia założonej liczby iteracji lub gdy przyrost funkcji wiarygodności jest już niewielki. Działanie poszczególnych kroków algorytmu EM przedstawiają poniższe zależności:

Proces estymacji

$$p(i \mid x_k,\lambda)=\frac{N(x_k,\mu_j,\Sigma_j)\cdot w_j}{\sum_{i=1}^{C} N(x_k,\mu_i,\Sigma_i)\cdot w_i} \tag{3.24}$$

Proces maksymalizacji

$$\bar{w}_i=\frac{1}{K}\sum_{k=1}^{K} p(i \mid x_k,\lambda) \tag{3.25}$$

$$\bar{\mu}_i=\frac{\sum_{k=1}^{K} p(i \mid x_k,\lambda)\cdot x_k}{\sum_{k=1}^{K} p(i \mid x_k,\lambda)} \tag{3.26}$$

$$\bar{\Sigma}_i=\frac{\sum_{k=1}^{K} p(i \mid x_k,\lambda)\cdot\left(x_k-\bar{\mu}_i\right)'\cdot\left(x_k-\bar{\mu}_i\right)}{\sum_{k=1}^{K} p(i \mid x_k,\lambda)} \tag{3.27}$$

Tak utworzone modele mówców λ_j, wykorzystywane są w procesie identyfikacji badanego obiektu. Ustalenie tożsamości na podstawie próbki głosu interpretować można jako największe prawdopodobieństwo przynależności fragmentu mowy, reprezentowanej przez zbiór wektorów cech osobniczych *X*, do modelu GMM znajdującego się w bazie danych. Aby możliwa była identyfikacja mówcy najpierw należy wyznaczyć dla każdego modelu w bazie prawdopodobieństwo warunkowe, które mówi o stopniu reprezentacji wektora cech dystynktywnych *X* przez model λ_j [28]. Realizowane jest to przy pomocy funkcji dyskryminacji

$$g_j(X)=p(\lambda_j \mid X) \tag{3.28}$$

Zgodnie z regułą Bayesa [19] powyższy wzór można przekształcić do postaci

$$p(\lambda_j \mid X)=\frac{p(\lambda_j)p(X \mid \lambda_j)}{p(X)} \tag{3.29}$$

gdzie $p(X)$ to prawdopodobieństwo wystąpienia wektora cech *X* w sygnale mowy (takie samo dla wszystkich modeli), $p(\lambda_j)$ stanowi statystyczną popularność głosu w bazie, która przy *N* rekordach równa jest $1/N$ (każdy głos w bazie jest tak samo prawdopodobny). Z kolei $p(X|\lambda_j)$ to funkcja wiarygodności bazująca na modelu mówcy i stanowi prawdopodobieństwo reprezentacji zbioru wektorów cech *X* przez model λ_j [28]. Dysponując rozkładem funkcji dyskryminacyjnej na składowe funkcje prawdopodobieństwa, można opisać proces wyboru najbardziej prawdopodobnego głosu w sposób rankingowy, wg zależności

$$k^* = \arg\max g_j(X) \tag{3.30}$$

przy czym należy mieć na uwadze to, że funkcja dyskryminacji za sprawą korzystania wyłącznie z podejścia rankingowego przyjmuje postać

$$g_j(X) = p(X \mid \lambda_j) \tag{3.31}$$

Ostatecznie w końcowej realizacji systemu obliczana jest wartość logarytmiczna funkcji wiarygodności (ang. *log-likelihood*), co pozwala na zastosowanie technik homomorficznych, czyli na zamianę zależności między kolejnymi obserwacjami z multiplikatywnych na addytywne. W związku z wykorzystaniem środowiska *Matlab* i jego wbudowanych funkcji zwracana wartość logarytmu prawdopodobieństwa jest ujemna (ang. *negative log-likelihood)*. Zatem model z najniższą wartością ujemnego logarytmu prawdopodobieństwa uznawany jest za najbardziej prawdopodobny i bliski rozpoznawanemu fragmentowi mowy. Ostatecznie równanie poszukujące modelu rozpoznawanego fragmentu głosowego przyjmuje postać

$$lk^* = \arg\min_{1 \le j \le N} \sum_{k=1}^{K} -\log p(x_k \mid \lambda_j) \tag{3.32}$$

3.5. Podsumowanie

W niniejszym rozdziale przedstawiono w sposób syntetyczny najważniejsze moduły systemu *ARMiA*. Opisane wyżej metody oraz zależności matematyczne są wynikiem długich badań i procesów optymalizacyjnych, których celem było stworzenie systemu ASR cechującego się wysoką skutecznością rozpoznawania mówcy.

Znajomość wszelkich mechanizmów zaimplementowanych w systemie *ARMiA* pozwoliła autorowi niniejszej rozprawy na stworzenie zbioru deskryptorów behawioralnych, które można wykorzystać do dalszego rozwoju rozwiązań z dziedziny automatycznego rozpoznawania mówcy.

4.

System oparty na cechach behawioralnych

W rozdziale 1. niniejszej rozprawy przeprowadzony został przegląd dostępnej literatury w dziedzinie systemów automatycznego rozpoznawania mówcy. W wielu dziełach autorzy bazują na cechach fizycznych przy jednoczesnej marginalizacji możliwości dystynktywnych drzemiących w behawioralnych charakterystykach głosu. Jednak niektórzy z badaczy wskazują na możliwe zastosowania cech wysokiego poziomu do tworzenia samodzielnych systemów ASR czy też wsparcia już istniejących rozwiązań [77], [106], [119]. Po zapoznaniu się z aktualnym stanem wiedzy w zakresie automatycznego rozpoznawania mówców autor niniejszej rozprawy zdecydował się sprawdzić możliwości dystynktywne deskryptorów behawioralnych. Pierwszym etapem było stworzenie autorskiego systemu ASR, który miał bazować na cechach wysokiego poziomu

i umożliwić ich dobór, a następnie selekcję. Za rozwiązanie referencyjne[1] przyjęto system *ARMiA* [60].

4.1. Wstępne przetwarzanie sygnału mowy na potrzeby systemu ASR

Proces wstępnego przetwarzania sygnału mowy w różnych systemach ASR, niezależnie od charakteru wykorzystywanych cech, jest zazwyczaj do siebie zbliżony. W przypadku rozwiązania bazującego na charakterystykach wysokiego poziomu wszystkie operacje przed procesem parametryzacji mowy zostały opisane w rozdziale 2. niniejszej rozprawy. W tab. 4.1 zestawiono etapy wstępnego przetwarzania sygnału mowy wraz z poszczególnymi parametrami i ich wartościami. Wielkości poszczególnych parametrów związanych z długością wycinanego fragmentu ciszy czy długością ramki podczas segmentacji, w początkowych etapach opracowywania rozwiązania, zostały wstępnie dobrane w sposób inspirowany przeglądem literatury. Następnie w toku badań poddano je procesowi optymalizacji. Tabela 4.1 zawiera zestawienie wszystkich parametrów po optymalizacji ich wartości.

Tab. 4.1. Zestawienie parametrów i ich wartości wykorzystanych podczas wstępnego przetwarzania sygnału mowy

Proces wchodzący w skład wstępnego przetwarzania sygnału	Parametr	Wartość parametru
Przepróbkowanie sygnału	Szybkość próbkowania	8 kS/s
Standaryzacja wartości	-	-
Wycinanie ciszy	Minimalna długość ciszy do wycięcia	2 s
Segmentacja mowy	Długość ramki	3 s
	Nakładanie się (ang. *overlapping*) ramek	2 s
	Przesunięcie ramki	1 s

Porównując sposób przygotowania sygnału mowy do etapu generacji cech osobniczych w systemach *ARMiA* oraz opartym na behawioralnych cechach sygnału mowy zauważyć można kilka znaczących różnic. Pierwszą z nich jest inne podejście do skalowania wartości próbek. Rozwiązanie autorstwa dr. Kamińskiego korzysta z normalizacji [60], podczas gdy autor niniejszej rozprawy zastosował w swoim algorytmie standaryzację wartości. W systemie *ARMiA* wycięciu podlega każdy segment niezawierający mowy (tj. „cała cisza”), zaś w „systemie behawioralnym” eliminowane są tylko fragmenty o długości ciszy przekraczającej 2 sekundy. Ostatni etap – segmentacji mowy – również jest odmienny w obu podejściach [60]. W pierwszym z nich do okienkowania wykorzystano okno Hamminga o długości nieprzekraczającej setek milisekund, zaś w przypadku systemu opartego na behawioralnych cechach sygnału mowy stosowane jest okno prostokątne o długości sięgającej 3 sekund.

[1] Poprzez rozwiązanie referencyjne rozumieć należy system, z którym porównywana będzie autorska aplikacja i do którego zostanie zastosowane wsparcie z wykorzystaniem cech behawioralnych.

4.2. Definicje i przykłady wykorzystanych w systemie cech behawioralnych

Pojęcie *behawior* zdefiniować można jako zachowanie, które pojawiło się w odpowiedzi na pewien czynnik. Dokładniej rzecz ujmując jest to kompleksowy układ zachowań organizmu w odpowiedzi na sygnały docierające ze środowiska zewnętrznego bądź z jego wnętrza [4]. W dziedzinie analizy głosu człowieka wspomniany wcześniej behawior to osobnicze różnice w sposobie artykulacji. Poprzez wyuczone nawyki w kontekście wypowiedzi rozumieć należy różnice w intonacji wypowiadanej frazy, czasu wypowiedzi czy poziomu głośności. Tak samo klasyfikuje się sposób doboru sekwencji słów przez człowieka – w tym przypadku wykorzystuje się omawiane wcześniej *N*-gramy [33]. Geneza pochodzenia cech behawioralnych ma swoje podłoże w wielu sferach, a dokładniejszy ich opis znajduję się w rozdziale 2.3. Poza wpływem czynników socjoekonomicznych, takich jak miejsce urodzenia czy środowisko dorastania wyróżnić należy również jedną z głównych składowych modyfikującą charakter wypowiedzi jaką jest aktualny stan emocjonalny człowieka. Zarówno z punktu widzenia mówcy, jak i słuchacza wpływ emocji na charakter wypowiedzi jest zauważalny [48]. Największy wpływ na charakter generowanej mowy mają różnego rodzaju stresory. Wyróżnia się ich następujące rodzaje [48], [112]:

- Stresory fizyczne (ang. *physical stressors*) – związane z oddziaływaniem czynników fizycznych, m.in. takich jak ruch, na składowe traktu głosowego.
- Podświadome stresory fizjologiczne (ang. *unconscious physiological stressors*) – zmiany mające swoje podłoże w chemicznych uwarunkowaniach organizmu, przejawiające się jako niekontrolowane napięcie mięśni czy zmiany szybkości oddychania.
- Świadome stresory fizjologiczne (ang. *conscious physiological stressors*) – ich źródłem jest środowisko wypowiedzi, tzn. wpływ hałasu czy grona słuchaczy. Są to świadome zmiany w charakterystyce mowy, takie jak zmiana głośności czy staranność artykulacji wypowiedzi.
- Stresory psychologiczne (ang. *psychological stressors*) – najtrudniejsze do jednoznacznego określenia, gdyż za ich pochodzenie odpowiada wiele czynników. Mogą nimi być zarówno sytuacja w jakiej znalazł się mówca, jak też emocje, które nim targają. Wynikiem ich działania jest zmiana rytmu artykulacji.

Mając na względzie opisane wyżej składowe wpływające na charakter wypowiedzi, dość trudnym zadaniem jest jednoznaczne określenie uniwersalnego zbioru cech behawioralnych zależnego wyłącznie od czynników opisanych w rozdziale 2.3. W niniejszych podrozdziałach autor opisał wstępny zestaw deskryptorów potencjalnie przydatnych w projektowanym systemie.

Aby możliwe było dokładne scharakteryzowanie wykorzystanych deskryptorów behawioralnych należy najpierw przybliżyć zastosowane metody parametryzacji sygnału mowy. W opracowanym rozwiązaniu zastosowano dwie dziedziny analizy przebiegów mowy, tj. czasową i częstotliwościową. Dokładny opis analizy widmowej sygnału głosowego opisany został w rozdziale 3.2.1. Z kolei przetwarzanie postaci czasowej sygnału w celu ekstrakcji cech w niniejszym systemie opiera się na analizie wcześniej wyznaczonych, 3 sekundowych ramek. Poza rozpatrywanymi fragmentami sygnału

o standardowej długości, dokonywany jest również podział na krótsze segmenty o czasie trwania rzędu setek milisekund.

Z postaci czasowej ekstrahowane są takie parametry jak:

- całkowity czas mowy w analizowanej ramce,
- unormowany czas trwania ramki mowy (w stosunku do długości całej wypowiedzi),
- skala głosu – cecha określająca zakres zmian częstotliwości podstawowej głosu, tzn. przedział od minimalnej do maksymalnej wysokości dźwięku,
- procent ramek o niskiej energii,
- liczba znaczących zjawisk impulsowych,
- liczba ramek dźwięcznych (w stosunku do całości sygnału),
- tempo (częstotliwość) przejść przez zero.

4.2.1. Całkowity czas mowy w analizowanej ramce oraz unormowany czas trwania ramki mowy

Całkowity czas mowy w analizowanej ramce to cecha pozwalająca na określenie szybkości artykulacji badanego obiektu, a dokładniej szybkości generacji głosek przez człowieka. Z kolei unormowany czas trwania ramki mowy odnosi się do stosunku między sumą wszystkich próbek reprezentujących mowę, a długością nagrania wyrażoną w próbkach.

4.2.2. Skala głosu

Skala głosu jest cechą, która mówi o tym, w jakim przedziale wysokości dźwięków znajduje się głos badanego mówcy. Sprowadzić ją można do zakresu od najniższej do najwyższej częstotliwości podstawowej generowanej przez narząd mowy [45].

4.2.3. Procent ramek o niskiej energii

Korzystając ze wzoru (2.2) możliwe jest wyznaczenie krótkoczasowej wartości energii. W przypadku generacji nowej cechy termin krótkoczasowa odnosi się do jednosekundowego wycinka ramki. Następnie wyznacza się średnią wartość krótkoczasowej energii w zbiorze wszystkich ramek. Ostatnim etapem jest porównanie powyższych parametrów. Ramki nie przekraczające połowy średniej zaliczane są do grona segmentów o niskiej energii, a ich suma w stosunku do wszystkich ramek w sygnale to ostatecznie ekstrahowana cecha [32].

4.2.4. Liczba ramek dźwięcznych

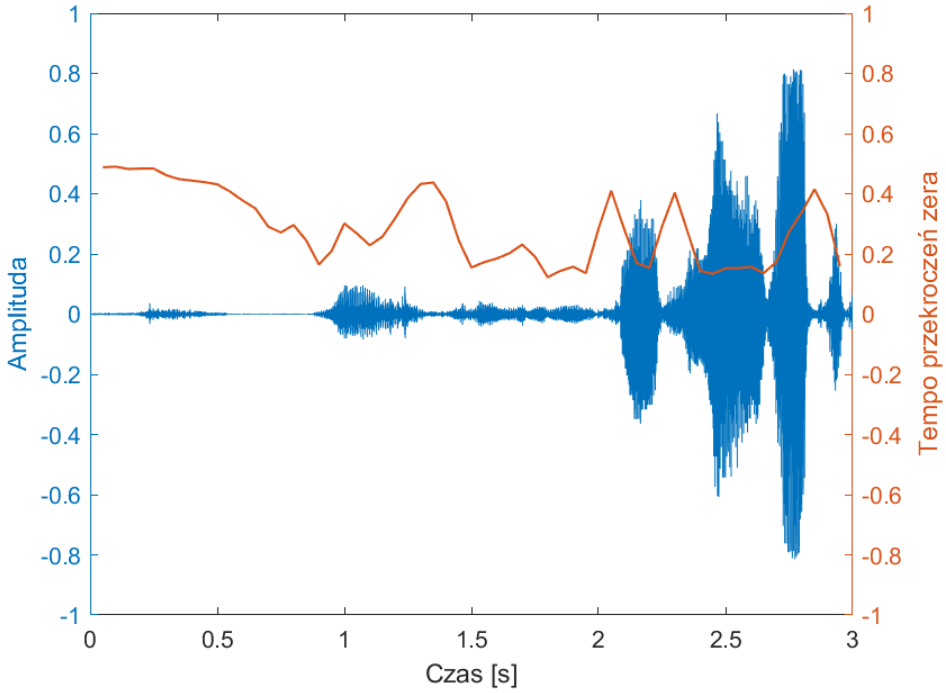
Analiza dźwięczności ramek dokonywana jest zgodnie ze wzorem (3.5), a sama klasyfikacja fragmentów mowy na dźwięczne i bezdźwięczne odbywa się poprzez detekcje odpowiednich maksimów funkcji autokorelacji (dokładniejszy opis znajduje się w podrozdziale 3.2.1). Mając liczbę wszystkich dźwięcznych ramek, można tą wartość podzielić przez liczbę wszystkich ramek składających się na sygnał. Tak powstały parametr mówi o udziale ramek dźwięcznych w analizowanym fragmencie głosu.

4.2.5. Tempo przejść przez zero

Tempo przekroczeń zera (ang. *Zero-Crossing Rate*), to cecha związana ze zmianami znaku sygnału w analizowanej ramce. Zliczane są wszystkie zmiany wartości poszczególnych próbek z dodatnich na ujemne i na odwrót, a otrzymana suma dzielona jest przez długość ramki [32], [127]. Cechę tą wyrazić można za pomocą poniższego wzoru:

$$ZCR_i = \frac{1}{2(N-1)} \sum_{n=0}^{N-1} \left(\left| \mathrm{sgn}(x_i(n)) - \mathrm{sgn}(x_i(n-1)) \right| \right), \qquad \mathrm{sgn}(x) = \begin{cases} 1 \ dla \ x \geq 0 \\ -1 \ dla \ x < 0 \end{cases} \tag{4.1}$$

gdzie N to długość ramki wyrażona w próbkach, $x_i(n)$ to n-ta próbka i-tej ramki, a ZCR_i stanowi tempo przekroczeń zera i-tej ramki. Na rys 4.1 pokazano przykładowy 3-sekundowy fragment analizowanego sygnału mowy wraz z naniesionym parametrem tempa przekroczeń zera.

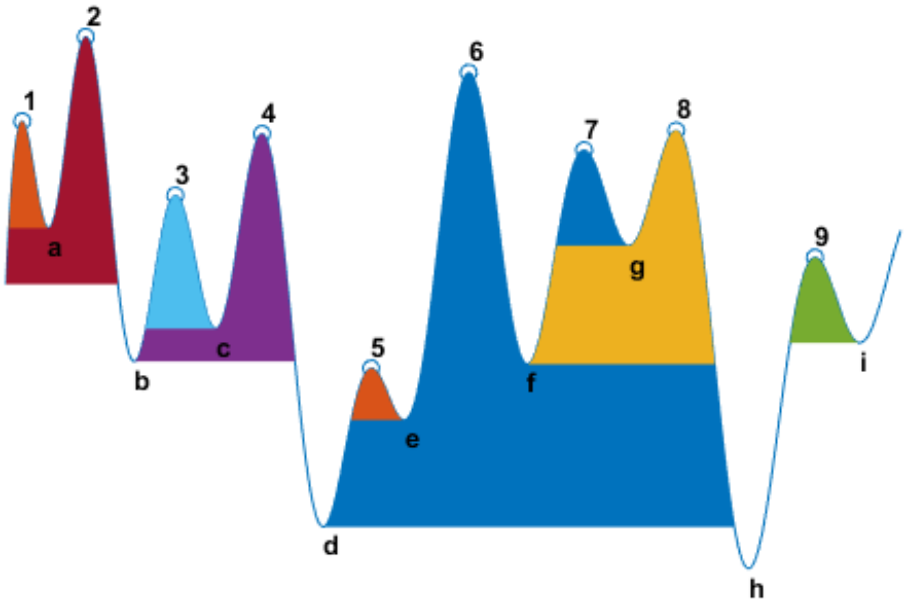


Rys. 4.1. Graficzne zobrazowanie parametru tempa przekroczeń zera

Jednocześnie parametr ten niesie również informacje o zaszumieniu sygnału, tzn. w przypadku dużej wartości *ZCR* mamy do czynienia z dużą wartością szumu – związane jest to z częstymi zmianami znaku następujących po sobie próbek sygnału szumowego [127].

4.2.6. Liczba znaczących zjawisk impulsowych

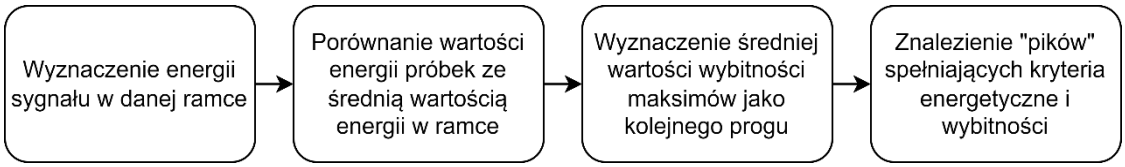
Jednym z autorskich parametrów, którego ekstrakcja została zaimplementowana w systemie ASR jest liczba znaczących zjawisk impulsowych. Poprzez znaczące rozumieć należy takie maksima, które przekraczają odpowiednio zdefiniowany próg, ale również wyróżniają się wśród swoich sąsiadów. Rys 4.2 stanowi graficzne zobrazowanie wyróżnienia maksimów.



Rys. 4.2. Przykładowe maksima sygnału i ich znamienność [156]

Ekstrema oznaczone od *a* do *i* to minima (najniższe punkty) w sąsiedztwie każdego z maksimów ponumerowanych od 1 do 9. Biorąc za przykład „pik" o numerze 5 widać, że w jego sąsiedztwie znajdują się minima oznaczone jako *d* oraz *e*. Znaczenie tego ekstremum (oznaczone kolorem pomarańczowym) jest wyznaczane przez najbliższe, najwyżej położone minimum, po którym następuję zmiana monotoniczności przebiegu funkcji (w tym przypadku jest „pik" *e*, po przekroczeniu którego następuję zmiana z malejącego charakteru sygnału na rosnący [156]). *Wybitność* tego ekstremum to różnica między jego wartością, a wartością najbliższego, najwyżej położonego minimum (zmieniającego charakter przebiegu sygnału). Innymi słowy jest to parametr *względnej wysokości „piku" w stosunku do najwyższego sąsiedniego minimum.*

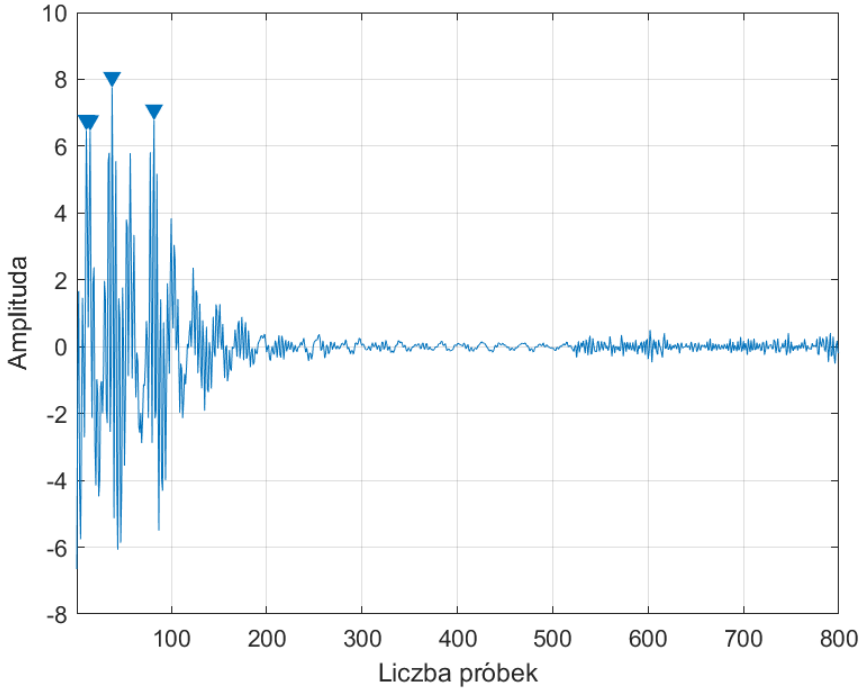
Tak określona istotność wyznaczana jest dla każdego maksimum w analizowanej ramce, a z niej wartość średnia $\overline{p}_p$. Przed tym obliczana jest średnia energia „pików" aktualnie przetwarzanego fragmentu sygnału $\overline{p}_h$. Całość procesu ekstrakcji cechy pokazana jest schematycznie na rys. 4.3.



Rys. 4.3. Uogólniony schemat wyznaczania najbardziej znaczących maksimów

Autor w oparciu o powyższe średnie parametry stworzył warunek klasyfikacji znaczenia *i*-tego maksimum (4.2) w zależności od jego wartości p_{h_i} oraz p_{p_i}. W pierwszym kroku „pik" musi spełnić kryterium energetyczne, aby następnie porównaniu podlegała jego *względna wysokość w stosunku do najwyższego okalającego minimum*. Na rys. 4.4 znajduje się przykładowa liczba znaczących zjawisk impulsowych analizowanej ramki.

$$p_{h_i} > \overline{p}_h \Leftrightarrow p_{p_i} > \overline{p}_p \tag{4.2}$$



Rys. 4.4. Znaczące zjawiska impulsowe w analizowanej ramce

4.2.7. Wartość średnia amplitudy w oktawach i tercjach

Cechy ekstrahowane z widma amplitudowego ramek stanowią przeważającą większość w puli wszystkich cech w opracowanym systemie ASR. Część parametrów definiowana jest
w *oktawach* i *tercjach*. Poprzez oktawę rozumieć należy pasmo częstotliwości o częstotliwości środkowej $f_{śr}$ (4.3) i stosunku między częstotliwością górną i dolną równym 2 (4.4) [34], [73].

$$\begin{cases} f_{śr} = f_d \cdot \sqrt[2]{2} \\ f_{śr} = f_g \cdot \sqrt[-2]{2} \end{cases} \tag{4.3}$$

$$\frac{f_g}{f_d} = 2 \tag{4.4}$$

gdzie f_g jest częstotliwością górną oktawy, f_d to częstotliwość dolna.

Z kolei tercje to pasma ⅓-oktawowe, gdyż w skład jednej oktawy wchodzą 3 tercje. W ich przypadku stosunek f_g do f_d równy jest $\sqrt[3]{2}$. Wzór (4.3) dla tercji przyjmuje postać [34], [73]:

$$\begin{cases} f_{śr} = f_d \cdot \sqrt[6]{2} \\ f_{śr} = f_g \cdot \sqrt[-6]{2} \end{cases} \tag{4.5}$$

Posługując się powyższymi wzorami w omawianym systemie zaimplementowany został podział na oktawy i tercje zgodnie z wartościami przedstawionymi w tabelach 4.2 i 4.3.

Tab. 4.2. Zestawienie pasm częstotliwości w oktawach

Numer oktawy	f_d [Hz]	$f_{śr}$ [Hz]	f_g [Hz]
1	1	1,4	2
2	2	2,8	4
3	4	5,7	8
4	8	11,3	16
5	16	22,6	32
6	32	45,3	64
7	64	90,5	128
8	128	181	256
9	256	362	512
10	512	724,1	1024
11	1024	1448,2	2048
12	2048	2896,3	4096

Tab. 4.3. Zestawienie pasm częstotliwości w tercjach

Numer tercji	f_d [Hz]	$f_{śr}$ [Hz]	f_g [Hz]
1	1	1,12	1,26
2	1,26	1,41	1,59
3	1,59	1,78	2
4	2	2,24	2,52
5	2,52	2,83	3,17
6	3,17	3,56	4
7	4	4,49	5,04
8	5,04	5,66	6,35
9	6,35	7,13	8
10	8	8,98	10,08
11	10,08	11,31	12,70
12	12,70	14,25	16
13	16	17,96	20,16
14	20,16	22,63	25,40
15	25,40	28,51	32
16	32	35,92	40,32
17	40,32	45,25	50,80
18	50,80	57,02	64
19	64	71,84	80,63
20	80,63	90,51	101,59
21	101,59	114,04	128
22	128	143,68	161,27
23	161,27	181,02	203,19
24	203,19	228,07	256
25	256	287,35	322,54
26	322,54	362,04	406,37
27	406,37	456,14	512
28	512	574,70	645,08
29	645,08	724,08	812,75
30	812,75	912,28	1024
31	1024	1149,40	1290,16
32	1290,16	1448,15	1625,50
33	1625,50	1824,56	2048
34	2048	2298,80	2580,32
35	2580,32	2896,31	3251
36	3251	3649,12	4096

Dla każdej z oktaw i tercji wyznaczane jest widmo amplitudowe, a następnie sumowane są wartości poszczególnych próbek widma i wskazywane są pasma częstotliwości o największej sumie (oktawa i tercja o największym polu pod krzywą opisującą widmo). Oktawa i tercja charakteryzujące się największym polem podlegają dalszej analizie, podczas której wyznaczane są wartości średnie ich amplitud. Tak otrzymana wartość średnia amplitudy w oktawach i tercjach stanowi kolejny deskryptor [1].

4.2.8. Pasmo częstotliwości zawierające 95% całkowitej mocy sygnału

Kolejną cechą związaną z widmem sygnału, która zaimplementowana została w opracowanym systemie ASR jest zajętość pasma (ang. *Occupied Bandwidth – OBW*). Parametr ten mówi o szerokości pasma częstotliwości, poza którym znajduję się maksymalnie β % mocy (po β/2 % po każdej stronie pasma) [29], [44]. Inaczej mówiąc jest to pewien przedział częstotliwości w którym zawiera się (100-β) % całości mocy sygnału. Moc całkowita w widmie obliczana jest za pomocą wzoru [2]:

$$P_c = \sum_{k=1}^{N} \left| S(k)^2 \right| \tag{4.6}$$

Następnie wyznaczane są wartości β/2 % i 100-(β/2) % mocy:

$$P_{\frac{\beta}{2}} = P_c \cdot \frac{\beta}{2} \cdot \frac{1}{100} \tag{4.7}$$

$$P_{\left[100-\left(\frac{\beta}{2}\right)\right]} = P_c \cdot \left[100 - \left(\frac{\beta}{2}\right)\right] \cdot \frac{1}{100} \tag{4.8}$$

Ostatnim krokiem jest kumulatywne sumowanie kwadratów próbek widma do momentu osiągnięcia opisanych wyżej progów:

$$P_{\frac{\beta}{2}} = \sum_{k=1}^{K_1} \left| S(k)^2 \right| \tag{4.9}$$

$$P_{\left[100-\left(\frac{\beta}{2}\right)\right]} = \sum_{k=1}^{K_2} \left| S(k)^2 \right| \tag{4.10}$$

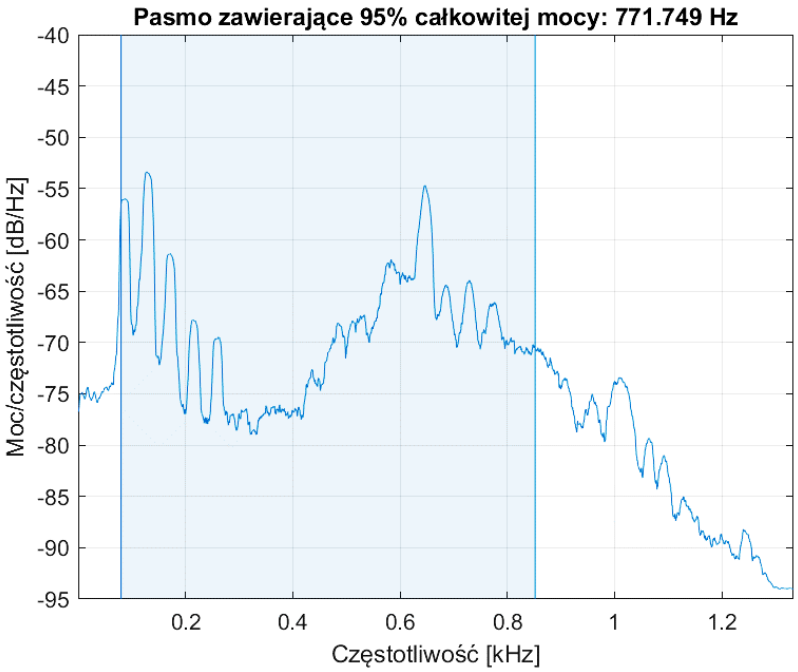
gdzie K_1 i K_2 to numery próbek, dla których moc wynosi odpowiednio β/2 i 100-(β/2) % mocy całkowitej. W oparciu o otrzymane indeksy oraz odstęp między próbkami widma Δf można określić częstotliwości f_1 (4.11) i f_2 (4.12) dla których widmo zawiera opisane wyżej części mocy.

$$f_1 = K_1 \cdot \Delta f \tag{4.11}$$

$$f_2 = K_2 \cdot \Delta f \tag{4.12}$$

Następnie różnica f_2 i f_1 to ekstrahowany parametr zajętości pasma (4.13). Graficzne zobrazowanie *OBW* przedstawiono na rys. 4.5.

$$OBW = f_2 - f_1 \tag{4.13}$$



Rys. 4.5. Przedział częstotliwości zawierający 95% mocy całkowitej

4.2.9. Miara płaskości widma

Widmo, będące wynikiem analizy fourierowskiej, reprezentuje wynik rozkładu sygnału na sumę składowych harmonicznych. Jedną z charakterystyk widma możliwych do interpretacji jest miara płaskości widma (ang. *Spectral Flatness Measure – SFM*). Określa ona podobieństwo analizowanego sygnału akustycznego do szumu białego. SFM zawiera się w granicach od 0 do 1, gdzie wartości graniczne charakteryzują odpowiednio „czysty ton" lub szum biały. Płaskość na poziomie 1 oznacza, że każda składowa częstotliwościowa w widmie charakteryzuję się podobną wartością mocy, a przebieg widma jest wygładzony i równy, przypominający szum biały. Z kolei SFM o wartości 0 mówi, że moc skupiona jest wokół niewielkiej liczby harmonicznych, a co za tym idzie widmo jest bardziej „szpiczaste" [109]. Stosunek parametrów średniej geometrycznej oraz arytmetycznej widma (4.14) to liczbowe wyrażenie miary płaskości widma [39].

$$SFM = \frac{\sqrt[K]{\prod_{k=0}^{K-1} S(k)}}{\frac{1}{K}\sum_{k=0}^{K-1} S(k)} \tag{4.14}$$

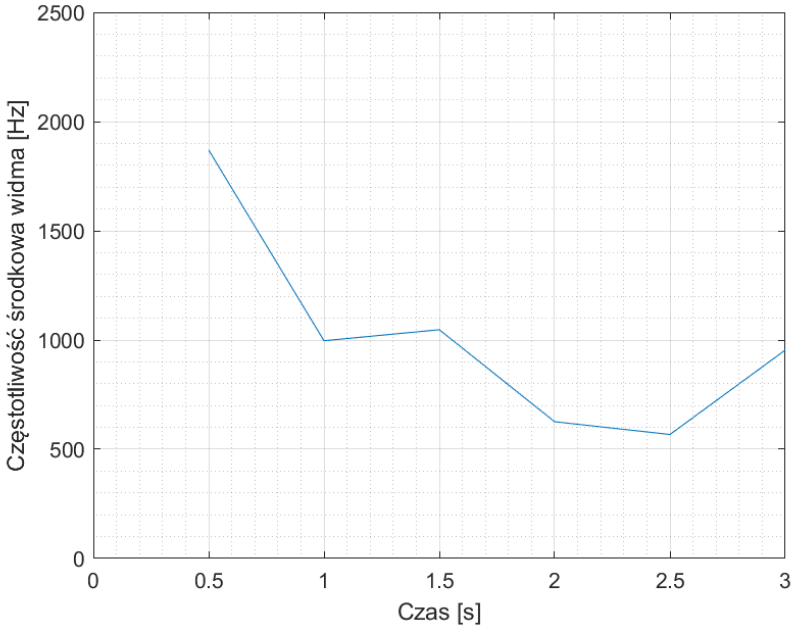
gdzie $S(k)$ to kolejne próbki widma, a K-1 to długość analizowanego sygnału.

4.2.10. Środek ciężkości widma i jego rozproszenie

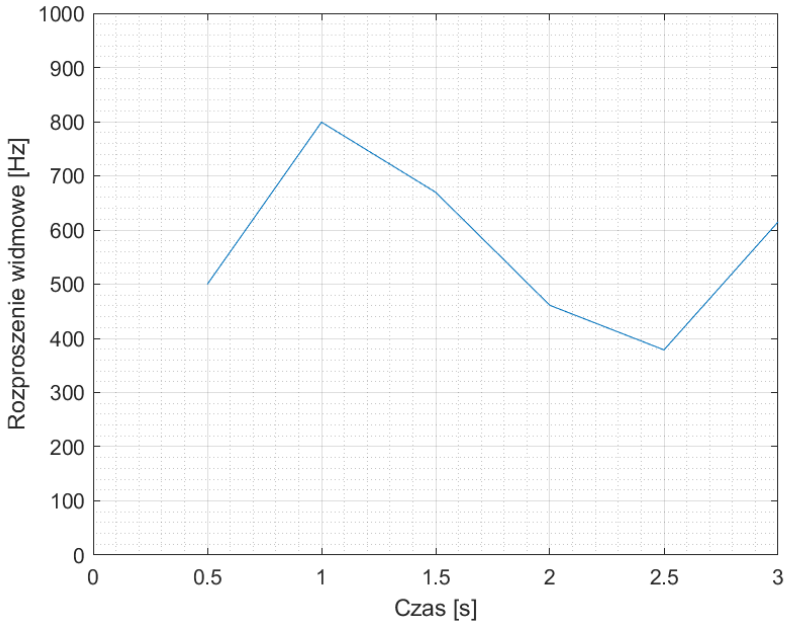
Środek ciężkości widma (ang. *Spectral Centroid*) oraz rozproszenie widmowe (ang. *Spectral Spread*) to parametry określające kształt oraz rozmieszczenie widma [127]. Pierwszy z nich definiuje częstotliwość środkową w widmie, czyli taką wartość, przy której spektrum podzielone jest na dwie równe połowy pod względem mocowym. Z kolei rozproszenie widmowe informuje o rozkładzie mocy wokół środka ciężkości. Właściwość

tą przyrównać można do statystycznego parametru jakim jest odchylenie standardowe mocy wokół częstotliwości środkowej widma [127]. Im bardziej sygnał jest skupiony wokół swojego środka ciężkości tym niższe są wartości rozproszenia widmowego. Cechy te skorelowane są z dominacją wysokich tonów w analizowanym sygnale dźwiękowym, co w środowisku muzycznym określa się mianem *jasnego brzmienia* [49]. Powyższe parametry opisać można przy pomocy zależności (4.15) oraz (2.3), a ich graficzne reprezentacje widoczne są na rysunkach 4.6 oraz 4.7.

$$SC_i = \frac{1}{P_c}\sum_{k=1}^{N}\left(k\cdot\Delta f\right)\cdot P_k \tag{4.15}$$



Rys. 4.6. Częstotliwości środkowe widma ramek o długości 500 ms



Rys. 4.7. Rozproszenia widmowe ramek o długości 500 ms

4.3. Metody ewolucyjne w zastosowaniu do selekcji zestawu cech

Proces generacji cech osobniczych we wstępnych etapach rozwoju systemów ASR ma za zadanie ekstrakcję możliwie jak największej liczby cech dystynktywnych z sygnału mowy. Jednak wykorzystanie maksymalnie licznego zbioru deskryptorów do rozwiązania zadania klasyfikacji nie zawsze pozwala na uzyskanie najlepszych wyników [59], [102], [105]. Odpowiedni dobór cech to złożony proces, który pozwala na redukcje wymiaru wektora charakterystyk osobniczych przy jednoczesnym zachowaniu takiej samej lub uzyskaniem większej skuteczności identyfikacji. Poza poprawą skuteczności klasyfikacji ważnym aspektem selekcji cech jest skrócenie czasu obliczeń wynikające ze zredukowanej liczby przetwarzanych parametrów. Mając to na uwadze dobór odpowiedniego algorytmu selekcji cech przez projektanta systemu ASR staje się kwestią bardzo istotną. Istnieje wiele metod doboru deskryptorów, spośród których wyróżnić można przykładowo metody filtracyjne, opakowane (ang. *wrapper methods*) czy hybrydowe [14], [113], [120]. Do najpowszechniej stosowanych należą: metoda dyskryminacyjna Fishera, metoda korelacji wzajemnej, metoda analizy składowych głównych czy algorytmy genetyczne. Autor w trakcie prowadzonych badań testował różne rozwiązania, ale swój wysiłek skupił głównie na ewolucyjnej metodzie selekcji cech jaką jest algorytm genetyczny (ang. *Genetic Algorithm*) [23].

Selekcja deskryptorów dystynktywnych poprzez algorytm genetyczny to alternatywna technika zaliczana do globalnych metod optymalizacji. Oznacza to, że równolegle przetwarzanych jest wiele punktów w przestrzeni rozwiązań, co pozwala uniknąć sytuacji utknięcia w minimum lokalnym [113]. Idea, która stoi za algorytmem genetycznym ma swoje podłoże w biologicznym zjawisku ewolucji organizmów na skutek ich interakcji z otoczeniem.

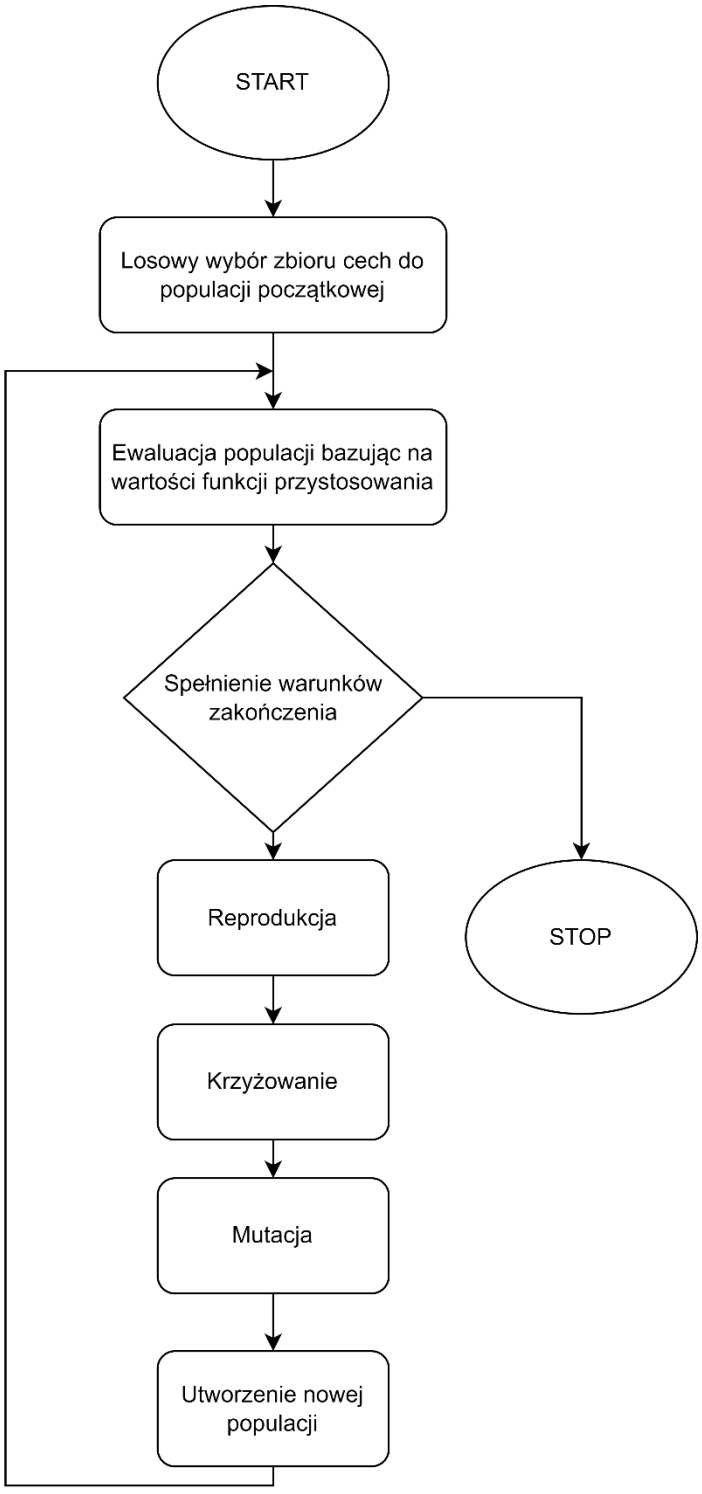
W opracowanym rozwiązaniu sposób wykorzystania algorytmu genetycznego pokazany został na rys. 4.8.

Pierwszy etap działania algorytmu genetycznego to dobór puli cech wchodzących w skład populacji początkowej. W opracowanym rozwiązaniu jeden podzbiór (osobnik) w populacji początkowej to zbiór wszystkich, pierwotnie zdefiniowanych deskryptorów. W algorytmie genetycznym poszczególnym cechom przyporządkowuje się wartości binarne, które stanowią wektor (chromosom) wykorzystywany w kolejnych operacjach genetycznych [113]. Następnym krokiem po wyborze populacji początkowej jest obliczenie wartości jej funkcji przystosowania. Funkcja przystosowania (ang. *fitness function*) to inaczej funkcja celu w procesie optymalizacji (minimalizacji) liczby cech. W procesie wyznaczania wartości funkcji przystosowania dla każdego osobnika obliczana jest sprawność systemu ASR. Następnie jest ona odejmowana od jedności (gdyż selekcja cech to zadanie minimalizacji), a otrzymana wartość stanowi miarę bliskości tego osobnika (rozwiązania) do optimum. W wykorzystanym przez autora wariancie algorytmu genetycznego wartość funkcji przystosowania opisana jest wzorem:

$$ff_i = 1 - Acc_i \tag{4.16}$$

gdzie ff_i stanowi wartość funkcji przystosowania i-tego osobnika, a Acc_i to sprawność identyfikacji systemu dla i-tego osobnika.

W związku z powyższym wyrażeniem, im niższa wartość parametru ff_i tym lepsze przystosowanie *i*-tego osobnika w procesie genetycznym.

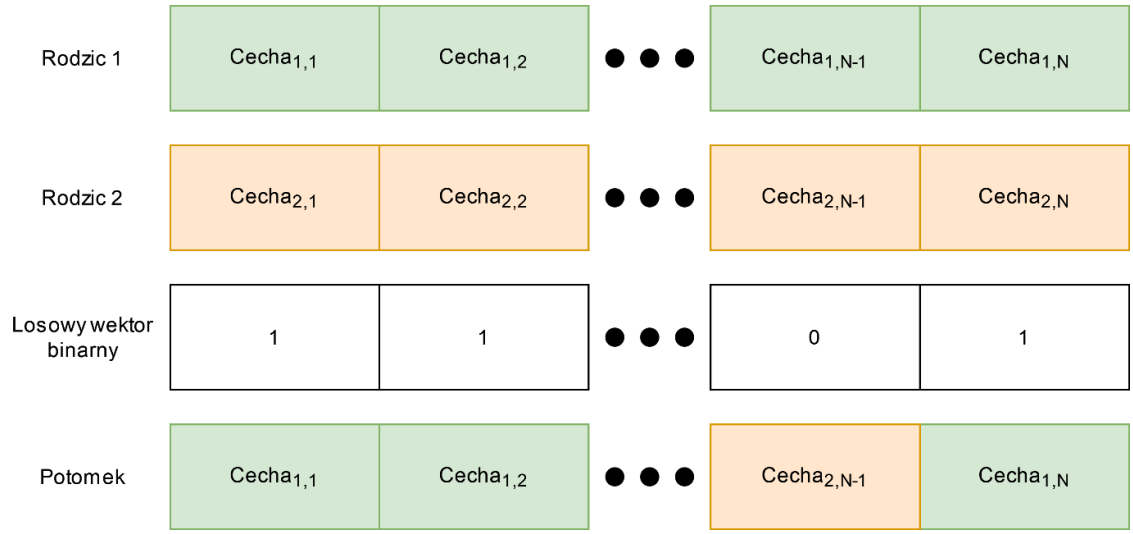


Rys. 4.8. Schemat działania algorytmu genetycznego w zastosowaniu do selekcji cech [23], [113]

Kolejnym etapem działania algorytmu genetycznego jest proces *selekcji* i *reprodukcji,* czyli powielenie konkretnego chromosomu w zależności od wartości funkcji przystosowania. Operacje te przyrównać można do zasady doboru naturalnego, gdzie najlepiej przystosowani (wyselekcjonowani rodzice) przeżywają i wydają potomstwo. Następny krok to *krzyżowanie*, które polega na połączeniu chromosomów rodziców

w celu utworzenia nowego osobnika. W opracowanym systemie wykorzystano *krzyżowanie rozproszone* (ang. *scattered crossover*), którego zasada działania w sposób graficzny przedstawiona została na rys 4.9.

Krzyżowanie to polega na doborze dwóch rodziców oraz losowego wektora binarnego o długości równej chromosomom rodziców. Następnie w zależności od wartości w wektorze binarnym poszczególne cechy od każdego z rodziców przekazywane są potomkowi.



Rys. 4.9. Schemat działania algorytmu genetycznego w zastosowaniu do selekcji cech [20]

W każdej kolejnej iteracji działania algorytmu genetycznego dzięki operacjom reprodukcji i krzyżowania powstają najlepiej przystosowani potomkowie (w ujęciu jakości materiału genetycznego – najbardziej dystynktywne cechy). Niekiedy jednak może dojść do eliminacji pewnych, potencjalnie cennych z punktu widzenia skuteczności identyfikacji cech. Aby temu zapobiec stosuje się *mutację,* czyli przypadkową, losową zmianę wartości genu. Mutacja Gaussa charakteryzuje się tym, że cecha podlegająca mutacji przyjmuje wartość losową z rozkładem Gaussa o wartości oczekiwanej równej wartości cechy przed mutacją. Operator mutacji Gaussa opisywany jest zależnością [87]

$$M_g(x) = \min\left(\max\left(N(x,\sigma),a\right),b\right) \wedge x \in \langle a,b \rangle \tag{4.17}$$

gdzie, x to gen przyjmujący wartości z przedziału $[a, b]$.

Analityczny wzór potomstwa po zastosowaniu mutacji Gaussa pokazany jest poniżej [87]:

$$x'_i = \sqrt{2}\sigma\left(b_i - a_i\right) erf^{-1}\left(u'_i\right) \tag{4.18}$$

$$erf(y) = \frac{2}{\sqrt{\pi}} \int_0^y e^{-t^2} dt \tag{4.19}$$

gdzie x'_i to i-ty potomek, $[a_i, b_i]$ to przedział wartości dla i-tego genu x_i, σ to odchylenie standardowe rozkładu Gaussa (zazwyczaj przyjmuje się wartość zależną od numeru pokolenia), erf^{-1} jest odwrotnością funkcji błędu Gaussa argumentu u'_i o losowej wartości z przedziału (0, 1).

W wyniku operacji genetycznych powstaje nowa populacja osobników, która poddawana jest ponownej ocenie przez wartość funkcji przystosowania. Wszystkie kroki powtarzane

są do momentu spełnienia warunku zakończenia obliczeń, tj. do momentu osiągnięcia maksymalnej liczby pokoleń lub określonej liczby pokoleń bez poprawy wartości funkcji przystosowania. Wynikiem działania całego algorytmu genetycznego jest otrzymanie najlepiej przystosowanego osobnika, czyli wektora cech najwyżej ocenionego w procesie genetycznym.

W związku z zależnością algorytmu genetycznego od losowo generowanych parametrów, etap selekcji cech był wielokrotnie powtarzany, aby dobrać jak najczęściej pojawiające się geny. Wynikiem procesu generacji cech był wyjściowy zbiór 71 cech osobniczych. W efekcie procesu długotrwałego wyboru deskryptorów doprowadzono do redukcji wymiaru cech o ponad 60% przy jednoczesnej, prawie 5% poprawie skuteczności identyfikacji mówcy. Ostatecznie do dalszych badań, opisanych w rozdziale 5, wykorzystano zbiór 27 wyselekcjonowanych algorytmem genetycznym cech, zawierającym cechy

$$\begin{bmatrix} C_3, & C_4, & C_6, & C_7, & C_8, & C_{10}, & C_{11}, & C_{12}, & C_{13}, & C_{15}, & C_{16}, & C_{20}, & C_{22}, & C_{25}, & C_{30}, & \cdots \\ \cdots & C_{33}, & C_{35}, & C_{36}, & C_{37}, & C_{48}, & C_{52}, & C_{57}, & C_{63}, & C_{67}, & C_{69}, & C_{70}, & C_{71} \end{bmatrix}$$

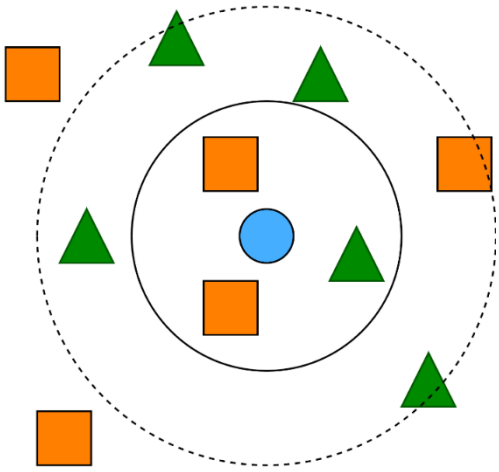
4.4. Zastosowany klasyfikator

Ostatni etap procesu rozpoznawania mówcy to *klasyfikacja,* czyli proces przyporządkowania odpowiedniej klasy do źródła sygnału [2]. Operator realizujący to nazywany jest *klasyfikatorem*. Jego projektowanie składa się z dwóch faz, tj. *fazy treningowej* (*uczenia*) i *fazy testowej* (*testowania*). Bazę dostępnych głosów (źródło cech dystynktywnych) dzieli się na dwa rozłączne podzbiory. Pierwszy z nich zwany *treningowym* służy do budowy (uczenia) klasyfikatora. Z kolei drugi to zbiór *testowy* służący, jak sama nazwa wskazuje, do testowania modelu klasyfikacji. Najprostszym sposobem klasyfikacji jest wykorzystanie odległości analizowanego wektora cech od wektorów przyporządkowanych do konkretnych klas. Metoda ta nosi nazwę *klasyfikacji minimalno-odległościowej*, gdyż zakłada bliskie położenie podobnych do siebie obiektów [2].

Jedną z podstawowych metod klasyfikujących opartych na odległościach między zbiorami cech jest *metoda najbliższego sąsiada*. Jej zasada działania opiera się na znalezieniu najbliższego wektora spośród danych uczących w stosunku do wektora testującego (czyli aktualnie rozpatrywanego). Rozwinięciem tej metody jest wariant opracowany w 1951 r. zakładający poszukiwanie *k* najbliższych wektorów [30]. Stąd też wzięła się jego nazwa, tj. *metoda k najbliższych sąsiadów* (ang. *k-nearest neighbors – k-nn*). W odróżnieniu od podejścia klasycznego, w klasyfikatorze *k-nn* wyznaczana jest suma odległości do *k* najbliższych wektorów, a nie tylko jednego. Przy założeniu, że zbiory cech x_i oraz x_j są wektorami, do obliczenia odległości stosuje się najczęściej *odległość euklidesową* [65]

$$d_{ij} = \left\| x_i - x_j \right\| = \sqrt{\left(x_i - x_j \right)\left(x_i - x_j \right)^T} = \sqrt{\sum_{k=1}^{N} \left(x_{ik} - x_{jk} \right)^2} \tag{4.20}$$

Na rysunku 4.10 zaprezentowano ideę działania metody *k-nn*. Dla *k* = 2 obiekt zostanie zaklasyfikowany do klasy kwadratów. Z kolei przy *k* = 5 do trójkątów.

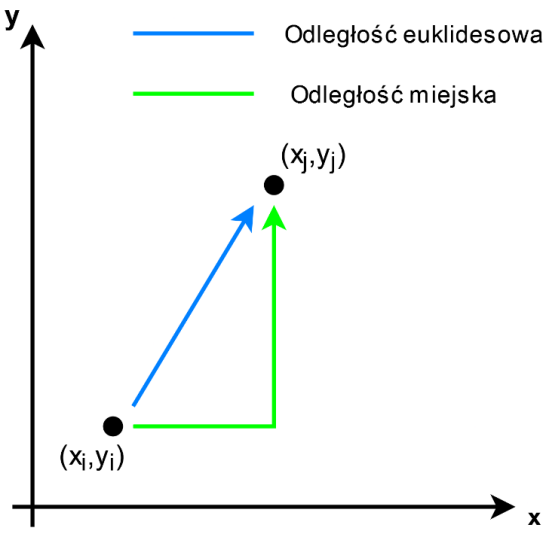


Rys. 4.10. Graficzne zobrazowanie idei działania metody *k* najbliższych sąsiadów

Zaletą metody najbliższych sąsiadów jest możliwość zastosowania innych metryk do obliczania odległości. Jedna z nich jest *metryka miejska* zamiennie nazywana metryką *Manhattan*, *Cityblock* bądź *Taxicab* [37]. Opisuje się ją wzorem

$$d_{ij} = \left\| x_i - x_j \right\| = \sum_{k=1}^{N} \left| x_{ik} - x_{jk} \right| \tag{4.21}$$

Odległość miejska definiowana jest jako suma różnic mierzonych wzdłuż konkretnych wymiarów (rys. 4.11). Jej nazwa wzięła się od podobieństwa do miast z kwadratową siatką ulic, gdzie taksówka może poruszać się tylko wzdłuż nich [151]. W odróżnieniu od najpopularniejszej metryki euklidesowej, odległość typu Manhattan minimalizuje wpływ pojedynczych dużych różnic, poprzez brak ich potęgowania [1].



Rys. 4.11. Metryki miejska i euklidesowa na płaszczyźnie kartezjańskiej

W opracowanym rozwiązaniu testowano wiele metryk (takich jak metryka euklidesowa, korelacyjna czy kosinusowa), ale metryka miejska okazała się najskuteczniejsza.

4.4.1. Metoda najbliższej średniej w zadaniu klasyfikacji

W toku prowadzonych badań autor zauważył, że walory dystynktywne cech behawioralnych ukazują się przy relatywnie dużej (w stosunku do cech fizycznych) ilości danych uczących. To spowodowało zastosowanie opisanego w rozdziale 2.1.5 sposobu segmentacji sygnału mowy, tj. cały przebieg czasowy głosu dzielony jest na trzysekundowe ramki z jednosekundowym przesunięciem. W takim przypadku dla minutowej próbki mowy otrzymano 58 fragmentów poddawanych analizie. W poniższej tabeli zaprezentowano schemat podziału całości przetwarzanego sygnału, tj. czas rozpoczęcia, zakończenia i trwania trzech pierwszych i trzech ostatnich ramek.

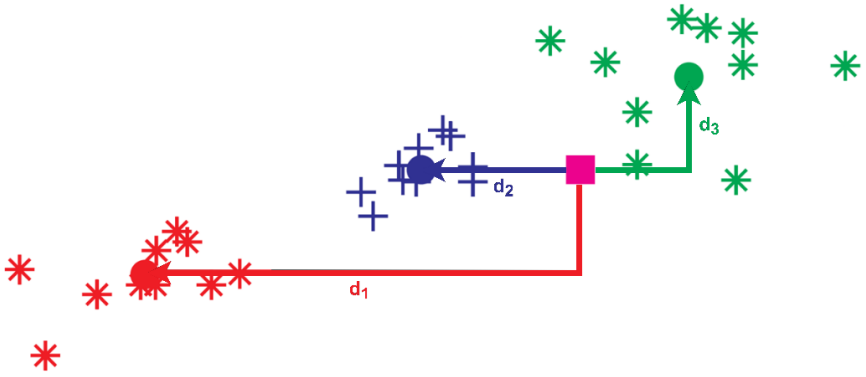
Tab. 4.4. Zestawienie czasów rozpoczęcia i zakończenia każdej z ramek

Numer ramki	*Czas rozpoczęcia* [s]	*Czas zakończenia* [s]	*Czas trwania* [s]
1	0	3	3
2	1	4	3
3	2	5	3
⋮	⋮	⋮	⋮
56	55	58	3
57	56	59	3
58	57	60	3

Taki sposób segmentacji powoduje, że jedna próbka głosu reprezentowana jest przez wiele obiektów (wektorów cech) wchodzących w skład tej samej klasy, tworząc przy tym swoisty obszar skupień [31]. W takim przypadku zastosowanie klasyfikatora *k-nn* (a konkretniej dobór odpowiedniej wartości *k*) nie jest łatwe. Rozwiązaniem tego problemu jest wykorzystanie *metody najbliższej średniej*. Polega ona na obliczeniu odległości (4.21) od aktualnie analizowanego wektora cech x_i do wektorów średnich $\overline{x}_j$ odpowiadających każdej *j*-tej klasie [38]. Stosując wspomnianą wcześniej metrykę miejską odległość do najbliższego, średniego zbioru cech oblicza się zgodnie z zależnością

$$d_{ij} = \sum_{k=1}^{N} \left| x_{ik} - \overline{x}_{jk} \right| \tag{4.22}$$

Idea metody najbliższej średniej widoczna jest na rysunku 4.12. Najbliżej testowego wektora cech (różowy kwadrat) jest klasa granatowa. W związku z wyznaczaniem wektorów średnich jako reprezentantów poszczególnych klas w zbiorze treningowym liczba danych potrzebna do zapamiętania na etapie uczenia uległa redukcji. Dla 1172 próbek głosu jest dokładnie tyle samo odległości co klas, podczas gdy w metodzie najbliższego sąsiada liczba ta rośnie do 67976.



Rys. 4.12. Metoda najbliższej średniej dla 3 różnych klas [2]

4.5. Podsumowanie

W niniejszym rozdziale scharakteryzowano opracowane rozwiązanie bazujące na behawioralnych cechach sygnału mowy w formie samodzielnego systemu ASR. Na wstępie autor odniósł się do procesu wstępnego przetwarzania sygnału mowy dokładniej opisanego w rozdziale 2. Następnie zwięźle scharakteryzował wpływ czynników zewnętrznych na parametry głosu i przedstawił przykłady cech, które zaimplementowane zostały do finalnej wersji algorytmu. W związku z dużą liczbą generowanych danych zdecydowano się na zastosowanie metod ewolucyjnych w celu redukcji wymiaru wektora cech, a co za tym idzie zmniejszenia kosztów obliczeniowych. W wyniku zastosowania algorytmu genetycznego nastąpiła znaczna redukcja wymiaru wektora cech przy jednoczesnej poprawie skuteczności klasyfikacji. Wraz z rozwojem całego systemu nastąpiła potrzeba rewizji, co przyczyniło się do zmiany wykorzystanego klasyfikatora. Pierwotnie zakładano zastosowanie metody najbliższych sąsiadów. Jednak badania eksperymentalne dowiodły, że przy tak dużej liczbie reprezentantów (58 wektorów cech osobniczych) pojedynczej klasy, to metoda najbliższej średniej daje lepsze rezultaty. Testowano również rozwiązanie bazujące na opisanych w rozdziale 3.4 modelach mieszanin gaussowskich, jednak bez powodzenia. Parametry niezbędne do prawidłowego działania systemu ASR, takie jak długość ramki czy czas trwania segmentów uczącego i testującego podlegały żmudnym procesom optymalizacji. Ostatecznie opracowane przez autora rozwiązanie zawiera 27 cech głosu, których wykorzystanie pozwala na ponad 68% poprawnych identyfikacji tożsamości mówcy w zbiorze 1172 głosów.

Aby móc obiektywnie ocenić zasadność opracowanego rozwiązania, należy je porównać z innym, dostępnym rozwiązaniem. Zgodnie z założeniem autora, za system referencyjny przyjęto rozwiązanie *ARMiA* [60]. Dodatkowo, aby wyniki porównań były rzetelne należy zastosować jednakowe czasy trwania segmentów uczących i testujących oraz te same bazy głosów. Do badań wykorzystano opisywaną wcześniej bazę SLR-12. Testy wykonano dla dwóch wariantów długości danych treningowych oraz testujących. W wariancie pierwszym przyjęto sumarycznie 30 sekund nagrania głosu podzielone odpowiednio na 25 sekund odcinka uczącego i 5 sekund odcinka testującego. Z kolei druga opcja to minuta mowy, gdzie 35 sekund przeznaczone było na dane treningowe, a pozostałe 25 sekund służy do testowania. Konkretne wartości wynikają z optymalizacji przeprowadzonej przez autorów tych rozwiązań. Optimum dla systemu *ARMiA* stanowi sumarycznie 30 sekund mowy, a dla algorytmu bazującego na behawioralnych cechach sygnału mowy, wartość optymalna to minuta próbki głosu. W tabeli 4.5 przedstawiono wyniki testów przeprowadzonych przez autora niniejszej rozprawy.

Tab. 4.5. Zestawienie sprawności systemów ASR w zależności od długości segmentów uczących i testujących

Czas uczenia / czas testowania [s]	***Sprawność systemu opartego na cechach behawioralnych [%]***	***Sprawność systemu ARMiA [%]***
25/5	24,91	86,26
35/25	68,43	98,21

Z przedstawionych wyżej wartości jasno widać wyższość rozwiązania bazującego na fizycznych charakterystykach głosu nad systemem opartym wyłącznie o cechy behawioralne. Dalsze wydłużanie czasów uczenia i testowania (aż do sumarycznych 5

minut mowy) poprawia skuteczność działania systemu opartego na behawioralnych cechach sygnału mowy do poziomu około 90% poprawnych identyfikacji tożsamości mówcy. Lepsze działanie systemu dopiero przy długich czasach wypowiedzi uniemożliwia praktyczną implementacje tego rozwiązania w warunkach rzeczywistych w systemach identyfikacji, ale może być brane pod uwagę w systemach weryfikacji, gdy nie ma ograniczeń czasowych, np. w zastosowaniach kryminalistycznych (badania fonoskopijne).

W kolejnych rozdziałach – w celu udowodnienia tezy rozprawy – przedstawiono możliwości wykorzystania opracowanych cech behawioralnych w roli wsparcia dla istniejącego już systemu *ARMiA*, w różnych konfiguracjach oraz z uwzględnieniem występowania różnych zewnętrznych zakłóceń i szumów.

5.

Fuzja systemów

Zaprezentowanie w poprzednim rozdziale wyniki badań miały na celu pokazanie możliwości dystynktywnych behawioralnych cech głosu w ujęciu automatycznego rozpoznawania mówcy. Osiągnięto sprawność autorskiego systemu na poziomie niecałych 70% przy analizie minutowych sygnałów mowy. Ponadto dokonano porównania liczby poprawnych identyfikacji tożsamości mówcy z istniejącym już systemem *ARMiA* [60]. Dla 30 i 60 sekund analizowanej mowy, system oparty na cechach behawioralnych jest odpowiednio o prawie 70 i 30% gorszy od rozwiązania bazującego na fizycznych charakterystykach głosu. Uzyskane wyniki jednoznacznie pokazują, że system ASR oparty jedynie na cechach behawioralnych nie jest wystarczająco skuteczny w warunkach rzeczywistych. W związku z tym, w niniejszym rozdziale przedstawiono rozwiązanie będące wynikiem fuzji opisanych wcześniej algorytmów automatycznego rozpoznawania mówcy.

5.1. Analiza rodzajów wykorzystywanych fuzji danych

Systemy automatycznego rozpoznawania mówcy traktować można jako systemy informacyjno-pomiarowe, których zadaniem jest zebranie możliwie największej liczby danych charakteryzujących badany obiekt (mówcę). Informacje z wielu takich „kanałów pomiarowych" można poddać procesom integracji, która może zostać przeprowadzona na wielu płaszczyznach [125]. Począwszy od warstwy sprzętowej i procesu akwizycji danych, poprzez sposoby przetwarzania informacji, kończąc na synergistycznym użyciu informacji pochodzących z różnych czujników. Ważnym aspektem jest odpowiednie zdefiniowanie procesu fuzji jako elementu składowego na każdej z powyższych płaszczyzn, gdzie dane zostają połączone w spójny, integralny zbiór [99], [100]. Dobór właściwego mechanizmu fuzji danych jest zadaniem skomplikowanym. W literaturze znaleźć można różne sposoby klasyfikacji procesu integracji [16], [94], [140]. Istnieje podział rozgraniczający trzy modele, które dotyczą:

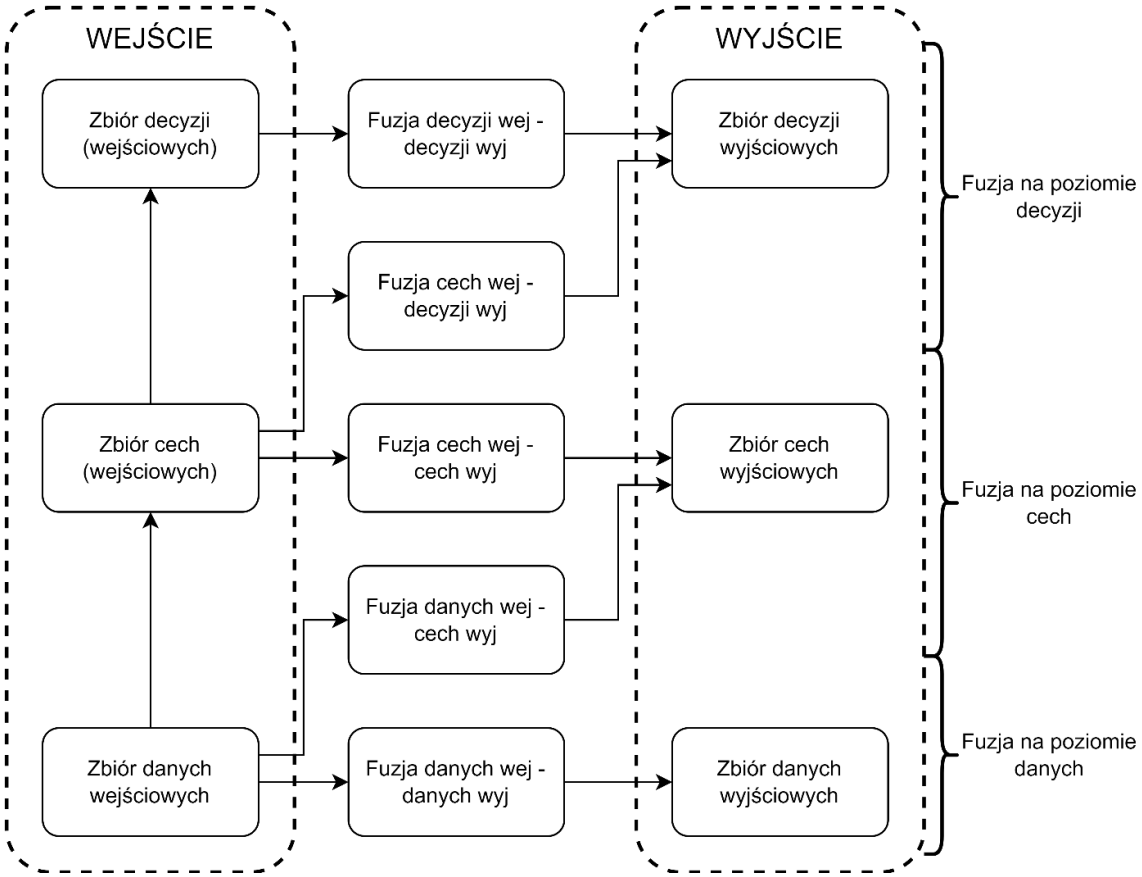
- poziomu abstrakcji integracji informacji [100],
- poziomu abstrakcji danych wejściowych lub/i wyjściowych [15],
- relacji między źródłami danych [40].

Pierwszy model wprowadza jeszcze dokładniejszy podział na fuzje *niskopoziomową*, *pośrednią* oraz *wysokopoziomową*. Wariant niskopoziomowy (ang. *low-level-fusion*) opiera się na operacji scalania danych pochodzących z wielu źródeł w jeden wyjściowy sygnał. Takie podejście pozwala na stworzenie danych o większym znaczeniu informacyjnym w stosunku do odrębnie analizowanych sygnałów pochodzących z integrowanych źródeł. Często zamiennie stosuję się tu nazwę *fuzja surowych danych* (ang. *raw-data-fusion*). Fuzja cech (ang. *feature-level-fusion*) jest drugą metodą integracji informacji, która bazuje na połączeniu parametrów wyekstrahowanych w trakcie przetwarzania sygnałów wejściowych przez różne „sensory". Podejście wysokiego poziomu (ang. *high-level-fusion*) zwane często fuzją decyzyjną (ang. *decision fusion*), to połączenie końcowych przewidywań (decyzji) dotyczących odpowiedniej klasyfikacji tego samego sygnału, ale uzyskanych za pomocą osobnych modułów decyzyjnych [17].

Model związany z abstrakcją danych wejściowych i/lub wyjściowych można przedstawić za pomocą rozszerzonego 5 stopniowego *schematu Dasarathy'ego*. Na rys. 5.1 zobrazowano możliwości fuzji danych według powyższego modelu. Pięć kategorii pokazanych na rys. 5.1 odnosi się szczegółowo do charakteru informacji na wejściu oraz wyjściu układu wraz ze sposobem ich odpowiedniej integracji [118]:

- Fuzja na poziomie surowych danych (ang. *Data In – Data Out*) – przyjmuje się, że surowe dane z różnych źródeł są bezpośrednio łączone, aby uzyskać bardziej kompletny zbiór danych. Jest to przykładowo agregacja danych z różnych sensorów przed jakąkolwiek wstępną obróbką, np. połączenie danych z kilku czujników temperatury w celu uzyskania bardziej precyzyjnego pomiaru średniej temperatury.
- Fuzja na poziomie cech i danych (ang. *Data In – Feature Out*) – w tym wariancie następuje przetwarzanie surowych danych w celu ekstrakcji cech, a następnie te cechy są łączone. Za przykład służyć może ekstrakcja cech z różnych zobrazowań medycznych, np.: RM (rezonans magnetyczny) i TK (tomografia komputerowa), a następnie ich połączenie w celu uzyskania bardziej szczegółowych informacji diagnostycznych.

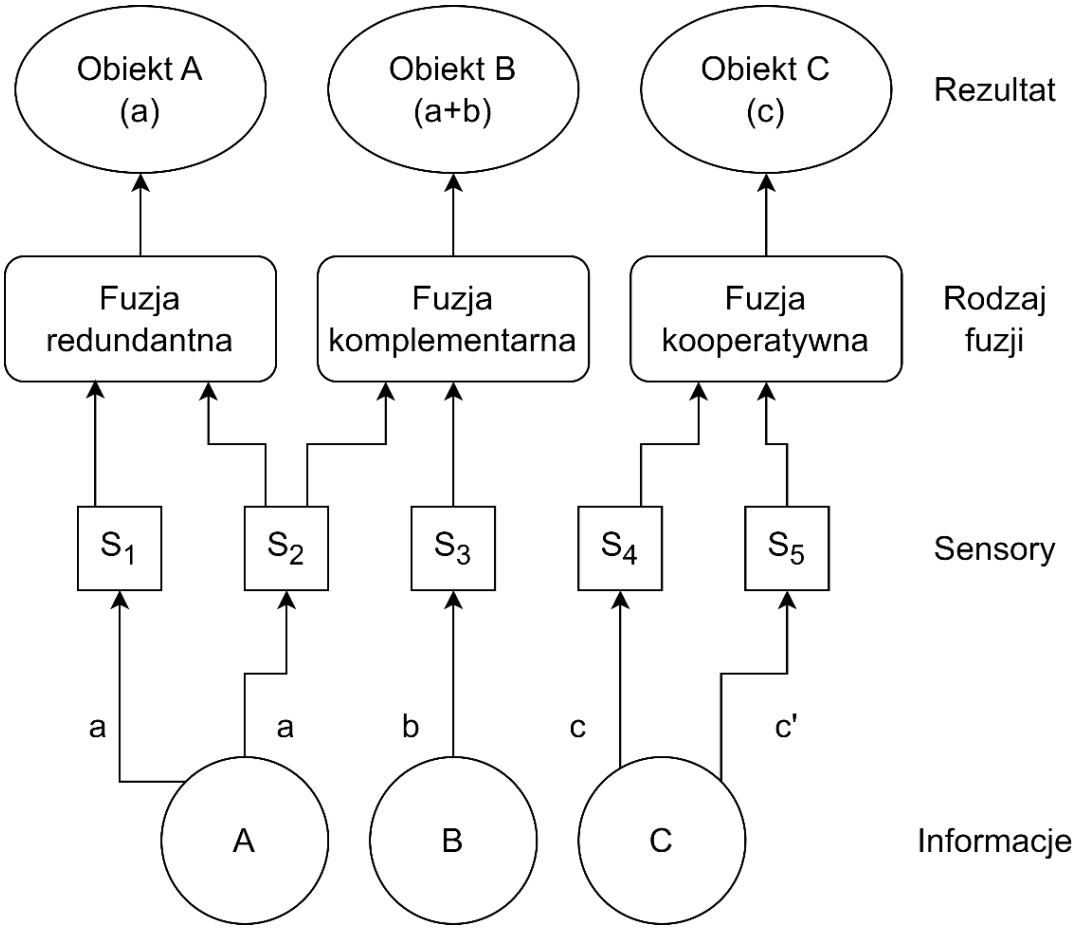
- Fuzja na poziomie cech (ang. *Feature In – Feature Out*) – wariant ten różni się od poprzedniego tym, iż wcześniej wyekstrahowane cechy łączone są bezpośrednio. Inaczej mówiąc następuje tylko sama integracja cech bez procesu ich ekstrakcji. Jest to poziom, na którym różne zestawy cech z różnych źródeł są integrowane w celu utworzenia bardziej kompleksowego zbioru cech. Przykładem może być proces łączenia parametrów pochodzących z różnych czujników wokół pojazdu, aby stworzyć dokładniejszą mapę otoczenia.
- Fuzja na poziomie decyzji (ang. *Feature In – Decision Out*) – poziom ten charakteryzuje się wykorzystaniem cech wejściowych do podjęcia decyzji niezależnie w każdym źródle danych (module decyzyjnym), a następnie decyzje te są łączone w celu uzyskania decyzji końcowej. Na przykład informacje z różnych systemów detekcji zagrożeń (analiza logów, podejrzana aktywność użytkownika itp.) wykorzystywane są do podejmowania decyzji o blokowaniu dostępu do usług.
- Fuzja na poziomie decyzji wyższego rzędu (ang. *Decision In – Decision Out*) – jest to najwyższy poziom fuzji, w którym gotowe, pośrednie decyzje z różnych źródeł są łączone w celu uzyskania końcowego wyboru. Ten sposób integracji wykorzystywany jest np. w systemach bezpieczeństwa łączących decyzje z różnych systemów detekcji intruzów, takich jak kamery czy czujniki ruchu, w celu podjęcia końcowej decyzji o alarmie.



Rys. 5.1. Model Dasarathy'ego wraz z kategoryzacją możliwości fuzji danych [17]

Inny model integracji danych – oparty o relacje między ich źródłami – opracowany został przez Durrant-Whyte'a [40]. Zaproponował on następujący podział metod fuzji danych (rys. 5.2):

- Fuzja komplementarna – w tym wariancie zakłada się, że istnieją różne, niezależne sensory, które dostarczają informacje dotyczącego tego samego zjawiska. Dzięki temu istnieje możliwość utworzenia zbioru danych dokładniej opisującego badany obiekt.
- Fuzja redundantna (zwana również kompetytywną) – u podstaw tej metody leży założenie, że kilka sensorów dostarcza informacje o tym samym parametrze badanego zjawiska.
- Fuzja kooperatywna – rodzaj metody integracji danych, w której informacja dotycząca obiektu pochodzi z więcej niż jednego sensora, jednocześnie tworząc bardziej skomplikowany zbiór danych.



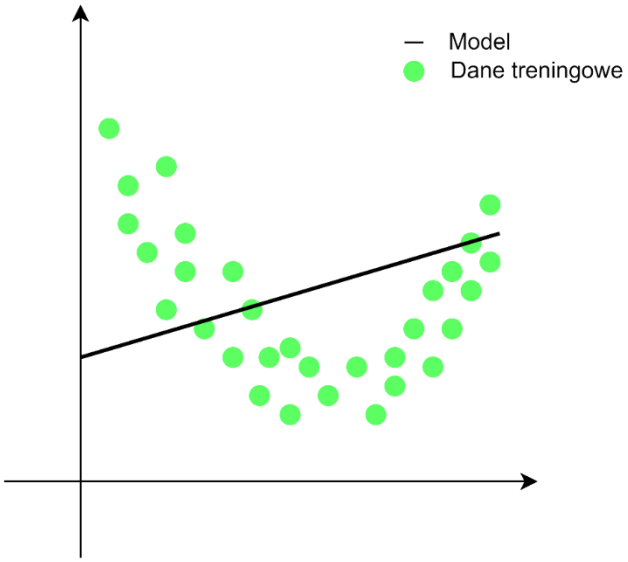
Rys. 5.2. Klasyfikacja metod fuzji według Durrant-Whyte'a [17], [40]

W celu zwiększenia liczby prawidłowych identyfikacji tożsamości mówcy przez system *ARMiA* autor zdecydował się zastosować techniki integracji danych. Jako że rozwiązanie autora oraz system autorstwa dr. Kamińskiego [60] różnią się pod względem charakteru (rodzaju) ekstrahowanych cech oraz sposobu klasyfikacji obiektów, autor niniejszej rozprawy postanowił opracować dwa, autorskie algorytmy fuzji danych. Pierwszy z nich działa na zasadzie integracji informacji w warstwie decyzji. Dokładniej rzecz ujmując, pierwszy etap działania algorytmu fuzji danych to wstępna klasyfikacja *N* najbliższych klas, które stanowią źródło informacji do podjęcia ostatecznej decyzji przy pomocy innego klasyfikatora. Drugie podejście bazuje na fuzji cech pochodzących z dwóch, niezależnych systemów automatycznego rozpoznawania mówcy. To znaczy, że na podstawie oddzielnych wektorów cech dystynktywnych mowy tworzony jest jeden

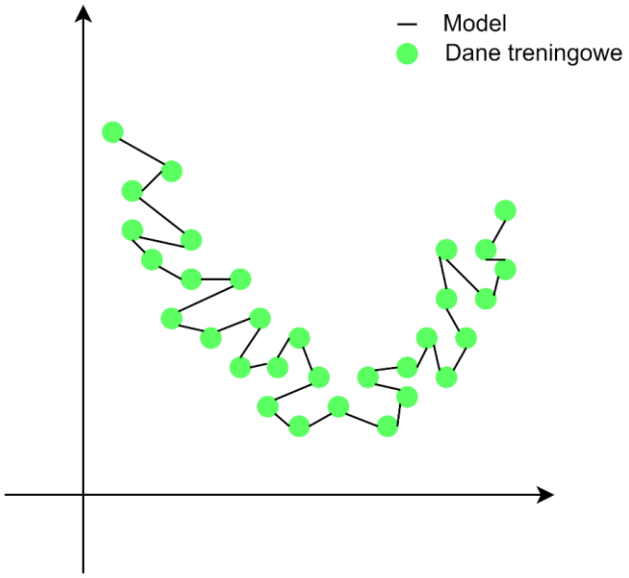
finalny zbiór, który stanowi podstawę procesu klasyfikacji obiektów. Poniżej opisane zostały oba rozwiązania.

5.2. Metody fuzji danych wsparte technikami sztucznej inteligencji

We współczesnym społeczeństwie kluczowe decyzje podejmowane są często po wcześniejszym zapoznaniu się z opiniami ekspertów w danym aspekcie. Wynika to z faktu, że jeden model decyzyjny (decydent) jest niedoskonały, tzn., że podjęta przez niego decyzja może być błędna. Zazwyczaj człowiek przed dokonaniem wyboru mającego istotne konsekwencje, szuka opinii różnych ekspertów, np. pacjent konsultuje wyniki badań z kilkoma lekarzami przed zgodą na poważną operację, a pracodawca sprawdza i analizuje referencje przed zatrudnieniem potencjalnego kandydata do pracy. Dzięki takim zabiegom możliwe jest dokonanie odpowiedniego (dla osoby decyzyjnej) wyboru. Takie podejścia znalazły odzwierciedlenie w technice sztucznej inteligencji o nazwie *ensemble learning*. Metoda ta charakteryzuje się lepszą zdolnością do generalizacji i predykcji w porównaniu do pojedynczego modelu decyzyjnego, gdyż finalna decyzja powstaje na skutek generacji i fuzji pomniejszych decyzji z wielu poszczególnych klasyfikatorów [126]. Każdy z nich charakteryzuje się dwoma parametrami, na które można wpłynąć, tj. błędem systematycznym (bias), czyli dokładnością klasyfikatora (rys. 5.3), oraz wariancją, czyli precyzja modelu decyzyjnego (rys. 5.4), gdy jest on trenowany na różnych zestawach danych. Często te dwa elementy pozostają w relacji kompromisowej: klasyfikatory o niskim błędzie systematycznym mają tendencję do wysokiej wariancji i odwrotnie. Wysoka wartość błędu systematycznego oznacza, że model zbytnio upraszcza problem (jest niedouczony) przez co nie może dopasować się do danych treningowych, a w rezultacie błędnie estymuje kolejne predykcje. Z kolei wysoka wariancja charakteryzuje model przeuczony, przez co nie może on przewidzieć następnych danych.



Rys. 5.3. Przypadek wysokiego błędu systematycznego modelu (high-bias)



Rys. 5.4. Przypadek wysokiej wariancji modelu (high-variance)

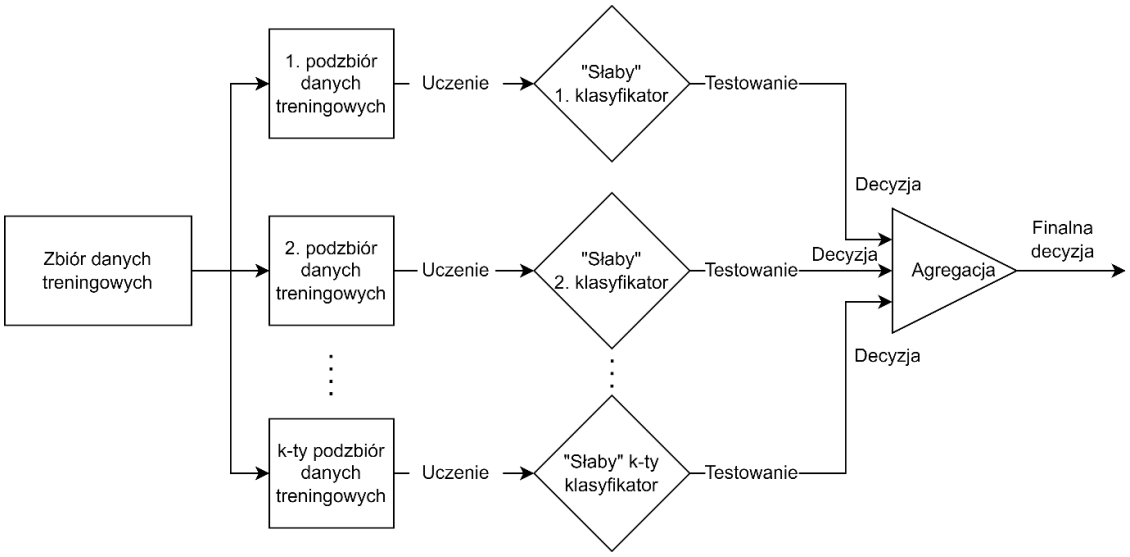
W związku z powyższym, celem rozwiązań bazujących na *ensemble learning* jest wykorzystanie kilku klasyfikatorów o relatywnie stałym (lub podobnym) błędzie systematycznym, a następnie połączenie ich wyników, (na przykład przez uśrednianie) w celu zmniejszenia wariancji [21]. Metody *ensemble learning* podzielić można na trzy główne grupy:

- bagging,
- boosting,
- stacking.

5.2.1. Bagging, boosting i stacking

Technika *baggingu*, a dokładniej ***b**ootstrap **agg**regat**ing*** opiera się na podziale danych treningowych na losowe podzbiory, gdzie każdy z nich jest wykorzystywany do trenowania innego klasyfikatora (tego samego typu). Następnie poszczególne modele decyzyjne są agregowane poprzez głosowanie większościowe, tj. decyzja podejmowana jest większością głosów [145]. Poniżej przedstawiony jest schemat procesu baggingu (rys. 5.5):

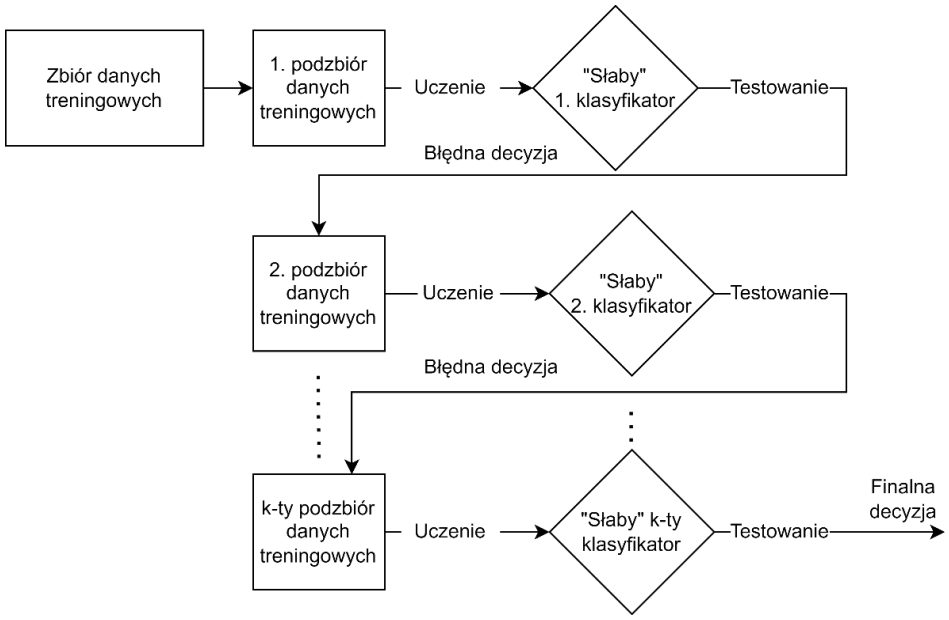
- Pierwszym krokiem działania algorytmu jest podział zbioru danych treningowych na *k* losowych podzbiorów. W każdym podzbiorze poszczególne dane mogą występować więcej niż raz (taki proces nosi nazwę *sample with replacement*).
- Wszystkie podzbiory stanowią dane uczące dla tzw.: „słabych” klasyfikatorów (wszystkie tego samego typu).
- Na wyjściu każdego z pomniejszych modeli, w wyniku procesu testowania powstaje wynik klasyfikacji (decyzja).
- Ostatnim etapem jest agregacja pomniejszych decyzji w jedną, finalną poprzez głosowanie większościowe czy uśrednianie.



Rys. 5.5. Proces baggingu [75]

Boosting to technika, której celem jest stworzenie optymalnego modelu decyzyjnego w określonej liczbie iteracji poprzez wykorzystanie mniejszych klasyfikatorów. Innymi słowy, sekwencyjnie budowany jest coraz bardziej precyzyjny system klasyfikacji, który opiera się głównie na poprawie błędów popełnionych przez wcześniejszy klasyfikator. Możliwe jest to poprzez przyporządkowanie wag danym treningowym na wejściu konkretnego klasyfikatora. Następnie po każdej iteracji procesu testowania współczynniki te są aktualizowane zgodnie z otrzymanymi wynikami, tj. czynniki wagowe błędnych klasyfikacji są zwiększane tak, aby kolejny model skupił się na trudnych przypadkach. Metoda ta zwana jest inaczej techniką *wzmocnienia* (*boosting*). Decyzja finalna podejmowana jest poprzez głosowane lub uśrednianie [21], [101]. Schemat procesu *boostingu* opisany jest poniżej (rys. 5.6):

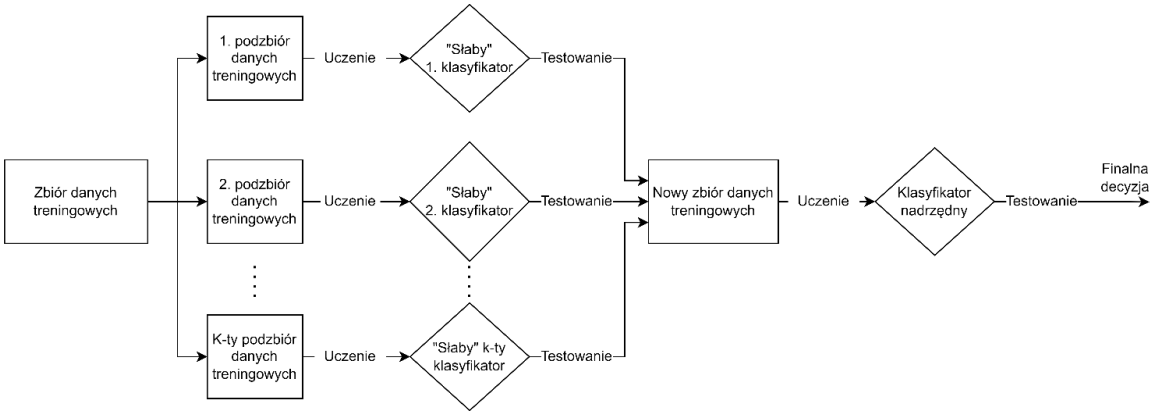
- Na początku następuje podział zbioru danych treningowych na *k* podzbiorów.
- Następnie rozpoczyna się proces uczenia oraz testowania pierwszego, „słabego” klasyfikatora.
- Błędne wyniki klasyfikacji są włączane do kolejnego podzbioru uczącego (który jest aktualizowany na podstawie rezultatów z wcześniejszych iteracji) i wykorzystywane przez następny klasyfikator.
- Proces uczenia i testowania jest powtarzany, (tym razem opierając się na nowym, zaktualizowanym podzbiorze) dla kolejnego klasyfikatora.
- Powyższe kroki są powtarzane, aż do osiągnięcia *k* iteracji.
- Wyniki uzyskane w *k*-tym kroku stanowią finalną decyzje, będącą wypadkową *k* poprzednich kroków uczenia i testowania klasyfikatorów.



Rys. 5.6. Proces fuzji danych na poziomie decyzji [75]

Podejście oparte na s*tackingu* zakłada wykorzystanie wielu różnych klasyfikatorów, gdzie każdy z nich trenowany jest na tym samym zbiorze danych. Uzyskane rezultaty podlegają agregacji
w celu utworzenia nowego zbioru danych uczących dla klasyfikatora nadrzędnego tzw.: *meta-model*. Ostateczna decyzja jest wynikiem testowania *meta-modelu* [24]. Poszczególne kroki procesu stackingu opisane są poniżej (rys. 5.7):

- Pierwszym etapem jest podział zbioru danych treningowych na *k* podzbiorów.
- Następnie rozpoczyna się proces uczenia oraz testowania *k* „słabych", różnych klasyfikatorów.
- Rezultaty uzyskane na drodze testowania wielu modeli decyzyjnych są agregowane tworząc nowy zbiór danych uczących dla *meta-modelu.*
- Klasyfikator nadrzędny (w oparciu o nowy podzbiór) poddawany jest procesom uczenia i testowania, a jego wynik stanowi finalną decyzję.

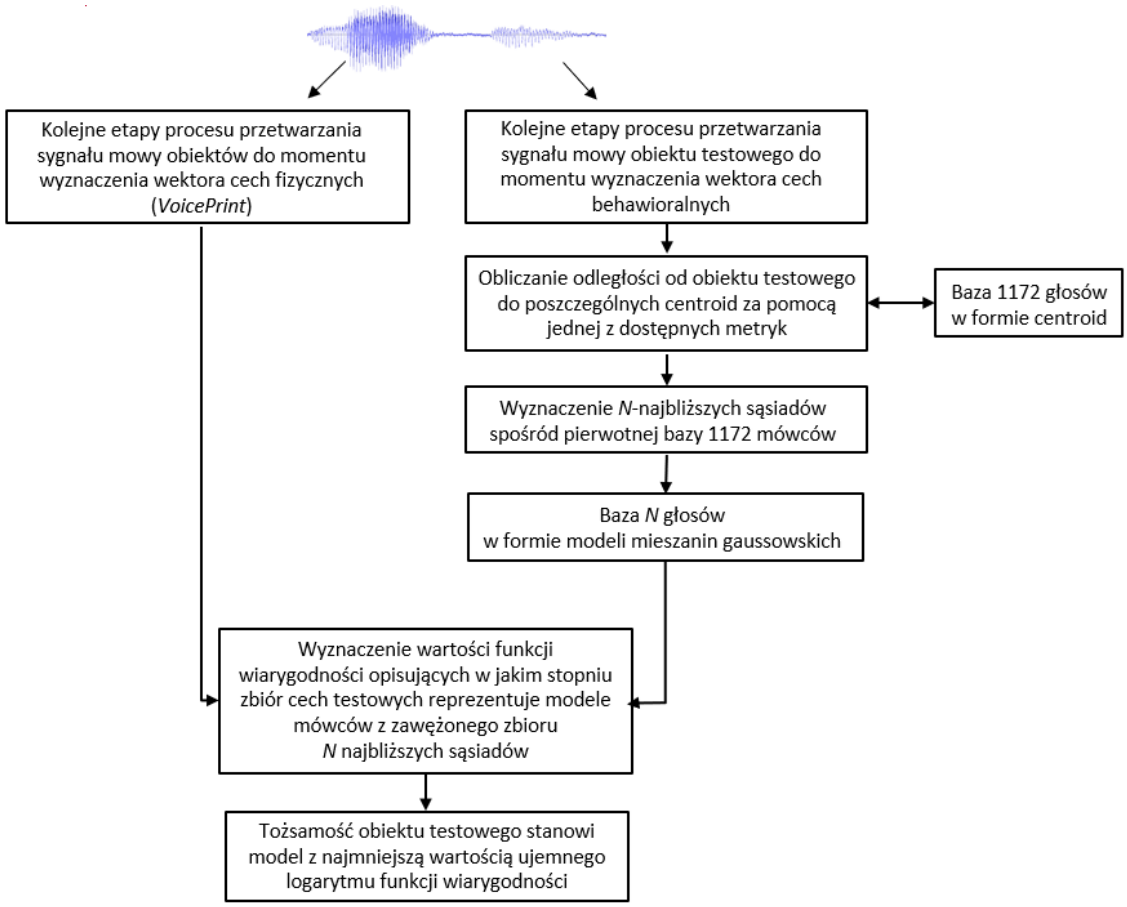


Rys. 5.7. Proces fuzji danych na poziomie decyzji [75]

W dziedzinie automatycznego rozpoznawania mówcy algorytmy te wykorzystywane są w głównej mierze wraz z sieciami neuronowymi [18], [43], [68], [71], [86], [88], [90], [108], [144], jednak nie wyklucza to możliwości ich stosowania z klasycznymi metodami uczenia maszynowego.

5.3. Metoda preselekcji najbliższej liczby sąsiadów – fuzja na poziomie decyzji

U podstaw tej metody fuzji danych leży zastosowanie odmiennych klasyfikatorów w obydwu systemach ASR. Rozwiązanie autorstwa dr. Kamińskiego [60] bazuje na modelach mieszanin gaussowskich (GMM) dokładniej opisanych w rozdziale 3.4. Z kolei algorytm autora niniejszej rozprawy do zadań klasyfikacji badanych obiektów wykorzystuje metodę najbliższej średniej (opis znajduję się w rozdziale 4.4). Podejście to pokazane jest na rys. 5.8.



Rys. 5.8. Proces fuzji danych na poziomie decyzji

Zgodnie z powyższym schematem proces fuzji rozpoczyna się już w momencie, gdy zostaną wyznaczone wektory cech fizycznych oraz behawioralnych badanego obiektu. Następnie tworzony jest model głosu badanego obiektu przy pomocy GMM. W tym samym czasie algorytm oparty na behawioralnych cechach sygnału mowy wyznacza *N* najbliższych centroid (poprzez obliczanie odległości od badanego obiektu do danych treningowych) spośród pełnego zbioru. Kolejnym krokiem jest redukcja pełnego zestawu modeli głosów (uczących) do wyznaczonych wcześniej *N* najbliższych sąsiadów. Dalej następuje już proces klasyfikacji zgodnie z implementacją w systemie *ARMiA* (moduł opisany dokładniej w rozdziale 3.4).

5.3.1. Wyniki działania fuzji na poziomie decyzji

W toku prowadzenia badań autor przetestował różne kombinacje sposobów integracji danych na poziomie decyzji. Finalnie jednak opisana wyżej metoda z wykorzystaniem modelowania i klasyfikacji mieszaninami Gaussa przyniosła najlepsze rezultaty. Ten sposób nazwany został techniką *preselekcji,* czyli wstępnego ograniczenia zbioru danych treningowych do *N* najbardziej podobnych do obiektu testowego (poprzez miarę odległościową). W tab. 5.1 oraz 5.2 pokazano wyniki zastosowania „preselektora" w wariantach 30 i 60 sekund mowy[1]. Na zielono zaznaczono liczbę poprawnych identyfikacji wyższą niż w podstawowym wariancie działania systemu *ARMiA* dla każdego z sumarycznych czasów (tj.: 86,26 oraz 98,21%).

Tab. 5.1. Wartości sprawności systemu ASR opartego na cechach fizycznych z zastosowanym behawioralnym preselektorem dla sumarycznego czasu mowy 30 sek

	Zastosowana metryka	
Liczba najbliższych sąsiadów	**Euklidesowa**	**Miejska**
25	66,89 %	70,99 %
50	76,79 %	78,58 %
75	79,27 %	81,74 %
100	82,68 %	84,90 %
125	84,47 %	86,52 %
150	86,26 %	87,12 %
175	86,69 %	87,63 %
200	87,37 %	88,05 %
225	87,63 %	88,48 %
250	87,80 %	88,31 %
275	87,88 %	88,74 %
300	88,23 %	88,65 %
325	88,48 %	88,74 %
350	88,40 %	88,65 %
375	88,57 %	88,82 %
400	88,57 %	88,82 %
425	88,31 %	88,48 %
450	88,05 %	88,40 %
475	87,71 %	88,23 %
500	87,71 %	87,97 %
525	87,71 %	88,05 %
550	87,54 %	87,88 %
575	87,29 %	87,71 %
600	87,46 %	87,80 %

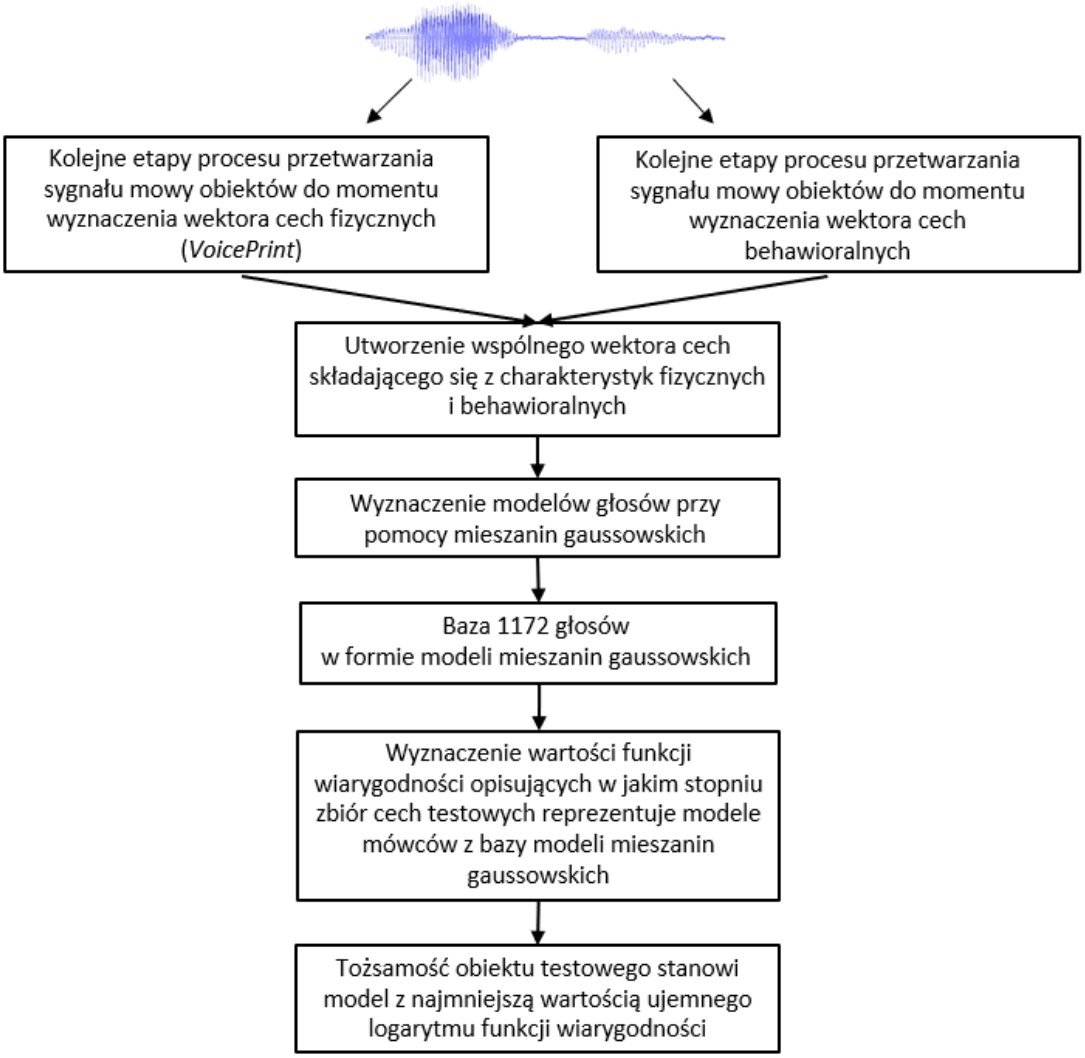
[1] W przypadku 30 sekundowego sygnału mowy, 25 sekund sygnału wykorzystano do uczenia systemu, a pozostałe 5 sekund do testowania. Wariant 60 sekund sygnału mowy wykorzystuję odpowiednio 35 i 25 sekund.

Tab. 5.2. Wartości sprawności systemu ASR opartego na cechach fizycznych z zastosowanym behawioralnym preselektorem dla sumarycznego czasu mowy 60 sek

	Zastosowana metryka	
Liczba najbliższych sąsiadów	**Euklidesowa**	**Miejska**
25	95,48 %	95,82 %
50	96,33 %	96,67 %
75	97,01 %	97,35 %
100	97,44 %	98,21 %
125	97,61 %	98,46 %
150	97,87 %	98,38 %
175	97,87 %	98,29 %
200	97,95 %	98,38 %
225	97,95 %	98,55 %
250	98,04 %	98,55 %
275	98,04 %	98,55 %
300	98,12 %	98,81 %
325	98,21 %	98,81 %
350	98,29 %	98,81 %
375	98,21 %	98,81 %
400	98,38 %	98,81 %
425	98,46 %	98,72 %
450	98,46 %	98,72 %
475	98,46 %	98,72 %
500	98,46 %	98,72 %
525	98,55 %	98,72 %
550	98,46 %	98,72 %
575	98,46 %	98,72 %
600	98,46 %	98,72 %

5.4. Metoda połączenia zestawów cech – fuzja na poziomie cech

Druga z metod integracji danych polega na fuzji dwóch wektorów cech dystynktywnych w celu otrzymania jednego, wypadkowego zbioru cech. Podyktowane jest to odmiennym charakterem cech fizycznych i behawioralnych. Ten sposób konsolidacji danych pozwala na utworzenie zasobniejszego w informacje dystynktywne wektora cech, który lepiej charakteryzuje badany obiekt. Proces fuzji danych na poziomie cech zobrazowany jest na rys. 5.9. W tym rozwiązaniu wszelkie etapy przetwarzania sygnału mowy wraz z jego parametryzacją następują w każdym z „podsystemów" (opartych na cechach behawioralnych i fizycznych) osobno. Następnie wyjściowe wektory charakterystyk dystynktywnych integrowane są ze sobą w celu stworzenia jednego, bogatszego zbioru cech badanego obiektu. Kolejne kroki działania systemu ASR, takie jak proces modelowania głosu mówcy, czy jego klasyfikacja są analogiczne jak w rozwiązaniu *ARMiA* [60].



Rys. 5.9. Proces fuzji danych na poziomie cech

5.4.1. Wyniki działania fuzji na poziomie cech

W ramach prowadzonych badań autor przetestował różne warianty przetwarzania następujące po fuzji danych, czyli warianty modelowania i klasyfikacji. Rozpatrywał m.in. czy zastosować techniki zaimplementowane w rozwiązaniu opartym na cechach behawioralnych, czy na cechach fizycznych, czy na ich jakiejś kombinacji. Najlepsze rezultaty uzyskano przy wykorzystaniu metod stosowanych w systemie *ARMiA* [60], tj. poprzez klasyfikator GMM. Jednak dodatkowym aspektem badań, który należało wziąć pod uwagę był sposób konsolidacji dwóch odmiennych wektorów cech. Ponieważ zbiory charakterystyk behawioralnych oraz fizycznych różnią się od siebie nie tylko charakterem cech, ale także ich licznością, wybór metody ich integracji stał się kluczowym aspektem procesu fuzji danych. Zbiór cech fizycznych opisać można jako macierz W_f (5.1) o wymiarach *M x N,* gdzie *M* stanowi liczbę analizowanych ramek, a *N*=23 liczbę cech.

$$W_f = \begin{bmatrix} C_{f1_1}, & C_{f2_1}, & \cdots & C_{fN_1} \\ C_{f1_2}, & C_{f2_2}, & \cdots & C_{fN_2} \\ \vdots & \vdots & \ddots & \vdots \\ C_{f1_M}, & C_{f2_M}, & \cdots & C_{fN_M} \end{bmatrix} \tag{5.1}$$

Z kolei cechy behawioralne przedstawić można w postaci wektora W_b składającego się z H=27 elementów reprezentujących kolejne cechy behawioralne

$$W_b = \begin{bmatrix} C_{b1}, & C_{b2}, & \cdots, & C_{bH} \end{bmatrix} \tag{5.2}$$

W związku z opisanymi wyżej różnicami w zbiorach cech autor przetestował dwie metody fuzji danych. Pierwsza z nich opiera się na połączeniu macierzy W_f oraz W_b w taki sposób, aby dołączony do macierzy cech fizycznych, wektor cech behawioralnych stanowił kolejne cechy badanego obiektu, ale tylko w jednej ramce, tj. charakterystyki behawioralne zawsze dotyczyć będą tylko pierwszej ramki. Odbywa się to poprzez mnożenie wektora pionowego W_{k_1}, którego liczba wierszy jest równa liczbie przetwarzanych w rozwiązaniu dr. Kamińskiego ramek, z wektorem poziomym W_b i dołączenie powstałego wyniku do macierzy W_f:

$$\begin{aligned} W_{k_1} &= \begin{bmatrix} 1, & 0, & \cdots, & 0 \end{bmatrix}^T \\ W_b &= \begin{bmatrix} C_{b1}, & C_{b2}, & \cdots, & C_{bH} \end{bmatrix} \\ W_{fb_1} &= \begin{bmatrix} C_{f1_1}, & C_{f2_1}, & \cdots & C_{fN_1} & C_{b1}, & C_{b2}, & \cdots & C_{bH} \\ C_{f1_2}, & C_{f2_2}, & \cdots & C_{fN_2} & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{f1_M}, & C_{f2_M}, & \cdots & C_{fN_M} & 0 & 0 & \cdots & 0 \end{bmatrix} \end{aligned} \tag{5.3}$$

gdzie: C_{fk} to k-ta cecha fizyczna, a C_{bk} to k-ta cecha behawioralna. Jej rzeczywista struktura pokazana została na rys. 5.10, na którym widać, że 23 pierwsze kolumny macierzy to cechy fizyczne, a kolejne 4 stanowią charakterystyki behawioralne.

21	22	23	24	25	26	27
-0.4922	-0.1269	0.1812	-0.1573	-0.1695	-0.3600	0.2406
-0.5188	0.0066	0.1770	0	0	0	0
-0.5220	-0.0866	-0.0477	0	0	0	0

Rys. 5.10. Postać macierzy W_{fb_1} (fragment)

Drugi sposób fuzji danych polegał na symulacji wyznaczania wektora cech behawioralnych dla każdej z ramek przetwarzanych w systemie *ARMiA* [60]. Odbywało to się poprzez mnożenie macierzy kolumnowej W_{k_2}, której każdy wiersz poza pierwszym posiada wartość losową z przedziału $(0{,}95 \div 1{,}05)$, z macierzą W_b. Następnie iloczyn wektorów dołączono do macierzy W_f. Wynikowa macierz W_{fb_2} opisana jest wzorem 5.4.

$$W_{k_2} = \left[1, \quad X_1, \quad \cdots \quad X_M\right]^T$$
$$W_b = \left[C_{b1}, \quad C_{b2}, \quad \cdots, \quad C_{bH}\right] \tag{5.4}$$
$$W_{fb_2} = \begin{bmatrix} C_{f1_1}, & C_{f2_1}, & \cdots & C_{fN_1} & C_{b1}, & C_{b2}, & \cdots & C_{bH} \\ C_{f1_2}, & C_{f2_2}, & \cdots & C_{fN_2} & X_1 \cdot C_{b1} & X_1 \cdot C_{b2} & \cdots & X_1 \cdot C_{bH} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{f1_M}, & C_{f2_M}, & \cdots & C_{fN_M} & X_M \cdot C_{b1} & X_M \cdot C_{b2} & \cdots & X_M \cdot C_{bH} \end{bmatrix}$$

Przykładowa macierz W_{fb_2} pokazana jest na rys 5.11.

21	22	23	24	25	26	27
-0.4922	-0.1269	0.1812	-0.1562	-0.1683	-0.3574	0.2390
-0.5188	0.0066	0.1770	-0.1530	-0.1648	-0.3501	0.2340
-0.5220	-0.0866	-0.0477	-0.1557	-0.1678	-0.3563	0.2382

Rys. 5.11. Postać macierzy W_{fb_2} (fragment)

W tab. 5.3 zestawiono wyniki działania systemu *ARMiA* [60] oraz rozwiązań bazujących na fuzji cech przy wykorzystaniu 30 i 60 sekund mowy. System *ARMiA* autorstwa dr. Kamińskiego [60] opiera się na cechach zawartych w macierzy W_f (5.1) i na ich podstawie uzyskano wyniki sprawności zawarte w drugiej kolumnie. Z kolei wartości znajdujące się w trzeciej kolumnie, to rezultat zastosowania zbioru cech dystynktywnych z macierzy W_{fb_1}. Ostatnia kolumna tabeli 5.3 przedstawia wyniki fuzji cech, podczas której symulowano proces wyznaczania cech behawioralnych dla wszystkich ramek, tj. za zbiór charakterystyk dystynktywnych przyjęto macierz W_{fb_2}.

Tab. 5.3. Zestawienie sprawności systemów ASR w zależności od długości segmentów uczących i testujących

Czas uczenia / czas testowania [s]	***Sprawność systemu ARMiA [%]***	***Sprawność fuzji na poziomie cech [%] – wariant 1 ramki***	***Sprawność fuzji na poziomie cech [%] – wariant wszystkich ramek***
25/5	86,26	57,25	23,38
35/25	98,21	97,25	69,03

Z powyższej tabeli jasno wynika, że wariant fuzji, gdzie symulowana jest ekstrakcja cech behawioralnych tylko dla jednej ramki sygnału ma znaczną przewagę pod kątem liczby prawidłowych identyfikacji mówcy. Pomimo stworzenia nowego zestawu deskryptorów służących do opisu mówcy, nie udało się osiągnąć sprawności wyższej niż w przypadku pierwotnego systemu *ARMiA* [60].

5.5. Podsumowanie

Na podstawie opisanych wyżej eksperymentów widać, że najlepszą (pod kątem wzrostu liczby poprawnych identyfikacji tożsamości) metodą integracji danych jest fuzja na poziomie decyzji. Zaproponowane rozwiązanie umożliwiło redukcję liczby błędów ze 161 do 132 w zbiorze 1172 mówców dla sumarycznego czasu mowy wynoszącego 30 sekund (25 sekund sygnału wykorzystano do uczenia systemu, a pozostałe 5 sekund do testowania). W przypadku 60 sekundowego sygnału mowy (odpowiednio 35 i 25 sekund)

liczba błędów została zredukowana z 21 do 14 w tym samym zbiorze 1172 osób. O ile rozwiązanie to pozwala na znaczną poprawę liczby poprawnych identyfikacji tożsamości o tyle nie jest ono pozbawione wad. Największym problemem jest określenie odpowiedniej liczby najbliższych sąsiadów przy dalszym douczaniu systemu – nie istnieje jedna granica, której zastosowanie w każdym wariancie zwiększa sprawność systemu. W powyższych badaniach zaproponowano liczbę 600 najbliższych sąsiadów (zawężenie przestrzeni analizowanych danych treningowych o około połowę – z 1172 do 600).

Rozwiązaniem niedogodności związanej z wyborem optymalnej liczby najbliższych sąsiadów jest zastosowanie fuzji na poziomie cech. Badania przeprowadzone dla wszystkich 50 cech (23 cechy fizyczne oraz 27 cech behawioralnych[1]) pokazują, że wyniki (poza krytycznym przypadkiem 25 sekund uczenia i 5 testowania) nie uległy znacznemu pogorszeniu. W związku z tym autor w kolejnym etapie badań ponownie wykorzystał metody ewolucyjne w celu selekcji najlepszych możliwych charakterystyk spośród pierwotnych 50 cech. Ponadto rozszerzył on pierwotną bazę głosów o ponad 500 dodatkowych próbek oraz sprawdził działanie fuzji w warunkach szumowych.

[1] Proces selekcji cech behawioralnych opisany został w rozdziale 4.3. Pierwotny zbiór zawierał 71 parametrów (ich dokładny opis znajduje się w załączniku Z.1), który przy pomocy metod ewolucyjnych został zredukowany do finalnego zbioru liczącego 27 deskryptorów.

6.

Badania eksploatacyjne systemu ASR

Proces tworzenia systemu ASR opartego na behawioralnych cechach sygnału mowy opisany został w rozdziałach 2-4. W rozdziale 4 przeprowadzono badania mające na celu ocenę możliwości dystynktywnych cech behawioralnych w zastosowaniu do automatycznego rozpoznawania mówcy. Na ich podstawie potwierdzono, że rozwiązanie bazujące wyłącznie na cechach behawioralnych nie jest w stanie konkurować z istniejącymi już systemami ASR [60]. Dlatego też autor niniejszej rozprawy zdecydował się na wykorzystanie wybranych i przetestowanych cech do wsparcia systemu *ARMiA* [60]. Sposób realizacji tego zagadnienia opisany został w rozdziale 5. Poniżej zaprezentowano wyniki wykorzystania technik integracji danych z dwóch niezależnych systemów ASR w bazie powiększonej z 1172 do 1688 głosów. Ponadto przeprowadzono badania w warunkach zaszumienia, tj. do analizowanych sygnałów mowy dodawany był szum utrudniający proces identyfikacji tożsamości.

6.1. Bazy głosów rozszerzające podstawowy zbiór SLR-12

Baza SLR-12 wykorzystywana do wstępnych badań opisana została w rozdziale 2.5. Zawiera ona 1172 próbki głosu różnych mówców. Autor postanowił rozszerzyć zbiór o ponad 500 dodatkowych sygnałów mowy pochodzących ze zróżnicowanych baz. W ich skład wchodzi: jedna baza komercyjna, jedna baza utworzona z polskojęzycznych audiobooków oraz trzy bazy głosów utworzone w Wojskowej Akademii Technicznej. Finalnie utworzony zbiór głosów zawiera 1688 próbek, spośród których:

- 1172 pochodzą z bazy SLR-12,
- 320 pochodzi z bazy NIST SRE 2002,
- 56 pochodzi z bazy utworzonej na podstawie audiobooków,
- 85 pochodzi z bazy autorstwa dr inż. Eweliny Majdy-Zdancewicz,
- 24 pochodzą z bazy autorstwa mgr. inż. Daniela Posiadały,
- 31 pochodzi z bazy autorstwa mgr. inż. Michała Bojszy.

Wszystkie dźwięki zostały odpowiednio przygotowane do dalszej obróbki, tj. poddane procesom opisanym w rozdziale 2.5, do których zaliczyć można:

- konwersję nagrań zapisanych w różnych formatach do nieskompresowanych monofonicznych plików *.wav,
- standaryzację sygnałów dla każdego mówcy z osobna,
- przepróbkowanie szybkości próbkowania do wartości 8 kS/s,
- usunięcie długich fragmentów ciszy,
- selekcję nagrań w taki sposób, aby uzyskać co najmniej minutę naturalnej mowy.

6.1.1. Baza głosów NIST 2002 SRE

Baza NIST 2002 SRE to zbiór zaprojektowany przez Linguistic Data Consortium (LDC) oraz National Institute of Standards and Technology (NIST) [149]. Zawiera on około 157 godzin nagrań mowy w języku angielskim. Źródłami danych są nagrane rozmowy telefoniczne, spotkania czy programy telewizyjne (wiadomości). Baza składa się z łącznie 9 153 plików mowy (6098 zostało zarejestrowanych z szybkością próbkowania 8 kS/s, natomiast pozostałe 3055 z szybkością 16 kS/s) zapisanych w formacie **.sph* (ang. *SPHere format*).

6.1.2. Baza stworzona na podstawie audiobooków

Jednym z wykorzystanych zbiorów głosów były nagrania dźwiękowe stanowiące odczytywane przez różnych lektorów teksty książek (ang. audiobooks) [9]. Są to prywatne zbiory autora niniejszej rozprawy, w skład których wchodzi 56 audiobooków. Pozycje literaturowe obecne w bazie danych to m.in.: *„Pan Wołodyjowski”* autorstwa Henryka Sienkiewicza czy „*Robinson Crusoe*” Daniela Defoe. Nagrania zrealizowane zostały w języku polskim oraz angielskim i zarejestrowane z szybkością 44,1 kS/s w formacie *.mp3.

6.1.3. Baza głosów autorstwa dr inż. Eweliny Majdy-Zdancewicz

Zbiór głosów autorstwa dr inż. Eweliny Majdy-Zdancewicz jest jedną z kilku baz utworzonych w Wojskowej Akademii Technicznej na potrzeby prowadzenia badań w zakresie przetwarzania mowy. Nagrania rejestrowane były jednokanałowo z szybkością próbkowania 25050 S/s przy 16-bitowej rozdzielczości amplitudowej

z użyciem mikrofonu dynamicznego Moncor DM-500. Dodatkowo zastosowano osłonę ograniczającą zniekształcenia, które powstają podczas artykulacji głosek świszczących (s, c, ś, ć, sz, cz) oraz wybuchowych (p, b). Autorka bazy głosów opracowała scenariusz składający się z czterech bloków o różnorodnym zabarwieniu emocjonalnym. Są to [31]:

- dialog, związany z podaniem fikcyjnych danych osobowych oraz celu podróży,
- oficjalna wypowiedź dot. ustawowych definicji z obszaru organizacji studiów wyższych,
- fragment „*Medalionów*" Z. Nałkowskiej, odzwierciedlający przygnębienie rozmówcy,
- ostatni blok to wypowiedź żartobliwa.

6.1.4. Baza głosów autorstwa mgr. inż. Daniela Posiadały

Zbiór głosów autorstwa mgr. inż. Daniela Posiadały to kolejna baza utworzona w Wojskowej Akademii Technicznej na potrzeby prowadzenia badań w zakresie przetwarzania mowy. Autor stworzył tzw. multisesyjną bazę głosów, czyli środowisko testowe pozwalające na weryfikację skuteczności rozpoznawania mówcy w warunkach zróżnicowanych torów akustycznych. Lektorzy odczytywali tekst oficjalny, żartobliwy, przygnębiający oraz dialog. Każdy z mówców zarejestrowany został 10-krotnie, z wykorzystaniem różnych torów nagraniowych. Nagrania zostały zarejestrowane jednokanałowo z 16-bitową rozdzielczością amplitudową oraz przy szybkości próbkowania 8 kS/s [60], [61].

6.1.5. Baza głosów autorstwa mgr. inż. Michała Bojszy

Zbiór głosów autorstwa mgr. inż. Michała Bojszy to najnowsza baza utworzona w Wojskowej Akademii Technicznej na potrzeby prowadzenia badań w zakresie przetwarzania mowy. Stworzona baza danych składa się z 62 nagrań należących do 31 użytkowników różniących się narodowością, płcią i wiekiem. Dodatkowo ze względu na ówcześnie panującą pandemię wirusa COVID-19 i faktu, że większość ludzi zasłaniała swoje usta maseczkami ochronnymi, każdy z mówców wykonał dodatkowo nagranie również w maseczce ochronnej. Czytano jednakowe teksty w języku polskim i angielskim, w zależności od narodowości mówcy. Ze względu na fakt konieczności zgromadzenia nagrań mówców z wielu zakątków świata i ujednolicenie formatów plików nagraniowych, konieczne było użycie środowiska do nagrywania głosów, które jest łatwo dostępne i wygodne z punktu widzenia użytkownika. W celu zgromadzenia nagrań autor wykorzystał platformę BITRES (rys. 6.1), która pozwala na nagranie 90 sekundowych plików dźwiękowych w formacie *.wav.

Rys. 6.1. Logo platformy BITRES

Z racji zdalnej rejestracji sygnałów mowy otrzymane nagrania znacząco różnią się od siebie jakością. Wynika to z faktu wykorzystania przez lektorów, zróżnicowanego

jakościowo sprzętu oraz często niesprzyjających warunków środowiskowych podczas realizacji nagrań [72].

6.2. Sposób prezentacji wyników testów

W celu oceny działania opracowanych rozwiązań w aspekcie wsparcia istniejącego systemu *ARMiA* [60] przeprowadzono szereg eksperymentów w różnych warunkach pracy. Badania prowadzone były dla różnych długości wypowiedzi uczących oraz testujących[1]. Badano kolejno system *ARMiA* [60], rozwiązanie oparte na cechach behawioralnych, fuzję obu systemów na poziomie cech oraz na poziomie decyzji dla 1000[2] najbliższych sąsiadów. Dodatkowo przed przeprowadzeniem wspomnianych wyżej testów autor niniejszej rozprawy jeszcze raz wykorzystał algorytm genetyczny (w tym rozdziale w skrócie nazywany *AG*) do selekcji cech dla wariantu fuzji danych na poziomie cech. Wynikiem tej operacji było otrzymanie nowego zbioru 36 charakterystyk będących zbiorem najbardziej dystynktywnych cech spośród wszystkich deskryptorów. Wśród nich 19 to cechy fizyczne, a pozostałe 17 pochodzą z systemu opartego na atrybutach behawioralnych.

Eksperymenty prowadzone były dla nagrań:

- niezakłóconych (wysokiej jakości),
- zniekształconych poprzez dodanie zewnętrznych zakłóceń,
- skompresowanych przez powszechnie wykorzystywane kodeki w kanałach telekomunikacyjnych.

Ponadto autor przetestował sprawność działania rozwiązań fuzji danych w przypadku zastosowania zróżnicowanych torów akustycznych.

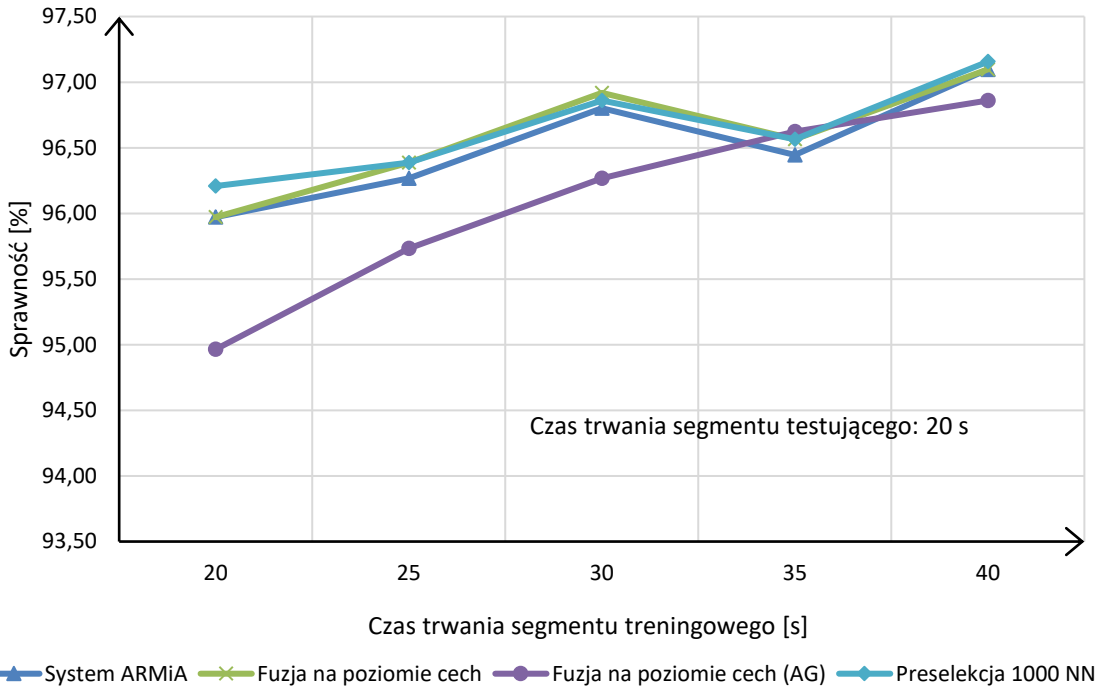
Ze względu na dużą liczbę uzyskanych rezultatów, w niniejszym rozdziale przedstawiono tylko te, które otrzymano dla najdłuższego czasu trwania segmentu testującego. Wyniki zaprezentowano w formie tabelarycznej wraz z odpowiadającymi im wykresami. Na zielono zaznaczono liczbę poprawnych identyfikacji tożsamości wyższą niż w systemie odniesienia, tj. w systemie *ARMiA* [60]. Z kolei kolorem pomarańczowym oznaczono wartości równe. We wszystkich zaprezentowanych poniżej przebiegach (poza rys 6.2 oraz 6.3) pominięto rezultaty uzyskane dla samodzielnego rozwiązania bazującego na cechach behawioralnych, ze względu na znacznie odbiegające od reszty wariantów wyniki (co zaburzałoby percepcję bardzo zbliżonych do siebie wartości sprawności uzyskanych przy pozostałych podejściach). Całość otrzymanych wyników zawarta została w formie tabel w załączniku 3.

[1] Każdy z eksperymentów (poza testem dla różnych urządzeń rejestrujących) przeprowadzony był dla długości segmentu testującego równej 10, 15 oraz 20 sekund. Ponadto czas przeznaczony na segment treningowy wynosił odpowiednio od 20 do 40 sekund z krokiem co 5 s.

[2] Liczbę najbliższych sąsiadów zwiększono z 600 do 1000 z powodu wyników dalszych badań eksperymentalnych, które wykazały, że w okolicach liczby 1000 najbliższych sąsiadów osiągano najwyższe wyniki sprawności.

6.3. Testy opracowanych rozwiązań w warunkach bezszumowych

Pierwszy eksperyment przeprowadzono dla pierwotnej, niezakłóconej bazy SLR-12 oraz jej rozszerzonej wersji opisanej w rozdziale 6.1. Rezultaty przedstawiono tylko dla poszerzonego zbioru głosów (rys. 6.2). Każdy z przebiegów dotyczy stałego czasu trwania segmentu testującego, który wynosi 20 sekund.



Rys. 6.2. Skuteczność identyfikacji mówcy różnych rozwiązań ASR dla zróżnicowanej długości segmentu treningowego

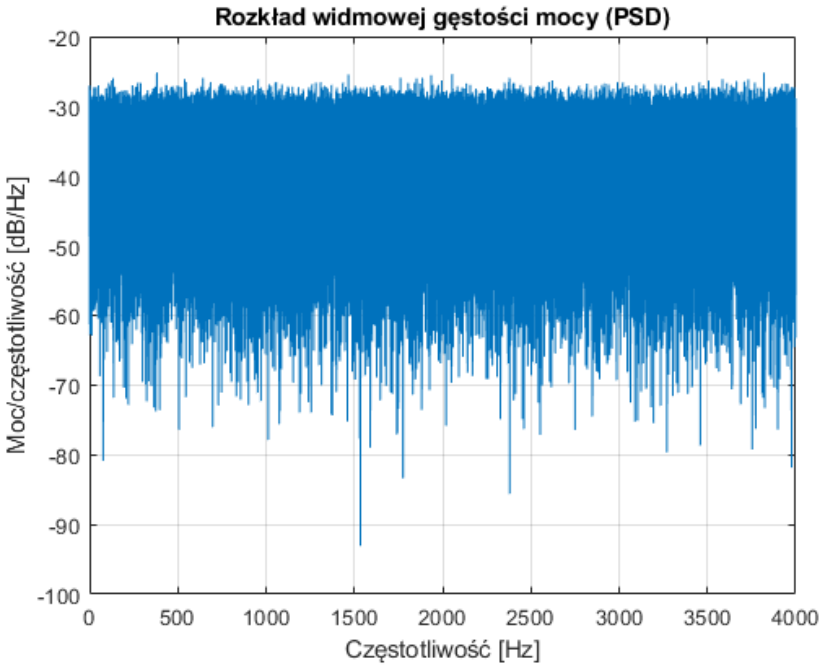
Bazując na powyższych wynikach widać przewagę podejść fuzji na poziomie cech oraz decyzji nad innymi wariantami systemów ASR (dla każdego wariantu długości segmentu treningowego).

6.4. Testy opracowanych rozwiązań w zróżnicowanych warunkach szumowych

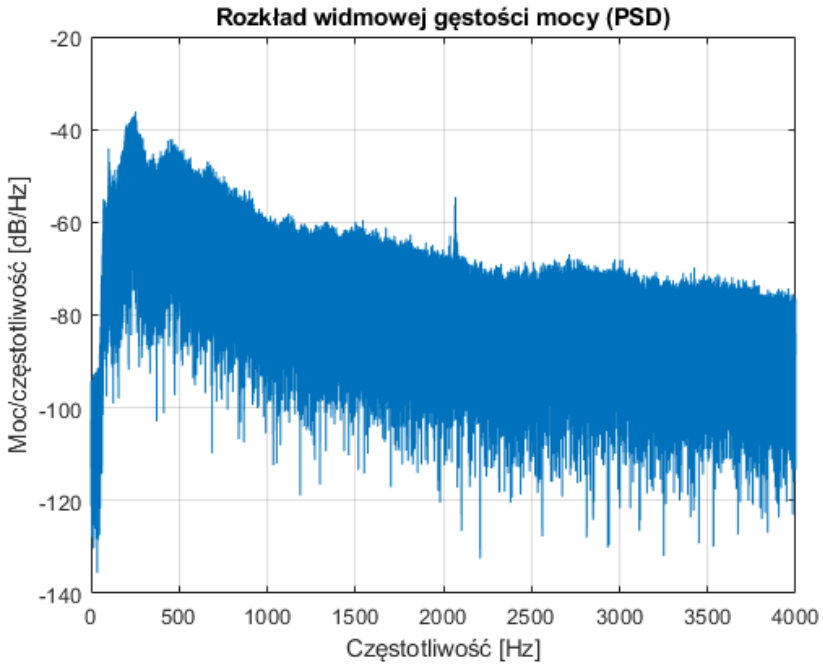
Poza bezszumowymi warunkami pracy systemów ASR przeprowadzono również testy, podczas których analizowane sygnały mowy zakłócone były przez zewnętrzne addytywne szumy lub zakłócenia. Autor opracował scenariusz wykorzystania 3 wariantów zakłóceń, tj.:

- wygenerowano szum biały (osobno dla segmentów uczących i testujących) i zakłócono nim pierwotny zbiór głosów,
- do każdego sygnału dodano zakłócenie w postaci tłumu, czyli zasymulowano rejestracje mowy w lokalu pełnym innych rozmawiających osób (w nomenklaturze zagranicznej taki rodzaj zakłócenia nosi nazwę *cocktail party*),
- ostatnim dodanym zakłóceniem był dźwięk ulewnego deszczu.

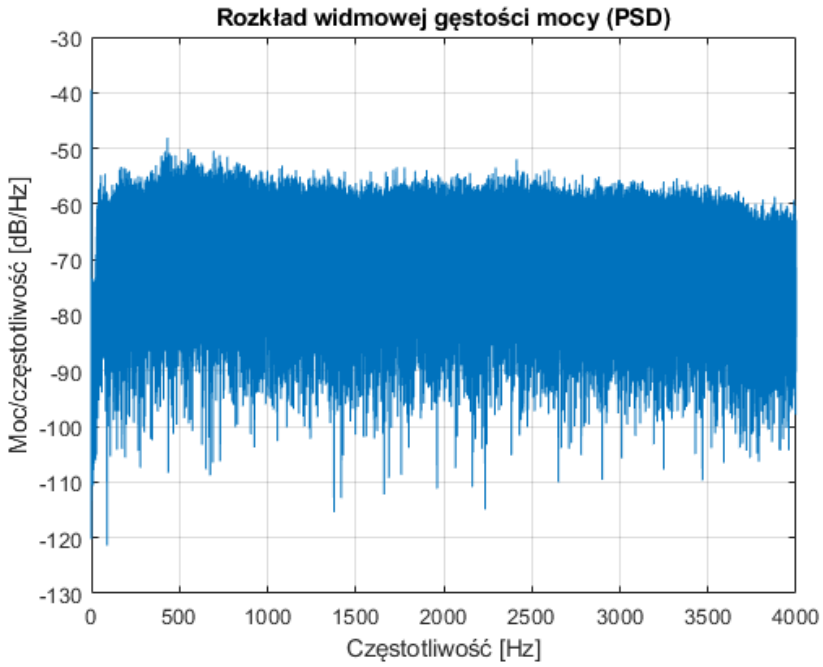
Taki dobór zakłóceń podyktowany jest potrzebą przetestowania opracowanych rozwiązań w warunkach jak najbardziej zbliżonych do rzeczywistych. Ponadto każdy z powyższych sygnałów cechuje się innym rozkładem mocy w widmie. Szum biały charakteryzuje się tym, iż na każdy herc w spektrum (rys. 6.3) przypada taka sama wartość mocy (równomierny rozkład widmowej gęstości mocy). W przypadku zakłócenia imitującego tłum rozmawiających ludzi przebieg widma (rys. 6.4) ma charakter funkcji wykładniczej (tj.: moc maleje wraz ze wzrostem częstotliwości). Z kolei wykres widmowej gęstości mocy dźwięku ulewnego deszczu jest zbliżony do przebiegu na rys. 6.3, ale z tendencją do powolnego spadku (rys. 6.5).



Rys. 6.3. Przebieg widmowej gęstości mocy szumu białego



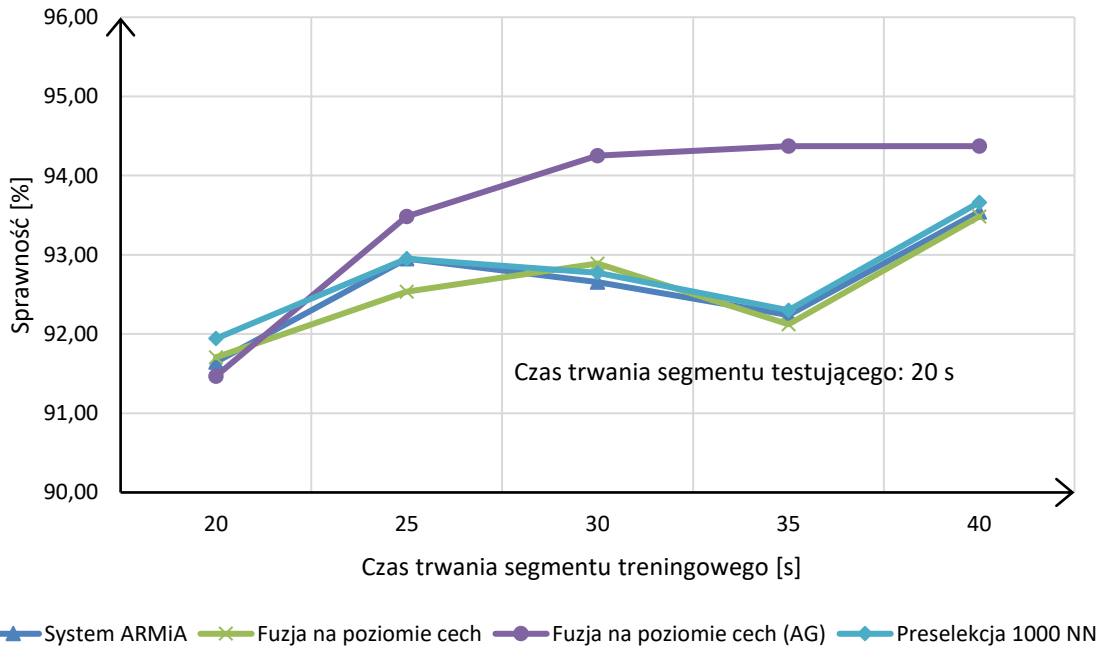
Rys. 6.4. Przebieg widmowej gęstości mocy zakłóceń o charakterze gwaru zarejestrowanego w tłumie ludzi



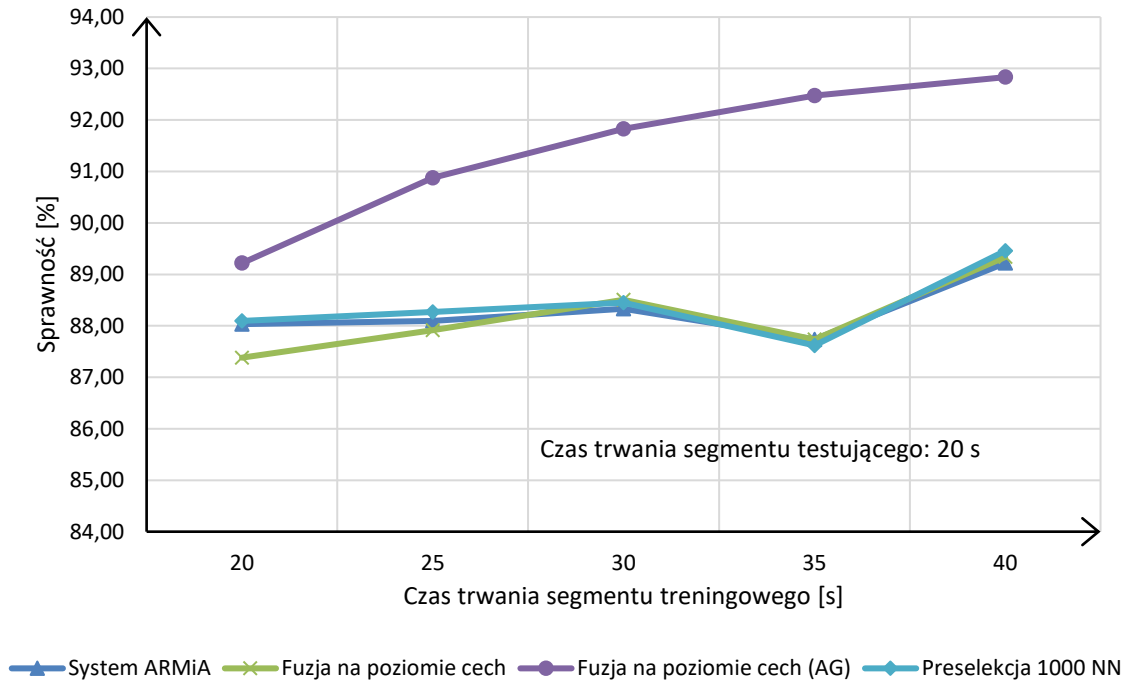
Rys. 6.5. Przebieg widmowej gęstości mocy zakłóceń o charakterze ulewnego deszczu

6.4.1. Badania opracowanych rozwiązań dla zakłóceń w postaci szumu białego

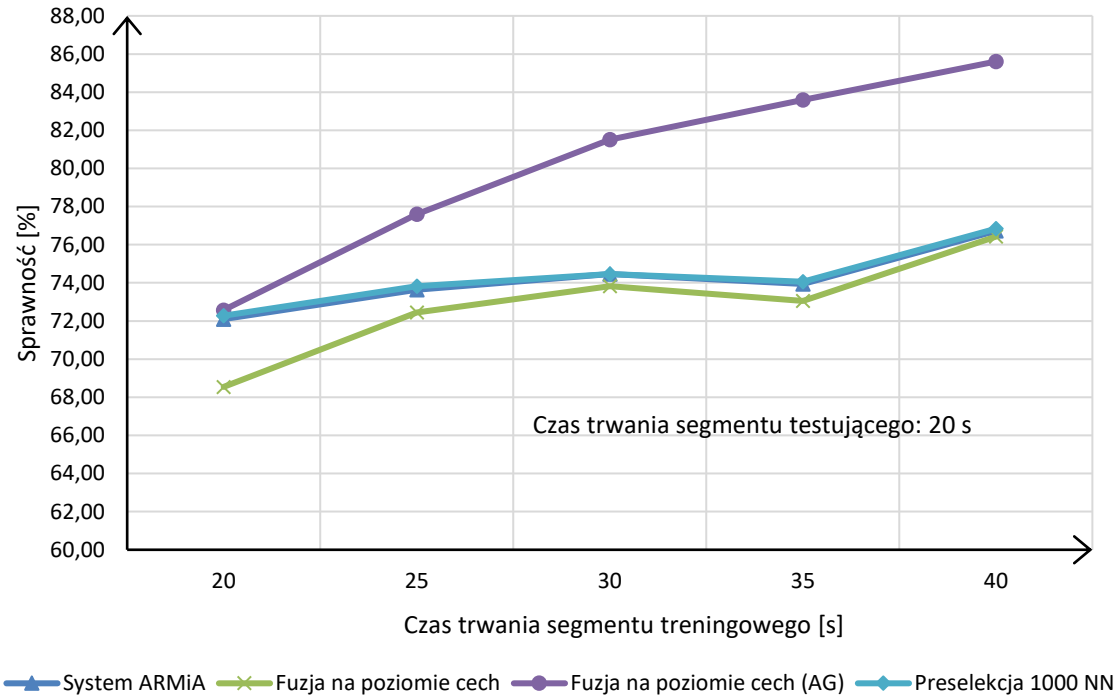
Pierwszym z testowanych zakłóceń był szum biały (ang. *AWGN – Additive White Gaussian Noise*). Zakłócenie wygenerowane zostało przy pomocy środowiska *Matlab*. Autor wykorzystał ten rodzaj zakłócenia ze względu na możliwość równomiernej degradacji całego sygnału. Przetestowane zostały 3 wartości stosunku sygnału do szumu (ang. *SNR – Signal-to-Noise Ratio*): 20, 15 oraz 10 dB. Rys. 6.6, 6.7 i 6.8, to graficzne reprezentacje rezultatów testów przeprowadzonych w rozszerzonym zbiorze 1688 głosów.



Rys. 6.6. Wyniki testów różnych wariantów systemów ASR przy degradacji szumem białym, SNR = 20 dB



Rys. 6.7. Wyniki testów różnych wariantów systemów ASR przy degradacji szumem białym, SNR = 15 dB



Rys. 6.8. Wyniki testów różnych wariantów systemów ASR przy degradacji szumem białym, SNR = 10 dB

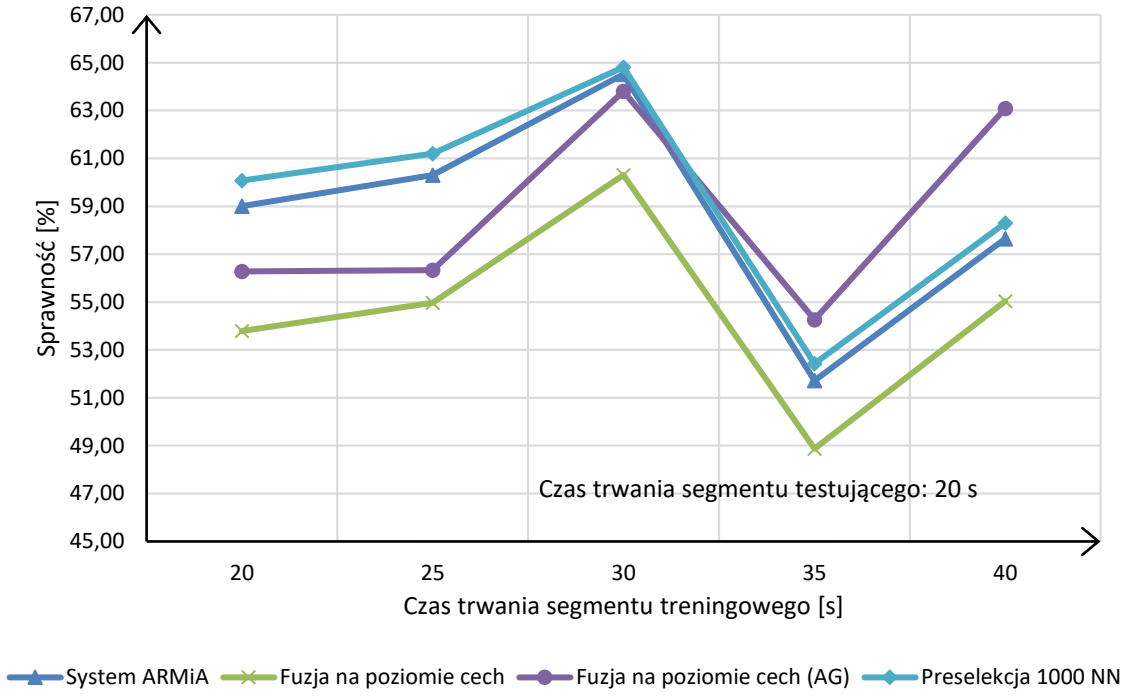
W przypadku zaburzenia sygnału mowy poprzez dodanie szumu białego liczba poprawnych identyfikacji tożsamości mówcy zmniejsza się w każdym z analizowanych wariantów systemów ASR. Jednak metody integracji danych zazwyczaj osiągają lepsze rezultaty niż system odniesienia [60]. Najlepiej sprawdza się technika fuzji danych na poziomie cech po zastosowaniu algorytmu genetycznego.

6.4.2. Badania opracowanych rozwiązań dla zakłóceń w postaci nagrania tłumu

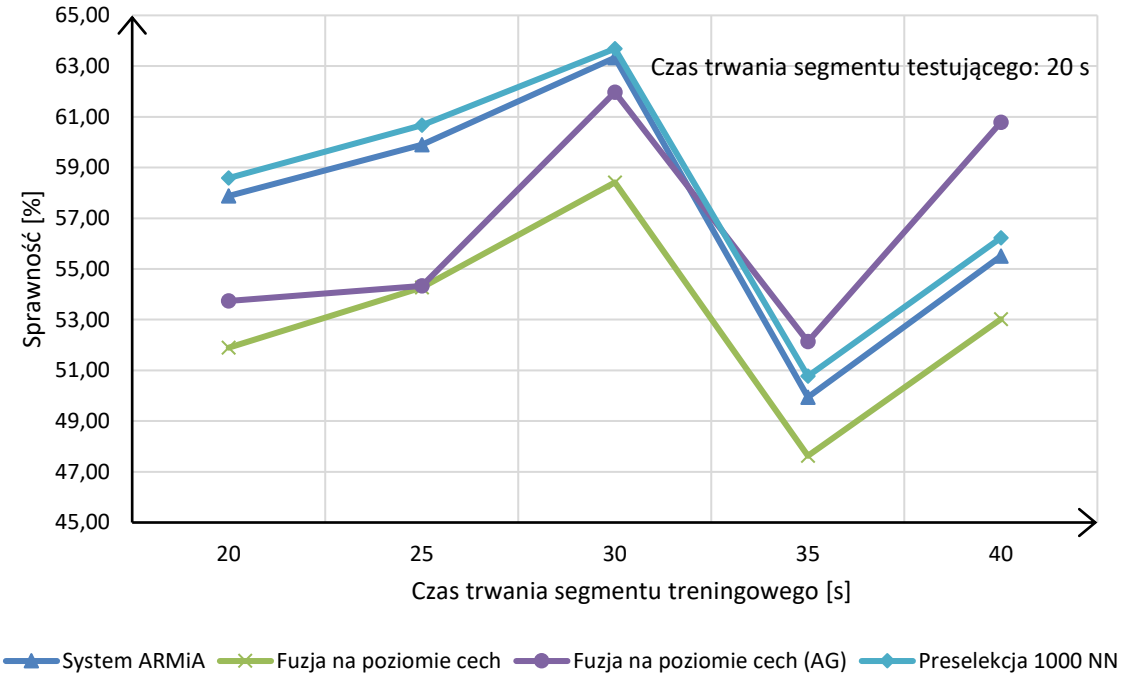
Kolejnym testowanym zakłóceniem było nagranie tłumu rozmawiających ludzi w tle. Ten scenariusz miał imitować sytuacje, gdy identyfikacja tożsamości mówcy następuję w kawiarni bądź środkach komunikacji publicznej (występują głosy innych osób). W związku z nierównomiernym rozkładem mocy w spektrum zakłócenia autor postanowił organoleptycznie dopasować jego amplitudę. W ten sposób za pomocą słuchu dobrano takie wartości amplitudy zakłócenia, aby reprezentowane były trzy stany jego występowania:

- zakłócenia słyszalne,
- zakłócenia uciążliwe,
- zakłócenia przeważające.

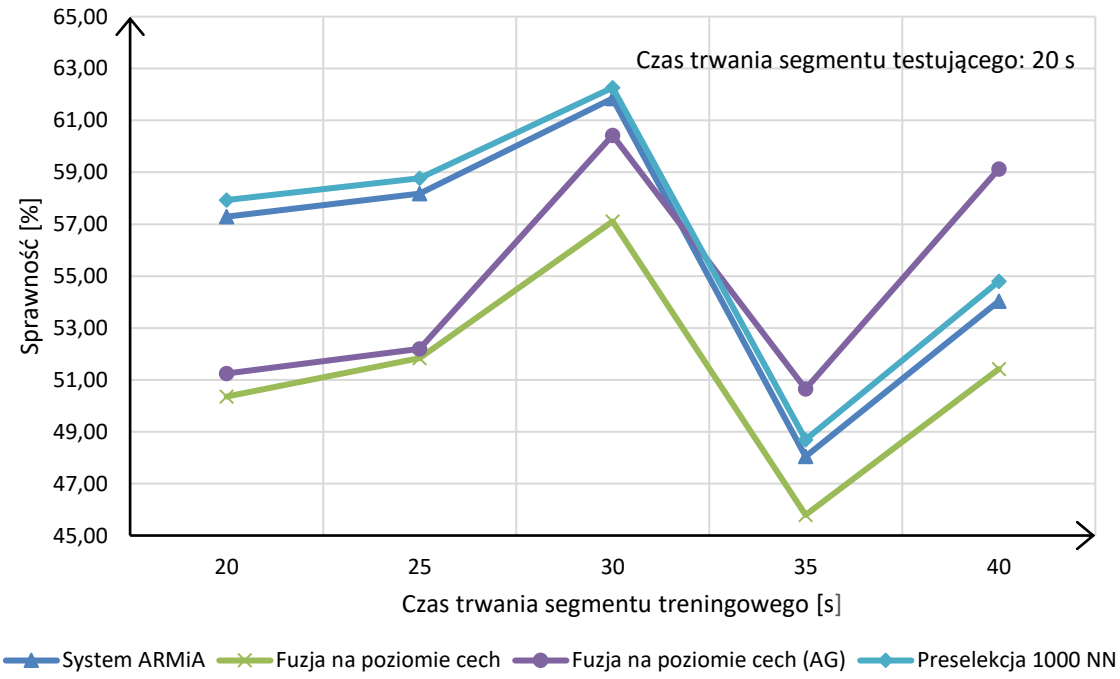
Rezultaty testów dla zbioru 1688 głosów przedstawiono na rys. 6.9, 6.10 i 6.11.



Rys. 6.9. Rezultaty testów różnych wariantów systemów ASR przy degradacji imitacją tłumu, szum słyszalny



Rys. 6.10. Rezultaty testów różnych wariantów systemów ASR przy degradacji imitacją tłumu, szum uciążliwy

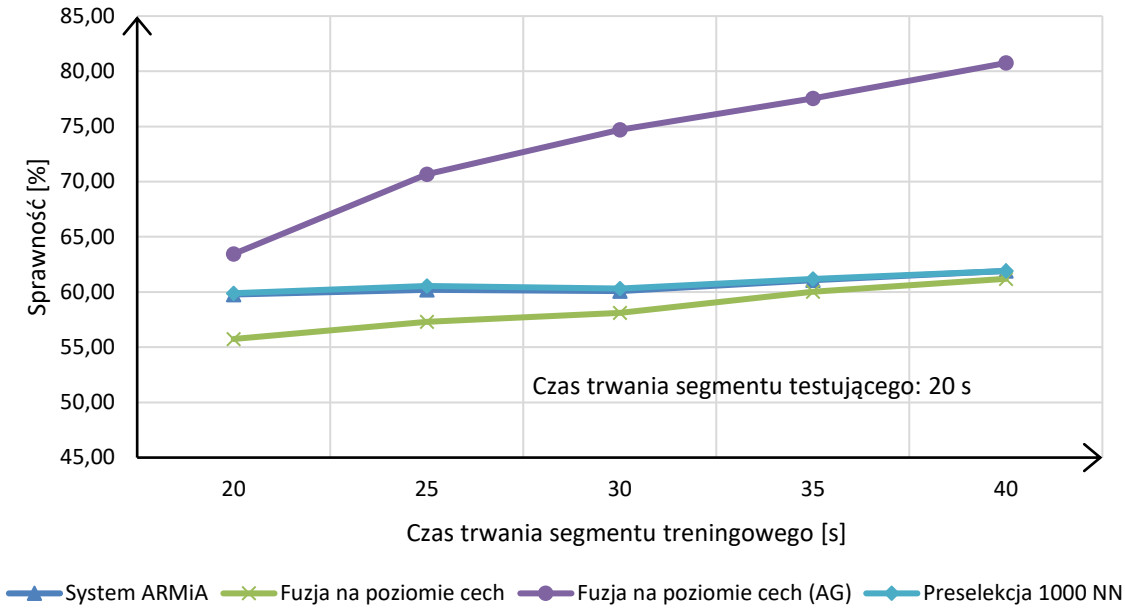


Rys. 6.11. Rezultaty testów różnych wariantów systemów ASR przy degradacji imitacją tłumu, szum przeważający

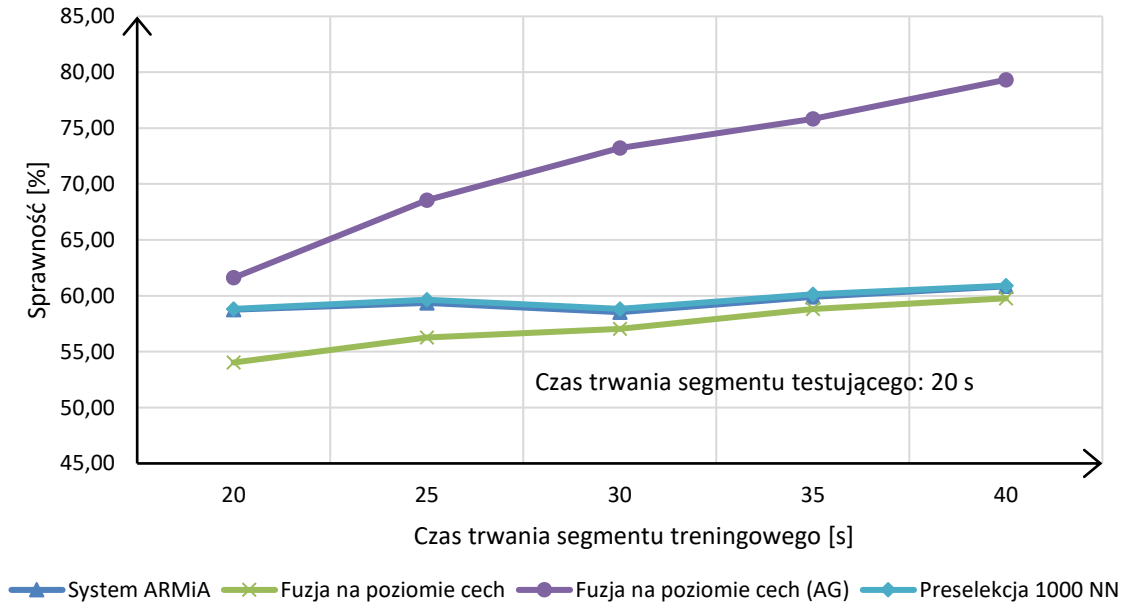
Niezależnie od amplitudy dodanego szumu oraz długości segmentu treningowego metoda preselekcji poprawia liczbę poprawnych identyfikacji tożsamości mówców. Porównując rezultaty uzyskane dla szumu białego i imitacji tłumu widać, iż nierównomierny rozkład mocy w widmie powoduje pogorszenie skuteczności działania wariantu integracji danych na poziomie cech.

6.4.3. Badania opracowanych rozwiązań dla zakłóceń w postaci nagrania ulewnego deszczu

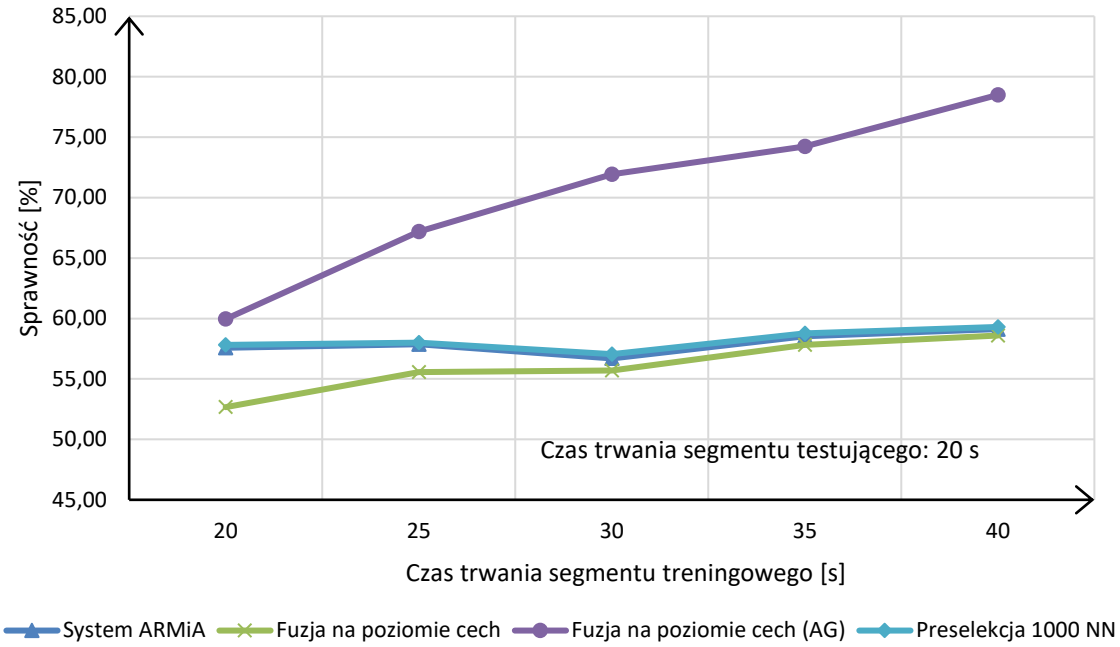
Ostatnim z testowanych szumów był dźwięk ulewnego deszczu. Samo zakłócenie podczas odsłuchu przypomina szum biały, ale analiza ich widm jasno pokazuje, że różnią się one rozkładem mocy. Autor postanowił wykorzystać ten rodzaj zakłóceń ze względu na wysokie prawdopodobieństwo występowania. Dopasowanie amplitudy zakłócenia realizowane było podobnie, jak w przypadku imitacji tłumu ludzi. Rys. 6.13, 6.14 i 6.15 przedstawiają wykresy wyników testów różnych wariantów systemów ASR dla zbioru 1688 głosów.



Rys. 6.13. Rezultaty testów różnych wariantów systemów ASR przy degradacji dźwiękiem ulewnego deszczu, szum słyszalny



Rys. 6.14. Rezultaty testów różnych wariantów systemów ASR przy degradacji dźwiękiem ulewnego deszczu, szum uciążliwy



Rys. 6.15. Rezultaty testów różnych wariantów systemów ASR przy degradacji dźwiękiem ulewnego deszczu, szum przeważający

Wykorzystanie zakłócenia imitujące dźwięk ulewnego deszczu pokazuję, iż niezależnie od zastosowanej metody integracji danych każda z nich pozwala na zwiększenie liczby poprawnych identyfikacji tożsamości mówcy względem systemu odniesienia [60].

6.5. Testy opracowanych rozwiązań w warunkach zróżnicowanej jakości nagrań

Autor postanowił również sprawdzić jak opracowane metody integracji danych z dwóch różnych systemów ASR sprawdzą się w warunkach zróżnicowanej jakości nagrań. Eksperymenty są analogiczne do tych opisanych w załączniku 2., a konkretniej w sekcji *Testy systemu przy degradacji jakości sygnału głosowego poprzez kodowanie mowy*. W powyższym załączniku zawarto również opis zastosowanych metod kodowania mowy. Testy przeprowadzone były dla każdego z kodeków w relacji *każdy z każdym.* Innymi słowy pojedynczy sposób kodowania był sprawdzany dla danych testujących poddanych przekształceniom przez wszystkie kodeki po kolei. Zaprezentowane zostały wyniki tylko dla kombinacji 40 sekund uczenia i 20 sekund testowania. Reszta rezultatów znajduję się w załączniku nr 3.

Tab. 6.1. Wartości sprawności systemu *ARMiA* dla różnych standardów kodowania

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,80 %**	94,85 %	91,47 %	92,89 %
SPEEX	94,61 %	**96,15 %**	93,19 %	91,82 %
G.723.1	90,70 %	93,13 %	**95,38** %	89,04 %
GSM 06.10	92,36 %	91,71 %	89,16 %	**96,33** %

Tab. 6.2. Wartości sprawności systemu opartego na cechach behawioralnych dla różnych standardów kodowania

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**57,82** %	27,55 %	34,42 %	37,50 %
SPEEX	26,13 %	**44,02** %	28,91 %	23,58 %
G.723.1	30,45 %	28,97 %	**44,14** %	36,55 %
GSM 06.10	35,60 %	24,70 %	37,74 %	**50,65** %

Tab. 6.3. Wartości sprawności wariantu fuzji danych na poziomie cech dla różnych standardów kodowania

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,80 %**	94,91 %	91,88 %	93,01 %
SPEEX	94,25 %	**96,39 %**	92,77 %	91,53 %
G.723.1	90,82 %	93,07 %	**95,38 %**	89,69 %
GSM 06.10	92,36 %	91,35 %	88,74 %	**96,39 %**

Tab. 6.4. Wartości sprawności wariantu fuzji danych na poziomie cech po zastosowaniu algorytmu genetycznego dla różnych standardów kodowania

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,74 %**	95,14 %	92,65 %	91,11 %
SPEEX	93,19 %	**95,50 %**	92,71 %	89,10 %
G.723.1	87,80 %	91,17 %	**94,61 %**	85,19 %
GSM 06.10	89,63 %	90,46 %	88,45 %	**95,97 %**

Tab. 6.5. Wartości sprawności wariantu fuzji danych na poziomie decyzji dla różnych standardów kodowania

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,86** %	95,02 %	91,65 %	93,07 %
SPEEX	94,02 %	**96,27** %	92,89 %	91,05 %
G.723.1	90,94 %	93,25 %	**95,44** %	89,34 %
GSM 06.10	92,54 %	91,94 %	89,40 %	**96,39** %

W przypadku analizy sygnałów mowy poddanych kodowaniu podejścia wykorzystujące fuzje rozwiązań bazujących na cechach behawioralnych i fizycznych są w stanie przewyższyć sprawność systemu *ARMiA* [60]. Zazwyczaj lepszym wariantem okazuję się metoda preselekcji, jednak nie w każdych warunkach (tj.: nie przy wszystkich standardach kodowania). Dlatego też nie można jednoznacznie wykazać wyższości jednego, konkretnego podejścia nad rozwiązaniem odniesienia [60].

6.6. Testy opracowanych rozwiązań przy wykorzystaniu zróżnicowanych torów akustycznych

W podrozdziale 6.1.4 opisana została baza głosów autorstwa mgr. inż. Daniela Posiadały. Charakterystyczną cechą tego zbioru było to, iż 50 mówców zarejestrowanych zostało przy pomocy 10 różnych urządzeń, tj. baza sumarycznie liczy 500 próbek głosu. Jako że każda z osób biorąca udział w nagraniach charakteryzuje się różną szybkością artykulacji niemożliwe było przetestowanie wszystkich kombinacji (w odniesieniu do opisanych powyżej testów). Najkrótsze nagranie w bazie ma długość 41 sekund. Dlatego też autor postanowił nie ingerować w strukturę dźwięków i przeprowadził testy dla kombinacji długości segmentów uczącego i testującego, których suma wynosi maksymalnie 40 sekund. Poniżej przedstawiono wyniki tylko dla wariantu: 30 sekund uczenia i 10 sekund testowania. Wyniki zestawiono w jednej tab. 6.6. Wytłuszczone zostały rezultaty otrzymane dla systemu odniesienia [60].

Przedstawione wyniki są mocno zależne od wykorzystanych do nagrań urządzeń. W swojej pracy autor [26] scharakteryzował urządzenia rejestrujące przy pomocy fotografii oraz podstawowych parametrów technicznych. Do rejestracji głosu wykorzystano następujące urządzenia:

1. mikrofon dynamiczny Pioneer DM-DV5,
2. mikrofon pojemnościowy Sekaku LM-10,
3. mikrofon komputerowy Media-Tech MT392,
4. mikrofon wbudowany w notebooku Lenovo Z570,
5. mikrofon wbudowany w telefonie Samsung Galaxy S Duos,
6. mikrofon wbudowany w telefonie Samsung Galaxy Mini 2,
7. mikrofon dynamiczny Electro-Voice ND-767/a,
8. mikrofon dynamiczny Shure Beta 58a,
9. mikrofon wbudowany w kamerce internetowej Esperanza TITANUM TC101 Onyx 3 Led Light,
10. mikrofon wbudowany w dyktafonie Olympus WS-812.

Autor niniejszej rozprawy (poza posiadanym spisem i charakterystyką rejestratorów głosu) dokonał też analizy organoleptycznej nagranych dźwięków, aby ocenić prawdziwość charakterystyk poszczególnych urządzeń. Wynikiem tej ekspertyzy jest fakt, iż urządzenia nr 4 oraz 9 charakteryzują się najgorszą jakością nagrań, tj. są stłumione i najmniej wyraźne.

Tab. 6.6. Wartości sprawności (wyrażone w %) różnych wariantów systemów ASR dla zróżnicowanych urządzeń rejestrujących

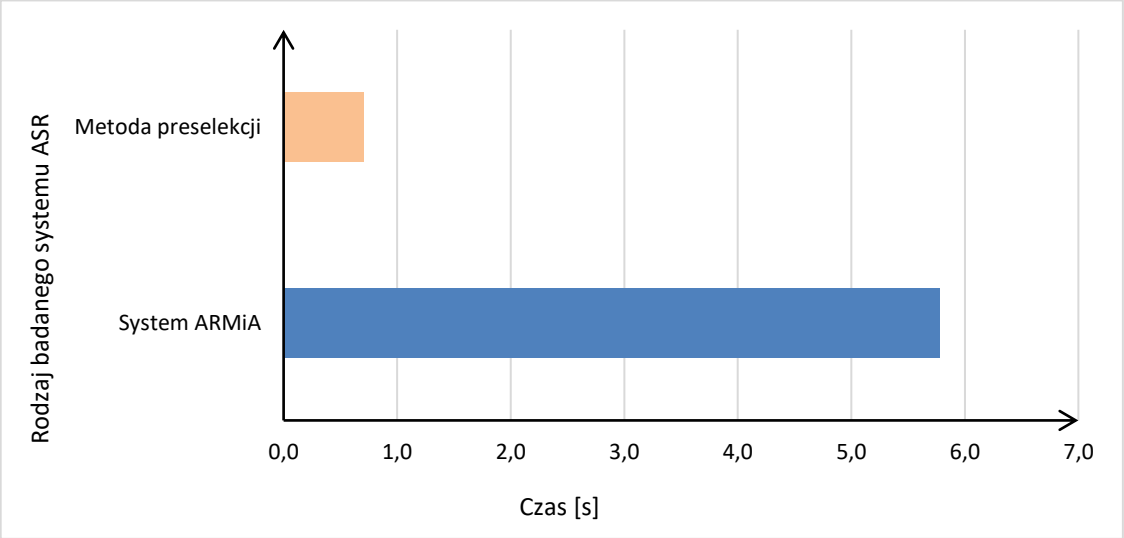
		Tor uczący									
	Tor testujący	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
System *ARMiA*	**1**	**100**	66	74	12	30	26	92	90	14	26
Fuzja danych na poziomie cech		**100**	70	74	18	30	26	88	90	14	28
Fuzja danych na poziomie cech (AG)		**98**	72	76	28	26	22	84	90	12	24
Fuzja danych na poziomie decyzji		**100**	52	72	16	40	30	92	90	12	32
System *ARMiA*	**2**	80	**98**	66	8	40	40	78	80	10	26
Fuzja danych na poziomie cech		76	**98**	66	8	42	36	78	80	10	26
Fuzja danych na poziomie cech (AG)		78	**98**	64	32	48	40	86	80	16	42
Fuzja danych na poziomie decyzji		60	**96**	50	28	36	28	84	66	8	30
System *ARMiA*	**3**	84	72	**98**	16	38	36	84	82	16	36
Fuzja danych na poziomie cech		86	74	**98**	16	38	36	82	80	16	36
Fuzja danych na poziomie cech (AG)		90	84	**98**	36	36	28	90	82	20	36
Fuzja danych na poziomie decyzji		90	68	**100**	22	42	40	92	86	10	34
System *ARMiA*	**4**	16	20	20	**76**	12	8	20	22	6	14
Fuzja danych na poziomie cech		16	18	16	**76**	12	8	22	20	6	14
Fuzja danych na poziomie cech (AG)		32	36	26	**100**	12	8	28	28	14	12
Fuzja danych na poziomie decyzji		22	26	34	**88**	12	14	36	24	6	10
System *ARMiA*	**5**	32	48	22	10	**100**	68	38	22	6	46
Fuzja danych na poziomie cech		32	48	22	8	**100**	70	38	26	6	44
Fuzja danych na poziomie cech (AG)		28	44	22	10	**100**	72	34	22	8	46
Fuzja danych na poziomie decyzji		34	36	30	18	**98**	66	42	36	12	48
System *ARMiA*	**6**	92	58	30	10	80	**96**	38	36	10	60
Fuzja danych na poziomie cech		88	58	34	8	80	**96**	36	36	10	60
Fuzja danych na poziomie cech (AG)		90	52	28	12	78	**98**	36	36	10	54
Fuzja danych na poziomie decyzji		90	36	34	16	72	**96**	34	42	8	52
System *ARMiA*	**7**	88	86	76	14	48	52	**94**	94	14	26
Fuzja danych na poziomie cech		84	86	74	14	48	54	**94**	94	14	26
Fuzja danych na poziomie cech (AG)		88	90	78	32	44	44	**92**	92	16	26
Fuzja danych na poziomie decyzji		84	66	72	20	44	36	**98**	92	8	30
System *ARMiA*	**8**	36	84	80	16	38	44	90	**96**	18	40
Fuzja danych na poziomie cech		36	82	80	16	40	44	90	**96**	18	40
Fuzja danych na poziomie cech (AG)		32	86	82	32	42	42	88	**96**	16	32
Fuzja danych na poziomie decyzji		26	66	70	22	40	32	86	**98**	10	30
System *ARMiA*	**9**	22	32	42	8	16	22	38	34	**94**	20
Fuzja danych na poziomie cech		24	32	42	8	16	24	38	32	**94**	22
Fuzja danych na poziomie cech (AG)		18	38	34	8	16	22	32	32	**96**	18
Fuzja danych na poziomie decyzji		20	32	38	10	14	32	30	36	**92**	14
System *ARMiA*	**10**	95	36	18	14	46	54	26	16	12	**100**
Fuzja danych na poziomie cech		75	36	20	14	48	54	26	14	12	**100**
Fuzja danych na poziomie cech (AG)		81	32	24	14	42	46	24	16	14	**98**
Fuzja danych na poziomie decyzji		97	30	18	14	50	52	24	28	10	**96**

Zgodnie ze spisem [26], to właśnie te urządzenia cechuje najniższa jakość rejestrowanych dźwięków. Taki wniosek również odzwierciedlony jest w powyższych wynikach, gdyż to dla tych mikrofonów uzyskane zostały najniższe wartości sprawności (przeważająca ich część wynosi poniżej 50% poprawnych identyfikacji tożsamości dla systemu *ARMiA* [60]). Reszta urządzeń charakteryzuje się zadowalającą jakością rejestracji. Znaczne rozbieżności uzyskanych sprawności dla różnych kombinacji rejestratorów dźwięków mogą wynikać z wysokiej odmienności sprzętów pod względem parametrów

technicznych. Analizując powyższe wyniki eksperymentów można wnioskować, że istnieje możliwość poprawy sprawności systemu odniesienia [60] poprzez zastosowanie fuzji danych.

6.7. Oszczędność obliczeniowa

Ostatnim testem przeprowadzonym przez autora było sprawdzenie średniego czasu identyfikacji mówcy w zbiorze 10000 modeli głosów. Przebadane zostały system *ARMiA* [60] oraz metoda fuzji danych na poziomie decyzji. Dla każdego z wariantów wykonano 100 pomiarów, a uzyskane wyniki czasowe zostały uśrednione. Pierwsze z analizowanych rozwiązań ASR charakteryzuje się średnim czasem identyfikacji mówcy wynoszącym 5,78 s. Natomiast przy zastosowaniu behawioralnego preselektora wystarczy zaledwie 0,7 s, aby zidentyfikować mówcę. Otrzymane wartości pokazują ponad 8 razy krótszy czas potrzebny do przetworzenia jednej próbki głosu przez rozwiązanie fuzji danych na poziomie decyzji. Tak znacząca redukcja czasu wynika z faktu, iż wstępny zbiór 10000 modeli głosów ograniczany jest przez klasyfikator behawioralny do 1000 najbliższych sąsiadów względem badanego (testowego) obiektu. Dzięki temu w kolejnym etapie moduł wykorzystujący cechy fizyczne [60] porównuje model testowy z dziesięć razy mniejszą liczbą modeli treningowych. Na rysunku 6.16 zobrazowano porównanie powyższych wyników.



Rys. 6.16. Średnie czasy identyfikacji mówcy w zbiorze 10000 modeli głosów

Ponad ośmiokrotny zysk czasu analizy wiąże się z niewielkim zwiększeniem rozmiaru danych na dysku (dla zbioru 10000 modeli) z około 32 MB do 34 MB. Zmiana ta wynika z konieczności przechowywania, oprócz modeli mieszanin gaussowskich, także zbioru wektorów cech behawioralnych danych treningowych. Jednakże uzyskany znaczący wzrost szybkości identyfikacji mówców pokazuje, że zastosowanie preselektora opartego na cechach behawioralnych istotnie zwiększa skalowalność dotychczasowego systemu *ARMiA* [60].

6.8. Podsumowanie

W niniejszym rozdziale zaprezentowano szereg eksperymentów przeprowadzonych przez autora w ostatnim etapie badań, które potwierdziły wstępne przypuszczenia, że samodzielny system ASR oparty wyłącznie na cechach behawioralnych nie jest w stanie konkurować z już istniejącymi rozwiązaniami wykorzystującymi zasadniczo cechy fizyczne. Uzyskane rezultaty pokazują, że wykorzystanie cech behawioralnych do wsparcia systemu *ARMiA* [60] poprawia jego sprawność, szczególnie w obecności zewnętrznych szumów i zakłóceń. Ponadto taka tendencja jest również zauważalna w przypadku analizy dźwięków kodowanych przez różne standardy kompresji (w medium transmisyjnym) czy przy wykorzystaniu różnych urządzeń rejestrujących. Zależnie od warunków w jakich prowadzona jest analiza sygnałów mowy możliwe jest dopasowanie sposobu przetwarzania, aby zmaksymalizować liczbę poprawnych identyfikacji tożsamości mówców. Dodatkowo wykazano, iż wykorzystanie behawioralnego preselektora w procesie identyfikacji mówcy pozwala na wielokrotną oszczędność czasu obliczeń względem systemu referencyjnego [60].

7.

Konkluzja

W niniejszej rozprawie podjęto problem wykorzystania cech behawioralnych głosu w rozwiązaniach automatycznego rozpoznawania mówcy. Celem pracy było utworzenie i implementacja w istniejącym systemie ASR pewnego zestawu cech behawioralnych, który pozwoliłby na zwiększenie liczby poprawnych identyfikacji tożsamości mówców, w szczególności w obecności zakłóceń o różnym charakterze.

Pierwszy etap badań obejmował przegląd literatury naukowej w dziedzinie rozwiązań automatycznego rozpoznawania mówcy oraz dokładną analizę istniejącego, opracowanego w Wojskowej Akademii Technicznej systemu *ARMiA* [60]. W literaturze rzadko występują wzmianki o cechach behawioralnych mowy, sposobie ich wykorzystania i implementacji. Przeważająca większość artykułów naukowych dotyczy analizy pojedynczych fonemów czy ich grup tworzących sylaby w celu ekstrakcji cennych informacji dystynktywnych. Dokładne przestudiowanie sposobu działania rozwiązania

autorstwa dr. Kamińskiego umożliwiło wyznaczenie głównego kierunku działań w zakresie utworzenia i sposobu implementacji wstępnego zbioru cech behawioralnych w osobnym, autorskim systemie automatycznego rozpoznawania mówcy. Rezultatem tych działań było utworzenie, selekcja (przy pomocy algorytmu genetycznego) oraz implementacja zbioru 27 cech behawioralnych głosu, stanowiących podstawę działania autorskiego systemu ASR.

Liczne dalsze testy oraz analiza porównawcza autorskiego rozwiązania z systemem *ARMiA* [60] pokazały, że podejście oparte tylko na cechach behawioralnych nie sprawdza się w warunkach przyjętych przez autora systemu opracowanego w WAT [60].

Dalsza część pracy obejmuje dokładny opis sposobów przeprowadzenia fuzji danych pochodzących z dwóch niezależnych systemów automatycznego rozpoznawania mówcy w celu utworzenia jednego, komplementarnego rozwiązania. Opracowano dwie autorskie metody integracji danych, tj. na poziomie cech oraz na poziomie decyzji. Następnie przeprowadzono testy ewaluacyjne tychże rozwiązań dla nagrań o różnej jakości (mowa niezakłócona oraz zakłócona trzema rodzajami szumów). Dodatkowo sprawdzono jak zastosowanie metody preselekcji wpłynie na średni czas identyfikacji pojedynczego mówcy. Finalnie otrzymano rezultaty potwierdzające postawioną we wprowadzeniu tezę, że ***możliwe jest zdefiniowanie i wyekstrahowanie z sygnału mowy takich cech behawioralnych, które w połączeniu z cechami fizycznymi zaimplementowanymi w istniejącym rozwiązaniu znacząco poprawią jego skuteczność w obecności zewnętrznych zakłóceń.***

Do najważniejszych osiągnięć autora uzyskanych w ramach realizacji niniejszej pracy można zaliczyć:

- wybór efektywnego zestawu cech behawioralnych, tj. selekcja najbardziej dystynktywnych cech spośród wszystkich zdefiniowanych wstępnie 71 cech, przy pomocy metod ewolucyjnych;
- autorską metodę ekstrakcji cech opartą na przetwarzaniu nakładających się (ang. *overlapping*) ramek mowy;
- autorskie metody integracji danych wykorzystujące fuzję na poziomie decyzji oraz na poziomie cech;
- eksperymentalną ocenę sprawności działania systemu w różnorodnych warunkach.

Zaproponowane w niniejszej rozprawie rozwiązanie fuzji systemów ASR można wciąż udoskonalać. Autor w dalszych etapach rozwoju algorytmu planuje wykorzystać sztuczne sieci neuronowe takie jak LSTM (ang. *Long Short-Term Memory*) do zadania klasyfikacji. Ponadto autor planuje przetestować system ASR dla nagrań pochodzących z interwencji policyjnych.

Reasumując, można stwierdzić, że postawiona we wstępnie teza została udowodniona, a założone cele osiągnięte. Przedstawione w niniejszej rozprawie rezultaty badań opublikowane zostały w [1] oraz zaprezentowane podczas *XXXIV Konferencji Naukowo-Technicznej EKOMILITARIS 2023*.

Literatura

[1] A. Dobrowolski and D. Mały, “Behavioral features of the speech signal as part of improving the effectiveness of the automatic speaker recognition system,” *Inżynieria Bezpieczeństwa Obiektów Antropogenicznych*, no. 4, pp. 26–34, Dec. 2023, DOI: 10.37105/iboa.187.

[2] A. Dobrowolski, *Transformacje sygnałów: od teorii do praktyki*, Legionowo: Wydawnictwo BTC, 2018.

[3] A. Janicki and T. Staroszczyk, "Klasyfikacja mówców oparta na modelowaniu GMM-UBM dla mowy o różnej jakości," *Przegląd Telekomunikacyjny - Wiadomości Telekomunikacyjne*, vol. 8–9, pp. 1469–1474, 2011.

[4] A. Kokocińska and T. Kaleta, "Behawioryzm i behawior - myśl filozoficzna i badania przyrodnicze," *Kosmos*, vol. 64, no. 2, pp. 221–227, 2015.

[5] A. Lerch, “An introduction to audio content analysis: Applications in signal processing and music informatics,” *An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics*, pp. 1–248, Oct. 2012, DOI: 10.1002/9781118393550.

[6] A. Phinyomark, S. Thongpanja, H. Hu, P. Phukpattaranont, and C. Limsakul, "The Usefulness of Mean and Median Frequencies in Electromyography Analysis," in *Computational Intelligence in Electromyography Analysis - A Perspective on Current Applications and Future Challenges*, G. R. Naik, Ed., IntechOpen, 2012.

[7] A. V. Oppenheim and R. W. Schafer, "From frequency to quefrency: a history of the cepstrum," in *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 95-106, Sept. 2004, DOI: 10.1109/MSP.2004.1328092.

[8] A. Waibel, *Prosody and Speech Recognition*, San Mateo, CA, USA: Morgan Kaufmann Publishers, 1988.

[9] A. Żbikowska-Migoń and M. Skalska-Zlat, Eds., *Encyklopedia książki. T. 2, K–Z*, Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego, 2017.

[10] B. G. Nagaraja, M. Anees, and T. Y. G, “Speech coding techniques and challenges: a comprehensive literature survey,” *Multimed Tools Appl*, vol. 83, no. 10, pp. 29859–29879, Sep. 2023, DOI: 10.1007/s11042-023-16665-3.

[11] B. Ma, D. Zhu, R. Tong, and H. Li, “Speaker cluster based GMM tokenization for speaker recognition,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2006, pp. 505–508. DOI: 10.21437/interspeech.2006-164.

[12] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The Quefrency Analysis of Time Series for Echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum, and Saphe Cracking," in *Proceedings of the Symposium on Time Series Analysis*, M. Rosenblatt, Ed., New York: Wiley, 1963, pp. 209–243.

[13] B. Peskin et al., “Using prosodic and conversational features for high-performance speaker recognition: Report from JHU WS’02,” in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2003, pp. 792–795. DOI: 10.1109/icassp.2003.1202762.

[14] B. Remeseiro and V. Bolon-Canedo, “A review of feature selection methods in medical applications,” *Comput Biol Med*, vol. 112, p. 103375, Sep. 2019, DOI: 10.1016/J.COMPBIOMED.2019.103375.

[15] B. V. Dasarathy, "Sensor fusion potential exploitation-innovative architectures and illustrative applications," in *Proceedings of the IEEE*, vol. 85, no. 1, pp. 24-38, Jan. 1997, DOI: 10.1109/5.554206.

[16] B. Wójtowicz and A. P. Dobrowolski, "Projekt integratora danych sensorycznych do detekcji niekontrolowanych upadków," *Biuletyn WAT*, vol. LXII, no. 4, pp. 230–240, 2013.

[17] B. Wójtowicz, “Detektor upadków wykorzystujący falkową generację cech dystynktywnych oraz klasyfikator SVM,” Rozprawa doktorska, Warszawa, 2015.

[18] B. Zagagy and M. Herman, “MESRS: Models Ensemble Speech Recognition System,” 2020, pp. 214–231. DOI: 10.1007/978-3-030-52246-9_15.

[19] B. Ziółko and M. Ziółko, *Przetwarzanie mowy*, Kraków: Wydawnictwa AGH, 2011.

[20] C. S. Ooi, M. H. Lim, and M. S. Leong, “Self-tune linear adaptive-genetic algorithm for feature selection,” IEEE Access, vol. 7, pp. 138211–138232, 2019, DOI: 10.1109/ACCESS.2019.2942962.

[21] C. Zhang and Y. Ma, Eds., *Ensemble Machine Learning: Methods and Applications*, Boston, MA: Springer, 2012.

[22] Cz. Basztura, *Komputerowe systemy diagnostyki akustycznej*, Warszawa: Wydawnictwo Naukowe PWN, 1996.

[23] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*., USA: Addison-Wesley Longman, 1989.

[24] D. H. Wolpert, “Stacked generalization,” *Neural Networks*, vol. 5, no. 2, pp. 241–259, Jan. 1992, DOI: 10.1016/S0893-6080(05)80023-1.

[25] D. O'Shaughnessy, *Speech Communications: Human and Machine*, 2nd ed., Wiley-IEEE Press, 2000.

[26] D. Posiadała, “Badania skuteczności systemu rozpoznawania mówcy w zróżnicowanym środowisku akustycznym,”, Praca inżynierska, Warszawa, 2015.

[27] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-Vectors: Robust DNN Embeddings for Speaker Recognition,” in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Institute of Electrical and Electronics Engineers Inc., Sep. 2018, pp. 5329–5333. DOI: 10.1109/ICASSP.2018.8461375.

[28] D. Tran, T. V. Le, and M. Wagner, "Fuzzy Gaussian Mixture Models for Speaker Recognition," in *Proc. 5th International Conference on Spoken Language Processing (ICSLP 1998)*, Sydney, Australia, Nov. 1998, pp. 1–4.

[29] E. Dahlman, S. Parkvall, and J. Sköld, *4G: LTE/LTE-Advanced for Mobile Broadband*, 2nd ed. Amsterdam, Netherlands: Academic Press, 2014.

[30] E. Fix and J. L. Hodges, “Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties,” *Int Stat Rev*, vol. 57, no. 3, p. 238, Dec. 1989, DOI: 10.2307/1403797.

[31] E. Majda, “Automatyczny system wiarygodnego rozpoznawania mówcy oparty na analizie cepstralnej sygnału mowy,” Rozprawa doktorska, Warszawa, 2013.

[32] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," 1997 IEEE International Conference on Acoustics,

Speech, and Signal Processing, Munich, Germany, 1997, pp. 1331-1334 vol.2, DOI: 10.1109/ICASSP.1997.596192.

[33] E. Shriberg, “Higher-level features in speaker recognition,” Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 4343 LNAI, pp. 241–259, 2007, DOI: 10.1007/978-3-540-74200-5_14.

[34] F. Alton, Everest and K. C. Pohlmann, *Master handbook of acoustics*. McGraw Hill, 2022.

[35] F. Brugnara, D. Falavigna, and M. Omologo, “Automatic segmentation and labeling of speech based on Hidden Markov Models,” *Speech Commun*, vol. 12, no. 4, pp. 357–370, Aug. 1993, DOI: 10.1016/0167-6393(93)90083-W.

[36] F. Jaroszyk, *Biofizyka: Podręcznik dla studentów*, Warszawa: Wydawnictwo Lekarskie PZWL, 2006.

[37] F. Szabo, *The Linear Algebra Survival Guide Illustrated with Mathematica*, Academic Press, 2015, 423 pp.

[38] G. Doddington, “Speaker Recognition based on Idiolectal Differences between Speakers,” in *EUROSPEECH 2001 - SCANDINAVIA - 7th European Conference on Speech Communication and Technology*, International Speech Communication Association, 2001, pp. 2521–2524. DOI: 10.21437/eurospeech.2001-417.

[39] G. Peeters, “A large set of audio features for sound description (similarity and classification) in the CUIDADO project,” Icram, 2004.

[40] H. F. Durrant-Whyte, “Sensor Models and Multisensor Integration,” *Int J Rob Res*, vol. 7, no. 6, pp. 97–113, Dec. 1988, DOI: 10.1177/027836498800700608.

[41] H. Gupta and D. Gupta, “LPC and LPCC method of feature extraction in Speech Recognition System,” in *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, IEEE, Jan. 2016, pp. 498–502. DOI: 10.1109/CONFLUENCE.2016.7508171.

[42] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *J Acoust Soc Am*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990, DOI: 10.1121/1.399423.

[43] I. Rebai, Y. Benayed, W. Mahdi, and J. P. Lorré, “Improving speech recognition using data augmentation and acoustic model fusion,” *Procedia Comput Sci*, vol. 112, pp. 316–322, Jan. 2017, DOI: 10.1016/J.PROCS.2017.08.003.

[44] ITU-R SM.328-11: Spectra and Bandwidth of Emissions,” International Telecommunication Union, Radiocommunication Sector, Recommendation ITU-R SM.328-11, 2011.

[45] J. Chaciński and K. Chacińska, *Podstawy emisji głosu w procesie kształcenia nauczycieli muzyki*., Słupsk: Wydawnictwo Uczelniane WSP, 1999.

[46] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," in *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314-323, Feb. 1988, DOI: 10.1109/49.608.

[47] J. H. Esling and H. Gaylord, “Computer Codes for Phonetic Symbols,” *Journal of the International Phonetic Association*, vol. 23, no. 2, pp. 83–97, 1993. doi:10.1017/S0025100300004898.

[48] J. H. L. Hansen and S. Patil, “Speech Under Stress: Analysis, Modeling and Recognition,” in Speaker Classification I, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 108–137. DOI: 10.1007/978-3-540-74200-5_6.

[49] J. M. Grey and J. W. Gordon, “Perceptual effects of spectral modifications on musical timbres,” *J Acoust Soc Am*, vol. 63, no. 5, pp. 1493–1500, May 1978, DOI: 10.1121/1.381843.

[50] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, “Robust speaker recognition in noisy conditions,” *IEEE Trans Audio Speech Lang Process*, vol. 15, no. 5, pp. 1711–1723, Jul. 2007, DOI: 10.1109/TASL.2007.899278.

[51] J. P. Campbell, "Speaker recognition: a tutorial," in Proceedings of the IEEE, vol. 85, no. 9, pp. 1437-1462, Sept. 1997, DOI: 10.1109/5.628714

[52] J. P. Campbell, D. A. Reynolds, and R. B. Dunn, "Fusing high- and low-level features for speaker recognition," in *Proceedings of Eurospeech 2003*, 2003, DOI: 10.21437/Eurospeech.2003-727.

[53] J. Son Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deep Speaker Recognition”, DOI: 10.21437/Interspeech.2018-1929.

[54] J. T. Hart, R. Collier, and A. Cohen, *A Perceptual Study of Intonation*. Cambridge University Press, 1990. DOI: 10.1017/CBO9780511627743.

[55] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, et al., “State-of-the-art speaker recognition for telephone and video speech: The JHU-MIT submission for NIST SRE18,” Proc. Interspeech, vol. 2019, pp. 1488–1492, 2019. DOI: 10.21437/Interspeech.2019.

[56] J.-P. Hosom, “Automatic phoneme alignment based on acoustic-phonetic modeling,” in *7th International Conference on Spoken Language Processing (ICSLP 2002)*, ISCA: ISCA, Sep. 2002, pp. 357–360. DOI: 10.21437/ICSLP.2002-154.

[57] K. A. Lee, O. Sadjadi, H. Li, and D. Reynolds, “Two decades into Speaker Recognition Evaluation - are we there yet?,” May 01, 2020, *Academic Press*. DOI: 10.1016/j.csl.2019.101058.

[58] K. Kamiński and A. Dobrowolski, "System automatycznego rozpoznawania mówcy z wykorzystaniem techniki cepstralnej i modeli mieszanin gaussowskich," Przegląd Elektrotechniczny, vol. 89, no. 9, pp. 83–93, 2013.

[59] K. Kamiński and A. P. Dobrowolski, “Automatic Speaker Recognition System Based on Gaussian Mixture Models, Cepstral Analysis, and Genetic Selection of Distinctive Features,” *Sensors*, vol. 22, no. 23, Dec. 2022, DOI: 10.3390/s22239370.

[60] K. Kamiński, „System automatycznego rozpoznawania mówcy oparty na analizie cepstralnej sygnału mowy i modelach mieszanin gaussowskich”, Rozprawa doktorska, Warszawa, 2018.

[61] K. Kamiński, A. Dobrowolski, E. Majda, and D. Posiadała, "Optymalizacja systemu automatycznego rozpoznawania mówcy w warunkach zróżnicowanych torów

akustycznych," *Przegląd Elektrotechniczny*, vol. 1, no. 9, pp. 91–94, Sep. 2015, DOI: 10.15199/48.2015.09.23.

[62] K. Kamiński, A. P. Dobrowolski, and E. Majda-Zdancewicz, "Selekcja cech osobniczych sygnału mowy z wykorzystaniem algorytmów genetycznych," *Biuletyn WAT*, vol. 65, no. 1, pp. 147–158, Mar. 2016, DOI: 10.5604/12345865.1197999.

[63] K. Kamiński, A. P. Dobrowolski, Z. Piotrowski, and P. Ścibiorek, "Enhancing Web Application Security: Advanced Biometric Voice Verification for Two-Factor Authentication," Electronics (Switzerland), vol. 12, no. 18, Sep. 2023, DOI: 10.3390/electronics12183791.

[64] K. Ślot, *Wybrane zagadnienia biometrii*, Warszawa: Wydawnictwo Naukowe PWN, 2008, 140 pp.

[65] K. Smith, *Precalculus: A Functional Approach to Graphing and Problem Solving*, 2nd ed., Jones & Bartlett Publishers, 2013.

[66] K. Zdziarski, *Emisja i higiena głosu*, Szczecińska Szkoła Wyższa Collegium Balticum, 2008.

[67] L. Burget, P. Matějka, P. Schwarz, O. Glembek, and J. H. Černocký, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," *IEEE Trans Audio Speech Lang Process*, vol. 15, no. 7, pp. 1979–1986, Sep. 2007, DOI: 10.1109/TASL.2007.902499.

[68] L. Deng and J. C. Platt, "Ensemble Deep Learning for Speech Recognition," in *Proceedings of Interspeech 2014*, Singapore, Sep. 2014, pp. 1915–1919. DOI: 10.21437/Interspeech.2014-433.

[69] L. Feng, "Speaker Recognition", M.Sc. thesis, Sept.

[70] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, 1st ed., Upper Saddle River, NJ, USA: Prentice Hall, 1993.

[71] M. Ali Humayun *et al.*, "Regularized Urdu Speech Recognition with Semi-Supervised Deep Learning," *Applied Sciences*, vol. 9, no. 9, p. 1956, May 2019, DOI: 10.3390/app9091956.

[72] M. Bojsza, „Ocena wpływu szumów i zakłóceń na skuteczność automatycznego systemu rozpoznawania mówcy w warunkach różnorodności językowej", Praca magisterska, Warszawa, 2021.

[73] M. Kirpluk, *Podstawy akustyki*, Pszczyna: NTL-M. Kirpluk, 2023.

[74] M. Kłaczyński, "Zjawiska wibroakustyczne w kanale głosowym człowieka," Rozprawa doktorska, Kraków, 2007.

[75] M. Tanveer *et al.*, "Ensemble deep learning in speech signal tasks: A review," *Neurocomputing*, vol. 550, p. 126436, Sep. 2023, DOI: 10.1016/J.NEUCOM.2023.126436.

[76] M. Tareq Hasan, A. Hadi Shakara, and A. Shakara, "A Signal processing approach to Music tutor," vol. 19, no. 6, pp. 13–25, DOI: 10.9790/0661-1906021325.

[77] N. Chauhan, T. Isshiki, and D. Li, “Text-Independent Speaker Recognition System Using Feature-Level Fusion for Audio Databases of Various Sizes,” SN Comput Sci, vol. 4, no. 5, Sep. 2023, DOI: 10.1007/s42979-023-02056-w.

[78] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788-798, May 2011, DOI: 10.1109/TASL.2010.2064307.

[79] N. Morgan *et al*., "Pushing the envelope - aside [speech recognition]," in *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 81-88, Sept. 2005, DOI: 10.1109/MSP.2005.1511826.

[80] N. N. Prachi, F. M. Nahiyan, M. Habibullah and R. Khan, "Deep Learning Based Speaker Recognition System with CNN and LSTM Techniques," *2022 Interdisciplinary Research in Technology and Management (IRTM)*, Kolkata, India, 2022, pp. 1-6, DOI: 10.1109/IRTM54583.2022.9791766.

[81] N. P. Jawarkar, R. S. Holambe and T. K. Basu, "Speaker Identification Using Whispered Speech," *2013 International Conference on Communication Systems and Network Technologies*, Gwalior, India, 2013, pp. 778-781, DOI: 10.1109/CSNT.2013.167.

[82] N. P. Jawarkar, R. S. Holambe and T. K. Basu, "Use of fuzzy min-max neural network for speaker identification," *2011 International Conference on Recent Trends in Information Technology (ICRTIT)*, Chennai, India, 2011, pp. 178-182, DOI: 10.1109/ICRTIT.2011.5972455.

[83] N. R. French and J. C. Steinberg, “Factors Governing the Intelligibility of Speech Sounds,” *J Acoust Soc Am*, vol. 19, no. 1, pp. 90–119, Jan. 1947, DOI: 10.1121/1.1916407.

[84] N. Singh, "A Study on Speech and Speaker Recognition Technology and its Challenges," *Proceedings of National Conference on Information Security Challenges*, 2014.

[85] N. Zheng, T. Lee and P. C. Ching, "Integration of Complementary Acoustic Features for Speaker Recognition," in *IEEE Signal Processing Letters*, vol. 14, no. 3, pp. 181-184, March 2007, DOI: 10.1109/LSP.2006.884031.

[86] N.-Q. Pham, T.-S. Nguyen, J. Niehues, M. Müller, S. Stüker, and A. Waibel, "Very Deep Self-Attention Networks for End-to-End Speech Recognition," *arXiv preprint arXiv:1904.13377*, 2019.

[87] O. Bell, “Applications of Gaussian Mutation for Self Adaptation in Evolutionary Genetic Algorithms,” *Journal of Machine Learning in Fundamental Sciences JMLFS-ID*, 2022, http://arxiv.org/abs/2201.00285.

[88] O. Caglayan, R. Sanabria, S. Palaskar, L. Barrault, and F. Metze, "Multimodal Grounding for Sequence-to-sequence Speech Recognition," in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 8648–8652.

[89] O. Novotný, O. Plchot, P. Matějka, L. Mošner, and O. Glembek, “On the use of X-vectors for Robust Speaker Recognition,” in *The Speaker and Language Recognition Workshop (Odyssey 2018)*, ISCA: ISCA, Jun. 2018, pp. 168–175. DOI: 10.21437/Odyssey.2018-24.

[90] O. Siohan and D. Rybach, "Multitask learning and system combination for automatic speech recognition," *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, USA, 2015, pp. 589-595, DOI: 10.1109/ASRU.2015.7404849.

[91] P. Kabal, *ITU-T G.723.1 Speech Coder: A Matlab Implementation*, TSP Lab Technical Report, Dept. of Electrical & Computer Engineering, McGill University, 2009.

[92] P. Kenny, P. Ouellet, N. Dehak, V. Gupta and P. Dumouchel, "A Study of Interspeaker Variability in Speaker Verification," in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980-988, July 2008, DOI: 10.1109/TASL.2008.925147.

[93] P. Král, "Discrete Wavelet Transform for automatic speaker recognition," *2010 3rd International Congress on Image and Signal Processing*, Yantai, China, 2010, pp. 3514-3518, DOI: 10.1109/CISP.2010.5646691.

[94] P. L. Rothman and R. V. Denton, “Fusion or confusion: knowledge or nonsense?” V. Libby, Ed., Aug. 1991, pp. 2–12. DOI: 10.1117/12.448351991.

[95] P. Matějka *et al.*, “13 years of speaker recognition research at BUT, with longitudinal analysis of NIST SRE,” *Comput Speech Lang*, vol. 63, Sep. 2020, DOI: 10.1016/j.csl.2019.101035.

[96] P. Matějka, O. Novotný, O. Plchot, L. Burget, M. D. Sánchez, and J. Černocký, “Analysis of Score Normalization in Multilingual Speaker Recognition,” in *Interspeech 2017*, ISCA: ISCA, Aug. 2017, pp. 1567–1571. DOI: 10.21437/Interspeech.2017-803.

[97] P. Staroniewicz, "Influence of Natural Voice Disguise Techniques on Automatic Speaker Recognition," *2018 Joint Conference - Acoustics*, Ustka, Poland, 2018, pp. 1-4, DOI: 10.1109/ACOUSTICS.2018.8502372.

[98] R. A. Makowski, *Automatyczne rozpoznawanie mowy: wybrane zagadnienia*, Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, 2011.

[99] R. C. Luo and M. G. Kay, "Multisensor integration and fusion in intelligent systems," in *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 5, pp. 901-931, Sept.-Oct. 1989, DOI: 10.1109/21.44007.

[100] R. C. Luo, Chih-Chen Yih and Kuo Lan Su, "Multisensor fusion and integration: approaches, applications, and future research directions," in *IEEE Sensors Journal*, vol. 2, no. 2, pp. 107-119, April 2002, DOI: 10.1109/JSEN.2002.1000251.

[101] R. E. Schapire, “The Boosting Approach to Machine Learning: An Overview,” 2003, pp. 149–171. DOI: 10.1007/978-0-387-21579-2_9.

[102] R. Loughran, A. Agapitos, A. Kattan, A. Brabazon, and M. O’Neill, “Feature selection for speaker verification using genetic programming,” *Evol Intell*, vol. 10, no. 1–2, pp. 1–21, Jul. 2017, DOI: 10.1007/s12065-016-0150-5.

[103] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior, *Tajemnica-Atak-Obrona, Biometria*. Warszawa: WNT, 2008.

[104] R. Mohd Hanifa, K. Isa, and S. Mohamad, “A review on speaker recognition: Technology and challenges,” *Computers & Electrical Engineering*, vol. 90, p. 107005, Mar. 2021, DOI: 10.1016/J.COMPELECENG.2021.107005.

[105] R. Quixtiano-Xicohtencatl, L. Flores-Pulido and O. F. Reyes-galaviz, "Feature Selection for a Fast Speaker Detection System with Neural Networks and Genetic Algorithms," *2006 15th International Conference on Computing*, Mexico City, Mexico, 2006, pp. 126-134, DOI: 10.1109/CIC.2006.38.

[106] R. Sharma, D. Govind, J. Mishra, A. K. Dubey, K. T. Deepak, and S. R. M. Prasanna, “Milestones in speaker recognition,” *Artif Intell Rev*, vol. 57, no. 3, Mar. 2024, DOI: 10.1007/s10462-023-10688-w.

[107] R. Tadeusiewicz, *Sygnał mowy*, Warszawa: Wydawnictwa Komunikacji i Łączności, 1988.

[108] S. Däubener, L. Schönherr, A. Fischer, and D. Kolossa, “Detecting Adversarial Examples for Speech Recognition via Uncertainty Quantification,” 2020, DOI: 10.21437/Interspeech.2020-2734.

[109] S. Dubnov, "Generalization of spectral flatness measure for non-Gaussian linear processes," in *IEEE Signal Processing Letters*, vol. 11, no. 8, pp. 698-701, Aug. 2004, DOI: 10.1109/LSP.2004.831663.

[110] S. Furui, “Cepstral analysis technique for automatic speaker verification,” *IEEE Trans Acoust*, vol. 29, no. 2, pp. 254–272, Apr. 1981, DOI: 10.1109/TASSP.1981.1163530.

[111] S. Krishnakumar, K. R. Prasanna Kumar, and N. Balakrishnan, “Pitch maxima for robust speaker recognition,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2003, pp. 201–204. DOI: 10.1109/icassp.2003.1202329.

[112] S. M. Monroe and G. M. Slavich, "Psychological Stressors: Overview," in *Stress: Concepts, Cognition, Emotion, and Behavior*, G. Fink, Ed., 1st ed., Academic Press, 2016, pp. 109–115.

[113] S. Osowski, *Metody i narzędzia eksploracji danych*, Legionowo: BTC, 2013.

[114] S. R. M. Prasanna and G. Pradhan, “Significance of Vowel-Like Regions for Speaker Verification Under Degraded Conditions,” *IEEE Trans Audio Speech Lang Process*, vol. 19, no. 8, pp. 2552–2565, Nov. 2011, doi: 10.1109/TASL.2011.2155061.

[115] S. R. Mahadeva Prasanna, C. S. Gupta, and B. Yegnanarayana, “Extraction of speaker-specific excitation information from linear prediction residual of speech,” *Speech Commun*, vol. 48, no. 10, pp. 1243–1261, Oct. 2006, DOI: 10.1016/J.SPECOM.2006.06.002.

[116] S. S. Tirumala, S. R. Shahamiri, A. S. Garhwal, and R. Wang, “Speaker identification features extraction methods: A systematic review,” Dec. 30, 2017, *Elsevier Ltd*. DOI: 10.1016/j.eswa.2017.08.015.

[117] S. Selva Nidhyananthan, R. Shantha Selva Kumari and G. Jaffino, "Robust speaker identification using vocal source information," *2012 International Conference on Devices, Circuits and Systems (ICDCS)*, Coimbatore, India, 2012, pp. 182-186, DOI: 10.1109/ICDCSyst.2012.6188700.

[118] S. Sharma, Nagpal C. K., "Data Fusion Reference Models: Basic Concept, Architecture and Comparison," *International Journal of Emerging Technologies and Innovative Research, vol. 6, no. 6, pp. 241–245, Jun. 2019.*

[119] S. Singh, "High level speaker specific features as an efficiency enhancing parameters in speaker recognition system," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 4, pp. 2443–2450, Aug. 2019, DOI: 10.11591/ijece.v9i4.pp2443-2450.

[120] S. Solorio-Fernández, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad, "A review of unsupervised feature selection methods," *Artif Intell Rev*, vol. 53, no. 2, pp. 907–948, Feb. 2020, DOI: 10.1007/s10462-019-09682-y.

[121] S. Sreedharan and C. Eswaran, "An Analysis of Feature Selection based on Optimization Algorithms for Speaker Verification," *2018 Tenth International Conference on Advanced Computing (ICoAC)*, Chennai, India, 2018, pp. 105-111, DOI: 10.1109/ICoAC44903.2018.8939099.

[122] S. Sujiya and E. Chandra, "A Review on Speaker Recognition," *International Journal of Engineering and Technology (IJET)*, vol. 9, no. 3, pp. 1592–1598, Jun.–Jul. 2017. DOI: 10.21817/ijet/2017/v9i3/170903513.

[123] S. V. Chougule and M. S. Chavan, "Robust Spectral Features for Automatic Speaker Recognition in Mismatch Condition," in *Procedia Computer Science*, Elsevier B.V., 2015, pp. 272–279. DOI: 10.1016/j.procs.2015.08.021.

[124] S. Young, "A review of large-vocabulary continuous-speech," in *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 45-, Sept. 1996, DOI: 10.1109/79.536824.

[125] T. C. Henderson, M. Dekhil, R. R. Kessler, and M. L. Griss, "Sensor fusion," in *Control Problems in Robotics and Automation*, London: Springer-Verlag, pp. 193–207. DOI: 10.1007/BFb0015084.

[126] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *Multiple Classifier Systems*, MCS 2000, Lecture Notes in Computer Science, vol. 1857, Springer, Berlin, Heidelberg, 2000. DOI: 10.1007/3-540-45014-9_1.

[127] T. Giannakopoulos and A. Pikrakis, *Introduction to Audio Analysis: A MATLAB Approach*. Elsevier Ltd, 2014. DOI: 10.1016/C2012-0-03524-7.

[128] T. Iliou and C. -N. Anagnostopoulos, "Statistical Evaluation of Speech Features for Emotion Recognition," *2009 Fourth International Conference on Digital Telecommunications*, Colmar, France, 2009, pp. 121-126, DOI: 10.1109/ICDT.2009.30.

[129] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun*, vol. 52, no. 1, pp. 12–40, Jan. 2010, DOI: 10.1016/j.specom.2009.08.009.

[130] T. Kinnunen and P. Alku, "On separating glottal source and vocal tract information in telephony speaker verification," *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, 2009, pp. 4545-4548, DOI: 10.1109/ICASSP.2009.4960641.

[131] T. Kinnunen, "Joint Acoustic-Modulation Frequency for Speaker Recognition," *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Toulouse, France, 2006, pp. I-I, DOI: 10.1109/ICASSP.2006.1660108.

[132] T. Kinnunen, B. Zhang, J. Zhu, and Y. Wang, "Speaker Verification with Adaptive Spectral Subband Centroids," in *Advances in Biometrics*, Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 58–66. DOI: 10.1007/978-3-540-74549-5_7.

[133] T. Kinnunen, C. Wei, E. Koh, L. Wang, H. Li, and E. S. Chng, "Temporal Discrete Cosine Transform: Towards Longer Term Temporal Features for Speaker Verification," in *Proc. ISCSLP 2006*, Singapore, 2006, pp. 547–558.

[134] T. Kinnunen, K.-A. Lee, and H. Li, "Dimension Reduction of the Modulation Spectrogram for Speaker Verification," in *Proc. The Speaker and Language Recognition Workshop*, South Africa, 2008.

[135] T. P. Zieliński, *Cyfrowe przetwarzanie sygnałów. Od teorii do zastosowań*, 2nd ed., Warszawa: Wydawnictwa Komunikacji i Łączności, 2007.

[136] T. Różniatowski, *Mała encyklopedia medycyny t. 1–3*, Warszawa: Wydawnictwo Naukowe PWN, 1989.

[137] V. Hautamäki, M. Tuononen, T. Niemi-Laitinen, and P. Fränti, "Improving Speaker Verification by Periodicity Based Voice Activity Detection," in *Proceedings of SPECOM 2007*, Russia, 2007, pp. 645–650.

[138] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, 2015, pp. 5206-5210, DOI: 10.1109/ICASSP.2015.7178964.

[139] V. R. Sreehari and L. Mary, "Automatic short utterance speaker recognition using stationary wavelet coefficients of pitch synchronised LP residual," *Int J Speech Technol*, vol. 25, no. 1, pp. 147–161, Mar. 2022, DOI: 10.1007/s10772-021-09895-z.

[140] W. Elmenreich, *Sensor Fusion in Time-Triggered Systems*, Ph.D. dissertation, Jan. 2002.

[141] W. M. Campbell, D. E. Sturim and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," in *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308-311, May 2006, DOI: 10.1109/LSP.2006.870086.

[142] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Comput Speech Lang*, vol. 20, no. 2–3, pp. 210–229, Apr. 2006, DOI: 10.1016/J.CSL.2005.06.003.

[143] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Upper Saddle River, NJ: Prentice Hall PTR, 2001.

[144] Y. H. Tu *et al.*, “An iterative mask estimation approach to deep learning based multi-channel speech recognition,” *Speech Commun*, vol. 106, pp. 31–43, Jan. 2019, DOI: 10.1016/J.SPECOM.2018.11.005.

[145] Y. Ren, L. Zhang and P. N. Suganthan, "Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article]," in *IEEE Computational Intelligence Magazine*, vol. 11, no. 1, pp. 41-53, Feb. 2016, DOI: 10.1109/MCI.2015.2471235.

[146] Y. Xing, H. Li and P. Tan, "Hierarchical fuzzy speaker identification based on FCM and FSVM," *2012 9th International Conference on Fuzzy Systems and Knowledge Discovery*, Chongqing, China, 2012, pp. 311-315, DOI: 10.1109/FSKD.2012.6233866.

[147] Z. Piotrowski, J. Wojtuń, and K. Kamiński, "Subscriber authentication using GMM and TMS320C6713DSP," *Przegląd Elektrotechniczny*, vol. 88, no. 12a, pp. 302–304, 2012.

[148] Z. Żyszkowski, *Podstawy elektroakustyki*, 2nd ed., Warszawa: Wydawnictwa Naukowo-Techniczne, 1984

Strony www:

[149] https://catalog.ldc.upenn.edu/LDC2004S04, dostęp: 01.11.2024 r.

[150] https://commonvoice.mozilla.org/pl/datasets, dostęp: 01.11.2024 r.

[151] https://en.wikipedia.org/wiki/Taxicab_geometry, dostęp: 01.11.2024 r.

[152] https://encyklopedia.pwn.pl/haslo/idiolekt;3913863.html, dostęp: 01.11.2024 r.

[153] https://only4music.com/czytelnia/18-musictech/704-mikrofony-czyli-jak-nagrac-swoj-glos, dostęp: 01.11.2024 r.

[154] https://web.stanford.edu/dept/linguistics/corpora/material/X_Speech_Corpora.pdf, dostęp: 01.11.2024 r.

[155] https://www.mathworks.com/help/matlab/ref/audiowrite.html, dostęp: 01.11.2024 r.

[156] https://www.mathworks.com/help/signal/ug/prominence.html, dostęp: 01.11.2024 r.

[157] https://www.openslr.org/index.html, dostęp: 01.11.2024 r.

[158] https://www.speex.org/, dostęp: 01.11.2024 r.

Z. 1

Zbiór definicji wszystkich cech behawioralnych występujących w systemie ASR

Do konstrukcji pierwszej wersji behawioralnego systemu automatycznego rozpoznawania mówcy zastosowano 71 – zdefiniowanych w niniejszym załączniku – cech. W celu uproszczenia zapisu definicji cech, Autor stosuje tu skrótowe zapisy takich podstawowych parametrów jak wartość minimalna, maksymalna i średnia oraz odchylenie standardowe, zgodnie ze składnią języka *Matlab*:

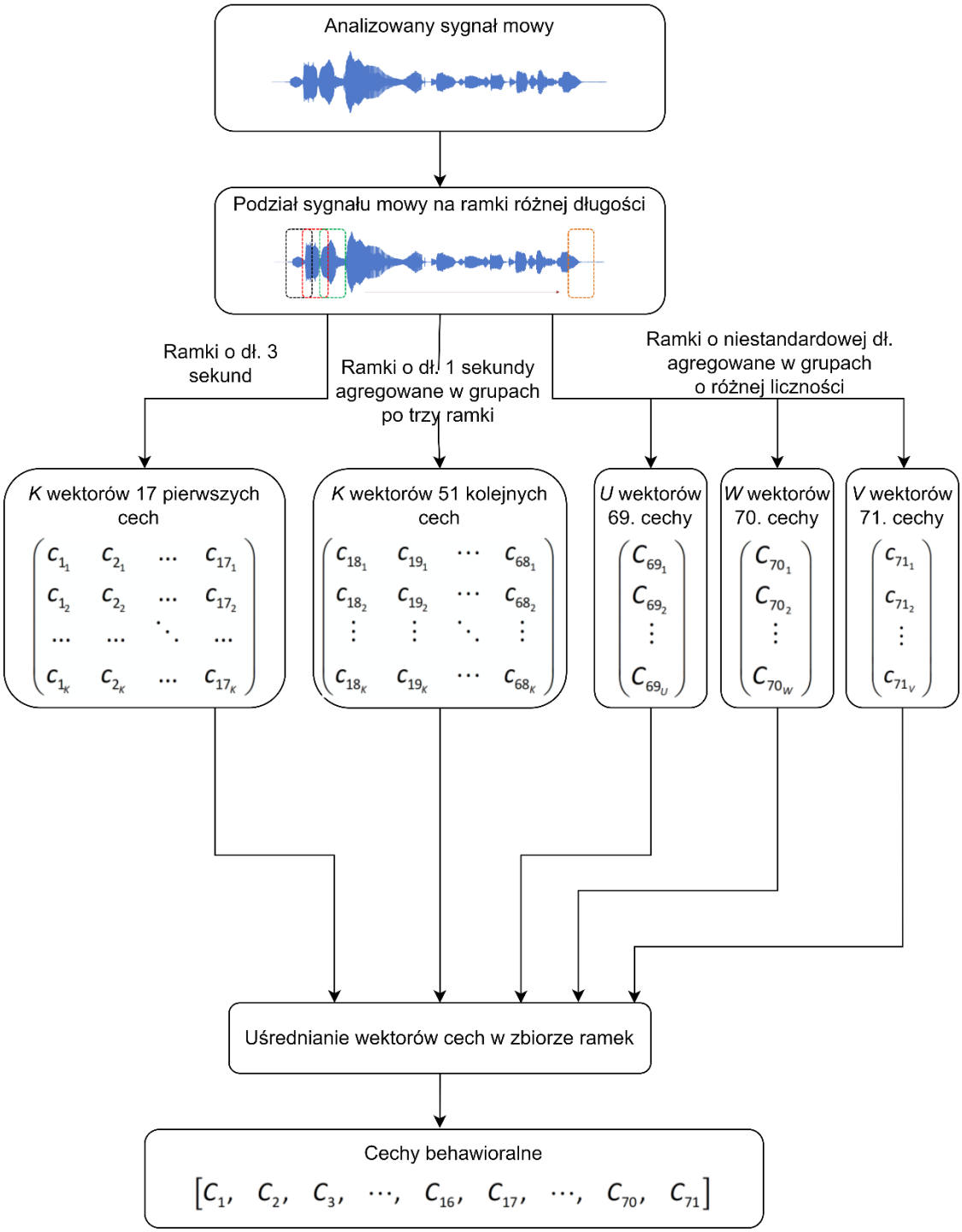
$\min(X)$ = najmniejsza spośród wszystkich wartości elementów wektora *X*

$\max(X)$ = największa spośród wszystkich wartości elementów wektora *X*

$$\mathrm{mean}(X) = \frac{1}{N}\sum_{n=1}^{N} x_n$$
$$\mathrm{std}(X) = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}\left(x_n - \mathrm{mean}(X)\right)^2} \qquad \text{(Z.1.1)}$$

gdzie X oznacza wektor o długości N: $X = [x_1, x_2, \ldots, x_N]^T$.

Na rys. Z.1.1 przedstawiono schemat wyznaczania 71 cech behawioralnych wykorzystywanych w pierwotnej wersji systemu.



Rys. Z.1.1. Uproszczony schemat blokowy wyznaczania cech behawioralnych

Cechy liczone z trzysekundowych ramek głosu:

17 pierwszych cech, tj. $[C_1, C_2, ..., C_{17}]$, są to cechy będące wartościami średnimi otrzymanymi po uśrednieniu cech otrzymanych z poszczególnych 3-sekundowych ramek[1]. Oznaczając cechy wyekstrahowane z *i*-tej ramki jako $\left[c_{1_i}, c_{2_i}, ..., c_{17_i}\right]$, pierwsze 17 cech określonych jest zależnością:

$$[C_1, C_2, ..., C_{17}] = \text{mean}\begin{pmatrix} c_{1_1} & c_{2_1} & ... & c_{17_1} \\ c_{1_2} & c_{2_2} & ... & c_{17_2} \\ ... & ... & \ddots & ... \\ c_{1_K} & c_{2_K} & ... & c_{17_K} \end{pmatrix} \tag{Z.1.2}$$

gdzie *K* oznacza liczbę 3-sekundowych ramek w całej rozpatrywanej wypowiedzi.

1. Stosunek czasu trwania mowy do czasu trwania ciszy w ramce:

$$c_1 = \frac{N_{speech}}{N_{silence}} \tag{Z.1.3}$$

gdzie: N_{speech} to liczba próbek sygnału zawierających mowę,

$N_{silence}$ to liczba próbek sygnału reprezentujących ciszę.

2. Czas trwania mowy w ramce unormowany względem liczby przedziałów mowy (średni czas trwania segmentu mowy):

$$c_2 = \frac{T_{speech}}{K_{speech}} \tag{Z.1.4}$$

gdzie: T_{speech} oznacza czas trwania mowy w ramce

K_{speech} oznacza liczbę segmentów mowy w ramce.

3. Unormowany średni czas trwania mowy w ramce:

$$c_3 = \frac{T_{speech}}{t_N} \tag{Z.1.5}$$

4. Liczba segmentów mowy w ramce:

$$c_4 = K_{speech} \tag{Z.1.6}$$

5. Maksymalna wartość sumy amplitud widma w oktawach:

$$c_5 = \max\left(SA_{MO}\right) \tag{Z.1.7}$$

gdzie: SA_{MO} oznacza wektor sum amplitud we wszystkich oktawach, przy czym pojedyncza suma dla *k*-tej oktawy wyznaczana jest ze wzoru

[1] Przy szybkości próbkowania f_p = 8 kSa/s, ramka o czasie trwania t_N = 3 s liczy N = 24000 próbek.

$$SA_{MO_k} = \sum_{n=0}^{N_k-1} A_{k_n} \tag{Z.1.8}$$

A_{k_n} to amplituda *n*-tej próbki widma *k*-tej oktawy,

N_k to liczba próbek w *k*-tej oktawie.

6. Maksymalna wartość sumy amplitud widma w tercjach:

$$c_6 = \max\left(SA_{MT}\right) \tag{Z.1.9}$$

gdzie: SA_{MT} oznacza wektor sum amplitud we wszystkich tercjach, przy czym pojedyncza suma dla *k*-tej tercji wyznaczana jest ze wzoru

$$SA_{MT_k} = \sum_{n=0}^{N_k-1} A_{k_n} \tag{Z.1.10}$$

A_{k_n} to amplituda *n*-tej próbki widma *k*-tej tercji,

N_k to liczba próbek w *k*-tej tercji.

7. Średnia wartość amplitudy w widmie oktawy o maksymalnej sumie amplitud:

$$c_7 = \mathrm{mean}\left(A_{oct_{\max}}\right) \tag{Z.1.11}$$

gdzie $A_{oct_{\max}}$ to zbiór amplitud wszystkich próbek w oktawie o maksymalnej sumie amplitud.

8. Odchylenie standardowe amplitudy w widmie oktawy o maksymalnej sumie amplitud:

$$c_8 = \mathrm{std}\left(A_{oct_{\max}}\right) \tag{Z.1.12}$$

9. Stosunek maksymalnej do minimalnej amplitudy w widmie oktawy o maksymalnej sumie amplitud:

$$c_9 = \frac{\max\left(A_{oct_{\max}}\right)}{\min\left(A_{oct_{\max}}\right)} \tag{Z.1.13}$$

10. Odchylenie standardowe amplitudy we wszystkich oktawach:

$$c_{10} = \mathrm{std}\left(A_{oct}\right) \tag{Z.1.14}$$

gdzie A_{oct} to zbiór sum amplitud widm we wszystkich oktawach.

11. Stosunek maksymalnej do minimalnej sumy amplitud we wszystkich oktawach:

$$c_{11} = \frac{\max\left(A_{oct}\right)}{\min\left(A_{oct}\right)} \tag{Z.1.15}$$

12. Średnia wartość amplitudy w widmie tercji o maksymalnej sumie amplitud:

$$c_{12} = \mathrm{mean}\left(A_{th_{\max}}\right) \tag{Z.1.16}$$

gdzie $A_{th_{\max}}$ to zbiór amplitud wszystkich próbek w tercji o maksymalnej sumie amplitud.

13. Odchylenie standardowe amplitudy w widmie tercji o maksymalnej sumie amplitud:

$$c_{13} = \mathrm{std}\left(A_{th_{\max}} \right) \tag{Z.1.17}$$

14. Stosunek maksymalnej do minimalnej amplitudy w widmie tercji o maksymalnej sumie amplitud:

$$c_{14} = \frac{\max\left(A_{th_{\max}} \right)}{\min\left(A_{th_{\max}} \right)} \tag{Z.1.18}$$

15. Odchylenie standardowe amplitudy we wszystkich tercjach:

$$c_{15} = \mathrm{std}\left(A_{th} \right) \tag{Z.1.19}$$

gdzie A_{th} to zbiór sum amplitud widm we wszystkich tercjach.

16. Stosunek maksymalnej do minimalnej sumy amplitud we wszystkich tercjach:

$$c_{16} = \frac{\max\left(A_{th} \right)}{\min\left(A_{th} \right)} \tag{Z.1.20}$$

17. Tempo przejść sygnału przez zero:

cecha c_{17} opisana jest wzorem (4.1), tj. $c_{17} = ZCR_i$.

Cechy liczone dla jednosekundowych ramek głosu:

Sposób wyznaczania kolejnych 51 cech, tj. $[C_{18}, C_{19}, ..., C_{68}]$ przebiega podobnie jak dla pierwszych 17 cech, tj. zgodnie z rys. Z.1.1, ale cechy te są dodatkowo wstępnie uśredniane w grupach zawierających kolejne trzy jednosekundowe ramki[1]. Oznaczając cechy wyekstrahowane z *i*-tej ramki jako $\left[c_{18_i}, c_{17_i}, ..., c_{68_i}\right]$, 51 kolejnych cech określonych jest zależnościami

$$[C_{18}, C_{19}, ..., C_{68}] = \text{mean}\begin{pmatrix} c_{18_1} & c_{19_1} & \cdots & c_{68_1} \\ c_{18_2} & c_{19_2} & \cdots & c_{68_2} \\ \vdots & \vdots & \ddots & \vdots \\ c_{18_K} & c_{19_K} & \cdots & c_{68_K} \end{pmatrix} \tag{Z.1.21}$$

gdzie *K* oznacza liczbę trójek jednosekundowych ramek w całej rozpatrywanej wypowiedzi.

18. Wartość częstotliwości środkowej w ramce [6]:

$$f_{med} = i_{med} \cdot \frac{f_p}{N} \tag{Z.1.22}$$

gdzie: i_{med} jest indeksem próbki widma określającej połowę mocy, tj.:

$$\sum_{i=1}^{i_{med}} P_i = \sum_{i=i_{med}}^{N} P_i = \frac{1}{2}\sum_{i=1}^{N} P_i \tag{Z.1.23}$$

N oznacza liczę próbek widma ramki (f_p/N to odstęp między próbkami widma),

P_i to moc *i*-tej próbki widma.

Wartość cechy wyznaczana jest ze wzoru:

$$c_{18} = \text{mean}\left(F_{med}\right) \tag{Z.1.24}$$

gdzie $F_{med} = \left[f_{med_1},\ f_{med_2},\ f_{med_3}\right]^T$ oznacza wektor częstości środkowych w trzech kolejnych ramkach tworzących grupę.

19. Odchylenie standardowe częstotliwości środkowej w grupie 3 ramek:

$$c_{19} = \text{std}\left(F_{med}\right) \tag{Z.1.25}$$

[1] Przy częstotliwości próbkowania f_p = 8 kSa/s, ramka o czasie trwania t_N = 1 s liczy *N* = 8000 próbek, a wartości kolejnych cech to średnie lub odchylenia standardowe liczone w zbiorze 3 jednosekundowych ramek.

20. Częstotliwość podstawowa:

$$\begin{gathered} i_{\max_{2,k}} = \arg\max_{i}{}_{2}\left(r_k\right) \\ f_{0_k} = \frac{f_p}{i_{\max_{2,k}}} \\ F_0 = \left[f_{0_1}, \quad f_{0_2}, \quad f_{0_3} \right]^T \\ c_{20} = \operatorname{mean}\left(F_0\right) \end{gathered} \tag{Z.1.26}$$

gdzie: r_k to wartość funkcji autokorelacji *k*-tej ramki opisana wzorem (3.5),

funkcja argmax_2 wyznacza drugie maksimum wektora będącego argumentem,

$i_{\max_k}$ to indeks próbki funkcji autokorelacji w *k*-tej ramce identyfikujący drugie maksimum, odpowiadające częstotliwości podstawowej,

f_{0_k} to częstotliwość podstawowa *k*-tej ramki,

F_0 to wektor częstotliwości podstawowych w trzech kolejnych ramkach tworzących grupę.

21. Stosunek maksymalnej do minimalnej częstotliwości podstawowej w grupie 3 ramek:

$$c_{21} = \frac{\max\left(F_0\right)}{\min\left(F_0\right)} \tag{Z.1.27}$$

22. Stosunek odchylenia standardowego częstotliwości podstawowej w grupie 3 ramek unormowany względem średniej częstotliwości podstawowej:

$$c_{22} = \frac{\operatorname{std}\left(F_0\right)}{c_{18}} \tag{Z.1.28}$$

23. Stosunek maksymalnej do minimalnej wartości krótkoczasowej energii w grupie 3 ramek:

$$c_{23} = \frac{\max\left(E\right)}{\min\left(E\right)} \tag{Z.1.29}$$

gdzie $E = \left[e_1, \quad e_2, \quad e_3 \right]^T$ to zbiór wartości krótkoczasowych energii ramek, opisanych wzorem (2.2) – przy czym tu dla porządku zmieniono oznaczenie z *E* na *e* – w trzech kolejnych ramkach tworzących grupę.

24. Odchylenie standardowe krótkoczasowej energii unormowane względem średniej krótkoczasowej energii w grupie 3 ramek:

$$c_{24} = \frac{\operatorname{std}\left(E\right)}{\operatorname{mean}\left(E\right)} \tag{Z.1.30}$$

25. Średnia wartość sumy próbek widma mocy w grupie 3 ramek:

$$c_{25} = \operatorname{mean}\left(P\right) \tag{Z.1.31}$$

gdzie: $P = \left[P_{\max_1},\ P_{\max_2},\ P_{\max_3} \right]^T$ to zbiór wartości mocy maksymalnych z widm ramek w trzech kolejnych ramkach tworzących grupę,

$P_{\max_i}$ to wartość maksimum mocy w *i*-tej ramce, tj.:

$$P_{\max_i} = \max\left(\sum_{l=1}^{N} \left(\left| S_i(l) \right|^2 \right) \right) \tag{Z.1.32}$$

$S_i(l)$ to wartość *l*-tej próbki widma *i*-tej ramki.

26. Stosunek maksimum do minimum sumy próbek widma mocy ramek w grupie 3 ramek:

$$c_{26} = \frac{\max(P)}{\min(P)} \tag{Z.1.33}$$

27. Odchylenie standardowe sumy próbek widma mocy w grupie 3 ramek unormowane względem wartości średniej spośród tych mocy:

$$c_{27} = \frac{\mathrm{std}(P)}{\mathrm{mean}(P)} \tag{Z.1.34}$$

28. Średnia wartość częstotliwości odpowiadającej maksimum mocy w widmie w grupie 3 ramek:

$$c_{28} = \mathrm{mean}(F) \tag{Z.1.35}$$

gdzie: $F = \left[f_{m_1},\ f_{m_2},\ f_{m_3} \right]^T$ to zbiór wartości częstotliwości odpowiadających maksimom mocy widm w trzech kolejnych ramkach tworzących grupę,

f_{m_i} to wartość częstotliwości odpowiadającej maksimum mocy, tj.:

$$f_{m_i} = i_{m_i} \cdot \frac{f_p}{N} \tag{Z.1.36}$$

i_{m_i} jest indeksem próbki widma określającej maksimum mocy w *i*-tej ramce

N oznacza liczę próbek widma ramki (f_p/N to odstęp między próbkami widma).

29. Stosunek maksymalnej do minimalnej częstotliwości odpowiadającej maksymalnej mocy w widmach ramek w grupie 3 ramek:

$$c_{29} = \frac{\max(F)}{\min(F)} \tag{Z.1.37}$$

30. Odchylenie standardowe częstotliwości odpowiadających maksimom mocy w grupie 3 ramek unormowane względem wartości średniej tej częstotliwości:

$$c_{30} = \frac{\mathrm{std}(F)}{\mathrm{mean}(F)} \tag{Z.1.38}$$

31. Stosunek maksymalnej do minimalnej wartości średniokwadratowej amplitudy ramek w grupie 3 ramek:

$$c_{31} = \frac{\max\left(A_{rms}\right)}{\min\left(A_{rms}\right)} \tag{Z.1.39}$$

gdzie: $A_{rms} = \left[a_{rms_1},\ a_{rms_2},\ a_{rms_3}\right]^T$ to zbiór wartości średniokwadratowych amplitud w trzech kolejnych ramkach tworzących grupę,

a_{rms_i} to wartość średniokwadratowa amplitudy *i*-tej ramki, tj.:

$$a_{rms_i} = \sqrt{\frac{1}{N}\sum_{k=1}^{N} x_k^{\,2}} \tag{Z.1.40}$$

x_k to wartość *k*-tej próbki.

32. Odchylenie standardowe wartości średniokwadratowych amplitud w grupie 3 ramek unormowane względem średniej spośród wartości średniokwadratowych amplitud:

$$c_{32} = \frac{\mathrm{std}\left(A_{rms}\right)}{\mathrm{mean}\left(A_{rms}\right)} \tag{Z.1.41}$$

33. Średnia wartość zajętości pasma zawierającego 95% mocy całkowitej w grupie 3 ramek:

$$c_{33} = \mathrm{mean}\left(OBW\right) \tag{Z.1.42}$$

gdzie: $OBW = \left[obw_1,\ obw_2,\ obw_3\right]^T$ to zbiór wartości zajętości pasma w trzech kolejnych ramkach tworzących grupę,

obw_i to wartość zajętości pasma *i*-tej ramki opisana wzorem (4.13), przy czym tu dla porządku zmieniono oznaczenie z *OBW* na *obw*.

34. Stosunek maksimum do minimum zajętości pasma w grupie 3 ramek:

$$c_{34} = \frac{\max\left(OBW\right)}{\min\left(OBW\right)} \tag{Z.1.43}$$

35. Odchylenie standardowe zajętości pasma w grupie 3 ramek unormowane względem wartości średniej zajętości pasma ramek:

$$c_{35} = \frac{\mathrm{std}\left(OBW\right)}{\mathrm{mean}\left(OBW\right)} \tag{Z.1.44}$$

36. Średnia wartość płaskości widma w grupie 3 ramek [46]:

$$c_{36} = \mathrm{mean}\left(SFM\right) \tag{Z.1.45}$$

gdzie: $SFM = \left[sfm_1,\ sfm_2,\ sfm_3\right]^T$ to zbiór wartości miar płaskości widma w trzech kolejnych ramkach tworzących grupę,

sfm_i to miara płaskości widma *i*-tej ramki opisana wzorem (4.14), przy czym tu dla porządku zmieniono oznaczenie z *SFM* na *sfm*.

37. Stosunek maksimum do minimum miary płaskości widma:

$$c_{37} = \frac{\max(SFM)}{\min(SFM)} \tag{Z.1.46}$$

38. Odchylenie standardowe płaskości widma w grupie 3 ramek unormowane do wartości średniej:

$$c_{38} = \frac{\mathrm{std}(SFM)}{\mathrm{mean}(SFM)} \tag{Z.1.47}$$

39. Średnia wartość widmowego współczynnika szczytu w grupie 3 ramek [39]:

$$c_{39} = \mathrm{mean}(SCR) \tag{Z.1.48}$$

gdzie: $SCR = [scr_1,\ scr_2,\ scr_3]^T$ to zbiór wartości widmowych współczynników szczytu widma trzech kolejnych ramkach tworzących grupę,

scr_i to wartość widmowego współczynnika szczytu *i*-tej ramki, tj.:

$$scr_i = \frac{\max(S_i)}{\frac{1}{L}\sum_{l=0}^{L-1} S_i(l)} \tag{Z.1.49}$$

L oznacza liczbę próbek w ramce,

$S_i(l)$ to kolejne próbki widma *i*-tej ramki,

S_i to zbiór wszystkich próbek widma *i*-tej ramki.

40. Stosunek maksimum do minimum widmowego współczynnika szczytu w grupie 3 ramek:

$$c_{40} = \frac{\max(SCR)}{\min(SCR)} \tag{Z.1.50}$$

41. Odchylenie standardowe widmowego współczynnika szczytu w grupie 3 ramek unormowane względem średniej:

$$c_{41} = \frac{\mathrm{std}(SCR)}{\mathrm{mean}(SCR)} \tag{Z.1.51}$$

42. Średnia wartość entropii widma w grupie 3 ramek [127]:

$$c_{42} = \mathrm{mean}(SE) \tag{Z.1.52}$$

gdzie: $SE = [se_1,\ se_2,\ se_3]^T$ to zbiór wartości entropii widm w trzech kolejnych ramkach tworzących grupę,

se_i to wartość entropii widma *i*-tej ramki, tj.:

$$se_i = -\sum_{n=0}^{N_s-1}\left(P_{Norm_i}(n)\cdot\log_2\left(P_{Norm_i}(n)\right)\right) \tag{Z.1.53}$$

$P_{Norm_i}(n)$ to unormowana wartość mocy *n*-tej składowej widma *i*-tej ramki, tj.:

$$P_{Norm_i}(n) = \frac{P_i(n)}{\sum_{n=0}^{N_s-1} P_i(n)} \tag{Z.1.54}$$

$P_i(n)$ to moc *n*-tej składowej widma *i*-tej ramki,

N_s to liczba próbek sygnału w analizowanej ramce.

43. Stosunek maksimum do minimum wartości entropii widma w grupie 3 ramek:

$$c_{43} = \frac{\max(SE)}{\min(SE)} \tag{Z.1.55}$$

44. Odchylenie standardowe entropii widma w grupie 3 ramek unormowane względem średniej:

$$c_{44} = \frac{\mathrm{std}(SE)}{\mathrm{mean}(SE)} \tag{Z.1.56}$$

45. Średnia wartość strumienia widmowego w grupie 3 ramek [5], [127]:

$$c_{45} = \mathrm{mean}(SFL) \tag{Z.1.57}$$

gdzie: $SFL = \left[\, sfl_{2,1},\ \ sfl_{3,2} \,\right]^T$ to zbiór wartości strumieni widmowych pomiędzy sąsiadującymi ramkami,

$sfl_{(i,i-1)}$ to wartość strumienia widmowego pomiędzy sąsiadującymi ramkami, tj.:

$$sfl_{(i,i-1)} = \sum_{l=0}^{L-1}\left(S_{Norm_i}(l) - S_{Norm_{i-1}}(l)\right)^2 \tag{Z.1.58}$$

L oznacza liczbę próbek w ramce,

$S_{Norm_i}(l)$ to unormowana wartość *l*-tej próbki widma *i*-tej ramki, tj.:

$$S_{Norm_i}(l) = \frac{S_i(l)}{\sum_{l=1}^{L-1} S_i(l)} \tag{Z.1.59}$$

46. Stosunek maksimum do minimum wartości strumienia widmowego w grupie 3 ramek:

$$c_{46} = \frac{\max(SFL)}{\min(SFL)} \tag{Z.1.60}$$

47. Odchylenie standardowe strumienia widmowego w grupie 3 ramek unormowane względem średniej:

$$c_{47} = \frac{\mathrm{std}(SFL)}{\mathrm{mean}(SFL)} \tag{Z.1.61}$$

48. Średnia wartość spadku widmowego w grupie 3 ramek [5]:

$$c_{48} = \mathrm{mean}(SD) \tag{Z.1.62}$$

gdzie: $SD = [sd_1,\ sd_2,\ sd_3]^T$ to zbiór wartości spadków widmowych w trzech kolejnych ramkach tworzących grupę,

sd_i to spadek widmowy *i*-tej ramki, tj.:

$$sd_i = \frac{\sum_{l=2}^{L-1} \frac{S_i(l) - S_i(1)}{l-1}}{\sum_{l=2}^{L-1} S_i(l)} \tag{Z.1.63}$$

$S_i(l)$ to wartość *l*-tej próbki widma *i*-tej ramki,

L oznacza liczbę próbek w ramce.

49. Stosunek maksimum do minimum wartości spadku widmowego w grupie 3 ramek:

$$c_{49} = \frac{\max(SD)}{\min(SD)} \tag{Z.1.64}$$

50. Odchylenie standardowe spadku widmowego w grupie 3 ramek unormowane względem średniej:

$$c_{50} = \frac{\mathrm{std}(SD)}{\mathrm{mean}(SD)} \tag{Z.1.65}$$

51. Średnia wartość widmowego punktu odcięcia w grupie 3 ramek [5], [127]:

$$c_{51} = \mathrm{mean}(SRP) \tag{Z.1.66}$$

gdzie: $SRP = [srp_1,\ srp_2,\ srp_3]^T$ to zbiór wartości widmowych punktów odcięcia w trzech kolejnych ramkach tworzących grupę,

srp_i to widmowy punkt odcięcia *i*-tej ramki, tj.:

$$srp_i = \frac{j \cdot f_s}{N} \Leftrightarrow \sum_{l=0}^{j} S_i(l) = C \sum_{l=0}^{L-1} S_i(l) \tag{Z.1.67}$$

$S_i(l)$ to wartość *l*-tej próbki widma *i*-tej ramki,

C to parametr określający procent mocy dla którego następuję „odcięcie” (w finalnej wersji systemu wartość ta została dobrana w sposób eksperymentalny),

j to numer próbki, dla której następuję „odcięcie”.

52. Stosunek maksimum do minimum wartości widmowego punktu odcięcia w grupie 3 ramek:

$$c_{52} = \frac{\max\left(SRP\right)}{\min\left(SRP\right)} \tag{Z.1.68}$$

53. Odchylenie standardowe wartości widmowego punktu odcięcia w grupie 3 ramek unormowane względem średniej:

$$c_{53} = \frac{\mathrm{std}\left(SRP\right)}{\mathrm{mean}\left(SRP\right)} \tag{Z.1.69}$$

54. Średnia wartość kurtozy widmowej w grupie 3 ramek [39]:

$$c_{54} = \mathrm{mean}\left(SK\right) \tag{Z.1.70}$$

gdzie: $SK = \left[sk_1,\ \ sk_2,\ \ sk_3\right]^T$ to zbiór wartości kurtoz widmowych w trzech kolejnych ramkach tworzących grupę,

sk_i to wartość kurtozy widmowej *i*-tej ramki, tj.:

$$sk_i = \frac{\sum_{l=0}^{L-1}\left(f_l - sc_i\right)^4 \cdot S_i(l)}{ss_i^{\,4} \cdot \sum_{l=0}^{L-1} S_i(l)} \tag{Z.1.71}$$

sc_i to środek ciężkości widma *i*-tej ramki (parametr środka ciężkości widma opisany jest wzorem (4.15), przy czym tu dla porządku zmieniono oznaczenie z *SC* na *sc*),

$f_l = l \cdot \left(f_p / N\right)$ to częstotliwość *l*-tej składowej widma sygnału,

ss_i to rozproszenie widma *i*-tej ramki (parametr rozproszenia widmowego opisany jest wzorem 2.3, przy czym tu dla porządku zmieniono oznaczenie z *SS* na *ss*),

L oznacza liczbę próbek w ramce,

$S_i(l)$ to wartość *l*-tej próbki widma *i*-tej ramki.

55. Stosunek maksimum do minimum wartości kurtozy widmowej w grupie 3 ramek:

$$c_{55} = \frac{\max\left(SK\right)}{\min\left(SK\right)} \tag{Z.1.72}$$

56. Odchylenie standardowe kurtozy widmowej w grupie 3 ramek unormowane względem średniej:

$$c_{56} = \frac{\mathrm{std}\left(SK\right)}{\mathrm{mean}\left(SK\right)} \tag{Z.1.73}$$

57. Średnia wartość rozproszenia widmowego w grupie 3 ramek [5], [39], [127]:

$$c_{57} = \mathrm{mean}\left(SS\right) \tag{Z.1.74}$$

gdzie, $SS = \left[ss_1,\ \ ss_2,\ \ ss_3\right]^T$ to zbiór wartości rozproszeń widmowych w trzech kolejnych ramkach tworzących grupę (parametr rozproszenia widmowego opisany jest wzorem 2.3, przy czym tu dla porządku zmieniono oznaczenie z *SS* na *ss*).

58. Stosunek maksimum do minimum wartości rozproszenia widmowego w grupie 3 ramek:

$$c_{58} = \frac{\max\left(SS\right)}{\min\left(SS\right)} \tag{Z.1.75}$$

59. Odchylenie standardowe wartości rozproszenia widmowego w grupie 3 ramek unormowane względem średniej:

$$c_{59} = \frac{\text{std}(SS)}{\text{mean}\left(SS\right)} \tag{Z.1.76}$$

60. Średnia wartość środków ciężkości widm w grupie 3 ramek [5], [39], [127]:

$$c_{60} = \text{mean}\left(SC\right) \tag{Z.1.77}$$

gdzie, $SC = \left[sc_1,\ \ sc_2,\ \ sc_3\right]^T$ to zbiór wartości środków ciężkości widma ramki w trzech kolejnych ramkach tworzących grupę (parametr środka ciężkości widma opisany jest wzorem 4.15, przy czym tu dla porządku zmieniono oznaczenie z *SC* na *sc*).

61. Stosunek maksimum do minimum wartości środków ciężkości widm w grupie 3 ramek:

$$c_{61} = \frac{\max\left(SC\right)}{\min\left(SC\right)} \tag{Z.1.78}$$

62. Odchylenie standardowe środka ciężkości widm w grupie 3 ramek unormowane względem średniej:

$$c_{62} = \frac{\text{std}\left(SC\right)}{\text{mean}\left(SC\right)} \tag{Z.1.79}$$

63. Średnia wartość skośności widmowej w grupie 3 ramek [39]:

$$c_{63} = \text{mean}\left(SSK\right) \tag{Z.1.80}$$

gdzie: $SSK = \left[ssk_1,\ \ ssk_2,\ \ ssk_3\right]^T$ to zbiór wartości skośności widmowych w trzech kolejnych ramkach tworzących grupę,

ssk_i to wartość skośności widmowej *i*-tej ramki, tj.:

$$ssk_i = \frac{\sum_{l=0}^{L-1}\left(f_l - sc_i\right)^3 \cdot S_i(l)}{ss_i^{\,3} \cdot \sum_{l=0}^{L-1} S_i(l)} \tag{Z.1.81}$$

sc_i to środek ciężkości widma *i*-tej ramki,

ss_i to rozproszenie widma *i*-tej ramki,

L oznacza liczbę próbek w ramce,

$S_i(l)$ to wartość *l*-tej próbki widma *i*-tej ramki.

64. Stosunek maksimum do minimum wartości skośności widmowej w grupie 3 ramek:

$$c_{64} = \frac{\max(SSK)}{\min(SSK)} \tag{Z.1.82}$$

65. Odchylenie standardowe wartości skośności widmowej w grupie 3 ramek unormowane względem średniej:

$$c_{65} = \frac{\mathrm{std}(SSK)}{\mathrm{mean}(SSK)} \tag{Z.1.83}$$

66. Średnia wartość nachylenia widmowego w grupie 3 ramek [5]:

$$c_{66} = \mathrm{mean}(SSL) \tag{Z.1.84}$$

gdzie: $SSL = [ssl_1,\ ssl_2,\ ssl_3]^T$ to zbiór wartości nachyleń widmowych w trzech kolejnych ramkach tworzących grupę,

ssl_i to wartość nachylenia widmowego *i*-tej ramki, tj.:

$$ssl_i = \frac{\sum_{l=0}^{L-1}(f_l - sc_i)\left(S_i(l) - \mathrm{mean}(S_i)\right)}{\sum_{l=0}^{L-1}(f_l - sc_i)^2} \tag{Z.1.85}$$

sc_i to środek ciężkości widma *i*-tej ramki,

$\mathrm{mean}(S_i)$ to średnia wartość amplitudy widma *i*-tej ramki,

L oznacza liczbę próbek w ramce,

$S_i(l)$ to wartość *l*-tej próbki widma *i*-tej ramki.

67. Stosunek maksimum do minimum wartości nachylenia widmowego w grupie 3 ramek:

$$c_{67} = \frac{\max(SSL)}{\min(SSL)} \tag{Z.1.86}$$

68. Odchylenie standardowe wartości nachylenia widmowego w grupie 3 ramek unormowane względem średniej:

$$c_{68} = \frac{\mathrm{std}(SSL)}{\mathrm{mean}(SSL)} \tag{Z.1.87}$$

Cechy liczone dla ramek o niestandardowej długości:

Sposób wyznaczania ostatnich 3 cech, tj.: $\left[C_{69}, C_{70}, C_{71}\right]$ przebiega podobnie jak dla poprzednich cech, tj. zgodnie z rys. Z.1.1, ale cechy te są dodatkowo wstępnie uśredniane w grupach zawierających ramki o różnej liczności. Oznaczając cechy wyekstrahowane z *i*-tej ramki jako $\left[c_{69_i}, c_{70_i}, c_{71_i}\right]$, 3 ostatnie cechy określone są zgodnie z zależnością

$$\left[C_{69}, C_{70}, C_{71}\right] = \text{mean}\begin{pmatrix} c_{69_1} & c_{70_1} & c_{71_1} \\ c_{69_2} & c_{70_2} & c_{71_2} \\ \vdots & \vdots & \vdots \\ c_{69_U} & c_{70_W} & c_{71_V} \end{pmatrix} \tag{Z.1.88}$$

69. Liczba znaczących zjawisk impulsowych, tj. zjawisk spełniających warunek:

$$c_{69} = \sum_{i=1}^{N} 1 \Leftrightarrow \begin{cases} p_{h_i} > \overline{p}_h \\ p_{p_i} > \overline{p}_p \end{cases} \tag{Z.1.89}$$

gdzie: *N* to liczba maksimów w analizowanej ramce,

p_{h_i} to wartość energii *i*-tego maksimum,

p_{p_i} to względna wysokość *i*-tego maksimum w stosunku do najwyższego okalającego minimum (szczegółowy opis przedstawiono w rozdziale 4.2.6),

$\overline{p}_h$ to średnia wartość energii w ramce,

$\overline{p}_p$ to średnia względna wysokość w stosunku do minimów *w* ramce.

70. Liczba ramek dźwięcznych (w stosunku do wszystkich ramek sygnału):

$$c_{70} = \frac{K_{dź}}{K} \tag{Z.1.90}$$

gdzie: $K_{dź}$ to liczba ramek dźwięcznych (sposób analizy dźwięczności ramek dokonywany jest zgodnie ze wzorem (3.5)),

K to liczba wszystkich ramek sygnału.

71. Średni procent ramek o niskiej energii w grupie 3 ramek [32]:

$$c_{71} = \frac{k_{le}}{3} \cdot 100\% \tag{Z.1.91}$$

gdzie: k_{le} to liczba ramek o niskiej energii, tj. ramek spełniających warunek:

$$k_{le} = \sum_{v=1}^{V} 1 \Leftrightarrow e_v < \frac{1}{2} \cdot \left(\frac{1}{N} \sum_{j=0}^{N-1} x_j^{\,2} \right) \tag{Z.1.92}$$

x_j to wartość *j*-tej próbki sygnału zawartego w pojedynczej ramce,

V to liczba wszystkich analizowanych ramek,

e_v to wartość krótkoczasowej energii *v*-tej ramki (parametr ten opisany jest wzorem (2.2) – przy czym tu dla porządku zmieniono oznaczenie z E na e).

Z. 2

Dokumentacja badawcza ewolucji projektowanego systemu ASR

W pierwszych 4 rozdziałach przedstawiony został proces tworzenia systemu ASR opartego na cechach behawioralnych, którego elementy stanowią wsparcie dla istniejącego rozwiązania *ARMiA* [60]. Zawarte tam wyniki uzyskane zostały na drodze wielokrotnie powtarzanych eksperymentów, których celem było znalezienie optymalnych (w rozumieniu maksymalizacji liczby poprawnych identyfikacji tożsamości mówcy) parametrów i warunków pracy autorskiego systemu automatycznego rozpoznawania mówcy bazującego na behawioralnych deskryptorach sygnału mowy. Wraz z postępem badań, a co za tym idzie rozwojem opracowanego rozwiązania zmieniały się warunki jego działania. To naturalnie skutkowało potrzebą przeprowadzenia wielu kolejnych testów. Dotyczyły one doboru odpowiednich czasów uczenia i testowania, metryki odległości czy opisanej w rozdziale 5. metody fuzji danych.

W niniejszym załączniku opisane zostały najważniejsze z przeprowadzonych procesów optymalizacyjnych mających na celu sprawdzenie możliwości wykorzystania opracowanego rozwiązania do zadań automatycznego rozpoznawania mówcy. Ponadto autor przedstawił wyniki kilku testów, których celem było sprawdzenie w jakim obszarze autorskie, czysto behawioralne rozwiązanie ASR, ma największe możliwości zastosowania.

Dobór odpowiedniej długości ramki w pierwszym wariancie systemu ASR

Po określeniu pierwotnego zbioru 71 cech ważnym aspektem stał się dobór optymalnej długości ramki, z której liczone będą poszczególne cechy. Autor przyjął rozpiętość czasów trwania analizowanej ramki od 2 do 15 sekund z krokiem co sekundę. Sumaryczny czas każdego z wykorzystanych nagrań wynosił 60 sekund. Na tym etapie badań zastosowano już opracowany sposób przetwarzania nakładających się fragmentów sygnału (rozdział 4.1). Ważnym aspektem procesu doboru optymalnej długości ramki był wybór czasów trwania segmentów uczącego i testującego.

Autor zdecydował się na ocenę skuteczności systemu dla 7 kombinacji czasów trwania segmentów uczącego i testującego, odpowiednio: 30 s / 30 s, 40 s / 20 s, 45 s / 15 s, 48 s / 12s oraz 50 s / 10 s. W tab. Z.2.1 przedstawiono wyniki otrzymane z wykorzystaniem wszystkich 71 cech.

Tab. Z.2.1. Zestawienie sprawności działania pierwszej wersji systemu ASR w zależności od długości ramki

	Stosunek długości czasu uczenia do czasu testowania [s]				
Długość ramki [s]	**30 / 30**	**40 / 20**	**45 / 15**	**48 / 12**	**50 / 10**
2	39,38 %	38,75 %	33,75 %	28,44 %	25,31 %
3	41,25 %	38,75 %	35,31 %	30,63 %	28,44 %
4	37,19 %	35,94 %	31,41 %	29,69 %	29,06 %
5	38,13 %	35,63 %	31,72 %	28,75 %	27,19 %
6	36,56 %	37,19 %	32,97 %	28,44 %	27,19 %
7	34,69 %	34,69 %	30,94 %	28,75 %	26,88 %
8	35,31 %	34,69 %	31,88 %	29,06 %	29,69 %
9	33,13 %	34,22 %	31,56 %	26,25 %	26,56 %
10	31,25 %	34,06 %	30,63 %	27,50 %	25,63 %
11	30,63 %	34,38 %	29,38 %	28,75 %	26,25 %
12	29,69 %	31,41 %	30,94 %	27,81 %	29,06 %
13	30,94 %	30,94 %	30,63 %	30,63 %	29,69 %
14	30,31 %	30,78 %	29,06 %	28,13 %	27,97 %
15	30,00 %	31,09 %	30,00 %	28,59 %	27,81 %

Otrzymane wyniki skłoniły autora do przyjęcia długości ramki równej 3 sekundy, gdyż dla tego czasu trwania otrzymano generalnie najwyższe wyniki sprawności.

Dobór odpowiedniej długości segmentów uczącego i testującego oraz metryki odległości

Pierwszym z przeprowadzonych testów było sprawdzenie, jak zmienia się sprawność systemu wykorzystującego cały pierwotny zbiór 71 deskryptorów behawioralnych dla różnych długości segmentów treningowego i testującego. Autor przyjął, że sumaryczny czas analizowanego sygnału mowy nie powinien przekroczyć 60 sekund oraz czas trwania segmentu testującego nie powinien być większy niż czas trwania segmentu uczącego. Jednocześnie zbadano wpływ zastosowanej metryki odległości na liczbę poprawnych identyfikacji tożsamości mówcy. Wyniki badań zaprezentowano w tab. Z.2.2. Na zielono zaznaczono warianty charakteryzujące się największą skutecznością poprawnych wykryć dla każdej z metryk odległościowych.

Tab. Z.2.2. Zestawienie sprawności (wyrażone w %) działania systemu ASR w zależności od zastosowanej metryki i długości segmentów treningowego i testującego

		Zastosowana metryka						
Czas uczenia [s]	**Czas testowania [s]**	**Eukli-desowa**	**Miej-ska**	**Czeby-szewa**	**Minkow-skiego**	**Mahala-nobisa**	**Kosinu-sowa**	**Korela-cyjna**
5	**5**	7,59	9,90	3,75	7,59	4,95	7,34	6,91
10	**5**	9,73	11,52	5,20	9,73	5,80	10,49	10,75
	10	15,36	18,00	9,22	15,36	10,32	16,04	16,21
15	**5**	12,80	13,82	7,00	12,80	8,19	12,46	12,37
	10	21,08	22,53	11,77	21,08	11,69	19,97	20,22
	15	26,54	28,24	14,25	26,54	14,25	25,34	25,51
20	**5**	14,68	15,02	6,83	14,68	6,06	14,59	14,33
	10	25,60	27,22	13,23	25,60	11,18	24,57	25,60
	15	32,34	32,85	18,43	32,34	17,92	32,51	33,19
	20	35,32	37,20	21,59	35,32	22,44	35,67	36,18
25	**5**	15,96	18,26	6,83	15,96	7,25	15,78	16,64
	10	28,50	30,29	15,96	28,50	14,93	26,71	28,24
	15	34,56	36,86	21,33	34,56	19,97	34,39	35,84
	20	41,30	43,00	24,91	41,30	25,17	41,64	42,15
	25	45,39	46,33	27,73	45,39	27,39	45,14	45,82
30	**5**	16,30	18,26	8,45	16,30	6,06	14,93	15,70
	10	29,78	30,55	16,89	29,78	14,93	27,56	28,84
	15	38,14	38,31	21,93	38,14	19,62	36,95	38,05
	20	44,62	44,28	27,22	44,62	23,98	43,43	45,31
	25	46,50	48,04	30,20	46,50	28,50	45,39	46,67
	30	49,57	50,43	33,28	49,57	30,97	48,04	49,06
35	**5**	17,24	18,34	7,85	17,24	6,06	16,13	16,81
	10	31,83	33,53	16,47	31,83	14,85	31,23	31,74
	15	40,44	41,64	23,63	40,44	21,59	39,85	40,19
	20	45,48	47,95	30,29	45,48	25,94	44,62	45,48
	25	51,54	51,88	33,79	51,54	29,95	50,00	50,60
40	**5**	17,32	20,05	7,34	17,32	6,83	17,41	17,58
	10	33,19	34,64	18,94	33,19	14,68	32,68	32,85
	15	41,21	43,17	24,32	41,21	22,18	40,53	41,30
	20	49,23	49,91	29,35	49,23	25,51	48,81	49,57
45	**5**	18,26	20,31	6,57	18,26	6,23	17,06	17,83
	10	34,22	34,98	18,34	34,22	14,33	33,11	33,36
	15	41,89	43,86	24,32	41,89	23,21	42,24	43,26
50	**5**	18,52	21,16	10,15	18,52	6,14	18,52	19,20
	10	33,79	34,39	17,49	33,79	13,91	33,11	34,22
55	**5**	17,15	19,11	8,62	17,15	7,00	17,41	17,41

Celem tego testu było określenie takiej kombinacji czasów trwania poszczególnych segmentów, przy których autorski system ASR osiąga największą liczbę poprawnych identyfikacji mówcy. Z wyników przedstawionych w powyższej tabeli widać, że dla prawie wszystkich metryk odległości (poza odległością Mahalanobisa) najlepszymi długościami segmentów treningowego i testującego są odpowiednio 35 i 25 sekund. Ponadto największą sprawnością spośród wszystkich, wykazał się wariant systemu, w którym wykorzystana została metryka miejska.

Autor powtórzył test, ale w wariancie fuzji danych na poziomie decyzji, tj. poprzez opisaną w rozdziale 5.2 metodę *preselekcji*. Eksperyment dotyczył wyboru najlepszej spośród 4 metryk odległości. Dla każdej z nich zastosował najpierw algorytm genetyczny, aby wyselekcjonować deskryptory najbardziej uwydatniające dystynktywny charakter cech behawioralnych. W tab. Z.2.3 zaprezentowano rezultaty eksperymentu[1].

Tab. Z.2.3. Wartości sprawności systemu ASR opartego na cechach fizycznych z zastosowanym behawioralnym preselektorem dla sumarycznego czasu mowy 60 sek i 4 różnych metryk odległości

	Zastosowana metryka			
Liczba najbliższych sąsiadów	**Euklidesowa [%]**	**Miejska [%]**	**Korelacyjna [%]**	**Kosinusowa [%]**
25	95,48	95,82	94,71	93.77
50	96.33	96.67	96.50	95.99
75	97.01	97.35	97.35	96.93
100	97.44	98.21	97.87	97.44
125	97,61	98,46	97,87	97,70
150	97,87	98,38	98,12	97,78
175	97,87	98,29	98,04	98,04
200	97,95	98,38	98,29	98,04
225	97,95	98,55	98,29	98,29
250	98,04	98,55	98,38	98,46
275	98,04	98,55	98,38	98,38
300	98,12	98,81	98,55	98,55
325	98,21	98,81	98,55	98,63
350	98,29	98,81	98,63	98,55
375	98,21	98,81	98,63	98,55
400	98,38	98,81	98,63	98,55
425	98,46	98,72	98,72	98,55
450	98,46	98,72	98,63	98,55
475	98,46	98,72	98,63	98,55
500	98,46	98,72	98,63	98,63
525	98,55	98,72	98,63	98,63
550	98,46	98,72	98,63	98,63
575	98,46	98,72	98,63	98,63
600	98,46	98,72	98,63	98,63

Wnioski uzyskane w tym teście pokrywają się z wnioskami wynikającymi z tab. Z.2.2, tj. najlepszą możliwą do zastosowania miarą odległości jest metryka miejska.

[1] W tym eksperymencie również wykorzystano 60 sekund sygnału mowy, czyli 35 i 25 sekund odpowiednio dla procesu uczenia i testowania.

Testy systemów dla różnej jakości zapisu sygnałów głosowych

Systemy oparte na cechach fizycznych i behawioralnych testowane były również dla sygnałów dźwiękowych zapisywanych w różnej jakości. Przez jakość rozumieć należy przepływność bitową (ang. *bitrate*), która określa jak wiele bitów transmitowanych jest w jednostce czasu. Najczęściej podawaną formą tego parametru jest liczba bitów na sekundę. Zapis plików dźwiękowych o zdegradowanej jakości został zrealizowany w środowisku *Matlab* poprzez parametr *Quality*. Zgodnie z dokumentacją [155] cecha ta przyjmuję wartości w zakresie od 0 do 100, gdzie im wyższa wartość tym większa przepływność. Autorzy dodatkowo wskazują, iż ta charakterystyka dotyczy stopnia kompresji formatów MP3, Vorbis bądź OPUS, gdzie im niższa wartość *Quality*, to tym wyższy stopień kompresji. Domyślnie parametr *Quality* ma wartość 75. Na potrzeby zapisu każdego z sygnałów z różnymi przepływnościami, autor zastosował 3 inne wielkości, tj.: *50, 25* oraz *0*. W tab. Z.2.4 zaprezentowano wyniki testów dla dwóch, analizowanych w pracy systemów ASR i czterech wartości parametru *Quality*. Warty odnotowania jest fakt, że system oparty na cechach behawioralnych w momencie przeprowadzania tego eksperymentu wykorzystywał metrykę euklidesową (dobór miary odległości realizowany był w późniejszym etapie badań).

Tab. Z.2.4. Wartości sprawności systemów ASR dla różnych jakości zapisu analizowanych sygnałów dźwiękowych

	Czas uczenia / czas testowania [s]		**Czas uczenia / czas testowania [s]**	
Wartość parametru *Quality*	**35/25**		**25/5**	
	Sprawność systemu opartego na cechach behawioralnych [%]	***Sprawność systemu ARMiA [%]***	***Sprawność systemu opartego na cechach behawioralnych [%]***	***Sprawność systemu ARMiA [%]***
75	65,61	98,21	21,42	86,26
50	63,31	98,12	22,53	84,47
25	63,91	98,12	22,18	84,30
0	60,07	98,12	20,90	83,53

Powyższe rezultaty pokazują, że degradacja jakości zapisu sygnałów dźwiękowych zazwyczaj zmniejsza skuteczność identyfikacji tożsamości mówcy w systemie opartym na cechach behawioralnych. Z kolei podejście bazujące na charakterystykach fizycznych dla odpowiednio długich czasów trwania segmentów uczącego i testującego nie wykazuje tej zależności. Autor dodatkowo przeprowadził test, gdzie sprawdził działanie rozwiązań ASR przy wykorzystaniu różnych jakości sygnałów dźwiękowych osobno dla danych treningowych i testujących. Eksperyment przeprowadzono tylko dla 60 sekund sygnału mowy (podział 35 do 25 sekund). Wyniki zaprezentowane zostały w tabelach Z.2.5 oraz Z.2.6.

Analizując rezultaty widać zależność, która pokazuje, że jeśli rejestracja poszczególnych segmentów odbywała się z różną przepływnością, to system bazujący na cechach behawioralnych ma trudności z utrzymaniem wysokiego poziomu liczby poprawnych identyfikacji tożsamości mówcy. W tym przypadku jednoznacznie wykazana jest wyższość rozwiązania dr. Kamińskiego [60].

Tab. Z.2.5. Wartości sprawności systemu *ARMiA* dla różnych jakości zapisu analizowanych sygnałów dźwiękowych

	Jakość segmentów testujących			
Jakość segmentów uczących	***75***	***50***	***25***	**0**
75	**98.21 %**	98,29 %	98,12 %	97,87 %
50	98,12 %	**98,12 %**	98,21 %	97,78 %
25	97,95 %	97,95 %	**98,12 %**	97,87 %
0	97,95 %	97,87 %	97,78 %	**98,12 %**

Tab. Z.2.6. Wartości sprawności systemu opartego na cechach behawioralnych dla różnych jakości zapisu analizowanych sygnałów dźwiękowych

	Jakość segmentów testujących			
Jakość segmentów uczących	***75***	***50***	***25***	**0**
75	**65,61 %**	0,43 %	0,43 %	0,17 %
50	0,34 %	**63,31 %**	62,71 %	9,22 %
25	0,34 %	62,46 %	**63,91 %**	9,39 %
0	0,34 %	31,66 %	31,91 %	**60,07 %**

Testy systemu przy degradacji jakości sygnału głosowego poprzez kodowanie mowy

Jednym z ważnych aspektów, który należy mieć na uwadze podczas użytkowania wybranego rozwiązania ASR, jest sposób transmisji głosu poprzez medium transmisyjne. Wykorzystany kanał telekomunikacyjny, to element modyfikujący przesyłany sygnał mowy, gdyż następuje w nim kodowanie mowy. Jest to proces kompresji głosu w celu umożliwienia jego transmisji w systemie telekomunikacyjnym [10]. Celem niniejszego eksperymentu było sprawdzenie liczby poprawnych identyfikacji tożsamości mówcy w warunkach kompresji mowy. Wykorzystano następujące standardy kodowania fonicznego [3]:

- G.711 – międzynarodowy standard kodowania mowy, stosowany w systemach telefonii stacjonarnej. Realizuje proces modulacji PCM z szybkością próbkowania 8 kS/s przy rozdzielczości kwantyzacji równej 8 bit/S. Transmisja mowy zakodowanej w tym standardzie odbywa się z prędkością 64 kbit/s. Powszechnie wykorzystywane są dwie odmiany kodeka: *A-law* stosowany w Europie (także w poniższym teście) oraz *μ-law* stosowany głównie w Ameryce Północnej i Japonii.
- GSM 06.10 – inaczej nazywany GSM Full Rate, standard kodowania mowy wykorzystywany w telefonii komórkowej GSM. Charakteryzuje się przepływnością 13,2 kbit/s oraz wykorzystaniem algorytmu RPE-LTP (ang. *Regular Pulse Excitation-Long Term Prediction*) do kompresji.
- G.723.1 – kodek stosowany do konwersji sygnału PCM na potrzeby telefonii internetowej VoIP (ang. *Voice over Internet Protocol*). Kodowanie odbywa się za pomocą jednego z dwóch algorytmów: MP-MLQ (ang. *MultiPulse Maximum Likelihood Quantization*) lub ACELP (ang. *Algebraic Code-Excited Linear Prediction*). Za ich pomocą uzyskuję się strumień o przepływności odpowiednio 6,3 lub 5,3 kbit/s. Podczas badań zastosowano algorytm MP-MLQ.

- SPEEX [158] – standard kodowania typu *open-source* znajdujący swoje zastosowanie w transmisji głosu poprzez sieć Internet. Utworzony na bazie algorytmu CELP (ang. *Code Excited Linear Prediction*). Umożliwia on pracę w 3 trybach kompresji, tj.: ultra szerokopasmowa (32 kS/s), szerokopasmowa (16 kS/s) i zastosowana w niniejszym eksperymencie wąskopasmowa (8 kS/s). Występują także warianty jakości kompresji (0-10), przy czym dla potrzeb eksperymentów przyjęta została domyślna wartość jakości kompresji na poziomie 9.

W tabelach Z.2.7 oraz Z.2.8 zaprezentowano wyniki dla systemów *ARMiA* [60] oraz opartego na cechach behawioralnych. Eksperyment przeprowadzono tylko dla 60 sekund sygnału mowy (podział 35 do 25 sekund).

Tab. Z.2.7. Wartości sprawności systemu *ARMiA* dla różnych standardów kodowania

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**98,04 %**	95,05 %	91,38 %	96,33 %
SPEEX	96,33 %	**97,35 %**	93,60 %	93,86 %
G.723.1	92,41 %	94,88 %	**96,76 %**	92,24 %
GSM 06.10	95,14 %	93,00 %	92,58 %	**98,21 %**

Tab. Z.2.8. Wartości sprawności systemu opartego na cechach behawioralnych dla różnych standardów kodowania

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**66,47 %**	5,63 %	18,26 %	23,98 %
SPEEX	23,63 %	**55,55 %**	23,72 %	20,48 %
G.723.1	5,46 %	0,09 %	**54,01 %**	36,60 %
GSM 06.10	7,68 %	0,17 %	37,12 %	**58,28 %**

Analizując powyższe wyniki widać, iż powtarza się schemat występujący przy testach różnej jakości zapisu danych (za pomocą parametru *Quality* środowiska *Matlab*), tj.: system oparty na cechach behawioralnych zaczyna znacząco tracić swoje właściwości dystynktywne (zwłaszcza w przypadku, gdy każdy z segmentów kodowany jest w innym standardzie). Zależność ta również występuje w przypadku rozwiązania dr. Kamińskiego, ale w znacznie mniejszym stopniu.

Testy systemu przy wykorzystaniu zróżnicowanych torów akustycznych

Eksperyment został przeprowadzony na wstępnym etapie badań (tj.: krótko po opracowaniu autonomicznego systemu ASR bazującego na behawioralnych cechach sygnału mowy). Do jego przeprowadzenia wykorzystano multisesyjną bazę 50 głosów autorstwa mgr inż. Daniela Posiadały (opisaną dokładniej w rozdziale 6.1.4). W trakcie testów analizowano 45 z 50 dostępnych głosów dla wszystkich 10 urządzeń rejestrujących sygnał mowy. Czas trwania nagrań przeznaczonych do utworzenia modeli głosów wynosił 25 s, natomiast segmenty testowe miały długość 5 s. W tabeli Z.2.9 przedstawione zostały wyniki (wyrażone w %) testów dla systemów ASR opartych na cechach fizycznych i behawioralnych dla wszystkich kombinacji urządzeń. System oparty na cechach behawioralnych opisany został w tabeli w skrócie jako *system BEHAWIORALNY.*

Tab. Z.2.9. Wartości sprawności systemów ASR dla różnych torów uczących i testujących (45 osób)

		Tor uczący									
	Tor testujący	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Sprawność systemu *ARMiA*	**1**	**95,56**	51,11	51,11	13,33	31,11	17,78	55,56	75,56	22,22	31,11
Sprawność systemu BEHAWIORALNEGO		**26,67**	2,22	4,44	2,22	2,22	4,44	13,33	11,11	2,22	4,44
Sprawność systemu *ARMiA*	**2**	68,89	**84,44**	53,33	6,67	35,56	24,44	68,89	82,22	24,44	33,33
Sprawność systemu BEHAWIORALNEGO		4,44	**11,11**	6,67	4,44	2,22	2,22	2,22	2,22	4,44	2,22
Sprawność systemu *ARMiA*	**3**	75,56	48,89	**86,67**	13,33	31,11	17,78	55,56	73,33	31,11	20,00
Sprawność systemu BEHAWIORALNEGO		11,11	2,22	**28,89**	8,89	6,67	8,89	11,11	20,00	2,22	2,22
Sprawność systemu *ARMiA*	**4**	20	15,56	15,56	**82,22**	8,89	8,89	20,00	15,56	8,89	17,78
Sprawność systemu BEHAWIORALNEGO		2,22	2,22	17,78	**44,44**	8,89	2,22	4,44	6,67	0,00	6,67
Sprawność systemu *ARMiA*	**5**	51,11	48,89	33,33	6,67	**100,00**	53,33	57,78	64,44	11,11	40,00
Sprawność systemu BEHAWIORALNEGO		2,22	2,22	4,44	2,22	**35,56**	4,44	6,67	6,67	2,22	11,11
Sprawność systemu *ARMiA*	**6**	64,44	48,89	40,00	8,89	53,33	**97,78**	55,56	51,11	13,33	33,33
Sprawność systemu BEHAWIORALNEGO		0,00	2,22	8,89	2,22	6,67	**24,44**	0,00	2,22	0,00	4,44
Sprawność systemu *ARMiA*	**7**	86,67	71,11	60,00	6,67	33,33	28,89	**100**	77,78	20	28,89
Sprawność systemu BEHAWIORALNEGO		15,56	2,22	8,89	4,44	2,22	4,44	**35,56**	17,78	2,22	4,44
Sprawność systemu *ARMiA*	**8**	80,00	64,44	53,33	8,89	31,11	22,22	73,33	**93,33**	24,44	24,44
Sprawność systemu BEHAWIORALNEGO		6,67	2,22	6,67	2,22	4,44	6,67	11,11	**22,22**	2,22	4,44
Sprawność systemu *ARMiA*	**9**	28,89	26,67	28,89	6,67	15,56	11,11	33,33	26,67	**82,22**	11,11
Sprawność systemu BEHAWIORALNEGO		6,67	2,22	2,22	6,67	4,44	2,22	4,44	6,67	**17,78**	4,44
Sprawność systemu ARMiA	10	31,11	31,11	28,89	8,89	22,22	40,00	17,78	40,00	26,67	**97,78**
Sprawność systemu BEHAWIORALNEGO		2,22	2,22	2,22	2,22	8,89	8,89	2,22	2,22	2,22	**20,00**

Powyższe wyniki jasno pokazują przewagę systemu *ARMiA* [60] nad rozwiązaniem bazującym na cechach behawioralnych. Tak niskie wartości sprawności rozwiązania opracowanego przez autora niniejszej rozprawy mogą być skutkiem wykorzystania kombinacji długości segmentów treningowego i testującego optymalnej dla podejścia dr. Kamińskiego [60].

Podsumowanie

Wszystkie z powyższych testów prowadzone były na etapie tworzenia systemu ASR bazującego wyłącznie na cechach behawioralnych głosu. Autor skrupulatnie sprawdził potencjalne obszary zastosowań, gdzie jego algorytm mógłby dorównać bądź przewyższyć sprawnością (w rozumieniu liczby poprawnych identyfikacji tożsamości mówcy) rozwiązanie *ARMiA* [60]. Mimo wielu przeprowadzonych eksperymentów, podejście opisane w niniejszej rozprawie ustępuje w każdym z testów systemowi dr. Kamińskiego. Dlatego też zdecydowano się na wykorzystanie elementów opracowanego algorytmu jako formy wsparcia rozwiązania dr. Kamińskiego. Cały proces realizacji tego zagadnienia opisany został w rozdziale 5, a otrzymane wyniki zaprezentowane zostały w rozdziale 6 oraz załączniku nr 3.

Z. 3

Pełne wyniki badań opracowanych rozwiązań fuzji danych

W związku z bardzo dużą liczbą wyników uzyskanych w trakcie prowadzenia badań, autor postanowił zaprezentować tylko część z nich we właściwej części rozprawy, a większość z pozostałych dołączyć w formie niniejszego załącznika. Przedstawione poniżej rezultaty stanowią rozwinięcie i uzupełnienie opisanych wcześniej eksperymentów (wraz z zachowaniem początkowych warunków każdego z testów[1]). Wyniki zaprezentowano tylko w formie tabelarycznej (tab. Z.3.1), począwszy od prezentacji uzyskanych sprawności dla wszystkich długości segmentu treningowego przy stałym czasie trwania odcinka testującego. Autor posiada również rezultaty uzyskane w bazie SLR-12, jednak

[1] Każdy z eksperymentów (poza testem dla różnych urządzeń rejestrujących) przeprowadzony był dla długości segmentu testującego równej 10, 15 oraz 20 sekund. Ponadto czas przeznaczony na segment treningowy wynosił odpowiednio od 20 do 40 sekund z krokiem co 5 s.

ze względu na ich liczność, a także małą wartość ewaluacyjną, postanowił ich nie uwzględniać w niniejszej rozprawie. W poniższych zestawieniach sprawności różnych wariantów systemów ASR kolorem zielonym zaznaczono liczbę poprawnych identyfikacji tożsamości wyższą niż w systemie odniesienia, tj. w systemie *ARMiA* [60]. Z kolei kolorem pomarańczowym oznaczono wartości równe.

Testy opracowanych rozwiązań w warunkach bezszumowych

Tab. Z.3.1. Zestawienie sprawności różnych wariantów systemów ASR dla niezakłóconej bazy 1688 głosów i przy każdej kombinacji długości segmentów uczących i testujących

		Wszystkie bazy, 1688 osób				
Czas uczenia [s]	**Czas testowania [s]**	**System *ARMiA* [%]**	**System oparty na cechach behawioralnych [%]**	**Fuzja na poziomie cech [%]**	**Fuzja na poziomie cech (AG) [%]**	**Preselekcja 1000 NN [%]**
20 sekund	**10 sekund**	93,84	32,58	71,03	75,36	94,19
25 sekund		94,49	36,49	79,44	82,70	94,49
30 sekund		93,90	38,63	79,44	83,77	94,02
35 sekund		94,61	39,93	80,33	83,65	94,67
40 sekund		95,38	40,58	81,58	84,60	95,56
20 sekund	**15 sekund**	95,73	39,57	93,13	93,07	95,97
25 sekund		95,32	44,37	94,91	94,61	95,56
30 sekund		95,79	47,99	95,44	94,61	95,91
35 sekund		96,09	50,59	95,62	95,91	96,21
40 sekund		96,27	49,76	96,03	95,91	96,33
20 sekund	**20 sekund**	95,97	46,09	95,97	94,96	96,21
25 sekund		96,27	50,24	96,39	95,73	96,39
30 sekund		96,80	55,04	96,92	96,27	96,86
35 sekund		96,45	57,35	96,56	96,62	96,56
40 sekund		97,10	56,40	97,10	96,86	97,16

Powyższe, pełne wyniki pokazują, że zastosowanie cech behawioralnych jako wsparcia systemu *ARMiA* [60] na poziomie decyzji pozwala na poprawę liczby poprawnych identyfikacji mówcy.

Testy opracowanych rozwiązań w zróżnicowanych warunkach szumowych

Zakłócenie w postaci szumu białego

W tabelach Z.3.2, Z.3.3 oraz Z.3.4 przedstawiono rezultaty testów dla rozszerzonego zbioru 1688 głosów.

Analiza przedstawionych poniżej wyników wskazuje, że zastosowanie metod fuzji danych pozwala na poprawę sprawności systemu odniesienia [60] przy odpowiednio długich czasach trwania poszczególnych segmentów. Dla najniższej wartości współczynnika SNR fuzja danych na poziomie cech (po zastosowaniu algorytmu genetycznego) poprawia wyniki przy każdej kombinacji długości segmentów uczenia i testowania. Z kolei metoda preselekcji w prawie każdym przypadku sprawdza się lepiej niż rozwiązanie *ARMiA* [60].

Tab. Z.3.2. Zestawienie sprawności różnych wariantów systemów ASR przy degradacji szumem białym (1688 mówców), SNR = 20 dB

		Wszystkie bazy, 1688 osób				
Czas uczenia [s]	**Czas testowania [s]**	**System *ARMiA* [%]**	**System oparty na cechach behawioralnych [%]**	**Fuzja na poziomie cech [%]**	**Fuzja na poziomie cech (AG) [%]**	**Preselekcja 1000 NN [%]**
20 sekund	**10 sekund**	87,26	27,96	65,64	72,22	87,62
25 sekund		87,97	30,98	71,56	79,86	87,91
30 sekund		86,49	33,29	71,80	79,68	86,49
35 sekund		86,43	32,94	71,09	80,21	86,55
40 sekund		88,74	35,96	73,10	81,34	88,98
20 sekund	**15 sekund**	90,46	36,02	88,09	90,23	90,70
25 sekund		90,76	38,57	90,05	92,30	91,05
30 sekund		90,82	40,64	90,34	92,77	90,88
35 sekund		91,11	42,54	90,28	93,54	91,11
40 sekund		91,65	44,02	91,05	93,60	91,71
20 sekund	**20 sekund**	91,65	39,87	91,71	91,47	91,94
25 sekund		92,95	42,77	92,54	93,48	92,95
30 sekund		92,65	46,39	92,89	94,25	92,77
35 sekund		92,24	48,64	92,12	94,37	92,30
40 sekund		93,54	50,36	93,48	94,37	93,66

Tab. Z.3.3. Zestawienie sprawności różnych wariantów systemów ASR przy degradacji szumem białym (1688 mówców), SNR = 15 dB

		Wszystkie bazy, 1688 osób				
Czas uczenia [s]	**Czas testowania [s]**	**System *ARMiA* [%]**	**System oparty na cechach behawioralnych [%]**	**Fuzja na poziomie cech [%]**	**Fuzja na poziomie cech (AG) [%]**	**Preselekcja 1000 NN [%]**
20 sekund	**10 sekund**	80,98	24,23	63,92	72,10	81,16
25 sekund		80,81	27,43	68,25	78,38	80,81
30 sekund		79,98	29,15	67,65	78,73	80,21
35 sekund		79,80	28,14	67,18	79,50	80,04
40 sekund		80,92	31,46	67,95	80,09	81,16
20 sekund	**15 sekund**	85,72	30,81	84,54	88,15	85,96
25 sekund		86,43	34,36	85,96	89,40	86,61
30 sekund		86,32	34,60	86,02	90,52	86,49
35 sekund		85,66	37,91	85,72	91,17	85,66
40 sekund		86,02	38,57	86,02	91,94	85,96
20 sekund	**20 sekund**	88,03	35,31	87,38	89,22	88,09
25 sekund		88,09	38,86	87,91	90,88	88,27
30 sekund		88,33	39,99	88,51	91,82	88,45
35 sekund		87,74	42,71	87,74	92,48	87,62
40 sekund		89,22	44,73	89,34	92,83	89,45

Tab. Z.3.4. Zestawienie sprawności różnych wariantów systemów ASR przy degradacji szumem białym (1688 mówców), SNR = 10 dB

		Wszystkie bazy, 1688 osób				
Czas uczenia [s]	**Czas testowania [s]**	**System *ARMiA* [%]**	**System oparty na cechach behawioralnych [%]**	**Fuzja na poziomie cech [%]**	**Fuzja na poziomie cech (AG) [%]**	**Preselekcja 1000 NN [%]**
20 sekund	**10 sekund**	63,86	17,65	63,09	69,73	63,80
25 sekund		64,28	19,85	63,39	73,64	64,57
30 sekund		61,32	21,98	61,67	74,29	61,85
35 sekund		60,37	20,85	60,13	77,01	60,72
40 sekund		61,61	21,50	62,32	79,38	62,03
20 sekund	**15 sekund**	69,49	20,68	66,65	71,62	69,49
25 sekund		70,20	23,82	69,25	76,66	70,68
30 sekund		70,79	26,66	70,44	78,44	70,91
35 sekund		68,78	27,84	68,36	80,81	68,90
40 sekund		71,45	29,98	71,27	82,35	71,50
20 sekund	**20 sekund**	72,10	24,76	68,54	72,57	72,27
25 sekund		73,64	28,73	72,45	77,61	73,82
30 sekund		74,47	29,32	73,82	81,52	74,47
35 sekund		73,93	32,52	73,05	83,59	74,05
40 sekund		76,72	33,53	76,42	85,60	76,84

Zakłócenie w postaci nagrania tłumu

Tabele Z.3.5, Z.3.6 oraz Z.3.7 stanowią rezultaty testów dla zbioru 1688 głosów.

Tab. Z.3.5. Zestawienie sprawności różnych wariantów systemów ASR przy degradacji imitacją tłumu (1688 mówców), szum słyszalny

		Wszystkie bazy, 1688 osób				
Czas uczenia [s]	**Czas testowania [s]**	**System *ARMiA* [%]**	**System oparty na cechach behawioralnych [%]**	**Fuzja na poziomie cech [%]**	**Fuzja na poziomie cech (AG) [%]**	**Preselekcja 1000 NN [%]**
20 sekund	**10 sekund**	46,27	16,05	45,68	43,78	47,33
25 sekund		56,52	18,72	54,98	56,34	56,93
30 sekund		56,34	17,48	55,39	65,70	56,93
35 sekund		45,32	13,74	45,38	49,17	46,86
40 sekund		45,50	14,10	46,56	52,19	45,91
20 sekund	**15 sekund**	57,17	20,97	53,79	52,73	57,94
25 sekund		59,00	20,73	55,69	60,43	59,77
30 sekund		60,66	18,19	57,29	62,86	61,08
35 sekund		53,85	17,83	53,50	55,15	55,15
40 sekund		48,05	11,67	47,04	54,68	48,87
20 sekund	**20 sekund**	59,00	20,73	53,79	56,28	60,07
25 sekund		60,31	21,27	54,98	56,34	61,20
30 sekund		64,51	20,20	60,31	63,80	64,81
35 sekund		51,72	14,16	48,87	54,27	52,43
40 sekund		57,64	13,33	55,04	63,09	58,29

Tab. Z.3.6. Zestawienie sprawności różnych wariantów systemów ASR przy degradacji imitacją tłumu (1688 mówców), szum uciążliwy

		Wszystkie bazy, 1688 osób				
Czas uczenia [s]	**Czas testowania [s]**	**System *ARMiA* [%]**	**System oparty na cechach behawioralnych [%]**	**Fuzja na poziomie cech [%]**	**Fuzja na poziomie cech (AG) [%]**	**Preselekcja 1000 NN [%]**
20 sekund	**10 sekund**	44,85	14,81	44,08	41,53	45,56
25 sekund		55,75	17,65	53,38	54,92	56,04
30 sekund		55,45	16,41	54,68	64,10	56,64
35 sekund		44,08	13,68	43,96	47,16	45,62
40 sekund		43,78	12,80	44,43	50,47	44,14
20 sekund	**15 sekund**	55,92	19,19	52,19	51,54	56,69
25 sekund		57,76	19,79	53,67	57,94	58,06
30 sekund		59,06	17,00	55,98	61,43	59,54
35 sekund		51,90	16,88	52,31	52,96	53,02
40 sekund		46,92	11,20	46,21	52,90	47,45
20 sekund	**20 sekund**	57,88	20,02	51,90	53,73	58,59
25 sekund		59,89	18,96	54,27	54,32	60,66
30 sekund		63,33	18,72	58,41	61,97	63,68
35 sekund		49,94	13,39	47,63	52,13	50,77
40 sekund		55,51	13,03	53,02	60,78	56,22

Tab. Z.3.7. Zestawienie sprawności różnych wariantów systemów ASR przy degradacji imitacją tłumu (1688 mówców), szum przeważający

		Wszystkie bazy, 1688 osób				
Czas uczenia [s]	**Czas testowania [s]**	**System *ARMiA* [%]**	**System oparty na cechach behawioralnych [%]**	**Fuzja na poziomie cech [%]**	**Fuzja na poziomie cech (AG) [%]**	**Preselekcja 1000 NN [%]**
20 sekund	**10 sekund**	44,96	13,92	43,25	39,87	45,73
25 sekund		53,79	17,06	51,84	53,02	53,97
30 sekund		54,03	15,64	54,03	61,67	54,92
35 sekund		43,13	12,26	43,01	45,26	44,37
40 sekund		41,53	11,61	42,54	48,34	42,06
20 sekund	**15 sekund**	55,45	17,89	50,53	48,76	56,10
25 sekund		56,75	18,19	51,54	55,51	57,05
30 sekund		58,29	15,40	55,15	60,19	59,00
35 sekund		51,01	15,58	50,95	51,24	51,72
40 sekund		45,32	10,72	44,43	51,78	45,73
20 sekund	**20 sekund**	57,29	18,31	50,36	51,24	57,94
25 sekund		58,18	18,60	51,84	52,19	58,77
30 sekund		61,85	17,12	57,11	60,43	62,26
35 sekund		48,05	12,32	45,79	50,65	48,70
40 sekund		54,03	11,97	51,42	59,12	54,80

Zakłócenie w postaci nagrania ulewnego deszczu

W tabelach Z.3.8, Z.3.9 oraz Z.3.10 znajdują się wartości sprawności różnych rozwiązań systemów ASR dla zbioru 1688 głosów.

Tab. Z.3.8. Zestawienie sprawności różnych wariantów systemów ASR przy degradacji dźwiękiem ulewnego deszczu (1688 mówców), szum słyszalny

		Wszystkie bazy, 1688 osób				
Czas uczenia [s]	**Czas testowania [s]**	**System *ARMiA* [%]**	**System oparty na cechach behawioralnych [%]**	**Fuzja na poziomie cech [%]**	**Fuzja na poziomie cech (AG) [%]**	**Preselekcja 1000 NN [%]**
20 sekund	**10 sekund**	51,78	28,67	51,90	64,63	52,37
25 sekund		50,47	28,20	50,89	70,14	50,53
30 sekund		49,11	31,10	49,88	72,63	49,41
35 sekund		49,23	33,23	50,83	74,17	49,29
40 sekund		48,28	31,10	49,70	77,13	48,76
20 sekund	**15 sekund**	56,87	34,54	54,68	62,97	57,41
25 sekund		56,58	36,85	55,33	71,33	56,75
30 sekund		56,04	37,38	55,39	75,06	56,34
35 sekund		56,99	39,81	56,58	76,60	57,29
40 sekund		57,76	41,29	57,76	79,32	57,94
20 sekund	**20 sekund**	59,77	38,86	55,75	63,45	59,89
25 sekund		60,19	41,23	57,29	70,68	60,55
30 sekund		60,07	39,87	58,12	74,70	60,31
35 sekund		61,08	45,62	60,01	77,55	61,20
40 sekund		61,91	46,09	61,20	80,75	61,91

Tab. Z.3.9. Zestawienie sprawności różnych wariantów systemów ASR przy degradacji dźwiękiem ulewnego deszczu (1688 mówców), szum uciążliwy

		Wszystkie bazy, 1688 osób				
Czas uczenia [s]	**Czas testowania [s]**	**System *ARMiA* [%]**	**System oparty na cechach behawioralnych [%]**	**Fuzja na poziomie cech [%]**	**Fuzja na poziomie cech (AG) [%]**	**Preselekcja 1000 NN [%]**
20 sekund	**10 sekund**	50,83	28,91	50,06	62,80	51,36
25 sekund		49,70	28,02	49,47	68,84	49,82
30 sekund		47,81	29,62	48,10	71,15	48,22
35 sekund		47,81	33,29	49,41	73,22	48,05
40 sekund		47,04	31,69	48,40	76,18	47,39
20 sekund	**15 sekund**	55,86	34,12	54,50	60,90	56,10
25 sekund		55,57	35,49	53,85	68,90	55,81
30 sekund		54,74	37,50	54,50	73,46	55,15
35 sekund		55,69	39,10	54,92	75,41	55,92
40 sekund		56,64	40,23	56,46	78,02	56,75
20 sekund	**20 sekund**	58,77	38,21	54,03	61,61	58,83
25 sekund		59,36	41,47	56,28	68,54	59,66
30 sekund		58,53	40,05	57,05	73,22	58,83
35 sekund		59,89	44,02	58,83	75,83	60,13
40 sekund		60,84	44,91	59,77	79,32	60,90

Tab. Z.3.10. Zestawienie sprawności różnych wariantów systemów ASR przy degradacji dźwiękiem ulewnego deszczu (1688 mówców), szum przeważający

		Wszystkie bazy, 1688 osób				
Czas uczenia [s]	**Czas testowania [s]**	**System *ARMiA* [%]**	**System oparty na cechach behawioralnych [%]**	**Fuzja na poziomie cech [%]**	**Fuzja na poziomie cech (AG) [%]**	**Preselekcja 1000 NN [%]**
20 sekund	**10 sekund**	50,18	28,55	49,35	61,49	50,83
25 sekund		48,46	28,38	48,10	66,88	48,64
30 sekund		46,92	29,03	46,92	69,61	47,33
35 sekund		47,33	32,58	48,40	72,16	47,45
40 sekund		46,33	31,58	46,98	75,00	46,62
20 sekund	**15 sekund**	54,80	33,47	52,67	59,60	55,33
25 sekund		54,50	35,07	53,08	67,30	54,62
30 sekund		53,97	36,97	52,73	72,22	54,15
35 sekund		54,15	38,39	53,85	73,99	54,44
40 sekund		55,09	39,57	55,45	76,66	55,15
20 sekund	**20 sekund**	57,58	37,62	52,67	59,95	57,82
25 sekund		57,88	41,41	55,57	67,18	58,00
30 sekund		56,69	39,22	55,69	71,92	57,05
35 sekund		58,53	44,08	57,82	74,23	58,77
40 sekund		59,12	44,85	58,59	78,50	59,30

Uzupełnione o wszystkie kombinacje czasów uczenia i testowania wyniki potwierdzają wniosek wysnuty we właściwej części rozprawy, tj.: *niezależnie od zastosowanej metody integracji danych każda z nich pozwala na zwiększenie liczby poprawnych identyfikacji tożsamości mówcy względem systemu odniesienia* [60].

Testy opracowanych rozwiązań w warunkach zróżnicowanej jakości nagrań

Znajdujące się poniżej tabele stanowią uzupełnienie wyników zawartych w podrozdziale 6.5. Dokładniej przedstawiono tu sprawności testowanych wcześniej rozwiązań dla rozszerzonego zbioru 1688 głosów. Tabele

- od Z.3.11 do Z.3.35 zawierają wyniki dla długości segmentu testującego równej 10 sekund.
- od Z.3.36 do Z.3.60 zawierają wyniki dla długości segmentu testującego równej 15 sekund.
- od Z.3.61 do Z.3.85 zawierają wyniki dla długości segmentu testującego równej 20 sekund.

Dodatkowo w poniższych opisach zastosowano skrótowy zapis w formacie *X/Y*, gdzie X odnosi się do czasu uczenia, a *Y* do czasu testowania.

Tab. Z.3.11. Wartości sprawności systemu *ARMiA* dla różnych standardów kodowania (20s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**93,54 %**	89,10 %	81,75 %	86,32 %
SPEEX	88,63 %	**91,41 %**	84,36 %	84,00 %
G.723.1	81,75 %	84,83 %	**88,92 %**	79,09 %
GSM 06.10	86,37 %	81,87 %	79,56 %	**91,82 %**

Tab. Z.3.12. Wartości sprawności systemu opartego na cechach behawioralnych dla różnych standardów kodowania (20s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**33,95 %**	13,63 %	18,66 %	20,68 %
SPEEX	14,40 %	**24,53 %**	13,98 %	14,63 %
G.723.1	15,94 %	13,63 %	**23,82 %**	18,90 %
GSM 06.10	19,25 %	12,56 %	19,67 %	**25,83 %**

Tab. Z.3.13. Wartości sprawności wariantu fuzji danych na poziomie cech dla różnych standardów kodowania (20s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**70,50 %**	66,47 %	61,26 %	64,34 %
SPEEX	64,57 %	**66,82 %**	61,14 %	60,90 %
G.723.1	57,94 %	59,60 %	**62,97 %**	55,81 %
GSM 06.10	62,20 %	59,12 %	57,46 %	**66,35 %**

Tab. Z.3.14. Wartości sprawności wariantu fuzji danych na poziomie cech po zastosowaniu algorytmu genetycznego dla różnych standardów kodowania (20s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**74,64 %**	70,56 %	64,93 %	65,23 %
SPEEX	66,65 %	**69,43 %**	63,74 %	60,13 %
G.723.1	58,41 %	61,08 %	**64,51 %**	53,79 %
GSM 06.10	63,03 %	61,73 %	59,66 %	**68,90 %**

Tab. Z.3.15. Wartości sprawności wariantu fuzji danych na poziomie decyzji dla różnych standardów kodowania (20s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**93,84 %**	89,57 %	82,64 %	86,85 %
SPEEX	88,09 %	**91,65 %**	83,95 %	83,41 %
G.723.1	82,29 %	85,49 %	**89,22 %**	79,68 %
GSM 06.10	86,67 %	82,41 %	80,15 %	**91,88 %**

Tab. Z.3.16. Wartości sprawności systemu *ARMiA* dla różnych standardów kodowania (25s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**93,72 %**	90,40 %	83,23 %	87,80 %
SPEEX	90,70 %	**92,65 %**	85,90 %	86,37 %
G.723.1	84,06 %	86,85 %	**90,17 %**	82,88 %
GSM 06.10	86,02 %	84,12 %	79,98 %	**91,88 %**

Tab. Z.3.17. Wartości sprawności systemu opartego na cechach behawioralnych dla różnych standardów kodowania (25s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**41,47 %**	17,06 %	22,04 %	23,82 %
SPEEX	18,25 %	**32,11 %**	18,19 %	17,42 %
G.723.1	20,26 %	19,19 %	**29,15 %**	24,82 %
GSM 06.10	25,06 %	17,71 %	23,46 %	**32,58 %**

Tab. Z.3.18. Wartości sprawności wariantu fuzji danych na poziomie cech dla różnych standardów kodowania (25s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**78,32 %**	74,64 %	67,30 %	73,16 %
SPEEX	72,51 %	**75,12 %**	67,48 %	69,85 %
G.723.1	65,82 %	68,13 %	**70,97 %**	66,00 %
GSM 06.10	68,42 %	67,06 %	63,86 %	**74,64 %**

Tab. Z.3.19. Wartości sprawności wariantu fuzji danych na poziomie cech po zastosowaniu algorytmu genetycznego dla różnych standardów kodowania (25s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**82,64 %**	79,03 %	75,00 %	75,65 %
SPEEX	74,64 %	**78,85 %**	73,34 %	72,10 %
G.723.1	65,76 %	70,08 %	**75,47 %**	65,05 %
GSM 06.10	68,54 %	67,95 %	66,59 %	**78,97 %**

Tab. Z.3.20. Wartości sprawności wariantu fuzji danych na poziomie decyzji dla różnych standardów kodowania (25s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**95,73 %**	92,54 %	86,20 %	89,34 %
SPEEX	90,82 %	**94,31 %**	88,74 %	86,55 %
G.723.1	84,12 %	87,26 %	**91,88 %**	83,12 %
GSM 06.10	89,28 %	86,85 %	83,53 %	**93,60 %**

Tab. Z.3.21. Wartości sprawności systemu *ARMiA* dla różnych standardów kodowania (30s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**93,48 %**	90,05 %	82,41 %	87,44 %
SPEEX	90,46 %	**91,59 %**	85,84 %	86,61 %
G.723.1	83,77 %	85,96 %	**89,34 %**	81,58 %
GSM 06.10	87,09 %	83,83 %	79,62 %	**92,48 %**

Tab. Z.3.22. Wartości sprawności systemu opartego na cechach behawioralnych dla różnych standardów kodowania (30s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**46,39 %**	19,96 %	25,47 %	29,38 %
SPEEX	20,73 %	**35,66 %**	21,50 %	18,48 %
G.723.1	23,46 %	20,91 %	**33,89 %**	27,90 %
GSM 06.10	28,02 %	18,90 %	27,07 %	**38,63 %**

Tab. Z.3.23. Wartości sprawności wariantu fuzji danych na poziomie cech dla różnych standardów kodowania (30s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**78,67 %**	74,88 %	67,77 %	73,64 %
SPEEX	72,87 %	**74,59 %**	67,89 %	70,68 %
G.723.1	65,05 %	67,77 %	**71,03 %**	65,94 %
GSM 06.10	69,37 %	67,12 %	63,51 %	**75,59 %**

Tab. Z.3.24. Wartości sprawności wariantu fuzji danych na poziomie cech po zastosowaniu algorytmu genetycznego dla różnych standardów kodowania (30s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**83,23 %**	78,67 %	74,11 %	77,19 %
SPEEX	76,48 %	**79,86 %**	74,17 %	72,33 %
G.723.1	63,74 %	69,61 %	**75,36 %**	64,16 %
GSM 06.10	69,37 %	69,14 %	67,00 %	**79,27 %**

Tab. Z.3.25. Wartości sprawności wariantu fuzji danych na poziomie decyzji dla różnych standardów kodowania (30s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,15 %**	93,19 %	87,09 %	89,75 %
SPEEX	91,53 %	**94,96 %**	89,57 %	86,49 %
G.723.1	84,60 %	89,16 %	**93,13 %**	83,59 %
GSM 06.10	90,28 %	86,85 %	85,13 %	**94,79 %**

Tab. Z.3.26. Wartości sprawności systemu *ARMiA* dla różnych standardów kodowania (35s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**94,25 %**	89,87 %	83,18 %	88,45 %
SPEEX	91,65 %	**93,07 %**	87,14 %	87,44 %
G.723.1	84,95 %	87,14 %	**91,00 %**	84,00 %
GSM 06.10	87,86 %	84,66 %	80,33 %	**92,89 %**

Tab. Z.3.27. Wartości sprawności systemu opartego na cechach behawioralnych dla różnych standardów kodowania (35s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**36,49 %**	15,28 %	20,26 %	22,87 %
SPEEX	14,69 %	**26,84 %**	16,65 %	14,22 %
G.723.1	18,90 %	16,47 %	**26,30 %**	21,68 %
GSM 06.10	22,87 %	14,10 %	22,51 %	**29,62 %**

Tab. Z.3.28. Wartości sprawności wariantu fuzji danych na poziomie cech dla różnych standardów kodowania (35s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**79,38 %**	74,88 %	67,54 %	74,94 %
SPEEX	73,05 %	**76,18 %**	67,89 %	71,39 %
G.723.1	67,24 %	68,78 %	**71,68 %**	67,83 %
GSM 06.10	69,14 %	65,94 %	62,80 %	**76,48 %**

Tab. Z.3.29. Wartości sprawności wariantu fuzji danych na poziomie cech po zastosowaniu algorytmu genetycznego dla różnych standardów kodowania (35s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**83,95 %**	79,80 %	75,59 %	77,84 %
SPEEX	76,36 %	**79,50 %**	74,23 %	73,76 %
G.723.1	67,30 %	71,68 %	**75,24 %**	67,42 %
GSM 06.10	70,91 %	70,91 %	67,77 %	**80,51 %**

Tab. Z.3.30. Wartości sprawności wariantu fuzji danych na poziomie decyzji dla różnych standardów kodowania (35s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**93,84 %**	90,40 %	83,71 %	88,09 %
SPEEX	89,93 %	**93,01 %**	85,43 %	85,55 %
G.723.1	84,72 %	87,03 %	**90,58 %**	83,83 %
GSM 06.10	86,26 %	84,48 %	80,51 %	**92,18 %**

Tab. Z.3.31. Wartości sprawności systemu *ARMiA* dla różnych standardów kodowania (40s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**94,37 %**	91,53 %	85,60 %	89,34 %
SPEEX	92,95 %	**93,66 %**	88,74 %	89,16 %
G.723.1	86,79 %	88,98 %	**92,59 %**	84,77 %
GSM 06.10	89,51 %	86,43 %	81,81 %	**94,14 %**

Tab. Z.3.32. Wartości sprawności systemu opartego na cechach behawioralnych dla różnych standardów kodowania (40s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**44,79 %**	18,60 %	23,99 %	28,79 %
SPEEX	18,96 %	**33,65 %**	21,09 %	17,65 %
G.723.1	23,40 %	19,55 %	**32,82 %**	26,30 %
GSM 06.10	27,43 %	18,25 %	28,85 %	**36,20 %**

Tab. Z.3.33. Wartości sprawności wariantu fuzji danych na poziomie cech dla różnych standardów kodowania (40s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**80,27 %**	76,30 %	69,79 %	76,54 %
SPEEX	75,47 %	**76,72 %**	70,08 %	73,46 %
G.723.1	69,37 %	70,79 %	**73,58 %**	69,31 %
GSM 06.10	71,27 %	68,90 %	64,87 %	**77,61 %**

Tab. Z.3.34. Wartości sprawności wariantu fuzji danych na poziomie cech po zastosowaniu algorytmu genetycznego dla różnych standardów kodowania (40s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**84,83 %**	80,69 %	76,78 %	78,26 %
SPEEX	77,67 %	**80,09 %**	75,71 %	74,29 %
G.723.1	68,31 %	73,70 %	**77,07 %**	67,83 %
GSM 06.10	70,97 %	72,33 %	69,61 %	**81,58 %**

Tab. Z.3.35. Wartości sprawności wariantu fuzji danych na poziomie decyzji dla różnych standardów kodowania (40s/10s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**95,26 %**	92,89 %	86,55 %	89,93 %
SPEEX	91,23 %	**94,55 %**	89,04 %	87,44 %
G.723.1	85,96 %	89,57 %	**92,48 %**	85,37 %
GSM 06.10	88,74 %	87,91 %	84,06 %	**94,31 %**

Tab. Z.3.36. Wartości sprawności systemu *ARMiA* dla różnych standardów kodowania (20s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**95,44 %**	92,24 %	85,37 %	88,74 %
SPEEX	91,77 %	**94,14 %**	89,34 %	87,56 %
G.723.1	83,53 %	86,73 %	**91,59 %**	82,70 %
GSM 06.10	89,10 %	86,55 %	82,88 %	**93,60 %**

Tab. Z.3.37. Wartości sprawności systemu opartego na cechach behawioralnych dla różnych standardów kodowania (20s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**50,59 %**	22,22 %	28,08 %	32,11 %
SPEEX	22,51 %	**39,40 %**	23,70 %	21,33 %
G.723.1	27,61 %	23,93 %	**37,86 %**	30,39 %
GSM 06.10	31,34 %	21,27 %	31,69 %	**41,11 %**

Tab. Z.3.38. Wartości sprawności wariantu fuzji danych na poziomie cech dla różnych standardów kodowania (20s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**93,13 %**	89,57 %	82,88 %	86,79 %
SPEEX	88,03 %	**90,76 %**	84,83 %	83,35 %
G.723.1	79,86 %	83,00 %	**87,09 %**	79,32 %
GSM 06.10	84,24 %	81,99 %	79,44 %	**89,45 %**

Tab. Z.3.39. Wartości sprawności wariantu fuzji danych na poziomie cech po zastosowaniu algorytmu genetycznego dla różnych standardów kodowania (20s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**92,71 %**	89,57 %	84,89 %	85,07 %
SPEEX	87,62 %	**90,88 %**	84,95 %	81,69 %
G.723.1	78,08 %	82,76 %	**86,55 %**	74,82 %
GSM 06.10	83,00 %	81,46 %	79,50 %	**89,45 %**

Tab. Z.3.40. Wartości sprawności wariantu fuzji danych na poziomie decyzji dla różnych standardów kodowania (20s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,45 %**	93,84 %	88,98 %	91,11 %
SPEEX	92,36 %	**95,97 %**	90,94 %	88,33 %
G.723.1	86,55 %	90,94 %	**94,67 %**	86,37 %
GSM 06.10	91,17 %	88,92 %	86,32 %	**95,85 %**

Tab. Z.3.41. Wartości sprawności systemu *ARMiA* dla różnych standardów kodowania (25s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**95,08 %**	92,71 %	85,90 %	89,51 %
SPEEX	92,00 %	**94,25 %**	89,22 %	88,33 %
G.723.1	85,13 %	88,92 %	**92,42 %**	84,72 %
GSM 06.10	88,39 %	87,26 %	83,53 %	**94,37 %**

Tab. Z.3.42. Wartości sprawności systemu opartego na cechach behawioralnych dla różnych standardów kodowania (25s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**38,09 %**	19,49 %	22,45 %	24,70 %
SPEEX	17,30 %	**30,15 %**	18,96 %	16,17 %
G.723.1	19,61 %	18,36 %	**27,67 %**	21,74 %
GSM 06.10	22,75 %	16,41 %	22,87 %	**31,58 %**

Tab. Z.3.43. Wartości sprawności wariantu fuzji danych na poziomie cech dla różnych standardów kodowania (25s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**95,02 %**	91,82 %	85,78 %	88,98 %
SPEEX	90,40 %	**93,66 %**	87,86 %	86,97 %
G.723.1	83,47 %	87,97 %	**90,76 %**	84,00 %
GSM 06.10	87,97 %	85,19 %	82,11 %	**93,19 %**

Tab. Z.3.44. Wartości sprawności wariantu fuzji danych na poziomie cech po zastosowaniu algorytmu genetycznego dla różnych standardów kodowania (25s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**94,43 %**	91,59 %	88,57 %	88,21 %
SPEEX	90,34 %	**93,13 %**	88,86 %	85,66 %
G.723.1	81,58 %	86,43 %	**91,17 %**	80,39 %
GSM 06.10	85,07 %	84,48 %	83,77 %	**93,54 %**

Tab. Z.3.45. Wartości sprawności wariantu fuzji danych na poziomie decyzji dla różnych standardów kodowania (25s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**93,54 %**	90,34 %	82,76 %	87,91 %
SPEEX	89,75 %	**91,71 %**	85,25 %	85,90 %
G.723.1	84,24 %	86,43 %	**89,51 %**	82,35 %
GSM 06.10	87,32 %	84,36 %	79,86 %	**92,48 %**

Tab. Z.3.46. Wartości sprawności systemu *ARMiA* dla różnych standardów kodowania (30s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**95,56 %**	92,42 %	87,14 %	90,11 %
SPEEX	92,65 %	**94,61 %**	90,28 %	89,69 %
G.723.1	86,55 %	89,81 %	**93,13 %**	86,08 %
GSM 06.10	89,99 %	87,20 %	84,42 %	**94,85 %**

Tab. Z.3.47. Wartości sprawności systemu opartego na cechach behawioralnych dla różnych standardów kodowania (30s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**48,64 %**	22,63 %	27,31 %	30,57 %
SPEEX	20,79 %	**36,91 %**	24,29 %	21,33 %
G.723.1	24,17 %	23,93 %	**36,85 %**	29,03 %
GSM 06.10	29,38 %	20,62 %	31,22 %	**39,99 %**

Tab. Z.3.48. Wartości sprawności wariantu fuzji danych na poziomie cech dla różnych standardów kodowania (30s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**95,26 %**	91,59 %	87,03 %	90,23 %
SPEEX	91,94 %	**94,31 %**	89,16 %	88,57 %
G.723.1	85,78 %	88,68 %	**91,41 %**	85,78 %
GSM 06.10	88,51 %	86,02 %	83,29 %	**94,43 %**

Tab. Z.3.49. Wartości sprawności wariantu fuzji danych na poziomie cech po zastosowaniu algorytmu genetycznego dla różnych standardów kodowania (30s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**94,96 %**	91,82 %	89,34 %	88,45 %
SPEEX	91,00 %	**92,83 %**	90,05 %	86,43 %
G.723.1	82,88 %	87,32 %	**91,47 %**	81,40 %
GSM 06.10	85,78 %	85,31 %	84,30 %	**93,19 %**

Tab. Z.3.50. Wartości sprawności wariantu fuzji danych na poziomie decyzji dla różnych standardów kodowania (30s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**95,62 %**	92,65 %	87,32 %	90,82 %
SPEEX	91,88 %	**94,85 %**	89,63 %	88,68 %
G.723.1	87,20 %	90,23 %	**93,13 %**	86,61 %
GSM 06.10	90,28 %	87,68 %	84,77 %	**95,08 %**

Tab. Z.3.51. Wartości sprawności systemu *ARMiA* dla różnych standardów kodowania (35s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**95,62 %**	92,83 %	88,21 %	90,52 %
SPEEX	93,60 %	**95,44 %**	91,00 %	89,81 %
G.723.1	87,50 %	90,46 %	**93,84 %**	86,37 %
GSM 06.10	91,00 %	87,86 %	86,08 %	**95,62 %**

Tab. Z.3.52. Wartości sprawności systemu opartego na cechach behawioralnych dla różnych standardów kodowania (35s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**56,16 %**	26,01 %	31,64 %	35,55 %
SPEEX	24,35 %	**42,59 %**	27,73 %	22,51 %
G.723.1	27,61 %	26,66 %	**41,82 %**	33,18 %
GSM 06.10	34,12 %	23,46 %	35,13 %	**46,33 %**

Tab. Z.3.53. Wartości sprawności wariantu fuzji danych na poziomie cech dla różnych standardów kodowania (35s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**95,26 %**	92,24 %	87,56 %	90,76 %
SPEEX	92,48 %	**94,67 %**	89,22 %	88,92 %
G.723.1	87,56 %	89,81 %	**92,77 %**	86,43 %
GSM 06.10	89,93 %	86,61 %	84,89 %	**94,43 %**

Tab. Z.3.54. Wartości sprawności wariantu fuzji danych na poziomie cech po zastosowaniu algorytmu genetycznego dla różnych standardów kodowania (35s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**95,62 %**	93,07 %	90,40 %	89,10 %
SPEEX	91,59 %	**94,08 %**	90,70 %	86,49 %
G.723.1	84,00 %	88,51 %	**92,54 %**	81,93 %
GSM 06.10	86,61 %	86,79 %	85,13 %	**94,85 %**

Tab. Z.3.55. Wartości sprawności wariantu fuzji danych na poziomie decyzji dla różnych standardów kodowania (35s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,62 %**	93,90 %	90,05 %	92,30 %
SPEEX	93,42 %	**95,68 %**	92,00 %	89,69 %
G.723.1	89,04 %	91,88 %	**94,67 %**	87,03 %
GSM 06.10	92,18 %	89,04 %	87,26 %	**96,09 %**

Tab. Z.3.56. Wartości sprawności systemu *ARMiA* dla różnych standardów kodowania (40s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,09 %**	93,84 %	88,86 %	91,59 %
SPEEX	94,08 %	**95,62 %**	91,59 %	90,70 %
G.723.1	89,51 %	92,30 %	**94,43 %**	88,03 %
GSM 06.10	91,94 %	88,86 %	86,14 %	**96,03 %**

Tab. Z.3.57. Wartości sprawności systemu opartego na cechach behawioralnych dla różnych standardów kodowania (40s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**39,63 %**	18,90 %	23,10 %	26,24 %
SPEEX	17,65 %	**30,33 %**	18,13 %	17,12 %
G.723.1	19,61 %	20,08 %	**28,44 %**	21,80 %
GSM 06.10	23,64 %	17,18 %	25,36 %	**32,64 %**

Tab. Z.3.58. Wartości sprawności wariantu fuzji danych na poziomie cech dla różnych standardów kodowania (40s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**95,56 %**	93,01 %	87,91 %	91,47 %
SPEEX	92,89 %	**94,73 %**	89,45 %	89,93 %
G.723.1	89,04 %	91,05 %	**93,01 %**	87,62 %
GSM 06.10	90,58 %	87,97 %	84,72 %	**95,32 %**

Tab. Z.3.59. Wartości sprawności wariantu fuzji danych na poziomie cech po zastosowaniu algorytmu genetycznego dla różnych standardów kodowania (40s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**95,97 %**	93,48 %	91,23 %	89,51 %
SPEEX	92,71 %	**94,37 %**	91,41 %	86,67 %
G.723.1	86,26 %	89,75 %	**93,25 %**	84,18 %
GSM 06.10	88,39 %	88,39 %	86,20 %	**95,14 %**

Tab. Z.3.60. Wartości sprawności wariantu fuzji danych na poziomie decyzji dla różnych standardów kodowania (40s/15s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**94,31 %**	90,11 %	83,65 %	88,92 %
SPEEX	90,82 %	**93,13 %**	86,79 %	86,85 %
G.723.1	85,90 %	87,56 %	**91,17 %**	84,95 %
GSM 06.10	88,51 %	85,31 %	80,86 %	**93,07 %**

Tab. Z.3.61. Wartości sprawności systemu *ARMiA* dla różnych standardów kodowania (20s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**95,91 %**	92,65 %	86,43 %	89,45 %
SPEEX	92,24 %	**94,91 %**	89,87 %	87,20 %
G.723.1	83,59 %	88,51 %	**92,89 %**	83,00 %
GSM 06.10	90,11 %	86,67 %	84,48 %	**94,73 %**

Tab. Z.3.62. Wartości sprawności systemu opartego na cechach behawioralnych dla różnych standardów kodowania (20s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**51,60 %**	24,47 %	29,21 %	33,53 %
SPEEX	23,70 %	**38,45 %**	24,59 %	22,04 %
G.723.1	25,06 %	23,87 %	**37,74 %**	29,86 %
GSM 06.10	30,63 %	21,68 %	31,40 %	**43,25 %**

Tab. Z.3.63. Wartości sprawności wariantu fuzji danych na poziomie cech dla różnych standardów kodowania (20s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**95,91 %**	92,54 %	86,37 %	89,57 %
SPEEX	91,05 %	**94,14 %**	88,45 %	86,43 %
G.723.1	83,59 %	87,68 %	**91,88 %**	82,94 %
GSM 06.10	88,45 %	85,43 %	83,71 %	**94,08 %**

Tab. Z.3.64. Wartości sprawności wariantu fuzji danych na poziomie cech po zastosowaniu algorytmu genetycznego dla różnych standardów kodowania (20s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**95,14 %**	91,88 %	87,50 %	87,38 %
SPEEX	89,75 %	**93,48 %**	88,09 %	84,36 %
G.723.1	81,28 %	86,02 %	**91,17 %**	79,09 %
GSM 06.10	85,78 %	84,89 %	83,41 %	**93,66 %**

Tab. Z.3.65. Wartości sprawności wariantu fuzji danych na poziomie decyzji dla różnych standardów kodowania (20s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**95,79 %**	93,01 %	88,33 %	91,35 %
SPEEX	92,89 %	**95,50 %**	90,34 %	88,92 %
G.723.1	88,27 %	90,94 %	**94,08 %**	87,26 %
GSM 06.10	91,41 %	88,21 %	86,49 %	**95,73 %**

Tab. Z.3.66. Wartości sprawności systemu *ARMiA* dla różnych standardów kodowania (25s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,39 %**	93,54 %	88,51 %	90,94 %
SPEEX	93,31 %	**95,91 %**	91,23 %	89,34 %
G.723.1	86,14 %	90,58 %	**94,49 %**	85,78 %
GSM 06.10	90,76 %	88,51 %	85,72 %	**95,91 %**

Tab. Z.3.67. Wartości sprawności systemu opartego na cechach behawioralnych dla różnych standardów kodowania (25s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**57,11 %**	27,49 %	32,82 %	38,27 %
SPEEX	26,18 %	**43,90 %**	27,49 %	24,35 %
G.723.1	29,56 %	27,43 %	**42,30 %**	34,30 %
GSM 06.10	35,72 %	23,82 %	36,37 %	**48,40 %**

Tab. Z.3.68. Wartości sprawności wariantu fuzji danych na poziomie cech dla różnych standardów kodowania (25s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,39 %**	93,42 %	88,86 %	90,82 %
SPEEX	92,42 %	**95,44 %**	90,88 %	88,57 %
G.723.1	85,96 %	89,69 %	**93,96 %**	85,66 %
GSM 06.10	90,76 %	88,27 %	85,37 %	**95,79 %**

Tab. Z.3.69. Wartości sprawności wariantu fuzji danych na poziomie cech po zastosowaniu algorytmu genetycznego dla różnych standardów kodowania (25s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**95,62 %**	93,25 %	89,81 %	89,16 %
SPEEX	91,65 %	**94,43 %**	90,70 %	86,37 %
G.723.1	83,35 %	88,51 %	**92,65 %**	81,93 %
GSM 06.10	87,80 %	86,79 %	85,49 %	**94,31 %**

Tab. Z.3.70. Wartości sprawności wariantu fuzji danych na poziomie decyzji dla różnych standardów kodowania (25s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,21 %**	94,02 %	89,40 %	92,59 %
SPEEX	93,25 %	**95,91 %**	91,71 %	89,87 %
G.723.1	89,63 %	92,06 %	**94,85 %**	88,15 %
GSM 06.10	91,94 %	89,81 %	88,68 %	**96,39 %**

Tab. Z.3.71. Wartości sprawności systemu *ARMiA* dla różnych standardów kodowania (30s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,56 %**	93,72 %	89,87 %	91,94 %
SPEEX	94,14 %	**95,62 %**	92,71 %	90,40 %
G.723.1	88,39 %	91,53 %	**94,79 %**	86,49 %
GSM 06.10	92,00 %	88,68 %	86,97 %	**96,03 %**

Tab. Z.3.72. Wartości sprawności systemu opartego na cechach behawioralnych dla różnych standardów kodowania (30s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**41,41 %**	20,50 %	24,47 %	27,61 %
SPEEX	16,88 %	**31,46 %**	18,96 %	17,36 %
G.723.1	20,97 %	20,44 %	**31,46 %**	25,24 %
GSM 06.10	25,83 %	18,48 %	26,36 %	**34,18 %**

Tab. Z.3.73. Wartości sprawności wariantu fuzji danych na poziomie cech dla różnych standardów kodowania (30s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,50 %**	93,54 %	89,51 %	91,23 %
SPEEX	93,19 %	**95,56 %**	91,71 %	89,81 %
G.723.1	88,80 %	91,17 %	**94,61 %**	87,56 %
GSM 06.10	91,59 %	88,92 %	87,09 %	**95,85 %**

Tab. Z.3.74. Wartości sprawności wariantu fuzji danych na poziomie cech po zastosowaniu algorytmu genetycznego dla różnych standardów kodowania (30s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,15 %**	93,78 %	90,76 %	89,63 %
SPEEX	92,24 %	**95,20 %**	91,29 %	87,20 %
G.723.1	84,95 %	89,81 %	**93,60 %**	82,58 %
GSM 06.10	88,33 %	87,50 %	87,09 %	**95,62 %**

Tab. Z.3.75. Wartości sprawności wariantu fuzji danych na poziomie decyzji dla różnych standardów kodowania (30s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**94,55 %**	91,77 %	86,14 %	89,87 %
SPEEX	92,12 %	**93,90 %**	88,51 %	88,33 %
G.723.1	87,44 %	89,28 %	**92,83 %**	85,60 %
GSM 06.10	89,69 %	86,85 %	82,29 %	**94,19 %**

Tab. Z.3.76. Wartości sprawności systemu *ARMiA* dla różnych standardów kodowania (35s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,21 %**	93,90 %	89,34 %	92,18 %
SPEEX	94,02 %	**95,73 %**	92,30 %	90,88 %
G.723.1	88,98 %	91,59 %	**94,49 %**	87,44 %
GSM 06.10	91,71 %	89,40 %	88,39 %	**96,21 %**

Tab. Z.3.77. Wartości sprawności systemu opartego na cechach behawioralnych dla różnych standardów kodowania (35s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**51,01 %**	24,70 %	29,98 %	33,89 %
SPEEX	23,28 %	**38,80 %**	24,70 %	21,27 %
G.723.1	26,78 %	24,76 %	**38,68 %**	31,64 %
GSM 06.10	29,92 %	21,62 %	33,71 %	**43,25 %**

Tab. Z.3.78. Wartości sprawności wariantu fuzji danych na poziomie cech dla różnych standardów kodowania (35s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,09 %**	93,72 %	89,63 %	91,88 %
SPEEX	93,54 %	**95,79 %**	91,47 %	89,75 %
G.723.1	89,87 %	91,53 %	**94,25 %**	88,09 %
GSM 06.10	91,47 %	89,51 %	87,80 %	**96,21 %**

Tab. Z.3.79. Wartości sprawności wariantu fuzji danych na poziomie cech po zastosowaniu algorytmu genetycznego dla różnych standardów kodowania (35s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,45 %**	94,67 %	92,12 %	90,46 %
SPEEX	93,01 %	**95,68 %**	92,18 %	88,21 %
G.723.1	85,96 %	90,23 %	**94,43 %**	83,83 %
GSM 06.10	88,74 %	89,69 %	87,20 %	**95,97 %**

Tab. Z.3.80. Wartości sprawności wariantu fuzji danych na poziomie decyzji dla różnych standardów kodowania (35s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,15 %**	93,84 %	89,16 %	92,06 %
SPEEX	93,42 %	**95,68 %**	91,35 %	89,81 %
G.723.1	89,81 %	92,65 %	**94,61 %**	88,51 %
GSM 06.10	92,18 %	89,16 %	86,49 %	**96,09 %**

Tab. Z.3.81. Wartości sprawności systemu *ARMiA* dla różnych standardów kodowania (40s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,80 %**	94,85 %	91,47 %	92,89 %
SPEEX	94,61 %	**96,15 %**	93,19 %	91,82 %
G.723.1	90,70 %	93,13 %	**95,38 %**	89,04 %
GSM 06.10	92,36 %	91,71 %	89,16 %	**96,33 %**

Tab. Z.3.82. Wartości sprawności systemu opartego na cechach behawioralnych dla różnych standardów kodowania (40s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**57,82 %**	27,55 %	34,42 %	37,50 %
SPEEX	26,13 %	**44,02 %**	28,91 %	23,58 %
G.723.1	30,45 %	28,97 %	**44,14 %**	36,55 %
GSM 06.10	35,60 %	24,70 %	37,74 %	**50,65 %**

Tab. Z.3.83. Wartości sprawności wariantu fuzji danych na poziomie cech dla różnych standardów kodowania (40s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,80 %**	94,91 %	91,88 %	93,01 %
SPEEX	94,25 %	**96,39 %**	92,77 %	91,53 %
G.723.1	90,82 %	93,07 %	**95,38 %**	89,69 %
GSM 06.10	92,36 %	91,35 %	88,74 %	**96,39 %**

Tab. Z.3.84. Wartości sprawności wariantu fuzji danych na poziomie cech po zastosowaniu algorytmu genetycznego dla różnych standardów kodowania (40s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,74 %**	95,14 %	92,65 %	91,11 %
SPEEX	93,19 %	**95,50 %**	92,71 %	89,10 %
G.723.1	87,80 %	91,17 %	**94,61 %**	85,19 %
GSM 06.10	89,63 %	90,46 %	88,45 %	**95,97 %**

Tab. Z.3.85. Wartości sprawności wariantu fuzji danych na poziomie decyzji dla różnych standardów kodowania (40s/20s)

	Kodek segmentów uczących			
Kodek segmentów testujących	**G.711 - a-law**	**SPEEX**	**G.723.1**	**GSM 06.10**
G.711 - a-law	**96,86 %**	95,02 %	91,65 %	93,07 %
SPEEX	94,02 %	**96,27 %**	92,89 %	91,05 %
G.723.1	90,94 %	93,25 %	**95,44 %**	89,34 %
GSM 06.10	92,54 %	91,94 %	89,40 %	**96,39 %**

Pełen obraz wyników pokazuje, że wykorzystanie metod integracji danych, a zwłaszcza techniki preselekcji pozwala na uzyskane wyższych wartości liczby poprawnych identyfikacji tożsamości. Przy długości segmentu testowego równej 20 sekund, fuzja danych zarówno na poziomie cech, jak i decyzji pozwala na uzyskane jednostkowych, wyższych wyników sprawności. Z kolei krótsze czasy testowania uwydatniają przewagę podejścia integracji poprzez wstępny wybór najbliższych sąsiadów. Dlatego też nie jest możliwe jednoznacznie określenie warunków (tj. kombinacji czasów uczenia i testowania, czy zastosowanych standardów kodowania), przy których którekolwiek rozwiązanie fuzji danych zawsze zapewni poprawę liczby poprawnych identyfikacji tożsamości mówcy.

Testy opracowanych rozwiązań przy wykorzystaniu zróżnicowanych torów akustycznych

Poniżej przedstawione zostały pełne wyniki testów systemów dla nagrań rejestrowanych przy pomocy zróżnicowanych urządzeń rejestrujących, tj.: w multisesyjnej bazie 50 głosów autorstwa mgr. inż. Daniela Posiadały. Tabele:

- od Z.3.86 do Z.3.88 zawierają wyniki (wyrażone w %) dla stałej długości segmentu testującego równej 10 sekund,
- od Z.3.89 do Z.3.90 zawierają wyniki (wyrażone w %) dla stałej długości segmentu testującego równej 15 sekund,
- Z.3.91 przedstawia wyniki (wyrażone w %) dla stałej długości segmentu testującego równej 20 sekund.

Przedstawione rezultaty podkreślają wnioski przytoczone w podrozdziale 6.6, tj.: poprawa liczby identyfikacji względem systemu odniesienia [60] jest silnie uzależniona od rodzaju (jakości) urządzeń, które służyły do rejestracji dźwięków. Niezależnie od długości segmentów uczącego i testującego widać tendencję do wzrostu sprawności przy wykorzystaniu metody fuzji danych na poziomie cech. Jednak mało liczna próba (50 osób – klas) oraz niejednoznaczne wyniki utrudniają wybór optymalnej techniki przetwarzania sygnałów mowy w przypadku zróżnicowanego środowiska akustycznego.

Tab. Z.3.86. Wartości sprawności różnych wariantów systemów ASR dla zróżnicowanych urządzeń rejestrujących (20 sekund uczenia i 10 sekund testowania)

		Tor uczący									
	Tor testujący	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
System *ARMiA*	**1**	**94**	70	66	18	30	26	90	88	10	24
Fuzja danych na poziomie cech		**92**	70	66	18	32	30	92	86	10	28
Fuzja danych na poziomie cech (AG)		**94**	68	68	26	22	26	84	82	10	22
Fuzja danych na poziomie decyzji		**94**	54	64	24	38	24	78	84	12	26
System *ARMiA*	**2**	72	**96**	60	18	40	30	82	66	10	20
Fuzja danych na poziomie cech		72	**96**	60	18	38	30	78	72	10	20
Fuzja danych na poziomie cech (AG)		84	**98**	56	28	36	38	94	76	12	30
Fuzja danych na poziomie decyzji		48	**94**	46	24	32	28	70	50	10	22
System *ARMiA*	**3**	90	76	**96**	22	38	36	80	82	18	32
Fuzja danych na poziomie cech		90	72	**94**	20	34	36	82	82	14	34
Fuzja danych na poziomie cech (AG)		90	76	**94**	28	36	38	84	82	20	22
Fuzja danych na poziomie decyzji		78	60	**96**	22	36	34	70	78	12	26
System *ARMiA*	**4**	28	28	30	**84**	8	10	26	26	10	16
Fuzja danych na poziomie cech		24	22	26	**84**	8	8	30	28	10	14
Fuzja danych na poziomie cech (AG)		22	28	30	**92**	10	8	28	34	8	16
Fuzja danych na poziomie decyzji		26	30	30	**84**	10	14	26	14	8	8
System *ARMiA*	**5**	34	40	28	10	**100**	72	40	34	10	46
Fuzja danych na poziomie cech		36	40	30	10	**100**	72	38	40	12	48
Fuzja danych na poziomie cech (AG)		36	34	22	14	**100**	70	42	32	8	44
Fuzja danych na poziomie decyzji		36	32	28	16	**94**	66	40	40	10	52
System *ARMiA*	**6**	80	50	34	10	64	**98**	36	36	10	40
Fuzja danych na poziomie cech		80	46	30	12	70	**98**	42	36	10	44
Fuzja danych na poziomie cech (AG)		82	40	28	12	54	**96**	34	32	8	40
Fuzja danych na poziomie decyzji		74	30	24	10	54	**84**	34	36	8	46
System *ARMiA*	**7**	78	82	70	20	48	42	**96**	90	10	28
Fuzja danych na poziomie cech		74	80	72	22	42	40	**96**	90	10	28
Fuzja danych na poziomie cech (AG)		82	88	74	30	32	30	**98**	86	14	14
Fuzja danych na poziomie decyzji		72	60	70	20	44	34	**96**	78	10	32
System *ARMiA*	**8**	32	74	70	18	42	38	86	**94**	12	32
Fuzja danych na poziomie cech		30	74	68	22	44	36	86	**94**	12	38
Fuzja danych na poziomie cech (AG)		34	74	70	28	36	42	88	**94**	10	30
Fuzja danych na poziomie decyzji		28	60	62	20	36	30	82	**92**	12	28
System *ARMiA*	**9**	16	34	34	14	22	22	44	38	**98**	12
Fuzja danych na poziomie cech		18	32	32	12	22	20	44	36	**98**	14
Fuzja danych na poziomie cech (AG)		20	26	24	14	10	16	30	26	**100**	18
Fuzja danych na poziomie decyzji		14	38	38	14	22	20	40	34	**96**	12
System *ARMiA*	**10**	96	22	24	12	40	50	26	22	12	**100**
Fuzja danych na poziomie cech		65	22	24	12	44	50	28	22	10	**100**
Fuzja danych na poziomie cech (AG)		72	18	20	14	36	40	26	16	16	**98**
Fuzja danych na poziomie decyzji		96	22	20	8	48	44	24	22	12	**90**

Tab. Z.3.87. Wartości sprawności różnych wariantów systemów ASR dla zróżnicowanych urządzeń rejestrujących (25 sekund uczenia i 10 sekund testowania)

		Tor uczący									
	Tor testujący	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
System *ARMiA*	**1**	**92**	66	68	14	32	22	88	84	10	28
Fuzja danych na poziomie cech		**92**	68	68	14	30	22	86	86	10	28
Fuzja danych na poziomie cech (AG)		**94**	68	80	28	36	20	80	84	12	28
Fuzja danych na poziomie decyzji		**100**	60	66	16	40	24	92	88	12	24
System *ARMiA*	**2**	68	**92**	58	10	46	32	76	76	8	30
Fuzja danych na poziomie cech		66	**92**	56	14	44	32	78	76	8	36
Fuzja danych na poziomie cech (AG)		74	**96**	64	30	48	36	84	80	12	32
Fuzja danych na poziomie decyzji		54	**96**	52	26	34	26	70	58	6	24
System *ARMiA*	**3**	84	66	**100**	16	42	36	78	84	14	32
Fuzja danych na poziomie cech		84	62	**100**	16	44	34	78	84	14	34
Fuzja danych na poziomie cech (AG)		86	68	**100**	30	46	28	82	82	18	28
Fuzja danych na poziomie decyzji		88	64	**98**	22	44	42	84	82	12	32
System *ARMiA*	**4**	20	22	30	**82**	18	18	24	32	10	16
Fuzja danych na poziomie cech		28	20	26	**82**	18	18	28	32	10	16
Fuzja danych na poziomie cech (AG)		38	32	30	**96**	12	14	28	34	10	14
Fuzja danych na poziomie decyzji		22	24	34	**86**	12	12	30	22	8	8
System *ARMiA*	**5**	38	44	28	10	**100**	72	38	38	8	54
Fuzja danych na poziomie cech		38	44	28	8	**100**	70	42	36	8	52
Fuzja danych na poziomie cech (AG)		30	36	24	12	**100**	62	38	28	8	48
Fuzja danych na poziomie decyzji		32	32	24	16	**96**	68	38	36	10	52
System *ARMiA*	**6**	88	56	30	8	74	**98**	40	46	12	52
Fuzja danych na poziomie cech		88	56	30	8	74	**98**	42	48	12	52
Fuzja danych na poziomie cech (AG)		86	40	28	10	72	**94**	42	38	12	42
Fuzja danych na poziomie decyzji		84	32	32	12	68	**88**	26	40	10	46
System *ARMiA*	**7**	78	80	78	16	40	40	**94**	96	14	32
Fuzja danych na poziomie cech		78	82	76	18	42	38	**94**	94	14	30
Fuzja danych na poziomie cech (AG)		86	86	82	32	32	38	**90**	94	12	26
Fuzja danych na poziomie decyzji		78	64	72	20	42	38	**100**	90	10	30
System *ARMiA*	**8**	34	80	68	16	42	36	86	**94**	12	30
Fuzja danych na poziomie cech		28	80	68	16	42	34	86	**94**	12	34
Fuzja danych na poziomie cech (AG)		30	82	66	26	46	44	88	**98**	12	26
Fuzja danych na poziomie decyzji		30	62	66	20	44	30	86	**96**	10	30
System *ARMiA*	**9**	18	36	44	6	20	16	38	42	**98**	16
Fuzja danych na poziomie cech		18	34	42	8	18	16	40	40	**96**	14
Fuzja danych na poziomie cech (AG)		22	38	28	10	16	20	30	28	**98**	18
Fuzja danych na poziomie decyzji		16	36	38	10	20	22	32	36	**96**	14
System *ARMiA*	**10**	94	34	28	12	48	48	28	22	6	**100**
Fuzja danych na poziomie cech		74	32	28	14	44	50	28	22	6	**100**
Fuzja danych na poziomie cech (AG)		80	30	24	14	44	40	26	18	12	**98**
Fuzja danych na poziomie decyzji		96	30	24	10	48	52	26	24	10	**96**

Tab. Z.3.88. Wartości sprawności różnych wariantów systemów ASR dla zróżnicowanych urządzeń rejestrujących (30 sekund uczenia i 10 sekund testowania)

		Tor uczący									
	Tor testujący	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
System *ARMiA*	**1**	**100**	66	74	12	30	26	92	90	14	26
Fuzja danych na poziomie cech		**100**	70	74	18	30	26	88	90	14	28
Fuzja danych na poziomie cech (AG)		**98**	72	76	28	26	22	84	90	12	24
Fuzja danych na poziomie decyzji		**100**	52	72	16	40	30	92	90	12	32
System *ARMiA*	**2**	80	**98**	66	8	40	40	78	80	10	26
Fuzja danych na poziomie cech		76	**98**	66	8	42	36	78	80	10	26
Fuzja danych na poziomie cech (AG)		78	**98**	64	32	48	40	86	80	16	42
Fuzja danych na poziomie decyzji		60	**96**	50	28	36	28	84	66	8	30
System *ARMiA*	**3**	84	72	**98**	16	38	36	84	82	16	36
Fuzja danych na poziomie cech		86	74	**98**	16	38	36	82	80	16	36
Fuzja danych na poziomie cech (AG)		90	84	**98**	36	36	28	90	82	20	36
Fuzja danych na poziomie decyzji		90	68	**100**	22	42	40	92	86	10	34
System *ARMiA*	**4**	16	20	20	**76**	12	8	20	22	6	14
Fuzja danych na poziomie cech		16	18	16	**76**	12	8	22	20	6	14
Fuzja danych na poziomie cech (AG)		32	36	26	**100**	12	8	28	28	14	12
Fuzja danych na poziomie decyzji		22	26	34	**88**	12	14	36	24	6	10
System *ARMiA*	**5**	32	48	22	10	**100**	68	38	22	6	46
Fuzja danych na poziomie cech		32	48	22	8	**100**	70	38	26	6	44
Fuzja danych na poziomie cech (AG)		28	44	22	10	**100**	72	34	22	8	46
Fuzja danych na poziomie decyzji		34	36	30	18	**98**	66	42	36	12	48
System *ARMiA*	**6**	92	58	30	10	80	**96**	38	36	10	60
Fuzja danych na poziomie cech		88	58	34	8	80	**96**	36	36	10	60
Fuzja danych na poziomie cech (AG)		90	52	28	12	78	**98**	36	36	10	54
Fuzja danych na poziomie decyzji		90	36	34	16	72	**96**	34	42	8	52
System *ARMiA*	**7**	88	86	76	14	48	52	**94**	94	14	26
Fuzja danych na poziomie cech		84	86	74	14	48	54	**94**	94	14	26
Fuzja danych na poziomie cech (AG)		88	90	78	32	44	44	**92**	92	16	26
Fuzja danych na poziomie decyzji		84	66	72	20	44	36	**98**	92	8	30
System *ARMiA*	**8**	36	84	80	16	38	44	90	**96**	18	40
Fuzja danych na poziomie cech		36	82	80	16	40	44	90	**96**	18	40
Fuzja danych na poziomie cech (AG)		32	86	82	32	42	42	88	**96**	16	32
Fuzja danych na poziomie decyzji		26	66	70	22	40	32	86	**98**	10	30
System *ARMiA*	**9**	22	32	42	8	16	22	38	34	**94**	20
Fuzja danych na poziomie cech		24	32	42	8	16	24	38	32	**94**	22
Fuzja danych na poziomie cech (AG)		18	38	34	8	16	22	32	32	**96**	18
Fuzja danych na poziomie decyzji		20	32	38	10	14	32	30	36	**92**	14
System *ARMiA*	**10**	94	36	18	14	46	54	26	16	12	**100**
Fuzja danych na poziomie cech		74	36	20	14	48	54	26	14	12	**100**
Fuzja danych na poziomie cech (AG)		80	32	24	14	42	46	24	16	14	**98**
Fuzja danych na poziomie decyzji		96	30	18	14	50	52	24	28	10	**96**

Tab. Z.3.89. Wartości sprawności różnych wariantów systemów ASR dla zróżnicowanych urządzeń rejestrujących (20 sekund uczenia i 15 sekund testowania)

		Tor uczący									
	Tor testujący	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
System *ARMiA*	**1**	**100**	72	72	16	34	28	94	88	10	30
Fuzja danych na poziomie cech		**98**	72	70	16	32	32	94	88	10	30
Fuzja danych na poziomie cech (AG)		**98**	70	76	26	22	28	88	88	10	26
Fuzja danych na poziomie decyzji		**92**	48	64	20	30	20	84	78	12	24
System *ARMiA*	**2**	70	**94**	56	16	40	26	78	74	10	28
Fuzja danych na poziomie cech		68	**94**	56	18	40	28	78	72	10	28
Fuzja danych na poziomie cech (AG)		82	**98**	60	30	40	36	94	82	12	32
Fuzja danych na poziomie decyzji		56	**88**	48	16	44	30	70	58	8	28
System *ARMiA*	**3**	92	80	**98**	20	42	40	90	84	14	34
Fuzja danych na poziomie cech		92	74	**98**	20	42	38	88	84	14	34
Fuzja danych na poziomie cech (AG)		94	78	**100**	30	34	40	88	86	16	26
Fuzja danych na poziomie decyzji		78	60	**100**	26	40	30	78	82	10	30
System *ARMiA*	**4**	26	24	34	**86**	12	12	32	30	8	18
Fuzja danych na poziomie cech		24	20	30	**86**	12	12	30	30	8	18
Fuzja danych na poziomie cech (AG)		30	38	30	**98**	8	8	26	34	12	16
Fuzja danych na poziomie decyzji		24	18	32	**80**	14	20	22	24	10	10
System *ARMiA*	**5**	34	40	24	10	**100**	76	38	32	10	46
Fuzja danych na poziomie cech		38	38	28	10	**100**	76	40	34	10	48
Fuzja danych na poziomie cech (AG)		34	36	26	14	**100**	76	40	30	8	44
Fuzja danych na poziomie decyzji		42	34	32	20	**96**	64	40	38	10	52
System *ARMiA*	**6**	90	52	32	12	76	**98**	38	42	12	50
Fuzja danych na poziomie cech		90	48	36	12	76	**98**	46	42	12	48
Fuzja danych na poziomie cech (AG)		86	42	26	14	68	**98**	42	40	12	38
Fuzja danych na poziomie decyzji		80	38	34	14	72	**88**	36	44	10	50
System *ARMiA*	**7**	80	84	74	18	46	40	**100**	98	10	30
Fuzja danych na poziomie cech		80	82	74	20	46	40	**100**	98	10	34
Fuzja danych na poziomie cech (AG)		86	88	78	32	32	34	**100**	94	10	22
Fuzja danych na poziomie decyzji		78	62	72	22	38	32	**92**	86	12	28
System *ARMiA*	**8**	36	80	74	16	48	42	92	**98**	12	32
Fuzja danych na poziomie cech		32	80	74	20	44	42	90	**98**	12	40
Fuzja danych na poziomie cech (AG)		32	78	72	28	44	44	94	**98**	12	26
Fuzja danych na poziomie decyzji		30	64	66	20	38	22	78	**90**	10	30
System *ARMiA*	**9**	22	36	34	8	18	24	40	40	**98**	12
Fuzja danych na poziomie cech		24	38	34	10	18	20	42	42	**98**	16
Fuzja danych na poziomie cech (AG)		24	36	22	14	16	16	34	24	**100**	18
Fuzja danych na poziomie decyzji		16	32	38	6	16	18	36	34	**90**	16
System *ARMiA*	**10**	96	26	24	12	42	52	30	24	14	**100**
Fuzja danych na poziomie cech		94	28	24	12	42	56	30	24	12	**100**
Fuzja danych na poziomie cech (AG)		94	24	20	12	40	42	30	16	14	**100**
Fuzja danych na poziomie decyzji		98	40	24	12	50	50	22	20	6	**98**

Tab. Z.3.90. Wartości sprawności różnych wariantów systemów ASR dla zróżnicowanych urządzeń rejestrujących (25 sekund uczenia i 15 sekund testowania)

		Tor uczący									
	Tor testujący	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
System *ARMiA*	**1**	**100**	70	74	14	32	26	90	92	10	28
Fuzja danych na poziomie cech		**100**	70	72	16	28	24	88	92	10	32
Fuzja danych na poziomie cech (AG)		**98**	78	80	28	30	24	90	92	14	26
Fuzja danych na poziomie decyzji		**98**	54	68	28	30	22	86	90	10	28
System *ARMiA*	**2**	72	**96**	58	10	42	32	78	76	8	32
Fuzja danych na poziomie cech		74	**96**	56	14	42	30	80	76	8	34
Fuzja danych na poziomie cech (AG)		82	**100**	70	32	52	38	90	82	14	36
Fuzja danych na poziomie decyzji		60	**94**	50	22	36	28	72	72	8	28
System *ARMiA*	**3**	88	76	**100**	18	44	36	84	84	14	34
Fuzja danych na poziomie cech		88	76	**100**	16	44	38	84	86	16	34
Fuzja danych na poziomie cech (AG)		92	78	**100**	34	44	36	88	84	22	34
Fuzja danych na poziomie decyzji		84	68	**100**	30	40	36	84	82	14	34
System *ARMiA*	**4**	26	26	24	**88**	12	14	26	30	8	16
Fuzja danych na poziomie cech		28	24	20	**86**	12	14	24	32	10	16
Fuzja danych na poziomie cech (AG)		38	36	34	**100**	12	10	26	34	10	16
Fuzja danych na poziomie decyzji		26	22	22	**88**	10	16	26	24	10	10
System *ARMiA*	**5**	34	46	28	8	**100**	78	42	28	8	52
Fuzja danych na poziomie cech		38	48	28	8	**100**	74	42	30	8	52
Fuzja danych na poziomie cech (AG)		26	40	24	14	**100**	74	36	30	6	50
Fuzja danych na poziomie decyzji		32	38	28	18	**100**	68	38	34	8	52
System *ARMiA*	**6**	90	54	34	8	84	**100**	40	46	12	54
Fuzja danych na poziomie cech		90	56	36	8	82	**100**	44	48	12	52
Fuzja danych na poziomie cech (AG)		88	44	34	14	78	**98**	48	38	8	50
Fuzja danych na poziomie decyzji		88	44	40	16	76	**96**	40	46	10	50
System *ARMiA*	**7**	88	82	80	14	48	46	**100**	96	12	30
Fuzja danych na poziomie cech		88	82	80	16	48	44	**100**	94	12	32
Fuzja danych na poziomie cech (AG)		90	90	84	32	40	44	**98**	98	14	28
Fuzja danych na poziomie decyzji		86	66	76	20	40	38	**100**	92	6	26
System *ARMiA*	**8**	32	84	72	16	44	42	92	**98**	16	32
Fuzja danych na poziomie cech		34	82	72	16	46	40	92	**98**	14	32
Fuzja danych na poziomie cech (AG)		36	80	76	30	46	42	90	**96**	12	28
Fuzja danych na poziomie decyzji		32	64	72	26	40	32	90	**96**	10	24
System *ARMiA*	**9**	24	38	44	8	18	20	36	40	**94**	16
Fuzja danych na poziomie cech		24	38	40	8	18	20	40	40	**92**	20
Fuzja danych na poziomie cech (AG)		24	36	34	8	16	20	30	26	**98**	20
Fuzja danych na poziomie decyzji		18	34	36	10	18	20	32	30	**88**	16
System *ARMiA*	**10**	98	38	28	12	44	58	28	22	8	**100**
Fuzja danych na poziomie cech		98	34	26	12	44	58	28	22	8	**100**
Fuzja danych na poziomie cech (AG)		96	36	26	14	44	46	24	22	12	**100**
Fuzja danych na poziomie decyzji		96	32	26	14	48	54	18	26	10	**98**

Tab. Z.3.91. Wartości sprawności różnych wariantów systemów ASR dla zróżnicowanych urządzeń rejestrujących (20 sekund uczenia i 20 sekund testowania)

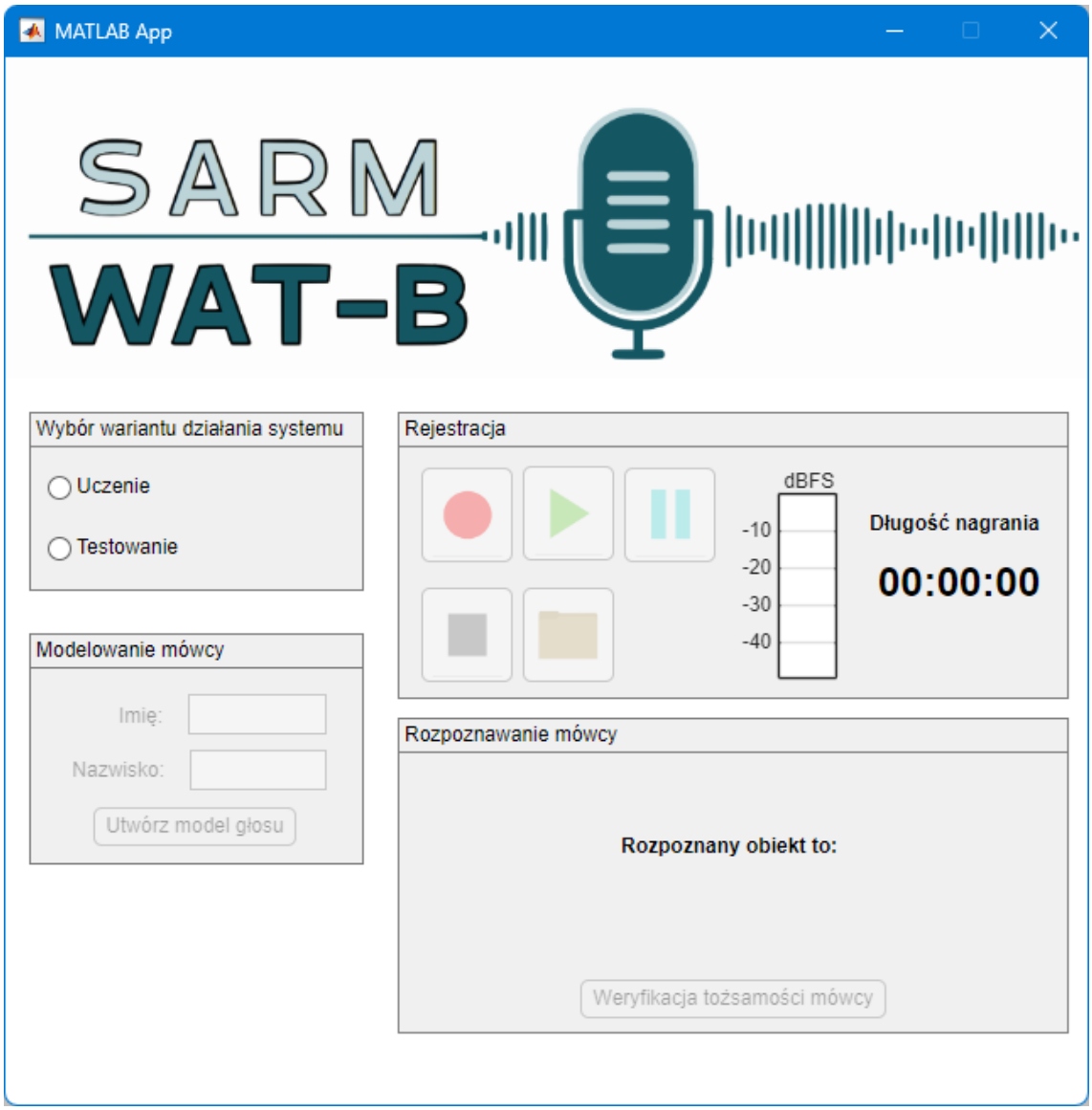
		Tor uczący									
	Tor testujący	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
System *ARMiA*	**1**	**100**	74	74	16	36	34	94	92	10	30
Fuzja danych na poziomie cech		**100**	72	74	14	34	32	94	90	10	32
Fuzja danych na poziomie cech (AG)		**100**	76	80	28	24	34	90	88	10	28
Fuzja danych na poziomie decyzji		**98**	60	72	26	28	26	90	86	14	26
System *ARMiA*	**2**	74	**94**	54	16	40	30	86	78	10	28
Fuzja danych na poziomie cech		70	**94**	54	16	38	30	86	78	10	30
Fuzja danych na poziomie cech (AG)		88	**100**	64	30	42	44	94	86	10	34
Fuzja danych na poziomie decyzji		72	**98**	56	18	38	28	70	74	10	28
System *ARMiA*	**3**	94	80	**100**	20	42	40	92	86	14	32
Fuzja danych na poziomie cech		94	78	**100**	20	40	40	92	88	14	36
Fuzja danych na poziomie cech (AG)		94	82	**100**	34	36	38	92	86	18	30
Fuzja danych na poziomie decyzji		84	66	**94**	24	34	32	80	76	16	36
System *ARMiA*	**4**	24	24	34	**88**	12	12	34	34	6	18
Fuzja danych na poziomie cech		22	22	30	**86**	12	12	28	30	6	18
Fuzja danych na poziomie cech (AG)		34	36	36	**98**	10	10	26	36	10	16
Fuzja danych na poziomie decyzji		20	20	18	**78**	12	10	20	26	6	12
System *ARMiA*	**5**	38	38	28	12	**100**	76	38	30	10	46
Fuzja danych na poziomie cech		38	40	28	12	**100**	74	38	32	10	44
Fuzja danych na poziomie cech (AG)		36	36	28	14	**100**	74	40	30	8	48
Fuzja danych na poziomie decyzji		34	42	22	16	**100**	62	32	20	6	42
System *ARMiA*	**6**	92	52	38	12	80	**98**	46	40	10	50
Fuzja danych na poziomie cech		92	48	36	12	78	**98**	48	40	10	46
Fuzja danych na poziomie cech (AG)		90	42	30	14	72	**100**	48	40	8	40
Fuzja danych na poziomie decyzji		90	42	34	14	80	**94**	34	42	8	52
System *ARMiA*	**7**	84	86	76	18	48	40	**98**	98	10	30
Fuzja danych na poziomie cech		84	86	74	20	50	38	**98**	98	10	30
Fuzja danych na poziomie cech (AG)		86	90	78	34	38	42	**100**	96	10	24
Fuzja danych na poziomie decyzji		84	74	76	24	44	44	**92**	92	12	28
System *ARMiA*	**8**	34	82	74	14	48	44	92	**98**	12	34
Fuzja danych na poziomie cech		32	80	74	14	46	42	92	**98**	12	40
Fuzja danych na poziomie cech (AG)		34	78	70	28	40	48	94	**94**	12	30
Fuzja danych na poziomie decyzji		28	74	80	22	36	30	88	**96**	10	34
System *ARMiA*	**9**	24	34	32	10	10	24	36	42	**94**	18
Fuzja danych na poziomie cech		24	32	32	10	12	22	38	44	**94**	16
Fuzja danych na poziomie cech (AG)		22	36	22	14	16	16	30	26	**100**	20
Fuzja danych na poziomie decyzji		20	24	36	14	12	22	32	22	**92**	18
System *ARMiA*	**10**	99	30	24	12	46	58	30	22	14	**100**
Fuzja danych na poziomie cech		98	30	24	12	44	58	30	22	14	**100**
Fuzja danych na poziomie cech (AG)		98	26	20	14	44	46	28	16	14	**100**
Fuzja danych na poziomie decyzji		96	36	20	16	48	52	18	26	8	**100**

Z. 4

Opis aplikacji

W ramach niniejszej pracy powstała aplikacja *SARM-WAT-B* (**S**ystem **A**utomatycznego **R**ozpoznawania **M**ówcy **W**sparty **AT**rybutami **B**ehawioralnymi). Program umożliwia pracę w dwóch trybach, tj. uczenia i testowania. Pierwszy z nich umożliwia dodawanie do istniejącej bazy głosów kolejnych próbek poprzez rejestracje dźwięku przy pomocy urządzeń wejściowych bądź wczytanie gotowego nagrania z dysku. Drugi tryb to testowanie, w którym na podstawie zarejestrowanego bądź wczytanego głosu realizowany jest proces identyfikacji tożsamości mówcy w istniejącej bazie danych.

Na rys. Z.4.1 przedstawiono interfejs aplikacji *SARM-WAT-B*.

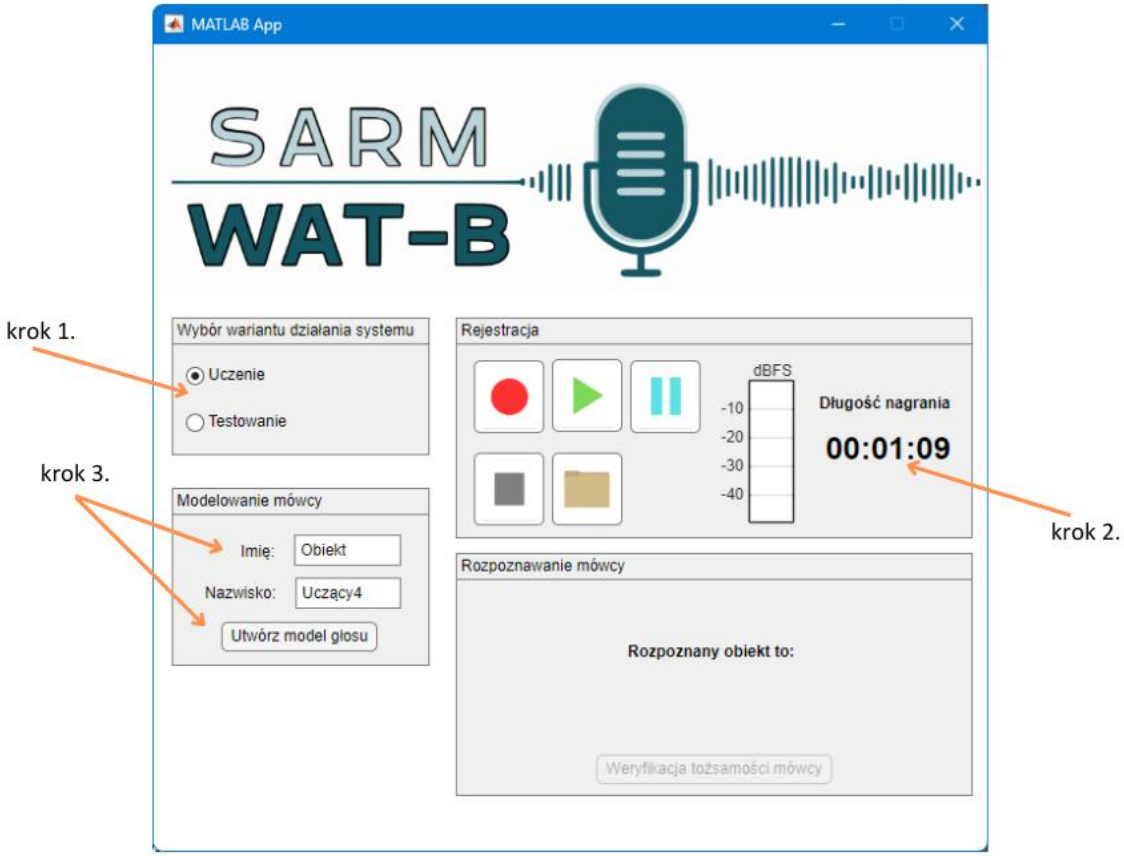


Rys. Z.4.1. Interfejs graficzny aplikacji *SARM-WAT-B* (**S**ystem **A**utomatycznego **R**ozpoznawania **M**ówcy **W**sparty **AT**rybutami **B**ehawioralnymi)

Proces uczenia - modelowanie głosu mówcy

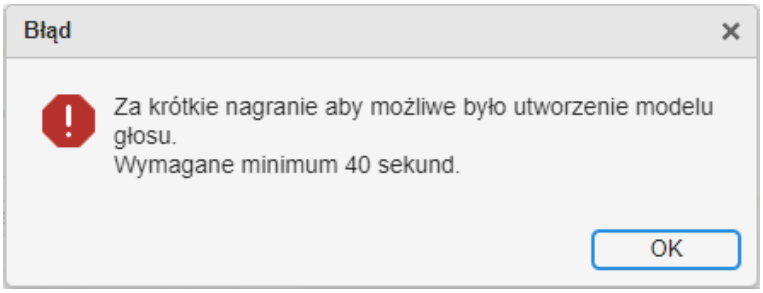
Na rysunku Z.4.2 pokazano kolejne kroki jakie należy wykonać w celu utworzenia modelu mówcy z przetwarzanej próbki głosu. Dodatkowo na rys. Z.4.3 oraz Z.4.4 przedstawiono monity, które mogą wyświetlić się podczas pracy z aplikacją. Proces modelowania głosu mówcy składa się z 3 głównych kroków:

- krok 1. – wybór wariantu pracy systemu (w tym wypadku *Uczenie*);
- krok 2. – wybór źródła przetwarzanego dźwięku (rejestracja przy pomocy mikrofonu bądź plik audio zapisany na dysku);
- krok 3. – wprowadzenie danych personalnych (wymaganych do zapisu modelu) oraz rozpoczęcie procesu modelowania głosu.

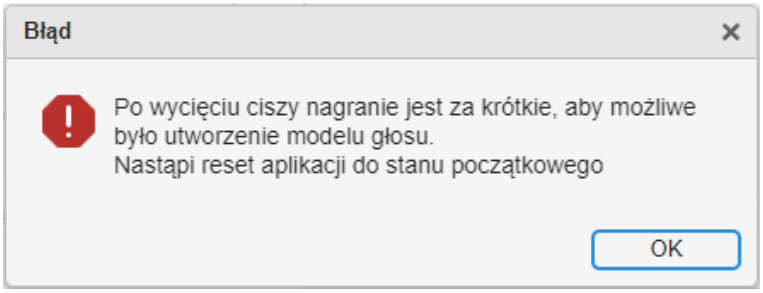


Rys. Z.4.2. Proces tworzenia modelu głosu mówcy w aplikacji *SARM-WAT-B*

W przypadku, gdy przetwarzany dźwięk jest krótszy niż 40 sekund, wyświetlony zostanie odpowiedni komunikat (rys. Z.4.3). Tak samo w przypadku, gdy po zakończonej, wstępnej analizie sygnału nie spełni on kryteriów dotyczących długości dla konkretnego trybu pracy (rys. Z.4.4).



Rys. Z.4.3. Komunikat o błędzie wynikającym ze zbyt krótkiego czasu nagrania



Rys. Z.4.4. Komunikat o błędzie wynikającym z niespełnionych kryteriów długości sygnału po wstępnej obróbce

Proces testowania – identyfikacja tożsamości mówcy

Na rysunku Z.4.5 zobrazowano kolejne kroki jakie należy wykonać w celu identyfikacji tożsamości mówcy na podstawie próbki jego głosu. Proces identyfikacji tożsamości mówcy składa się z 4 głównych kroków (dwa pierwsze są tożsame z etapami występującymi podczas procesu uczenia):

- krok 1. – wybór wariantu pracy systemu (w tym wypadku *Testowanie*);
- krok 2. – wybór źródła przetwarzanego dźwięku (rejestracja przy pomocy mikrofonu bądź plik audio zapisany na dysku);
- krok 3. – rozpoczęcie procesu modelowania głosu (w tym wariancie pracy systemu SARM-WAT-B nie jest wymagane wprowadzenie danych personalnych, gdyż utworzony model głosu nie będzie zapisany);
- krok 4. – wybór wariantu fuzji danych oraz przystąpienie do procesu identyfikacji tożsamości mówcy (gdy użytkownik rozpocznie etap identyfikacji mówcy, to wszystkie pozostałe elementy interfejsu zostają wyłączone).

Rys. Z.4.5. Proces identyfikacji tożsamości mówcy w aplikacji *SARM-WAT-B*