POLITECHNIKA ŚLĄSKA
WYDZIAŁ INŻYNIERII BIOMEDYCZNEJ

Opracowanie metodyki przetwarzania sygnałów akustycznych i danych obrazowych dla celów komputerowego wspomagania diagnostyki logopedycznej z wykorzystaniem technik sztucznej inteligencji

— ROZPRAWA DOKTORSKA—

AUTOR
mgr inż. Agata Sage

PROMOTOR
dr hab. inż. Paweł Badura, prof. PŚ

Zabrze 2024

Składam serdeczne podziękowania Panu dr. hab. inż. Pawłowi Badurze, prof. PŚ za okazane wsparcie, zaangażowanie i cierpliwość.

Szczególne wyrazy wdzięczności kieruję również do dr inż. Zuzanny Miodońskiej oraz dr. inż. Michała Kręcichwosta za życzliwość i chęć dzielenia się wiedzą.

Słowa uznania należą się także Pani dr Joannie Trzaskalik oraz Pani mgr Ewie Kwaśniok za owocną współpracę w zespole badawczym.

Dziękuję Pani prof. dr hab. inż. Ewie Piętce oraz Koleżankom i Kolegom z katedry Informatyki Medycznej i Sztucznej Inteligencji za chęć pomocy oraz serdeczność.

Dziękuję moim przyjaciołom oraz rodzinie, zwłaszcza mojej siostrze Marcie.

Najbardziej jednak dziękuję mojej Mamie.

Spis treści

Spis rysunków

Spis tabel

Spis skrótów i oznaczeń

Ważniejsze skróty

ASR automatyczne rozpoznawanie mowy (ang. *automated speech recognition*)

ASPP piramida ASPP (ang. *atrous spatial pyramid pooling*)

BF test Browna-Forsythe'a

CALL komputerowe wspomaganie nauki języków (ang. *computer-assisted language learning*)

CAPT komputerowe wspomaganie ćwiczenia artykulacji (ang. *computer-assisted pronunciation training*)

CLAHE adaptacyjne wyrównywanie histogramu z ograniczeniem wzmocnienia kontrastu (ang. *contrast-limited adaptive histogram equalization*)

CNN splotowa sieć neuronowa (ang. *convolutional neural network*)

DFT dyskretna transformata Fouriera (ang. *discrete Fourier transform*)

DRLSE metoda ewolucji zbioru poziomic (ang. *distance regularized level set evolution*)

FBE energie zespołu filtrów (ang. *filter bank energies*)

GLCM macierz współwystępowania poziomów szarości (ang. *gray level co-ocurrence matrix*)

GLRLM macierz jednorodnych ciągów pikseli (ang. *gray level run length matrix*)

GLSZM macierz jednorodnych stref pikseli (ang. *gray level size zone matrix*)

IQR rozstęp ćwiartkowy (ang. *interquartile range*)

IPA międzynarodowy alfabet fonetyczny (ang. International Phonetic Alphabet)

KW test Kruskala-Wallisa

MFCC współczynniki melcepstralne (ang. *Mel-frequency cepstral coefficients*)

NGTDM macierz różnic poziomów szarości w sąsiedztwie (ang. *neighbouring gray tone difference matrix*)

NPR wskaźnik wymowy normatywnej (ang. *normal pronunciation rate*)

ROI obszar zainteresowania (ang. *region of interest*)

SGD metoda stochastycznego spadku wzdłuż gradientu (ang. *stochastic gradient descent*)

STFT krótkoczasowa transformata Fouriera (ang. *short-time Fourier transform*)

SW test Shapiro-Wilka

U MW test U Manna-Whitneya

YOLO sieć YOLO (ang. *you only look once*)

Ważniejsze oznaczenia

Acc dokładność (ang. *accuracy*)

AP średnia precyzja (ang. *average precision*)

DSC współczynnik Dice'a (ang. *Dice similarity coefficient*)

$F1$ wskaźnik F1 (ang. *F1 score*)

IoU współczynnik Jaccarda, część wspólna do całości (ang. *intersection over union*)

mAP uśredniona średnia precyzja (ang. *mean average precision*)

P precyzja (ang. *precision*)

R czułość (ang. *recall*)

t_{IoU} próg współczynnika Jaccarda w detekcji obiektów

Uwagi na temat oznaczeń cech wizualnych i akustycznych

Ze względu na mnogość cech wizualnych i akustycznych używanych w badaniach konieczne było wprowadzenie jednolitej, zrozumiałej i łatwo intepretowalnej nomenklatury oznaczeń. W tym celu zastosowano system indeksów dolnych i górnych oraz przedrostków w symbolach cech, jednoznacznie identyfikujących cechę w ramach jej grupy. Przyjęte zasady zestawiono poniżej.

Cechy wizualne

Cechy kształtu 2D są oznaczane przez dodanie indeksu górnego „$2D$", np. A^{2D}, Ax^{2D}_{major}. Szczegółowe oznaczenia: tab. 4.5.

Cechy kształtu 3D są oznaczane przez dodanie indeksu górnego „$3D$", np. V^{3D}, D^{3D}_{Feret}. Szczegółowe oznaczenia: tab. 4.6.

Cechy teksturowe I rzędu bazujące na histogramie są oznaczane przez dodanie indeksu dolnego „h", np. E_h, σ^2_h. Szczegółowe oznaczenia: tab. 4.7.

Cechy teksturowe I rzędu bazujące na poziomach szarości obrazu są oznaczane przez literę „I" z dowolnymi indeksami, np. I_E, I_R. Szczegółowe oznaczenia: tab. 4.8.

Cechy teksturowe z rodziny GLCM są oznaczane przez dodanie indeksu górnego „$GLCM$", np. ASM^{GLCM}, Hom^{GLCM}. Szczegółowe oznaczenia: tab. 4.9.

Cechy teksturowe z rodziny GLRLM są oznaczane przez dodanie indeksu górnego „$GLRLM$", np. SRE^{GLRLM}, RP^{GLRLM}. Szczegółowe oznaczenia: tab. 4.10.

Cechy teksturowe z rodziny GLSZM są oznaczane przez dodanie indeksu górnego „$GLSZM$", np. SZE^{GLSZM}, ZP^{GLSZM}. Szczegółowe oznaczenia: tab. 4.11.

Cechy teksturowe z rodziny NGTDM są oznaczane przez dodanie indeksu górnego „$NGTDM$", np. $Coar^{NGTDM}$, TS^{NGTDM}. Szczegółowe oznaczenia: tab. 4.12.

Do wskazania wymiarowości cech teksturowch użyto indeksu dolnego „$2D$" lub „$3D$", np. GLV^{GLSZM}_{2D} lub Bus^{NGTDM}_{3D}.

Cechy akustyczne

Cechy w dziedzinie czasu są oznaczane przez dodanie indeksu dolnego „t", np. ZCR_t, HR_t. Szczegółowe oznaczenia: tab. 5.1.

Cechy w dziedzinie częstotliwości są oznaczane przez dodanie indeksu dolnego „f”, np. $SCen_f$, SE_f. Wyjątek stanowią współczynniki melcepstralne, oznaczane za pomocą $MFCC$ z indeksem dolnym wskazującym numer współczynnika. Szczegółowe oznaczenia: tab. 5.2.

Cechy szumowe są oznaczane przez dodanie przedrostka „N”, np. NFF_i, NPA. Szczegółowe oznaczenia: tab. 5.3.

1. Wprowadzenie

Wady wymowy stanowią istotną barierę w zrównoważonym rozwoju dziecka. Wpływają na trudności w nauce czytania i pisania, stają się źródłem kompleksów i wycofania społecznego. To przeszkody nie tylko językowe, ale również psychologiczne, socjologiczne i dydaktyczne. Zaniedbanie wad wymowy, które pojawiają się w wieku dziecięcym może skutkować ich dalszym pogłębianiem, a w efekcie rzutować na życie dorosłe. Badania przeprowadzone w latach 80. ubiegłego wieku raportowały występowanie zaburzeń u około 20-30% sześciolatków [98], podczas gdy w na początku drugiej dekady XXI wieku liczbę tę szacowano już na 48% [98, 146]. Wśród patologii mowy specjaliści mówią o dominacji jednej z jej rodzajów — dyslalii. Są to odstępstwa od normy w artykulacji fonemów. Z kolei najczęściej występującym typem dyslalii wśród dzieci jest seplenienie (sygmatyzm). Seplenieniem określa się niepoprawną realizcję głosek dentalizowanych (inaczej sybilantów): /s/, /z/, /ts/, /dz/, /s̪/, /z̪/, /ts̪/, /dz̪/, /ɕ/, /ʑ/, /tɕ/, /dʑ/ (tab. 1.1)[1]. Patologiczna może być realizacja zarówno jednej głoski spośród dentalizowanych, jak i kilku jednocześnie lub nawet wszystkich.

Tab. 1.1: Zestawienie sybilantów w transkrypcji międzynarodowego alfabetu fonetycznego IPA i w zapisie ortograficznym.

Szereg	syczący				szumiący				ciszący			
Symbol IPA	/s/	/z/	/ts/	/dz/	/s̪/	/z̪/	/ts̪/	/dz̪/	/ɕ/	/ʑ/	/tɕ/	/dʑ/
Zapis ortograficzny	s	z	c	dz	sz	ż	cz	dż	ś	ź	ć	dź

Diagnostyka logopedyczna jest procesem złożonym — zaburzenia mogą mieć odmienną etiologię względem siebie, a ocenie poddaje się nie tylko mowę swobodną (m.in. zasób słownictwa, poprawność budowania zdań, płynność i prozodię), ale także wybrane aspekty anatomiczne i fizjologiczne (w tym: sprawność narządów artykulacyjnych, oddychanie i połykanie, zgryz i uzębienie, budowę jamy nosowej i ustnej czy kwestie związane ze słuchem fizycznym i fonemowym) [138]. Wielowarstwowość procesu stawiania diagnozy oraz jego oparcie na, często subiektywnej, obserwacji pracy narządów artykulacyjnych wymaga doświadczenia specjalisty i bywa czasochłonne. Z kolei właściwie dobrana ścieżka

[1] Transkrypcja fonetyczna w rozprawie została zapisana z użyciem międzynarodowego alfabetu fonetycznego IPA (ang. *International Phonetic Alphabet*) [54].

terapeutyczna zwiększa skuteczność leczenia i skraca czas jego trwania. Opracowywanie komputerowych metod wspierających diagnostykę logopedyczną jest zatem kluczowe z wielu wymienionych wyżej względów.

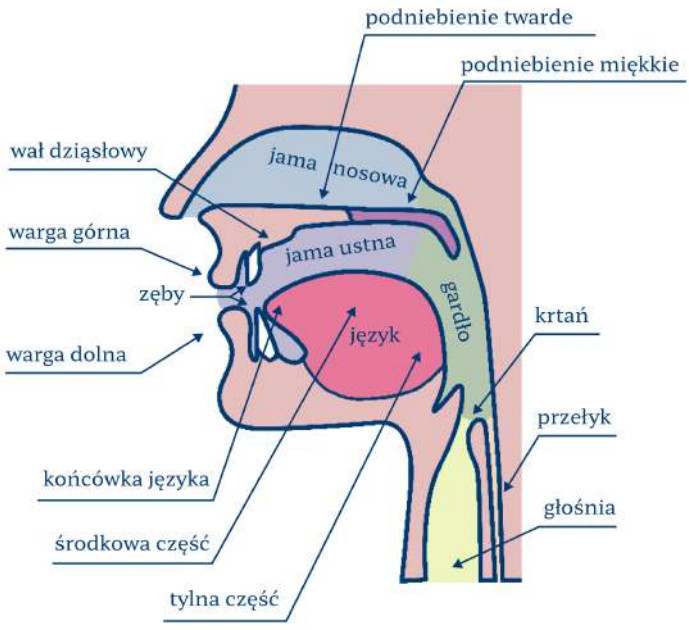
1.1 Artykulacja głosek dentalizowanych

Głoska jest najmniejszym i niepodzielnym elementem mowy, który da się wyodrębnić. Poszczególne głoski wyróżniają się ze względu na strukturę akustyczną (wysokość, siła, tembr, długość) i artykulacyjną (związanych z ułożeniem narządów mowy). Najczęściej do kategoryzacji artykulacyjnej głosek wykorzystuje się [146, 150]:

- kierunek przepływu powietrza przez narządy mowy (głoski ekspiracyjne i inspiracyjne);
- zachowanie się więzadeł głosowych (głoski dźwięczne, kiedy więzadła są zsunięte i wibrujące; bezdźwięczne, kiedy więzadła są rozsunięte, niewibrujące);
- położenie podniebienia miękkiego (głoski ustne oraz nosowe);
- stopień zbliżenia narządów mowy (głoski: otwarte, nosowe, boczne, drżące, półsamogłoskowe, szczelinowe, zwarto-szczelinowe, zwarto-wybuchowe);
- miejsce artykulacji (głoski: wargowe, przedniojęzykowe, środkowojęzykowe, tylnojęzykowe);
- pionowe i poziome ruchy języka (przy ruchach pionowych: głoski wysokie, średnie, niskie; przy ruchach poziomych: głoski przednie, środkowe, tylne).

W kontekście konkretnych systemów językowych wprowadzono także pojęcie fonemu. Fonem jest najmniejszą jednostką rozróżnialną dla mówców danego języka i jest realizowany w mowie przez głoski[2]. Opis fonemu obejmuje jedynie cechy artykulacyjne, które w wybranym systemie językowym pozwalają na jego rozróżnienie od innych fonemów (w przeciwieństwie do opisu głoski, który może być obszerny). W języku polskim większość fonemów definiuje się wykorzystując trzy aspekty: miejsce artykulacji, sposób artykulacji oraz dźwięczność. Niemniej jednak w przypadku patologicznej wymowy zaburzeniom mogą ulegać też inne cechy głoski, które nie są istotne w kontekście fonemu, ale wpływają na jej brzmienie [146, 150]. Uproszczony schemat traktu głosowego przedstawia rys. 1.1 [146].

[2] Ponieważ głoski są dźwiękową realizacją fonemów, w opisywanej pracy przyjęto uproszczenie stosowane często w tematyce przetwarzania mowy i wymiennie stosowano pojęcia głoski oraz fonemu.



Rys. 1.1: Schemat traktu głosowego człowieka. Rysunek opracowano na podstawie [146].

Podstawową wspólną cechą sybilantów jest dentalizacja — w trakcie realizacji głosek górne i dolne siekacze znacznie się do siebie zbliżają. Tarcie wypływającego prądu powietrza o krawędzie zębów powoduje wytworzenie szumu [111, 140]. Głoski dentalizowane mogą być klasyfikowane ze względu na: dźwięczność, miejsce artykulacji i sposób artykulacji. W kontekście dźwięczności rozróżnia się fonemy dźwięczne lub bezdźwięczne (w zależności od udziału więzadeł głosowych). Sposób artykulacji dzieli sybilanty na szczelinowe (frykatywne, trące) i zwarto-szczelinowe (afrykaty, przytarte). Podczas artykulacji głosek szczelinowych jedyną barierą, na jaką napotyka prąd powietrza jest szczelina pomiędzy narządami mowy (dzięki temu artykulację fonemu można przedłużać bez zmiany jej barwy) [150, 164]. Z kolei spółgłoski zwarto-szczelinowe składają się z dwóch faz artykulacyjnych — w pierwszej dochodzi do blokady przepływu przez jamę ustną i nosową (zwarcie), po czym, w kolejnym etapie, tworzy się szczelina, która umożliwia powstanie tarcia. Szum, który towarzyszy głoskom szczelinowym narasta stopniowo, podczas gdy w drugim przypadku pojawia się gwałtownie [150, 164]. W ostatnim podziale, tj. pod kątem miejsca artykulacji (miejsca tworzenia się szczeliny), głoski dentalizowane grupuje się w trzy zbiory: przedniojęzykowo-zębowe (/s, z, ts, dz/; szczelina tworzy się między czubkiem języka a górnymi zębami), przedniojęzykowo-dziąsłowe (/ʂ, ʐ, tʂ, dʐ/; szczelina tworzy się między czubkiem języka a górnymi dziąsłami; przykład zaprezen-

(a) **(b)**

Rys. 1.2: Ilustracja artykulacji głoski /ʂ/ w słowie „*Leszek*": (a) normatywna realizacja, w której szczelina tworzy się między końcówką języka a wałem dziąsłowym; (b) przykład jednej z nienormatywnych realizacji (artykulacja laminalna). Rysunek opracowano na podstawie [157].

towano na rys. 1.2), środkowojęzykowo-prepalatalne (/ɕ, ʑ, t͡ɕ, d͡ʑ/; szczelina tworzy się poprzez uwypuklenie środka języka ku podniebieniu twardemu w kierunku przednim). Rodzaj szczeliny warunkuje charakter wysokoczęstotliwościowego szumu, który towarzyszy sybilantom i różnicuje strumień wypływającego powietrza [138, 140, 146, 150].

Jak wspomniano wcześniej, logopeda poddaje ocenie wiele aspektów dotyczących mowy, anatomii i fizjologii badanego. Pojęcie cech artykulacyjnych głoski wykorzystywane w niniejszej pracy obejmuje cechy opisujące stan i sprawność narządów artykulacyjnych, ocenę słuchu fonemowego oraz cechy dotyczące realizacji poszczególnych głosek [112, 113, 138]. Do cech związanych z motoryką i budową artykulatorów zaliczyć można np. stopień skrócenia wędzidełka językowego, stan uzębienia, zgryz, sprawność stawu skroniowo-żuchwowego, budowę podniebienia oraz jamy nosowej. Cechy związane z fizjologią badanego obejmują ocenę słuchu fonemowego i fizycznego, a także funkcje oddychania i połykania. Ostatnia z grup dotyczy opisu cech związanych z realizacją poszczególnych głosek, głównie dotyczących sposobu oraz miejsc artykulacji i położenia artykulatorów. Są one zależne od realizowanej głoski. Przykładowe cechy, które wyróżnia się dla głosek dentalizowanych umieszczono w tab. 1.2 [112, 113].

1.2 Nienormatywna realizacja sybilantów

Nienormatywna realizacja głosek dentalizowanych jest jedną z najczęściej występujących wad w obrębie dyslalii [138, 140, 146, 150]. Seplenienie może przyjąć różne formy [56]:

Tab. 1.2: Zbiór przykładowych cech artykulacyjnych związanych z realizacją głosek dentalizowanych [112, 113].

Cecha normatywnej artykulacji	**Głoski**	**Opis cechy**
Dentalność	/s/, /z/, /ts/, /dz/	kontakt przedniej części języka z górnymi siekaczami
Dentalizacja	wszystkie sybilanty	dodatkowe przewężenie w postaci szczeliny zgryzowej
Postdentalność	/s̪/, /z̪/, /ts̪/, /dz̪/	zazębowy kontakt przedniej części języka
Apikalność	wszystkie sybilanty	kontakt wierzchołkowej części języka
Dorsalność	/s/, /ɕ/, /ʑ/, /tɕ/, /dʑ/	kontakt grzbietowej powierzchni języka
Sonorność	/z/, /dz/, /ʑ/, /dʑ/	drgający ruch wiązadeł głosowych
Nonsonorność	/s/, /ts/, /ʂ/, /tʂ/, /ɕ/, /tɕ/	niedrgający ruch wiązadeł głosowych
Palatalność	/ɕ/, /ʑ/, /tɕ/, /dʑ/	itowy układ języka
Nonpalatalność	/s/, /z/, /ts/, /dz/, /ʂ/, /z̻/, /tʂ/, /dz̻/	nieitowy układ języka
Nonnazalność	wszystkie sybilanty	wyłączenie rezonatora nosowego
Frykatywność	/s/, /z/, /ʂ/, /z̻/, /ɕ/, /ʑ/	trące pokonanie kontaktu typu szczelinowego
Afrykatywność	/ts/, /dz/, /tʂ/, /dz̻/, /tɕ/, /dʑ/	przytarte pokonanie kontaktu typu zwartego
Medialność języka	wszystkie sybilanty	kontakt w części ustnej nasady w linii pośrodkowej

- deformacje (sygmatyzm właściwy), które polegają na niepoprawnej artykulacji sybilantów i są skutkiem zmiany miejsca artykulacji;
- substytucje (parasygmatyzm), czyli zastępowanie wybranych głosek dentalizowanych innymi, które realizowane są normatywnie (jest to zjawisko, które, w trakcie rozwoju mowy i do pewnego momentu, jest normą rozwojową);
- elizje (mogisygmatyzm), który polega na opuszczaniu dźwięków (w czasie rozwojowym również jest dopuszczalne).

W ramach sygmatyzmu właściwego rozróżnia się jeszcze bardziej szczegółowe podgrupy (np. sygmatyzm międzyzębowy, przyzębowy, wargowo-zębowy, boczny). Zdarza się również, że u pacjenta logopedycznego sygmatyzm jest złożony i więcej cech artykulacyjnych równocześnie wykazuje nienormatywność [138, 140, 146].

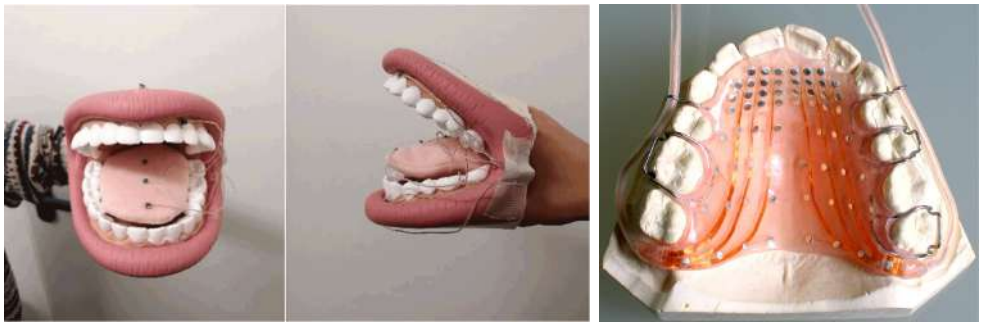
Wśród czynników, które sprzyjają występowaniu niepoprawnej realizacji głosek dentalizowanych można wyróżnić wiele aspektów fizjologicznych i anatomicznych [138, 140, 146]. Seplenieniu sprzyja nieprawidłowa budowa narządów artykulacyjnych, głównie języka — może być zbyt duży, zbyt gruby lub

charakteryzować się krótkim wędzidełkiem podjęzykowym [56]. Istotne w tym zakresie są również występujące wady zgryzu, zwłaszcza te, które prowadzą do zaniku dentalizacji (przede wszystkim zgryz otwarty, ale również przodozgryz i tyłozgryz), anomalie zębowe czy rozszczep podniebienia (skutkuje niedostatecznym zamknięciem jamy nosowej przez podniebienie miękkie; seplenienie nosowe). Kolejnym istotnym czynnikiem jest niska sprawność narządów artykulacyjnych, zwłaszcza języka. Brak pionizacji języka jest przyczyną infantylnego połykania, a to z kolei może prowadzić do wymowy międzyzębowej. Ważne są również kwestie związane ze słuchem, w tym upośledzenie słuchu w zakresie tonów wysokich (prowadzi do niedostatecznego rozróżniania głosek dentalizowanych) czy obniżenie słyszalności w okresie rozwoju mowy. Nie bez znaczenia pozostają doświadczenia środowiskowe, w tym naśladowanie nieprawidłowych wzorców i wadliwych artykulacji, ale także zbyt długie używanie smoczka przez dziecko (lub stosowanie niewłaściwych smoczków) oraz ssanie palca, które mogą być przyczyną wad zgryzu i niemowlęcej pozycji języka [56, 138, 140].

1.3 Komputerowa analiza wymowy

Komputerowa analiza wymowy jest zagadnieniem szerokim, zarówno pod kątem przeznaczenia takich rozwiązań, wykorzystywanych danych, jak i stosowanych metod. Przeważająca część dotychczasowych rozwiązań skupiała się na analizie wymowy normatywnej — m.in. nauce języków obcych, rozpoznawaniu mowy, identyfikacji mówcy, ale też rozpoznawaniu i klasyfikacji pojedynczych głosek. Systemy komputerowe wspomagające rozwój kompetencji lingwistycznych (ang. *computer-assisted language learning*, CALL; ang. *computer-aided pronunciation training*, CAPT) to przede wszystkim e-learningowe platformy oferujące użytkownikom ćwiczenia wykorzystujące materiał multimedialny [86, 92]. Większość rozwiązań CALL nie posiada rozbudowanej analizy sygnałów akustycznych wymowy, stąd w niewielu przypadkach spotkać się można z informacją zwrotną dotyczącą jej poprawności. Ocena umiejętności odbywa się głównie przez analizę ćwiczeń niewymagających wypowiadania żadnych zwrotów. Z kolei narzędzia CAPT często oferują wskazówki korygujące wymowę, m.in. wykorzystując narzędzia rozpoznawania mowy, są jednak z założenia opracowywane do nauki języka przez nienatywnych mówców. Błąd w takim kontekście nie jest tożsamy z wadą wymowy — źródłem wymowy nienormatywnej są nieprawidłowości w budowie i funkcjonowaniu narządów artykulacyjnych, podczas gdy błędom w trakcie nauki języka takie zaburzenia towarzyszyć nie muszą [44].

Narzędzia rozpoznawania mowy zyskują zresztą coraz większą popularność i użyteczność, nie tylko w opisywanych rozwiązaniach, ale także jako ułatwienie wielu czynności życia codziennego i zawodowego. Na przykład, automatyczne



(a) **(b)**

Rys. 1.3: Przykłady urządzeń do rejestracji artykulacji: (a) rozłożenie czujników do artykulografii elektromagnetycznej na fantomie [121], (b) nakładka z elektrodami do elektropalatografii [4].

rozpoznawanie mowy (ang. *automated speech recognition*, ASR), bez jej szczegółowej analizy pod kątem logopedycznym, znajduje zastosowanie w generowaniu raportów medycznych [156] czy terapii afazji [55]. W literaturze istnieje niewiele doniesień o wykorzystaniu ASR do diagnozy i terapii sygmatyzmu u dzieci. Wśród znalezionych zastosowań metody jest platforma zawierająca zestaw gier multimedialnych przeznaczona do wsparcia terapii sepleniących dzieci [93]. Niemniej są to przede wszystkim narzędzia, które nie zwracają informacji diagnostycznej i mogą stanowić wyłącznie dodatkowe wsparcie terapeutyczne.

1.3.1 Systemy rejestracji danych

Rozwiązania z założenia ściśle ukierunkowane na wsparcie diagnostyki i terapii logopedycznej stanowią dużo węższe spektrum. Niektóre z dostępnych koncepcji charakteryzują się wysoką dokładnością przestrzenną i czasową przy jednoczesnej inwazyjności, znacznych wymaganiach lub kosztach eksperymentalnych. Dotyczy to m.in. artykulografii elektromagnetycznej [61, 163], wykorzystywanej do obserwacji artykulatorów w zmiennym polu magnetycznym (rys. 1.3a), czy elektropalatografii [165], która monitoruje kontakt języka z podniebieniem w trakcie wymowy (rys. 1.3b). W przypadku artykulografii elektromagnetycznej, istnieją pojedyncze udostępnione bazy danych [163]. Obie techniki nie są całkowicie bezkontaktowe i wymagają ingerencji w jamę ustną badanego. Może to dyskwalifikować ich wykorzystanie do obserwacji artykulacji dzieci w wieku przedszkolnym. Istotne jest zatem poszukiwanie sposobów rejestracji mowy, które będą zapewniać komfort badanego i charakteryzować się jak najmniejszą inwazyjnością.

Wielu badaczy wykorzystuje sygnał akustyczny zarejestrowany za pomocą jednego lub wielu mikrofonów w różnej konfiguracji [9, 66, 68, 85]. Literatura

z zakresu lingwistyki oraz fonetyki oferuje duży zasób informacji dotyczących akustyki głosek szczelinowych i zwarto-szczelinowych. Bazując na tej wiedzy, badacze analizują możliwości zastosowań sygnału akustycznego np. w automatycznym rozpoznawaniu głosek (choć do tej pory przede wszystkim normatywnej ich wymowy). Liczne badania skupiają się na poszukiwaniu parametrów akustycznych, które umożliwią rozróżnienie między poszczególnymi fonemami frykatywnymi [15, 118, 141, 176] — ze względu na specyfikę sybilantów, badania dotyczą zazwyczaj ograniczonego podzbioru dźwięków występujących w danym języku. Co więcej, stosunkowo niewielka liczba prac opisująca analizę akustyczną dotyczy mowy dziecięcej [66, 75, 99, 100, 109, 176].

Analiza głosek dentalizowanych w literaturze często opiera się na przetwarzaniu widma sygnału. W wielu pracach pojawia się wykorzystanie momentów widmowych [118, 141]. Badacze raportowali, że środek ciężkości widma jest przesunięty dla sybilantów w zależności od miejsca artykulacji [15]. Inną z grup parametrów akustycznych opisujących głoski dentalizowane są cechy związane z szumem tarcia. Część prac skupiła się na poszukiwaniu różnic między fonemami w częstotliwościach i amplitudach formantów szumowych, które pojawiają się w widmie powyżej 2 lub 3 kHz, inne wykorzystywały szerokość pasma szumu i jego dolną granicę, różnice energii w poszczególnych pasmach częstotliwości, czas trwania tarcia czy współczynniki cepstralne w pasmie szumu i stosunki formantów szumowych [99, 155, 183].

Autorka pracy nie znalazła jednak badań, które w analogiczny sposób wykorzystują potencjał danych obrazowych reprezentujących wymowę dzieci. Można przypuszczać, że niektóre błędne wzorce związane z ruchem lub ułożeniem narządów mogą być widoczne na nagraniach wideo. Zarejestrowany materiał składa się z serii kolejnych klatek (obrazów) pobranych z zadaną częstotliwością. Podobieństwa można doszukać się w pomocach wykorzystywanych przez logopedów. Specjaliści w trakcie terapii logopedycznej stosują zestawy fotografii (lub rysunków) prezentujących kolejne etapy wymawiania poszczególnych głosek nazywane labiogramami [146]. Plansze pomagają przede wszystkim ćwiczyć prawidłowe ułożenie narządów. Wykorzystywanie takich materiałów w praktyce sugeruje użyteczność budowania metod komputerowych bazujących na tej modalności.

Obserwowany w ostatnich latach wzrost popularności rozwiązań telemedycznych również wskazuje, że wykorzystanie danych wideo jest właściwym kierunkiem rozwoju. Prowadzone do tej pory eksperymenty dotyczyły synchronicznego i asynchronicznego podejścia do zdalnej terapii logopedycznej. Tryb synchroniczny obejmuje przede wszystkim wideokonferencje prowadzone pomiędzy terapeutą a pacjentem za pomocą laptopów i smartfonów. Zdalną terapię uważa się za porównywalnie skuteczną z podejściem tradycyjnym [28, 120]. Z kolei tryb asynchroniczny jest bardziej użyteczny w leczeniu niż procesie diagnostycznym. Wykorzystuje dedykowane platformy internetowe, aby zapewnić

pacjentom ćwiczenia przygotowane przez specjalistę na podstawie ich wcześniejszej diagnozy [1, 47]. W obu przypadkach istnieje duży potencjał gromadzenia danych do projektowania narzędzi do komputerowych metod diagnostyki mowy i słuchu, obejmujących wstępną ocenę wybranych zaburzeń [9, 67, 69, 125]. Ponadto logopedzi często korzystają z filmów nagranych smartfonem podczas badań indywidualnych, aby udokumentować swoje ustalenia, umożliwić późniejszą weryfikację lub monitorować postęp terapii. Mimo że takie nagrania mogą być niestabilne lub mieć niską jakość, nadal można je uznać za pomocne w zautomatyzowanym wsparciu diagnostycznym.

1.3.2 Segmentacja narządów mowy

Etapem pośrednim w budowaniu rozwiązań wykorzystujących dane obrazowe często jest segmentacja wybranych narządów. Metody obserwacji i analizy obrazów twarzy obecne są nie tylko w literaturze naukowej, ale i wielu rozwiązaniach komercyjnych. Zakres tematyki obejmuje m.in. automatyczne rozpoznawanie mowy na podstawie ruchu warg, systemy komputerowego wsparcia stomatologii oraz rehabilitację i monitorowanie stanu zdrowia. Zaledwie kilka prac znalezionych przez autorkę tej rozprawy dotyczy diagnostyki i terapii mowy. Badacze podejmujący tematykę segmentacji ust i warg proponowali rozwiązania bazujące na różnych zaawansowanych metodach, np. grupowaniu z wykorzystaniem algorytmu maksymalizacji oczekiwań [89], klasteryzacji rozmytej opartej na kształcie [72, 162] lub wykrywaniu krawędzi falką wieloskalową [46].

W ciągu ostatnich kilku lat zaproponowano badania o istotnym wkładzie w rozwój tematyki dotyczącej segmentacji ust/warg, głównie wykorzystując koncepcje głębokiego uczenia (ang. *deep learning*), a zwłaszcza konwolucyjnych (splotowych) sieci neuronowych (ang. *convolutional neural network*, CNN). Choć żadne z nich nie było opracowywane z bezpośrednim założeniem wsparcia logopedii, to ich przedmiot jest w pewnym zakresie spójny z tematyką niniejszej rozprawy. Wśród proponowanych rozwiązań raportowano na przykład wykorzystanie głębokiej sieci CNN do podziału obrazu termowizyjnego twarzy w podczerwieni na dziewięć klas, w tym ust [104]. Struktura modelu opierała się na wieloklasowej sieci U-Net zamkniętej w warunkowej generatywnej sieci przeciwstawnej (ang. *conditional generative adversarial nets*, cGAN). Średni współczynnik podobieństwa Jaccarda (ang. *intersection over union*, IoU) dla segmentacji ust wyniósł tam 0,7, który odpowiada wartości ok. 0,82 współczynnika podobieństwa Dice'a (ang. *Dice similarity coefficient*, DSC). Dwa inne badania skupiały się na rozpoznawaniu słów na podstawie nagrań wideo prezentujących ruch warg. Zaproponowano na przykład wielostopniowy detektor zmian ułożenia ust jako część systemu do klasyfikacji słów w języku amharskim [12]. Do klasyfikacji wykorzystano metody uczenia maszynowego, jednak sama segmentacja opierała się na bardziej konwencjonalnych metodach: wykrywaniu

obiektów w oparciu o algorytm Viola-Jones, wzmocnienie kontrastu, wykrywanie krawędzi Sobela i progowanie. Ogólna dokładność uzyskana w zbiorze 14 słów osiągnęła 66,43% przy użyciu maszyny wektorów nośnych (ang. *support vector machine*, SVM), badanie jednak nie raportuje skuteczności segmentacji ust. W kolejnej pozycji [97] opisano koncepcję czasoprzestrzennego modelu segmentacji warg w sekwencji klatek wideo przy użyciu kombinacji sieci CNN i dwukierunkowych bramkowanych jednostek rekurencyjnych (ang. *bidirectional GRU*, Bi-GRU). Podejście zakładało wstępne przetwarzanie w celu znalezienia zgrubnej lokalizacji ust i warg z wykorzystaniem kaskady Haara oraz algorytmu hybrydowych aktywnych konturów. Autorzy skupili się na ilościowych wynikach rozpoznawania słów (dokładność ponad 90% w publicznej bazie danych zawierającej fragmenty nagrań telewizyjnych), ograniczając ocenę segmentacji obiektów do nielicznych ilustracji. Kolejnym proponowanym podejściem było porównanie dziewięciu nowoczesnych sieci CNN nauczonych na dwóch licznych bazach danych [23]. Spośród proponowanych modeli najwyższą skuteczność wykazała architektura Mobile DeepLabV3, a w zależności od zbioru testowego wartość DSC w przypadku segmentacji warg oscylowała wokół 0,93.

Lista prac, które dotyczą przetwarzania obrazów barwnych w celu wyodrębnienia innych narządów mowy, np. zębów i języka, jest jeszcze krótsza. Analiza zębów jest w tej dziedzinie prawie nieobecna. Stanowi jedynie niewielką część bardziej ogólnych badań obrazu twarzy [70, 180], jednak praktycznie bez wiarygodnej oceny numerycznej. Segmentację języka, z kolei, często przeprowadza się w specyficznym obszarze, gdzie jego ułożenie prawie w ogóle nie przypomina ruchu podczas swobodnej mowy — np. w rozwiązaniach tradycyjnej medycyny chińskiej [52, 79, 178, 179]. W przypadku tych badań, większość obrazów w swojej centralnej części przedstawia duży obszar języka wysuniętego na brodę, a wartości IoU segmentacji osiągają wartości aż 0,97 [52, 178]. Nie ma to jednak przełożenia na zagadnienia logopedyczne — podczas wymowy język jest przeważnie częściowo przysłonięty zębami lub wargami, widoczny w znacznie mniejszym stopniu. Spośród pojedynczych prac dotyczących bezpośrednio segmentacji artykulatorów dla celów wsparcia terapii logopedycznej, obiecujące wydają się badania, w których zaproponowano segmentację narządów mowy z wykorzystaniem modeli głębokiego uczenia (U-Net) [10, 11]. Opisane metody nie zostały jednak ocenione ilościowo, trudno zatem odwoływać się lub porównywać do uzyskanych rezultatów. Autorka niniejszej rozprawy również podjęła się próby segmentacji języka w ramach wstępnych eksperymentów ukierunkowanych na rozwój metodyki opisanej w pracy [126]. Pilotażowa próba analizy języka była przeprowadzana w dedykowanej konfiguracji, przy użyciu dobrze oświetlonych obrazów o wysokiej rozdzielczości. Przygotowany wtedy model prostej sieci CNN zwrócił wyniki świadczące o ich dużej zależności od rozmiaru obiektu, przy średniej wartości IoU sięgającej 0,74.

Przegląd literatury w zakresie tematyki niniejszej rozprawy sugeruje, że problem wad wymowy wśród dzieci jest powszechny i ze względu na swoje skutki nie powinien być bagatelizowany. Mimo stałego wdrażania rozwiązań komputerowych usprawniających proces diagnostyczny, który obserwowany jest w ostatnich latach, istnieje wciąż potrzeba rozwoju w tym kierunku. Analiza wad wymowy, zwłaszcza wśród dzieci, wymaga poszukiwania bezkontaktowych i komfortowych sposobów rejestracji danych, szczególnie zakładając, że mowa w trakcie pomiarów powinna być możliwie najbardziej swobodna. Można przypuszczać, że przetwarzanie nagrań wideo zarejestrowanych w trakcie artykulacji przyniesie dodatkowe informacje diagnostyczne związane z ruchem i ułożeniem poszczególnych narządów. Co więcej, hybrydyzacja wskaźników bazujących na danych obrazowych i danych dźwiękowych (zarejestrowanych w tym samym czasie) może poszerzyć informacje diagnostyczne i wzajemnie się uzupełniać. Przed wdrożeniem rozwiązań do wykorzystania praktycznego konieczne jest jednak przeprowadzenie badań wstępnych, które skupią się na poszukiwaniu istotnych statystycznie różnic w parametrach uzyskanych na podstawie sygnałów akustycznych lub danych obrazowych prezentujących różne cechy wymowy sybilantów. Przegląd literatury i aktualnych rozwiązań sugeruje również brak rzetelnych (popartych metrykami jakościowymi) metod pośrednich, m.in. algorytmów segmentacji narządów mowy, które mogłyby zostać wykorzystane do dalszych kroków analizy wad wymowy. W tym aspekcie zasadne wydaje się wykorzystanie metod sztucznej inteligencji, zwłaszcza głębokiego uczenia, które wykazują skuteczność w przypadku dużych zbiorów danych (zwłaszcza obrazowych) i dążą do generalizacji problemu.

2. Zakres pracy

2.1 Tezy i cel pracy

Na podstawie analizy literatury dotyczącej specyfiki realizacji sybilantów oraz dostępnych rozwiązań w zakresie komputerowego wsparcia diagnostyki wad wymowy, w których brakuje raportowania zależności między parametrami reprezentacji wizualnej i akustycznej artykulacji a cechami wymowy fonemów dentalizowanych, sformułowano następującą **tezę główną**:

> **Istnieją istotne statystycznie różnice w cechach sygnałów akustycznych i danych obrazowych prezentujących mowę dzieci z różnymi (normatywnymi i nienormatywnymi) cechami wymowy.**

Aby zweryfikować tezę główną, sformułowano również **tezy pomocnicze**:

> **Teza pomocnicza nr 1:** Możliwa jest wiarygodna segmentacja wybranych artykulatorów w obrazach twarzy z wykorzystaniem metod sztucznej inteligencji.
>
> **Teza pomocnicza nr 2:** Ekstrakcja i analiza cech obrazowych 2D i 3D oraz parametrów akustycznych pozwala na określenie różnic między grupami w wybranych cechach artykulacyjnych.

Potwierdzenie tezy głównej będzie możliwe dzięki

> opracowaniu metodyki przetwarzania sygnałów akustycznych i danych obrazowych z wykorzystaniem metod sztucznej inteligencji,

które stanowi **cel niniejszej rozprawy**.

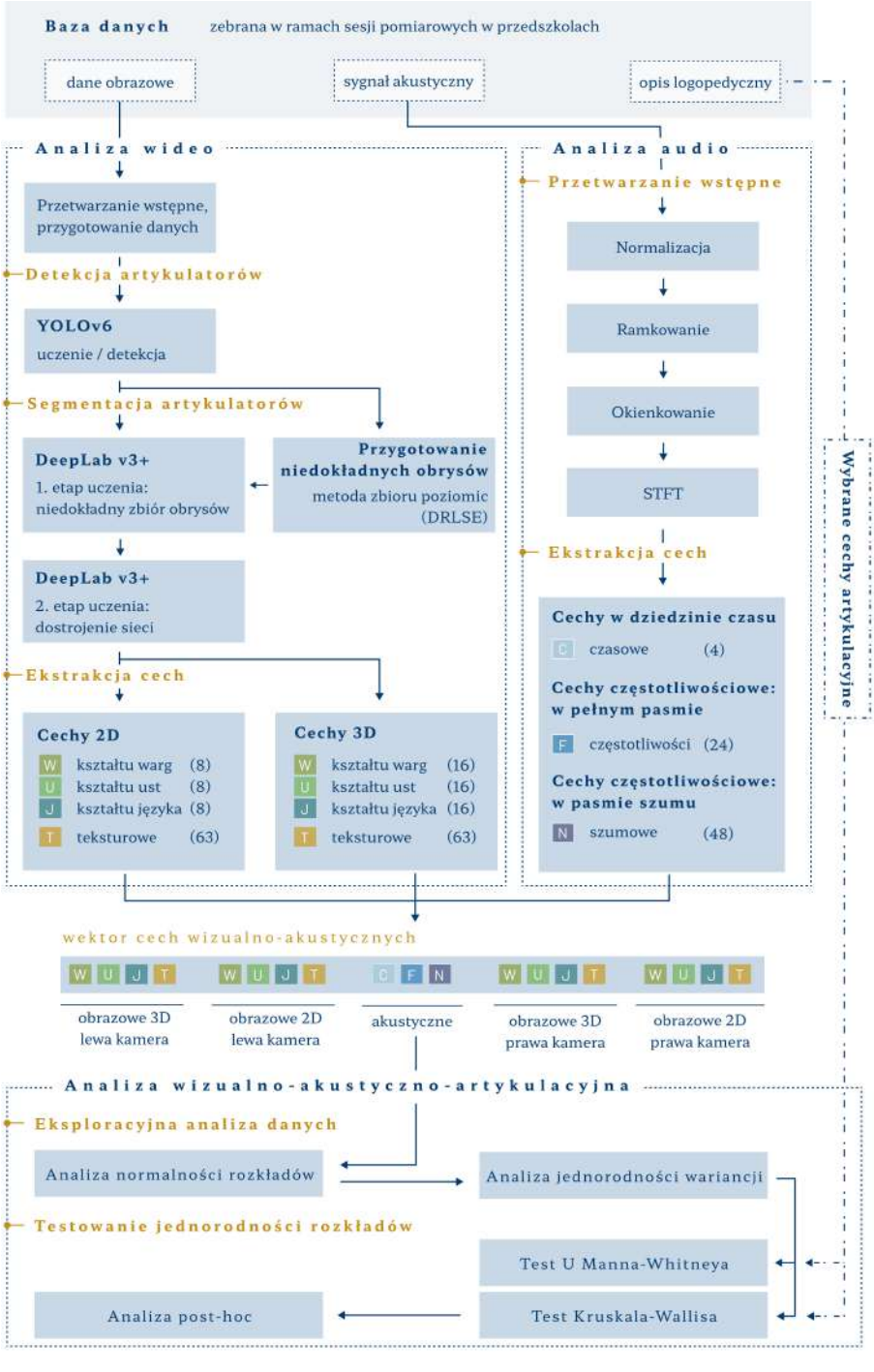
Elementy autorskie proponowanego podejścia obejmują:

- opracowanie dwuetapowej, automatycznej metody segmentacji artykulatorów (warg, ust, zębów, języka), która bazuje na konwolucyjnych sieciach neuronowych do detekcji obiektów oraz segmentacji;
- w ramach drugiego kroku automatycznej metody segmentacji, opracowanie metodyki dwuetapowego uczenia sieci słabo nadzorowanej: (1) w oparciu o liczny zbiór niedokładnych obrysów przygotowanych automatycznie za pomocą wstępnej metody, która bazuje na analizie zbiorów poziomic i obliczeniach rozmytych oraz (2) dostrojenie sieci za pomocą niewielkiego zbioru eksperckiego;
- hybrydyzację danych obrazowych i dźwiękowych w celu poszukiwania zależności między parametrami wizualnymi i akustycznymi a opisem logopedycznym;
- propozycję ekstrakcji cech obrazowych bazujących na trójwymiarowych wolumenach, w których trzeci wymiar jest związany z czasem.

2.2 Opis metodyki pracy

W celu zweryfikowania słuszności postawionej tezy opracowano metodykę przetwarzania sygnałów akustycznych i danych obrazowych mowy dziecięcej (rys. 2.1). Pierwszym działaniem było zgromadzenie materiału w ramach sesji pomiarowych w kilku przedszkolach. W efekcie zebrano bazę danych obejmującą nagrania 201 dzieci: dane obrazowe oraz sygnał akustyczny zarejestrowane jednocześnie w trakcie wymowy oraz, dla każdego dziecka, opis diagnostyczny sporządzony przez dwóch niezależnych logopedów.

Kolejny etap dotyczył opracowania ścieżki przetwarzania danych obrazowych i sygnału dźwiękowego w celu przygotowania do ekstrakcji cech. Działania na obu typach danych wykonano niezależnie i równolegle, uprzednio je synchronizując. W ramach analizy wideo opracowano dwuetapową metodę prowadzącą do segmentacji artykulatorów (warg, ust, zębów, języka) za pomocą metod głębokiego uczenia. Pierwszy krok stanowiła detekcja z wykorzystaniem konwolucyjnych sieci neuronowych typu YOLO (ang. *you only look once*). Drugi etap dotyczył segmentacji obiektów stosując sieć DeepLabv3+. Ponieważ wykonanie obrysów eksperckich dużych zbiorów danych jest zadaniem czasochłonnym, przygotowano metodę słabo nadzorowanego uczenia sieci, która wykorzystywała trzystopniowy proces. W pierwszym kroku, na podstawie informacji o prostokątach okalających artykulatory uzyskanych siecią YOLO oraz obszarach ust, przygotowano duży zbiór zgrubnie wysegmentowanych danych za pomocą metody zbioru poziomic oraz z zastosowaniem rozmycia obrazów. Wygenerowane dane posłużyły do wstępnego uczenia sieci DeepLabv3+. Ostatnim



Rys. 2.1: Schemat blokowy metodyki opisanej w pracy.

krokiem było dostrojenie sieci za pomocą niewielkiego zbioru eksperckiego. Po zweryfikowaniu jakości działania metody możliwa była ekstrakcja cech obrazowych. Obliczano dwa typy cech: dwuwymiarowe i trójwymiarowe. W przypadku pierwszego rodzaju (2D), dla każdej ramki nagrania, która obejmowała wymowę głoski, uzyskiwano wektor 24 cech kształtu (po 8 opisujących wargi, usta i język) oraz 63 cechy teksturowe. Dla drugiego rodzaju (3D), najpierw generowano model 3D, który był złożeniem kolejnych ramek w czasie, i na tej podstawie obliczano wektor: 48 cech kształtu (po 16 dla warg, ust i języka) oraz 63 cechy teksturowe.

Sygnał audio tego samego fragmentu nagrania był przetwarzany w odmienny sposób. W pierwszym kroku był poddany przetwarzaniu wstępnemu, które obejmowało normalizację, ramkowanie, okienkowanie oraz obliczenie krótkoczasowej transformaty Fouriera. Następnie ekstrahowano zestaw cech dla każdej ramki sygnału: 4 cechy w dziedzinie czasu, 24 cechy częstotliwościowe w pełnym pasmie i 48 parametrów częstotliwościowych w pasmie szumu.

Cech obrazowe z lewej i prawej kamery oraz parametry audio zostały następnie poddane konkatenacji. Materiał wykorzystano do przeprowadzenia testów statystycznych, które miały na celu zbadanie występowania istotnych różnic między grupami w wybranych cechach artykulacyjnych (będących częścią opisu logopedycznego) na podstawie parametrów sygnału akustycznego i danych wizualnych.

2.3 Układ rozprawy

Rozprawę ułożono w dziewięć rozdziałów oraz dwa dodatki. Rozdział 1 przybliża podstawy teoretyczne artykulacji głosek, wskazuje na powszechność problemu seplenienia i zawiera przegląd aktualnych rozwiązań z zakresu komputerowego wsparcia logopedii. Bieżący rozdział opisuje sformułowane tezy i cele pracy oraz pokrótce przybliża zakres opracowanej metodyki. Rozdział 3 obejmuje informacje dotyczące gromadzenia materiału w ramach sesji pomiarowych oraz danych wykorzystanych do realizacji poszczególnych etapów dalszego przetwarzania. Rozdziały 4 i 5 szczegółowo prezentują ścieżki przetwarzania, odpowiednio, danych obrazowych i sygnałów akustycznych. Każde z istotnych zagadnień rozpoczyna się jest od wprowadzenia teoretycznego. W rozdziale 6 przybliżono szczegóły wykorzystanej metodyki analizy wizualno-akustyczno-artykulacyjnej. Rozdział 7 raportuje wyniki otrzymane we wszystkich przeprowadzonych eksperymentach i testach, z kolei w rozdziale 8 skomentowano rezultaty i przeprowadzono dyskusję na temat zagadnień podjętych w pracy oraz wyciągnięto wnioski. Rozwinięciem rozdziału dotyczącego wyników są dodatki A i B, które zawierają tabele ze szczegółowymi wynikami analiz. Rozdział 9 podsumowuje wnioski i opisuje możliwe kierunki dalszego rozwoju.

3. Materiały

Założeniem pracy była analiza artykulacji głosek dentalizowanych dzieci w wieku przedszkolnym. W oparciu o przegląd literatury oraz dostępnych rozwiązań stwierdzono brak adekwatnej bazy danych dla języka polskiego. Z tego względu, przed przystąpieniem do realizacji opracowanej metodyki, konieczne było przygotowanie bazy wielomodalnych danych wraz z odpowiadającymi im opisami logopedycznymi.

3.1 Projekt badawczy „Hybrydowy system akwizycji i przetwarzania sygnału wielomodalnego w analizie sygmatyzmu u dzieci"

Prace opisywane w dysertacji były realizowane w ramach stypendium doktoranckiego w projekcie nr 2018/30/E/ST7/00525 o tytule: „Hybrydowy system akwizycji i przetwarzania sygnału wielomodalnego w analizie sygmatyzmu u dzieci", finansowanym ze środków Narodowego Centrum Nauki (konkurs: SONATA BIS 8) w latach 2019–2024. Organizacja pomiarów badawczych dzieci w placówkach przedszkolnych i szkolnych uzyskała pozytywną opinię komisji bioetycznej (Uchwała nr 3/2021 Uczelnianej Komisji Bioetycznej ds. Badań Naukowych przy Akademii Wychowania Fizycznego im. Jerzego Kukuczki w Katowicach). Przed dopuszczeniem dziecka do udziału w badaniach konieczna była zgoda rodziców lub opiekunów prawnych oraz słowne wyrażenie woli udziału przez samego badanego.

Interdyscyplinarny zespół badawczy, w skład którego wchodzili inżynierowie biomedyczni (w tym autorka niniejszej rozprawy) oraz specjaliści logopedii, przeprowadził rejestrację danych w sześciu placówkach przedszkolnych i szkolnych w okresie od października 2021 r. do czerwca 2023 r. Badania objęły 201 dzieci w wieku 4–8 lat (tab. 3.1). Największą grupę zarejestrowanych stanowiły dzieci w szóstym oraz siódmym roku życia (odpowiednio 60 i 96).

Badanie składało się z dwóch etapów: w pierwszym kroku, specjalnie zaprojektowane urządzenie pomiarowe rejestrowało swobodną mowę dziecka oraz przebieg ćwiczeń logopedycznych, natomiast część druga moderowana była przez logopedę i polegała na badaniu poprawności wymowy według uprzednio opracowanego protokołu [158]. Wynikiem wymienionych etapów były kolejno: dane

Tab. 3.1: Zestawienie wieku i płci osób badanych.

Wiek	Średnia (miesiąc życia)	Liczba dzieci		
		Dziewczęta	Chłopcy	Łącznie
5. rok życia	58.7 ± 0.6	1	2	3
6. rok życia	66.6 ± 3.4	37	23	60
7. rok życia	77.7 ± 3.6	48	48	96
8. rok życia	87.9 ± 3.5	21	21	42
Łącznie	76.2 ± 8.6	107	94	201

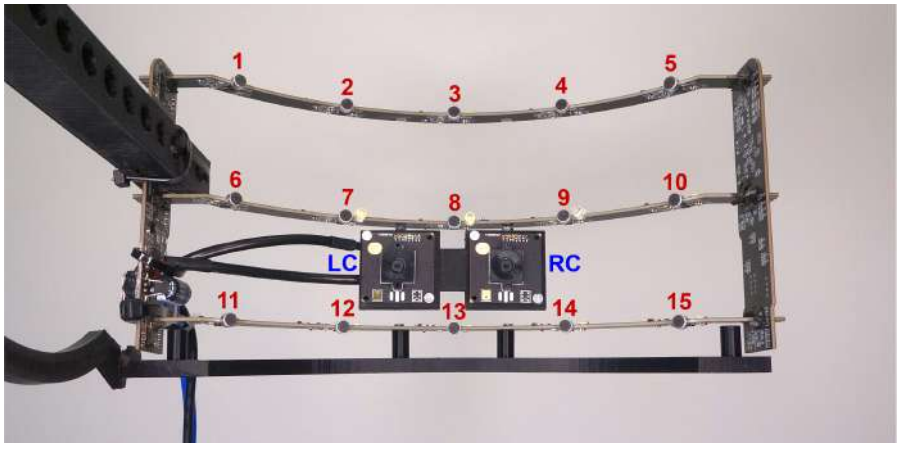
wielomodalne (przestrzenne dane akustyczne oraz stereowizyjne dane wizualne) oraz opis ekspercki (diagnoza) przypadku. Dodatkowo, dziecko było badane w innym terminie przez kolejnego logopedę zgodnie z protokołem drugiego etapu badania, ponieważ jednym z założeń projektu było pozyskanie dwóch niezależnych diagnoz.

3.2 Urządzenie pomiarowe

Dane poddane analizie zostały zebrane z wykorzystaniem wielomodalnego urządzenia pomiarowego (maski akustyczno-wizyjnej). Aparatura była wynikiem prac zespołu projektowego [65, 67], a główną rolą autorki niniejszej pracy było dostosowywanie parametrów kamer i sceny. Narzędzie pozwala na rejestrację sygnału akustycznego w 15 przestrzennie rozłożonych kanałach (półcylindryczna macierz mikrofonów) oraz stereowizyjnych danych wideo uzyskanych za pomocą dwóch kamer (rys. 3.1). Konstrukcję zaadaptowano do rejestracji danych pochodzących od dzieci w wieku przedszkolnym. Uwzględniono w tym zakresie budowę oraz wagę stelaża, wykonanego technikami druku 3D i przypominającego dziecięcy kask rowerowy. Wnętrze nakrycia zostało wyłożone miękkimi gąbkami zwiększającymi komfort użytkowania. Zapewniono łatwość montażu oraz możliwość dostosowywania pozycji części ruchomych i odległości sensorów od źródła dźwięku. Mimo mobilności w tym zakresie, urządzenie cechuje się w trakcie pomiarów mechaniczną stabilnością względem źródła dźwięku i obiektu obrazowania. Konstrukcja gwarantuje również bezpieczeństwo oraz higienę użytkowania. Całość uatrakcyjniono kolorowymi elementami np. pozorującymi uszy królika lub pióropusz.

Urządzenie pomiarowe składa się z jednostki centralnej (zasilanej napięciem 5 V) oraz trzech łuków rejestrujących. Główne parametry techniczne opisujące narzędzie zebrano w tab. 3.2. Jednostka centralna wykorzystuje interfejs USB do komunikacji z komputerem, natomiast przesył danych z łuków rejestrujących wymaga interfejsu SPI (ang. *serial peripheral interface*). Dwie płytki drukowane PCB (ang. *printed circuit board*) z jednostką centralną sta-

(a) **(b)**



(c)

Rys. 3.1: Budowa urządzenia pomiarowego składającego się ze stelaża oraz łuku rejestracyjnego wyposażonego w macierz mikrofonów oraz dwie kamery: (a) pierwotna wersja maski o konstrukcji zamkniętej, (b) zmodyfikowana wersja urządzenia o konstrukcji otwartej, (c) struktura wewnętrzna urządzenia (czerwone liczby odpowiadają numerom kolejnych mikrofonów, a LC i RC odnoszą się, odpowiednio, do lewej i prawej kamery) [67].

nowią podporę dla łuków. Pojedynczy łuk wyposażony jest w pięć mikrofonów Panasonic WM-61a [116], poprzedzonych przedwzmacniaczem TS472 oraz wzmacniaczem TLV6741 [153]. Całość tworzy półcylindryczną macierz o wymiarach 3×5 z 5-centymetrowymi odległościami pomiędzy sąsiednimi mikrofonami. Narzędzie rejestruje 16-bitowy sygnał akustyczny zsynchronizowany w czasie z częstotliwością próbkowania 44,1 kHz. Pomiędzy dwoma dolnymi łukami rejestrującymi zamontowano parę kamer Arducam 8MP 1080P Auto

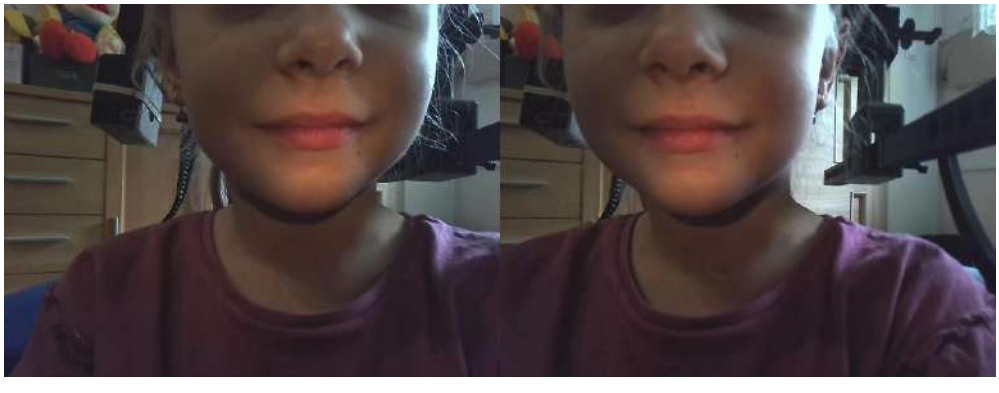
Tab. 3.2: Specyfikacja techniczna wielomodalnego urządzenia pomiarowego [65].

Urządzenie	
Liczba kanałów	15
Częstotliwość próbkowania	44,1 kHz
Liczba kamer	2
Mikrofon	
Model	Panasonic WM-61A
Rodzaj przetwornika	Elektretowy
Pasmo przenoszenia	20 Hz – 16 kHz
Poziom ciśnienia akustycznego	120 dB
SNR	62 dB
Czułość (1 kHz, 94 dB SPL)	-35 ± 4 dB
Kamera	
Model	Arducam 8MP 1080P Auto Focus
Rozdzielczość	640×480 VGA
Liczba klatek na sekundę	30

Focus. Dodano również oświetlenie LED w celu poprawy jakości nagrań wideo oraz rozświetlenia obszaru ust mówcy. Przyjęcie takiej konfiguracji pozwala na otrzymanie niezakłóconego obrazu pracy artykulatorów rejestrowanego z niewielkiej odległości (ok. 15 cm). Charakter urządzenia pozwala na wykonywanie ruchów głowy, zachowując jednocześnie stabilność położenia czujników względem narządów mowy.

Dane akustyczne wydają się być naturalnym źródłem informacji w diagnostyce logopedycznej. Stereowizyjne nagrania wideo również mogą stanowić istotny element wsparcia pracy logopedów, zarówno jako podstawa rozwoju algorytmów wspierających zautomatyzowaną diagnozę, jak i materiał archiwalny, do którego specjaliści mogą wracać w celu monitorowania progresu czy wątpliwości diagnostycznych. Koniecznym jest zwizualizowanie obszarów istotnych diagnostycznie — nie tylko samych narządów artykulacyjnych i wnętrza jamy ustnej, ale również żuchwy czy mięśni, które otaczają szparę ust. Kamery stereowizyjne na stelażu opisywanego urządzenia zamocowano tak, aby rejestrowały twarz mówcy ograniczoną do jej dolnego obszaru (od nosa do ramion). Oprócz rejestracji najistotniejszych regionów aktywnych w trakcie wymowy, takie rozwiązanie wspierało anonimizację danych poprzez pominięcie oczu dziecka. Przykładowe widoki z kamer przedstawiono na rys. 3.2. Gromadzony materiał charakteryzował się dużą różnorodnością wynikającą z naturalnych różnic w budowie anatomicznej pomiędzy mówcami oraz czynnikami zewnętrznymi (wśród których wymienić można m.in.: oświetlenie, cienie, zasłonięcie ust dłonią, kolejnymi wersjami urządzenia pomiarowego). Dodatkowym atutem była możliwość

(a)



(b)

Rys. 3.2: Widoki z dwóch kamer zarejestrowane dla przykładowych mówców — zestawienie uwypukla różnorodność danych oraz wpływ czynników zewnętrznych, jak oświetlenie sceny, na obraz.

dopasowywania widoku.

Użyteczność maski pomiarowej była na bieżąco poddawana walidacji wynikającej z przeprowadzania rejestracji w placówkach przedszkolnych. Pierwotna wersja urządzenia (rys. 3.1a) charakteryzowała się zamkniętą konstrukcją stelaża, który spajał kask i jednostkę centralną wraz z macierzą mikrofonów i kamerami. W kolejnym modelu (rys. 3.1b) zastosowano stelaż otwarty. Zmniejszono w ten sposób masę urządzenia, a ograniczona liczba elementów konstrukcyjnych podniosła poziom przyjaznego wrażenia wizualnego — obie kwestie są istotne dla rejestracji mówców w wieku przedszkolnym. Dodatkowo, na łuku zamontowano źródło światła, które miało rozjaśnić obszar artykulacji osoby badanej, zwłaszcza uwypuklić wnętrze jamy ustnej. Wsparło to możliwość ponownego odtworzenia materiałów przez specjalistów oraz zwiększyło precyzję zdalnej obserwacji i wnioskowania o poprawności wymowy, np. w przypadku wątpliwych diagnoz. Podkreślenie artykulatorów zewnętrznym źródłem świa-

(a) **(b)** **(c)**

Rys. 3.3: Dwuetapowy protokół pomiarowy: (3.3a-3.3b) rejestracja wymowy oraz ćwiczeń logopedycznych z wykorzystaniem maski, (3.3c) badanie logopedyczne według założonego protokołu.

tła było również znaczące dla opracowywanej metodyki przetwarzania obrazów poprzez zwiększenie kontrastu między artykulatorami i uwypuklenie narządów trudniej zauważalnych w trakcie mowy, zwłaszcza języka.

3.3 Protokół rejestracji danych wielomodalnych

Protokół badania był trzyetapowy, przy czym w dwóch pierwszych etapach wykorzystywano urządzenie wizyjno-akustycznego: (1) rejestracja swobodnej wymowy dziecka w trakcie nazywania grafik widocznych na monitorze (rys. 3.3a), (2) rejestracja zestawu ćwiczeń logopedycznych oraz wybranych słów i logotomów powtarzanych za logopedą (rys. 3.3b) oraz (3) badanie logopedyczne według opracowanego protokołu (rys. 3.3c). Zarówno materiał słowny, jego kolejność, jak i wykonywanie poleceń logopedy, były usystematyzowane i charakteryzowały się powtarzalnością względem mówców.

3.3.1 Materiał słowny

Materiał słowny zgromadzony w bazie danych składał się z 51 pojedynczych wyrazów (słów) oraz 12 jednosylabowych logotomów, zawierających poszczególne sybilanty, zorganizowane w trzech szeregach: /s/, /z/, /ʦ/, /dz/, /ʂ/, /ʐ/, /ʈʂ/, /dʐ/, /ɕ/, /ʑ/, /ʨ/, /dʑ/ (tab. 3.3; patrz także tab. 1.1). Słownik organizował wyizolowane wyrazy z głoskami dentalizowanymi w różnej konfi-

Tab. 3.3: Zbiór wyrazów z wyróżnionymi fonemami: niebieskie pola zawierają słowa prezentowane na monitorze (etap I badania), zielone i czerwone pola przedstawiają odpowiednio słowa i logotomy powtarzane przez dziecko za logopedą (etap II).

Szereg I (syczący)			
/s/	/z/	/ts/	/dz/
pie**s**	ko**z**a	**c**ebula	**dz**wonek
strażak	**z**egar	owo**c**e	sa**dz**awka
samolot	**z**abawki	widele**c**	**dz**a
sałata	ma**z**aki	ta**c**a	
para**s**ol	**z**a	paja**c**	
la**s**		**c**a	
cia**s**tka			
sadzawka			
sa			
Szereg II (szumiący)			
/ʂ/	/ʐ/	/tʂ/	/dʐ/
szafa	**ż**arówka	**cz**apka	**dż**okej
szufelka	**rz**eka	ka**cz**ka	ra**dż**a
kos**z**yk	je**ż**e	biega**cz**	**dż**a
kalo**sz**e	ró**ż**a	**cz**a	
nó**ż**	stra**ż**ak		
wą**ż**	**ż**yrafa		
książka	**ż**aba		
leka**rz**	war**z**ywa		
sznurek	**rz**a		
kucha**rz**			
szalik			
ka**sz**a			
sza			
Szereg III (ciszący)			
/ɕ/	/ʑ/	/tɕ/	/dʑ/
k**si**ążka	**zi**arno	**ci**astka	**dzi**adek
siatka	ba**zi**e	bo**ci**an	ło**dzi**e
w pa**si**e	**zi**a	łokie**ć**	**dzi**a
pa**ź**		**ci**a	
sia			

guracji, otoczeniu i w odmiennych fazach artykulacji: na początku, w środku oraz na końcu wyrażenia.

Znaczną część słów (38) wyświetlano w postaci graficznej na monitorze w pierwszym etapie badania — każdemu z poszczególnych słów odpowiadała

Tab. 3.4: Zestaw ćwiczeń logopedycznych.

Ćwiczenie logopedyczne
Swobodna pozycja warg.
Szeroki uśmiech z widocznymi zębami.
Powtarzanie głosek: /*i*/, /*u*/, /*a*/.
Powtarzanie sekwencji: /*i*-*u*/, /*i*-*a*/.
Kląskanie językiem.
Ustawienie języka w pozycji „kobry”.
Zakrywanie górnej wargi językiem.
Sięganie językiem do ostatniego dolnego zęba po lewej stronie.
Sięganie językiem do ostatniego dolnego zęba po prawej stronie.
Wysuwanie języka możliwie najdalej na brodę.
Przełknięcie śliny.

pojedyncza ilustracja. Zadaniem dziecka było nazwanie obiektu, który widzi. Wybrane słowa musiały być zarówno znane i jednoznaczne dla dziecka w wieku przedszkolnym, jak i łatwo reprezentowalne graficznie. Logotomy (12) oraz pozostałe słowa (13), ze względu na ich trudność i niejednoznaczność graficzną, były wypowiadane przez logopedę w drugim etapie badania. Zadaniem mówcy było powtórzenie usłyszanej frazy. Kolejność wyrazów była jednakowa we wszystkich pomiarach. Ze względu na to, że cztery słowa zawiera po dwa sybilanty (strażak, książka, sadzawka, ciastka), łączna liczba unikalnych wystąpień sybilantów w materiale słownym wynosi 67 (55 + 12).

3.3.2 Ćwiczenia logopedyczne

W drugiej części sesji pomiarowej rejestrowano zestaw ćwiczeń logopedycznych, które zostały zebrane w tab. 3.4. Logopeda zwięźle i jednoznacznie tłumaczył zadania, a w wybranych przypadkach demonstrował oczekiwane ruchy narządów. Kolejność ćwiczeń była określona w protokole i tym samym jednakowa dla wszystkich pomiarów. Celem etapu była obserwacja sprawności i wzorców ruchowych poszczególnych narządów artykulacyjnych, długości wędzidełka językowego oraz realizacji głosek w poszczególnych sylabach.

3.4 Materiały eksperckie

Opracowanie rzetelnych metod komputerowego wspomagania diagnostyki medycznej wymaga analizy jakościowej poprzez porównanie z opisami przygotowanymi przez ekspertów (ang. *ground truth*) oraz wyznaczenia właściwie dobranych metryk oceny. Etykiety eksperckie to także element niezbędny do

przeprowadzenia nadzorowanego treningu modeli uczenia maszynowego. Projekt wymagał przygotowania trzech rodzajów opisów eksperckich: (1) dla kroków pośrednich algorytmu przetwarzania danych wideo (detekcji regionów zainteresowania oraz segmentacji semantycznej artykulatorów), (2) segmentacji sygnału audio w czasie na słowa i fonemy oraz (3) opisów logopedycznych określających cechy artykulacyjne wymowy dziecka.

3.4.1 Etykiety eksperckie danych wizualnych

Dane eksperckie dla sygnału wideo obejmowały dwie kategorie etykiet wykorzystywanych do uczenia oraz walidacji proponowanych modeli konwolucyjnych sieci neuronowych:

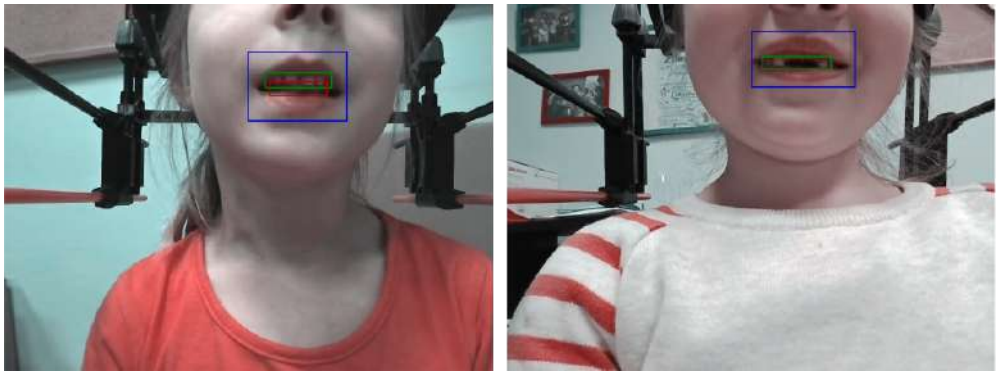
- prostokątne regiony zainteresowania (ang. *region of interest*, ROI) do detekcji poszczególnych artykulatorów (warg, zębów, języka),
- obrysy każdego z segmentowanych obiektów (warg, zębów, języka, ust).

Każdemu z obiektów przypisano właściwą klasę. Obrysy segmentacyjne obejmowały wszystkie piksele należące do danego artykulatora, a każdemu z pikseli przypisana została odpowiednia klasa (tj. wargi, zęby, język; usta generowano na podstawie maski warg). Rys. 3.4 ilustruje przykłady etykiet eksperckich dla obu przypadków.

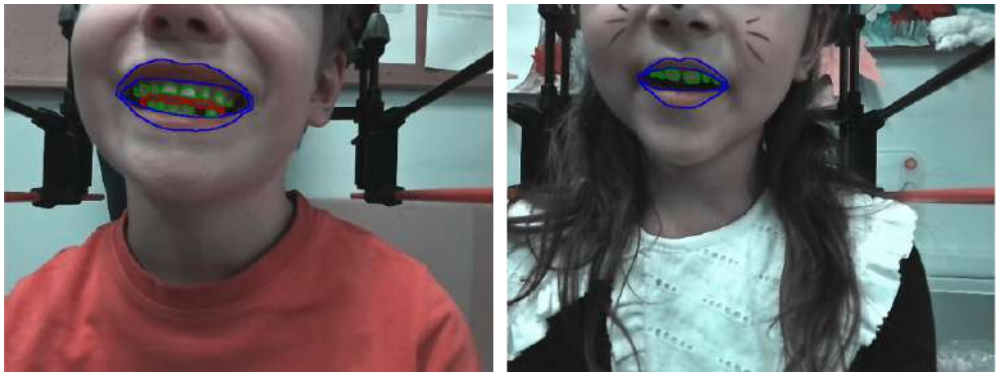
Dla celów badawczych zostały przygotowane etykiety przeznaczone do detekcji artykulatorów dla 17 913 obrazów wyekstrahowanych z nagrań 76 dzieci. Na 17 832 ramkach zaznaczono usta, a liczby wystąpień zębów i języka wynosiły odpowiednio 11 684 oraz 7 267. W przypadku segmentacji artykulatorów, baza danych eksperckich liczyła 1 757 przypadków, w tym 1 753 wystąpień warg, 946 wystąpień zębów oraz 501 wystąpień języka.

3.4.2 Segmentacja danych akustycznych

Z bazy danych na wstępie odrzucono dane z sygnałami widocznie uszkodzonymi (m.in. brak sygnału akustycznego lub wideo) oraz nagrania dzieci ze schorzeniami zaburzającymi dane w znacznym stopniu (np. obecność drenów usznych). Nagrania kwalifikujące się do dalszego przetwarzania zostały właściwie przygotowane. Oryginalne dane obejmowały zarówno wypowiadane przez dziecko słowa, fragmenty ciszy, tłumaczenia ćwiczeń, jak i zewnętrzne zakłócenia. W sygnale akustycznym pochodzącym z centralnego mikrofonu wskazano czas początku oraz końca poszczególnych wyrazów, a w każdym z wyrazów dodatkowo czas występowania pojedynczych fonemów (rys. 3.5).



(a)

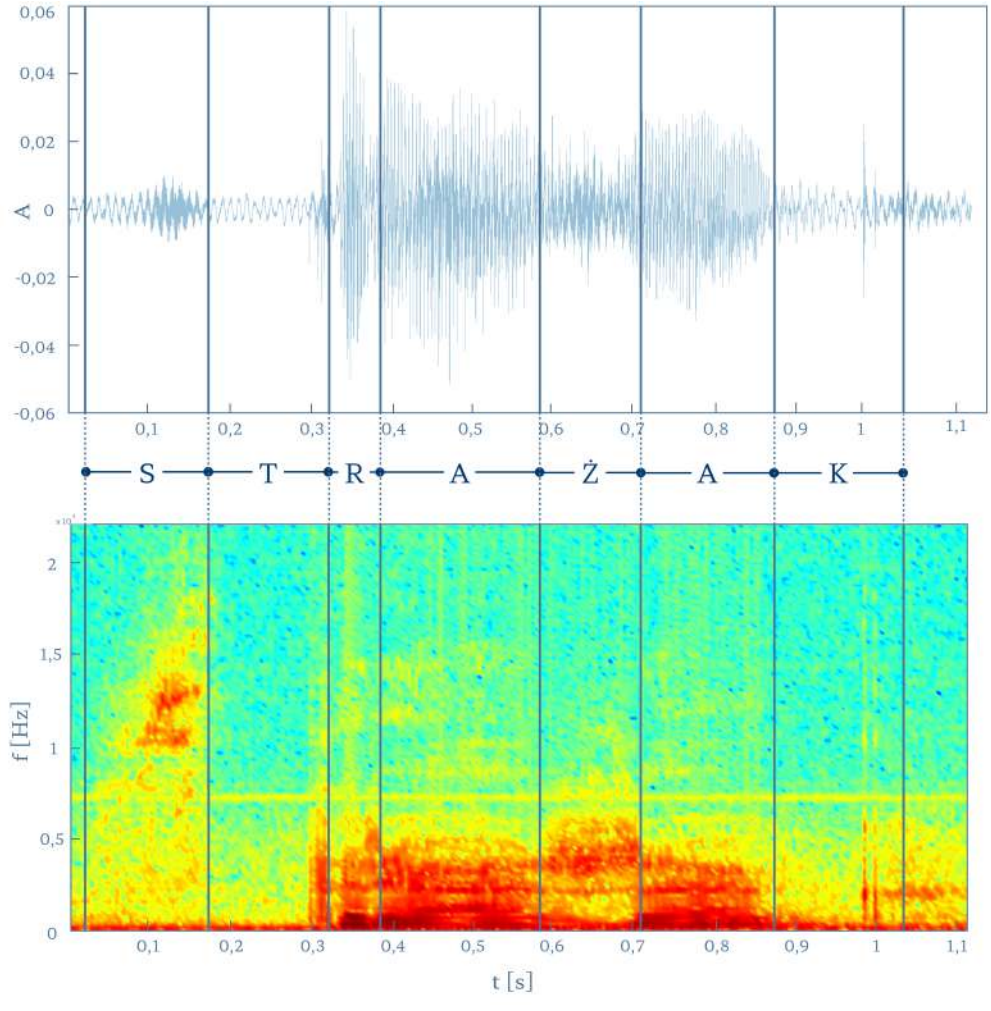


(b)

Rys. 3.4: Przykłady ręcznych obrysów eksperckich danych wizualnych: (3.4a) etykiety w formie prostokątnych ROI, przeznaczone do detekcji artykulatorów, (3.4b) obrysy dedykowane segmentacji semantycznej, wyróżniające piksele należące do wybranych obiektów. Linią niebieską zaznaczono obszar warg, zieloną — zębów, a czerwoną — języka.

3.4.3 Protokół badania logopedycznego

Dla każdej z badanych osób przygotowany został opis logopedyczny na podstawie protokołu opracowanego w projekcie badawczym przez zespół ekspertów w dziedzinie logopedii. Celem opisu był szczegółowy zarys: stanu wymowy dziecka (zwłaszcza sybilantów), cech oraz zdolności fizjologicznych (w tym połykania, oddychania czy mobilności języka), cech anatomicznych narządów artykulacyjnych (m.in. określenie długości wędzidełka podjęzykowego i wargi górnej, budowa podniebienia, uzębienie). Logopedzi stawiali również diagnozę dotyczącą normatywności realizacji poszczególnych głosek dentalizowanych. Ocenie poddano również cechy artykulacyjne: apikalność, dentalność/postdentalność, nonsonorność/sonorność, nonpalatalność/palatalność, dorsalność, nonzazalność, frykatywność/afrykatywność, medialność języka, medialność warg, medialność



Rys. 3.5: Reprezentacja czasowa i odpowiadająca jej postać czasowo-częstotliwościowa (spektrogram) słowa „*strażak*" z podziałem na poszczególne głoski.

żuchwy, medialność wypływu powietrza, dentalizację [112, 113]. Przygotowany protokół badania obejmował łącznie 196 pól, w tym elementy opisowe. Do zadań autorki rozprawy należało przygotowanie elektronicznej wersji protokołu, który był uzupełniany danymi diagnostycznymi badanych dzieci przez specjalistów-logopedów, oraz kontrola poprawności danych. Oceniane cechy są w niniejszej rozprawie równoznaczne z pojęciem cech artykulacyjnych.

Opis logopedyczny przygotowało dwoje ekspertów. W zależności od wybranego fonemu dentalizowanego, od 64% do 80% dzieci charakteryzowało się nienormatywną realizacją sybilantów (tab. 3.5) [158]. Realizację uznawano za normatywną, jeśli każda z 11-12 cech artykulacyjnych i fonetycznych danego fo-

Tab. 3.5: Wskaźniki wymowy normatywnej (NPR, ang. *normal pronunciation rate*) fonemów dentalizowanych [158].

Fonem	NPR	Fonem	NPR	Fonem	NPR
/s/	20,1%	/s̪/	25,8%	/s̪ʲ/	35,3%
/z/	20,7%	/z̪/	25,5%	/z̪ʲ/	34,2%
/ts/	20,4%	/ts̪/	26,2%	/ts̪ʲ/	35,6%
/dz/	19,6%	/dz̪/	25,0%	/dz̪ʲ/	35,1%

nemu była normatywna. Najniższym poziomem poprawności wymowy charakteryzował się fonem */dz/* (19,6%). Otrzymane statystyki potwierdzają istotność oraz powszechność problemu wśród polskich dzieci w wieku przedszkolnym. W niniejszej rozprawie korzystano z diagnoz jednego eksperta w celu uniknięcia ewentualnych niejednoznaczności interpretacyjnych — badanie stopnia zgodności między terapeutami nie było celem tej części badań.

4. Analiza danych wideo

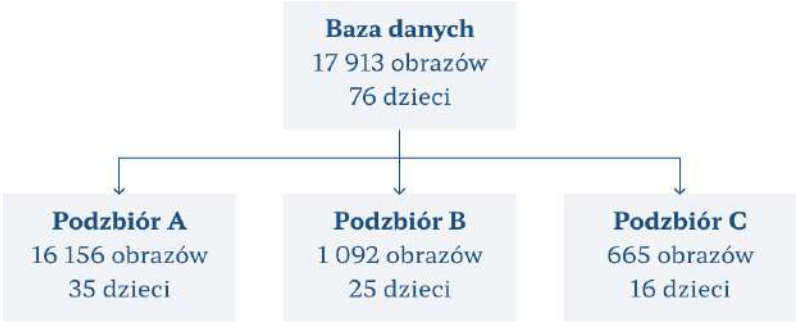
Ścieżka przetwarzania danych wideo (rys. 4.1) stanowi składową celu niniejszej pracy. W opisywanym podejściu opracowano hybrydowy algorytm segmentacji semantycznej artykulatorów (warg, ust, zębów, języka). W pierwszej kolejności ograniczony jest region zainteresowania ROI do prostokątów okalających narządy artykulacyjne (detekcja artykulatorów), następnie w pomniejszonym obszarze wyróżnione są piksele należące do poszczególnych obiektów (segmentacja semantyczna artykulatorów). Na podstawie pogrupowanych pikseli możliwe jest przeprowadzenie ekstrakcji cech.

Rys. 4.1: Schemat blokowy kolejnych kroków przetwarzania danych wideo.

4.1 Baza danych do detekcji i segmentacji artykulatorów

Baza danych przeznaczona do prac nad algorytmem detekcji i segmentacji semantycznej artykulatorów gromadziła 17 913 losowo wybranych obrazów pochodzących z nagrań wideo 76 dzieci. Zdjęcia z wybranych momentów nagrania ograniczono do lewej lub prawej kamery, tak, aby w zbiorze uczącym i testowym nie było dwóch widoków zarejestrowanych w tej samej chwili. Nagrania, z których ekstrahowano ramki, stanowiły część bazy danych opisanej w rozdziale 3.

Projektowana metoda charakteryzowała się wieloetapowością, dlatego procedury wstępnego przetwarzania oraz podziału danych były różne dla poszczególnych części procesu. Rys. 4.2 ilustruje liczebność danych w poszczególnych



Rys. 4.2: Podział danych wykorzystanych do przygotowania modelu do segmentacji artykulatorów — (1) podzbiór A: podstawowy zbiór treningowy i walidacyjny sieci do detekcji oraz sieci do segmentacji, (2) podzbiór B: przeznaczony na dostrojenie wag sieci segmentacyjnej, (3) podzbiór C: dane do testowania algorytmu.

Tab. 4.1: Rozkład klas w bazie danych.

	Ramki	**Dzieci**	**Ramki na mówcę** mediana (min–max)	**Wargi**	**Zęby**	**Język**
Podzbiór A	16,156	35	407 (290–880)	16,079	10,738	6,766
Podzbiór B	1,092	25	45 (29–47)	1,089	525	289
Podzbiór C	665	16	45 (16–54)	664	421	212
Suma	17,913	76		17,832 (99.5%)	11,684 (65.2%)	7,267 (40.6%)

podzbiorach. Podzbiór A składał się z 16 156 ramek pobranych od 35 dzieci i stanowił podstawowy zbiór treningowy oraz walidacyjny dla wszystkich etapów przetwarzania. Podzbiór B (1 092 ramek, 25 mówców) składał się z obrazów oraz odpowiadających im obrysów eksperckich przeznaczonych do dostrojenia sieci dedykowanej segmentacji semantycznej. Pozostałe 665 obrazów (16 dzieci, podzbiór C) poświęcono testowaniu algorytmu. W tab. 4.1 zebrano statystyki dotyczące liczebności obiektów poszczególnych klas oraz informację o rozkładzie liczby ramek od poszczególnych mówców w każdej z grup.

Wszystkie wykorzystywane obrazy posiadały etykiety eksperckie w postaci prostokątnych zaznaczeń regionów zainteresowania. Podział bazy danych na poszczególne podzbiory uwzględniał wystąpienia pacjentów, tj. dane pojedynczego dziecka pojawiały się tylko w jednym z podzbiorów. Takie działanie wspierało dążenie do uniwersalności modelu oraz miało zapobiec fałszywie zawyżonym metrykom jakościowym w przypadku analizy zdjęć podobnych do tych, na których sieć dokonywała modyfikacji swoich parametrów.

Rys. 4.3: Zawężenie regionu zainteresowania do obszaru twarzy dziecka na obrazie z lewej i prawej kamery.

4.2 Przetwarzanie wstępne

Na przetwarzanie wstępne składało się wyeliminowanie części brzegowych obrazu oraz adaptacyjne wyrównanie histogramu z ograniczeniem wzmocnienia kontrastu [105] (ang. *contrast-limited adaptive histogram equalization*, CLAHE). Ze względu na stabilność urządzenia pomiarowego na głowie dziecka oraz względnie nieduże różnice anatomiczne pomiędzy badanymi, obszar ust na obrazach występował w powtarzalnej lokalizacji. Umożliwiało to bezpieczne usunięcie nieistotnej informacji obrazowej jaką stanowiło tło. Prowadziło to do zmniejszenia złożoności obliczeniowej. Region zainteresowania zawężano o stałą liczbę pikseli, co skutkowało otrzymaniem ramki o wymiarach 300×440 — wartość dobrano doświadczalnie dla obu kamer (rys. 4.3).

W celu poprawy kontrastu mającej szczególne znaczenie w przypadku wpływu oświetlenia zewnętrznego (np. ciemne pomieszczenie, w którym przeprowadzano pomiary) zredukowany obraz był następnie poddawany wyrównywaniu histogramu algorytmem CLAHE. Eksperymentalnie wybrano siatkę o wymiarach 8×8 oraz ograniczenie wzmocnienia kontrastu na wartość 6.

Kolejny krok przygotowania danych wynikał z etapu uczenia modelu przeznaczonego do detekcji artykulatorów. W opracowanym podejściu na wejście modelu wprowadzano czteroramkowe mozaiki o wymiarach 2×2. Na macierz składały się wstępnie przetworzone obrazy. Pozwoliło to na zwiększenie liczby wystąpień poszczególnych klas na pojedynczym przypadku podawanym na wejście sieci. Takie podejście stanowi formę augmentacji danych oraz, w odniesieniu do literatury, usprawnia skuteczność treningu sieci typu *you only look once* (YOLO) [49, 170]. Porównanie jakości działania modelu uczonego jednoklatkowymi danymi potwierdziło słuszność wykorzystania mozaik. Stąd podzbiór

Rys. 4.4: Czteroelementowa mozaika złożona z losowych ramek różnych mówców.

treningowy A (rys. 4.2) liczący 16 156 obrazów został losowo podzielony na czteroelementowe macierze skutkując 4 039 mozaikami o wymiarach 600×880. Wynikowe dane gromadzą ramki od różnych mówców (rys. 4.4) — zróżnicowanie danych wspiera dążenie do uniwersalności oraz niezawodności metody, zwłaszcza w przypadku obrazów zawierających artefakty.

4.3 Detekcja artykulatorów

Identyfikacja oraz umiejscowienie wybranych obiektów w przestrzeni stanowią zasadniczy czynnik wiedzy człowieka o otaczającym świecie. Zdrowy ludzki układ wzrokowy pracuje z dużą szybkością przy jednoczesnej wysokiej precyzji — naukowcy opracowujący algorytmy dążą do uzyskania jak najwyższych wyników w obu kwestiach [53]. Wykrywanie (również: rozpoznawanie, detekcja) określonych obiektów jest istotnym i wciąż rozwijanym obszarem zagadnień wizji komputerowej [122, 169].

Detekcja obiektów polega na umiejscowieniu obiektu w przestrzeni obrazu (zagadnienie lokalizacji) oraz przypisaniu kategorii do jakiej wybrany obiekt należy (problem klasyfikacji) [175]. Konwencjonalne podejście do wykrywania obiektów składa się z trzech etapów: selekcji obszarów niosących istotne informacje, ekstrakcji cech oraz klasyfikacji. Obiekty mogą występować w różnych konfiguracjach, pozycjach i rozmiarach, dlatego prześledzenie powierzchni całego obrazu z wykorzystaniem przesuwającego się okna o różnych wymiarach wydaje się być rozsądną koncepcją. Zastosowanie takiej techniki wiąże się jednak z wysoką złożonością obliczeniową i długim czasem przetwarzania.

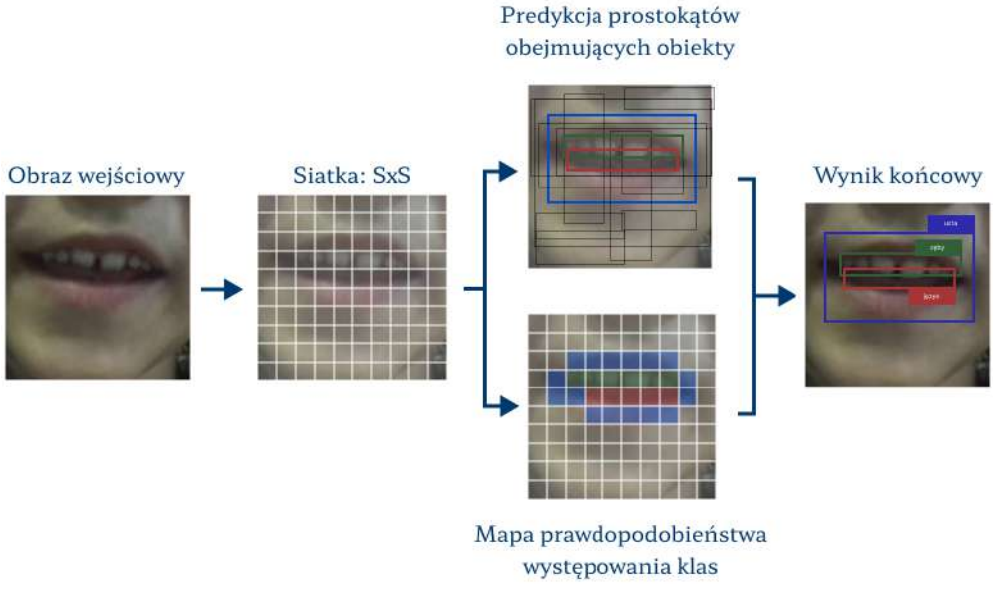
W kolejnym z etapów następuje ekstrakcja cech wizualnych, które mają różnicować obiekty pomiędzy sobą. Do metod wykorzystywanych w tym celu zaliczyć można m.in.: SIFT (ang. *scale invariant feature transform*) [87], HOG (ang. *histogram of oriented gradients*) [30] czy cechy Haara [77]. Ostatnim z kroków pośrednich jest zastosowanie klasyfikatora dzielącego przestrzeń cech na zadane kategorie — wykorzystać w tym celu można: maszynę wektorów nośnych (ang. *support vector machine*, SVM), AdaBoost, lasy losowe (ang. *random forest*) oraz wiele innych.

Poprawę szybkości oraz jakości detekcji przyniosło wprowadzenie algorytmów opartych na metodach głębokiego uczenia, zwłaszcza konwolucyjnych sieciach neuronowych (ang. *convolutional neural network*, CNN [142], m.in.: SSD (ang *single-shot detection*) [81], R-CNN (ang. *region-based CNN*) [42], Fast R-CNN [43], R-FCN (ang. *region-based fully convolutional networks*) [29] czy YOLO [122].

4.3.1 Sieć YOLO

Architektura sieci *you only look once* pozwala na uproszczenie problemu detekcji do pojedynczego problemu regresji [122]. Współrzędne obszarów ograniczających (ang. *bounding-box*) oraz prawdopodobieństwa klas przypisanych do znalezionych regionów wynikają bezpośrednio z pikseli obrazu (rys. 4.5). Wystarczy jedna analiza całego obrazu (stąd geneza nazwy modelu), aby sieć mogła lokalizować obiekty i przypisywać im klasy. Przewagą architektury YOLO nad większością dotychczasowych rozwiązań są: krótki czas przetwarzania związany z prostotą algorytmu, globalne spojrzenie na obraz podczas prognozowania oraz wysoki poziom efektywności nauki uniwersalnych reprezentacji obiektów [122]. Schematyczna zasada działania sieci YOLO została przedstawiona na rys. 4.6.

Kwadratowy obraz przyjmowany przez model na wejściu dzielony jest na siatkę $S \times S$ — jeżeli środek danego obiektu znajduje się w jednej z komórek siatki, komórka ta odpowiedzialna jest za detekcję przedmiotu. Każdy z elementów przewiduje B regionów zainteresowania oraz współczynnik pewności sc (ang. *confidence score*) dla tych obszarów. Ostatni z wymienionych parametrów niesie informację zarówno o prawdopodobieństwie wystąpienia danego obiektu w wybranym obszarze, jak i o dokładności lokalizacji i wymiarów zawężonego regionu. Dla każdego z obszarów ograniczających znajdywanych jest zatem 5 wartości (rys. 4.7): x, y, w, h, sc, gdzie symbole oznaczają kolejno: współrzędną x środka prostokąta, współrzędną y środka prostokąta, wysokość prostokąta, szerokość prostokąta oraz współczynnik pewności. Pewność sc liczona jest jako część wspólna do całości (ang. *intersection over union*, IoU) pomiędzy przewidywanym prostokątnym zaznaczeniem a odpowiadającym mu zaznaczeniem eksperta. Ponadto, każda z komórek siatki przewiduje również C warunkowych prawdopodobieństw klas, zgodnie ze wzorem: $Pr(Klasa_i|Obiekt)$ [122]. War-



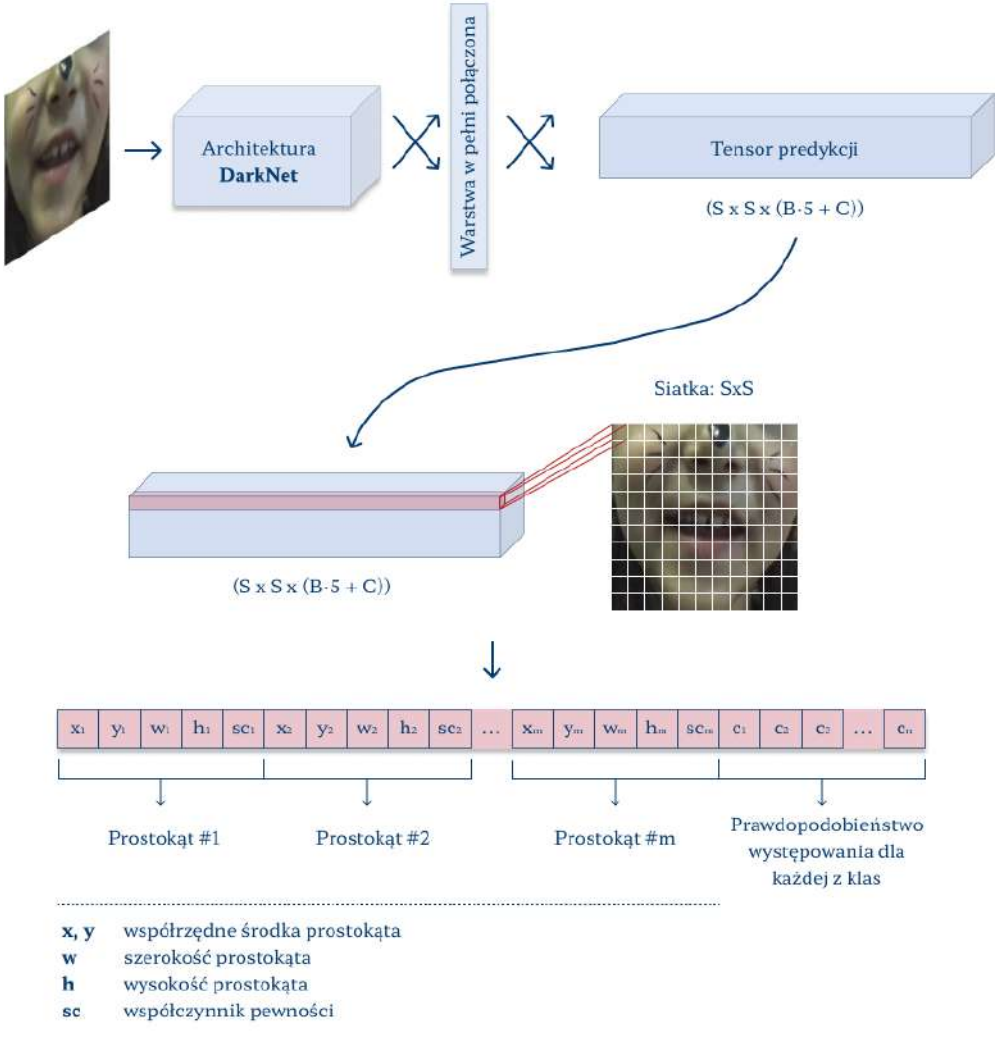
Rys. 4.5: Ilustracja analizy obrazu za pomocą sieci YOLO. Sieć poszukuje jednocześnie obszarów ograniczających obiekty oraz wartości prawdopodobieństwa występowania klas na kwadratowej siatce.

tości te są liczone jednokrotnie dla wybranej komórki siatki, bez względu na określoną liczebność B. Rozmiar siatki oraz liczba prostokątów ograniczających nie jest założona odgórnie i może być dostosowana przez osobę projektującą model. Przykład formatu etykiet wejściowych do uczenia sieci YOLO, który jest jednakowy dla wszystkich wersji, przedstawia rys. 4.7. Wektor wyjściowy, w zależności od ustawień, często poszerzony jest o wartość *sc* dla każdego z prostokątów ograniczających.

Ze względu na sukces wprowadzonego modelu, rodzina YOLO zaczęła poszerzać się o kolejne wersje architektury, przejawiające coraz wyższą skuteczność działania. W niniejszej pracy porównano efektywność czterech wybranych wariantów (v3, v5, v6 oraz v7). Opis przeprowadzonych eksperymentów znajduje się w rozdziale 7. Ich wyniki wskazały na wybór YOLOv6. Architekturę tej wersji zaprojektowano z myślą o wymaganiach sprzętowych (ang. *hardware-aware architecture*) wykorzystując struktury Rep-Pan i EfficientRep [74].

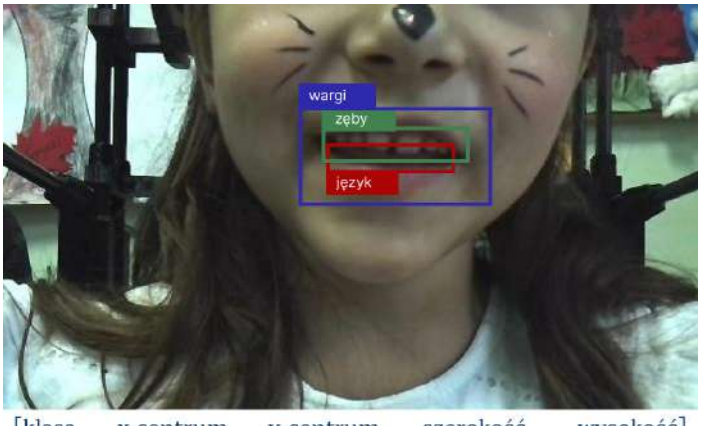
4.3.2 Hiperparametry sieci YOLOv6

Zaimplementowany model sieci wymagał obrazu wejściowego o kształcie kwadratu, dlatego obrazy poddano operacji skalowania. Odpowiadający rozmiar mozaik (320×320) uzyskano dzięki interpolacji dwuliniowej. Na danych treningowych zastosowano również wymuszone poszerzenie (tzw. augmentacja)



Rys. 4.6: Ilustacja zasady działania sieci YOLO. Sieć rozwiązuje problem regresji: dla każdej komórki siatki $S \times S$ przygotowywany jest tensor przechowujący dane B prostokątów, ich współczynników pewności oraz C prawdopodobieństw występowania klas.

w postaci: losowych zmian w modelu barw HSV, translacji horyzontalnej oraz wertykalnej, a także losowemu odbiciu lustrzanemu w poziomie. W efekcie obrazy zostały podzielone na zbiór treningowy liczący 3 888 mozaik (15 552 obrazów, 25 dzieci) oraz 151 mozaik (604 obrazów, 10 dzieci) w zbiorze walidacyjnym. Jako metodę optymalizacji uczenia wykorzystano metodę stochastycznego spadku gradientu (ang. *stochastic gradient descent*, SGD). Eksperymentalnie dobrano 300 epok procesu treningowego oraz wielkość mini-paczki (ang. *mini-*



[klasa	x-centrum	y-centrum	szerokość	wysokość]
[0	0.63	0.40	0.27	0.22]
[1	0.62	0.31	0.11	0.04]
[2	0.62	0.35	0.17	0.08]

Rys. 4.7: Przykład etykiety w formie wektora wartości, która opisuje obiekty zlokalizowane na obrazie. Klasy 0, 1 i 2 reprezentują odpowiednio: wargi, zęby oraz język. Wartości współrzędnych środka oraz wymiarów prostokąta są znormalizowane.

Tab. 4.2: Hiperparametry modelu YOLOv6 oraz augmentacji danych.

Hiperparametry	
Liczba epok	300
Rozmiar mini-paczki	16
Optymalizator	SGD
Funkcja straty	ważona entropia krzyżowa
Współczynnik szybkości uczenia	0,02
Regularyzacja L2	0,0005
Augmentacja danych	
Losowe zmiany w modelu barw HSV	H: 0,015; S: 0,7; V: 0,4
Translacja horyzontalna	±10% szerokości obrazu
Translacja wertykalna	±10% wysokości obrazu
Lustrzane odbicie w poziomie	50% prawdopodobieństwa
Zmiana skali	±50%

batch) równą 16. Hiperparametry modelu oraz parametry augmentacji danych zestawiono w tab. 4.2.

4.4 Segmentacja semantyczna artykulatorów

Pojęcie segmentacji stanowi jedno z najbardziej popularnych zagadnień badawczych wizji komputerowej. Proces ten dzieli obraz na określone regiony według kryteriów, które rozróżniają piksele należące do oddzielnych klas. W przypadku obrazów medycznych segmentacja często dotyczy rozdzielenia obszarów na poszczególne narządy lub tkanki. Operacja nierzadko stanowi punkt wyjścia do dalszego przetwarzania, np. rozpoznawania wzorców, ekstrakcji cech w celu numerycznego opisu obiektów czy możliwości przeprowadzania klasyfikacji [108, 168].

W zależności od spodziewanego rezultatu, wśród rodzajów technik wyodrębniania obiektów w obrazie, wyróżnia się m.in. segmentację semantyczną (ang. semantic segmentation), segmentację instancji oraz segmentację panoptyczną [25, 35]. Zgodnie z metodą segmentacji semantycznej, każdemu z pikseli obrazu przypisywana jest właściwa klasa [35]. Z kolei, segmentacja instancji identyfikuje oraz oddziela pojedyncze wystąpienia obiektu [35]. Połączeniem obu wymienionych technik jest segmentacja panoptyczna — pikselom obrazu przypisywana jest klasa, jednak równocześnie rozróżniane są także poszczególne wystąpienia obiektu w znalezionych obszarach [25].

4.4.1 Segmentacja semantyczna z wykorzystaniem technik głębokiego uczenia

Analizując literaturę przedmiotu z ostatniej dekady można wnioskować, że jedną z częściej i chętniej rozwijanych dziedzin segmentacji semantycznej, zwłaszcza w porównaniu z klasycznymi metodami, jest wykorzystywanie metod głębokiego uczenia [103, 133, 167]. Przyczyn takiego trendu można poszukiwać w większej dokładności działania metod popartej raportowanymi przez badaczy wynikami, krótszemu czasowi działania algorytmów oraz — co istotne również z punktu widzenia niniejszej pracy — możliwości wykorzystania niepełnego lub niedokładnie przygotowanego zbioru etykiet eksperckich niezbędnych do przeprowadzenia procesu uczenia sieci.

Sieci głębokie wymagają bardzo dużych zbiorów danych uczących. Przygotowanie etykiet eksperckich jest zadaniem czasochłonnym, zwłaszcza dla baz charakteryzujących się wysoką różnorodnością. Trening sieci może być w pełni nadzorowany (ang. *supervised learning*), słabo nadzorowany (ang. *weakly-supervised learning*), połowicznie nadzorowany (ang. *semi-supervised learning*) lub nienadzorowany (ang. *unsupervised learning*) [167]. Forma etykiet różni się w zależności od wybranej metody. W przypadku całkowicie nadzorowanych technik wymagany jest pełny zbiór możliwie dokładnych obrysów. Słabo oraz połowicznie nadzorowane metody wykorzystują inne postaci etykiet. Dla pierwszego podejścia mogą to być: prostokąty okalające obiekty, punkty lub niedo-

kładne zbiory punktów zawierające się w obiekcie (ang. *points, scribbles*) czy etykiety na poziomie całego obrazu, nazywające obiekt, lecz niewskazujące na jego położenie (ang. *image-level labels*). Drugie podejście przeważnie zakłada wykorzystanie zbioru, na który składają się etykiety dokładnie przygotowane oraz obrysy zgrubne (uproszczone lub nieprecyzyjne) [167].

4.4.2 Konwolucyjne sieci neuronowe w segmentacji

Ciągły rozwój zaawansowanych algorytmów głębokiego uczenia do segmentacji semantycznej widoczny jest w rozwiązaniach raportowanych w literaturze dotyczącej różnych obszarów badawczych [21, 35, 167]. Zadania bazujące na analizie obrazów w dużej mierze wykorzystują koncepcje wychodzące z konwolucyjnych (splotowych) sieci neuronowych CNN [152]. Idea architektury CNN inspirowana jest procesami zachodzącymi w korze wzrokowej żywych organizmów. Sieci zbudowane są z: warstwy wejściowej, wyjściowej oraz warstw ukrytych pomiędzy nimi. Ostatnie z wymienionych zbudowane są z kombinacji warstw o różnym przeznaczeniu. Do podstawowych zalicza się: warstwy konwolucyjne (ang. *convolutional layers*), w których zdefiniowane filtry za pomocą operacji splotu wykrywają cechy obrazu; warstwy redukujące (ang. *pooling layers*), które poprzez zastosowanie funkcji agregacji ograniczają wymiary map cech, pozostawiając najistotniejsze parametry; warstwy aktywacji (ang. *activation layers*, które wprowadzają nieliniowość do sieci. Sieci CNN są chętnie wykorzystywane do rozwiązywania problemów klasyfikacji, detekcji czy segmentacji obrazów (w tym segmentacji semantycznej). W zależności od typu zadania, różnić się będzie struktura architektury, budowa bazy danych oraz ustawienia hiperparametrów [152, 159].

Metody segmentacji semantycznej można zróżnicować na poszczególne podgrupy pod kątem odmiennej struktury sieci. Wśród nich można wyróżnić: metody oparte na podejściu proponowania regionów (ang. *region proposal based*) i metody oparte na sieciach całkowicie konwolucyjnych (ang. *fully convolutional networks*). Do pierwszej z podgrup zalicza się m.in.: R-CNN (ang. *regional CNN*) [42], Fast R-CNN [43], RANet (ang. *region attention network*) [136]. Z kolei, druga z kategorii obejmuje kilka różniących się grup algorytmów [35]. Należą do nich:

- sieci typu koder-dekoder (ang. *encoder-decoder*), m.in.: DeconvNet [110], SegNet [6], DeepLabv3+ [20], HadNet [88], IIE-SegNet [76];
- sieci wykorzystujące poszerzoną/dziurawą konwolucję (ang. *dilated/atrous convolution*), m.in.: DeepLabv1-v3 [17–19], DeepLabv3+ [20], EncNet [172], SETR [177];
- sieci stosujące fuzję cech, m.in.: ParseNet [82], ExFuse [174], SA-FFNet [147];

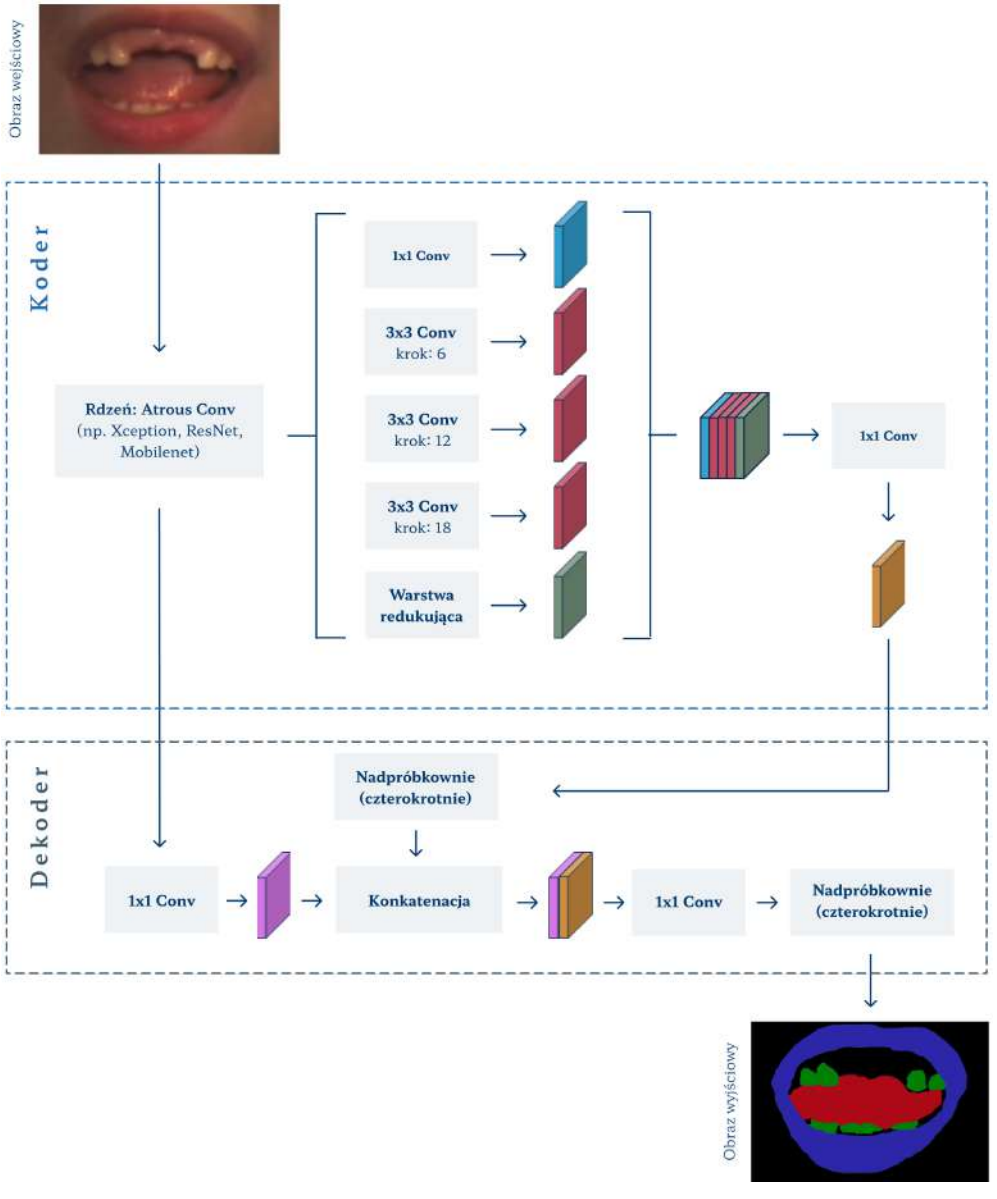
- sieci wykorzystujące piramidową architekturę oraz wieloskalowe cechy, m.in.: Multiscale ConvNet [38], DeepLabv2 [17], DeepLabv3 [19], SDN [39], HRNet [88];
- metody wykorzystujące rekurencyjne sieci neuronowe (ang. *recurrent neural networks*, RNN), m.in.: DD-RNN [36], ML-CRNN [37], MGCRNN [58].

4.4.3 Sieć DeepLab

Zasadniczą część algorytmu opracowanego w ramach doktoratu stanowi sieć DeepLabv3+ przeznaczona do segmentacji semantycznej — jest to czwarta i, na czas pisania rozprawy, ostatnia wersja architektury zaliczana do rodziny sieci DeepLab [17, 18, 20]. Dwuetapową architekturę typu koder-dekoder sieci zaprezentowano na rys. 4.8.

Większość dotychczasowych rozwiązań wykorzystujących metody głębokiego uczenia do segmentacji semantycznej zakładała dwuczłonową architekturę składającą się z kodera (tworzącego skompresowany wektor cech) oraz dekodera (rekonstruującego mapę cech do pożądanego wyniku oraz wymiarów wejściowych obrazu) [21]. Segmentacja semantyczna w oparciu o głębokie sieci konwolucyjne dzieli się na trzy gałęzie: (1) wykorzystująca kaskadową segmentację obrazu, po której następuje klasyfikacja regionów z użyciem CNN; (2) aplikująca sieci CNN do ekstrakcji cech w celu etykietowania obiektów na obrazie oraz przypisująca etykiety do niezależnie wysegmentowanych obszarów; (3) wykorzystująca sieci CNN do bezpośredniego znalezienia etykiet kategorycznych odpowiadających pojedynczym pikselom.

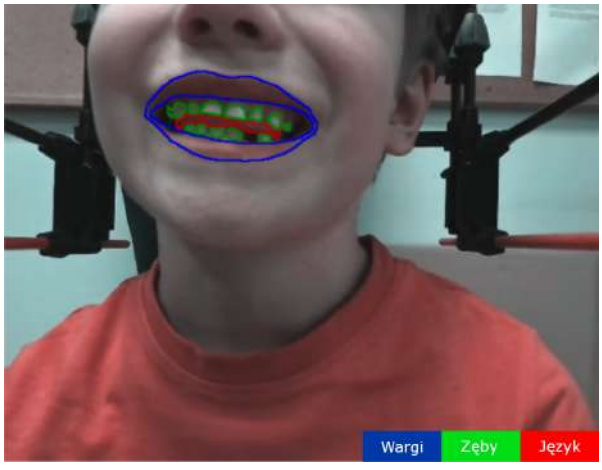
Stosowanie głębokich sieci konwolucyjnych do etykietowania danych wiąże się z technicznymi utrudnieniami, głównie podpróbkowaniem sygnału oraz przestrzenną niezmiennością [18]. Pierwszy problem odnosi się do wykorzystania warstw redukujących w konwencjonalnych sieciach — zamiast tego, autorzy sieci DeepLab zastosowali konwolucję "rozszerzoną, dziurawą" (ang. *atrous convolution*) pozwalającą na wydajną ekstrakcję cech głębokich. Stosowane w tym podejściu filtry obejmują większy obszar pola recepcyjnego (ang. *receptive field*) bez konieczności zmniejszania rozdzielczości przestrzennej lub zwiększania rozmiaru jądra. Druga z technicznych przeszkód wiąże się z wymogiem niewrażliwości na transformacje przestrzenne w przypadku działania klasyfikatorów obiektowo-centrycznych, co z natury ogranicza przestrzenną dokładność głębokich sieci konwolucyjnych. Architektura DeepLab w pierwszych wariantach aplikuje w pełni połączone warunkowe pole losowe (ang. *fully-connected conditional random field*). Rdzeń modelu DeepLab stanowi głęboka sieć konwolucyjna (np. szkielety sieci VGG, ResNet czy Xception). Standardowo jest ona przeszkolona w zadaniu klasyfikacji obrazów, lecz na potrzeby rozwiązania zadanego problemu jest adaptowana do rozwiązywania zagadnień segmentacji semantycznej poprzez: (1) przekształcenie warstw w pełni połączonych w warstwy



Rys. 4.8: Architektura sieci DeepLabv3+ do segmentacji semantycznej artykulatorów — ilustrację przygotowano w oparciu o [20].

splotowe oraz (2) zwiększenie rozdzielczości cech poprzez zastosowanie rozszerzonych warstw splotowych [17, 18].

Pierwotny wariant sieci DeepLab był sukcesywnie modyfikowany w celu poprawy jakości jego działania [17–20]. Zmianą zastosowaną w drugiej wersji, DeepLabv2, było wykorzystanie piramid ASPP (ang. *atrous spatial pyramid po-*



Rys. 4.9: Założony efekt działania opracowywanego algorytmu do segmentacji obszaru wybranych artykulatorów: warg, zębów, języka

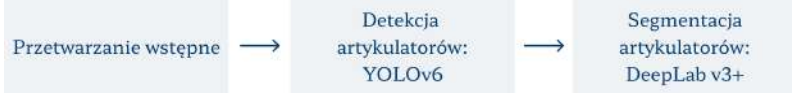
oling) w warstwach konwolucyjnych. Zgodnie z założeniami ASPP, na wejściową mapę cech aplikuje się równolegle kilka „dziurawych" konwolucji różniących się krokiem, a następnie rezultaty poddawane są fuzji. Obiekty tej samej klasy mogą być odmiennej skali nawet na pojedynczym obrazie, dlatego ASPP ma ograniczyć wpływ tego problemu i tym samym możliwie zwiększać jakość działania modelu [17]. Wersja trzecia (DeepLabv3) zachowała strukturę poprzedniej edycji, jednak została zredukowana o warstwę CRF, co pozwoliło na uzyskanie możliwości kompleksowego uczenia (ang. *end-to-end learning*). Ostatnia z wersji, DeepLabv3+, implementowana w opisywanej metodzie, została przekształcona w strukturę koder-dekoder. Koder wykazuje podobieństwo do struktury DeepLabv3, jednak z różnicą wykorzystania rozdzielnych rozszerzonych warstw konwolucyjnych (ang. *separable atrous convolution*), podczas gdy dekoder charakteryzuje się prostotą składając się głównie z operacji nadpróbkowywania celem uzyskania większej rozdzielczości przestrzennej.

4.4.4 Metoda segmentacji semantycznej artykulatorów

Celem opisywanego etapu pracy była segmentacja obszaru warg, ust, zębów oraz języka w każdej ramce zarejestrowanych danych wideo (rys. 4.9). Model segmentacji operował na obrazach zredukowanych do obszaru ust. Ograniczenie było możliwe dzięki wykorzystaniu sieci YOLO zaprezentowanej w poprzednich fragmentach pracy (sekcja 4.3.1).

Faza uczenia algorytmu składała się z trzech etapów (rys. 4.10): (1) wstępnej segmentacji podzbioru A z wykorzystaniem metody zbioru poziomic (ang. *level set method*) w celu uzyskania zgrubnych obrysów, (2) słabo nadzorowanego uczenia wstępnie przeszkolonego (transfer wiedzy, ang. *transfer learning*) mo-

(a)



(b)

Rys. 4.10: Budowa proponowanej metody segmentacji semantycznej wybranych artykulatorów w trybie uczenia (a) oraz działania (b).

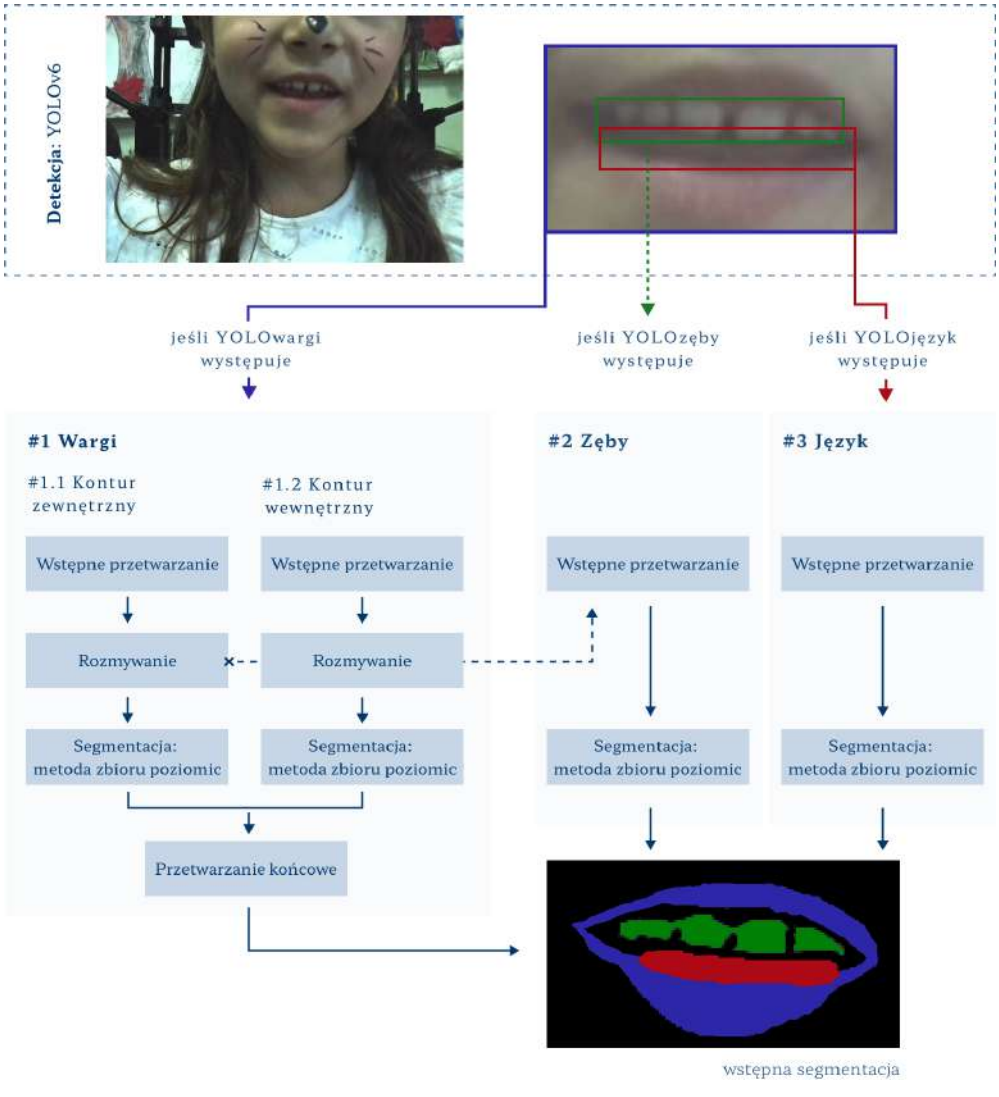
delu DeepLabv3+ z użyciem niedokładnych etykiet, (3) dostrojenie modelu z wykorzystaniem ręcznych obrysów eksperckich (podzbiór B). Idea częściowo nadzorowanego dwuetapowego treningu miała zniwelować problem czasochłonności przygotowania ręcznych etykiet przez ekspertów.

Obrazy podawane na wejście modelu opierają się na wynikach detekcji artykulatorów w danej ramce wideo. Określony został minimalny prostokąt okalający regiony zainteresowana (ROI) wszystkich obiektów zwracany przez sieć YOLO (w przypadku uczenia słabo nadzorowanego sieci) lub poprzez etykiety eksperckie (w przypadku dostrajania modelu). W znaczącej części przypadków ROI odzwierciedla obszar ust, z wyjątkiem stosunkowo niewielkiej liczby obrazów, na których język wystawał poza wargi. Obszar był redukowany do zadanego ROI pozostawiając trzypikselowy margines z każdej strony.

Wstępna segmentacja metodą zbioru poziomic

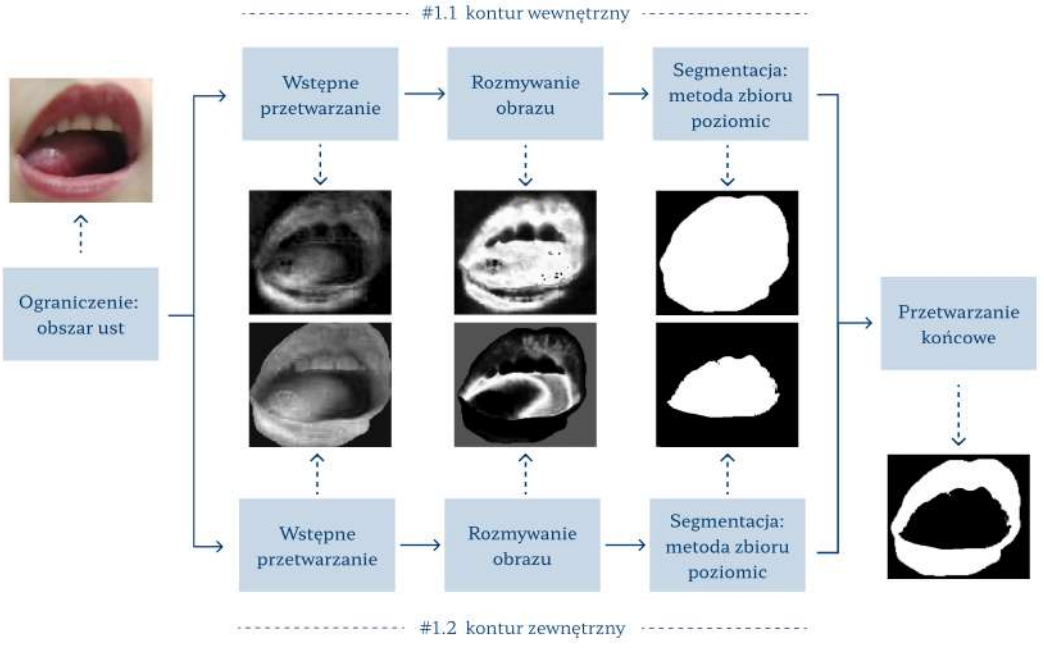
Wstępna segmentacja z wykorzystaniem elementów logiki rozmytej oraz metody zbioru poziomic DRLSE (ang. *distance regularized level set evolution*) [73] miała przygotować zgrubne obrysy do procesu uczenia słabo nadzorowanej sieci neuronowej, która w kolejnych etapach będzie rozwijana w kierunku dokładnej segmentacji. Zaletą takiego podejścia było zredukowanie etapu czasochłonnego przygotowywania ręcznych etykiet przez ekspertów. Zgodnie z rys. 4.11, obszar każdego z artykulatorów wyznaczony był według osobnej ścieżki przetwarzania.

Segmentacja warg dzieliła się na dwie gałęzie (rys. 4.12): poszukiwanie ich zewnętrznego (#1.1) oraz wewnętrznego konturu (#1.2). Wyznaczenie wierzchnich krawędzi przeprowadzano na obrazie różnicy czerwonego oraz zielonego



Rys. 4.11: Schemat blokowy wstępnej segmentacji metodą zbioru poziomic: obszar każdego z artykulatorów wyznaczany jest według osobnej ścieżki przetwarzania.

kanału obrazu RGB ust. Takie przekształcenie skutkowało zintensyfikowaniem wartości pikseli należących do obszaru ust. W kolejnym kroku, na podstawie miar statystycznych intensywności początkowego obszaru ust (średnia i odchylenie standardowe) definiowane były parametry opisujące postać funkcji przynależności stosowanej do rozmywania obrazu. Wykorzystano rozmytą funkcję przynależności Gaussa, aby przekształcić obraz do wartości z zakresu 0–1 dla metody DRLSE [7]. Efekt końcowy uzyskano po poprawie wyników wykorzystując operację otwarcia morfologicznego elementem strukturalnym w kształcie



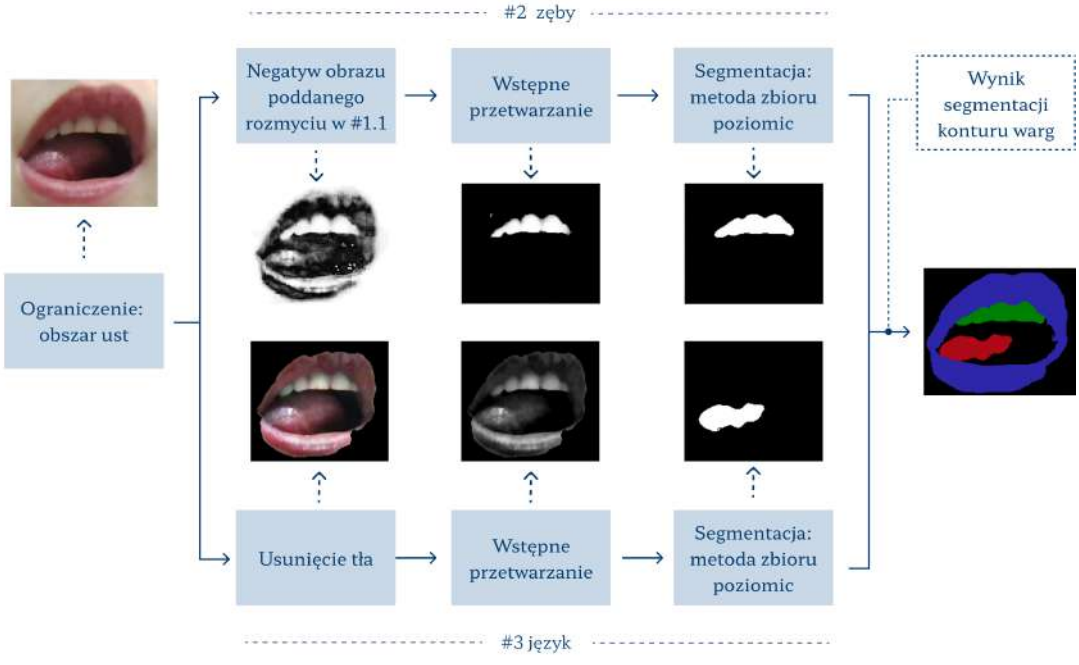
Rys. 4.12: Schemat blokowy dwugałęziowej metody wstępnej segmentacji warg.

dysku o średnicy pięciu pikseli. Otrzymany rezultat mógł być w tym momencie postrzegany jako maska dla całego obszaru ust.

Poszukiwanie wewnętrznego konturu warg bazowało jedynie na czerwonym kanale obrazu oryginalnego. Miary statystyczne centralnej części ROI determinowały przebieg krzywej rozmycia Gaussa. W efekcie piksele nienależące do obszaru warg zostały wyróżnione. Przetworzony obraz stanowił dane wejściowe dla metody DRLSE, której działanie zwracało wewnętrzny kontur ust (lub bardziej precyzyjnie — zewnętrzne granice przestrzeni pomiędzy wargami). Finalna maska warg stanowiła różnicę maski odzwierciedlającej kontur wewnętrzny i maski dla krawędzi zewnętrznych.

Ścieżka segmentacji zębów (#2, rys. 4.13) operowała na skorygowanym operacjami morfologicznymi negatywie wyniku, który został otrzymany w gałęzi #1.1. Jako kontur początkowy przyjęto prostokąt ograniczający zęby (uzyskany w ramach detekcji siecią YOLO). Analogicznie, w przypadku segmentacji języka (#3, rys. 4.13), obrazem wejściowym do metody DRLSE był kanał czerwony ograniczony przez ROI okalający wargi, a początkowy kontur stanowił prostokąt obejmujący obszar języka — rezultat działania modelu YOLO.

Każda z opisanych gałęzi segmentacji wstępnej wykorzystywała algorytm DRLSE, jednak ścieżki przetwarzania różniły się wartościami parametrów metody (tab. 4.3). Dobrano je eksperymentalnie, a różnice, w zależności od narządu, wynikały m.in. z konieczności kurczenia lub rozszerzania się konturu



Rys. 4.13: Schemat blokowy wstępnej segmentacji zębów oraz języka.

Tab. 4.3: Parametry metody DRLSE w zależności od ścieżki przetwarzania.

Oznaczenia:						
	α: współczynnik ważonej powierzchni					
	λ: współczynnik ważonej długości					
	ϵ: szerokość delty Diraca					
	σ: szerokość krzywej Gaussa					
	I: liczba iteracji algorytmu					
Parametry DRLSE						
		α	λ	ϵ	σ	I
#1.1	kontur wewnętrzny	-5	5	2,0	0,5	125
#1.2	kontur zewnętrzny	4	3	1,5	1,5	100
#2	zęby	6	4	2,0	1,5	125
#3	język	5	3	1,5	1,5	125

początkowego, z odmiennych przeważających odcieni szarości, z różnego kontrastu, a także z niejednoznaczności przebiegu konturów (szczególnie w przypadku języka).

Dwuetapowe uczenie sieci DeepLabv3+

Ideą dwuetapowego treningu sieci DeepLabv3+ było wykorzystanie stosunkowo niewielkiej liczby dokładnie obrysowanych etykiet do uzyskania modelu charakteryzującego się satysfakcjonującą jakością i uniwersalnością działania. Dlatego w pierwszym kroku do uczenia sieci wykorzystano duży zbiór różnorodnych danych z odpowiadającymi im zgrubnymi etykietami przygotowanymi w etapie poprzednim (sekcja 4.4.4). Oba kroki treningowe wykorzystywały tę samą architekturę sieci CNN.

W ramach eksperymentów przetestowano cztery rdzenie sieci DeepLabv3+ wybrane na podstawie przeglądu literatury: ResNet-101 [51], ResNet-152 [51], Xception [22] oraz MobileNetv2 [128]. Porównano także jakość ich działania wykorzystując obiektywne metryki. Ze względu na najkorzystniejsze rezultaty, rdzeń docelowej metody stanowiła architektura Xception [22] wstępnie przeszkolona na zbiorze danych ImageNet [31].

Obrazy wejściowe oraz odpowiadające im obrysy przeskalowano do wymiarów zgodnych z wymaganiami modelu DeepLabv3+: 224×224. W przypadku etapu dostrajania sieci — która pomija automatyczną detekcję siecią YOLO i wstępną segmentację — konieczne było ograniczenie oryginalnego obrazu na podstawie etykiet przygotowanych przez ekspertów (podzbiór C). Schemat postępowania był analogiczny i zakładał minimalny okalający prostokąt z uwzględnieniem dodatkowego trzypikselowego marginesu.

Oba etapy uczenia modelu wykorzystywały zbliżone ustawienia oraz hiperparametry (tab. 4.4). Zastosowano augmentację danych, uwzględniającą losowe zmiany kontrastu, translacje, rotację, skalowanie oraz obrót w osi horyzontalnej. Eksperymenty treningowe wskazywały optymalizator Adam (ang. *adaptive moment estimation*) i ważoną funkcję błędu entropii krzyżowej (ang. *weighted cross-entropy loss*) jako najbardziej wydajne. Rozmiar paczki treningowej ustawiono na 16, a maksymalna liczba epok wyniosła 150 z możliwością wczesnego zatrzymania, aby zminimalizować zagrożenie przeuczenia się sieci. Ponadto, 10% zbioru treningowego wykorzystano jako podzbiór walidacyjny.

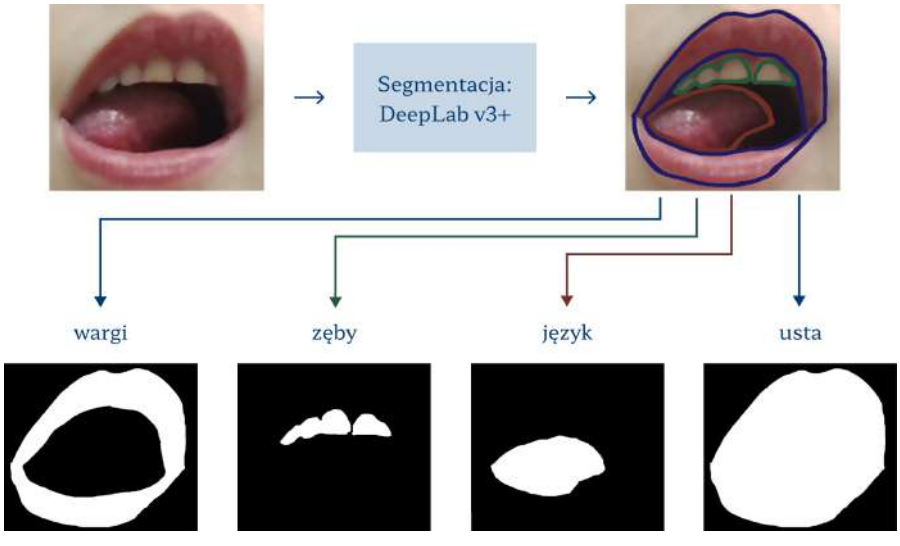
Model został przygotowany do ekstrakcji trzech obiektów: warg, zębów oraz języka. Mając taką konfigurację możliwe jest zdefiniowanie czwartego — oraz przypuszczalnie najbardziej istotnego — obiektu: całkowitego obszaru ust (rys. 4.14). Uzyskano go poprzez połączenie wszystkich klas uzupełnione o zalanie ewentualnych otworów.

4.5 Ekstrakcja cech obiektów

Obiektywne cechy opisujące dane stanowią istotne zagadnienie wizji komputerowej oraz analizy obrazów. Nierzadko stanowią podstawę do identyfikacji istotnych informacji oraz rozwiązywania problemów klasyfikacyjnych. Przed

Tab. 4.4: Hiperparametry modelu DeepLabv3+ oraz augmentacji danych.

Hiperparametry	
Liczba epok	150
Rozmiar mini-paczki	16
Optymalizator	Adam
Funkcja straty	ważona entropia krzyżowa
Współczynnik szybkości uczenia	0,001
Rdzeń	Xception, ResNet-101, ResNet-152, MobileNetv2
Augmentacja danych	
Losowe zmiany kontrastu	współczynnik kontrastu 0,1
Przesunięcie pionowe	±10% wysokości obrazu
Przesunięcie poziome	±10% szerokości obrazu
Rotacja	$\pm 36^o$
Zmiana skali	±10%
Lustrzane odbicie w poziomie	50% prawdopodobieństwa



Rys. 4.14: Obiekty wyodrębniane w procesie segmentacji semantycznej.

procesem estymacji parametrów często konieczne jest wstępne przetworzenie obrazu — np. filtracja, normalizacja, segmentacja — celem poprawy jakości danych lub wyróżnienia istotnych regionów (wyznaczenia regionu zaintereso-

Rys. 4.15: Podział cech obrazowych.

wania) [24, 127, 144, 148]. Na podstawie literatury omawiającej zagadnienia radiomiki oraz wykorzystywania estymacji parametrów obrazów, można rozpisać podział cech obrazowych zaprezentowany na rys. 4.15. Podstawowa klasyfikacja rozróżnia cechy związane z kształtem obiektu oraz jego teksturą [127, 148, 171].

W skład parametrów związanych z geometrią obiektu zalicza się wskaźniki oparte na konturze lub na obszarze. Wybór techniki zależy od podstawy obliczania cechy — czy jest nią wyłącznie krawędź, czy wszystkie piksele należące do obiektu. Te cechy dzielą się z kolei na globalne oraz strukturalne w zależności od tego, czy kształt jest brany pod uwagę całościowo, czy podzielony na segmenty [171].

W przeciwieństwie do cech związanych z kształtem, parametry teksturowe zamiast pojedynczych pikseli wykorzystują ich zbiorowiska [53, 148, 171]. Ekstrakcja cech teksturowych jest ściśle związana z zagadnieniami radiomiki, które w ostatnich latach zyskują coraz większą popularność [117]. Radiomika wykorzystuje ekstrakcję parametrów ilościowych z obrazów medycznych. Uzyskiwane wartości tworzą wielowymiarowy zestaw danych, który poddaje się eksploracji w celu potencjalnego wsparcia decyzji diagnostycznych. Cechy radiomiczne można ogólnie podzielić na: statystyczne, w tym zależne od histogramu i tekstury; związane z modelowaniem matematycznym; widmowe oraz parametry dotyczące kształtu [94, 95].

4.5.1 Cechy kształtu

Cechy bazujące na geometrii obiektu są deskryptorami rozmiaru i kształtu dwu- lub trójwymiarowego regionu zainteresowania. Charakteryzują się niezależnością względem rozkładu intensywności poziomów szarości w danym regionie i skutkują ilościowym opisem geometrycznych cech analizowanego obszaru [78, 130, 148]. Cechy kształtu w zagadnieniach radiomiki stanowią uzupeł-

nienie wskaźników teksturowych i mogą nieść istotne informacje dotyczące badanych obiektów. Różnice w tym zakresie pomiędzy obiektami mogą wskazywać na ich przynależność do różnych kategorii, a zmiany zachodzące w czasie często wiążą się z rozwijającą się patologią. Przykładowo, w dziedzinie onkologicznej cechy związane z geometrią są wykorzystywane m.in. do opisu agresywności guzów [78]. Wybrane w niniejszej pracy cechy geometryczne z uwzględnieniem wymiarowości przestrzeni zebrano w tab. 4.5 oraz tab. 4.6[1].

Tab. 4.5: Wybrane cechy geometryczne w przestrzeni dwuwymiarowej [8, 45, 78, 80, 106, 143, 181].

Cecha	Symbol	Wzór/definicja
Pole powierzchni	A^{2D}	$A^{2D} = \sum_{k=1}^{N_p} A_k$ (4.1) N_p — liczba pikseli obiektu; A_k — pole powierzchni piksela.
Obwód	P^{2D}	$P^{2D} = \sum_{i=1}^{N_f} P_i$ (4.2) $P_i = \sqrt{(a_i - b_i)^2}$ (4.3) a_i, b_i — wierzchołki i-tego spośród N_f odcinków siatki obwodu.
Sferyczność	S^{2D}	$S^{2D} = \frac{2\sqrt{\pi A^{2D}}}{P^{2D}}$ (4.4)
Dysproporcja sferyczna	DS^{2D}	$DS^{2D} = \frac{P^{2D}}{2\sqrt{\pi A^{2D}}}$ (4.5)
Długość osi głównej	Ax^{2D}_{major}	$Ax^{2D}_{major} = 4\sqrt{\lambda_{major}}$ (4.6) λ_{major} — wartość własna obiektu o maksymalnym module.

Kontynuacja tabeli na następnej stronie

[1] Przed wprowadzeniem symboli cech wizualnych i akustycznych należy zaznaczyć, że w pracy przyjęto zasady wyróżniania grup cech w szczególny sposób. Przyjęte reguły zostały opisane w Spisie skrótów i oznaczeń na str. xvii.

Cecha	Symbol	Wzór/definicja
Długość osi małej	Ax^{2D}_{minor}	$Ax^{2D}_{minor} = 4\sqrt{\lambda_{minor}}$ (4.7) λ_{major} — druga wartość własna obiektu.
Wydłużenie	E^{2D}	$E^{2D} = \sqrt{\frac{\lambda_{minor}}{\lambda_{major}}}$ (4.8)
Maksymalna średnica Fereta	D^{2D}_{Feret}	Długość najdłuższego rzutu obiektu.

Tab. 4.6: Wybrane cechy geometryczne w przestrzeni trójwymiarowej [13, 45, 129].

Cecha	Symbol	Wzór
Objętość	V^{3D}	$V^{3D}_v = \sum_{k=1}^{N_v} V_k$ (4.9) N_v — liczba wokseli obiektu; V_k — objętość woksela.
Pole powierzchni	A^{3D}	$A^{3D} = \sum_{i=1}^{N_f} A_i$ (4.10) $A_i = \frac{1}{2}\left\|a_i b_i \times a_i c_i\right\|$ (4.11) a_i, b_i, c_i — wierzchołki i-tego spośród N_f trójkątów siatki powierzchni.
Stosunek pola powierzchni do objętości	SVR^{3D}	$SVR^{3D} = \frac{A^{3D}}{V^{3D}}$ (4.12)
Sferyczność	S^{3D}	$S^{3D} = \frac{\sqrt[3]{36\pi\left(V^{3D}\right)^2}}{A^{3D}}$ (4.13)
Zwartość #1	C^{3D}_1	$C^{3D}_1 = \frac{V}{\sqrt{\pi\left(A^{3D}\right)^3}} = \frac{1}{6\pi}\sqrt{\left(S^{3D}\right)^3}$ (4.14)

Kontynuacja tabeli na następnej stronie

Cecha	Symbol	Wzór
Zwartość #2	C_2^{3D}	$C_2^{3D} = 36\pi \frac{\left(V^{3D}\right)^2}{\left(A^{3D}\right)^3} = \left(S^{3D}\right)^3$ (4.15)
Dysproporcja sferyczna	DS^{3D}	$DS^{3D} = \frac{A^{3D}}{\sqrt[3]{36\pi \left(V^{3D}\right)^2}}$ (4.16)
Długość osi głównej	Ax_{major}^{3D}	$Ax_{major}^{3D} = 4\sqrt{\lambda_{major}}$ (4.17)
Długość osi mniejszej	Ax_{minor}^{3D}	$Ax_{minor}^{3D} = 4\sqrt{\lambda_{minor}}$ (4.18)
Długość osi najmniejszej	Ax_{least}^{3D}	$Ax_{least}^{3D} = 4\sqrt{\lambda_{least}}$ (4.19) λ_{least} — trzecia wartość własna obiektu.
Wydłużenie	E^{3D}	$E^{3D} = \sqrt{\frac{\lambda_{minor}}{\lambda_{major}}}$ (4.20)
Płaskość	F^{3D}	$F^{3D} = \sqrt{\frac{\lambda_{least}}{\lambda_{major}}}$ (4.21)
Maksymalna średnica Fereta	D_{Feret}^{3D}	Długość najdłuższego rzutu obiektu.
Maksymalna średnica w płaszczyźnie osiowej	D_{XY}^{3D}	Długość rzutu obiektu w płaszczyźnie XY. Kierunki: X, Y — poziomy i pionowy wymiar ramki 2D; Z — trzeci wymiar związany z czasem ramki.
Maksymalna średnica w płaszczyźnie czołowej	D_{YZ}^{3D}	Długość rzutu obiektu w płaszczyźnie YZ.
Maksymalna średnica w płaszczyźnie strzałkowej	D_{XZ}^{3D}	Długość rzutu obiektu w płaszczyźnie XZ.

4.5.2 Cechy związane z teksturą

Tekstura obrazu [27, 94] zawiera informacje o budowie strukturalnej jego powierzchni i relacjach względem otoczenia. Należy do cech, które umożliwiają ludzkiemu wzrokowi identyfikację regionów zainteresowania w polu widzenia.

Metody wizji komputerowej, po części inspirowane działaniem wzroku człowieka, wykorzystują analizę tekstury w licznych zagadnieniach (m.in. analizie obrazów medycznych, teledetekcji, segmentacji obrazów). W radiologii, tekstura obrazu odnosi się do różnic w skali szarości w danym ROI — np. gładki (w kontekście tekstury) materiał charakteryzowałby się niską wartością entropii, a szorstki wykazywałby przeciwną zależność [160]. Metody ekstrakcji cech wykorzystywane w obszarze radiomiki dzielą się na trzy główne kategorie, które obejmują dane statystyczne, filtracyjne oraz morfologiczne [117, 130, 173]. Rezultatem opisywanych technik jest wygenerowanie wielowymiarowej przestrzeni cech zdolnej do poddania dalszej analizie.

Wskaźniki, które dotyczą opisu tekstury obejmują cechy pierwszego rzędu, drugiego rzędu oraz parametry wyższych rzędów [16, 83, 95, 106, 117, 144, 148, 173]. Pierwsze z wymienionych opisują rozkład wartości wokseli nie uwzględniając relacji przestrzennych. Właściwości te wyznaczane są na podstawie histogramu poziomów szarości, który jest funkcją przedstawiającą liczbę pikseli o danej intensywności dla każdego z poziomów (tab. 4.7) lub bezpośrednio na podstawie poziomów szarości pikseli obrazu (tab. 4.8). Ponieważ histogram obejmuje informacje dotyczące pojedynczych pikseli, tym samym zawiera on parametry statystyczne pierwszego rzędu. Dzieląc wartości histogramu przez całkowitą liczbę wokseli obrazu można uzyskać przybliżoną gęstość prawdopodobieństwa wystąpienia poziomów poszczególnych intensywności [94].

Tab. 4.7: Zestawienie cech pierwszego rzędu wyliczanych na podstawie histogramu [83, 94, 106, 117].

Cecha	Symbol	Wzór
Energia histogramu	E_h	$E_h = \sum_{i=0}^{N_g-1} p_i^2$ (4.22) $p_i = \frac{h_i}{N_p}$ (4.23) N_p — liczba wokseli; N_g — liczba poziomów szarości obrazu; h_i — i-ty przedział histogramu.
Entropia histogramu	H_h	$H_h = -\sum_{i=0}^{N_g-1} p_i \log_2(p_i)$ (4.24)

Kontynuacja tabeli na następnej stronie

Cecha	Symbol	Wzór
Wariancja histogramu	σ_h^2	$\sigma_h^2 = \sum_{i=0}^{N_g-1} (i-\mu_h)^2 p_i$ (4.25)
Odchylenie standardowe histogramu	σ_h	$\sigma_h = \sqrt{\sigma^2}$ (4.26)
Skośność histogramu	S_h	$S_h = \mu_3 = \sigma^{-3} \sum_{i=0}^{N_g-1} (i-\mu_h)^3 p_i$ (4.27)
Kurtoza histogramu	K_h	$K_h = \mu_4 = \sigma^{-4} \sum_{i=0}^{N_g-1} (i-\mu_h)^4 p_i - 3$ (4.28)

Tab. 4.8: Zestawienie cech pierwszego rzędu wyliczanych na podstawie poziomów szarości (intensywności) pikseli obrazu [83, 94, 106, 117].

Cecha	Symbol	Wzór
Energia obrazu	I_E	$I_E = \sum_{i=1}^{N_p} I_i^2$ (4.29) I_i — poziom szarości i-tego woksela obiektu.
Wariancja rozkładu poziomów szarości	I_{σ^2}	$I_{\sigma^2} = \frac{1}{N_p} \sum_{i=1}^{N_p} (I_i - \mu_h)^2$ (4.30)
Odchylenie standardowe rozkładu poziomów szarości	I_σ	$I_\sigma = \sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} (I_i - \mu_h)^2}$ (4.31)
Minimalny poziom szarości	I_{min}	$I_{min} = \min(I)$ (4.32) I — zbiór poziomów szarości wokseli;

Kontynuacja tabeli na następnej stronie

Cecha	Symbol	Wzór
Maksymalny poziom szarości	I_{max}	$I_{max} = \max(I)$ (4.33)
Średni poziom szarości	I_{μ}	$I_{\mu} = \frac{\sum_{i=1}^{N_p} I_i}{N_p}$ (4.34)
Mediana rozkładu poziomów szarości	I_{med}	$I_{med} = \mathrm{med}(I)$ (4.35)
Zakres poziomów szarości	I_R	$I_R = I_{max} - I_{min}$ (4.36)
Średnie odchylenie bezwzględne rozkładu poziomów szarości	I_{MAD}	$I_{MAD} = \frac{1}{N_p} \sum_{i=1}^{N_p} \lvert I_i - \mu_h \rvert$ (4.37)
Średnia kwadratowa rozkładu poziomów szarości	I_{RMS}	$I_{RMS} = \sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} I_i^2}$ (4.38)
10. percentyl rozkładu poziomów szarości	I_{p10}	$I_{p10} =$ 10. percentyl zbioru I (4.39)
90. percentyl rozkładu poziomów szarości	I_{p90}	$I_{p90} =$ 90. percentyl zbioru I (4.40)
Rozstęp ćwiartkowy rozkładu poziomów szarości	I_{IQR}	$I_{IQR} = I_{p75} - I_{p25}$ (4.41) I_{p75}, I_{p25} — 75. i 25. percentyl zbioru I.
Odporne średnie odchylenie bezwzględne rozkładu poziomów szarości	I_{rMAD}	$I_{rMAD} = \frac{1}{N_{10-90}} \sum_{i=1}^{N_{10-90}} \left\lvert I_{10-90}(i) - \bar{I}_{10-90} \right\rvert$ (4.42) N_{10-90} — liczba wokseli o poziomie szarości równym lub pomiędzy 10. oraz 90. percentylem; $\bar{I}_{10-90}$ — średni poziom szarości w tym zbiorze.

Kontynuacja tabeli na następnej stronie

Cecha	Symbol	Wzór
Skośność rozkładu poziomów szarości	I_S	$I_S = \frac{\frac{1}{N_p}\sum_{i=1}^{N_p}(I_i-\mu_h)^3}{\left(\sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p}(I_i-\mu_h)^2}\right)^3}$ (4.43)
Kurtoza rozkładu poziomów szarości	I_K	$I_K = \frac{\frac{1}{N_p}\sum_{i=1}^{N_p}(I_i-\mu_h)^4}{\left(\frac{1}{N_p}\sum_{i=1}^{N_p}(I_i-\mu_h)^2\right)^2}$ (4.44)

Cechy bazujące na parametrach statystycznych pierwszego rzędu pomijają zależności występujące między pikselami. Do ilościowego opisu relacji międzywokselowych wykorzystać można m.in. macierz współwystępowania poziomów szarości (ang. *gray level co-occurence matrix*, GLCM), macierz jednorodnych ciągów pikseli (ang. *gray level run length matrix*, GLRLM), macierz jednorodnych stref pikseli (ang. *gray level size zone matrix*, GLSZM) i macierz różnic poziomów szarości w sąsiedztwie (ang. *neighbouring gray tone difference matrix*, NGTDM).

Macierz współwystępowania poziomów szarości GLCM [16, 48, 50, 95, 117, 182] pozwala na opis przestrzennej dystrybucji poziomów szarości występujących w obrazie. Dane konstruuje się uwzględniając relacje pomiędzy parami wokseli oraz częstotliwością występowania każdej pary intensywności w obszarze obrazu lub wybranego regionu zainteresowania. Zależność między parą wokseli jest opisywana przez dwa parametry: odległość d oraz kąt θ. Jeśli liczba poziomów szarości w obrazie wynosi N_g, możliwych jest $N_g \times N_g$ par pikseli i taki jest rozmiar macierzy GLCM. Macierz GLCM obrazu o wymiarach $M \times N$ oraz N_g poziomach szarości jest opisana następującym wzorem matematycznym [16, 48, 50, 117, 182]:

$$GLCM_d^{\theta}(i,j) = \left|\left\{\left((r,s),(t,v)\right) : I(r,s) = i, I(t,v) = j\right\}\right| \forall i,j \in \{1,2,3,\cdots,N_g\}, \tag{4.45}$$

gdzie:

$$(r,s),(t,v) \in M \times N, (t,v) = \begin{cases} r+d, s & \text{jeżeli } \theta = 0^\circ \\ r+d, s+d & \text{jeżeli } \theta = 45^\circ \\ r, s+d & \text{jeżeli } \theta = 90^\circ \\ r-d, s+d & \text{jeżeli } \theta = 135^\circ \end{cases}, \tag{4.46}$$

I stanowi przetwarzany obraz jako funkcję $N_x \times N_y \rightarrow \{1,2,\cdots,N_g\}$, a $|\cdot|$ oznacza moc zbioru. Dla obrazu dwuwymiarowego zakres parametrów wynosi:

$\theta \in \{0°, 45°, 90°, 135°\}$, $d \in \{1, 2, 3, \cdots, n\}$, podczas gdy w przypadku przestrzeni trójwymiarowej zbiór θ rozszerza się do 13 kątów. Miary obliczane na podstawie macierzy GLCM zebrano w tab. 4.9[2] Zdefiniowano je dla pojedynczej macierzy zdarzeń obliczonej dla odległości d i kąta θ. Ostateczna wartość cechy powstaje przez uśrednienie składowych dla każdej pary (d, θ).

Tab. 4.9: Zestawienie cech otrzymywanych na podstawie macierzy GLCM [16, 45, 48, 50, 117, 182].

Cecha	Symbol	Wzór
Moment zwykły drugiego rzędu (ang. *angular second moment*)	ASM^{GLCM}	$ASM^{GLCM} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_{i,j}^2$ (4.47) $p_{i,j}$ — znormalizowana macierz GLCM.
Kontrast (ang. *contrast*)	Con^{GLCM}	$Con^{GLCM} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \lvert i-j\rvert^2 p_{i,j}$ (4.48)
Entropia (ang. *entropy*)	Ent^{GLCM}	$Ent^{GLCM} = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_{i,j}\log_2(p_{i,j})$ (4.49)
Średnia (ang. *mean*)	$Mean^{GLCM}$	$Mean^{GLCM} = \frac{(\mu_x+\mu_y)}{2}$ (4.50) $\mu_x = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} i\ p_{i,j}$ (4.51) $\mu_y = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} j p_{i,j}$ (4.52)

Kontynuacja tabeli na następnej stronie

[2] Nomenklatura polskojęzyczna nazw poszczególnych cech teksturowych w grupach GLCM, GLRLM, GSZDM i NGTDM nie wydaje się być oficjalnie ujednolicona. Bazując na nielicznych pozycjach literaturowych, m.in. [145], w niniejszej rozprawie podjęto próbę usystematyzowanego tłumaczenia nazw tych cech. Tab. 4.9–4.12 zawierają propozycje polskich nazw wraz z podaniem nazw oryginalnych.

Cecha	Symbol	Wzór
Wariancja (ang. *variance*)	Var^{GLCM}	$Var^{GLCM} = \frac{\left(\sigma_x^2 + \sigma_y^2\right)}{2}$ (4.53) $\sigma_x^2 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_{i,j}\,(i-\mu_x)^2$ (4.54) $\sigma_y^2 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_{i,j}\,(j-\mu_y)^2$ (4.55)
Korelacja (ang. *correlation*)	Cor^{GLCM}	$Cor^{GLCM} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{p_{i,j}\,(i-\mu_x)\,(j-\mu_y)}{\sqrt{\sigma_i^2\sigma_j^2}}$ (4.56)
Jednorodność (ang. *homogeneity*)	Hom^{GLCM}	$Hom^{GLCM} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{p_{i,j}}{1+(i-j)^2}$ (4.57)
Odmienność (ang. *dissimilarity*)	Dis^{GLCM}	$Dis^{GLCM} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_{i,j}\,\lvert i-j\rvert$ (4.58)
Autokorelacja (ang. *autocorrelation*)	AC^{GLCM}	$AC^{GLCM} = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} i\;jp_{i,j}$ (4.59)
Wartość średnia rozkładu sumacyjnego (ang. *sum average*)	SA^{GLCM}	$SA^{GLCM} = \sum_{n=2}^{2N_g} p_{x+y}(n)$ (4.60) $p_{x+y}(n) = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_{i,j}$ gdzie $i+j=n; \quad n=2,3,\ldots 2N_g$ (4.61)

Macierz jednorodnych ciągów pikseli GLRLM [40, 95, 151] niesie informacje o przestrzennym rozkładzie ciągów pikseli o jednakowym poziomie szarości, w jednym lub kilku kierunkach oraz w dwóch lub trzech wymiarach. W ujęciu praktycznym wykorzystuje się najczęściej cztery podstawowe kierunki, odpowiadające kątom θ równym 0°, 45°, 90°oraz 135°. Podobnie jak w przypadku GLCM, na podstawie macierzy pozyskiwane są specyficzne cechy (tab. 4.10; podano wzory cech dla pojedynczego kąta θ, ponownie ostateczna wartość cechy powstaje przez uśrednienie).

Tab. 4.10: Zestawienie cech otrzymywanych na podstawie macierzy GLRLM [40, 45, 95, 151].

Cecha	Symbol	Wzór
Współczynnik zawartości krótkich ciągów (ang. *short run emphasis*)	SRE^{GLRLM}	$SRE^{GLRLM} = \frac{1}{n_r}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{p_{i,j}}{j^2}$ (4.62) $p_{i,j}$ — znormalizowana macierz GLRLM dla kierunku θ; N_r — maksymalna długość jednorodnego ciągu pikseli w obrazie; n_r — liczba jednorodnych ciągów pikseli dla kierunku θ.
Współczynnik zawartości długich ciągów (ang. *long run emphasis*)	LRE^{GLRLM}	$LRE^{GLRLM} = \frac{1}{n_r}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} j^2 p_{i,j}$ (4.63)
Niejednorodność rozkładu poziomów szarości (ang. *gray-level nonuniformity*)	GLN^{GLRLM}	$GLN^{GLRLM} = \frac{1}{n_r}\sum_{i=1}^{N_g}\left(\sum_{j=1}^{N_r} p_{i,j}\right)^2$ (4.64)
Niejednorodność rozkładu długości ciągów (ang. *run length nonuniformity*)	RLN^{GLRLM}	$RLN^{GLRLM} = \frac{1}{n_r}\sum_{j=1}^{N_r}\left(\sum_{i=1}^{N_g} p_{i,j}\right)^2$ (4.65)

Kontynuacja tabeli na następnej stronie

Cecha	Symbol	Wzór
Odsetek liczby ciągów (ang. *run percentage*)	RP^{GLRLM}	$RP^{GLRLM} = \frac{n_r}{N_p}$ (4.66)
Współczynnik zawartości ciemnych ciągów[3] (ang. *low gray-level run emphasis*)	$LGRE^{GLRLM}$	$LGRE^{GLRLM} = \frac{1}{n_r}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{p_{i,j}}{i^2}$ (4.67)
Współczynnik zawartości jasnych ciągów (ang. *high gray-level run emphasis*)	$HGRE^{GLRLM}$	$HGRE^{GLRLM} = \frac{1}{n_r}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} i^2 p_{i,j}$ (4.68)
Współczynnik zawartości krótkich ciemnych ciągów (ang. *short run low gray-level emphasis*)	$SRLGE^{GLRLM}$	$SRLGE^{GLRLM} = \frac{1}{n_r}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{p_{i,j}}{i^2 j^2}$ (4.69)
Współczynnik zawartości krótkich jasnych ciągów (ang. *short run high gray-level emphasis*)	$SRHGE^{GLRLM}$	$SRHGE^{GLRLM} = \frac{1}{n_r}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{i^2 p_{i,j}}{j^2}$ (4.70)
Współczynnik zawartości długich ciemnych ciągów (ang. *long run low gray-level emphasis*)	$LRLGE^{GLRLM}$	$LRLGE^{GLRLM} = \frac{1}{n_r}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r}\frac{j^2 p_{i,j}}{i^2}$ (4.71)
Współczynnik zawartości długich jasnych ciągów (ang. *long run high gray-level emphasis*)	$LRHGE^{GLRLM}$	$LRHGE^{GLRLM} = \frac{1}{n_r}\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} i^2 j^2 p_{i,j}$ (4.72)

Kontynuacja tabeli na następnej stronie

[3] W pracy zastosowano tłumaczenie sformułowania „low/high gray-level" w postaci „ciemny/jasny" zamiast „niskich/wysokich poziomów szarości", ze względu na zwięzłość, jednoznaczność i łatwość interpretacji.

Cecha	Symbol	Wzór
Wariancja rozkładu poziomów szarości (ang. *gray level variance*)	GLV^{GLRLM}	$GLV^{GLRLM} = \frac{1}{N_g N_r} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p_{i,j} \, (i - \mu_i)^2$ (4.73) $\mu_i = \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} i \; p_{i,j}$ (4.74)
Wariancja rozładu długości ciągów (ang. *run length variance*)	RLV^{GLRLM}	$RLV^{GLRLM} = \frac{1}{N_g N_r} \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} p_{i,j} \, (j - \mu_j)^2$ (4.75) $\mu_j = \sum_{i=1}^{N_g} \sum_{j=1}^{N_r} j p_{i,j}$ (4.76)

Macierz jednorodnych stref pikseli GLSZM [45, 151, 154] wykazuje analogię do poprzednio opisywanej macierzy GLRLM, jednakże w przypadku GLSZM podstawę macierzy stanowią zliczenia liczby grup (stref, regionów) połączonych ze sobą sąsiadujących pikseli lub wokseli o jednakowym poziomie szarości. Tekstura charakteryzująca się wyższą homogenicznością skutkuje szerszą i bardziej płaską macierzą. W przeciwieństwie do poprzednich macierzy, GLSZM nie jest wyliczana dla różnych kierunków, może być jednak wyznaczana dla różnych odległości pikseli definiujących sąsiedztwo. Ponadto, cechy GLSZM mogą być uzyskiwane dla dwóch wymiarów (8-sąsiedztwo) lub trzech wymiarów (26-sąsiedztwo). Metryki obliczane na podstawie macierzy GLSZM zebrano w tab. 4.11.

Tab. 4.11: Zestawienie cech otrzymywanych na podstawie macierzy GLSZM [45, 151, 154].

Cecha	Symbol	Wzór
Współczynnik zawartości małych stref (ang. *small zone emphasis*)	SZE^{GLSZM}	$SZE^{GLSZM} = \frac{1}{n_z}\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{p_{i,j}}{j^2}$ (4.77) $p_{i,j}$ — znormalizowana macierz GLSZM; N_s — maksymalny rozmiar jednorodnej strefy pikseli w obrazie; n_z — liczba jednorodnych stref pikseli w teksturze.
Współczynnik zawartości dużych stref (ang. *large zone emphasis*)	LZE^{GLSZM}	$LZE^{GLSZM} = \frac{1}{n_z}\sum_{i=1}^{N_g}\sum_{j=1}^{N_s} j^2 p_{i,j}$ (4.78)
Niejednorodność rozkładu poziomów szarości (ang. *gray-level nonuniformity*)	GLN^{GLSZM}	$GLN^{GLSZM} = \frac{1}{n_z}\sum_{i=1}^{N_g}\left(\sum_{j=1}^{N_s} p_{i,j}\right)^2$ (4.79)
Niejednorodność rozkładu rozmiarów stref (ang. *zone size nonuniformity*)	ZSN^{GLSZM}	$ZSN^{GLSZM} = \frac{1}{n_z}\sum_{i=1}^{N_g}\left(\sum_{j=1}^{N_s} p_{i,j}\right)^2$ (4.80)
Odsetek liczby stref (ang. *zone percentage*)	ZP^{GLSZM}	$ZP^{GLSZM} = \frac{n_z}{N_p}$ (4.81)
Współczynnik zawartości ciemnych stref (ang. *low gray-level zone emphasis*)	$LGZE^{GLSZM}$	$LGZE^{GLSZM} = \frac{1}{n_z}\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{p_{i,j}}{i^2}$ (4.82)
Współczynnik zawartości jasnych stref (ang. *high gray-level zone emphasis*)	$HGZE^{GLSZM}$	$HGZE^{GLSZM} = \frac{1}{n_z}\sum_{i=1}^{N_g}\sum_{j=1}^{N_s} i^2 p_{i,j}$ (4.83)

Kontynuacja tabeli na następnej stronie

Cecha	Symbol	Wzór
Współczynnik zawartości małych ciemnych stref (ang. *small zone low gray-level emphasis*)	$SZLGE^{GLSZM}$	$SZLGE^{GLSZM} = \frac{1}{n_z}\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{p_{i,j}}{i^2 j^2}$ (4.84)
Współczynnik zawartości małych jasnych stref (ang. *small zone high gray-level emphasis*)	$SZHGE^{GLSZM}$	$SZHGE^{GLSZM} = \frac{1}{n_z}\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{i^2 p_{i,j}}{j^2}$ (4.85)
Współczynnik zawartości dużych ciemnych stref (ang. *large zone low gray-level emphasis*)	$LZLGE^{GLSZM}$	$LZLGE^{GLSZM} = \frac{1}{n_z}\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}\frac{j^2 p_{i,j}}{i^2}$ (4.86)
Współczynnik zawartości dużych jasnych stref (ang. *large zone high gray-level emphasis*)	$LZHGE^{GLSZM}$	$LZHGE^{GLSZM} = \frac{1}{n_z}\sum_{i=1}^{N_g}\sum_{j=1}^{N_s} i^2 j^2 p_{i,j}$ (4.87)
Wariancja rozkładu poziomów szarości (ang. *gray level variance*)	GLV^{GLSZM}	$GLV^{GLSZM} = \frac{1}{N_g N_s}\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}(i p_{i,j} - \mu_i)^2$ (4.88) $\mu_i = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r} i\ p_{i,j}$ (4.89)
Niejednorodność rozkładu rozmiarów stref (ang. *zone size variance*)	ZSV^{GLSZM}	$ZSV^{GLSZM} = \frac{1}{N_g N_s}\sum_{i=1}^{N_g}\sum_{j=1}^{N_s}(j p_{i,j} - \mu_j)^2$ (4.90) $\mu_j = \sum_{i=1}^{N_g}\sum_{j=1}^{N_r} j p_{i,j}$ (4.91)

Macierz różnic poziomów szarości w sąsiedztwie NGTDM określa ilościowo sumę różnic pomiędzy poziomem szarości danego piksela lub woksela a średnią intensywnością sąsiadujących pikseli lub wokseli znajdujących się w zadanej odległości [3]. Podstawowe cechy bazujące na macierzy NGTDM zamieszczono w tab. 4.12.

Tab. 4.12: Zestawienie cech otrzymywanych na podstawie macierzy NGTDM [3, 45].

Cecha	Symbol	Wzór
Zgrubność (ang. *coarseness*)	$Coar^{NGTDM}$	$Coar^{NGTDM} = \frac{1}{\sum_{i=1}^{N_g} p_i s_i}$ (4.92) p_i — prawdopodobieństwo wystąpienia w teksturze poziomu szarości $i \in \langle 1, N_g \rangle$; s_i — suma bezwzględnych różnic poziomów szarości dla pikseli o poziomie szarości i i ich sąsiadów.
Kontrast (ang. *contrast*)	Con^{NGTDM}	$Con^{NGTDM} = \left(\frac{1}{N_{g,p}\left(N_{g,p}-1\right)} \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_i p_j \left(i-j\right)^2 \right) \cdot \left(\frac{1}{N_{v,p}} \sum_{i=1}^{N_g} s_i \right)$, gdzie $p_i \neq 0, p_j \neq 0$ (4.93) $N_{g,p}$ — liczba poziomów szarości, dla których $p_i \neq 0$; $N_{v,p}$ — liczba pikseli z kompletnym sąsiedztwem.
Zmienność (ang. *busyness*)	Bus^{NGTDM}	$Bus^{NGTDM} = \frac{\sum_{i=1}^{N_g} p_i s_i}{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \lvert i p_i - j p_j \rvert}$, (4.94) gdzie $p_i \neq 0, p_j \neq 0$
Złożoność (ang. *complexity*)	Com^{NGTDM}	$Com^{NGTDM} = \frac{1}{N_{v,p}} \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \lvert i-j \rvert \frac{p_i s_i + p_j s_j}{p_i + p_j}$, gdzie $p_i \neq 0, p_j \neq 0$ (4.95)
Siła tekstury (ang. *texture strength*)	TS^{NGTDM}	$TS^{NGTDM} = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \left(p_i + p_j\right)\left(i-j\right)^2}{\sum_{i=1}^{N_g} s_i}$, (4.96) gdzie $p_i \neq 0, p_j \neq 0$

Tab. 4.13: Zestawienie cech obrazowych analizowanych w pracy.

Typ	2D/3D	Grupa/obiekt	#	Łącznie	Definicje
Kształtu	2D	Usta	8	24	Tab. 4.5
		Wargi	8		
		Język	8		
	3D	Usta	16	48	Tab. 4.6
		Wargi	16		
		Język	16		
Teksturowe	2D	I rzędu (histogram)	6	63	Tab. 4.7
		I rzędu (obraz)	16		Tab. 4.8
		GLCM	10		Tab. 4.9
		GLRLM	13		Tab. 4.10
		GLSZM	13		Tab. 4.11
		NGTDM	5		Tab. 4.12
	3D	I rzędu (histogram)	6	63	Tab. 4.7
		I rzędu (obraz)	16		Tab. 4.8
		GLCM	10		Tab. 4.9
		GLRLM	13		Tab. 4.10
		GLSZM	13		Tab. 4.11
		NGTDM	5		Tab. 4.12

Łącznie unikalne cechy kształtu	72
Łącznie unikalne cechy teksturowe	126
Łącznie unikalne cechy obrazowe	198
Łącznie cechy obrazowe (kamera L+P)	**396**

4.5.3 Ekstrakcja cech obrazowych dla celów komputerowego wsparcia diagnostyki logopedycznej

W pracy zdecydowano się analizować: cechy teksturowe, które mają pokazać ogólne wzorce w obszarze ust, oraz parametry geometryczne warg, ust (suma warg i przestrzeni międzywargowej) oraz języka (tab. 4.13). Po konsultacji z logopedami, w analizie pominięto cechy kształtu związane z zębami — wśród dzieci przedszkolnych ich brak jest często uwarunkowany biologicznie, a ich włączenie do analizy mogłoby skomplikować proces lub wymagać uproszczeń w założeniach.

W trakcie sesji pomiarowych dane obrazowe rejestrowano w synchronizacji z sygnałem akustycznym. Na podstawie segmentacji eksperckiej nagrań audio na głoski (rozdział 3.4.2) nagrania wideo zostały ograniczone do ramek obejmowanych przez wyodrębnione segmenty danej głoski. W pracy analizowano zarówno głoski frykatywne (przedłużające się, ciągnące), jak i afrykaty (w których występuje moment zwarcia oraz późniejszego tworzenia się szczeliny). Ze

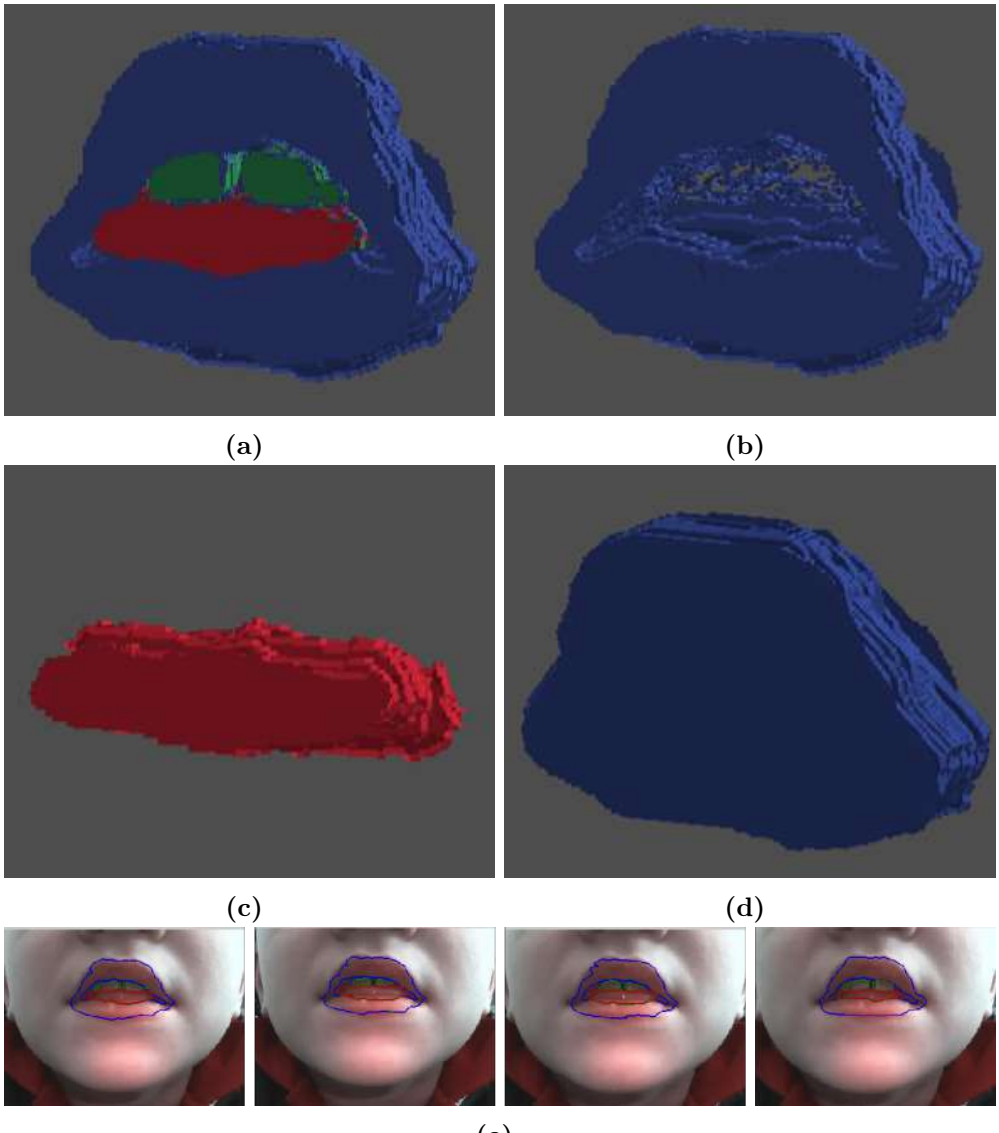
względu na specyfikę afrykatów oraz możliwość występowania ciszy lub szumu na brzegach nagrania, dodatkowo odrzucono 25% skrajnych ramek wideo z początku oraz końca segmentu głoski[4]. Zasada ta nie obowiązywała w przypadku krótkich segmentów (składających się z maksymalnie 3 ramek).

Dla danych wideo ekstrahowano zestawy dwu- oraz trójwymiarowych cech (tab. 4.13). Dwuwymiarowe podejście zakładało wykorzystanie serii kolejnych ramek (2D) nagrań. Dla każdego z obrazów wyznaczano 24 cechy geometryczne (po 8 cech dla warg, ust, języka) oraz 63 cechy teksturowe. Obrazy w skali szarości uzyskiwano z obrazów barwnych jako składową luminancyjną modelu barw HSL (ang. *hue, saturation, luminance*). Ponieważ założono, że analizowanie tekstury obszaru ust można ograniczyć do badania relatywnie zgrubnych relacji, parametry teksturowe obliczano dla $N_g = 32$ poziomów szarości.

Drugie podejście (cechy 3D) wymagało przygotowania trójwymiarowego wolumenu ruchu artykulatorów, w którym na kolejne przekroje składały się następujące po sobie ramki wideo (rys. 4.16). Pojedynczy model prezentował wymowę wyizolowanej głoski. Na podstawie wolumenu wyznaczano 63 cechy teksturowe oraz po 16 cech kształtu dla każdego z narządów (wargi, język, usta).

Łączna liczba unikalnych cech kształtu wyniosła 72, natomiast cech teksturowych 126. Razem dało to 198 niepowtarzających się cech wizualnych. W pracy analizowano jednak wektory parametrów uzyskanych z dwóch kamer (396 cech).

[4] Analogiczny zabieg przeprowadzono dla stosownych ramek w ekstrakcji cech akustycznych.



Rys. 4.16: Wizualizacja trójwymiarowego modelu segmentacji fonemu /ts/ w słowie *taca*: (a) pełny model (wargi, zęby, język), (b) model warg, (c) model języka, (d) model ust, (e) wybrane ramki wideo z naniesionymi obrysami segmentacji.

5. Analiza sygnału audio

Drugą ścieżką przetwarzania opracowaną w ramach pracy była analiza sygnału akustycznego. Specyfika danych rejestrowanych z wykorzystaniem 15 przestrzennie rozłożonych mikrofonów umożliwia wdrożenie różnych wariantów przetwarzania, np. w formie jednokanałowego układu (1-CH) wykorzystującego centralny mikrofon; jako 15-kanałowy układ (15-CH), w którym indywidualnie przetwarzane są dane z każdego z mikrofonów; jako pięciokanałowy układ (5-CH), gdzie pięć sygnałów zostaje zagregowanych w pięciu zorientowanych pionowo matrycach (ang. uniform linear arrays, ULA) [66]. W niniejszej pracy wykorzystano jednokanałowe podejście przetwarzania sygnału

5.1 Przetwarzanie wstępne

Na podstawie segmentacji eksperckiej opisanej w rozdziale 3.4.2 nagrania zostały ograniczone do czasu trwania głosek dentalizowanych w poszczególnych słowach (tj. podzielone na segmenty). Kolejne kroki przetwarzania sygnału, które prowadziły do ekstrakcji cech, zostały zilustrowane na rys. 5.1.

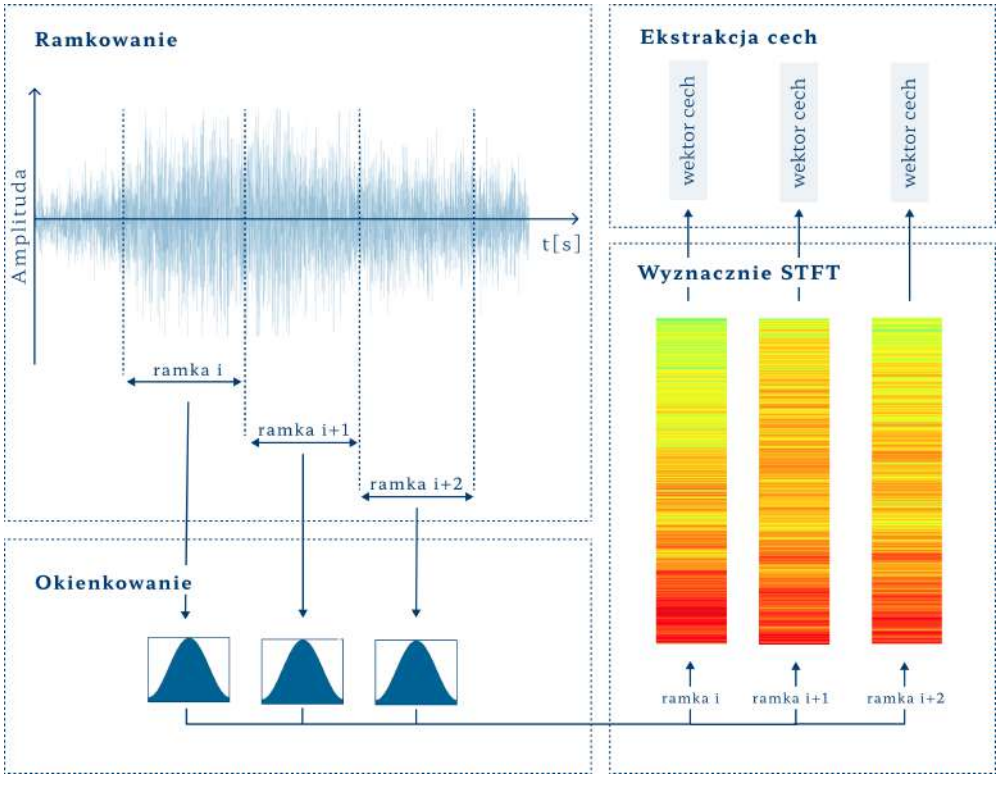
Pierwszym etapem przetwarzania była normalizacja sygnału akustycznego w ramach wysegmentowanych głosek do wartości z zakresu $[0, 1]$ zgodnie z:

$$x(n) = \frac{x_o(n) - x_{min}}{x_{max} - x_{min}}, \tag{5.1}$$

gdzie x_o jest sygnałem podanym na wejście, x_{min} i x_{max} najmniejszą i największą jego wartością w ramach głoski. Aby zminimalizować występowanie zniekształceń widmowych, każda z ramek kształtowana była oknem Hamminga:

$$w(n) = 0,54 - 0,46 \cos \frac{2\pi n}{N-1}, \qquad 0 \leqslant n \leqslant N-1, \quad N = 1470. \tag{5.2}$$

W procesie powstawania mowy kształtowana jest przede wszystkim obwiednia amplitudowo-częstotliwościowa sygnału, podczas gdy struktury ucha wewnętrznego w trakcie jego percepcji (jeszcze przed analizą w mózgu) rozkładają dane na składowe o poszczególnych częstotliwościach — dlatego z biologicznego punktu widzenia przetwarzane jest głównie widmo sygnału [149]. Ze względu na użyteczność widmowej reprezentacji sygnału mowy, jednym z pod-

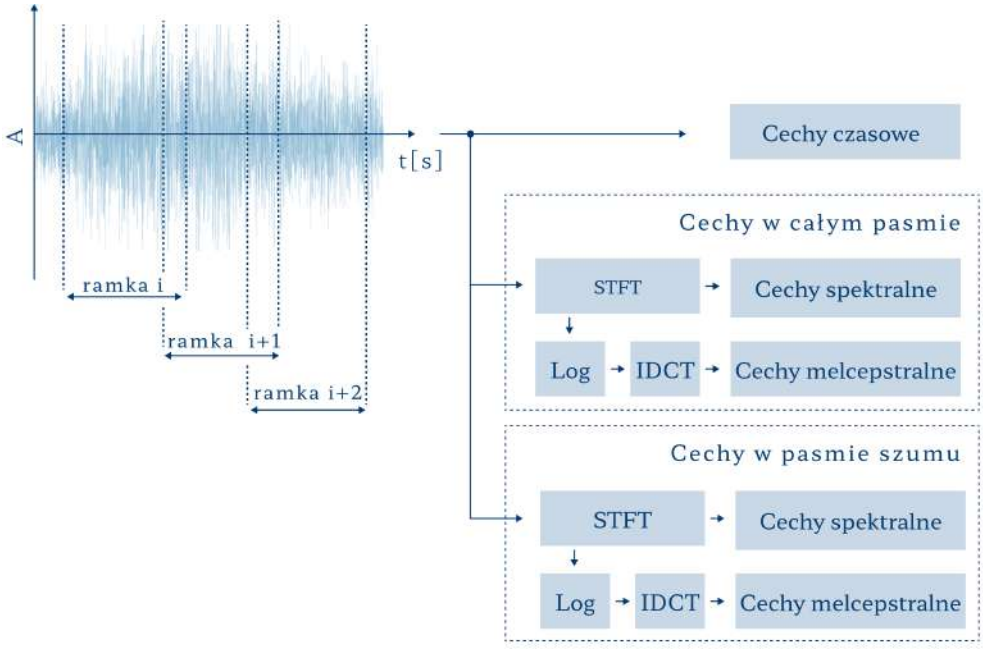


Rys. 5.1: Kolejne etapy przygotowujące sygnał do ekstrakcji cech akustycznych.

stawowych kroków jej analizy jest przeniesienie danych z dziedziny czasu do dziedziny częstotliwości. Dyskretna transformata Fouriera (ang. *discrete Fourier transform*, DFT) zakłada stacjonarność przetwarzanych danych. Ponieważ sygnał mowy nie spełnia tego wymagania, zamiast wymienionego przekształcenia stosuje się krótkoczasową transformatę Fouriera (ang. *short-time Fourier transform*, STFT). Niemniej, aby zachować niezbędne dla analizy widmowej założenie o (pseudo) stacjonarności sygnału, analizowane dane były dzielone na ramki o długości 33 ms każda. Częstotliwość ramkowania odpowiadała częstotliwości klatkowania w sygnale wideo. Nie zastosowano nakładkowania, a częstotliwość próbkowania wynosiła 44,1 kHz. Dla każdego pasma k oraz każdej ramki czasowej m widmo STFT równa się:

$$X(k,m) = \sum_{n=0}^{N-1} w(n)x(n+mN)e^{-\frac{i2\pi kn}{N}}, \tag{5.3}$$

Liczba próbek w ramce N jest rozmiarem DFT. Wynikiem działania STFT są widma kolejnych ramek, które zestawione ze sobą reprezentują czasowo-często-



Rys. 5.2: Podział cech akustycznych wyznaczanych w kolejnych ramkach z uwzględnieniem dziedziny sygnału.

tliwościowy przebieg sygnału. Jego postacią graficzną jest spektrogram (por. rys. 3.5).

5.2 Ekstrakcja cech

Cechy sygnału akustycznego znajdują wiele zastosowań w dziedzinie przetwarzania mowy: jej rozpoznawania, identyfikacji mówcy czy syntezy mowy. Ich wykorzystanie raportowane jest też w literaturze dotyczącej systemów komputerowego wsparcia diagnostyki i terapii logopedycznej [64, 99]. Biorąc pod uwagę dziedzinę, cechy akustyczne można podzielić na: czasowe oraz częstotliwościowe (rys. 5.2) [41, 57, 107, 135]. Parametry spektralne, z kolei, można uzyskiwać na podstawie pełnego pasma lub częściowo ograniczonego. Cechy czasowe wyznaczane są w dziedzinie czasu i wykazują najmniejszą złożoność obliczeniową, cechy widmowe wyodrębnia się na podstawie częstotliwościowej reprezentacji sygnału [134].

5.2.1 Cechy w dziedzinie czasu

Jedną z zalet parametrów wyznaczanych w dziedzinie czasu jest to, że nie wymagają przekształceń oryginalnych danych, a ich obliczenia przeprowadzane

są bezpośrednio na próbkach sygnału. W literaturze można znaleźć podział cech na zależne od: liczby przejść przez zero, amplitudy, mocy oraz rytmu sygnału [2, 57, 107]. Zestawienie wybranych parametrów czasowych z różnych kategorii przedstawiono w tab. 5.1. Niemniej, mimo, że ekstrakcja opisywanych cech charakteryzuje się prostotą oraz szybkością, parametry te nie są wystarczające do precyzyjnego opisu zmian w sygnale (zwłaszcza, jeśli są one subtelne). Cechy w przestrzeni czasu wykazują również wrażliwość na szum występujący w tle. Większą niezawodnością w modelowaniu zmian w sygnale akustycznym charakteryzują się cechy widmowe [57].

Tab. 5.1: Zestawienie cech akustycznych w dziedzinie czasu [2, 5, 41, 63, 91, 102, 107].

Cecha	**Symbol**	**Wzór/definicja**
Liczba przejść przez zero (ang. *zero-crossing rate*)	ZCR_t	$ZCR_t = \frac{1}{2N}\sum_{n=1}^{N-1}\left\|sgn\left[x\left(n\right)\right] - sgn\left[x\left(n-1\right)\right]\right\|,$ (5.4) $sgn\left[x\left(n\right)\right] = \begin{cases} 1, & x\left(n\right) \geq 0, \\ -1, & x\left(n\right) < 0. \end{cases},$ (5.5) x — N-próbkowy sygnał czasowy ramki.
Krótkoczasowa energia (ang. *short-term energy*)	STE_t	$STE_t = \frac{1}{N}\sum_{n=0}^{N-1}\left\|x\left(n\right)\right\|^2$ (5.6)
Częstotliwość podstawowa (ang. *pitch*)	P_t	Częstotliwość podstawowa głosu (ton podstawowy, formant zerowy), obliczona metodą znormalizowanej funkcji korelacji [5].
Stosunek składowych harmonicznych (ang. *harmonic ratio*)	HR_t	$HR_t(m) = \frac{\sum_{n=1}^{N} x(n)x(n-m)}{\sqrt{\sum_{n=1}^{N} x(n)^2 \sum_{n=1}^{N} x(n-m)^2}},$ gdzie $1 \leq m \leq M$ (5.7) M — maksymalne opóźnienie w obliczeniach (maksymalna wartość to 40, która odpowiada częstotliwości podstawowej 25 kHz).

5.2.2 Cechy częstotliwościowe w pełnym pasmie

Cechy spektralne i cepstralne (tab. 5.2) stanowią istotny element analizy sygnału mowy. Dziedzina częstotliwości jest analizowana w celu obserwacji okresowości i składu widmowego sygnału. Ponieważ parametry spektralne uważa się za odzwierciedlenie związku pomiędzy zmieniającym się traktem głosowym a dźwiękiem, analiza tych cech może być użyteczna w kontekście przetwarzania sygnału mowy. W literaturze można znaleźć wykorzystanie podstawowych cech spektralnych, zwłaszcza: środka ciężkości widma, skośności widmowej czy szerokości, rozproszenia i płaskości widma.

Grupą parametrów częstotliwościowych, której użyteczność jest szczególnie istotna w akustyce mowy, są współczynniki melcepstralne (ang. *mel-frequency cepstral coefficients*, MFCC). Współczynniki MFCC uwzględniają charakterystykę ludzkiego słuchu poprzez wykorzystanie filtrów rozłożonych równomiernie na skali melowej. Aby je uzyskać, należy przeprowadzić pięć operacji:

1) ramkowanie sygnału;
2) zastosowanie dyskretnej transformaty Fouriera DFT;
3) przekształcenie z wykorzystaniem banku filtrów melowych w celu otrzymania współczynników Y_k energii banku filtrów (ang. *filter bank energies*, FBE): typowy zestaw filtrów melowych składa się z 40 filtrów trójkątnych, a zmiana jednostek częstotliwości z herców na mele jest zgodna z:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right), \tag{5.8}$$

   gdzie f to wartości w hercach;
4) obliczenie logarytmów wartości współczynników Y_k;
5) zastosowanie dyskretnej transformaty kosinusowej (ang. *discrete cosine transform*, DCT) w celu otrzymania współczynników MFCC:

$$MFCC_i = \sqrt{\frac{2}{N_F}} \sum_{k=1}^{N_F-1} \left(\ln Y_k \cos\left(\frac{(2k-1)\, i\pi}{2N_F}\right), \quad 0 \leqslant i \leqslant N_F - 1, \right. \tag{5.9}$$

   gdzie i jest numerem współczynnika, a N_F to liczba użytych filtrów. W pracy zastosowano 40 filtrów trójkątnych pokrywających częstotliwości od 0 do 22 kHz. Do analizy wybrano 13 współczynników o najniższych indeksach: $MFCC_0$–$MFCC_{12}$.

Tab. 5.2: Zestawienie cech akustycznych w dziedzinie częstotliwości [60, 71, 101, 119, 131].

Cecha	Symbol	Wzór
Środek ciężkości widma (ang. *spectral centroid*)	$SCen_f$	$SCen_f = \mu_1 = \frac{\sum_{k=b_1}^{b_2} f_k s_k}{\sum_{k=b_2}^{b_1} s_k}$, (5.10) s_k — k-ty współczynnik amplitudy widma dla częstotliwości f_k, b_1, b_2 — numery skrajnych współczynników widma dla analizowanego pasma częstotliwości; tu: $b_1 = 0$, $b_2 = N - 1$ (pełne widmo).
Rozproszenie widmowe (ang. *spectral spread*)	$SSpr_f$	$SSpr_f = \sqrt{\frac{\sum_{k=b_1}^{b_2} (f_k - \mu_1)^2 s_k}{\sum_{k=b_1}^{b_2} s_k}}$. (5.11)
Skośność widmowa (ang. *spectral skewness*)	SSk_f	$SSk_f = \frac{\sum_{k=b_1}^{b_2} (f_k - \mu_1)^3 s_k}{(\mu_2)^3 \sum_{k=b_1}^{b_2} s_k}$. (5.12)
Grzbiet widma (ang. *spectral crest*)	SCr_f	$SCr_f = \frac{\max_{k \in [b_1, b_2]} (s_k)}{\frac{1}{b_2 - b_1} \sum_{k=b_1}^{b_2} s_k}$. (5.13)
Współczynnik spadku widma (ang. *spectral decrease*)	SD_f	$SD_f = \frac{\sum_{k=b_1+1}^{b_2} \frac{s_k - s_{b_1}}{k-1}}{\sum_{k=b_1+1}^{b_2} s_k}$. (5.14)
Entropia widma (ang. *spectral entropy*)	SE_f	$SE_f = \frac{-\sum_{k=b_1}^{b_2} s_k \log(s_k)}{\log(b_2 - b_1)}$. (5.15)

Kontynuacja tabeli na następnej stronie

Cecha	Symbol	Wzór
Płaskość widma (ang. *spectral flatness*)	$SFla_f$	$SFla_f = \frac{\left(\prod_{k=b_1}^{b_2} s_k\right)^{\frac{1}{b_2-b_1}}}{\frac{1}{b_2-b_1}\sum_{k=b_1}^{b_2} s_k}.$ (5.16)
Strumień widmowy (ang. *spectral flux*)	$SFlx_f$	$SFlx_f(t) = \left(\sum_{k=b_1}^{b_2} \lvert s_k(t) - s_k(t-1)\rvert^p\right)^{\frac{1}{p}},$ (5.17) tu: $p = 2$.
Szerokość pasma (ang. *spectral rolloff point*)	SRP_f	$SRP_f = i$, gdzie $\sum_{k=b_1}^{i} \lvert s_k\rvert = \kappa \sum_{k=b_1}^{b_2} s_k,$ (5.18) κ — próg łącznej energii pasma; tu: $\kappa = 0,95$.
Kurtoza widma (ang. *specral kurtosis*)	$SKurt_f$	$SKurt_f = \frac{\sum_{k=b_1}^{b_2}(f_k - SCen_f)^4 s_k}{(SSpr_f)^4 \sum_{k=b_1}^{b_2} s_k}$ (5.19)
Współczynnik nachylenia widma (ang. *spectral slope*)	SSl_f	$SSl_f = \frac{\sum_{k=b_1}^{b_2}(f_k - \mu_f)(s_k - \mu_s)}{\sum_{k=b_1}^{b_2}(f_k - \mu_f)^2},$ (5.20) μ_f — średnia częstotliwość, μ_s — średnia wartość amplitudy widma.
Współczynniki MFCC	$MFCC_i$	$MFCC_i = \sqrt{\frac{2}{N_F}} \sum_{k=1}^{N_F-1} (\ln Y_k \cos\left(\frac{(2k-1)\,i\pi}{2N_F}\right),$ (5.21) $0 \leqslant i \leqslant N_F - 1,$ Y_k — k-ty współczynnik energii banku filtrów, N_F — liczba filtrów trójkątnych; tu: $N_F = 40$.

5.2.3 Cechy częstotliwościowe w pasmie szumu

Oprócz podstawowych parametrów akustycznych do opisu wymowy sybilantów zaczęto wykorzystywać właściwości związane z występowaniem szumu turbulentnego (pasmo powyżej 2 kHz [90] lub 3 kHz [155]) charakterystycznego dla tych głosek [99, 124, 166]. Z początku opis dotyczył głównie ich wymowy normatywnej, dopiero później badacze zaczęli rozpatrywać parametry szumowe jako elementy istotne dla zagadnień diagnostyki i terapii logopedycznej wad wymowy. W literaturze można znaleźć prace obejmujące różne aspekty akustyczne wymowy u polskich dzieci w wieku przedszkolnym [90, 99, 100]. W jednej z prac [90] badano akustykę sybilantów w mowie dzieci bez stwierdzonych zaburzeń mowy. Analizę oparto na formantach szumowych, a jej wynikiem było stwierdzenie wolniejszego doskonalenia przez dzieci artykulacji afrykatów w porównaniu z fonemami frykatywnymi. W kolejnych badaniach eksperymenty wykazały istotne statystycznie różnice pomiędzy realizacjami zębowymi i międzyzębowymi w przypadku niektórych parametrów akustycznych, uwzględniając głoski /ʂ/, /ʐ/ [99] oraz /s/ i /ts/ [100]. W drugim przypadku zaobserwowano również występowanie różnic w rozkładach cech pomiędzy głoską szczelinową a zwarto-szczelinową. Użyteczność wykorzystania parametrów szumowych jest zatem poparta literaturą. Wybrane przykłady parametrów tej grupy wymieniono w rozdziale 1.3.1, a zbiór, który wykorzystano w pracy przedstawiono w tab. 5.3.

Tab. 5.3: Zestawienie cech akustycznych związanych z szumem towarzyszącym głoskom dentalizowanym [96, 99, 139].

Cecha	Symbol	Wzór/definicja
Częstotliwości pierwszych formantów szumowych	NFF_i	Częstotliwości czterech pierwszych formantów w pasmie szumu powyżej 1900 Hz, obliczone metodą opartą o liniowe kodowanie predykcyjne (ang. *linear predictive coding*, LPC) [96, 99, 139].
Amplitudy pierwszych formantów szumowych	$NFFL_i$	$$NFFL_i = \frac{1}{3}\sum_{n=k-1}^{k+1} \lvert X(n)\rvert, \quad 1 \le i \le 4, \qquad (5.22)$$ $X(n)$ — n-ty prążek widma DFT sygnału w ramce, n — prążek odpowiadający częstotliwości NFF_i.
Stosunek częstotliwości formantów	$NFFR_{ij}$	$$NFFR_{ij} = \frac{FF_i}{FF_j}, \quad 1 \le i \le 3, \quad i+1 \le j \le 4, \qquad (5.23)$$

Kontynuacja tabeli na następnej stronie

Cecha	Symbol	Wzór/definicja
Stosunek amplitud formantów	$NFFLR_{ij}$	$NFFLR_{ij} = \frac{FFL_i}{FFL_j}, \quad 1 \leq i \leq 3, \quad i+1 \leq j \leq 4,$ (5.24)
Współczynniki cepstralne szumu	NCC_k	$NCC_k = \sqrt{\frac{2}{N_n}} \sum_{n=n_d}^{n_g} \ln\left(\lvert X(n)\rvert\right) \cos\frac{\pi k\,(2n+1)}{2N_n}, \quad k = 0, 1, ..., n_g - n_d$ (5.25) n_d, n_g — numery prążków o częstotliwościach najbliższych f_d, f_g, $f_d = FF_1 - f_b, \quad f_d = FF_4 + f_b$, f_b — margines o szerokości 500 Hz, $N_n = n_g - n_d + 1$ — liczba prążków widma szumu.
Energie pasm szumu	NE_k	$NE_k = \sum_{n=n_{d,k}}^{n_{g,k}} \lvert X(n)\rvert^2, k = 0, 1, 2, ..., K-1$ (5.26) K — liczba podpasm o szerokości 500 Hz w zakresie od 2000 do 7000 Hz, $n_{g,k}, n_{d,k}$ — numery prążków DFT reprezentujące częstotliwości graniczne k-tego podpasma.
Odległości pomiędzy formantami szumowymi	$NFFD_{ij}$	$NFFD_{i,j} = NFF_i - NFF_j, \quad 1 \leq i \leq 3, j = i+1$ (5.27)
Najwyższa amplituda (w zakresie 2-7 kHz)	NPA	$NPA = \max_{f \in (2,7)kHz} X(f)$ (5.28)
Częstotliwość odpowiadająca NPA	NPF	$NPF = \arg \max_{f \in (2,7)kHz} X(f)$ (5.29)

5.2.4 Ekstrakcja cech akustycznych dla celów komputerowego wsparcia diagnostyki logopedycznej

Zgodnie z zasadą wspomnianą w rozdziale 4, z analizy odrzucono wektory cech dla skrajnych ramek (25% początkowych i 25% końcowych) we wszystkich segmentach. Dla każdej przetwarzanej ramki w segmencie mowy ekstrahowano

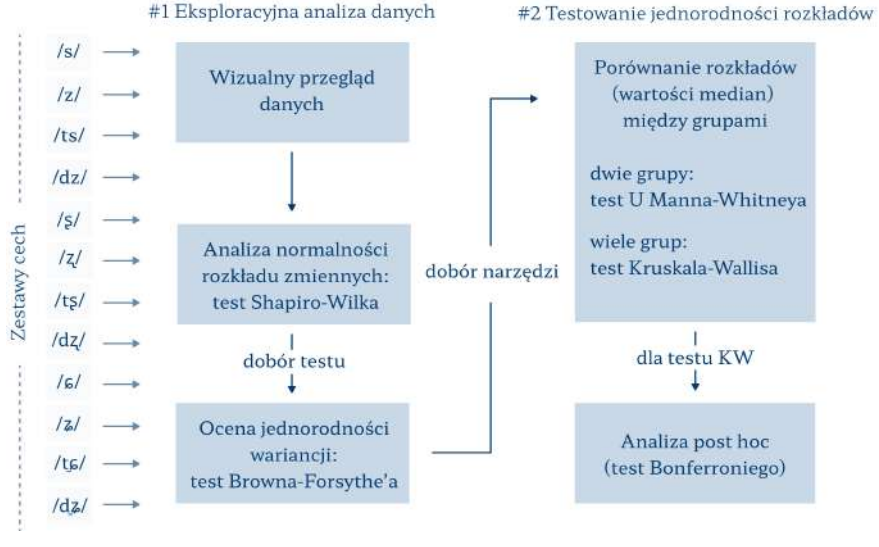
Tab. 5.4: Zestawienie cech akustycznych analizowanych w pracy.

Typ	**Grupa**	#	**Łącznie**	**Definicje**
Czasowe			4	Tab. 5.1
Częstotliwościowe	Spektralne	11	24	Tab. 5.2
	Melcepstralne	13		
Szumowe			48	Tab. 5.3
		Łącznie	**76**	

wektor 76 cech akustycznych: 4 cech czasowych, 11 cech spektralnych, 13 współczynników MFCC oraz 48 cech wynikających z szumu (tab. 5.4).

6. Analiza wizualno-akustyczno-artykulacyjna

Ostatni z etapów niniejszej pracy obejmował określanie cech sygnałowych niosących istotną statystycznie informację dotyczącą artykulacji. Wykorzystana ścieżka analizy była dwuetapowa (rys. 6.1). Najpierw przeprowadzono eksplorację danych w celu określenia rozkładów zmiennych i opracowania hipotez dotyczących wartości cech sygnałów wizualno-akustycznych w wybranych grupach. Drugi etap obejmował weryfikację hipotez z wykorzystaniem testów statystycznych. Poziom istotności α dla wszystkich opisanych w tym rozdziale eksperymentów wynosił 0,05, a interpretację założonego w pracy stopnia korelacji Spearmana ρ oraz przyjętych poziomów wielkości efektów (w przypadku testu Kruskala-Wallisa obliczanej za pomocą η^2, a dla testu U Manna-Whitneya — współczynnika korelacji dwuseryjnej rb) zebrano w tab. 6.1 [26, 132].



Rys. 6.1: Schemat dwuetapowej analizy statystycznej dotyczącej zależności między parametrami wizualno-akustycznymi i cechami artykulacyjnymi. Analizę oparto na metodach nieparametrycznych z uwagi na brak normalności rozkładu wielu zmiennych.

Tab. 6.1: Interpretacja współczynnika korelacji Spearmana (ρ) oraz wielkości efektu w wykorzystywanych testach statystycznych: współczynnik korelacji dwuseryjnej (rb) dla testu U Manna-Whitneya i η^2 w przypadku testu Kruskala-WallisaT.

	Współczynnik korelacji Spearmana ρ*	**Wielkość efektu**	
		rb*	η^2
mała	<0,39	<0,39	<0,01
średnia	0,40–0,59	0,40–0,59	0,02–0,06
duża	>0,60	>0,60	>0,14
* rozpatrywany jest moduł wartości			

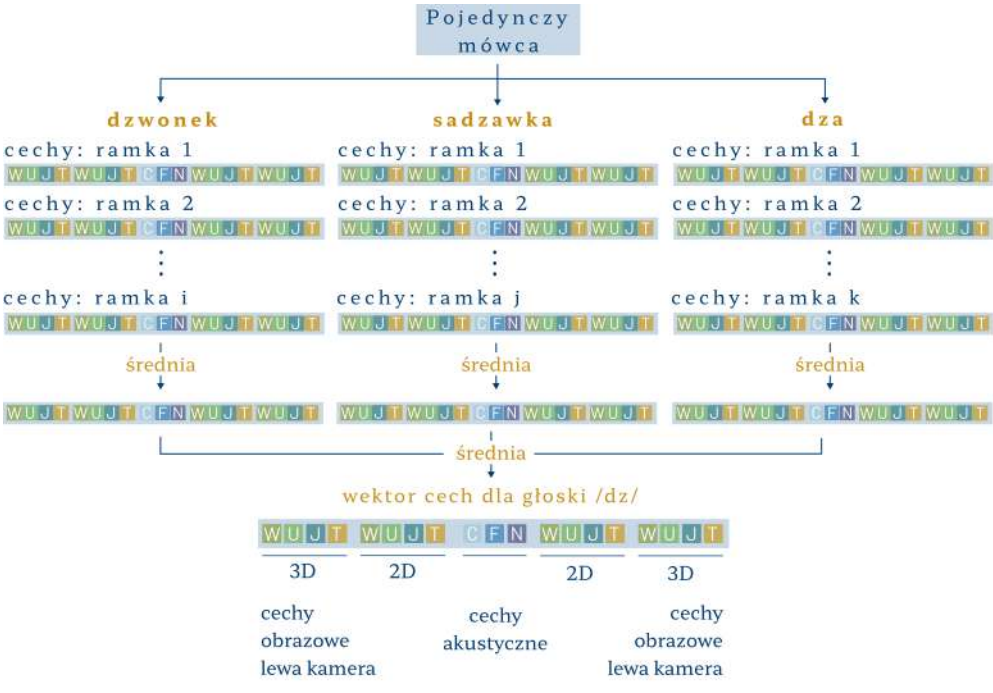
6.1 Zbiór danych

Analizę przeprowadzono osobno dla każdej z 12 głosek dentalizowanych. W każdym przypadku podstawą poszukiwania zależności były wektory 472 cech wizualno-akustycznych: 174 parametrów wizualnych 2D (po 87 dla każdej z kamer), 228 wizualnych 3D (podobnie, po 114 cechy dla pojedynczych kamer) oraz 76 parametrów akustycznych dla środkowego mikrofonu macierzy. Zgodnie z opisem przedstawionym w rozdziałach 4 i 5, parametry dotyczyły środkowych ramek głoski (przedział: 25%-75% wszystkich ramek dla danego fragmentu).

Obserwacja statystyczna odnosi się do pojedynczej realizacji zmiennej — jednostki, pojedynczego zjawiska lub jednego punktu danych — które są wykorzystywane do analizy statystycznej. W pracy jako obserwacje rozumie się zestaw cech dla pojedynczego mówcy. Badani wypowiadali każdą z głosek kilkukrotnie, co więcej, w przypadku cech akustycznych oraz obrazowych-dwuwymiarowych, wektory liczone były dla każdej ramki danego fonemu. Chcąc zapewnić niezależność danych, parametry dla poszczególnych głosek zostały uśrednione względem mówców na etapie przygotowywania danych do analizy (rys. 6.2). Zastosowanie funkcji agregującej gwarantowało, że poszczególne obserwacje nie pochodziły od tego samego mówcy.

Liczba mówców różniła się pomiędzy poszczególnymi głoskami (tab. 6.2). Dysproporcje wynikały z odrzucenia obserwacji, które wykazywały niewystarczającą jakość segmentacji. Największą liczbą obserwacji pochodzących od niezależnych mówców charakteryzowała się głoska /s/ (183), a najmniejszą — fonem /ʑ/ (149). Łączna liczba unikalnych mówców w analizowanej bazie wyniosła 195. Głoską, która najczęściej występowała (tab. 6.3) dla wymienionej liczby mówców było /z/, najmniej — /dʑ/ (odpowiednio 2 047 i 332). Są to sumy wystąpień poszczególnych fonemów dla wszystkich dzieci, które finalnie poddano agregacji tak, aby dla każdego dziecka i danej głoski otrzymać jeden wektor cech wizualnych i akustycznych.

Dla każdej głoski wybrano także cechy artykulacyjne. Sprawdzono liczebność poszczególnych grup w przypadku każdego z parametrów logopedycz-



Rys. 6.2: Schemat agregacji cech na przykładzie pojedynczego mówcy i głoski /dz/.

	/s/	/z/	/ts/	/dz/
Liczba mówców	183	166	178	167
	/ʂ/	/ʐ/	/tʂ/	/dʐ/
Liczba mówców	178	167	162	156
	/ɕ/	/ʑ/	/tɕ/	/dʑ/
Liczba mówców	164	149	154	155
Liczba unikalnych mówców				195

Tab. 6.2: Liczba analizowanych mówców dla każdej z głosek.

nych. Odrzucono cechy, w których zdecydowanie dominowała jedna grupa. Nie uwzględniano również mało licznych grup wewnątrz parametrów (poniżej 10 obserwacji). Badanie logopedyczne, według protokołu opisanego w rozdziale 3, obejmowało zbiór kilkudziesięciu cech dotyczących motoryki narządów mowy i poprawności artykulacji. Po konsultacjach z logopedami do analizy wytypowano 7 parametrów dotyczących wymiaru artykulacyjnego oraz jedną cechę opisującą aspekt anatomiczny (stopień skrócenia wędzidełka językowego). Wybrane cechy uznano za istotne z punktu widzenia wstępnych badań w zakresie opisywanej tematyki. Zgodnie z opisem zawartym w rozdziale 1.1 (tab. 1.2), kryterium wyłączającym część parametrów była specyfika głosek, zwłaszcza

	/s/	/z/	/ts/	/dz/	Razem
Liczba głosek	1 334	832	1 049	545	3 760
	/ʂ/	/ʐ/	/tʂ/	/dʐ/	Razem
Liczba głosek	2 047	1 310	627	522	4 506
	/ɕ/	/ʑ/	/tɕ/	/dʑ/	Razem
Liczba głosek	756	513	335	332	1 936
Suma wszystkich analizowanych głosek					10 202

Tab. 6.3: Liczba wystąpień analizowanych głosek.

w kontekście miejsca ich artykulacji. Zgodnie z tym, dentalność oceniana jest jedynie w przypadku głosek, których miejscem artykulacji są zęby (głoski dentalne, szereg syczący), postdentalność charakteryzuje głoski dziąsłowe (szereg szumiący), a palatalność — głoski palatalne (szereg ciszący). Palatalność jednak nie była analizowana ze względu na przewagę normatywności artykulacji wśród badanych dzieci. Uproszczoną charakterystykę cech wybranych do analizy zamieszczono w tab. 6.4 [112, 113], a podsumowanie rozkładu ich wartości w zbiorze danych umieszczono w tab. 6.5.

6.2 Eksploracyjna analiza danych

W pierwszej kolejności przeprowadzono wizualną analizę danych. Ocenie poddano histogramy każdej ze zmiennych dla poszczególnych głosek. Zgrubna analiza wykazała przewagę rozkładów niesymetrycznych (lewostronnych i prawostronnych). Ponieważ dobór narzędzi przetwarzania w kolejnych krokach analizy opierał się na własnościach rozkładów badanych zmiennych, przeprowadzono serię eksperymentów, w zamierzeniu mających zweryfikować poprawność subiektywnych wniosków. Dla każdej rozpatrywanej zmiennej przeprowadzono test Shapiro-Wilka (SW) [137] w celu określenia normalności rozkładu. Poziom istotności α wynosił 0,05. Testowi poddano wszystkie zmienne dla każdej głoski, wybranych cech artykulacyjnych oraz ich grup. Ze względu na dużą liczbę eksperymentów (łączna liczba cech obrazowo-akustycznych wynosiła 472, a cech artykulacyjnych 8), podsumowanie wyników testu Shapiro-Wilka umieszczono w dodatku A.1 (tab. A.2). Dla większości przypadków test SW dał podstawę do odrzucenia hipotezy zerowej o normalności rozkładu. W przypadku cech o dużej skośności, w dalszej analizie wykorzystywano logarytm ich wartości. Ze względu na dominację rozkładów asymetrycznych, w dalszych krokach założono przeprowadzanie wyłącznie testów nieparametrycznych.

Tab. 6.4: Opis wybranych cech artykulacyjnych oraz analizowanych w pracy grup [112, 113].

	Cecha	**Opis cechy**	**Grupy**	**Opis grup**
#1	Dentalizacja	zbliżenie górnych i dolnych siekaczy	norma	dentalizacja; szczelina między łukami zębowymi
			dysdentalizacja pionowa	cecha niepożądana; zbyt słaba dentalizacja
#2	Dentalność	miejsce artykulacji	norma	dentalność; kontakt przedniej (apikalnej) części języka z górnymi siekaczami
			międzyzębowość	cecha niepożądana; kontakt górnej i dolnej powierzchni języka z krawędziami zębów
#3	Postdentalność	miejsce artykulacji	norma	postdentalność; kontakt przedniej części języka z dziąsłami za górnymi siekaczami
			międzyzębowość	cecha niepożądana; kontakt górnej i dolnej powierzchni języka z krawędziami zębów
			zadziąsłowość	kontakt apikalnej części języka z granicą tylnej części dziąseł i początkiem prepalatum
			zębowość	kontakt przedniej części języka z górnymi siekaczami
#4	Apikalność	pozycja języka w trakcie artykulacji	norma	apikalność; kontakt apeksu języka z podniebienną powierzchnią górnych siekaczy
			dorsalność	kontakt grzbietowej części języka
#5	Skrócenie wędzidełka języka	stopień skrócenia wędzidełka języka	norma	ruch języka nie jest zauważalnie ograniczony
			nieznacznie	ruch języka jest nieznacznie ograniczony
			średnio	ruch języka jest ograniczony w średnim stopniu
#6	Medialność wypływu powietrza	kierunek wypływu powietrza	norma	pośrodkowy przepływ powietrza
			lewostronny	lewostronny wypływ powietrza
			prawostronny	prawostronny wypływ powietrza
#7	Medialność języka	ułożenie języka	norma	symetria kontaktujących się narządów
			dysmedialność lewostronna	lewostronne ułożenie języka (szpara ust szersza po prawej stronie)
			dysmedialność prawostronna	prawostronne ułożenie języka (szpara ust szersza po lewej stronie)
#8	Medialność żuchwy	ułożenie żuchwy	norma	symetria kontaktujących się narządów
			dysmedialność lewostronna	lewostronne ruchy żuchwy w trakcie artykulacji
			dysmedialność prawostronna	prawostronne ruchy żuchwy w trakcie artykulacji

Tab. 6.5: Wybrane cechy artykulacyjne i rozkład ich wartości w zbiorze danych.

	Cecha	**Grupy**	**Obserwacje**											
			/s/	/z/	/ʦ/	/ʣ/	/ʂ/	/ʐ/	/ʈʂ/	/ʈʐ/	/ɕ/	/ʑ/	/ʨ/	/ʥ/
#1	Dentalizacja	norma	115	104	111	104	125	119	112	105	109	101	105	105
		dysdentalizacja pionowa	38	37	38	35	32	29	28	29	35	30	30	31
#2	Dentalność	norma	113	103	112	104								
		międzyzębowość	31	27	31	28								
#3	Postdentalność	norma					106	97	97	91	144			
		międzyzębowość							24		13			
		zadziąsłowość					29	28	25	24				
		zębowość					27	26		24				
#4	Apikalność	norma	63	59	66	62	130	124	118	116	78	75	82	81
		dorsalność	116	104	108	102	37	33	34	28	81	71	68	70
#5	Skrócenie wędzidełka języka	norma	61	57	58	58	62	62	55	54	58	50	52	51
		nieznacznie	70	63	68	63	69	60	61	55	63	55	39	59
		średnio	45	40	45	41	40	41	40	41	41	39	37	40
#6	Medialność wypływu powietrza	norma	119	110	116	113					137	123	129	129
		lewostronna	23	19	20	18						18		
		prawostronna	36	33	37	32					18		16	18
#7	Medialność języka	norma	118	111	115	106	149		135		136	125	129	129
		dysmedialność lewostronna	23	19	21	19	12		12					
		dysmedialność prawostronna	38	33	38	39	12		11		20	17	17	18
#8	Medialność żuchwy	norma	120	111	118	112	129	120	119	115	119	110	112	
		dysmedialność lewostronna	23	21	21	18	15	16	15	16	16	13	15	
		dysmedialność prawostronna	28	24	28	27	29	25	21	20	26	22	23	

Kolejny krok analizy dotyczył oceny jednorodności skal pomiędzy grupami za pomocą testu homogeniczności wariancji. Wybrano nieparametryczny test Browna-Forsythe'a (BF) [14]. Chociaż przeważały wartości p powyżej założonego progu 0,05, które dawały podstawę do przyjęcia hipotezy zerowej (o jednorodności wariancji), dla niektórych cech wynik sugerował heterogeniczność tej miary. W tych sytuacjach, obliczano stosunek wariancji parametrów pomiędzy każdym z wariantów danej cechy artykulacyjnej. Sprawdzono w ten sposób stopień różnorodności skal. Z dalszej analizy odrzucono cechy, dla których wartość stosunków wariancji między grupami przekroczyła 10 lub była mniejsza niż 0,1 (w przypadku wielogrupowych analiz, odrzucano parametr nawet, jeśli zasada została złamana tylko pomiędzy jedną parą). Dodatkowe obliczenia potwierdziły, że pomimo heterogeniczności wariancje dla poszczególnych grup są wartościami tego samego rzędu. Zebrane wyniki testów umieszczono w dodatku A.2 i ograniczono je jedynie do cech, które okazały się istotne statystycznie w dalszych etapach analizy.

6.3 Testowanie jednorodności rozkładów

Aby ocenić zdolności dyskryminacyjne poszczególnych cech, wykorzystano zestaw testów statystycznych sprawdzających równość rozkładów zmiennych. Ze względu na przewagę asymetrycznych rozkładów zmiennych, nie było możliwości zastosowania analizy średnich. Analizę parametryczną zastąpiono testami nieparametrycznymi, pozwalającymi na ocenę równości median: w przypadku binarnym jest to test U Manna-Whitneya (U MW) [34, 137], a dla wieloklasowych problemów realizowany jest test Kruskala-Wallisa (KW) [33, 137]. Istotny statystycznie wynik testu KW zwraca jednak jedynie informację o tym, że co najmniej jedna z badanych grup różni się od innej. Aby wyłonić, które dokładnie grupy różnią się między sobą, należy przeprowadzić test post-hoc. W przypadku tej pracy zastosowano test Bonferroniego [32]. Analiza wyników wymienionych testów pozwala na zbadanie istotności różnic pomiędzy medianami parametrów w grupach mówców o odmiennych sposobach realizacji głosek dentalizowanych. Na tej podstawie można wnioskować, czy istnieją cechy obrazowe lub akustyczne, które mogłoby skutecznie i rzetelnie różnicować różne warianty wymowy, zwłaszcza normatywne i patologiczne.

7. Eksperymenty i wyniki

Etapy pośrednie, opisane w poprzednich rozdziałach, wymagały wykonania serii eksperymentów i walidacji jakości wyników. W celu zweryfikowania dokładności działania opracowanej metody detekcji i segmentacji artykulatorów przeprowadzono zestaw rozłącznych eksperymentów dla obu algorytmów wykorzystując jednakowy zbiór danych. Analiza identycznych obrazów zawartych w zbiorze testowym (rozdział 4.1, rys. 4.2) symulowała realne i zamierzone działanie metody oraz zapewniała wiarygodność walidacji. Ocenie poddano:

- algorytm detekcji artykulatorów oparty na YOLO (rozdział 7.1), gdzie
  - porównano efektywność działania różnych modeli sieci YOLO (rozdział 7.1.1) oraz
  - zbadano wpływ IoU na wyniki detekcji (rozdział 7.1.2);
- algorytm segmentacji artykulatorów opartej na sieci DeepLabv3+ (rozdział 7.2) oraz
  - poszczególnych etapów pośrednich słabo-etykietowanej metody uczenia (rozdział 7.2.1), a także
  - jakości działania sieci w zależności od architektury rdzenia sieci DeepLabv3+ (rozdział 7.2.2).

Finalizacją ścieżek przetwarzania obrazów i sygnałów akustycznych była ekstrakcja cech. Wektory parametrów poddano analizie statystycznej w celu poszukiwania zależności pomiędzy ich elementami a opisem logopedycznym (rozdział 7.3).

7.1 Detekcja artykulatorów w obrazach

Trening modelu przeznaczonego do detekcji artykulatorów był powtarzany kilkukrotnie z uwzględnieniem zmian wybranych parametrów. Poszukiwano w ten sposób ustawień umożliwiających optymalne działanie sieci poparte metrykami jakościowymi. Modyfikowano dwa aspekty: wybór architektury sieci YOLO oraz dopasowanie progu współczynnika IoU (t_{IoU}) dla ramek wynikowych.

Opis kolejnych kroków walidacyjnych wymaga przybliżenia wykorzystanych metryk jakościowych. Wśród miar skuteczności detekcji znaleźć można [62, 114]: krzywą precyzji względem czułości (ang. *precision vs. recall curve*), średnią precyzję (ang. *average precision*, AP) dla każdej z klas, uśrednioną średnią precyzję (ang. *mean average precision*, mAP), wskaźnik F1 (ang. *F1 score*), współczynnik części wspólnej do sumy ramek IoU (ang. *intersection over union*; znany również jako indeks Jaccarda). Rozwinięcie poszczególnych miar wykorzystanych w pracy zestawiono w tab. 7.1.

Tab. 7.1: Zestawienie wybranych miar jakości działania sieci do detekcji obiektów [115].

Miara	**Symbol**	**Wzór**
Precyzja (ang. *precision*)	P	$P = \frac{TP}{TP + FP}$, (7.1) TP, FP — liczba prawdziwie i fałszywie pozytywnych detekcji.
Czułość (ang. *recall*)	R	$R = \frac{TP}{TP + FN}$, (7.2) FN — liczba fałszywie negatywnych detekcji.
Wskaźnik F1 (ang. *F1 score*)	$F1$	$F1 = \frac{2PR}{P + R}$ (7.3)
Średnia precyzja (ang. *average precision*)	AP	$AP = \sum_n (R_{n+1} - R_n)\, P_{interp}(R_{n+1})$, (7.4) n — iterator po przedziałach czułości przy obliczaniu powierzchni pod krzywą precyzji względem czułości, $P_{interp}(R_{n+1}) = \max_{\tilde{R}:\tilde{R} \gtrless R_{n+1}} P\left(\tilde{R}\right)$. (7.5)
Uśredniona średnia precyzja (ang. *mean average precision*)	mAP	$mAP = \frac{1}{N} \sum_{i=0}^{N} AP_i$, (7.6) AP_i — średnia precyzja detekcji i-tego spośród N typów obiektów.

Kontynuacja tabeli na następnej stronie

Miara	Symbol	Wzór
Współczynnik części wspólnej do sumy ramek (ang. *intersection over union*)	IoU	$IoU = \frac{\lvert\text{Obiekt} \bigcap \text{Wynik detekcji}\rvert}{\lvert\text{Obiekt} \bigcup \text{Wynik detekcji}\rvert}$, (7.7) Obiekt i Wynik detekcji to stosowne prostokątne ramki obejmujące dany obiekt, a wykorzystane operacje mierzą powierzchnię ich części wspólnej i sumy.

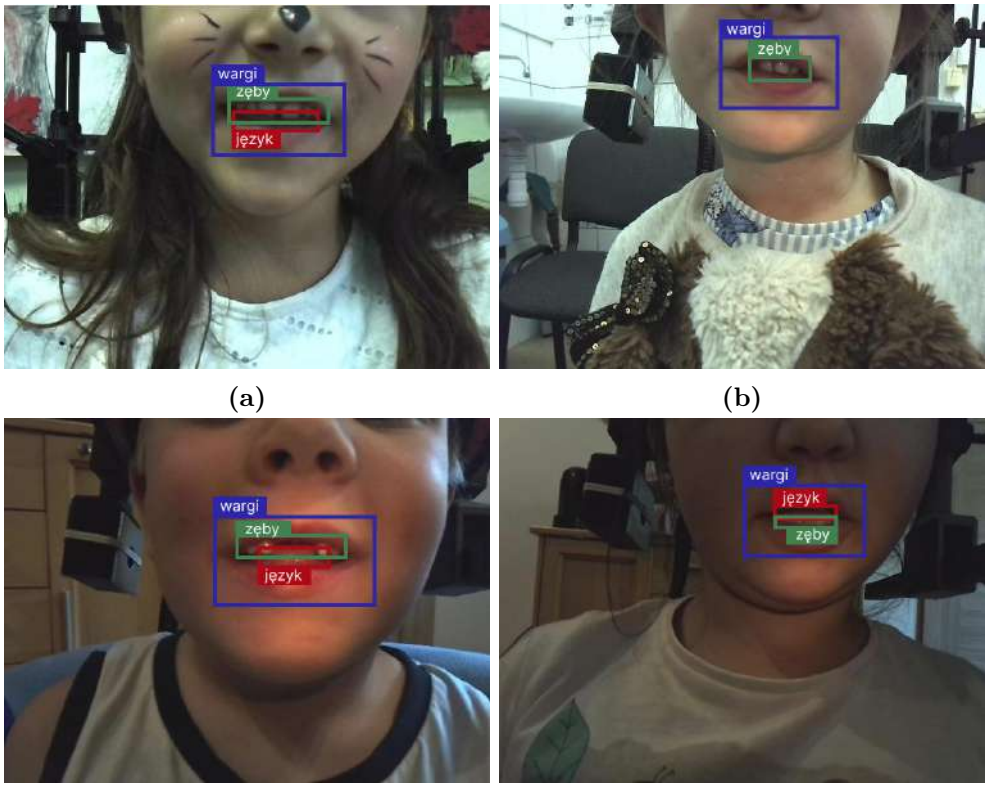
Zadaniem modelu była skuteczna detekcja trzech artykulatorów na obrazach twarzy dzieci: warg, zębów oraz języka. Wykorzystanie wykrywania dwóch ostatnich narządów ograniczało się jedynie do przygotowania zgrubnie etykietowanego zbioru danych do wstępnego uczenia sieci DeepLabv3+. Detekcja warg była wykorzystywana zarówno do opracowania niedokładnych obrysów, jak i w ostatecznym działaniu algorytmu. Przykładowe wyniki detekcji na losowych ramkach przedstawiono na rys. 7.1.

7.1.1 Wpływ architektury YOLO na skuteczność działania

W pracy wykonano serię eksperymentów oceniających jakość działania czterech wybranych wersji sieci YOLO. Wybrano jedną ze starszych wariantów oraz trzy nowsze (na czas opracowywania metody): YOLO v3 [123], v5 [59], v6 [74] oraz v7 [161]. Sieci wytypowano bazując na przeglądzie literatury, zwłaszcza na podstawie prac porównujących działanie różnych modeli. Tab. 7.2 podsumowuje wartości metryk (średnią precyzję oraz współczynnik $F1$) dla wybranych modeli charakteryzujących się zbliżonym ustawieniem hiperparametrów dla progu $t_{IoU} = 0,5$. Wartości AP uzyskane dla szóstej wersji architektury (YOLOv6) w przypadku języka oraz wartości średniej dla wszystkich klas są nieznacznie wyższe względem pozostałych modeli (odpowiednio 0,651 i 0,798). Niższa wartość dla detekcji zębów nie jest istotna ze względu na najmniejszą istotność i wkład tego artykulatora w kolejne etapy przetwarzania. Z tego względu do dalszych eksperymentów oraz rozwoju metody segmentacji wybrano model YOLOv6.

7.1.2 Wpływ współczynnika IoU na wyniki detekcji

Współczynnik IoU pozwala na określenie stopnia poprawności, z jakim model potrafi odróżnić obiekty od tła (lub innych obiektów) na obrazie. Próg IoU (t_{IoU}) określa minimalną wartość indeksu, aby prostokąt będący wynikiem detekcji nie został odrzucony. Dobór progu pozwalana na podjęcie próby

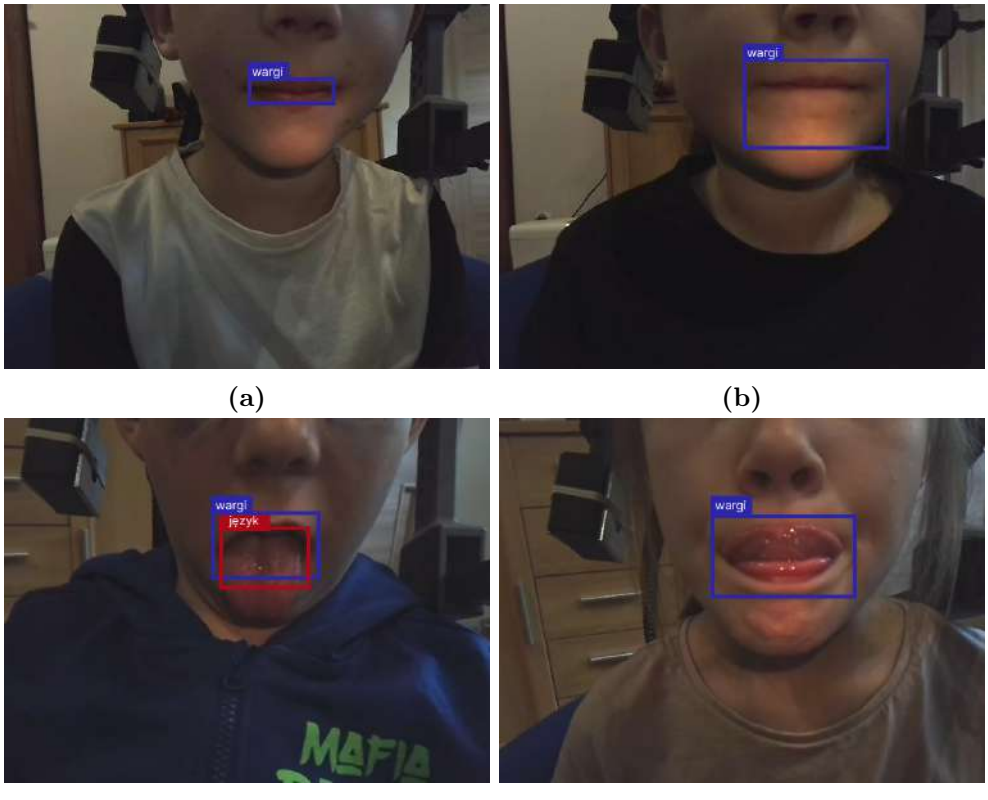


Rys. 7.1: Przykłady detekcji artykulatorów.

Tab. 7.2: Zestawienie działania wybranych modeli architektury YOLO przy progu $t_{IoU} = 0,50$. Wartości wyróżnione kolorem niebieskim wskazują na najlepsze wyniki.

	Model	Wargi	Zęby	Język	Średnia
AP	YOLOv3	0,999	0,612	0,276	0,629
	YOLOv5	0,999	**0,774**	0,532	0,768
	YOLOv6	0,999	0,745	**0,651**	**0,798**
	YOLOv7	**1,000**	0,744	0,624	0,789
*F*1	YOLOv3	**0,999**	0,645	0,336	0,641
	YOLOv5	**0,999**	0,749	0,543	0,753
	YOLOv6	**0,999**	0,794	**0,664**	0,798
	YOLOv7	**0,999**	**0,822**	0,605	**0,800**

ustalenia kompromisu pomiędzy dokładnością wykrywania a stopą fałszywych detekcji. W przypadku zagadnienia detekcji narządów artykulacyjnych jakość pokrywania właściwego obszaru rzutuje na kolejne kroki przetwarzania — zbyt duży region zainteresowania będzie skutkował przetwarzaniem nadmiaru nie-

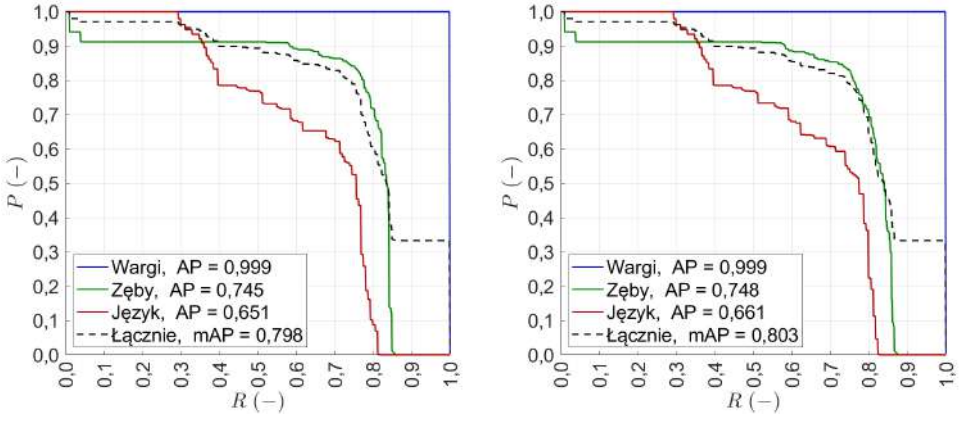


Rys. 7.2: Przykłady błędnej detekcji artykulatorów, która może mieć wpływ na dalsze kroki przetwarzania lub uczenia sieci: (a) zaznaczenie zbyt małego obszaru warg, (b) zaznaczenie zbyt dużego obszaru warg, (c) poprawna detekcja obszaru ust przy niewłaściwym wskazaniu regionu mocno wysuniętego języka, (d) brak detekcji języka, który pojawia się w obszarze górnej wargi.

istotnego tła, zbyt mały, z kolei, nie pozwoli na poprawne wysegmentowanie artykulatorów (rys. 7.2).

Rys. 7.3 przedstawia krzywe precyzji względem czułości dla progu t_{IoU} wynoszącego 0,5 oraz 0,75. Zaprezentowano wynik dla każdego z obiektów indywidualnie oraz średnią wartość AP. W tab. 7.3 zaprezentowano wskaźnik $F1$ oraz AP dla każdej z klas z podziałem na dwa wcześniej wymienione progi t_{IoU}: 0,5 oraz 0,75. Wykres pudełkowy prezentujący rozkład wartości IoU pozytywnych detekcji zastosowanych w dalszych krokach do uczenia sieci do segmentacji artykulatorów przedstawiono na rys. 7.4.

Detekcja warg charakteryzowała się najwyższą skutecznością spośród wszystkich klas. Wysoką jakość procesu poparły prawie idealne metryki wskaźnika $F1$ oraz średniej precyzji AP — w obu przypadkach wynoszącej 0,999 zarówno



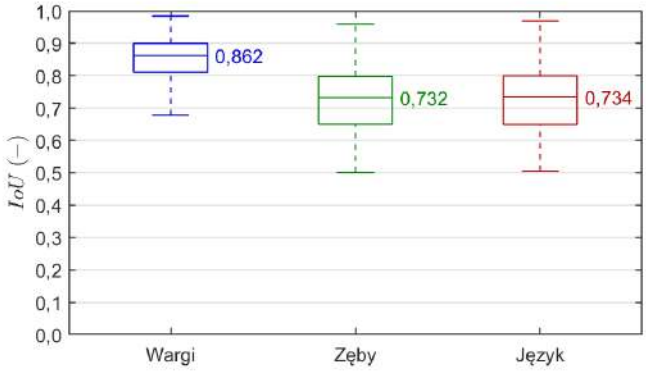
(a) $t_{IoU} = 0.50$ **(b)** $t_{IoU} = 0.75$

Rys. 7.3: Krzywa precyzji względem czułości otrzymana dla modelu YOLOv6 dla dwóch progów t_{IoU}: (a) 0,50 i (b) 0,75.

Tab. 7.3: Zestawienie wyników detekcji artykulatorów z uwzględnieniem różnych wartości progu t_{IoU}. Kolorami zaznaczono najwyższe wartości pomiędzy progami t_{IoU} dla poszczególnych artykulatorów: zielonym metrykę AP, niebieskim wartość $F1$.

t_{IoU}	Metryka	**Wargi**	**Zęby**	**Język**	**Średnia**
0,50	AP	**0,999**	0,745	0,651	0,798
	F1	**0,999**	**0,794**	**0,664**	**0,798**
0,75	AP	**0,999**	**0,748**	**0,661**	**0,803**
	F1	**0,999**	0,788	0,659	**0,798**

dla t_{IoU} równej 0,5, jak i 0,75. W trakcie eksperymentów zaobserwowano też, że wykrywanie warg jest w najmniejszym stopniu wrażliwe na zmiany progu t_{IoU} oraz innych ustawień. Niewiele przypadków wykazywało wartości IoU poniżej 0,8 (rys. 7.4). Na wysoką zgodność wykrywania wskazuje również mediana równa 0,862. Jest to istotne w ogólnym ujęciu problemu — obszar okalający wargi/usta jest kluczowy dla dalszego przetwarzania, m.in. ze względu na określanie danych wejściowych do etapu segmentacji (zarówno przygotowania zgrubnego zbioru danych do wstępnego uczenia sieci, jak i działania gotowego modelu). Pozostałe artykulatory (zęby, język) są wykrywane z dokładnością, która jest wystarczająco wysoka, aby umożliwić przygotowanie niedokładnych etykiet do słabo nadzorowanego uczenia sieci, otrzymanych za pomocą metody DRLSE (rozdział 4.4.4).



Rys. 7.4: Wykres pudełkowy wartości IoU w przypadku detekcji pojedynczych klas ($t_{IoU} = 0,50$). Każde z pudełek obejmuje rozstęp ćwiartkowy (IQR) z zaznaczeniem mediany wskazanej przez linię poprowadzoną wewnątrz. Wąsy obejmują $1,5 \cdot \mathrm{IQR}$.

7.2 Segmentacja artykulatorów

Kolejnym krokiem pośrednim, a zarazem realizacją tezy pomocniczej, jest przygotowanie wiarygodnej metody segmentacji narządów mowy. Na testowanie modelu segmentacyjnego składały się dwa eksperymenty: po pierwsze, każdy z kroków pośrednich metody segmentacji narządów mowy (z modelem finalnym włącznie) został oceniony, aby zweryfikować słuszność zastosowania wieloetapowości; po drugie, porównano skuteczność działania różnych rdzeni sieci DeepLabv3+ w celu wybrania najbardziej efektywnego.

Do ewaluacji modeli segmentacyjnych wykorzystuje się inne metryki oceny jakości niż w przypadku problemu detekcji. W pracy zastosowano zestaw podstawowych miar (tab. 7.4): współczynnik podobieństwa Dice'a (ang. *Dice similarity coefficient*, DSC) oraz dokładność (ang. *accuracy*, Acc). Sprawdzono skuteczność działania zarówno finalnej wersji modelu, jak i każdego z etapów pośrednich wykorzystując jednakowy zbiór testowy.

7.2.1 Ocena etapów metody

Za pomocą podzbioru danych C (rys. 4.2), na który składało się 665 obrazów zebranych od 16 mówców, sprawdzono jakość segmentacji każdego z trzech bloków stanowiących algorytm segmentacji (rys. 4.10a): (1) wstępnej segmentacji z wykorzystaniem metody DRLSE, (2) zgrubnie uczonego modelu sieci DeepLabv3+ oraz (3) dostrojonej sieci DeepLabv3+ z poprzedniego kroku (wersji ostatecznej modelu).

Wyniki testowania zaprezentowano na rys. 7.5. Oprócz warg, zębów oraz języka uwzględniono również usta, które obejmowały także piksele należące do szpary międzywargowej (rys. 4.14). Wyniki segmentacji konsekwentnie ulegały

Tab. 7.4: Zestawienie wybranych miar jakości działania sieci do segmentacji.

Miara	Symbol	Wzór
Dokładność (ang. *accuracy*)	Acc	$Acc = \frac{TP + TN}{TP + TN + FP + FN}$, (7.8) TP, TN, FP, FN — liczba pikseli prawdziwie pozytywnych, prawdziwie negatywnych, fałszywie pozytywnych i fałszywie negatywnych w wyniku segmentacji.
Współczynnik podobieństwa Dice'a (ang. *Dice similarity coefficient*)	DSC	$DSC = \frac{2TP}{2TP + FP + FN}$ (7.9)

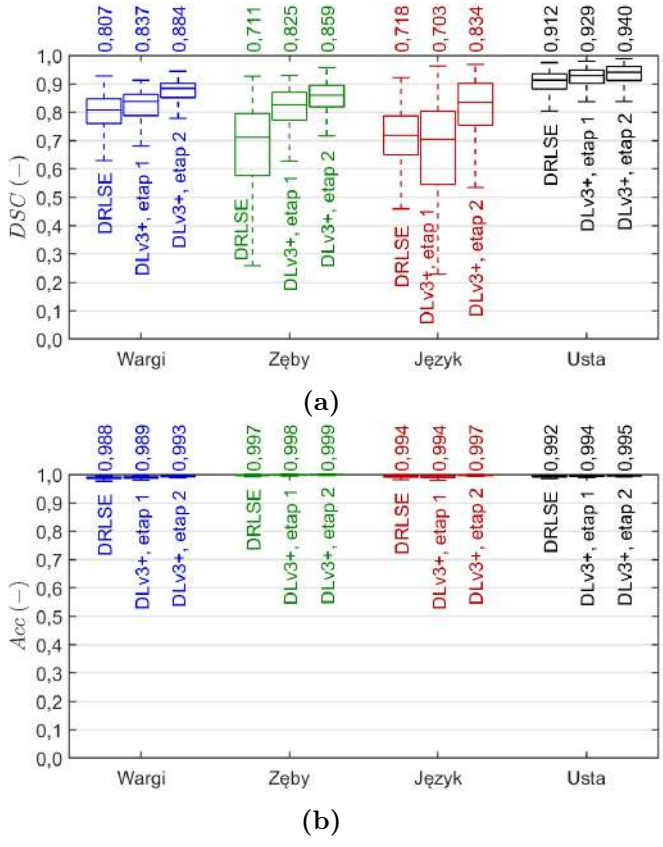
poprawie w kolejnych etapach algorytmu. Wyjątek stanowi język, dla którego jakość procesu nieznacznie spada w przypadku etapu uczenia słabo nadzorowanego (2) i znacznie wzrasta po dostrojeniu sieci (3).

Na rys. 7.6 przedstawiono rezultaty przetwarzania identycznego zdjęcia dla modeli z różnych opisywanych etapów metody. Poprawa jakości idzie w parze z postępami algorytmu. Najniższą skutecznością charakteryzowała się segmentacja języka. Trudność w segmentacji języka może wynikać z jego stosunkowo rzadkiej obecności na nagraniach, nieregularnych kształtach czy niewielkim obszarze w trakcie wymowy. Ponadto często jego obecność i dokładny obszar — ze względu na niewielkie wymiary i brak możliwości dobrego oświetlenia wnętrza jamy ustnej — może różnić się między ekspertami. Proces segmentacji, bez względu na artykulator, stawał się problematyczny w nietypowym oświetleniu (np. prześwietlone lub bardzo ciemne nagrania, nadmierne jasne lub ciemne obszary na twarzy), przy niewyraźnych krawędziach czy częściowo przysłoniętych narządach. Biorąc pod uwagę liczbę wszystkich dostępnych obrazów, błędnie lub w ogóle niewysegmentowanych przypadków było stosunkowo niewiele.

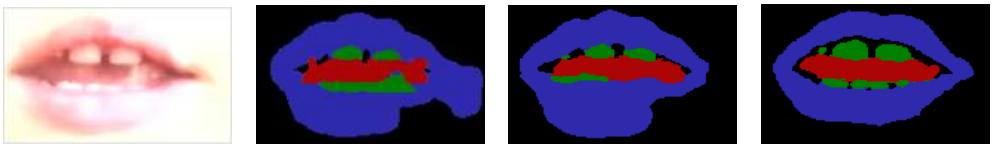
7.2.2 Wpływ architektury rdzenia DeepLabv3+ na jakość działania sieci

Budowa sieci DeepLabv3+ pozwala na wykorzystanie różnych modeli sieci CNN zawartych wewnątrz sieci nadrzędnej. Ich głównym zadaniem jest ekstrakcja cech. Dobór architektury zależy od rozpatrywanego problemu oraz charakteru dostępnego zbioru danych. W pracy porównano cztery konstrukcje: ResNet-101, ResNet-152, Xception oraz MobileNet v2 [22, 51, 128]. W doborze sieci kierowano się raportowaną w literaturze skutecznością poszczególnych modeli.

Wpływ modeli na jakość działania zestawiono na rys. 7.7 względem uzy-



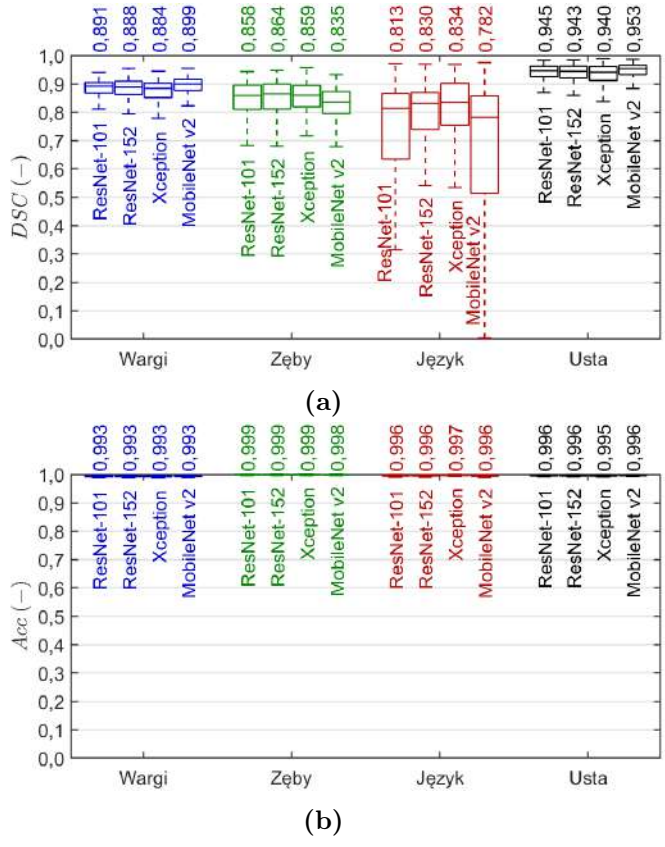
Rys. 7.5: Wykresy pudełkowe wyników segmentacji dla kolejnych etapów przetwarzania: (a) współczynnik podobieństwa Dice'a i (b) dokładność dla każdego z etapów procesu. Każde z pudełek obejmuje rozstęp ćwiartkowy (IQR) z zaznaczeniem mediany wskazanej przez linię poprowadzoną wewnątrz. Wąsy obejmują $1,5 \cdot \mathrm{IQR}$.



(a) **(b)** **(c)** **(d)**

Rys. 7.6: Przykłady wyników segmentacji dla kolejnych etapów metody: a) ROI ograniczone do ust po zastosowaniu YOLO, b) wynik segmentacji metodą DRLSE, c) wynik segmentacji DLv3+ po uczeniu zgrubnym zbiorem danych, d) wynik segmentacji DLv3+ po dostrojeniu sieci.

skanego współczynnika Dice'a oraz dokładności. Najwyższa mediana *DSC* wyniosła 0,953 dla ust, 0,899 dla warg, 0,864 dla zębów i 0,834 w przypadku języka (przy czym dokładność przekroczyła 0,99 dla każdego z artykulatorów).



Rys. 7.7: Wykresy pudełkowe wyników segmentacji dla różnych rdzeni modelu DeepLabv3+: (a) współczynnik podobieństwa Dice'a oraz (b) dokładność po ukończeniu pełnego słabo nadzorowanego uczenia sieci DeepLabv3+ z różnymi architekturami rdzenia. Każde z pudełek obejmuje rozstęp ćwiartkowy (IQR) z zaznaczeniem mediany wskazanej przez linię poprowadzoną wewnątrz. Wąsy obejmują $1,5 \cdot \mathrm{IQR}$.

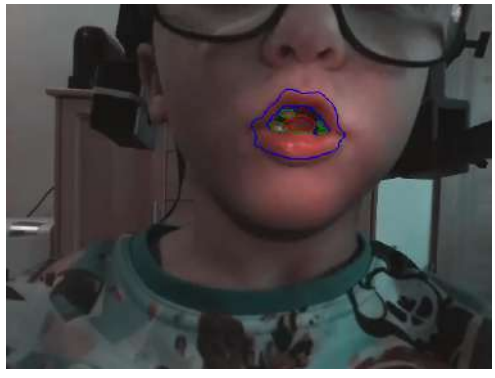
Wyniki uzyskano dla obrazów, w których odnotowano udane wykrycie obiektów, tj. liczba pikseli oznaczonych jako prawdziwie poprawne (ang. *true positive*) była niezerowa. W przeciwnym przypadku, uwzględniając również niewłaściwą detekcję, mediana współczynnika Dice'a pozostaje jednakowa dla ust oraz warg, nieznacznie spada dla zębów (0,857), jednak zauważalnie obniża się w przypadku języka (0,796). W porównaniu z innymi artykulatorami, jego segmentacja jest szczególnie problematyczna. Stosunkowo niska jasność szpary między zębami górnymi a dolnymi, nieregularność kształtu, kolor zbliżony do warg, trudność w określeniu jednoznacznych krawędzi, liczne artefakty (związane np. z występowaniem śliny odbijającej światło), rzadkość pojawiania się języka oraz brak powtarzającego się wzorca wyglądu oraz motoryki wśród różnych mówców sprawiają, że przygotowanie modelu, który wykazywałby wysoką

jakość działania w różnorodnych przypadkach jest złożonym zadaniem. Uczenie sieci wymagałoby większej liczby przypadków w różnorodnych warunkach w zbiorze treningowym sieci. Mimo wszystko, warto zauważyć, że w ewaluacji jakości działania sieci w bazie danych pojawiały się także słabo widoczne wystąpienia języka.

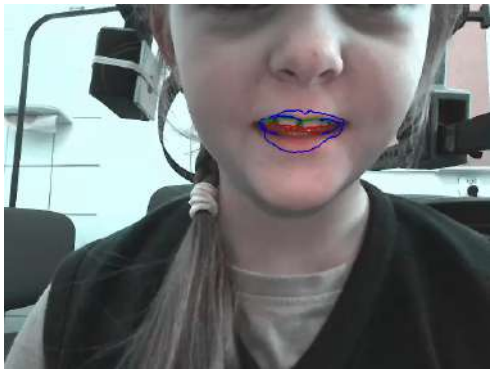
Nauczone modele przejawiają w swoim działaniu kompromis pomiędzy wydajnością dla dużych i małych obiektów (rys. 7.8). Biorąc pod uwagę rozważane aspekty oraz otrzymane wyniki, architekturą rdzenia, która została ostatecznie wdrożona, była Xception. Zestawienie wartości DSC dla poszczególnych obiektów i architektur (rys. 7.7a) pokazuje, że różnice median pomiędzy sieciami są stosunkowo niewielkie, z najlepszymi wynikami dla: ResNet-152 i Xception. Przewaga modelu Xception jest widoczna w przypadku segmentacji języka. Xception został opracowany w oparciu o ideę sieci ResNet, co sprawia, że oprócz dziedziczenia ich zalet, dodaje również swoje, m.in. związane ze zamianą standardowych warstw splotowych na warstwy rozdzielne w głąb (ang. *depthwise separable convolution*). Ponadto, jest architekturą płytszą niż ResNet-152 i Resnet-101, co zmniejsza liczbę parametrów i złożoność obliczeniową oraz może stanowić zaletę w przypadku charakterystyki wykorzystanej bazy danych.

Opisane do tej pory eksperymenty oraz wyniki opracowanych metod detekcji i segmentacji sugerują, że algorytm był wystarczająco wiarygodny i stabilny. Pozwala to na przeprowadzenie kolejnych kroków prowadzących do weryfikacji tez postawionych na początku pracy. Zdarzały się przypadki, w których algorytm segmentacji wskazywał błędne obszary, włączając w to: obiekty wysegmentowane niecałkowicie, o bardzo nieregularnych kształtach nieoddających rzeczywistości, pojawianie się niewłaściwej etykiety w danym regionie (np. piksele opisane jako język w sferze ust) czy rozlanie segmentacji. Nierzadko przyczyną trudności była specyfika nagrania, np. prześwietlone lub za ciemne obrazy, cień na twarzy czy przypadkowo poruszony obraz. Wpływ mogły mieć też modyfikacje wprowadzane w urządzeniu pomiarowym. Do uczenia modelu sieci neuronowej wykorzystano obrazy z pierwszej wersji maski, a w kolejnych wariantach wprowadzono m.in. punktowe oświetlenie obszaru okolic ust. Niedokładne rezultaty segmentacji mogłyby w efekcie zafałszować wyniki analizy statystycznej poprzez ekstrakcję cech w niewłaściwym obszarze lub nieoddającym rzeczywistości kształcie obiektu. Dlatego przed przystąpieniem do wyznaczania parametrów wizualno-dźwiękowych i późniejszej eksploracji uzyskanych danych, oceniono segmentację każdej realizacji głosek dentalizowanych u wszystkich dzieci. Do dalszych kroków przetwarzania wykorzystano jedynie fonemy, których obrysy były subiektywnie zakwalifikowane jako co najmniej bardzo dobre (poziom 0,9 w skali 0–1, ok. 86% przypadków). Umożliwiło to rzetelną analizę ewentualnych relacji pomiędzy parametrami wizualno-akustycznymi a opisem artykulacyjnym.

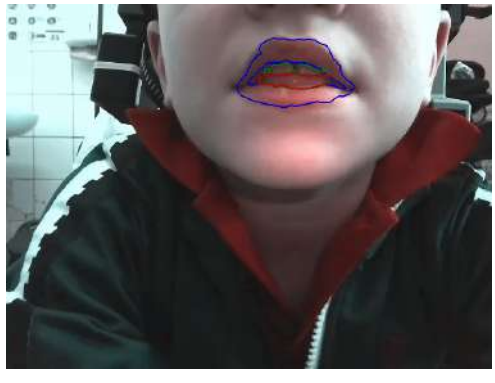
(a) /ʂ/ w słowie „*szufelka*”



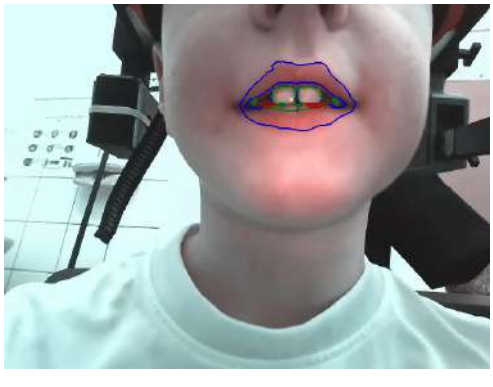
(b) /ʂ/ w słowie „*koszyk*”



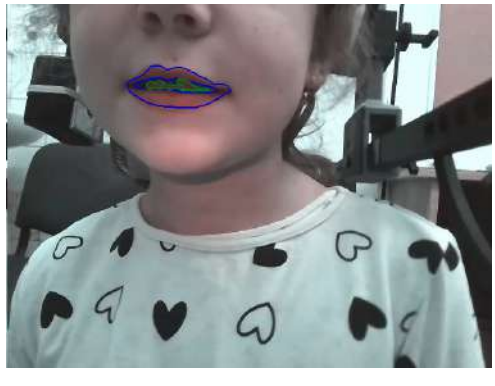
(c) /s/ w słowie „*pies*”



(d) /ts/ w słowie „*taca*”



(e) /s/ w słowie „*pies*”



(f) /ts/ w słowie „*pajac*”

Rys. 7.8: Przykłady wyników segmentacji dla losowych ramek w trakcie realizacji różnych fonemów dentalizowanych: kolor niebieski obramowuje wargi, zielony zęby, a czerwony język.

7.3 Analiza statystyczna

Testom poddano wektory wszystkich otrzymanych parametrów wizualnych i akustycznych oraz poszczególne cechy artykulacyjne (tab. 6.5). Kryteria wyboru charakterystyk artykulacyjnych opisano w rozdziale 6. Analiza składała się z dwóch etapów. W pierwszej kolejności przeprowadzono eksplorację danych (rozdział 7.3.1), która wykazała przewagę rozkładów asymetrycznych wśród cech. W kolejnym kroku testowano jednorodność rozkładów pomiędzy grupami dla wybranych cech artykulacyjnych (rozdział 7.3.2) — wyniki wcześniejszego etapu narzuciły wykorzystanie testów nieparametrycznych (test U Manna-Whitneya oraz test Kruskala-Wallisa).

7.3.1 Analiza eksploracyjna danych

Wstępna (wizualna) eksploracja danych sugerowała, że większość zmiennych wizualno-akustycznych nie podlegała rozkładowi normalnemu, a dodatkowo część z nich charakteryzowała się dużą skośnością. Aby potwierdzić przypuszczenia, wykonano analizę normalności rozkładu zmiennych dla każdej z głosek wykorzystując test Shapiro-Wilka. Wyniki umieszczono w dodatku A.1. Dominowały cechy, w przypadku których odrzucono hipotezę zerową o normalności rozkładu.

Kolejny krok analizy obejmował ocenę jednorodności skal pomiędzy grupami za pomocą testu homogeniczności wariancji Browna-Forsythe'a. Jedynie w nielicznych przypadkach rezultaty sugerowały odrzucenie hipotezy zerowej o jednorodności tej miary. Ponieważ w takich przypadkach heterogeniczność wariancji i brak normalności rozkładów ogranicza możliwości analizy międzygrupowej median (lub średnich), sprawdzono dla nich stopień różnorodności skal za pomocą stosunku wariancji zmiennych. Eksperyment wykazał, że wariancje dla poszczególnych grup są wartościami tego samego rzędu. Zebrane wyniki testów umieszczono w dodatku A.2.

Przed przygotowaniem wektorów wejściowych do opracowywania i testowania hipotez sprawdzono zależności pomiędzy odpowiadającymi sobie cechami w obrazach zarejestrowanych przez kamerę lewą i prawą. Ze względu na przewagę rozkładów niesymetrycznych zastosowano korelację Spearmana. Średni współczynnik korelacji ρ i procent cech o wysokiej lub średniej zależności sugerują, że parametry pomiędzy kamerami są do siebie zbliżone (tab. 7.5). Jest to widoczne zwłaszcza dla cech teksturowych. Wśród cech związanych z kształtem obiektów wysoką korelacją charakteryzują się parametry ust i warg (2D i 3D). Niższe wartości współczynników ρ wykazuje język, szczególnie w kontekście parametrów trójwymiarowych. Prawie połowa cech (43,8%) jest skorelowana w niewielkim stopniu. Trudność w segmentacji języka wynikająca m.in. z niejednoznaczności jego konturu oraz niewielkiego (często przysłoniętego) obszaru

Tab. 7.5: Podsumowanie analizy korelacji Spearmana pomiędzy jednakowymi cechami obrazowymi uzyskanymi dla kamery lewej i prawej. Interpretację zakresów umieszczono w tab. 6.1.

W	cechy kształtu warg					
U	cechy kształtu ust					
J	cechy kształtu języka					
T	cechy teksturowe					
	Rodzaj cechy		μ_ρ	Odsetek cech		
				wysokie ρ	średnie ρ	małe ρ
Cechy kształtu	U	2D	0,96 ± 0,02	100%		
	W	2D	0,90 ± 0,07	100%		
	J	2D	0,47 ± 0,16	63,1%	36,8%	
	U	3D	0,94 ± 0,05	100%		
	W	3D	0,88 ± 0,12	100%		
	J	3D	0,41 ± 0,19	18,8%	37,5%	43,8%
Cechy teksturowe	T	2D	0,96 ± 0,05	100%		
	T	3D	0,94 ± 0,08	100%		

mogła skutkować różnicami w uzyskanych kształtach masek pomiędzy kamerami. Uwzględniając rezultaty testu istnieje możliwość ograniczenia wybranych parametrów teksturowych do jednej kamery. Zdecydowano się jednak poddać analizie wszystkie cechy mające wystarczającą liczbę obserwacji dla danego parametru artykulacyjnego.

7.3.2 Testowanie jednorodności rozkładów

Kolejny etap obejmował weryfikację postawionych hipotez. Przeprowadzono porównanie rozkładów (wartości median) pomiędzy grupami dla wybranych cech artykulacyjnych. W przypadku binarnym (dwie grupy), zastosowano test U Manna-Whitneya. Dla analiz wielogrupowych wykorzystano test Kruskala-Wallisa, po którym dodatkowo wykonywano analizę post hoc testem Bonferroniego. Szczegółowe rezultaty dla każdej z głosek oraz każdej z charakterystyk artykulacyjnych można znaleźć w dodatku B. W niniejszym rozdziale zdecydowano się zaprezentować jedynie zbiorcze tabele podsumowujące uzyskane wyniki. Podział na poziomy wielkości efektów (mały, średni, wysoki rozmiar efektu) umieszczono w tab. 6.1.

Szereg syczący

W przypadku szeregu syczącego (tab. 7.6) dominują cechy związane z kształtem artykulatorów (zakładając jedynie rezultaty o co najmniej średniej wielkości efektu). Wśród nich przeważają parametry opisujące język (23), choć

w przypadku głoski /z/ pojawiają się również dotyczące geometrii warg (2). Większość cech wideo jest trójwymiarowa (15 cech 3D i 10 cech 2D). Wystąpiły również pojedyncze cechy spośród pozostałych kategorii: jedna teksturowa (I_h^{3D}, w przypadku głoski /z/) oraz akustyczna-szumowa (NCC_{10}, dla głoski /dz/).

Tab. 7.6: Zestawienie występowania cech akustyczno-wizualnych o wartości p poniżej 0,05 i co najmniej średniej wielkości efektu w przypadku głosek z szeregu syczącego. Cechy akustyczne i obrazowe zaznaczono symbolami: *A* i *V*. Kolorami rozróżniono rodzaje cech obrazowych i akustycznych. Szczegółowe wyniki porównania rozkładów dla poszczególnych głosek zawierają tab. B.2–B.9 w dodatku B.1.

Kategorie cech:

obrazowe:
- W kształtu warg
- U kształtu ust
- J kształtu języka
- T teksturowe

akustyczne:
- C czasowe
- F częstotliwościowe
- N szumowe

Cechy artykulacyjne:
- #1 dentalizacja
- #2 dentalność
- #3 postdentalność
- #4 apikalność
- #5 skrócenie wędzidełka językowego
- #6 medialność wypływu powietrza
- #7 medialność języka
- #8 medialność żuchwy

Cecha	Dane	2D/3D	Rodzaj	/s/			/z/				/ts/					/dz/					
				#1	#2	#6	#1	#2	#5	#7	#1	#2	#4	#5	#7	#1	#2	#4	#6	#7	#8
A^{2D}	V	2D	W						x												
A^{2D}	V	2D	J		x	x				x						x	x	x			
Ax^{2D}_{major}	V	2D	J																x		
Ax^{2D}_{minor}	V	2D	J		x											x	x				
D^{2D}_{Feret}	V	3D	J		x												x		x		x
DS^{2D}	V	2D	W						x												
DS^{2D}	V	2D	J													x	x	x			
E^{2D}	V	2D	J											x				x			
P^{2D}	V	2D	J			x				x						x	x	x			
S^{2D}	V	2D	J													x	x	x			
A^{3D}	V	3D	J									x	x		x		x			x	x
Ax^{3D}_{least}	V	3D	J	x	x							x					x			x	
Ax^{3D}_{major}	V	3D	J		x						x	x	x				x				
Ax^{3D}_{minor}	V	3D	J	x	x							x	x		x		x				

Kontynuacja tabeli na następnej stronie

Cecha	Dane	2D/3D	Rodzaj	/s/			/z/				/ts/					/dz/					
				#1	#2	#6	#1	#2	#5	#7	#1	#2	#4	#5	#7	#1	#2	#4	#6	#7	#8
C_1^{3D}	V	3D	J											x		x	x	x			
C_2^{3D}	V	3D	J											x		x	x	x			
D_{YZ}^{3D}	V	3D	J	x	x							x	x		x		x				
D_{Feret}^{3D}	V	3D	J		x							x	x				x				
D_{XZ}^{3D}	V	3D	J		x			x				x	x				x				
D_{XY}^{3D}	V	3D	J		x			x				x	x				x				
DS^{3D}	V	3D	J									x		x			x			x	
E^{3D}	V	3D	J												x						
S^{3D}	V	3D	J											x							
SVR^{3D}	V	3D	J		x																
V^{3D}	V	3D	J	x	x		x	x				x	x		x		x				
I_h^{3D}	V	3D	T						x												
NCC_{10}	A		N																		x

Dostrzec można pewne powtarzające się wzorce pomiędzy różnymi głoskami, zwłaszcza w przypadku dentalności (#2). W tym przypadku, dla głosek /s/, /ts/ i /dz/ powtarza się 9 cech kształtu 3D (w przypadku głoski /z/ obserwuje się jedynie 3 wspólne parametry), pomiędzy /s/ i /dz/ dodatkowo jeszcze 4 cechy kształtu 2D. Występowanie tych samych parametrów można również zauważyć pomiędzy różnymi cechami artykulacyjnymi — do najliczniej występujących cech można zaliczyć objętość wokseli V^{3D} (8), później pole powierzchni pikseli A^{3D}, długość osi mniejszej Ax_{minor}^{3D}, maksymalna średnica w płaszczyźnie czołowej D_{YZ}^{3D} i pole powierzchni A^{2D} (wszystkie po 6 wystąpień).

Analiza post hoc (poprzedzona testem Kruskala-Wallisa) wykonana dla istotnych statystycznie cech wizualno-akustycznych również sugerowała występowanie różnic w rozkładach par poszczególnych grup. W przypadku głoski /s/ (tab. B.3) i medialności wypływu powietrza, różnice obserwowane były pomiędzy parami: *norma-wypływ lewostronny* oraz *wypływ prawostronny-wypływ lewostronny*. Dla tej samej cechy artykulacyjnej, ale w przypadku głoski /dz/ (tab. B.9), różnice zauważalne są jedynie dla pierwszej pary (*norma-wypływ lewostronny*). W przypadku cechy dotyczącej medialności języka, dla głoski /z/ różnice można zauważyć tylko pomiędzy wymową normatywną (*norma*) a *dysmedialnością prawostronną*, podczas gdy dla głoski /ts/ pojawiają się we wszystkich parach, choć dominuje wskazanie różnic pomiędzy *dysmedialnością lewostronną* a *dysmedialnością prawostronną*. Dla głoski /dz/ różnice obserwuje

się tylko pomiędzy obiema dysmedialnościami. Medialność żuchwy analizowana była jedynie przy realizacji głoski /dz/, a różnice obserwowane są między każdą z branych pod uwagę par. Podobnie w przypadku głoski /z/ i oceny stopnia skrócenia wędzidełka językowego — tu również każde z zestawień wykazywało co najmniej jedną istotną statystycznie różnicę w rozkładach median między grupami.

Szereg szumiący

W szeregu szumiącym (tab. 7.7) przeważają cechy akustyczne (łącznie 39 cech: 26 szumowych, 12 częstotliwościowe i 1 czasowa). Cech obrazowych o co najmniej średniej wielkości efektu zgromadzono 25: 16 trójwymiarowych i 9 dwuwymiarowych. W tej grupie dominują parametry kształtu opisujące język, choć w pojedynczych przypadkach pojawiają się również cechy warg i ust (głoska /tʂ/). Cechy teksturowe (5) pojawiają się głównie w przypadku dentalności (#2) głoski /tʂ/. Jednorazowe wystąpienie zauważa się w przypadku stopnia skrócenia wędzidełka językowego (#5) głoski /ʐ/.

Tab. 7.7: Zestawienie występowania cech akustyczno-wizualnych o wartości p poniżej 0,05 i co najmniej średniej wielkości efektu w przypadku głosek z szeregu szumiącego. Cechy akustyczne i obrazowe zaznaczono symbolami: *A* i *V*. Kolorami rozróżniono rodzaje cech obrazowych i akustycznych. Szczegółowe wyniki porównania rozkładów dla poszczególnych głosek zawierają tab. B.10–B.14 w dodatku B.2.

Kategorie cech:

obrazowe:
- W — kształtu warg
- U — kształtu ust
- J — kształtu języka
- T — teksturowe

akustyczne:
- C — czasowe
- F — częstotliwościowe
- N — szumowe

Cechy artykulacyjne:
- #1 dentalizacja
- #2 dentalność
- #3 postdentalność
- #4 apikalność
- #5 skrócenie wędzidełka językowego
- #6 medialność wypływu powietrza
- #7 medialność języka
- #8 medialność żuchwy

Cecha	Dane	2D/3D	Rodzaj	/ʂ/			/ʐ/		/tʂ/			/dʐ/			
				#3	#5	#8	#3	#5	#3	#5	#7	#1	#3	#4	#5
A^{2D}	V	2D	J		×					×					
Ax^{2D}_{major}	V	2D	J					×						×	

Kontynuacja tabeli na następnej stronie

Cecha	Dane	2D/3D	Rodzaj	/ʂ/			/ʐ/		/tʂ/			/dʐ/			
				#3	#5	#8	#3	#5	#3	#5	#7	#1	#3	#4	#5
D^{2D}_{Feret}	V	2D	J					x							
DS^{2D}	V	2D	J		x										
P^{2D}	V	2D	J		x										
A^{3D}	V	3D	J		x										
Ax^{3D}_{least}	V	3D	J					x				x	x		
Ax^{3D}_{least}	V	3D	U						x						
Ax^{3D}_{least}	V	3D	W						x						
Ax^{3D}_{major}	V	3D	J								x	x			
Ax^{3D}_{minor}	V	3D	J					x				x			
C^{3D}_1	V	3D	J					x							
C^{3D}_2	V	3D	J					x							
D^{3D}_{Feret}	V	3D	J									x			
D^{3D}_{YZ}	V	3D	J					x				x			
D^{3D}_{XZ}	V	3D	J			x						x	x		
D^{3D}_{XY}	V	3D	J		x							x	x		
DS^{3D}	V	3D	J				x	x							
S^{3D}	V	3D	J					x							
V^{3D}	V	3D	J		x							x			
Cor^{GLCM}_{2D}	V	2D	T					x							
$LRHGE^{GLRLM}_{2D}$	V	2D	T						x						
GLV^{GLSZM}_{2D}	V	2D	T						x						
H^{2D}_h	V	2D	T						x						
Bus^{NGTDM}_{3D}	V	3D	T						x						
ZCR_t	A		C	x			x		x						x
$MFCC_0$	A		F						x						
$MFCC_1$	A		F						x						
$MFCC_2$	A		F	x			x								
$MFFC_3$	A		F												x
$MFFC_4$	A		F												x
$MFCC_5$	A		F	x											x
$MFCC_8$	A		F	x			x								
$MFCC_{10}$	A		F	x			x								
$MFCC_{11}$	A		F	x			x								
$Sfla_f$	A		F				x								
$Skurt_f$	A		F	x			x		x						
P_f			F									x			

Kontynuacja tabeli na następnej stronie

Cecha	Dane	2D/3D	Rodzaj	/ʂ/			/ʐ/		/tʂ/			/dʐ/			
				#3	#5	#8	#3	#5	#3	#5	#7	#1	#3	#4	#5
NFF_2	A		N	x			x		x						
$NFFD_2$	A		N						x						
$NFFD_{12}$	A		N	x			x								
$NFFD_{23}$	A		N	x			x								
$NFFL_1$	A		N	x					x						
$NFFL_2$	A		N												x
$NFFL_3$	A		N	x					x						
$NFFL_4$	A		N	x											
$NFFR_{12}$	A		N	x			x		x						
$NFFR_{23}$	A		N	x			x		x				x		
$NFFR_{24}$	A		N	x			x								
$NFFRL_{13}$	A		N				x		x						
$NFFRL_{14}$	A		N	x			x		x			x			
$NFFRL_{23}$	A		N	x			x								
$NFFRL_{24}$	A		N				x		x						
$NFFRL_{34}$	A		N									x			
NE_0	A		N	x					x						
NE_1	A		N	x			x		x						
NE_2	A		N	x			x		x						x
NE_3	A		N	x											x
NE_4	A		N												x
NE_5	A		N	x											x
NE_6	A		N	x			x		x						x
NE_7	A		N				x								x
NE_8	A		N												x
NPF	A		N	x			x		x						

Powtarzalności cech można doszukiwać się w przypadku głosek /ʂ/ i /ʐ/ pod kątem postdentalności. Wyróżnić w tym przypadku można aż 15 wspólnych parametrów. Z kolei dentalność u głoski /dʐ/ charakteryzuje się większym udziałem cech wizualnych, głównie trójwymiarowych języka. Te same cechy (dwu- oraz trójwymiarowe) dominują również w przypadku opisu stopnia skrócenia wędzidełka językowego (#5). Zależność nie jest prawdziwa jedynie w przypadku ostatniej z głosek szeregu szumiącego, u której znajduje się różnice w rozkładach median cech akustycznych. Najliczniej występujące cechy dla wszystkich możliwości (pojawiające się czterokrotnie) to liczba przejść przez zero (ZCR_t),

stosunek częstotliwości formantów 2 i 3 ($NFFR_{23}$) i stosunek amplitud formantów 1 i 4 ($NFFRL_{14}$).

Przeważająca część analizy w szeregu szumiącym dotyczyła cech wielogrupowych. Po teście Kruskala-Wallisa, dla cech istotnych statystycznie, wykonano porównanie post hoc. Postdentalność dotyczyła testów we wszystkich głoskach. W przypadku głoski /ʂ/ różnice w rozkładach obserwowano pomiędzy wszystkimi grupami, jednak dominowały pomiędzy *normą* a *zębowością* oraz pomiędzy *zębowością* i *zadziąsłowością*. Warto nadmienić, że występowały tu tylko parametry akustyczne, z przewagą szumowych (o wysokiej wielkości efektu w ośmiu przypadkach). U głoski /ʐ/, z kolei, różnice zaobserwowano tylko pomiędzy wymową: normatywną (*norma*) a zębową (*zębowość*) oraz pomiędzy wymową zadziąsłową (*zadziąsłowość*) i zębową (*zębowość*). Tu również większość cech istotnych statystycznie (o co najmniej średniej wielkości efektu) jest akustyczna, przede wszystkim bazująca na szumie. W przypadku postdentalności u głosek /tʂ/ oraz /dʐ/ pojawiają się cechy obrazowe — dla pierwszej z nich głównie dotyczące tekstury (choć, mimo wszystko, przeważają cechy akustyczne), dla drugiej opisujące kształt języka. Różnice pojawiają się między wszystkimi parami, choć dominują *norma-zębowość* oraz *zadziąsłowość-zębowość*. Kolejną cechą artykulacyjną obejmującą analizę w całym szeregu szumiącym jest stopień skrócenia wędzidełka językowego. Oprócz fonemu /dʐ/, przeważają cechy związane z geometrią języka (głównie 3D). W przypadku odbiegającej głoski, dominują cechy akustyczne (tu również szumowe). Różnice obserwuje się pomiędzy każdą z analizowanych par, choć dla głosek /ʂ/ i /ʐ/ głównie w zestawieniach: *norma-średnio skrócone* oraz *nieznacznie skrócone-średnio-skrócone*, podczas gdy dla głoski /dʐ/ przede wszystkim: *norma-nieznacznie skrócone*.

Szereg ciszący

Cechy obrazowe stanowią liczniejszą grupę niż akustyczne (20 vs. 4) w szeregu ciszącym (tab. 7.8). Trójwymiarowe cechy związane z kształtem języka stanowią dominującą podkategorię parametrów wizualnych. Można zauważyć, że są głównymi elementami analizy w przypadku głoski /ɕ/ — przede wszystkim oceniając różnice rozkładów median biorąc pod uwagę: dentalność, stopień skrócenia wędzidełka językowego, medialność żuchwy. Medialność żuchwy w przypadku głoski /ɕ/ również wykazuje cechy o co najmniej średniej wielkości efektu spośród zbioru trójwymiarowych cech opisujących kształt języka. Dla fonemów /tɕ/ i /dʑ/ dominują raczej cechy dwuwymiarowe tego samego artykulatora.

Tab. 7.8: Zestawienie występowania cech akustyczno-wizualnych o wartości p poniżej 0,05 i co najmniej średniej wielkości efektu w przypadku głosek z szeregu ciszącego. Cechy akustyczne i obrazowe zaznaczono symbolami: *A* i *V*. Kolorami rozróżniono rodzaje cech obrazowych i akustycznych. Szczegółowe wyniki porównania rozkładów dla poszczególnych głosek zawierają tab. B.15–B.21 w dodatku B.3.

Kategorie cech:

obrazowe:
- W kształtu warg
- U kształtu ust
- J kształtu języka
- T teksturowe

akustyczne:
- C czasowe
- F częstotliwościowe
- N szumowe

Cechy artykulacyjne:
- #1 dentalizacja
- #2 dentalność
- #3 postdentalność
- #4 apikalność
- #5 skrócenie wędzidełka językowego
- #6 medialność wypływu powietrza
- #7 medialność języka
- #8 medialność żuchwy

Cecha	Dane	2D/3D	Rodzaj	/ɕ/				/ʑ/			/tɕ/			/dʑ/
				#1	#3	#5	#8	#5	#7	#8	#1	#5	#8	#1
E^{2D}	V	2D	J	x				x	x			x		
P^{2D}	V	2D	J		x						x			
Ax^{2D}_{minor}	V	2D	J					x					x	
A^{2D}	V	2D	J								x			
DS^{2D}	V	2D	J								x			x
D^{2D}_{Feret}	V	2D	J								x			
D^{3D}_{XZ}	V	3D	J		x					x				
V^{3D}	V	3D	J		x	x								
D^{3D}_{XY}	V	3D	J		x					x				
Ax^{3D}_{major}	V	3D	J		x					x				
C^{3D}_{1}	V	3D	J		x									
D^{3D}_{Feret}	V	3D	J		x	x				x				
E^{3D}	V	3D	J			x								
Ax^{3D}_{least}	V	3D	J			x								
Ax^{3D}_{minor}	V	3D	J			x								
D^{3D}_{YZ}	V	3D	J			x								
A^{3D}	V	3D	W					x						
A^{3D}	V	3D	J		x		x	x						
S^{3D}	V	3D	J		x									
C^{3D}_{2}	V	3D	J		x						x			

Kontynuacja tabeli na następnej stronie

Cecha	Dane	2D/3D	Rodzaj	/ɕ/				/ʑ/			/tɕ/			/dʑ/
				#1	#3	#5	#8	#5	#7	#8	#1	#5	#8	#1
ZCR_t	A		C					×						
$MFCC_4$	A		F					×					×	
NCC_6	A		N										×	
NE_9	A		N				×							

Analogicznie do analizy poprzednich szeregów, cechy, które wykazywały istotność statystyczną w teście Kruskala-Wallisa zostały dodatkowo poddane ocenie porównaniem post hoc. W trzech pierwszych głoskach szeregu ciszącego powtarzała się ocena stopnia skrócenia wędzidełka językowego. W przypadku głoski /ɕ/ dominują różnice pomiędzy parą *norma-nieznaczne skrócenie*, a dla głosek /ʑ/ i /tɕ/ jest to zestawienie grup mówców charakteryzujących się prawidłowym wędzidełkiem (*norma*) i jego *średnim skróceniem*. W każdym z wymienionych przypadków przeważają cechy dotyczące kształtu języka (głównie 3D). Zbadano również ewentualne różnice w rozkładach median cech biorąc pod uwagę medialność żuchwy (głoski /ʑ/ i /tɕ/). Najczęściej pojawiały się w tej analizie cechy obrazowe związane z kształtem języka, wyłącznie 3D, choć w przypadku fonemu /tɕ/ pojawił się również jeden ze współczynników melcepstralnych.

Podsumowanie zależności wizualno-akustyczno-artykulacyjnych

Przyglądając się opisanym do tej pory rezultatom analizy statystycznej, można wnioskować, że na podstawie części zaproponowanych parametrów wizualnych i akustycznych istnieje możliwość rozróżniania grup mówców charakteryzujących się wybranymi cechami opisującymi sposób artykulacji głosek dentalizowanych.

Każdy z przeprowadzonych testów statystycznych zwracał istotne statystycznie cechy wizualno-akustyczne, które wskazują na widoczność różnic w rozkładach pomiędzy danymi zmiennymi (lub parami grup) w danej cesze artykulacyjnej (tab. 7.9). W niektórych przypadkach znaleziono cechy o $p \leq 0{,}05$ charakteryzujące się jedynie niską wartością wskaźnika rozmiaru efektu. Minimalna liczba znalezionych cech to pojedynczy parametr (głoska /s/, cechy artykulacyjne: #5 i #7), maksymalna liczba to 127 (w tym 23 parametry o średniej WE i 4 o wysokiej; głoska /tʂ/, cecha artykulacyjna: postdentalność). Najwięcej cech charakteryzujących się dużą skalą uzyskanego efektu zaobserwowano dla

głoski /ʂ/ i postdentalności (#3; 8 wystąpień). Warto zauważyć, że w szeregu szumiącym dla każdej głoski w analizie postdentalności znaleziono minimum trzy cechy o wysokiej WE. Najczęściej duże wielkości efektu pojawiały się natomiast dla analizy głoski /dz/ (w czterech cechach artykulacyjnych: #2, #6-8). W przypadku 5 głosek spośród wszystkich możliwych nie znaleziono ani jednego parametru o dużym efekcie. Co więcej, w przypadku głoski /dʑ/ znaleziono też tylko jeden parametr o średnim rozmiarze efektu w przypadku dentalizacji (#1).

Tab. 7.9: Podsumowanie liczby cech o istotności statystycznej poniżej 0,05 z trzystopniowym podziałem wielkości efektu (WE). Szczegółowe rozkłady w podziale na grupy cech akustycznych i wizualnych zawierają tab. B.23–B.30 w dodatku B.4.

			#1 dentalizacja #2 dentalność #3 postdentalność #4 apikalność			#5 skrócenie wędzidełka językowego #6 medialność wypływu powietrza #7 medialność języka #8 medialność żuchwy				
			#1	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**
Szereg syczący										
/s/	**p<0,05**		44	65		68	20	10	3	1
	WE	**mała**	36	49		68	20	9	3	1
		średnia	8	16				1		
		wysoka								
/z/	**p<0,05**		46	86		31	33	8	2	3
	WE	**mała**	45	82		31	30	8		3
		średnia	1	4			3		2	
		wysoka								
/ts/	**p<0,05**		34	62		94	24	18	13	3
	WE	**mała**	33	48		85	19	18	7	3
		średnia	1	14		9	4		6	
		wysoka					1			
/dz/	**p<0,05**		22	39		60	22	5	10	4
	WE	**mała**	15	15		56	22	3	7	1
		średnia	7	20		4			2	2
		wysoka		4				2	1	1
Szereg szumiący										
/ʂ/	**p<0,05**		8		48	49	12		16	5
	WE	**mała**	8		22	49	5		16	4
		średnia			18		7			1
		wysoka			8					

Kontynuacja tabeli na następnej stronie

			#1	#2	#3	#4	#5	#6	#7	#8
/ʐ/	p<0,05		15		60	52	23			2
	WE	mała	15		37	52	13			2
		średnia			20		10			
		wysoka			3					
/tʂ/	p<0,05		9		127	36	2		2	4
	WE	mała	9		100	36	1		1	4
		średnia			23		1		1	
		wysoka			4					
/dʐ/	p<0,05		11		48	18	27			15
	WE	mała	3		44	17	15			15
		średnia	8		1	1	12			
		wysoka			3					
Szereg ciszący										
/ɕ/	p<0,05		4		43	5	10	12	38	13
	WE	mała	3		30	5	3	12	38	11
		średnia	1		13		7			2
		wysoka								
/ʑ/	p<0,05		6			17	26	42	59	14
	WE	mała	6			17	20	42	58	10
		średnia					4		1	4
		wysoka					2			
/tɕ/	p<0,05		23			9	4	21	34	20
	WE	mała	18			9	3	21	34	18
		średnia	5							1
		wysoka					1			1
/dʑ/	p<0,05		23			13	5	19	63	16
	WE	mała	22			13	5	19	63	16
		średnia	1							
		wysoka								

W dodatku B.4 umieszczono zestawienia dla wszystkich możliwych rodzajów parametrów obrazowych i akustycznych. Ze względu na fakt, że cechy czasowe i częstotliwościowe występowały stosunkowo rzadko, parametry sygnału dźwiękowego zebrano łącznie w jedną tabelę. Na podstawie tab. B.22–B.30 można zauważyć, że cechy akustyczne dominowały w przypadku postdentalności (#3) w szeregu szumiącym. Wśród często pojawiających się parametrów obrazowych wyróżnić można cechy związane z opisem kształtu języka (zarówno 2D, jak i 3D), najrzadziej parametry odnoszące się do geometrii ust. Wysokim rozmiarem efektu charakteryzowało się 16 cech akustycznych w szeregu

szumiącym (wszystkie w przypadku postdentalności), 13 cech 2D związanych z kształtem języka (8 w szeregu syczącym przy dentalności, stopniu skrócenia wędzidełka językowego, medialności wypływu powietrza i medialności żuchwy; 1 w szeregu szumiącym przy medialności żuchwy; 4 w szeregu ciszącym przy stopniu skrócenia wędzidełka językowego oraz medialności żuchwy) oraz 4 cechy 3D geometrii języka (1 w szeregu syczącym przy medialności języka oraz 3 w szeregu szumiącym dla postdentalności). Dla trzech głosek — /t̪s̪/, /c̪/, /d̪z̪/, dwuwymiarowych parametrów opisujących kształt ust oraz warg nie zaobserwowano ani razu. Otrzymane zestawienia zaprezentowane w tabelach sugerują więc, że największe znaczenie dla poszukiwania różnic w rozkładach w omawianym zagadnieniu mają cechy akustyczne, szczególnie szumowe, oraz parametry związane z kształtem języka.

8. Dyskusja

Dotychczasowe rozwiązania z zakresu analizy artykulacji dzieci w wieku przedszkolnym — budowane na podstawie danych zarejestrowanych w sposób niewymagający ingerencji w jamie ustnej — bazowały przede wszystkim na ocenie sygnałów akustycznych. W poprzednich rozdziałach opisano metodykę przetwarzania nagrań wideo i danych dźwiękowych do komputerowej analizy wybranych cech związanych z artykulacją sybilantów. Proponowane metody oraz eksperymenty stanowią wstępne badania dotyczące hybrydyzacji parametrów wymienionych modalności w kontekście wad wymowy i wymagają komentarza.

8.1 Detekcja i segmentacja artykulatorów w obrazach

Dwuetapowa metoda segmentacji artykulatorów, której opis znajduje się w rozdziale 4, wykazywała zadowalającą skuteczność w przypadku wszystkich rozróżnianych narządów. Maski, otrzymywane jako wynik działania algorytmu, mogły stanowić podstawę ekstrakcji cech. Założenia opracowanej metody obejmowały próbę ograniczenia ręcznego przygotowywania obrysów eksperckich, które w przypadku licznych zbiorów danych wymagają dużego nakładu czasu i pracy. Częściowo nadzorowana metoda łącząca uczenie maszynowe oraz idee technik inteligencji obliczeniowej przebiegała zgodnie z zasadami opisywanymi w literaturze, choć została wzbogacona o własne rozwiązania. Oryginalność opisywanej koncepcji dotyczy: (1) równoczesnej segmentacji wielu artykulatorów dla celów komputerowego wsparcia diagnozy logopedycznej, szczególnie w przypadku zębów i języka; (2) integracji metod z zakresu głębokiego uczenia oraz obliczeń rozmytych i propozycji podejścia częściowo nadzorowanego, który został zbudowany w oparciu o obszerny i zróżnicowany zbiór danych. Odnosząc się do pierwszego z wymienionych punktów należy dodać, że — zgodnie ze stanem wiedzy autorki — język jest kluczowy dla poprawnej realizacji mowy i nie był wcześniej wykorzystywany w aspekcie komputerowej analizy jego udziału w procesie realizacji głosek. Zęby również stanowią ważny element wymowy, jednak po konsultacji ze specjalistami logopedii nie były wykorzystywane do analizy artykulacyjnej w niniejszej pracy. Wiarygodność algorytmu potwierdziły wyniki eksperymentów zaprezentowanych w poprzednim rozdziale.

Skuteczność sieci YOLOv6 w detekcji obiektów była na poziomie wystarczająco wysokim, aby realizować dalsze kroki. Umiarkowaną trudnością w odnajdywaniu na obrazach charakteryzowały się zęby oraz język. Problematyczność detekcji języka wynika z jego kolorystyki (często zbliżonej do warg), raczej niewielkiej powierzchni widocznej na nagraniach wideo w trakcie mowy swobodnej, niejednoznacznych krawędzi, niewielkiej ilości światła, która dostaje się do jamy ustnej oraz stosunkowo rzadkiego występowania na nagraniach. Zęby, z kolei, chociaż przeważnie wyróżniają się kolorystycznie spośród reszty narządów, to ich braki w kompletności, różnice w rozmieszczeniu, wielkości i kształcie czy zasłonięcie przez wargi i język powodują, że ich detekcja nie jest prostym zadaniem. Niemniej skuteczność detekcji wszystkich obiektów była na tyle wysoka, że umożliwiała przygotowanie na ich podstawie niedokładnych etykiet eksperckich do wstępnego uczenia sieci segmentacyjnej. Istotne jest również to, że ostateczny model segmentacyjny nie bazował bezpośrednio na wykrywaniu zębów lub języka. To detekcja warg stanowiła kluczowy element zarówno w praktycznym działaniu metody, jak i procesie uczenia sieci. Wyniki zaprezentowane w poprzednim rozdziale (tab. 7.3) wskazują na detekcje bliskie idealnym (wartości AP oraz współczynnika $F1$ wyniosły 0,999) i o wysokim stopniu wiarygodności (mediana IoU równa 0,862). Na tej podstawie można wnioskować, że wszystkie założenia dotyczące poszukiwania obszarów artykulatorów na obrazach zostały zrealizowane. Dodatkowo w ramach eksperymentów porównano wybrane wersje architektury sieci YOLO w celu znalezienia wariantu o najwyższej jakości działania. Największą wydajność w ramach testów wykazała sieć YOLOv6, która na wejście przyjmowała dane w formie czteroelementowych mozaik.

Słuszność kolejnego kroku proponowanego podejścia, czyli dwuetapowego uczenia modelu sieci DeepLabv3+, uzasadniają wyniki przedstawione na rys. 7.5. Wzrost efektywności sieci dla wszystkich artykulatorów w kolejnych krokach był zauważalny i konsekwentny. Właściwy punkt wyjścia stanowiło wykorzystanie wstępnej (zgrubnej) segmentacji, która bazowała na algorytmie DRLSE: wartość współczynnika DSC wyniosła dla tej metody od 0,711 w przypadku zębów do 0,912 dla ust. Wyniki znacząco poprawił drugi etap, który obejmował słabo nadzorowane uczenie sieci DeepLabv3+ z transferem wiedzy. Do treningu wykorzystano ponad 16 tysięcy niedokładnych etykiet przygotowanych w poprzednim kroku. Niewielki spadek mediany DSC zaobserwowano po tym etapie jedynie dla języka. Końcowe dostrajanie sieci, z użyciem ok. 6% wszystkich klatek przeznaczonych do treningu, skutkowało kolejną znaczną poprawą jakości działania.

Decyzja o wyborze szkieletu (rdzenia) sieci, który był wykorzystywany do finalnej, właściwej segmentacji, nie była oczywista. Rys. 7.7 ilustruje kompromis widoczny w jakości segmentacji pomiędzy mniejszymi i bardziej nieregularnymi obiektami (zęby i język) a większymi (usta i wargi). Te ostatnie były skutecz-

niej wyodrębniane przez sieci ResNet i MobileNet v2 (górna granica *DSC* na poziomie 0,953), ale ich przewaga w różnicach bezwzględnych była nieznaczna. Z drugiej strony, wyniki wskazują, że język stanowił najtrudniejszy artykulator do wyodrębnienia, a różnice między modelami były pod tym kątem najbardziej znaczące. Xception charakteryzował się najwyższymi i najmniej rozrzuconymi wynikami, dlatego to on został wybrany jako rdzeń ostatecznej metody. Co więcej, spośród dwóch modeli wykazujących najwyższą skuteczność w segmentacji zębów i języka, Xception jest siecią mniejszą i szybszą w porównaniu z ResNet-152.

Zęby i język stanowią obiekty trudne w detekcji i segmentacji z kilku wymienionych wcześniej powodów (zęby: wiele małych, szarych lub białych obiektów; język: zmienny, rzadko powtarzalny kształt, nierzadko przysłaniany przez zęby oraz wargi; oba artykulatory: problemy z oświetleniem). Z tego względu średni współczynnik DSC przekraczający w obu przypadkach 0,83 uzyskany za pomocą metody w pełni zautomatyzowanej jest satysfakcjonujący. Jest to podobny poziom dokładności do jakości uzyskanej w badaniach rozpoczynających prace w kierunku przygotowania niniejszej metody. Eksperymenty bazowały na lepiej oświetlonych obrazach przedstawiających statyczne języki o większej powierzchni [126]. Niemniej jednak, w niniejszym rozwiązaniu, język nie zawsze był wykrywany prawidłowo, szczególnie w przypadku małych, zakrytych lub słabo oświetlonych obszarów. Z kolei wartości DSC powyżej 0,88 w przypadku warg i powyżej 0,94 dla ust są zadowalające i wskazują na jakość działania, która jest wystarczająca do przeprowadzenia dalszych etapów. Wyniki są porównywalne lub lepsze od dostępnych w literaturze rozwiązań opartych na sieciach uczonych pod pełnym nadzorem i kompletnymi bazami eksperckimi [12, 23, 97, 126].

Opracowana metoda miała też pewne ograniczenia. Po pierwsze, przetwarzanie było wrażliwe na jakość wykrywania ust. W przypadku, kiedy ROI było niedokładne, zwłaszcza niedoszacowane, segmentacji nie przeprowadzano na pełnym obszarze. Ponieważ fałszywie negatywne rezultaty są niepożądane, wyniki detekcji poszerzano o dodatkowy margines, który miał zapewnić właściwą segmentację ust nawet pomimo niedostatecznie dużego ROI. Po drugie, zauważono wrażliwość metody na oświetlenie otoczenia. Do bazy danych zaliczały się ramki o różnej jasności i kontraście, m.in. ze względu na oświetlenie zewnętrzne czy odmienny odcień skóry wśród dzieci. Rozwiązaniem tego problemu w przyszłości może być użycie większego zbioru treningowego, który obejmie więcej różnorodnych przypadków. W ten sposób model przypuszczalnie nabierze większej odporności na czynniki zewnętrzne związane z oświetleniem i zwiększy uniwersalność działania. Mimo wszystko, liczba mówców w bazie danych stanowiła pokaźną wartość względem rozwiązań dostępnych w literaturze, w których metody opracowuje się przeważnie na zaledwie kilkunastu przypadkach. Język stanowi jednak największe wyzwanie spośród narządów segmentowanych w pracy

i podejście do jego wyodrębniania wymaga dalszego ulepszania (np. poprzez bardziej rozbudowane przetwarzanie wstępne).

Przy okazji omawiania wyników detekcji oraz segmentacji obrazów warto wspomnieć o samej rejestracji sygnałów oraz skonstruowanej bazie danych. Wykorzystano narzędzie pomiarowe, które nie ingerowało w proces artykulacji i zapewniło stabilną, niezakłóconą i nieruchomą scenę. Tak skonstruowany zbiór danych stanowił odpowiednią podstawę do przeprowadzenia wstępnych badań dotyczących wykorzystania nagrań wideo do poszukiwania różnic w normatywnych i patologicznych realizacjach głosek. Cała zarejestrowana w ramach projektu baza mówców stanowi jedną z niewielu tak dużych zbiorów danych dla celów rozwoju komputerowego wsparcia logopedii. Wśród innych baz znaleziono jedynie dotyczącą wykorzystania artykulografii elektromagnetycznej, wielokanałowego sygnału akustycznego i stereoskopowych danych obrazowych [163]. Niemniej jednak obejmowała ona osoby dorosłe i wykorzystana konfiguracja raczej nie pozwalała na rejestrację danych w przedszkolu. Dane wykorzystane w niniejszej pracy są różnorodne, towarzyszą im opisy logopedyczne o szerokim zakresie (zarówno dotyczące sprawności i budowy narządów mowy, jak i oceny realizacji głosek) i mogą stanowić punkt wyjścia do dalszego rozwoju metod komputerowych. Należy dodać, że w momencie składania niniejszej rozprawy baza danych jest w trakcie opracowywania i opisywania w celu upublicznienia i umożliwienia prowadzenia badań w rozpatrywanej dziedzinie.

Segmentacja artykulatorów stanowiła podstawę do ekstrakcji cech obrazowych. Wiarygodność analizy statystycznej w dużej mierze była zależna od jakości wyodrębniania obiektów (tj. zafałszowane segmentacje nie oddają wiernie rzeczywistości). Dlatego oprócz testowania dokładności działania wykorzystywanych modeli, zadbano również o weryfikację poprawności segmentacji ramek obejmujących każdą z głosek. Mimo odrzucenia przypadków o niewystarczającej jakości, do analizy statystycznej wykorzystano 195 mówców, od których pochodziły łącznie 10 202 fonemy. Stanowi to dużą, różnorodną liczbę obserwacji.

8.2 Analiza wizualno-akustyczno-artykulacyjna

Na wstępie warto zaznaczyć, że w pracy przyjęto pewne założenia dotyczące opisu logopedycznego. Do badań kwalifikowano dzieci, u których analizowane realizacje wybranych fonemów są na etapie doskonalenia i odróżnienie normy (w tym rozwojowej) od zaburzeń nie jest zadaniem łatwym. Co więcej, zdarza się, że pewne cechy głoski mogą być realizowane w kilku poprawnych wariantach — np. głoska /s/ może być artykułowana apikalnie lub dorsalnie i obie opcje są akceptowalne. W pracy wykorzystano opis logopedyczny, który wprowadzał wyraźny podział na cechy normatywne i nienormatywne. Należy jednak mieć na

uwadze, że określanie normy zależy od wielu czynników (w tym biologicznych, funkcjonalnych czy językoznawczych [44, 84]) i w przypadku wykorzystanej bazy danych podział na różne warianty realizacji fonemów (zamiast na cechy pożądane i niepożądane) mógłby okazać się bardziej wiarygodny.

W wyniku ekstrakcji cech uzyskano wektory 472 cech: 48 cech kształtu 2D, 96 cech kształtu 3D, po 126 cech teksturowych 2D oraz 3D i 76 parametrów akustycznych. Suma cech obrazowych (396) w równej części pochodziła z dwóch wykorzystywanych kamer. Wysokie współczynniki korelacji pomiędzy parami tych samych parametrów, ale uzyskanych z kamery prawej i lewej, dawały możliwość ograniczenia wektora do cech obrazowych jedynie z jednego źródła. Nie zdecydowano się jednak na taką redukcję zakładając, że informacje uzyskane z dwóch widoków mogą się wzajemnie uzupełniać. Wykorzystanie opisywanej konfiguracji sprzętowej w zamierzeniu miało pozwolić na przygotowanie danych stereowizyjnych [67]. W ramach badań (które stanowią wstępne eksperymenty do poszukiwania zależności wizualno-artykulacyjnych) zdecydowano się pozostać przy konwencjonalnych danych obrazowych. Zaproponowano jednak wykorzystanie trójwymiarowych wolumenów, w których trzeci wymiar stanowił czas. Model był złożeniem serii następujących po sobie wysegmentowanych ramek sygnału (rys. 4.16). Powinien zatem odzwierciedlać zmiany w ułożeniu narządów w trakcie realizacji głosek. Zgodnie z przeglądem aktualnej literatury, nie znaleziono doniesień o wykorzystaniu takiego podejścia dla celów komputerowego wsparcia logopedii. Autorka natomiast założyła, że obserwacja ruchu poszczególnych artykulatorów, zwłaszcza warg oraz języka, może odzwierciedlać niepoprawne wzorce motoryczne lub związane z nieprawidłowym położeniem i tym samym rozróżniać poprawną i błędną wymowę. Potencjalnie, wykorzystanie cech bazujących na trójwymiarowych modelach może nieść większe korzyści dla opracowywanego tematu i zwracać istotniejszą informację (szczególnie w przypadku parametrów geometrycznych) niż uzyskiwanie macierzy cech dwuwymiarowych (złożenie wektorów dla każdej ramki obejmującej fonem) i ich agregacja w ramach danych jednego mówcy i głoski przed analizą statystyczną.

Analiza statystyczna wykazała, że w każdym z szeregów głosek wśród cech obrazowych przeważają parametry geometryczne-trójwymiarowe. W szeregu syczącym i ciszącym dominują też wśród wszystkich istotnych statystycznie zmiennych akustycznych i wizualnych. Najczęściej występującym artykulatorem, na podstawie którego ekstrahowano istotne statystycznie cechy obrazowe, był język. Można było oczekiwać, że wargi i usta są obiektami, w ułożeniu których najłatwiej dostrzegalne będą błędne wzorce ruchowe (nawet nie będące skutkiem niewłaściwej motoryki tych narządów, a np. języka czy żuchwy). Ich dużą zaletą była też stała widoczność. Z kolei język ze względu na rzadkość pojawiania się i trudność w segmentacji, o której pisano wcześniej, nie wydawał się być najskuteczniejszym obiektem, na podstawie którego możliwe byłoby

różnicowanie sposobów realizacji fonemów. Być może jednak, błędna wymowa wiązała się z większą widocznością języka na obrazach lub pozycjonowaniem go w nietypowych miejscach. To założenie wydaje się mieć odzwierciedlenie w ocenie dentalności — w pracy sprawdzano obecność różnic w rozkładach zmiennych pomiędzy wymową normatywną i międzyzębową. Dla wszystkich głosek szeregu syczącego w analizie dentalności najliczniejszą grupą parametrów istotnych statystycznie są te, które opisują kształt wolumenu języka. Na częstszą obecność i większą powierzchnię narządu w porównaniu z realizacjami normatywnymi mogą również wskazywać najwyższe wielkości efektu dla parametrów opisujących objętość modelu, jego pole powierzchni czy średnice. Z drugiej strony, proces oceny postdentalności występującej jako normatywna cecha realizacji głosek szumiących (odpowiada dentalności w szeregu syczącym) sugerował, że najliczniejszą grupą parametrów istotnych statystycznie są cechy akustyczne bazujące na pasmie szumu. Analiza post hoc wykazała najwięcej różnic pomiędzy normą a zębowością i zadziąsłowością a zębowością. Może to wynikać z przesunięcia się pasma szumu w przypadku każdego ze sposobów realizacji oraz mniejszej widoczności języka na obrazach. Sugeruje to również, że hybrydyzacja modalności może nieść korzyści w postaci bogatszej informacji diagnostycznej.

Analiza liczby wszystkich cech, które w testach wykazywały wartość p poniżej 0,05 (bez względu na wielkość efektu), sugeruje większą użyteczność parametrów ekstrahowanych na bazie trójwymiarowych modeli. Mniejszą liczbę parametrów o co najmniej średniej wielkości efektu obserwowano dla cech teksturowych (w porównaniu z liczebnością parametrów geometrycznych). Możliwe, że różnice osobnicze oraz wpływ warunków zewnętrznych (głównie oświetlenia) pomiędzy poszczególnymi mówcami miały wpływ na brak powtarzających się wzorców rozróżniających wybrane cechy mowy normatywnej lub patologicznej. Stało się tak mimo poszukiwania zgrubnych, ogólnych relacji teksturowych (32 poziomy szarości w analizie). Można też przypuszczać, że cechy kształtu związane z wargami powinny wykazywać podobne zależności jak parametry opisujące usta, gdyż obrys ust jest segmentacją warg poszerzoną o zalanie przestrzeni między wargami. Wyniki analizy nie wskazują jednak na taką zależność. Liczba istotnych statystycznie cech opisujących kształt warg (2D i 3D, o dowolnej wielkości efektu) jest większa niż liczba parametrów opisujących usta. Co więcej, niewielka liczba istotnych statystycznie wskaźników geometrycznych tych dwóch artykulatorów w przestrzeni dwuwymiarowej sugeruje ich niewielką użyteczność w zadanym problemie.

Szerokie badania prowadzone w kilku placówkach przedszkolnych pokazały, że nienormatywna realizacja głosek dentalizowanych (seplenienie) jest częsta, nasilona w różnym stopniu i wynikająca często z odmiennych przyczyn. Biorąc pod uwagę skalę problemu i fakt, że dzieci na tym etapie mogą wiele wypracować pod okiem specjalisty, rozwój komputerowych metod wspomagania

diagnostyki logopedycznej (np. zaprezentowany w tej pracy) jest istotny i powinien być kontynuowany. Zaprezentowane w pracy wyniki wskazują na możliwość poszukiwania różnic między różnymi (normatywnymi i patologicznymi) realizacjami głosek w kontekście wybranych cech artykulacyjnych. Wiarygodność testów została uzyskana dzięki bazowaniu na wystarczająco skutecznej metodzie segmentacji artykulatorów. Metryki dotyczące działania opracowanej metody segmentacji narządów oraz selekcja obrysów o odpowiednio wysokiej jakości zwiększały rzetelność analizy statystycznej. Przedstawione rezultaty mogą stanowić punkt wyjścia do prac nad budową systemów eksperckich wspierających diagnostykę logopedyczną sygmatyzmu.

9. Podsumowanie

Badania prowadzone przez specjalistów logopedii sugerują, że wady wymowy — zwłaszcza seplenienie, na którym koncentruje się ta rozprawa — są problemem powszechnym wśród dzieci w wieku przedszkolnym. Ze względu na możliwość skutecznej terapii na tym etapie życia i uwzględniając negatywne skutki zaniedbania diagnozy, cenne jest wdrażanie komputerowych rozwiązań, które potencjalnie poszerzają informację diagnostyczną. W niniejszej pracy opisano wyniki badań dotyczących użyteczności parametrów wizualnych i akustycznych, które bazują na danych reprezentujących artykulację głosek dentalizowanych przez przedszkolaków, do oceny wybranych cech artykulacyjnych.

Zaproponowano wieloetapową strukturę badań, która obejmowała następujące kroki: rejestrację bazy danych w ramach sesji pomiarowych w przedszkolach; opracowanie metodyki przetwarzania danych obrazowych z wykorzystaniem sieci głębokich: YOLO do detekcji artykulatorów oraz modelu DeepLabv3+ do segmentacji; opracowanie ścieżki analizy sygnałów akustycznych; przeprowadzenie testów statystycznych, które sprawdzały relacje między ekstrahowanymi parametrami obrazowymi i akustycznymi a wybranymi cechami artykulacyjnymi głosek dentalizowanych. Pełna ścieżka przetwarzania zaproponowana w poprzednich rozdziałach rozprawy stanowi realizację **celu niniejszej rozprawy**, który obejmował:

> opracowanie metodyki przetwarzania sygnałów akustycznych i danych obrazowych z wykorzystaniem metod sztucznej inteligencji.

Weryfikacja jakości działania metody przetwarzania danych wideo wskazała na wystarczająco wysoką skuteczność. Raportowana jakość umożliwiła wykorzystanie wyników segmentacji do kolejnych kroków metodyki. Stanowiło to **potwierdzenie 1. tezy pomocniczej**, która została sformułowana następująco:

> Możliwa jest wiarygodna segmentacja wybranych artykulatorów w obrazach twarzy z wykorzystaniem metod sztucznej inteligencji.

W przypadku ekstrakcji cech obrazowych, oprócz konwencjonalnego podejścia dwuwymiarowego, zaproponowano wykorzystanie trójwymiarowych wolumenów, w których trzeci wymiar stanowił czas. Otrzymane zestawy parametrów wizualnych (2D i 3D: teksturowe, kształtu warg, ust i języka) i akustycznych (czasowych, częstotliwościowych w pełnym pasmie i częstotliwościowych w pasmie szumu) stanowiły podstawę poszukiwania różnic w ich rozkładach pomiędzy wymową normatywną a zaburzoną. Analiza statystyczna wykazała, że w każdym z szeregów głosek wśród cech obrazowych przeważają parametry geometryczne-trójwymiarowe, zwłaszcza dotyczące kształtu języka. Być może błędna wymowa wiązała się z większą widocznością języka na obrazach lub pozycjonowaniem go w nieoczekiwanych miejscach. W przypadku oceny postdentalności w szeregu szumiącym dosyć liczną grupę istotnych statystycznie parametrów stanowiły cechy szumowe. Zaprezentowane w pracy wyniki wskazują na możliwość poszukiwania różnic między różnymi (normatywnymi i patologicznymi) realizacjami głosek w kontekście wybranych cech artykulacyjnych. Wiarygodność testów została uzyskana dzięki bazowaniu na rzetelnej metodzie segmentacji artykulatorów. Stanowi to **potwierdzenie tezy pomocniczej nr 2**:

> Ekstrakcja i analiza cech obrazowych 2D i 3D oraz parametrów akustycznych pozwala na określenie różnic między grupami w wybranych cechach artykulacyjnych.

Wykazanie przytoczonych tez pomocniczych oraz opracowanie metodyki będącej celem tej rozprawy, pozwoliły na **potwierdzenie tezy głównej**:

> **Istnieją istotne statystycznie różnice w cechach sygnałów akustycznych i danych obrazowych prezentujących mowę dzieci z różnymi (normatywnymi i nienormatywnymi) cechami wymowy.**

Metodyka opisana w pracy oraz zaprezentowane rezultaty mogą stanowić punkt wyjścia do rozwoju systemów eksperckich wspierających diagnostykę logopedyczną sygmatyzmu. Ze względu na złożoność i wieloaspektowość procesu diagnostycznego obecnie nie byłoby możliwe — lub byłoby bardzo trudne —opracowanie automatycznego systemu eksperckiego do kompleksowej oceny wad wymowy (lub nawet jednego jej wariantu). Realizacja mniejszych projektów z zakresu analizy artykulacji stanowi cenny wkład w ogólny rozwój inżynieryjnego wsparcia specjalistów terapii mowy. Opisywane podejście wykazuje jednak różne ograniczenia. Istotne jest poświęcenie uwagi różnicom pomiędzy głoskami frykatywnymi (szczelinowymi) a afrykatami (zwarto-szczelinowymi).

W pracy wszystkie głoski, bez względu na ich sposób artykulacji, przetwarzano w jednakowy sposób — eliminowano początkowe oraz końcowe ramki. Konstrukcja głosek zwarto-szczelinowych (dwustopniowość ich realizacji) może sugerować konieczność opracowania indywidualnej ścieżki przetwarzania, która będzie ukierunkowana założeniami uwzględniającymi specyfikę tej grupy fonemów. Możliwe jest również poszerzenie analiz o dodatkowe cechy artykulacyjne, w tym związane z anatomią i sprawnością narządów mowy. Przypuszczalnie możliwe jest także wykorzystanie algorytmów klasyfikacji danych, które mogą bazować na cechach wizualnych i akustycznych, a których głównym celem będzie wskazywanie odchyleń od normy w mowie. Ponadto zaproponowana metodyka, zwłaszcza propozycja wykorzystania hybrydowego zestawu cech, ma obiecujący potencjał do badania innych zaburzeń z obszaru dyslalii (np. kappacyzm, lambdacyzm czy rotacyzm). Metoda detekcji i segmentacji artykulatorów zaproponowana w niniejszej pracy, po właściwej adaptacji, może również posłużyć w innych dziedzinach medycznych, np. jako wsparcie rehabilitacji po urazach lub interwencjach chirurgicznych w obrębie twarzy. Niezależnie od kierunku rozwoju, pewnym jest, że poszerzanie dostępnym rozwiązań o kolejne badania stanowi wartościowy wkład w usprawnianie diagnostyki i terapii logopedycznej.

Bibliografia

[1] B. Ahmed, P. Monroe, A. Hair, C. T. Tan, R. Gutierrez-Osuna i K. J. Ballard. „Speech-driven mobile games for speech therapy: User experiences and feasibility". W: *International Journal of Speech-Language Pathology* vol. 20. no. 6 (2018), s. 644–658. DOI: 10.1080/17549507.2018.1513562 (cytowane na stronie 9).

[2] F. Alías, J. Socoró i X. Sevillano. „A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds". W: *Applied Sciences* vol. 6. no. 5 (2016). DOI: 10.3390/app6050143 (cytowane na stronie 70).

[3] M. Amadasun i R. King. „Textural features corresponding to textural properties". W: *IEEE Transactions on Systems, Man, and Cybernetics* vol. 19. no. 5 (1989), s. 1264–1274. DOI: 10.1109/21.44046 (cytowane na stronie 63).

[4] Articulate Instruments. *EPG products (EPG palate)*. http://www.articulateinstruments.com/epg-products/. Dostęp: 19.06.2024 (cytowane na stronie 7).

[5] B. Atal. „Automatic Speaker Recognition Based on Pitch Contours". W: *Journal of the Acoustical Society of America* vol. 52 (1969), s. 1687–1697 (cytowane na stronie 70).

[6] V. Badrinarayanan, A. Handa i R. Cipolla. „SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling". W: (2015). arXiv: 1505.07293 [cs.CV] (cytowane na stronie 38).

[7] P. Badura i W. Wieclawek. „Calibrating level set approach by granular computing in computed tomography abdominal organs segmentation". W: *Applied Soft Computing* vol. 49 (2016), s. 887–900. DOI: 10.1016/j.asoc.2016.09.028 (cytowane na stronie 43).

[8] R. D. Beemer, L. Li, A. Leonti, J. Shaw, J. Fonseca, I. Valova, M. Iskander i C. H. Pilskaln. „Comparison of 2D Optical Imaging and 3D Microtomography Shape Measurements of a Coastal Bioclastic Calcare-

ous Sand". W: *Journal of Imaging* vol. 8. no. 3 (2022). DOI: 10.3390/jimaging8030072 (cytowane na stronie 49).

[9] R. Bilibajkić, M. Vojnović i Z. Šarić. „Detection of Lateral Sigmatism using Support Vector Machine". W: *Speech and Language* (2019), s. 322–328 (cytowane na stronach 7 i 9).

[10] Z. Bilkova, M. Bartos, A. Dominec, S. Gresko, A. Novozamsky, B. Zitova i M. Paroubkova. „ASSISLT: Computer-aided speech therapy tool". W: *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, s. 598–602. DOI: 10.23919/eusipco55093.2022.9909627 (cytowane na stronie 10).

[11] Z. Bílková, A. Novozámský, A. Domínec, Š. Greško, B. Zitová i M. Paroubková. „Automatic Evaluation of Speech Therapy Exercises Based on Image Data". W: *Lecture Notes in Computer Science*. Springer International Publishing, 2019, s. 397–404. DOI: 10.1007/978-3-030-27202-9_36 (cytowane na stronie 10).

[12] M. Birara i G. B. Gebremeskel. „Augmenting machine learning for Amharic speech recognition: a paradigm of patient's lips motion detection". W: *Multimedia Tools and Applications* vol. 81. no. 17 (2022), s. 24377–24397. DOI: 10.1007/s11042-022-12399-w (cytowane na stronach 9 i 113).

[13] E. Bribiesca. „A measure of compactness for 3D shapes". W: *Computers & Mathematics with Applications* vol. 40. no. 10 (2000), s. 1275–1284. DOI: 10.1016/S0898-1221(00)00238-8 (cytowane na stronie 50).

[14] M. B. Brown i A. B. Forsythe. „Robust Tests for the Equality of Variances". W: *Journal of the American Statistical Association* vol. 69. no. 346 (1974), s. 364–367. DOI: 10.1080/01621459.1974.10482955 (cytowane na stronie 83).

[15] V. Bukmaier i J. Harrington. „The articulatory and acoustic characteristics of Polish sibilants and their consequences for diachronic change". W: *Journal of the International Phonetic Association* vol. 46. no. 3 (2016), s. 311–329. DOI: 10.1017/S0025100316000062 (cytowane na stronie 8).

[16] J. Chaki i N. Dey. „Statistical Texture Features". W: *Texture Feature Extraction Techniques for Image Recognition*. Singapore: Springer Singapore, 2020, s. 7–23. DOI: 10.1007/978-981-15-0853-0_2 (cytowane na stronach 52, 55 i 56).

[17] L.-C. Chen, G. Papandreou, I. Kokkinos, K. P. Murphy i A. L. Yuille. „DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs". W: *IEEE Transactions*

on Pattern Analysis and Machine Intelligence vol. 40 (2016), s. 834–848 (cytowane na stronach 38, 39, 40 i 41).

[18] L.-C. Chen, G. Papandreou, I. Kokkinos, K. P. Murphy i A. L. Yuille. „Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs". W: *CoRR* vol. abs/1412.7062 (2014) (cytowane na stronach 38, 39 i 40).

[19] L. Chen, G. Papandreou, F. Schroff i H. Adam. „Rethinking Atrous Convolution for Semantic Image Segmentation". W: *CoRR* vol. abs/1706.05587 (2017). arXiv: 1706.05587 (cytowane na stronach 38, 39 i 40).

[20] L. Chen, Y. Zhu, G. Papandreou, F. Schroff i H. Adam. „Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation". W: *CoRR* vol. abs/1802.02611 (2018). arXiv: 1802.02611 (cytowane na stronach 38, 39 i 40).

[21] J. Cheng, H. Li, D. Li, S. Hua i V. S. Sheng. „A Survey on Image Semantic Segmentation Using Deep Learning Techniques". W: *Computers, Materials & Continua* vol. 74. no. 1 (2023), s. 1941–1957. DOI: 10.32604/cmc.2023.032757 (cytowane na stronach 38 i 39).

[22] F. Chollet. „Xception: Deep learning with depthwise separable convolutions". W: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, s. 1251–1258. DOI: 10.1109/CVPR.2017.195 (cytowane na stronach 46 i 92).

[23] K. Chotikkakamthorn, P. Ritthipravat, W. Kusakunniran, P. Tuakta i P. Benjapornlert. „A lightweight deep learning approach to mouth segmentation in color images". W: *Applied Computing and Informatics* (2022). DOI: 10.1108/aci-08-2022-0225 (cytowane na stronach 10 i 113).

[24] C. L. Chowdhary i D. Acharjya. „Segmentation and Feature Extraction in Medical Imaging: A Systematic Review". W: *Procedia Computer Science* vol. 167 (2020). International Conference on Computational Intelligence and Data Science, s. 26–36. DOI: 10.1016/j.procs.2020.03.179 (cytowane na stronie 48).

[25] Y. Chuang, S. Zhang i X. Zhao. „Deep learning-based panoptic segmentation: Recent advances and perspectives". W: *IET Image Processing* vol. 17. no. 10 (2023), s. 2807–2828. DOI: 10.1049/ipr2.12853 (cytowane na stronie 37).

[26] J. Cohen. „A power primer". W: *Psychological Bulletin* vol. 112. no. 1 (1992), s. 155–159. DOI: 10.1037//0033-2909.112.1.155 (cytowane na stronie 77).

[27] M. Cote, A. Dash i A. B. Albu. „Semantic segmentation of textured mosaics". W: *EURASIP Journal on Image and Video Processing* vol. 2023 (2023), s. 1–26. DOI: 10.1186/s13640-023-00613-0 (cytowane na stronie 51).

[28] K. Coufal, D. Parham, M. Jakubowitz, C. Howell i J. Reyes. „Comparing Traditional Service Delivery and Telepractice for Speech Sound Production Using a Functional Outcome Measure". W: *American Journal of Speech-Language Pathology* vol. 27. no. 1 (2018), s. 82–90. DOI: 10.1044/2017_AJSLP-16-0070 (cytowane na stronie 8).

[29] J. Dai, Y. Li, K. He i J. Sun. „R-FCN: Object Detection via Region-based Fully Convolutional Networks". W: *CoRR* vol. abs/1605.06409 (2016). arXiv: 1605.06409 (cytowane na stronie 33).

[30] N. Dalal i B. Triggs. „Histograms of oriented gradients for human detection". W: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. T. 1. 2005, 886–893 vol. 1. DOI: 10.1109/CVPR.2005.177 (cytowane na stronie 33).

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li i L. Fei-Fei. „Imagenet: A large-scale hierarchical image database". W: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2009, s. 248–255. DOI: 10.1109/CVPR.2009.5206848 (cytowane na stronie 46).

[32] A. Dinno. „Nonparametric Pairwise Multiple Comparisons in Independent Groups using Dunn's Test". W: *Stata Journal* vol. 15 (2015), s. 292–300. DOI: 10.1177/1536867X1501500117 (cytowane na stronie 83).

[33] Y. Dodge. „Kruskal-Wallis Test". W: *The Concise Encyclopedia of Statistics*. New York, NY: Springer New York, 2008, s. 288–290. DOI: 10.1007/978-0-387-32833-1_216 (cytowane na stronie 83).

[34] Y. Dodge. „Mann–Whitney Test". W: *The Concise Encyclopedia of Statistics*. New York, NY: Springer New York, 2008, s. 327–329. DOI: 10.1007/978-0-387-32833-1_243 (cytowane na stronie 83).

[35] B. Emek Soylu, M. S. Guzel, G. E. Bostanci, F. Ekinci, T. Asuroglu i K. Acici. „Deep-Learning-Based Approaches for Semantic Segmentation of Natural Scene Images: A Review". W: *Electronics* vol. 12. no. 12 (2023). DOI: 10.3390/electronics12122730 (cytowane na stronach 37 i 38).

[36] H. Fan i H. Ling. „Dense Recurrent Neural Networks for Scene Labeling". W: *CoRR* vol. abs/1801.06831 (2018). arXiv: 1801.06831 (cytowane na stronie 39).

[37] H. Fan, X. Mei, D. V. Prokhorov i H. Ling. „Multi-level Contextual RNNs with Attention Model for Scene Labeling". W: *CoRR* vol. abs/1607.02537 (2016). arXiv: 1607.02537 (cytowane na stronie 39).

[38] C. Farabet, C. Couprie, L. Najman i Y. LeCun. „Learning Hierarchical Features for Scene Labeling". W: *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol. 35. no. 8 (2013), s. 1915–1929. DOI: 10.1109/TPAMI.2012.231 (cytowane na stronie 39).

[39] J. Fu, J. Liu, Y. Wang i H. Lu. „Stacked Deconvolutional Network for Semantic Segmentation". W: *CoRR* vol. abs/1708.04943 (2017). arXiv: 1708.04943 (cytowane na stronie 39).

[40] M. M. Galloway. „Texture analysis using gray level run lengths". W: *Computer Graphics and Image Processing* vol. 4. no. 2 (1975), s. 172–179. DOI: 10.1016/S0146-664X(75)80008-6 (cytowane na stronie 58).

[41] T. Giannakopoulos i A. Pikrakis. „Chapter 4 - Audio Features". W: *Introduction to Audio Analysis*. Red. T. Giannakopoulos i A. Pikrakis. Oxford: Academic Press, 2014, s. 59–103. DOI: 10.1016/B978-0-08-099388-1.00004-2 (cytowane na stronach 69 i 70).

[42] R. Girshick, J. Donahue, T. Darrell i J. Malik. „Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation". W: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, s. 580–587. DOI: 10.1109/CVPR.2014.81 (cytowane na stronach 33 i 38).

[43] R. B. Girshick. „Fast R-CNN". W: *CoRR* vol. abs/1504.08083 (2015). arXiv: 1504.08083 (cytowane na stronach 33 i 38).

[44] S. Grabias. „Mowa i jej zaburzenia". W: *Audiofonologia* vol. 10 (1997), s. 9–36 (cytowane na stronach 6 i 115).

[45] J. J. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper i H. J. Aerts. „Computational Radiomics System to Decode the Radiographic Phenotype". W: *Cancer Research* vol. 77. no. 21 (2017), e104–e107. DOI: 10.1158/0008-5472.CAN-17-0339 (cytowane na stronach 49, 50, 56, 58, 60, 61 i 63).

[46] Y.-P. Guan. „Automatic extraction of lips based on multi-scale wavelet edge detection". W: *IET Computer Vision* vol. 2 (1 2008), 23–33(10) (cytowane na stronie 9).

[47] A. Hair, P. Monroe, B. Ahmed, K. J. Ballard i R. Gutierrez-Osuna. „Apraxia World: A Speech Therapy Game for Children with Speech Sound Disorders". W: *Proceedings of the 17th ACM Conference on Interaction Design and Children*. IDC '18. Trondheim, Norway: Association for Computing Machinery, 2018, s. 119–131. DOI: 10.1145/3202185.3202733 (cytowane na stronie 9).

[48] M. Hall-Beyer. „Practical guidelines for choosing GLCM textures to use in landscape classification tasks over a range of moderate spatial scales". W: *International Journal of Remote Sensing* vol. 38. no. 5 (2017), s. 1312–1338. DOI: 10.1080/01431161.2016.1278314 (cytowane na stronach 55 i 56).

[49] W. Hao i S. Zhili. „Improved mosaic: algorithms for more complex images". W: *Journal of Physics: Conference Series*. T. 1684. 1. IOP Publishing. 2020, s. 012094. DOI: 10.1088/1742-6596/1684/1/012094 (cytowane na stronie 31).

[50] R. M. Haralick, K. Shanmugam i I. Dinstein. „Textural Features for Image Classification". W: *IEEE Transactions on Systems, Man, and Cybernetics* vol. SMC-3. no. 6 (1973), s. 610–621. DOI: 10.1109/TSMC.1973.4309314 (cytowane na stronach 55 i 56).

[51] K. He, X. Zhang, S. Ren i J. Sun. „Deep Residual Learning for Image Recognition". W: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, s. 770–778. DOI: 10.1109/CVPR.2016.90 (cytowane na stronach 46 i 92).

[52] Z. Huang, J. Miao, H. Song, S. Yang, Y. Zhong, Q. Xu, Y. Tan, C. Wen i J. Guo. „A novel tongue segmentation method based on improved U-Net". W: *Neurocomputing* vol. 500 (2022), s. 73–89. DOI: 10.1016/j.neucom.2022.05.023 (cytowane na stronie 10).

[53] A. Humeau-Heurtier. „Texture Feature Extraction Methods: A Survey". W: *IEEE Access* vol. 7 (2019), s. 8975–9000. DOI: 10.1109/ACCESS.2018.2890743 (cytowane na stronach 32 i 48).

[54] International Phonetic Association. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, 1999 (cytowane na stronie 1).

[55] N. Jamal, S. Shanta, F. Mahmud i M. Sha'abani. „Automatic speech recognition (ASR) based approach for speech therapy of aphasic patients: A review". W: *AIP Conference Proceedings* vol. 1883. no. 1 (2017), s. 020028. DOI: 10.1063/1.5002046 (cytowane na stronie 7).

[56] G. Jastrzębowska. *Podstawy teorii i diagnozy logopedycznej*. Wydawnictwo Uniwersytetu Opolskiego, 1998 (cytowane na stronach 4 i 6).

[57] C. Ji, T. B. Mudiyanselage, Y. Gao i Y. Pan. „A review of infant cry analysis and classification". W: *EURASIP Journal on Audio, Speech, and Music Processing* vol. 2021. no. 8 (2021). DOI: 10.1186/s13636-021-00197-5 (cytowane na stronach 69 i 70).

[58] D. Jiang, H. Qu, J. Zhao, J. Zhao i W. Liang. „Multi-level graph convolutional recurrent neural network for semantic image segmentation". W: *Telecommun Syst.* no. 77 (2021), s. 563–576. DOI: 10.1007/s11235-021-00769-y (cytowane na stronie 39).

[59] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, K. Michael, TaoXie, J. Fang, imyhxy i in. *ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation*. Wer. v7.0. 2022. DOI: 10.5281/zenodo.7347926 (cytowane na stronie 87).

[60] J. D. Johnston. „Transform coding of audio signals using perceptual noise criteria". W: *IEEE J. Sel. Areas Commun.* vol. 6 (1988), s. 314–323 (cytowane na stronie 72).

[61] W. Katz, S. Mehta, M. Wood i J. Wang. „Using Electromagnetic Articulography with a Tongue Lateral Sensor to Discriminate Manner of Articulation". W: *The Journal of the Acoustical Society of America* vol. 141. no. 1 (2017), s. 57–63. DOI: 10.1121/1.4973907 (cytowane na stronie 7).

[62] R. Kaur i S. Singh. „A comprehensive review of object detection with deep learning". W: *Digital Signal Processing* vol. 132 (2023), s. 103812. DOI: 10.1016/j.dsp.2022.103812 (cytowane na stronie 86).

[63] S. Koolagudi, Y. Srinivasa Murthy i S. Bhaskar. „Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition". W: *International Journal of Speech Technology* vol. 21 (2018), s. 1–17. DOI: 10.1007/s10772-018-9495-8 (cytowane na stronie 70).

[64] M. Kręcichowst. „Analiza przestrzennych modeli akustycznych głosek dentalizowanych w diagnostyce sygmatyzmu". rozprawa doktorska. Gliwice: Politechnika Śląska, 2020 (cytowane na stronie 69).

[65] M. Kręcichwost, Z. Miodońska, J. Trzaskalik i P. Badura. „Multichannel speech acquisition and analysis for computer-aided sigmatism diagnosis in children". W: *IEEE Access* vol. 8 (2020), s. 98647–98658. DOI: 10.1109/ACCESS.2020.2996413 (cytowane na stronach 18 i 20).

[66] M. Kręcichwost, N. Moćko i P. Badura. „Automated detection of sigmatism using deep learning applied to multichannel speech signal". W: *Biomedical Signal Processing and Control* vol. 68 (2021), s. 1–11. DOI: 10.1016/j.bspc.2021.102612 (cytowane na stronach 7, 8 i 67).

[67] M. Kręcichwost, A. Sage, Z. Miodońska i P. Badura. „4D Multimodal Speaker Model for Remote Speech Diagnosis". W: *IEEE Access* vol. 10 (2022), s. 93187–93202. DOI: 10.1109/ACCESS.2022.3203572 (cytowane na stronach 9, 18, 19 i 115).

[68] D. Król, A. Lorenc i R. Święciński. „Detecting Laterality and Nasality in Speech with the use of a Multi-channel Recorder". W: *IEEE International Conference on Acoustics, Speech and Signal Processing*. ICASSP"15. 2015, s. 5147–5151. DOI: 10.1109/ICASSP.2015.7178952 (cytowane na stronie 7).

[69] Y.-M. Kuo, S.-J. Ruan, Y.-C. Chen i Y.-W. Tu. „Deep-learning-based automated classification of Chinese speech sound disorders". W: *Children* vol. 9. no. 7 (2022), s. 996. DOI: 10.3390/children9070996 (cytowane na stronie 9).

[70] S. Lee i J.-E. Kim. „Evaluating the Precision of Automatic Segmentation of Teeth, Gingiva and Facial Landmarks for 2D Digital Smile Design Using Real-Time Instance Segmentation Network". W: *Journal of Clinical Medicine* vol. 11. no. 3 (2022), s. 852. DOI: 10.3390/jcm11030852 (cytowane na stronie 10).

[71] A. Lerch. „Instantaneous Features". W: *An Introduction to Audio Content Analysis*. John Wiley & Sons, Ltd, 2012. Rozd. 3, s. 31–69. DOI: 10.1002/9781118393550.ch3 (cytowane na stronie 72).

[72] S.-H. Leung, S.-L. Wang i W.-H. Lau. „Lip Image Segmentation Using Fuzzy Clustering Incorporating an Elliptic Shape Function". W: *IEEE Transactions on Image Processing* vol. 13. no. 1 (2004), s. 51–62. DOI: 10.1109/tip.2003.818116 (cytowane na stronie 9).

[73] C. Li, C. Xu, C. Gui i M. D. Fox. „Distance Regularized Level Set Evolution and Its Application to Image Segmentation". W: *IEEE Transactions on Image Processing* vol. 19. no. 12 (2010), s. 3243–3254. DOI: 10.1109/TIP.2010.2069690 (cytowane na stronie 42).

[74] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie i in. „YOLOv6: A single-stage object detection framework for industrial applications". W: *arXiv preprint arXiv:2209.02976* (2022). DOI: 10.48550/ARXIV.2209.02976 (cytowane na stronach 34 i 87).

[75] F. Li i B. Munson. „The development of voiceless sibilant fricatives in Putonghua-speaking children". W: *Journal of Speech, Language, and Hearing Research* vol. 59. no. 4 (2016), s. 699–712. DOI: 10.1044/2016_JSLHR-S-14-0142 (cytowane na stronie 8).

[76] Q. Li, H. Wang, B.-y. Li, T. Yanghua i J. Li. „IIE-SegNet: Deep Semantic Segmentation Network With Enhanced Boundary Based on Image Information Entropy". W: *IEEE Access* vol. PP (2021), s. 40612–40622. DOI: 10.1109/ACCESS.2021.3064346 (cytowane na stronie 38).

[77] R. Lienhart i J. Maydt. „An extended set of Haar-like features for rapid object detection". W: *Proceedings. International Conference on Image Processing*. T. 1. 2002, s. I–I. DOI: 10.1109/ICIP.2002.1038171 (cytowane na stronie 33).

[78] E. J. Limkin, S. Reuzé, A. Carré, R. Sun, A. Schernberg, A. Alexis, E. Deutsch, C. Ferté i C. Robert. „The complexity of tumor shape, spiculatedness, correlates with tumor radiomic shape features". W: *Scientific Reports* vol. 9. no. 4329 (2019), s. 2045–2322. DOI: 10.1038/s41598-019-40437-5 (cytowane na stronach 48 i 49).

[79] B. Lin, J. Xle, C. Li i Y. Qu. „Deeptongue: Tongue Segmentation Via Resnet". W: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, s. 1035–1039. DOI: 10.1109/icassp.2018.8462650 (cytowane na stronie 10).

[80] J. Liu i Y. Shi. „Image Feature Extraction Method Based on Shape Characteristics and Its Application in Medical Image Analysis". W: *Applied Informatics and Communication*. Red. D. Zeng. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, s. 172–178 (cytowane na stronie 49).

[81] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu i A. C. Berg. „SSD: Single Shot MultiBox Detector". W: *Computer Vision – ECCV 2016*. Red. B. Leibe, J. Matas, N. Sebe i M. Welling. Cham: Springer International Publishing, 2016, s. 21–37. DOI: 10.1007/978-3-319-46448-0_2 (cytowane na stronie 33).

[82] W. Liu, A. Rabinovich i A. C. Berg. „ParseNet: Looking Wider to See Better". W: *CoRR* vol. abs/1506.04579 (2015). arXiv: 1506.04579 (cytowane na stronie 38).

[83] T. Löfstedt, P. Brynolfsson, T. Asklund, T. Nyholm i A. Garpebring. „Gray-level invariant Haralick texture features". W: *PLOS ONE* vol. 14. no. 2 (2019), s. 1–18. DOI: 10.1371/journal.pone.0212110 (cytowane na stronach 52 i 53).

[84] A. Lorenc. „Kryteria diagnostyczne normy wymawianiowej". W: *Logopedia artystyczna* (2016), s. 168–193 (cytowane na stronie 115).

[85] A. Lorenc, D. Król i K. Klessa. „An acoustic camera approach to studying nasality in speech: The case of Polish nasalized vowels". W: *The Journal of the Acoustical Society of America* vol. 144. no. 6 (2018), s. 3603–3617. DOI: 10.1121/1.5084038 (cytowane na stronie 7).

[86] M. Lounis, B. Dendani i H. Bahi. „Mispronunciation detection and diagnosis using deep neural networks: a systematic review". W: *Multimedia Tools and Applications* (2024). DOI: 10.1007/s11042-023-17899-x (cytowane na stronie 6).

[87] D. G. Lowe. „Distinctive Image Features from Scale-Invariant Keypoints". W: *International Journal of Computer Vision* vol. 60 (2004), s. 91–110 (cytowane na stronie 33).

[88] Y. Lu i H. Liu. „Semantic segmentation with step-by-step upsampling of the fusion context". W: *2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*. 2021, s. 156–161. DOI: 10.1109/ICAICA52286.2021.9497923 (cytowane na stronach 38 i 39).

[89] S. Lucey, S. Sridharan i V. Chandran. „Adaptive mouth segmentation using chromatic features". W: *Pattern Recognition Letters* vol. 23. no. 11 (2002), s. 1293–1302. DOI: 10.1016/s0167-8655(02)00078-8 (cytowane na stronie 9).

[90] P. Łobacz i K. Dobrzańska. „Opis akustyczny glosek sybilantnych w wymowie dzieci przedszkolnych". W: *Audiofonologia* vol. 15 (1999), s. 7–26 (cytowane na stronie 74).

[91] K. El-Maleh, M. Klein, G. Petrucci i P. Kabal. „Speech/music discrimination for multimedia applications". W: *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*. T. 4. 2000, s. 2445–2448. DOI: 10.1109/ICASSP.2000.859336 (cytowane na stronie 70).

[92] V. Martinez-Paricio i J. Koreman. „The Spanish Computer-Assisted Listening and Speaking Tutor: A multilingual approach to pronunciation training". W: *Revista Espanola de Linguistica Aplicada/Spanish Journal of Applied Linguistics* (2024). DOI: 10.1075/resla.22046.mar (cytowane na stronie 6).

[93] S. Martins i S. Cavaco. „Customizable Serious Speech Therapy Games with Dynamic Difficulty Adjustment for Children with Sigmatism". W: *Studies in Health Technology and Informatics* vol. 290 (2022), s. 924–928. DOI: 10.3233/SHTI220215 (cytowane na stronie 7).

[94] A. Materka i M. Strzelecki. „Texture Analysis Methods - A Review". W: 1998 (cytowane na stronach 48, 51, 52 i 53).

[95] M. E. Mayerhoefer, A. Materka, G. Langs, I. Häggström, P. Szczypiński, P. Gibbs i G. Cook. „Introduction to Radiomics". W: *Journal of Nuclear Medicine* vol. 61. no. 4 (2020), s. 488–495. DOI: 10.2967/jnumed.118.222893 (cytowane na stronach 48, 52, 55 i 58).

[96] Z. B. Messaoud, D. Gargouri, S. Zribi i A. B. Hamida. „Formant Tracking Linear Prediction Model using HMMs for Noisy Speech Processing". W: *World Academy of Science, Engineering and Technology, International*

Journal of Electrical and Computer Engineering vol. 3. no. 11 (2009), s. 2102–2107 (cytowane na stronie 74).

[97] M. Miled, M. A. B. Messaoud i A. Bouzid. „Lip reading of words with lip segmentation and deep learning". W: *Multimedia Tools and Applications* vol. 82. no. 1 (2022), s. 551–571. DOI: 10.1007/s11042-022-13321-0 (cytowane na stronach 10 i 113).

[98] E. M. Minczakiewicz. „Dyslalia na tle innych wad i zaburzeń mowy u dzieci w wieku przedszkolnym i szkolnym". W: *Konteksty Pedagogiczne*. no. 8 (2017), s. 149–169. DOI: 10.19265/KP.2017.018149 (cytowane na stronie 1).

[99] Z. Miodonska, P. Badura i N. Mocko. „Noise-based acoustic features of Polish retroflex fricatives in children with normal pronunciation and speech disorder". W: *Journal of Phonetics* vol. 92 (2022), s. 101149. DOI: 10.1016/j.wocn.2022.101149 (cytowane na stronach 8, 69 i 74).

[100] Z. Miodońska, M. Kręcichwost, E. Kwaśniok, A. Sage i P. Badura. „Frication noise features of Polish voiceless dental fricative and affricate produced by children with and without speech disorder". W: *INTERSPEECH 2024*. (w druku). ISCA, 2024 (cytowane na stronach 8 i 74).

[101] H. Misra, S. Ikbal, H. Bourlard i H. Hermansky. „Spectral entropy based feature for robust ASR". W: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. T. 1. 2004. Rozd. I, s. 193–196. DOI: 10.1109/ICASSP.2004.1325955 (cytowane na stronie 72).

[102] D. Mitrović, M. Zeppelzauer i C. Breiteneder. „Chapter 3 - Features for Content-Based Audio Retrieval". W: *Advances in Computers: Improving the Web*. T. 78. Advances in Computers. Elsevier, 2010, s. 71–150. DOI: 10.1016/S0065-2458(10)78003-7 (cytowane na stronie 70).

[103] Y. Mo, Y. Wu, X. Yang, F. Liu i Y. Liao. „Review the state-of-the-art technologies of semantic segmentation based on deep learning". W: *Neurocomputing* vol. 493 (2022), s. 626–646. DOI: 10.1016/j.neucom.2022.01.005 (cytowane na stronie 37).

[104] D. Müller, A. Ehlen i B. Valeske. „Convolutional Neural Networks for Semantic Segmentation as a Tool for Multiclass Face Analysis in Thermal Infrared". W: *Journal of Nondestructive Evaluation* vol. 40. no. 1 (2021). DOI: 10.1007/s10921-020-00740-y (cytowane na stronie 9).

[105] P. Musa, F. A. Rafi i M. Lamsani. „A Review: Contrast-Limited Adaptive Histogram Equalization (CLAHE) methods to help the application of face recognition". W: *2018 Third International Conference on Informatics and Computing (ICIC)*. 2018, s. 1–6. DOI: 10.1109/IAC.2018.8780492 (cytowane na stronie 31).

[106] W. K. Mutlag, S. K. Ali, Z. M. Aydam i B. H. Taher. „Feature Extraction Methods: A Review". W: *Journal of Physics: Conference Series* vol. 1591. no. 1 (2020), s. 012028. DOI: 10.1088/1742-6596/1591/1/012028 (cytowane na stronach 49, 52 i 53).

[107] N. E. Naal-Ruiz, E. A. Gonzalez-Rodriguez, G. Navas-Reascos, R. Romo-De Leon, A. Solorio, L. M. Alonso-Valerdi i D. I. Ibarra-Zarate. „Mouth Sounds: A Review of Acoustic Applications and Methodologies". W: *Applied Sciences* vol. 13. no. 7 (2023). DOI: 10.3390/app13074331 (cytowane na stronach 69 i 70).

[108] V. Narayan, M. Faiz, P. K. Mall i S. Srivastava. „A Comprehensive Review of Various Approach for Medical Image Segmentation and Disease Prediction". W: *Wireless Pers Commun.* no. 132 (2023), s. 1819–1848. DOI: 10.1007/s11277-023-10682-z (cytowane na stronie 37).

[109] S. L. Nissen i R. A. Fox. „Acoustic and spectral characteristics of young children's fricative productions: a developmental perspective". W: *J Acoust Soc Am* vol. 118. no. 4 (2005), s. 2570–2578 (cytowane na stronie 8).

[110] H. Noh, S. Hong i B. Han. „Learning Deconvolution Network for Semantic Segmentation". W: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, s. 1520–1528. DOI: 10.1109/ICCV.2015.178 (cytowane na stronie 38).

[111] M. Osowicka-Kondratowicz i A. Serowik. „Defektywne realizacje spółgłosek palatalnych dentalizowanych przy prawidłowych i nieprawidłowych warunkach zgryzowych. Wskazówki do terapii logopedycznej". W: *Prace Językoznawcze* vol. 11 (2009), s. 155–172 (cytowane na stronie 3).

[112] B. Ostapiuk. *Dyslalia: o badaniu jakosci wymowy w logopedii*. Wydawnictwo Naukowe Uniwersytetu Szczecinskiego, 2013. ISBN: 978-83-7241-893-7 (cytowane na stronach 4, 5, 27, 80 i 81).

[113] B. Ostapiuk. „Zaburzenia dźwiękowej realizacji fonemów języka polskiego – propozycja terminów i klasyfikacji". W: *Audiofonologia* vol. 10 (1997), s. 117–130 (cytowane na stronach 4, 5, 27, 80 i 81).

[114] R. Padilla, W. Lobato Passos, T. Dias, S. Netto i E. da Silva. „A Comparative Analysis of Object Detection Metrics with a Companion Open-Source Toolkit". W: *Electronics* vol. 10 (2021), s. 279–306. DOI: 10.3390/electronics10030279 (cytowane na stronie 86).

[115] R. Padilla, S. L. Netto i E. A. B. da Silva. „A Survey on Performance Metrics for Object-Detection Algorithms". W: *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. 2020, s. 237–242. DOI: 10.1109/IWSSIP48289.2020.9145130 (cytowane na stronie 86).

[116] Panasonic. *Omnidirectional Back Electret Condenser Microphone Cartridge, Series: WM-61A, WM-61B*. http://konektor.nazwa.pl/serwisowe/panasonic-wm-61a.pdf. Dostęp: 04.11.2023. (cytowane na stronie 19).

[117] V. Parekh i M. A. Jacobs. „Radiomics: a new application from established techniques". W: *Expert Review of Precision Medicine and Drug Development* vol. 1. no. 2 (2016), s. 207–226. DOI: 10.1080/23808993.2016.1164013 (cytowane na stronach 48, 52, 53, 55 i 56).

[118] C. Patgiri, M. Sarma i K. Sarma. „A Class of Neuro-Computational Methods for Assamese Fricative Classification". W: *Journal of Artificial Intelligence and Soft Computing Research* vol. 5 (2015). DOI: 10.1515/jaiscr-2015-0019 (cytowane na stronie 8).

[119] G. Peeters. „A large set of audio features for sound description (similarity and classification) in the CUIDADO project". W: *CUIDADO Ist Project Report* vol. 54. no. 0 (2004), s. 1–25 (cytowane na stronie 72).

[120] N. Raman, R. Nagarajan, L. Venkatesh, D. S. Monica, V. Ramkumar i M. Krumm. „School-based language screening among primary school children using telepractice: A feasibility study from India". W: *International Journal of Speech-Language Pathology* vol. 21. no. 4 (2019), s. 425–434. DOI: 10.1080/17549507.2018.1493142 (cytowane na stronie 8).

[121] T. Rebernik, J. Jacobi, R. Jonkers, A. Noiray i M. Wieling. „A review of data collection practices using electromagnetic articulography". W: *Laboratory Phonology: Journal of the Association for Laboratory Phonology* vol. 12 (2021), s. 6. DOI: 10.5334/labphon.237 (cytowane na stronie 7).

[122] J. Redmon, S. Divvala, R. Girshick i A. Farhadi. „You Only Look Once: Unified, Real-Time Object Detection". W: t. abs/1506.02640. 2016, s. 779–788. DOI: 10.1109/CVPR.2016.91 (cytowane na stronach 32 i 33).

[123] J. Redmon i A. Farhadi. „YOLOv3: An Incremental Improvement". W: *CoRR* vol. abs/1804.02767 (2018). arXiv: 1804.02767 (cytowane na stronie 87).

[124] P. F. Reidy. „Spectral dynamics of sibilant fricatives are contrastive and language specific". W: *The Journal of the Acoustical Society of America* vol. 140. no. 4 (2016). DOI: 10.1121/1.4964510 (cytowane na stronie 74).

[125] J. Rusz, J. Hlavnička, T. Tykalová, M. Novotný, P. Du šek, K. Šonka i E. ž. Rùžička. „Smartphone Allows Capture of Speech Abnormalities Associated With High Risk of Developing Parkinson's Disease". W: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* vol. 26.

no. 8 (2018), s. 1495–1507. DOI: 10.1109/TNSRE.2018.2851787 (cytowane na stronie 9).

[126] A. Sage, Z. Miodońska, M. Kręcichwost, J. Trzaskalik, E. Kwaśniok i P. Badura. „Deep Learning Approach to Automated Segmentation of Tongue in Camera Images for Computer-Aided Speech Diagnosis". W: *Advances in Intelligent Systems and Computing*. Springer International Publishing, 2020, s. 41–51. DOI: 10.1007/978-3-030-49666-1_4 (cytowane na stronach 10 i 113).

[127] A. O. Salau i S. Jain. „Feature Extraction: A Survey of the Types, Techniques, Applications". W: *2019 International Conference on Signal Processing and Communication (ICSC)*. 2019, s. 158–164. DOI: 10.1109/ICSC45622.2019.8938371 (cytowane na stronie 48).

[128] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov i L. Chen. „Inverted Residuals and Linear Bottlenecks: Mobile Networks for Classification, Detection and Segmentation". W: *CoRR* vol. abs/1801.04381 (2018). arXiv: 1801.04381 (cytowane na stronach 46 i 92).

[129] P. Sanghani, A. B. Ti, N. K. Kam King i H. Ren. „Evaluation of tumor shape features for overall survival prognosis in glioblastoma multiforme patients". W: *Surgical Oncology* vol. 29 (2019), s. 178–183. DOI: 10.1016/j.suronc.2019.05.005 (cytowane na stronie 50).

[130] C. Scapicchio, M. Gabelloni, A. Barucci, D. Cioni i E. Saba Luca Neri. „A deep look into radiomics". W: *La Radiologia Medica* vol. 126 (2021), s. 1296–1311. DOI: 10.1007/s11547-021-01389-x (cytowane na stronach 48 i 52).

[131] E. Scheirer i M. Slaney. „Construction and evaluation of a robust multifeature speech/music discriminator". W: *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*. T. 2. 1997, 1331–1334 vol.2. DOI: 10.1109/ICASSP.1997.596192 (cytowane na stronie 72).

[132] P. Schober, C. Boer i L. A. Schwarte. „Correlation Coefficients: Appropriate Use and Interpretation". W: *Anesthesia & Analgesia* vol. 126. no. 5 (2018), s. 1763–1768. DOI: 10.1213/ANE.0000000000002864 (cytowane na stronie 77).

[133] U. Sehar i M. Naseem. „How deep learning is empowering semantic segmentation". W: *Multimed Tools Appl.* no. 81 (2022), s. 30519–30544. DOI: 10.1007/s11042-022-12821-3 (cytowane na stronie 37).

[134] R. Serizel, V. Bisot, S. Essid i G. Richard. „Acoustic Features for Environmental Sound Analysis". W: *Computational Analysis of Sound Scenes and Events*. Cham: Springer International Publishing, 2018, s. 71–101. DOI: 10.1007/978-3-319-63450-0_4 (cytowane na stronie 69).

[135] G. Sharma, K. Umapathy i S. Krishnan. „Trends in audio signal feature extraction methods". W: *Applied Acoustics* vol. 158 (2020), s. 107020. DOI: 10.1016/j.apacoust.2019.107020 (cytowane na stronie 69).

[136] D. Shen, Y. Ji, P. Li, Y. Wang i D. Lin. „RANet: Region Attention Network for Semantic Segmentation". W: *Neural Information Processing Systems*. 1168. 2020, s. 13927–13938 (cytowane na stronie 38).

[137] D. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC Press, 2000. ISBN: 9780584881332 (cytowane na stronach 80 i 83).

[138] E. Skorek. *Oblicza wad wymowy*. Wydawnictwo Akademickie „Żak", 2001 (cytowane na stronach 1, 4, 5 i 6).

[139] R. C. Snell i F. Milinazzo. „Formant location from LPC analysis data." W: *IEEE Trans. Speech and Audio Processing* vol. 1. no. 2 (1993), s. 129–134. DOI: 10.1109/89.222882 (cytowane na stronie 74).

[140] A. Sołtys-Chmielowicz. *Zaburzenia artykulacji. Teoria i praktyka*. Oficyna Wydawnicza Impuls, 2016 (cytowane na stronach 3, 4, 5 i 6).

[141] L. Spinu i J. Lilley. „A comparison of cepstral coefficients and spectral moments in the classification of Romanian fricatives". W: *Journal of Phonetics* vol. 57 (2016), s. 40–58. DOI: 10.1016/j.wocn.2016.05.002 (cytowane na stronie 8).

[142] S. Srivastava, A. Divekar, C. Anilkumar, I. Naik, V. Kulkarni i P. V. *Comparative Analysis of Deep Learning Image Detection Algorithms*. 2021. DOI: 10.1186/s40537-021-00434-w (cytowane na stronie 33).

[143] M. Stojmenović i J. Žunić. „Measuring Elongation from Shape Boundary". W: *Journal of Mathematical Imaging and Vision* vol. 30 (2008), s. 73–85. DOI: 10.1007/s10851-007-0039-0 (cytowane na stronie 49).

[144] M. Strzelecki, P. Szczypinski, A. Materka i A. Klepaczko. „A software tool for automatic classification and segmentation of 2D/3D medical images". W: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* vol. 702 (2013), s. 137–140. DOI: 10.1016/j.nima.2012.09.006 (cytowane na stronach 48 i 52).

[145] M. Strzelecki i A. Materka. *Tekstura obrazów biomedycznych: Metody analizy komputerowej*. Wydawnictwo Naukowe PWN, 2017 (cytowane na stronie 56).

[146] I. Styczek. *Logopedia*. Wydawnictwo Naukowe PWN, 1980 (cytowane na stronach 1, 2, 3, 4, 5 i 8).

[147] J. Sun i Y. Li. „Multi-feature fusion network for road scene semantic segmentation". W: *Computers & Electrical Engineering* vol. 92 (2021), s. 107155. DOI: 10.1016/j.compeleceng.2021.107155 (cytowane na stronie 38).

[148] P. M. Szczypiński, M. Strzelecki, A. Materka i A. Klepaczko. „MaZda—A software package for image texture analysis". W: *Computer Methods and Programs in Biomedicine* vol. 94. no. 1 (2009), s. 66–76. DOI: 10.1016/j.cmpb.2008.08.005 (cytowane na stronach 48 i 52).

[149] R. Tadeusiewicz. *Sygnał mowy*. Wydawnictwa Komunikacji i Łączności, 1988. ISBN: 83-206-0705-1 (cytowane na stronie 67).

[150] J. Tambor i D. Ostaszewska. *Fonetyka i fonologia współczesnego języka polskiego*. Wydawnictwo Naukowe PWN, 2009 (cytowane na stronach 2, 3 i 4).

[151] X. Tang. „Texture information in run-length matrices". W: *IEEE Transactions on Image Processing* vol. 7. no. 11 (1998), s. 1602–1609. DOI: 10.1109/83.725367 (cytowane na stronach 58, 60 i 61).

[152] J. Teuwen i N. Moriakov. „Chapter 20 - Convolutional neural networks". W: *Handbook of Medical Image Computing and Computer Assisted Intervention*. The Elsevier and MICCAI Society Book Series. Academic Press, 2020, s. 481–501. DOI: 10.1016/B978-0-12-816176-0.00025-9 (cytowane na stronie 38).

[153] Texas Instruments. *TLV6741, TLV6742, TLV6744 10-MHz, Low Broadband Noise, RRO, Operational Amplifier*. https://www.ti.com/lit/ds/symlink/tlv6741.pdf. Dostęp: 29.11.2023 (cytowane na stronie 19).

[154] G. Thibault, B. Fertil, C. Navarro, S. Pereira, N. Lévy, J. Sequeira i J.-L. Mari. „Texture Indexes and Gray Level Size Zone Matrix Application to Cell Nuclei Classification". W: 2009 (cytowane na stronach 60 i 61).

[155] M. Toda, S. Maeda i K. Honda. „Formant-cavity affiliation in sibilant fricatives". W: *Turbulent Sounds*. Red. S. Fuchs, M. Toda i M. Zygis. Interface Explorations. Berlin: De Gruyter Mouton, 2010, s. 341–371. DOI: 10.1515/9783110226584 (cytowane na stronach 8 i 74).

[156] B. D. Tran, M. Tai-Seale, R. Mangu, J. Lafata i K. Zheng. „Automatic speech recognition performance for digital scribes: a performance comparison between general-purpose and specialized models tuned for patient-clinician conversations". W: *AMIA 2022, American Medical Informatics Association Annual Symposium*. AMIA, 2022 (cytowane na stronie 7).

[157] A. Trochymiuk i R. Święciński. „Artykulograficzne badanie wymowy grzbietowej. Studium przypadku". W: *Logopedia* vol. 38 (2009), s. 173–201 (cytowane na stronie 4).

[158] J. Trzaskalik, E. Kwaśniok, Z. Miodońska, M. Kręcichwost, A. Sage i P. Badura. „Hybrid System for Acquisition and Processing of Multimodal Signal: Population Study on Normal and Distorted Pronunciation of Sibilants in Polish Preschool Children". W: *XXIII Polish Conference on Biocybernetics and Biomedical Engineering, Lodz, September 25-27, 2023. Abstract Book*. 2023, s. 81 (cytowane na stronach 17, 27 i 28).

[159] M. Vakalopoulou, S. Christodoulidis, N. Burgos, O. Colliot i V. Lepetit. „Deep Learning: Basics and Convolutional Neural Networks (CNNs)". W: *Machine Learning for Brain Disorders*. New York, NY: Springer US, 2023, s. 77–115. DOI: 10.1007/978-1-0716-3195-9_3 (cytowane na stronie 38).

[160] B. A. Varghese, S. Y. Cen, D. H. Hwang i V. A. Duddalwar. „Texture Analysis of Imaging: What Radiologists Need to Know". W: *American Journal of Roentgenology* vol. 212. no. 3 (2019), s. 520–528. DOI: 10.2214/AJR.18.20624 (cytowane na stronie 52).

[161] C.-Y. Wang, A. Bochkovskiy i H.-Y. M. Liao. „YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors". W: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, s. 7464–7475. DOI: 10.1109/CVPR52729.2023.00721 (cytowane na stronie 87).

[162] S.-L. Wang, W.-H. Lau, A. W.-C. Liew i S.-H. Leung. „Robust lip region segmentation for lip images with complex background". W: *Pattern Recognition* vol. 40. no. 12 (2007), s. 3481–3491. DOI: 10.1016/j.patcog.2007.03.016 (cytowane na stronie 9).

[163] R. Wielgat, R. Jędryka, A. Lorenc, Ł. Mik i D. Król. „POLEMAD–A database for the multimodal analysis of Polish pronunciation". W: *Speech Communication* vol. 127 (2021), s. 29–42. DOI: 10.1016/j.specom.2020.12.005 (cytowane na stronach 7 i 114).

[164] B. Wierzchowska. *Fonetyka i fonologia języka polskiego*. Zakład Narodowy im. Ossolińskich, 1980 (cytowane na stronie 3).

[165] S. Wood, J. Wishart, W. Hardcastle, J. Cleland i C. Timmins. „The use of Electropalatography (EPG) in the Assessment and Treatment of Motor Speech Disorders in Children with Down's Syndrome: Evidence from two Case Studies". W: *Developmental Neurorehabilitation* vol. 12. no. 2 (2009), s. 66–75. DOI: 10.1080/17518420902738193 (cytowane na stronie 7).

[166] J. Yang i L. Xu. „Acoustic characteristics of sibilant fricatives and affricates in Mandarin-speaking children with cochlear implants". W: *The Journal of the Acoustical Society of America* vol. 153. no. 6 (2023), s. 3501–3512. DOI: 10.1121/10.0019803 (cytowane na stronie 74).

[167] H. Yu, Z. Yang, L. Tan, Y. Wang, W. Sun, M. Sun i Y. Tang. „Methods and datasets on semantic segmentation: A review". W: *Neurocomputing* vol. 304 (2018), s. 82–103. DOI: 10.1016/j.neucom.2018.03.037 (cytowane na stronach 37 i 38).

[168] Y. Yu, C. Wang, Q. Fu, R. Kou, F. Huang, B. Yang, T. Yang i M. Gao. „Techniques and Challenges of Image Segmentation: A Review". W: *Electronics* vol. 12. no. 5 (2023). DOI: 10.3390/electronics12051199 (cytowane na stronie 37).

[169] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar i B. Lee. „A survey of modern deep learning based object detection models". W: *Digital Signal Processing* vol. 126 (2022), s. 103514. ISSN: 1051-2004. DOI: 10.1016/j.dsp.2022.103514 (cytowane na stronie 32).

[170] G. Zeng, W. Yu, R. Wang i A. Lin. „Research on Mosaic Image Data Enhancement for Overlapping Ship Targets". W: *arXiv preprint arXiv:2105.05090* (2021). DOI: 10.48550/arXiv.2105.05090 (cytowane na stronie 31).

[171] D. Zhang i G. Lu. „Review of shape representation and description techniques". W: *Pattern Recognition* vol. 37. no. 1 (2004), s. 1–19. DOI: 10.1016/j.patcog.2003.07.008 (cytowane na stronie 48).

[172] H. Zhang, K. J. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi i A. Agrawal. „Context Encoding for Semantic Segmentation". W: *CoRR* vol. abs/1803.08904 (2018). arXiv: 1803.08904 (cytowane na stronie 38).

[173] W. Zhang, Y. Guo i Q. Jin. „Radiomics and Its Feature Selection: A Review". W: *Symmetry* vol. 15. no. 10 (2023). DOI: 10.3390/sym15101834 (cytowane na stronie 52).

[174] Z. Zhang, X. Zhang, C. Peng, D. Cheng i J. Sun. „ExFuse: Enhancing Feature Fusion for Semantic Segmentation". W: *CoRR* vol. abs/1804.03821 (2018). arXiv: 1804.03821 (cytowane na stronie 38).

[175] Z.-Q. Zhao, P. Zheng, S.-T. Xu i X. Wu. „Object Detection With Deep Learning: A Review". W: *IEEE Transactions on Neural Networks and Learning Systems* vol. PP (2019), s. 1–21. DOI: 10.1109/TNNLS.2018.2876865 (cytowane na stronie 32).

[176] N. Zharkova, W. J. Hardcastle i F. E. Gibbon. „The dynamics of voiceless sibilant fricative production in children between 7 and 13 years old: An ultrasound and acoustic study". W: *J Acoust Soc Am* vol. 144. no. 3 (2018), s. 1454. DOI: 10.1121/1.5053585 (cytowane na stronie 8).

[177] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr i L. Zhang. „Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers". W: *CoRR* vol. abs/2012.15840 (2020). arXiv: 2012.15840 (cytowane na stronie 38).

[178] C. Zhou, H. Fan i Z. Li. „Tonguenet: Accurate Localization and Segmentation for Tongue Images Using Deep Neural Networks". W: *IEEE Access* vol. 7 (2019), s. 148779–148789. DOI: 10.1109/access.2019.2946681 (cytowane na stronie 10).

[179] J. Zhou, Q. Zhang, B. Zhang i X. Chen. „TongueNet: A Precise and Fast Tongue Segmentation System Using U-Net with a Morphological Processing Layer". W: *Applied Sciences* vol. 9. no. 15 (2019), s. 3128. DOI: 10.3390/app9153128 (cytowane na stronie 10).

[180] G. Zhu, Z. Piao i S. C. Kim. „Tooth Detection and Segmentation with Mask R-CNN". W: *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE, 2020, s. 070–072. DOI: 10.1109/icaiic48513.2020.9065216 (cytowane na stronie 10).

[181] J. Žunić. „Shape Descriptors for Image Analysis". W: *Zbornik Radova*. no. 23 (2012), s. 5–38 (cytowane na stronie 49).

[182] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. W. L. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard i in. „The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping". W: *Radiology* vol. 295. no. 2 (2020). PMID: 32154773, s. 328–338. DOI: 10.1148/radiol.2020191145 (cytowane na stronach 55 i 56).

[183] M. Żygis i J. Padgett. „A perceptual study of Polish fricatives, and its implications for historical sound change". W: *Journal of Phonetics* vol. 38. no. 2 (2010), s. 207–226. DOI: 10.1016/j.wocn.2009.10.003 (cytowane na stronie 8).

Dodatek A: Wyniki analizy eksploracyjnej danych

Dodatek zawiera szczegółowe wyniki testów przeprowadzonych w ramach analizy eksploracyjnej danych. Tab. A.1 gromadzi symbole cech wykorzystane w dalszych zestawieniach. Tab. A.2 przedstawia podsumowanie rezultatów analizy normalności rozkładów zmiennych za pomocą testu Shapiro-Wilka dla każdej głoski i cechy artykulacyjnej, a tab. A.3-A.5 prezentują wyniki testu Browna-Forsythe'a oraz, w nielicznych przypadkach, stosunków wariancji dla kolejnych szeregów fonemów (zamieszczono jedynie najistotniejsze cechy).

Tab. A.1: Legenda oznaczeń rodzajów cech w tabelach dodatku A.

Cechy obrazowe	W	kształtu warg
	U	kształtu ust
	J	kształtu języka
	T	teksturowe
Cechy akustyczne	C	czasowe
	F	częstotliwościowe
	N	szumowe

A.1 Wyniki analizy normalności rozkładu cech (test Shapiro-Wilka)

Tab. A.2: Podsumowanie testu Shapiro-Wilka dla każdej głoski i wybranych cech artykulacyjnych (analizowane cechy zawierały maksymalnie 3 grupy). W tabeli zaznaczono, dla ilu cech spośród wszystkich możliwych (472) odrzucono hipotezę zerową o normalności rozkładu w każdej z grup.

#1 dentalizacja	#5 skrócenie wędzidełka języka
#2 dentalność	#6 medialność wypływu powietrza
#3 postdentalność	#7 medialność języka
#4 apikalność	#8 medialność żuchwy

			#1	#2	#3	#4	#5	#6	#7	#8
Szereg syczący										
/s/	p<0,05	**Grupa 1**	324	313		240	268	331	326	332
		Grupa 2	197	183		344	268	169	120	169
		Grupa 3					234	191	194	143
/z/	p<0,05	**Grupa 1**	303	299		260	272	331	327	319
		Grupa 2	191	188		336	261	136	117	162
		Grupa 3					215	174	151	158
/ts/	p<0,05	**Grupa 1**	315	326		271	225	314	312	322
		Grupa 2	197	196		339	274	160	115	163
		Grupa 3					250	219	211	176
/dz/	p<0,05	**Grupa 1**	312	319		244	266	331	325	335
		Grupa 2	188	180		328	258	171	95	145
		Grupa 3					210	180	213	162
Szereg szumiący										
/ʂ/	p<0,05	**Grupa 1**	336		331	351	293		362	354
		Grupa 2	189		206	228	284		82	105
		Grupa 3			187		223		53	187
/ʐ/	p<0,05	**Grupa 1**	346		321	336	299			344
		Grupa 2	166		192	230	253			128
		Grupa 3			218		244			159
/tʂ/	p<0,05	**Grupa 1**	324		318	337	268		338	321
		Grupa 2	187		213	225	260		72	125
		Grupa 3			208		236		62	146
/dʐ/	p<0,05	**Grupa 1**	318		300	342	251			318
		Grupa 2	194		170	201	261			194
		Grupa 3			192		234			

Kontynuacja tabeli na następnej stronie

			#1	#2	#3	#4	#5	#6	#7	#8
Szereg ciszący										
/ɕ/	**p<0,05**	**Grupa 1**	328		354	301	282	334	347	333
		Grupa 2	204		147	298	284	101	122	120
		Grupa 3					224			167
/ʑ/	**p<0,05**	**Grupa 1**	319			297	248	332	328	320
		Grupa 2	176			272	249	82	79	119
		Grupa 3					222			134
/tɕ/	**p<0,05**	**Grupa 1**	331			313	286	357	367	334
		Grupa 2	180			270	270	91	80	120
		Grupa 3					225			166
/dʑ/	**p<0,05**	**Grupa 1**	316			285	233	347	342	327
		Grupa 2	164			279	265	95	98	88
		Grupa 3					215			150

A.2 Wyniki analizy jednorodności wariancji (test Browna-Forsythe'a)

Tab. A.3: Wyniki analizy jednorodności wariancji dla głosek szeregu syczącego testem Browna-Forsythe'a. Dla cech, w których odrzucono hipotezę zerową, dodatkowo przedstawiono stosunki wariancji pomiędzy grupami.

		Cecha	**Dane**	**Rodzaj**	**2D/3D**	**Kamera**	BF p	σ^2_{12}	σ^2_{13}	σ^2_{23}
Głoska /s/	**#1**	Ax^{3D}_{minor}	V	3D	J	L	0,627			
		Ax^{3D}_{least}	V	3D	J	L	0,317			
		D^{3D}_{YZ}	V	3D	J	L	0,458			
		Ax^{3D}_{least}	V	3D	J	R	0,400			
		Ax^{3D}_{minor}	V	3D	J	R	0,518			
		D^{3D}_{YZ}	V	3D	J	R	0,819			
		V^{3D}	V	3D	J	R	0,144			
		V^{3D}	V	3D	J	L	0,207			
	#2	V^{3D}	V	3D	J	L	0,701			
		Ax^{3D}_{least}	V	3D	J	L	0,481			
		Ax^{3D}_{minor}	V	3D	J	L	0,355			
		D^{3D}_{XY}	V	3D	J	L	0,444			
		D^{3D}_{Feret}	V	3D	J	L	0,767			
		D^{3D}_{YZ}	V	3D	J	L	0,328			
		D^{3D}_{XZ}	V	3D	J	L	0,389			
		Ax^{3D}_{major}	V	3D	J	L	0,230			
		D^{2D}_{Feret}	V	2D	J	R	0,817			
		A^{2D}	V	2D	J	R	0,467			
		Ax^{3D}_{least}	V	3D	J	R	0,587			
		Dia^{3D}_{Feret}	V	3D	J	L	0,679			
		Ax^{3D}_{major}	V	3D	J	R	0,679			
		SVR^{3D}	V	3D	J	L	0,941			
		Ax^{2D}_{minor}	V	2D	J	L	0,517			
		V^{3D}	V	3D	J	R	0,620			
	#6	P^{2D}	V	2D	J	L	0,615			

Kontynuacja tabeli na następnej stronie

		Cecha	**Dane**	**Rodzaj**	**2D/3D**	**Kamera**	BF p	σ^2_{12}	σ^2_{13}	σ^2_{23}
Głoska /z/	**#1**	V^{3D}	V	3D	J	R	0,403			
	#2	D^{3D}_{XZ}	V	3D	J	R	0,668			
		D^{3D}_{XY}	V	3D	J	R	0,623			
		V^{3D}	V	3D	J	L	0,225			
	#5	I^{3D}_{h}	V	3D	T	R	0,237			
		A^{2D}	V	2D	W	R	0,055			
		DS^{2D}	V	2D	W	R	0,191			
	#7	A^{2D}	V	2D	J	L	0,688			
		P^{2D}	V	2D	J	L	0,871			
Głoska /ts/	**#1**	Ax^{3D}_{major}	V	3D	J	L	0,388			
	#2	Ax^{3D}_{major}	V	3D	J	L	0,872			
		A^{3D}	V	3D	J	L	0,598			
		D^{3D}_{XZ}	V	3D	J	L	0,845			
		D^{3D}_{XY}	V	3D	J	L	0,514			
		V^{3D}	V	3D	J	L	0,932			
		D^{3D}_{Feret}	V	3D	J	L	0,415			
		DS^{3D}	V	3D	J	L	0,285			
		Ax^{3D}_{least}	V	3D	J	L	0,619			
		Ax^{3D}_{minor}	V	3D	J	L	0,609			
		Ax^{3D}_{minor}	V	3D	J	L	0,923			
		V^{3D}	V	3D	J	R	0,765			
		D^{3D}_{YZ}	V	3D	J	R	0,952			
		A^{3D}	V	3D	J	R	0,516			
		D^{3D}_{YZ}	V	3D	J	L	0,417			
	#4	V^{3D}	V	3D	J	L	0,444			
		Ax^{3D}_{least}	V	3D	J	L	0,963			
		D^{3D}_{YZ}	V	3D	J	L	0,046	2,260		
		D^{3D}_{XZ}	V	3D	J	L	0,409			
		D^{3D}_{XY}	V	3D	J	L	0,931			
		D^{3D}_{Feret}	V	3D	J	L	0,611			
		Ax^{3D}_{minor}	V	3D	J	L	0,083			
		Ax^{3D}_{major}	V	3D	J	L	0,768			
		A^{3D}	V	3D	J	L	0,575			

Kontynuacja tabeli na następnej stronie

		Cecha	Dane	Rodzaj	2D/3D	Kamera	BF p	σ^2_{12}	σ^2_{13}	σ^2_{23}
	#5	E^{2D}	V	2D	J	R	0,282			
		S^{3D}	V	3D	J	L	0,503			
		C_1^{3D}	V	3D	J	L	0,539			
		C_2^{3D}	V	3D	J	L	0,510			
		DS^{3D}	V	3D	J	L	0,590			
	#7	$Dia_{col}x^{3D}$	V	3D	J	L	0,739			
		Ax^{3D}_{minor}	V	3D	J	L	0,225			
		A^{3D}	V	3D	J	L	0,280			
		Ax^{3D}_{minor}	V	3D	J	R	0,536			
		E^{3D}	V	3D	J	R	0,334			
		V^{3D}	V	3D	J	L	0,770			
Głoska /dz/	#1	A^{2D}	V	2D	J	R	0,608			
		DS^{2D}	V	2D	J	R	0,159			
		DS^{2D}	V	2D	J	L	0,625			
		S^{2D}	V	2D	J	R	0,134			
		P^{2D}	V	2D	J	L	0,196			
		S^{2D}	V	2D	J	L	0,350			
		A^{2D}	V	2D	J	L	0,408			
	#2	A^{2D}	V	2D	J	L	0,135			
		A^{2D}	V	2D	J	R	0,791			
		P^{2D}	V	2D	J	L	0,229			
		DS^{2D}	V	2D	J	L	0,546			
		V^{3D}	V	3D	J	R	0,042	0,701		
		Ax^{2D}_{minor}	V	2D	J	R	0,931			
		DS^{2D}	V	2D	J	R	0,828			
		A^{3D}	V	3D	J	R	0,194			
		D^{3D}_{XZ}	V	3D	J	R	0,084			
		S^{2D}	V	2D	J	R	0,314			
		Ax^{3D}_{major}	V	3D	J	R	0,500			
		D^{3D}_{Feret}	V	3D	J	R	0,055			
		D^{3D}_{XY}	V	3D	J	R	0,092			
		DS^{3D}	V	3D	J	R	0,649			
		D^{2D}_{Feret}	V	2D	J	R	0,920			
		Ax^{3D}_{minor}	V	3D	J	R	0,249			
		D^{3D}_{YZ}	V	3D	J	R	0,170			
		S^{3D}	V	3D	J	R	0,883			
		C_1^{3D}	V	3D	J	R	0,917			

Kontynuacja tabeli na następnej stronie

		Cecha	**Dane**	**Rodzaj**	**2D/3D**	**Kamera**	BF p	σ_{12}^2	σ_{13}^2	σ_{23}^2
		C_2^{3D}	V	3D	J	R	0,958			
		S^{2D}	V	2D	J	L	0,387			
		D_{Feret}^{2D}	V	2D	J	L	0,296			
		Ax_{least}^{3D}	V	3D	J	R	0,888			
		V^{3D}	V	3D	J	L	0,674			
	#4	DS^{2D}	V	2D	J	R	0,175			
		A^{2D}	V	2D	J	R	0,630			
		S^{2D}	V	2D	J	R	0,175			
		E^{2D}	V	2D	J	L	0,586			
	#6	D_{Feret}^{2D}	V	3D	J	L	0,291			
		Ax_{major}^{2D}	V	2D	J	L	0,128			
	#7	A^{3D}	V	3D	J	R	0,085			
		DS^{3D}	V	3D	J	R	0,100			
		Ax_{least}^{3D}	V	3D	J	R	0,342			
	#8	D_{Feret}^{2D}	V	2D	J	L	0,622			
		A^{3D}	V	3D	J	L	0,702			
		NCC_{10}	A		N		0,081			

Tab. A.4: Wyniki analizy jednorodności wariancji dla głosek szeregu szumiącego testem Browna-Forsythe'a. Dla cech, w których odrzucono hipotezę zerową, dodatkowo przedstawiono stosunki wariancji pomiędzy grupami.

		Cecha	**Dane**	**Rodzaj**	**2D/3D**	**Kamera**	BF p	σ^2_{12}	σ^2_{13}	σ^2_{23}
Głoska /ʂ/		NPF	A		N		0,202			
		FFR_{23}	A		N		0,064			
		FFD_{23}	A		N		0,020	0,818	1,156	1,413
		$FFRL_{14}$	A		N		0,550			
		ZCR_t	A		C		0,445			
		$FFRL_{14}$	A		N		0,717			
		$Skurt_f$	A		C		0,060			
		NE_1	A		N		0,712			
		FFD_{12}	A		N		0,587			
		$MFCC_{10}$	A		F		0,216			
		NE_2	A		N		0,807			
		FFR_{12}	A		N		0,960			
	#3	FFR_{24}	A		N		0,318			
		FFL_3	A		N		0,968			
		FFL_1	A		N		0,453			
		FFL_4	A		N		0,312			
		FF_2	A		N		0,619			
		$MFCC_2$	A		N		0,215			
		NE_0	A		N		0,274			
		$MFCC_{11}$	A		N		0,384			
		NE_6	A		N		0,874			
		$FFRL_{23}$	A		N		0,052			
		NE_5	A		N		0,513			
		NE_3	A		N		0,284			
		$MFCC_8$	A		N		0,178			
		$MFCC_5$	A		N		0,235			
		P^{2D}	V	2D	J	R	0,417			
		A^{2D}	V	2D	J	R	0,352			
		D^{3D}_{XY}	V	3D	J	L	0,085			
	#5	DS^{2D}	V	2D	N	R	0,145			
		A^{3D}	V	3D	N	L	0,741			
		V^{3D}	V	3D	N	L	0,464			
		A^{2D}	V	2D	N	L	0,501			
	#8	D^{3D}_{XZ}	V	3D	N	R	0,967			

Kontynuacja tabeli na następnej stronie

		Cecha	**Dane**	**Rodzaj**	**2D/3D**	**Kamera**	BF p	σ^2_{12}	σ^2_{13}	σ^2_{23}
Głoska /z/	**#3**	$Skurt_f$	A		N		0,639			
		NPF	A		N		0,863			
		FFD_{12}	A		N		0,305			
		FFR_{12}	A		N		0,731			
		$FFRL_{14}$	A		N		0,202			
		NE_2	A		N		0,496			
		FFR_{23}	A		N		0,986			
		$FFRL_{13}$	A		N		0,961			
		$MFCC_{10}$	A		F		0,370			
		FF_2	A		N		0,494			
		NE_1	A		N		0,337			
		$FFRL_{23}$	A		N		0,367			
		$MFCC_8$	A		F		0,266			
		$FFRL_{24}$	A		N		0,623			
		ZCR_t	A		N		0,226			
		FFD_{23}	A		N		0,676			
		$MFCC_2$	A		F		0,634			
		DS^{3D}	V	3D	J	L	0,598			
		$Sfla_f$	A		F		0,071			
		NE_6	A		N		0,003	0,636	0,650	1,021
		$MFCC_{11}$	A		F		0,498			
		FFR_{24}	A		N		0,571			
		NE_7	A		N		0,024	0,728	0,536	0,735
	#5	D^{2D}_{Feret}	V	2D	J	R	0,133			
		C^{3D}_1	V	3D	J	L	0,406			
		C^{3D}_2	V	3D	J	L	0,470			
		S^{3D}	V	3D	J	L	0,330			
		Ax^{2D}_{major}	V	2D	J	R	0,330			
		D^{3D}_{YZ}	V	3D	J	L	0,517			
		DS^{3D}	V	3D	J	L	0,530			
		Ax^{3D}_{minor}	V	3D	J	L	0,746			
		Cor^{GLCM}_{2D}	V	2D	T	L	0,811			
		Ax^{3D}_{least}	V	3D	J	R	0,510			

Kontynuacja tabeli na następnej stronie

		Cecha	Dane	Rodzaj	2D/3D	Kamera	BF p	σ^2_{12}	σ^2_{13}	σ^2_{23}
Głoska /tʂ/		NE_1	A		N		0,712			
		NE_2	A		N		0,753			
		$Skurt_f$	A		F		0,083			
		$FFRL_{13}$	A		N		0,246			
		$FFRL_{14}$	A		N		0,230			
		FFR_{24}	A		N		0,619			
		NE_0	A		N		0,381			
		FFL_1	A		N		0,273			
		$MFCC_1$	A		F		0,284			
		FFD_2	A		N		0,663			
		FFR_{23}	A		N		0,786			
		FFL_3	A		N		0,644			
		ZCR_t	A		C		0,561			
	#3	NPF	A		N		0,344			
		$MFCC_0$	A		F		0,030	0,760	0,724	0,952
		FFR_{12}	A		N		0,394			
		Bus^{NGTDM}_{3D}	V	3D	T	L	0,981			
		$LRHGE^{GLRLM}_{2D}$	V	2D	T	L	0,144			
		Bus^{NGTDM}_{3D}	V	3D	T	R	0,441			
		$LRHGE^{GLRLM}_{2D}$	V	2D	T	R	0,154			
		Ax^{3D}_{least}	V	3D	U	L	0,925			
		GLV^{GLSZM}_{2D}	V	2D	T	L	0,526			
		H^{2D}_h	V	2D	T	L	0,276			
		FF_2	A		N		0,648			
		NE_6	A		N		0,340			
		GLV^{GLSZM}_{2D}	V	2D	T	R	0,952			
		Ax^{3D}_{least}	V	3D	W	L	0,881			
	#5	A^{2D}	V	2D	J	R	0,902			
	#7	Ax^{3D}_{major}	V	3D	J	R	0,411			

Kontynuacja tabeli na następnej stronie

		Cecha	**Dane**	**Rodzaj**	**2D/3D**	**Kamera**	BF p	σ^2_{12}	σ^2_{13}	σ^2_{23}
Głoska /dz/	**#1**	Ax^{3D}_{least}	V	3D	J	R	0,760			
		Ax^{3D}_{max}	V	3D	J	R	0,301			
		Ax^{3D}_{row}	V	3D	J	R	0,956			
		V^{3D}	V	3D	J	R	0,820			
		Ax^{3D}_{minor}	V	3D	J	R	0,241			
		Ax^{3D}_{major}	V	3D	J	R	0,224			
		D^{3D}_{YZ}	V	3D	J	R	0,211			
		D^{3D}_{XY}	V	3D	J	R	0,248			
	#4	Ax^{2D}_{major}	V	2D	J	R	0,343			
	#3	D^{3D}_{XZ}	V	3D	J	R	0,286			
		D^{3D}_{XY}	V	3D	J	R	0,640			
		Ax^{3D}_{least}	V	3D	J	L	0,112			
		FFR_{23}	A		N		0,966			
	#5	$MFFC_4$	A		F		0,401			
		NE_3	A		N		0,511			
		NE_4	A		N		0,380			
		FFL_2	A		N		0,190			
		$MFFC_3$	A		F		0,121			
		NE_7	A		N		0,095			
		NE_6	A		N		0,254			
		$MFFC_5$	A		F		0,100			
		ZCR_t	A		C		0,508			
		NE_5	A		N		0,421			
		NE_2	A		N		0,487			
		NE_8	A		N		0,072			

Tab. A.5: Wyniki analizy jednorodności wariancji dla głosek szeregu ciszącego testem Browna-Forsythe'a. Dla cech, w których odrzucono hipotezę zerową, dodatkowo przedstawiono stosunki wariancji pomiędzy grupami.

		Cecha	Dane	Rodzaj	2D/3D	Kamera	BF p	σ^2_{12}	σ^2_{13}	σ^2_{23}
/Głoska ɕ/	#1	E^{2D}	V	2D	J	L	0,521			
	#3	D^{3D}_{XZ}	V	3D	J	L	0,567			
		V^{3D}	V	3D	J	L	0,810			
		A^{3D}	V	3D	J	R	0,253			
		D^{3D}_{XZ}	V	3D	J	R	0,514			
		D^{3D}_{XY}	V	3D	J	L	0,780			
		D^{3D}_{XY}	V	3D	J	R	0,855			
		Ax^{3D}_{major}	V	3D	J	R	0,305			
		A^{3D}	V	3D	J	L	0,576			
		S^{3D}	V	3D	J	R	0,825			
		C^{3D}_{1}	V	3D	J	R	0,774			
		C^{3D}_{2}	V	3D	J	R	0,687			
		V^{3D}	V	3D	J	R	0,272			
		D^{3D}_{Feret}	V	3D	J	R	0,244			
	#5	D^{3D}_{Feret}	V	3D	J	R	0,640			
		E^{3D}	V	3D	J	L	0,596			
		Ax^{3D}_{least}	V	3D	J	R	0,332			
		Ax^{3D}_{minor}	V	3D	J	R	0,612			
		Ax^{3D}_{minor}	V	3D	J	L	0,092			
		D^{3D}_{YZ}	V	3D	J	L	0,466			
		V^{3D}	V	3D	J	R	0,755			
	#8	NE_9	A		N		0,274			
		A^{3D}	V	3D	J	L	0,305			
Głoska /ʑ/	#5	E^{2D}	V	2D	J	R	0,488			
		Ax^{2D}_{minor}	V	2D	J	L	0,623			
		A^{3D}	V	3D	J	L	0,576			
		SVR^{3D}	V	3D	W	L	0,507			
		$MFCC_4$	A		F		0,142			
		ZCR_t	A		C		0,067			
	#7	E^{2D}	V	2D	J	L	0,543			
	#8	D^{3D}_{Feret}	V	3D	J	L	0,526			
		D^{3D}_{XZ}	V	3D	J	L	0,819			
		Ax^{3D}_{major}	V	3D	J	L	0,172			
		D^{3D}_{XY}	V	3D	J	L	0,810			

Kontynuacja tabeli na następnej stronie

		Cecha	**Dane**	**Rodzaj**	**2D/3D**	**Kamera**	BF p	σ^2_{12}	σ^2_{13}	σ^2_{23}
Głoska /ʧ/	**#1**	A^{2D}	V	2D	J	L	0,472			
		DS^{2D}	V	2D	J	L	0,210			
		P^{2D}	V	2D	J	L	0,116			
		C_2^{2D}	V	2D	J	L	0,271			
		D_{Feret}^{2D}	V	2D	J	L	0,479			
	#5	E^{2D}	V	2D	J	R	0,427			
	#8	Ax_{minor}^{2D}	V	3D	J	L	0,234			
		$MFCC_4$	A		F		0,328			
		NCC_6	A		N		0,618			
/ʤ/	**#1**	DS_{2D}	V	2D	J	L	0,527			

Dodatek B: Wyniki testowania jednorodności rozkładów

Dodatek zawiera szczegółowe wyniki testów przeprowadzonych w ramach analizy jednorodności rozkładów. Tab. B.2–B.9 prezentują rezultaty testów U Manna-Whitneya i Kruskala-Wallisa (w tym przypadku również analizy post hoc) dla głosek z szeregu syczącego. Tab. B.10–B.14 przedstawiają wyniki tej samej analizy dla szeregu szumiącego, a tab. B.15–B.21 dla fonemów szeregu ciszącego.

Sekcja B.4 zawiera wyniki analizy liczebności cech wizualnych i akustycznych z podziałem na różne poziomy wielkości efektu (na podstawie wyników testów U Manna-Whitneya i Kruskala-Wallisa) dla każdego z rodzajów parametrów: kształtu 2D ust, warg i języka (tab. B.22–B.24), kształtu 3D (tab. B.25–B.27), teksturowych 2D i 3D (tab. B.28-B.29) oraz akustycznych (tab. B.30).

Tab. B.1: Legenda oznaczeń rodzajów cech w tabelach dodatku B.

Cechy obrazowe	W	kształtu warg
	U	kształtu ust
	J	kształtu języka
	T	teksturowe
Cechy akustyczne	C	czasowe
	F	częstotliwościowe
	N	szumowe

B.1 Porównanie rozkładów: szereg syczący

B.1.1 Głoska /s/

Tab. B.2: Wyniki testu U Manna-Whitneya dla wybranych cech artykulacyjnych w przypadku głoski /s/.

	#1 dentalizacja	(1) norma, (2) dysdentalizacja pionowa						
	#2 dentalność	(1) norma, (2) międzyzębowość						
	Cecha	**Dane**	**Rodzaj**	**2D/3D**	**Kamera**	**Test U M-W**		
						p	**U**	**WE**
#1	Ax^{3D}_{minor}	V	J	3D	L	<0,001	2084,5	0,473
	Ax^{3D}_{least}	V	J	3D	L	<0,001	2105,0	0,470
	D^{3D}_{YZ}	V	J	3D	L	<0,001	2103,5	0,452
	Ax^{3D}_{least}	V	J	3D	R	<0,001	1977,5	0,452
	Ax^{3D}_{minor}	V	J	3D	R	<0,001	1976,5	0,433
	D^{3D}_{YZ}	V	J	3D	R	<0,001	1985,0	0,425
	V^{3D}	V	J	3D	R	<0,001	1985,0	0,421
	V^{3D}	V	J	3D	L	<0,001	2136,0	0,412
#2	V^{3D}	V	J	3D	L	<0,001	1569,5	0,543
	Ax^{3D}_{least}	V	J	3D	L	<0,001	1595,0	0,524
	Ax^{3D}_{minor}	V	J	3D	L	<0,001	1602,0	0,501
	D^{3D}_{XY}	V	J	3D	L	<0,001	1606,0	0,494
	D^{3D}_{Feret}	V	J	3D	L	<0,001	1607,0	0,493
	D^{3D}_{YZ}	V	J	3D	L	<0,001	1611,5	0,489
	D^{3D}_{XZ}	V	J	3D	L	<0,001	1619,5	0,476
	Ax^{3D}_{major}	V	J	3D	L	<0,001	1642,0	0,446
	D^{2D}_{Feret}	V	J	2D	R	0,003	611,5	0,439
	A^{2D}	V	J	2D	R	0,004	614,5	0,429
	Ax^{3D}_{least}	V	J	3D	R	<0,001	1593,0	0,427
	D^{2D}_{Feret}	V	J	3D	L	0,003	580,0	0,426
	Ax^{3D}_{major}	V	J	3D	R	0,004	616,0	0,424
	SVR^{3D}	V	J	3D	L	<0,001	2289,0	-0,418
	Ax^{2D}_{minor}	V	J	2D	L	0,004	583,5	0,416
	V^{3D}	V	J	3D	R	<0,001	1597,0	0,411

Tab. B.3: Wynik testu Kruskala-Wallisa i analizy post hoc dla wybranych cech artykulacyjnych w przypadku głoski /s/.

#6 medialność wypływu powietrza						(1) norma (2) lewostronna (3) prawostronna					
	Cecha	**Dane**	**Rodzaj**	**2D/3D**	**Kamera**	**Test K-W**			**Post hoc**		
						p	**H**	**WE**	**1-2**	**1-3**	**2-3**
#6	P^{2D}	V	J	2D	L	0,015	8,446	0,117	0,034	1,000	0,016

B.1.2 Głoska /z/

Tab. B.4: Wyniki testu U Manna-Whitneya dla wybranych cech artykulacyjnych w przypadku głoski /z/.

#1 dentalizacja **#2 dentalność**						(1) norma, (2) dysdentalizacja pionowa (1) norma, (2) międzyzębowość		
	Cecha	**Dane**	**Rodzaj**	**2D/3D**	**Kamera**	**Test U M-W**		
						p	**U**	**WE**
#1	V^{3D}	V	J	3D	R	0,005	683,0	0,408
#2	D^{3D}_{XZ}	V	J	3D	R	0,001	522,5	0,517
	V^{3D}	V	J	3D	R	0,002	533,0	0,479
	D^{3D}_{XY}	V	J	3D	R	0,004	544,5	0,438
	V^{3D}	V	J	3D	L	0,009	399,0	0,402

Tab. B.5: Wynik testu Kruskala-Wallisa i analizy post hoc dla wybranych cech artykulacyjnych w przypadku głoski /z/.

	#5 skrócenie wędzidełka języka					(1) norma (2) nieznacznie (3) średnio					
	#7 medialność języka					(1) norma (2) dysmedialność lewostronna (3) dysmedialność prawostronna					
	Cecha	**Dane**	**Rodzaj**	**2D/3D**	**Kamera**	**Test K-W**			**Post hoc**		
						p	**H**	**WE**	**1-2**	**1-3**	**2-3**
#5	DS^{2D}	V	W	2D	R	0,001	13,852	0,085	0,459	0,001	0,049
	A^{2D}	V	W	2D	R	0,003	11,568	0,069	0,680	0,004	0,069
	I_h^{3D}	V	T	3D	R	0,002	12,417	0,067	0,046	1,0	0,002
#7	P^{2D}	V	J	2D	L	0,045	6,184	0,116	1,0	0,039	0,414
	A^{2D}	V	J	2D	L	0,047	6,126	0,115	1,0	0,040	0,276

B.1.3 Głoska /ts/

Tab. B.6: Wyniki testu U Manna-Whitneya dla wybranych cech artykulacyjnych w przypadku głoski /ts/.

	#1 dentalizacja					(1) norma, (2) dysdentalizacja pionowa		
	#2 dentalność					(1) norma, (2) międzyzębowość		
	#4 apikalność					(1) norma, (2) dorsalność		
	Cecha	**Dane**	**Rodzaj**	**2D/3D**	**Kamera**	**Test U M-W**		
						p	**U**	**WE**
#1	Ax^{3D}_{major}	V	J	3D	L	0,001	1167,0	0,415
#2	Ax^{3D}_{major}	V	J	3D	L	< 0,001	907,0	0,570
	A^{3D}	V	J	3D	L	< 0,001	915,0	0,555
	D^{3D}_{XZ}	V	J	3D	L	< 0,001	917,0	0,550
	D^{3D}_{XY}	V	J	3D	L	< 0,001	919,5	0,545
	V^{3D}	V	J	3D	L	< 0,001	927,5	0,529
	D^{3D}_{Feret}	V	J	3D	L	< 0,001	928,5	0,527
	DS^{3D}	V	J	3D	L	< 0,001	961,0	0,461
	Ax^{3D}_{least}	V	J	3D	L	< 0,001	977,0	0,455
	Ax^{3D}_{minor}	V	J	3D	L	0,001	969,0	0,450
	Ax^{3D}_{minor}	V	J	3D	L	0,002	711,0	0,435
	V^{3D}	V	J	3D	R	0,002	712,0	0,432
	D^{3D}_{YZ}	V	J	3D	R	0,004	722,0	0,407
	A^{3D}	V	J	3D	R	0,004	722,0	0,406
	D^{3D}_{YZ}	V	J	3D	L	0,002	992,5	0,404
#4	V^{3D}	V	J	3D	L	< 0,001	1787,0	-0,543
	Ax^{3D}_{least}	V	J	3D	L	< 0,001	1743,0	-0,529
	D^{3D}_{YZ}	V	J	3D	L	< 0,001	1764,5	-0,523
	D^{3D}_{XZ}	V	J	3D	L	< 0,001	1763,5	-0,515
	D^{3D}_{XY}	V	J	3D	L	< 0,001	1752,5	-0,502
	D^{3D}_{Feret}	V	J	3D	L	< 0,001	1741,0	-0,489
	Ax^{3D}_{minor}	V	J	3D	L	< 0,001	1735,0	-0,486
	Ax^{3D}_{major}	V	J	3D	L	< 0,001	1736,0	-0,483
	A^{3D}	V	J	3D	L	< 0,001	1735,0	-0,482

Tab. B.7: Wynik testu Kruskala-Wallisa i analizy post hoc dla wybranych cech artykulacyjnych w przypadku głoski /ts/.

#5 skrócenie wędzidełka języka
(1) norma
(2) nieznacznie
(3) średnio

#7 medialność języka
(1) norma
(2) dysmedialność lewostronna
(3) dysmedialność prawostronna

	Cecha	Dane	Rodzaj	2D/3D	Kamera	Test K-W			Post hoc		
						p	H	WE	1-2	1-3	2-3
#5	E^{2D}	V	J	2D	R	0,025	7,351	0,162	0,097	0,041	1,000
	S^{3D}	V	J	3D	L	0,005	10,525	0,117	0,005	1,000	0,138
	C_1^{3D}	V	J	3D	L	0,006	10,300	0,114	0,007	1,000	0,125
	C_2^{3D}	V	J	3D	L	0,007	9,981	0,109	0,006	1,000	0,092
	DS^{3D}	V	J	3D	L	0,011	9,080	0,097	0,008	0,801	0,345
#7	D_{YZ}^{3D}	V	J	3D	L	0,007	9,842	0,106	0,283	0,073	0,007
	Ax_{minor}^{3D}	V	J	3D	L	0,008	9,749	0,105	0,679	0,031	0,013
	A^{3D}	V	J	3D	L	0,011	8,983	0,094	0,283	0,110	0,010
	Ax_{minor}^{3D}	V	J	3D	R	0,035	6,732	0,078	0,088	0,213	1,000
	E^{3D}	V	J	3D	R	0,035	6,692	0,077	0,029	1,000	0,601
	V^{3D}	V	J	3D	L	0,029	7,110	0,069	0,726	0,106	0,037

B.1.4 Głoska /dz/

Tab. B.8: Wyniki testu U Manna-Whitneya dla wybranych cech artykulacyjnych w przypadku głoski /dz/.

	#1 dentalizacja				(1) norma, (2) dysdentalizacja pionowa			
	#2 dentalność				(1) norma, (2) międzyzębowość			
	#4 apikalność				(1) norma, (2) dorsalność			
	Cecha	**Dane**	**Rodzaj**	**2D/3D**	**Kamera**	**Test U M-W**		
						p	**U**	**WE**
#1	A^{2D}	V	J	2D	R	0,019	170,0	0,508
	DS^{2D}	V	J	2D	R	0,019	170,0	0,508
	DS^{2D}	V	J	2D	L	0,021	157,0	0,499
	S^{2D}	V	J	2D	R	0,023	237,0	-0,493
	P^{2D}	V	J	2D	L	0,025	158,0	0,485
	S^{2D}	V	J	2D	L	0,025	226,0	-0,484
	A^{2D}	V	J	2D	L	0,035	160,0	0,456
#2	A^{2D}	V	J	2D	L	0,002	94,0	0,657
	A^{2D}	V	J	2D	R	0,003	107,0	0,648
	P^{2D}	V	J	2D	L	0,003	96,0	0,630
	DS^{2D}	V	J	2D	L	0,004	97,0	0,617
	V^{3D}	V	J	3D	R	<0,001	322,5	0,598
	Ax^{2D}_{minor}	V	J	2D	R	0,006	111,0	0,595
	DS^{2D}	V	J	2D	R	0,007	112,0	0,582
	A^{3D}	V	J	3D	R	0,001	326,0	0,580
	D^{3D}_{XZ}	V	J	3D	R	0,001	327,0	0,575
	S^{2D}	V	J	2D	R	0,008	199,0	-0,569
	Ax^{3D}_{major}	V	J	3D	R	0,001	334,0	0,538
	D^{3D}_{Feret}	V	J	3D	R	0,001	335,0	0,533
	D^{3D}_{XY}	V	J	3D	R	0,002	338,0	0,517
	DS^{3D}	V	J	3D	R	0,004	344,0	0,486
	D^{2D}_{Feret}	V	J	2D	R	0,026	119,5	0,483
	Ax^{3D}_{minor}	V	J	3D	R	0,004	346,0	0,477
	D^{3D}_{YZ}	V	J	3D	R	0,005	346,5	0,476
	S^{3D}	V	J	3D	R	0,005	528,0	-0,475
	C^{3D}_{1}	V	J	3D	R	0,005	528,0	-0,475
	C^{3D}_{2}	V	J	3D	R	0,005	528,0	-0,475
	S^{2D}	V	J	2D	L	0,029	180,0	-0,472
	D^{2D}_{Feret}	V	J	2D	L	0,034	109,0	0,459

Kontynuacja tabeli na następnej stronie

	Cecha	Dane	Rodzaj	2D/3D	Kamera	Test U M-W		
						p	U	WE
	Ax^{3D}_{least}	V	J	3D	R	0,008	357,0	0,444
	V^{3D}	V	J	3D	L	0,013	298,0	0,417
#4	DS^{2D}	V	J	2D	R	0,004	277,0	-0,553
	A^{2D}	V	J	2D	R	0,007	148,0	-0,511
	S^{2D}	V	J	2D	R	0,007	148,0	0,511
	E^{2D}	V	J	2D	L	0,027	140,0	0,430

Tab. B.9: Wynik testu Kruskala-Wallisa i analizy post hoc dla wybranych cech artykulacyjnych w przypadku głoski /dz/.

#6 medialność wypływu powietrza	(1) norma (2) lewostronna (3) prawostronna
#7 medialność języka	(1) norma (2) dysmedialność lewostronna (3) dysmedialność prawostronna
#8 medialność żuchwy	(1) norma (2) dysmedialność lewostronna (3) dysmedialność prawostronna

	Cecha	Dane	Rodzaj	2D/3D	Kamera	Test K-W			Post hoc		
						p	H	WE	1-2	1-3	2-3
#6	D^{2D}_{Feret}	V	J	3D	L	0,033	6,852	0,194	0,042	1,000	0,054
	Ax^{2D}_{major}	V	J	2D	L	0,039	6,463	0,179	0,033	1,000	0,206
#7	A^{3D}	V	J	3D	R	0,010	9,216	0,154	0,486	0,072	0,008
	DS^{3D}	V	J	3D	R	0,019	7,918	0,126	0,913	0,068	0,021
	Ax^{3D}_{least}	V	J	3D	R	0,039	6,481	0,095	1,000	0,085	0,054
#8	D^{2D}_{Feret}	V	J	2D	L	0,039	6,505	0,188	0,297	0,283	0,041
	A^{3D}	V	J	3D	L	0,038	6,538	0,106	0,416	0,227	0,033
	NCC_{10}	A	N			0,002	12,234	0,066	0,011	0,048	1,000

B.2 Porównanie rozkładów: szereg szumiący

B.2.1 Głoska /ṣ/

Tab. B.10: Wynik testu Kruskala-Wallisa i analizy post hoc dla wybranych cech artykulacyjnych w przypadku głoski /ṣ/.

#3 postdentalność	(1) norma (2) zadziąsłowość (3) zębowość
#5 skrócenie wędzidełka języka	(1) norma (2) nieznacznie (3) znacznie
#8 medialność żuchwy	(1) norma (2) dysmedialność lewostronna (3) dysmedialność prawostronna

	Cecha	Dane	Rodzaj	2D/3D	Kamera	Test K-W			Post hoc		
						p	H	WE	1-2	1-3	2-3
#3	NPF	A	N			<0,001	35,575	0,211	0,014	<0,001	<0,001
	FFR_{23}	A	N			<0,001	34,603	0,205	1,000	<0,001	<0,001
	FFD_{23}	A	N			<0,001	28,196	0,165	1,000	<0,001	<0,001
	$FFRL_{14}$	A	N			<0,001	27,991	0,163	0,036	<0,001	<0,001
	ZCR_t	A	C			<0,001	27,628	0,161	0,038	<0,001	<0,001
	$FFRL_{14}$	A	N			<0,001	27,571	0,161	0,012	<0,001	0,001
	$Skurt_f$	A	F			<0,001	25,331	0,147	0,859	0,004	<0,001
	NE_1	A	N			<0,001	24,294	0,140	1,000	<0,001	<0,001
	FFD_{12}	A	N			<0,001	23,844	0,137	0,619	0,010	<0,001
	$MFCC_{10}$	A	F			<0,001	21,496	0,123	0,112	0,004	0,149
	NE_2	A	N			<0,001	21,445	0,122	1,000	0,004	<0,001
	FFR_{12}	A	N			<0,001	21,426	0,122	0,346	0,041	<0,001
	FFR_{24}	A	N			<0,001	20,940	0,119	0,617	<0,001	<0,001
	FFL_3	A	N			<0,001	20,116	0,114	0,040	<0,001	0,005
	FFL_1	A	N			<0,001	19,272	0,109	0,313	<0,001	0,001
	FFL_4	A	N			<0,001	18,764	0,105	0,014	<0,001	0,026
	FF_2	A	N			<0,001	17,904	0,100	1,000	0,002	<0,001
	$MFCC_2$	A	F			<0,001	17,856	0,100	0,035	<0,001	0,016
	NE_0	A	N			<0,001	16,313	0,090	0,024	<0,001	0,047

Kontynuacja tabeli na następnej stronie

	Cecha	Dane	Rodzaj	2D/3D	Kamera	Test K-W			Post hoc		
						p	H	WE	1-2	1-3	2-3
	$MFCC_{11}$	A	F			0,001	15,033	0,082	1,000	0,006	<0,001
	NE_6	A	N			0,001	14,523	0,079	0,025	<0,001	0,097
	$FFRL_{23}$	A	N			0,001	14,148	0,076	1,000	0,007	0,001
	NE_5	A	N			0,001	14,124	0,076	0,091	0,001	0,032
	NE_3	A	N			0,001	13,477	0,072	1,000	0,004	<0,001
	$MFCC_8$	A	F			0,002	12,644	0,067	0,459	0,002	0,012
	$MFCC_5$	A	F			0,003	11,474	0,060	0,477	0,003	0,020
#5	P^{2D}	V	J	2D	R	0,016	8,284	0,114	0,786	0,015	0,167
	A^{2D}	V	J	2D	R	0,020	7,853	0,106	0,953	0,020	0,158
	D^{3D}_{XY}	V	J	3D	L	0,007	9,848	0,085	0,120	1,000	0,007
	DS^{2D}	V	J	2D	R	0,045	6,189	0,076	1,000	0,053	0,222
	A^{3D}	V	J	3D	L	0,013	8,710	0,073	0,342	1,000	0,010
	V^{3D}	V	J	3D	L	0,014	8,514	0,071	0,389	1,000	0,011
	A^{2D}	V	J	2D	L	0,042	6,331	0,070	1,000	0,058	0,187
#8	D^{3D}_{XZ}	V	J	3D	R	0,018	8,079	0,065	0,029	0,462	0,348

B.2.2 Głoska /z̨/

Tab. B.11: Wynik testu Kruskala-Wallisa i analizy post hoc dla wybranych cech artykulacyjnych w przypadku głoski /z̨/.

	#3 postdentalność					(1) norma (2) zadziąsłowość (3) zębowość					
	#5 skrócenie wędzidełka języka					(1) norma (2) nieznacznie (3) średnio					
	Cecha	**Dane**	**Rodzaj**	**2D/3D**	**Kamera**	**Test K-W**			**Post hoc**		
						p	**H**	**WE**	**1-2**	**1-3**	**2-3**
#3	$Skurt_f$	A	F			<0,001	27,964	0,175	1,000	0,001	<0,001
	NPF	A	N			<0,001	24,685	0,153	0,642	<0,001	<0,001
	FFD_{12}	A	N			<0,001	23,945	0,148	1,000	0,001	<0,001
	FFR_{12}	A	N			<0,001	22,460	0,138	0,989	0,006	<0,001
	$FFRL_{14}$	A	N			<0,001	21,011	0,128	0,438	<0,001	<0,001
	NE_2	A	N			<0,001	20,942	0,128	1,000	0,004	<0,001
	FFR_{23}	A	N			<0,001	20,762	0,127	0,892	<0,001	<0,001
	$FFRL_{13}$	A	N			<0,001	18,805	0,114	0,722	<0,001	<0,001
	$MFCC_{10}$	A	F			<0,001	17,744	0,106	1,000	0,001	<0,001
	FF_2	A	N			<0,001	16,992	0,101	1,000	0,002	<0,001
	NE_1	A	N			<0,001	16,877	0,101	1,000	0,002	<0,001
	$FFRL_{23}$	A	N			<0,001	16,093	0,095	0,867	0,001	0,001
	$MFCC_8$	A	F			<0,001	15,249	0,090	1,000	0,008	<0,001
	$FFRL_{24}$	A	N			0,001	14,841	0,087	1,000	0,002	0,001
	ZCR_t	A	C			0,001	14,825	0,087	0,185	<0,001	0,012
	FFD_{23}	A	N			0,001	14,628	0,085	0,584	0,001	0,004
	$MFCC_2$	A	F			0,001	13,765	0,079	0,405	0,001	0,008
	DS^{3D}	V	J	3D	L	0,042	6,327	0,073	0,710	0,039	0,153
	$Sfla_f$	A	F			0,002	12,787	0,073	0,357	0,001	0,015
	NE_6	A	N			0,002	12,054	0,068	0,118	0,002	0,068
	$MFCC_{11}$	A	F			0,003	11,926	0,067	1,000	0,025	0,002
	$FFR_{24]}$	A	N			0,003	11,575	0,065	1,000	0,006	0,006
	NE_7	A	N			0,004	11,259	0,063	0,061	0,003	0,188

Kontynuacja tabeli na następnej stronie

	Cecha	Dane	Rodzaj	2D/3D	Kamera	Test K-W			Post hoc		
						p	H	WE	1-2	1-3	2-3
#5	D^{2D}_{Feret}	V	J	2D	R	0,022	7,603	0,119	0,112	0,038	1,000
	C^{3D}_{1}	V	J	3D	L	0,008	9,617	0,119	1,000	0,038	0,015
	C^{3D}_{2}	V	J	3D	L	0,008	9,553	0,118	1,000	0,036	0,017
	S^{3D}	V	J	3D	L	0,009	9,365	0,115	1,000	0,044	0,016
	Ax^{2D}_{major}	V	J	2D	R	0,028	7,149	0,110	0,165	0,040	1,000
	D^{3D}_{YZ}	V	J	3D	L	0,022	7,644	0,088	0,017	0,293	0,718
	DS^{3D}	V	J	3D	L	0,031	6,923	0,077	1,000	0,212	0,034
	Ax^{3D}_{minor}	V	J	3D	L	0,038	6,528	0,071	0,032	0,313	0,981
	Cor^{GLCM}_{2D}	V	T	2D	L	0,003	11,716	0,069	1,000	0,009	0,014
	Ax^{3D}_{least}	V	J	3D	R	0,045	6,191	0,065	0,052	1,000	0,336

B.2.3 Głoska /tṣ/

Tab. B.12: Wynik testu Kruskala-Wallisa i analizy post hoc dla wybranych cech artykulacyjnych w przypadku głoski /tṣ/.

#3 postdentalność	(1) norma (2) zadziąsłowość (3) zębowość
#5 skrócenie wędzidełka języka	(1) norma (2) nieznacznie (3) średnio
#7 medialność języka	(1) norma (2) dysmedialność lewostronna (3) dysmedialność prawostronna

	Cecha	Dane	Rodzaj	2D/3D	Kamera	Test K-W			Post hoc		
						p	H	WE	1-2	1-3	2-3
#3	NE_1	A		N		<0,001	30,769	0,201	1,000	<0,001	<0,001
	NE_2	A		N		<0,001	26,122	0,169	0,903	<0,001	<0,001
	$Skurt_f$	A		F		<0,001	24,782	0,159	0,157	<0,001	<0,001
	$FFRL_{13}$	A		N		<0,001	22,430	0,143	1,000	0,001	<0,001
	$FFRL_{14}$	A		N		<0,001	21,179	0,134	0,719	<0,001	<0,001
	FFR_{24}	A		N		<0,001	20,970	0,133	1,000	0,002	<0,001
	NE_0	A		N		<0,001	16,776	0,103	0,109	<0,001	0,007
	FFL_1	A		N		<0,001	16,686	0,103	1,000	0,001	0,001
	$MFCC_1$	A		F		<0,001	16,398	0,101	0,217	<0,001	0,004
	FFD_2	A		N		0,001	15,142	0,092	1,000	0,043	<0,001
	FFR_{23}	A		N		0,001	14,982	0,091	1,000	0,012	<0,001
	FFL_3	A		N		0,001	14,822	0,090	1,000	0,047	<0,001
	ZCR_t	A		T		0,001	14,036	0,084	1,000	0,014	0,001
	NPF	A		N		0,001	13,912	0,083	1,000	0,034	0,001
	$MFCC_0$	A		F		0,001	13,442	0,080	0,888	0,002	0,004
	FFR_{12}	A		N		0,001	13,134	0,078	1,000	0,102	0,001
	Bus_{3D}^{NGTDM}	V	3D	T	L	0,002	12,972	0,077	0,005	1,000	0,055
	$LRHGE_{2D}^{GLRLM}$	V	2D	T	L	0,003	11,436	0,074	0,053	1,000	0,014
	Bus_{3D}^{NGTDM}	V	3D	T	R	0,003	11,600	0,068	0,009	1,000	0,077
	$LRHGE_{2D}^{GLRLM}$	V	2D	T	R	0,005	10,412	0,066	0,036	1,000	0,039
	Ax_{least}^{3D}	V	3D	U	L	0,004	11,291	0,065	0,067	1,000	0,012
	GLV_{2D}^{GLSZM}	V	2D	T	L	0,007	9,947	0,063	0,154	1,000	0,015

Kontynuacja tabeli na następnej stronie

	Cecha	Dane	Rodzaj	2D/3D	Kamera	Test K-W			Post hoc		
						p	H	WE	1-2	1-3	2-3
	H_h	V	2D	T	L	0,009	9,423	0,062	0,018	1,000	0,180
	FF_2	A		N		0,004	10,902	0,062	1,000	0,061	0,003
	NE_6	A		N		0,005	10,797	0,062	1,000	0,092	0,003
	GLV_{2D}^{GLSZM}	V	2D	T	R	0,008	9,758	0,061	0,121	1,000	0,020
	Ax_{least}^{3D}	V	3D	W	L	0,005	10,634	0,061	0,082	1,000	0,015
#5	A^{2D}	V	2D	J	R	0,050	6,003	0,108	0,045	1,000	0,609
#7	Ax_{major}^{3D}	V	3D	J	R	0,042	6,317	0,071	1	0,036	0,187

B.2.4 Głoska /dz̨/

Tab. B.13: Wyniki testu U Manna-Whitneya dla wybranych cech artykulacyjnych w przypadku głoski /dz̨/.

	#1 dentalizacja (1) norma, (2) dysdentalizacja pionowa **#4 apikalność** (1) norma, (2) dorsalność							
	Cecha	Dane	Rodzaj	2D/3D	Kamera	Test U M-W		
						p	U	WE
#1	Ax_{least}^{3D}	V	J	3D	R	0,001	349,0	0,593
	D_{Feret}^{3D}	V	J	3D	R	0,001	348,0	0,572
	D_{XZ}^{3D}	V	J	3D	R	0,002	350,0	0,560
	V^{3D}	V	J	3D	R	0,002	352,0	0,543
	Ax_{minor}^{3D}	V	J	3D	R	0,003	353,0	0,537
	Ax_{major}^{3D}	V	J	3D	R	0,003	353,0	0,535
	D_{YZ}^{3D}	V	J	3D	R	0,003	354,0	0,528
	D_{XY}^{3D}	V	J	3D	R	0,007	360,0	0,484
#4	Ax_{major}^{2D}	V	J	2D	R	0,046	233,0	-0,433

Tab. B.14: Wynik testu Kruskala-Wallisa i analizy post hoc dla wybranych cech artykulacyjnych w przypadku głoski /dz̨/.

#3 postdentalność
(1) norma
(2) zadziąsłowość
(3) zębowość

#5 skrócenie wędzidełka języka
(1) norma
(2) nieznacznie
(3) średnio

	Cecha	Dane	Rodzaj	2D/3D	Kamera	Test K-W			Post hoc		
						p	H	WE	1-2	1-3	2-3
#3	D_{XZ}^{3D}	V	J	3D	R	0,024	7,422	0,194	0,032	0,068	1,000
	D_{XY}^{3D}	V	J	3D	R	0,040	6,442	0,159	0,035	0,271	1,000
	Ax_{least}^{3D}	V	J	3D	L	0,042	6,352	0,145	0,185	1,000	0,045
	FFR_{23}	A	N			0,004	10,860	0,065	0,785	0,006	0,016
#5	$MFFC_4$	A	F			<0,001	15,988	0,095	0,068	<0,001	0,196
	NE_3	A	N			0,001	13,721	0,080	0,020	0,001	1,000
	NE_4	A	N			0,001	13,696	0,080	0,022	0,001	0,942
	FFL_2	A	N			0,002	12,962	0,075	0,017	0,002	1,000
	$MFFC_3$	A	F			0,002	12,848	0,074	0,089	0,001	0,397
	NE_7	A	N			0,002	12,228	0,070	0,074	0,001	0,557
	NE_6	A	N			0,002	12,103	0,069	0,045	0,002	0,863
	$MFFC_5$	A	F			0,003	11,905	0,067	0,155	0,002	0,321
	ZCR_t	A	C			0,003	11,774	0,066	0,151	0,002	0,343
	NE_5	A	N			0,004	11,159	0,062	0,058	0,003	0,927
	NE_2	A	N			0,004	11,071	0,062	0,015	0,006	1,000
	NE_8	A	N			0,004	10,960	0,061	0,070	0,003	0,847

B.3 Porównanie rozkładów: szereg ciszący

B.3.1 Głoska /ɕ/

Tab. B.15: Wyniki testu U Manna-Whitneya dla wybranych cech artykulacyjnych w przypadku głoski /ɕ/.

	#1 dentalizacja (1) norma, (2) dysdentalizacja pionowa **#3 postdentalność** (1) norma, (2) międzyzębowość							
	Cecha	**Dane**	**Rodzaj**	**2D/3D**	**Kamera**	**Test U M-W**		
						p	**U**	**WE**
#1	E^{2D}	V	J	2D	L	0,015	515,5	-0,410
#3	D_{XZ}^{3D}	V	J	3D	L	<0,001	1549,0	0,480
	V^{3D}	V	J	3D	L	<0,001	1549,5	0,479
	A^{3D}	V	J	3D	R	<0,001	1902,5	0,462
	D_{XZ}^{3D}	V	J	3D	R	<0,001	1906,0	0,453
	D_{XY}^{3D}	V	J	3D	L	<0,001	1563,0	0,448
	D_{XY}^{3D}	V	J	3D	R	<0,001	1915,0	0,435
	Ax_{major}^{3D}	V	J	3D	R	0,001	1923,0	0,418
	A^{3D}	V	J	3D	L	0,001	1576,5	0,416
	S^{3D}	V	J	3D	R	0,001	2335,0	-0,414
	C_1^{3D}	V	J	3D	R	0,001	2335,0	-0,414
	C_2^{3D}	V	J	3D	R	0,001	2334,0	-0,412
	V^{3D}	V	J	3D	R	0,001	1926,5	0,411
	D_{Feret}^{3D}	V	J	3D	R	0,001	1927,0	0,410

Tab. B.16: Wynik testu Kruskala-Wallisa i analizy post hoc dla wybranych cech artykulacyjnych w przypadku głoski /ɕ/.

#5 skrócenie wędzidełka języka						(1) norma (2) nieznacznie (3) średnio					
#8 medialność żuchwy						(1) norma (2) dysmedialność lewostronna (3) dysmedialność prawostronna					
	Cecha	**Dane**	**Rodzaj**	**2D/3D**	**Kamera**	**Test K-W**			**Post hoc**		
						p	**H**	**WE**	**1-2**	**1-3**	**2-3**
#5	D^{3D}_{Feret}	V	J	3D	R	0,006	10,089	0,117	0,012	0,015	1
	E^{3D}	V	J	3D	L	0,020	7,844	0,093	0,024	0,074	1
	Ax^{3D}_{least}	V	J	3D	R	0,018	8,040	0,088	0,014	0,168	0,981
	Ax^{3D}_{minor}	V	J	3D	R	0,023	7,576	0,081	0,037	0,046	1
	Ax^{3D}_{minor}	V	J	3D	L	0,032	6,912	0,078	0,029	0,843	0,29
	D^{3D}_{YZ}	V	J	3D	L	0,037	6,574	0,073	0,032	0,666	0,426
	V^{3D}	V	J	3D	R	0,040	6,417	0,064	0,046	0,122	1
#8	NE_9	A	N			0,002	12,457	0,066	0,004	0,007	0,066
	A^{3D}	V	J	3D	L	0,046	6,169	0,065	0,546	0,042	0,065

B.3.2 Głoska /ʑ/

Tab. B.17: Wyniki testu U Manna-Whitneya dla wybranych cech artykulacyjnych w przypadku głoski /ʑ/.

#7 medialność języka						(1) norma, (2) dysmedialność prawostronna		
	Cecha	**Dane**	**Rodzaj**	**2D/3D**	**Kamera**	**Test U M-W**		
						p	**U**	**WE**
#7	E^{2D}	V	J	2D	L	0,024	521,5	-0,404

Tab. B.18: Wynik testu Kruskala-Wallisa i analizy post hoc dla wybranych cech artykulacyjnych w przypadku głoski /ʑ/.

#5 skrócenie wędzidełka języka	(1) norma (2) nieznacznie (3) średnio
#8 medialność żuchwy	(1) norma (2) dysmedialność lewostronna (3) dysmedialność prawostronna

	Cecha	Dane	Rodzaj	2D/3D	Kamera	Test K-W			Post hoc		
						p	H	WE	1-2	1-3	2-3
#5	E^{2D}	V	J	2D	R	0,012	8,917	0,256	0,462	0,145	0,009
	Ax^{2D}_{minor}	V	J	2D	R	0,038	6,543	0,168	0,208	0,856	0,062
	A^{3D}	V	J	3D	L	0,033	6,803	0,089	0,042	1,000	0,133
	SVR^{3D}	V	W	3D	L	0,005	10,745	0,062	1,000	0,047	0,007
	$MFCC_4$	A	F			0,005	10,465	0,060	0,174	0,004	0,491
	ZCR_t	A	C			0,005	10,442	0,060	0,063	0,004	1,000
#8	D^{3D}_{Feret}	V	J	3D	L	0,033	6,837	0,090	0,071	1,000	0,030
	D^{3D}_{XZ}	V	J	3D	L	0,039	6,509	0,084	0,047	1,000	0,054
	Ax^{3D}_{major}	V	J	3D	L	0,046	6,157	0,077	0,093	1,000	0,045
	D^{3D}_{XY}	V	J	3D	L	0,048	6,066	0,075	0,049	1,000	0,087

B.3.3 Głoska /tɕ/

Tab. B.19: Wyniki testu U Manna-Whitneya dla wybranych cech artykulacyjnych w przypadku głoski /tɕ/.

#1 dentalizacja	(1) norma, (2) dysdentalizacja pionowa							
	Cecha	Dane	Rodzaj	2D/3D	Kamera	Test U M-W		
						p	U	WE
#1	A^{2D}	V	J	2D	L	0,003	253,5	0,568
	DS^{2D}	V	J	2D	L	0,004	255,5	0,549
	P^{2D}	V	J	2D	L	0,005	256,5	0,540
	C^{2D}_2	V	J	2D	L	0,022	362,5	-0,439
	D^{2D}_{Feret}	V	J	2D	L	0,029	270,0	0,416

Tab. B.20: Wynik testu Kruskala-Wallisa i analizy post hoc dla wybranych cech artykulacyjnych w przypadku głoski /tɕ/.

#5 skrócenie wędzidełka języka						(1) norma (2) nieznaczny (3) średni					
#8 medialność żuchwy						(1) norma (2) dysmedialność lewostronna (3) dysmedialność prawostronna					
	Cecha	**Dane**	**Rodzaj**	**2D/3D**	**Kamera**	**Test K-W**			**Post hoc**		
						p	**H**	**WE**	**1-2**	**1-3**	**2-3**
#5	E^{2D}	V	J	2D	R	0,024	7,460	0,195	0,127	0,020	0,814
#8	Ax^{2D}_{minor}	V	J	3D	L	0,030	7,036	0,174	0,212	0,266	0,030
	$MFCC_4$	A	F			0,001	14,523	0,085	0,047	0,003	1,000

B.3.4 Głoska /dʑ/

Tab. B.21: Wyniki testu U Manna-Whitneya dla wybranych cech artykulacyjnych w przypadku głoski /dʑ/.

#1 dentalizacja					(1) norma, (2) dysdentalizacja pionowa			
	Cecha	**Dane**	**Rodzaj**	**2D/3D**	**Kamera**	**Test U M-W**		
						p	**U**	**WE**
#1	DS_{2D}	V	J	2D	L	0,033	87,0	0,497

B.4 Wyniki analizy liczebności cech wizualno-akustycznych z podziałem na różne poziomy wielkości efektu

Tab. B.22: Podsumowanie liczby cech kształtu ust (2D) o wartości p poniżej 0,05 z podziałem na różne stopnie wielkości efektu.

			#1 dentalizacja	#2 dentalność	#3 postdentalność	#4 apikalność	#5 skrócenie wędzidełka języka	#6 medialność wypływu powietrza	#7 medialność języka	#8 medialność żuchwy
			#1	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**
Szereg syczący										
/s/	**p<0,05**					4				
	WE	**mała**				4				
		średnia								
		wysoka								
/z/	**p<0,05**						10			
	WE	**mała**					10			
		średnia								
		wysoka								
/ts/	**p<0,05**					10		1		
	WE	**mała**				10		1		
		średnia								
		wysoka								
/dz/	**p<0,05**					4				
	WE	**mała**				4				
		średnia								
		wysoka								
Szereg szumiący										
/ʂ/	**p<0,05**		2			1				
	WE	**mała**	2			1				
		średnia								
		wysoka								
/ʐ/	**p<0,05**		3			3				
	WE	**mała**	3			3				
		średnia								
		wysoka								

Kontynuacja tabeli na następnej stronie

			#1	#2	#3	#4	#5	#6	#7	#8
/tʂ/		**p<0,05**								
	WE	**mała**								
		średnia								
		wysoka								
/dʐ/		**p<0,05**				4				
	WE	**mała**				4				
		średnia								
		wysoka								
Szereg ciszący										
/ɕ/		**p<0,05**								
	WE	**mała**								
		średnia								
		wysoka								
/ʑ/		**p<0,05**				2				
	WE	**mała**				2				
		średnia								
		wysoka								
/tɕ/		**p<0,05**	1			2				
	WE	**mała**	1			2				
		średnia								
		wysoka								
/dʑ/		**p<0,05**								
	WE	**mała**								
		średnia								
		wysoka								

Tab. B.23: Podsumowanie liczby cech kształtu warg (2D) o wartości p poniżej 0,05 z podziałem na różne stopnie wielkości efektu.

#1 dentalizacja
#2 dentalność
#3 postdentalność
#4 apikalność
#5 skrócenie wędzidełka języka
#6 medialność wypływu powietrza
#7 medialność języka
#8 medialność żuchwy

			#1	#2	#3	#4	#5	#6	#7	#8
Szereg syczący										
/s/	**p<0,05**					9	3			
	WE	**mała**				9	3			
		średnia								
		wysoka								
/z/	**p<0,05**						6			1
	WE	**mała**					4			1
		średnia					2			
		wysoka								
/ts/	**p<0,05**					8	2	2		
	WE	**mała**				8	2	2		
		średnia								
		wysoka								
/dz/	**p<0,05**					9	5			
	WE	**mała**				9	5			
		średnia								
		wysoka								
Szereg szumiący										
/ʂ/	**p<0,05**		2			1	1			
	WE	**mała**	2			1	1			
		średnia								
		wysoka								
/ʐ/	**p<0,05**		1			3				
	WE	**mała**	1			3				
		średnia								
		wysoka								
/tʂ/	**p<0,05**									
	WE	**mała**								
		średnia								
		wysoka								

Kontynuacja tabeli na następnej stronie

			#1	#2	#3	#4	#5	#6	#7	#8
/dʐ/		**p<0,05**				4				
	WE	**mała**				4				
		średnia								
		wysoka								
Szereg ciszący										
/ɕ/		**p<0,05**								
	WE	**mała**								
		średnia								
		wysoka								
/ʑ/		**p<0,05**				2				
	WE	**mała**				2				
		średnia								
		wysoka								
/tɕ/		**p<0,05**	1			2				
	WE	**mała**	1			2				
		średnia								
		wysoka								
/dʑ/		**p<0,05**								
	WE	**mała**								
		średnia								
		wysoka								

Tab. B.24: Podsumowanie liczby cech kształtu języka (2D) o wartości p poniżej 0,05 z podziałem na różne stopnie wielkości efektu.

			#1 dentalizacja	#2 dentalność	#3 postdentalność	#4 apikalność	#5 skrócenie wędzidełka języka	#6 medialność wypływu powietrza	#7 medialność języka	#8 medialność żuchwy
			#1	#2	#3	#4	#5	#6	#7	#8
Szereg syczący										
/s/		**p<0,05**	6	7		1		1		
	WE	**mała**	6	2		1				
		średnia		5				1		
		wysoka								

Kontynuacja tabeli na następnej stronie

			#1	#2	#3	#4	#5	#6	#7	#8
/z/	p<0,05			1					2	
	WE	mała		1						
		średnia							2	
		wysoka								
/ts/	p<0,05					1	1			
	WE	mała				1				
		średnia								
		wysoka					1			
/dz/	p<0,05		7	10		5		2		1
	WE	mała				1				
		średnia	7	6		4				
		wysoka		4				2		1
Szereg szumiący										
/ʂ/	p<0,05					1	4			
	WE	mała				1				
		średnia					4			
		wysoka								
/ʐ/	p<0,05		1				2			
	WE	mała	1							
		średnia					2			
		wysoka								
/tʂ/	p<0,05					2				
	WE	mała				2				
		średnia								
		wysoka								
/dʐ/	p<0,05					1				
	WE	mała								
		średnia				1				
		wysoka								
Szereg ciszący										
/ɕ/	p<0,05		1		5					
	WE	mała			5					
		średnia	1							
		wysoka								
/ʑ/	p<0,05					2	2	1	2	
	WE	mała				2		1	1	
		średnia							1	
		wysoka					2			

Kontynuacja tabeli na następnej stronie

			#1	#2	#3	#4	#5	#6	#7	#8
/tɕ/	p<0,05		4				1			1
	WE	mała								
		średnia	4							
		wysoka					1			1
/dʑ/	p<0,05		1							
	WE	mała								
		średnia	1							
		wysoka								

Tab. B.25: Podsumowanie liczby cech kształtu ust (3D) o wartości p poniżej 0,05 z podziałem na różne stopnie wielkości efektu.

#1 dentalizacja
#2 dentalność
#3 postdentalność
#4 apikalność
#5 skrócenie wędzidełka języka
#6 medialność wypływu powietrza
#7 medialność języka
#8 medialność żuchwy

			#1	#2	#3	#4	#5	#6	#7	#8
Szereg syczący										
/s/	p<0,05			1		9				
	WE	mała		1		9				
		średnia								
		wysoka								
/z/	p<0,05		5	10		6		1		
	WE	mała	5	10		6		1		
		średnia								
		wysoka								
/ts/	p<0,05		7	5		13		1		
	WE	mała	7	5		13		1		
		średnia								
		wysoka								
/dz/	p<0,05					8	4			
	WE	mała				8	4			
		średnia								
		wysoka								

Kontynuacja tabeli na następnej stronie

			#1	#2	#3	#4	#5	#6	#7	#8
Szereg szumiący										
/ʂ/	**p<0,05**					6			3	
	WE	**mała**				6			3	
		średnia								
		wysoka								
/ʐ/	**p<0,05**		1			6				
	WE	**mała**	1			6				
		średnia								
		wysoka								
/tʂ/	**p<0,05**				7				2	
	WE	**mała**			6				2	
		średnia			1					
		wysoka								
/dʐ/	**p<0,05**					1				1
	WE	**mała**				1				1
		średnia								
		wysoka								
Szereg ciszący										
/ɕ/	**p<0,05**				1			9	7	
	WE	**mała**			1			9	7	
		średnia								
		wysoka								
/ʑ/	**p<0,05**					1		14	14	
	WE	**mała**				1		14	14	
		średnia								
		wysoka								
/tɕ/	**p<0,05**		1			2		6	6	
	WE	**mała**	1			2		6	6	
		średnia								
		wysoka								
/dʑ/	**p<0,05**		1					6	7	
	WE	**mała**	1					6	7	
		średnia								
		wysoka								

Tab. B.26: Podsumowanie liczby cech kształtu warg (3D) o wartości p poniżej 0,05 z podziałem na różne stopnie wielkości efektu.

#1 dentalizacja
#2 dentalność
#3 postdentalność
#4 apikalność
#5 skrócenie wędzidełka języka
#6 medialność wypływu powietrza
#7 medialność języka
#8 medialność żuchwy

			#1	#2	#3	#4	#5	#6	#7	#8
Szereg syczący										
/s/	**p<0,05**			4		17	1	1		
	WE	**mała**		4		17	1	1		
		średnia								
		wysoka								
/z/	**p<0,05**		10	10		8	2	1		
	WE	**mała**	10	10		8	2	1		
		średnia								
		wysoka								
/ts/	**p<0,05**		5	4		17	3	2		
	WE	**mała**	5	4		17	3	2		
		średnia								
		wysoka								
/dz/	**p<0,05**			1		10	3			
	WE	**mała**		1		10	3			
		średnia								
		wysoka								
Szereg szumiący										
/ʂ/	**p<0,05**					3			6	
	WE	**mała**				3			6	
		średnia								
		wysoka								
/ʐ/	**p<0,05**					5				
	WE	**mała**				5				
		średnia								
		wysoka								
/tʂ/	**p<0,05**				8				3	
	WE	**mała**			7				3	
		średnia			1					
		wysoka								

Kontynuacja tabeli na następnej stronie

			#1	#2	#3	#4	#5	#6	#7	#8
/dʐ/	p<0,05									
	WE	mała								
		średnia								
		wysoka								
Szereg ciszący										
/ɕ/	p<0,05				1		1	3	1	
	WE	mała			1		1	3	1	
		średnia								
		wysoka								
/ʑ/	p<0,05					2	4	14	14	
	WE	mała				2	3	14	14	
		średnia					1			
		wysoka								
/tɕ/	p<0,05		5			2		7	6	
	WE	mała	5			2		7	6	
		średnia								
		wysoka								
/dʑ/	p<0,05		4					6	7	
	WE	mała	4					6	7	
		średnia								
		wysoka								

Tab. B.27: Podsumowanie liczby cech kształtu języka (3D) o wartości p poniżej 0,05 z podziałem na różne stopnie wielkości efektu.

#1 dentalizacja					#5 skrócenie wędzidełka języka					
#2 dentalność					#6 medialność wypływu powietrza					
#3 postdentalność					#7 medialność języka					
#4 apikalność					#8 medialność żuchwy					
			#1	#2	#3	#4	#5	#6	#7	#8
Szereg syczący										
/s/	p<0,05		20	20		9				
	WE	mała	12	9		9				
		średnia	8	11						
		wysoka								

Kontynuacja tabeli na następnej stronie

			#1	#2	#3	#4	#5	#6	#7	#8
/z/	p<0,05		6	11		1				
	WE	mała	5	7		1				
		średnia	1	4						
		wysoka								
/ts/	p<0,05		13	20		16	4		7	
	WE	mała	12	6		7			1	
		średnia	1	14		9	4		6	
		wysoka								
/dz/	p<0,05		4	16		3			3	1
	WE	mała	4	2		3				
		średnia		14					2	1
		wysoka							1	
Szereg szumiący										
/ʂ/	p<0,05		1				6		1	3
	WE	mała	1				3		1	2
		średnia					3			1
		wysoka								
/ʐ/	p<0,05		2		1	5	7			
	WE	mała	2			5				
		średnia			1		7			
		wysoka								
/tʂ/	p<0,05								1	
	WE	mała								
		średnia							1	
		wysoka								
/dʐ/	p<0,05		8		3					
	WE	mała								
		średnia	8							
		wysoka			3					
Szereg ciszący										
/ɕ/	p<0,05				27		8		5	1
	WE	mała			14		1		5	
		średnia			13		7			1
		wysoka								
/ʑ/	p<0,05					3	1	2	6	4
	WE	mała				3		2	6	
		średnia					1			4
		wysoka								

Kontynuacja tabeli na następnej stronie

			#1	#2	#3	#4	#5	#6	#7	#8
/tɕ/	p<0,05		4					2		
	WE	mała	4					2		
		średnia								
		wysoka								
/dʑ/	p<0,05		2			1				
	WE	mała	2			1				
		średnia								
		wysoka								

Tab. B.28: Podsumowanie liczby cech teksturowych (2D) o wartości p poniżej 0,05 z podziałem na różne stopnie wielkości efektu.

			#1	#2	#3	#4	#5	#6	#7	#8
#1 dentalizacja #2 dentalność #3 postdentalność #4 apikalność			#5 skrócenie wędzidełka języka #6 medialność wypływu powietrza #7 medialność języka #8 medialność żuchwy							
Szereg syczący										
/s/	p<0,05		2	17		5				
	WE	mała	2	17		5				
		średnia								
		wysoka								
/z/	p<0,05		6	22		4	2			
	WE	mała	6	22		4	2			
		średnia								
		wysoka								
/ts/	p<0,05			7		4	1			
	WE	mała		7		4	1			
		średnia								
		wysoka								
/dz/	p<0,05					9	1			
	WE	mała				9	1			
		średnia								
		wysoka								

Kontynuacja tabeli na następnej stronie

			#1	#2	#3	#4	#5	#6	#7	#8
Szereg szumiący										
/ʂ/	**WE**	**p<0,05**	1		6					
		mała	1		6					
		średnia								
		wysoka								
/ʐ/	**WE**	**p<0,05**			20	5	4			
		mała			20	5	3			
		średnia					1			
		wysoka								
/tʂ/	**WE**	**p<0,05**			56	8			4	3
		mała			51	8			4	3
		średnia			5					
		wysoka								
/dʐ/	**WE**	**p<0,05**			27	2				6
		mała			27	2				6
		średnia								
		wysoka								
Szereg ciszący										
/ɕ/	**WE**	**p<0,05**							7	
		mała							7	
		średnia								
		wysoka								
/ʑ/	**WE**	**p<0,05**	2			2			4	4
		mała	2			2			4	4
		średnia								
		wysoka								
/tɕ/	**WE**	**p<0,05**							8	
		mała							8	
		średnia								
		wysoka								
/dʑ/	**WE**	**p<0,05**							27	4
		mała							27	4
		średnia								
		wysoka								

Tab. B.29: Podsumowanie liczby cech teksturowych (3D) o wartości p poniżej 0,05 z podziałem na różne stopnie wielkości efektu.

#1 dentalizacja
#2 dentalność
#3 postdentalność
#4 apikalność
#5 skrócenie wędzidełka języka
#6 medialność wypływu powietrza
#7 medialność języka
#8 medialność żuchwy

			#1	#2	#3	#4	#5	#6	#7	#8
/s/	p<0,05		12	5		3	7			
	WE	mała	12	5		3	7			
		średnia								
		wysoka								
/z/	p<0,05		13	18		4	9	4		2
	WE	mała	13	18		4	8	4		2
		średnia					1			
		wysoka								
/ts/	p<0,05		4	13		4	7	3		2
	WE	mała	4	13		4	7	3		2
		średnia								
		wysoka								
/dz/	p<0,05		2	1		3	2			
	WE	mała	2	1		3	2			
		średnia								
		wysoka								
Szereg szumiący										
/ʂ/	p<0,05		1		1	3			5	1
	WE	mała	1		1	3			5	1
		średnia								
		wysoka								
/ʐ/	p<0,05		3			2	3			1
	WE	mała	3			2	3			1
		średnia								
		wysoka								
/tʂ/	p<0,05		7		22	15			3	
	WE	mała	7		20	15			3	
		średnia			2					
		wysoka								
/dʐ/	p<0,05				5	1	1			1
	WE	mała			5	1	1			1
		średnia								
		wysoka								

Kontynuacja tabeli na następnej stronie

			#1	#2	#3	#4	#5	#6	#7	#8
Szereg ciszący										
/ɕ/	**p<0,05**		2		5		1		17	3
	WE	**mała**	2		5		1		17	3
		średnia								
		wysoka								
/ʑ/	**p<0,05**		1				1	9	15	
	WE	**mała**	1				1	9	15	
		średnia								
		wysoka								
/tɕ/	**p<0,05**		2				3	1	8	
	WE	**mała**	2				3	1	8	
		średnia								
		wysoka								
/dʑ/	**p<0,05**						1		16	
	WE	**mała**					1		16	
		średnia								
		wysoka								

Tab. B.30: Podsumowanie liczby cech akustycznych o wartości p poniżej 0,05 z podziałem na różne stopnie wielkości efektu.

	#1 dentalizacja				#5 skrócenie wędzidełka języka					
	#2 dentalność				#6 medialność wypływu powietrza					
	#3 postdentalność				#7 medialność języka					
	#4 apikalność				#8 medialność żuchwy					
			#1	#2	#3	#4	#5	#6	#7	#8
Szereg syczący										
/s/	**p<0,05**		4	11		9	7	8	3	1
	WE	**mała**	4	11		9	7	8	3	1
		średnia								
		wysoka								
/z/	**p<0,05**		6	14		8		2		
	WE	**mała**	6	14		8		2		
		średnia								
		wysoka								

Kontynuacja tabeli na następnej stronie

			#1	#2	#3	#4	#5	#6	#7	#8
/ts/	p<0,05		5	13		15	6	7	6	1
	WE	mała	5	13		15	6	7	6	1
		średnia								
		wysoka								
/dz/	p<0,05		9	11		7	5	3	7	2
	WE	mała	9	11		7	5	3	7	1
		średnia								1
		wysoka								
Szereg szumiący										
/ʂ/	p<0,05		1		40	34	1		1	1
	WE	mała	1		15	34	1		1	1
		średnia			18					
		wysoka			8					
/ʐ/	p<0,05		3		39	24	7			1
	WE	mała	3		17	24	7			1
		średnia			19					
		wysoka			3					
/tʂ/	p<0,05		2		34	9				1
	WE	mała	2		16	9				1
		średnia			14					
		wysoka			4					
/dʐ/	p<0,05		3		13		26			7
	WE	mała	3		12		14			7
		średnia			1		12			
		wysoka								
Szereg ciszący										
/ɕ/	p<0,05		1		4	5			1	9
	WE	mała	1		4	5			1	8
		średnia								1
		wysoka								
/ʑ/	p<0,05		3			3	17	2	4	6
	WE	mała	3			3	16	2	4	6
		średnia					2			
		wysoka								
/tɕ/	p<0,05		4			1		5	6	19
	WE	mała	4			1		5	6	18
		średnia								1
		wysoka								

Kontynuacja tabeli na następnej stronie

			#1	#2	#3	#4	#5	#6	#7	#8
/dʑ/	**p<0,05**		15			12	4	7	6	12
	WE	**mała**	15			12	4	7	6	12
		średnia								
		wysoka								

Streszczenie

Niniejsza rozprawa podejmuje temat komputerowego wsparcia diagnostyki logopedycznej z wykorzystaniem metod sztucznej inteligencji. Opisane badania skupiają się na weryfikacji istnienia relacji pomiędzy normatywnymi i nienormatywnymi cechami głosek dentalizowanych a cechami wizualnymi obrazów i akustycznymi sygnałów prezentujących mowę dzieci w wieku przedszkolnym. Zaproponowano wieloetapową strukturę badań, która obejmowała: rejestrację bazy danych, opracowanie metodyki równoległego przetwarzania danych obrazowych i sygnałów akustycznych oraz przeprowadzenie testów statystycznych, które sprawdzały relacje wizualno-akustyczno-artykulacyjne. Badania bazowały na danych zarejestrowanych podczas sesji pomiarowych w przedszkolach. Materiał obejmował nagrania wideo i sygnał akustyczny mowy dzieci oraz opis logopedyczny przygotowany przez specjalistów terapii mowy.

Kolejny etap dotyczył przygotowania dwugałęziowej metody przetwarzania nagrań, osobno dla obrazów i dźwięku. W ramach ścieżki dotyczącej materiału wideo opracowano dwuetapową metodę segmentacji artykulatorów (warg, ust, zębów, języka): detekcji obiektów za pomocą sieci YOLO i ich segmentacji z wykorzystaniem modelu DeepLabv3+. Do wstępnego uczenia sieci zaproponowano wykorzystanie niedokładnych etykiet uzyskanych za pomocą metody zbioru poziomic i obrazów rozmytych. W kolejnym kroku model został dostrojony za pomocą mniejszego zbioru obrazów z obrysami eksperckimi. Równolegle przetwarzano sygnał akustyczny tego samego fragmentu nagrania.

Na podstawie segmentacji wyekstrahowano parametry obrazowe: dwu oraz trójwymiarowe. Zestaw parametrów obrazowych obejmował cechy teksturowe i parametry kształtu ust, warg i języka. W przypadku sygnałów akustycznych dla każdej ramki wyodrębniono zestaw cech, który obejmował cechy w dziedzinie czasu, cechy częstotliwościowe w pełnym pasmie i w pasmie szumu.

Wyniki analizy wykazały istnienie różnic między normatywnymi i patologicznymi realizacjami głosek w kontekście wybranych cech artykulacyjnych. Rezultaty sugerują największą użyteczność cech obrazowych bazujących na wolumenach, głównie opisujących kształt języka. Metodyka opisana w pracy oraz zaprezentowane rezultaty mogą stanowić punkt wyjścia do rozwoju systemów eksperckich wspierających diagnostykę logopedyczną sygmatyzmu.

Słowa kluczowe: inżynieria biomedyczna, komputerowe wspomaganie diagnostyki logopedycznej, sygmatyzm, detekcja i segmentacja artykulatorów, uczenie głębokie

Abstract

This doctoral thesis focuses on processing video data and speech signal in computer-aided speech diagnosis using artificial intelligence. The main objective was to verify relationships between normal and disordered sibilant pronunciation features and the visual and audio features based on data presenting the speech of preschool children. The proposed multi-stage study included dataset registration, development of methods for processing acoustic signals and image data, and statistical analysis to verify visual-acoustic-articulation relations. The study involved data recorded in kindergartens, including video recordings and acoustic signals of child speech, both annotated by speech therapy experts.

In the following stage, a double-branch data processing method was proposed separately for images and sounds. The video processing path included a two-stage framework for segmenting articulators (lips, mouth, teeth, tongue): object detection using the YOLO model and segmentation with a DeepLabv3+ model. The model was initially trained using weak labels produced by rough segmentation based on the distance-regularized level set evolution over fuzzified images. Next, the model was fine-tuned using a portion of manual ground-truth delineations. The audio processing branch worked in parallel on the corresponding data.

A set of visual and acoustic features was then extracted. Image parameters included oral-region textural features and shape features of the mouth, lips, and tongue. They were determined based on 2D images and 3D volumes, with time as the third dimension. The acoustic features included time-domain features, full-band, and noise-band frequency features.

The results indicated differences between normal and disordered realizations of sibilants in the context of selected articulation features. The experimental results suggest that image features based on 3D volumes are most useful, mainly the parameters describing the tongue shape. The methodology described in this dissertation and the presented results can be a starting point for developing expert systems supporting speech diagnosis and therapy in sigmatism.

Keywords: biomedical engineering, computer-aided speech diagnosis, sigmatism, detection and segmentation of articulators, deep learning